Contributions à l'inférence statistique de modèles complexes

Sophie Donnet

Université Paris-Sud

École doctorale de mathématiques Hadamard (ED 574)
MIA Paris (UMR MIA518, INRA/AgroParisTech)

Mémoire présenté pour l'obtention du

Diplôme d'habilitation à diriger les recherches

Discipline : Mathématiques

par

Sophie DONNET

Contributions à l'inférence statistique de modèles complexes

Rapporteurs :
Gilles CELEUX
Susanne DITLEVSEN
Franck PICARD

Date de soutenance : 14 juin 2018

Composition du jury :

Gilles CELEUX	(Rapporteur)
Christophe GIRAUD	(Examinateur)
Sylvie HUET	(Présidente du jury)
Jean-Michel MARIN	(Examinateur)
Nathalie PEYRARD	(Examinatrice)
Franck PICARD	(Rapporteur)
Christian P. ROBERT	(Examinateur)

Remerciements

Je remercie sincèrement Susanne Ditlevsen, Gilles Celeux et Franck Picard pour leurs rapports et l'intérêt qu'ils ont porté à mon travail. Je sais que votre temps est précieux et j'en suis d'autant plus honorée. Un merci spécial à Gilles qui a rendu son rapport dans des conditions chaotiques: je mesure ton sens du devoir.

Merci également à Gilles et Franck pour le temps que vous consacrez à mon jury.

Je remercie Sylvie Huet d'avoir accepté de participer à mon jury, je suis heureuse d'avoir pu soutenir cette habilitation avant ton départ pour d'autres aventures.

Je remercie Christophe Giraud d'avoir accepté un n-ième jury d'HDR. Je te sais très sollicité et je te remercie d'avoir répondu à mon invitation avec autant d'enthousiasme.

Je remercie Nathalie Peyrard d'avoir aussi accepté avec enthousiasme et d'avoir fait le déplacement. J'espère que nous pourrons poursuivre les discussions autour des modèles graphiques et des réseaux dans d'autres occasions.

Je remercie Jean-Michel Marin et Christian Robert de leur présence aujourd'hui. Par largement, d'abord à Jean-Michel, merci de m'avoir poussée vers le CEREMADE lors de mon recrutement comme maître de conférences. C'était définitivement une très bonne décision. Et merci à Christian pour ton accueil au CEREMADE. J'y ai passé des années enrichissantes autant scientifiquement qu'humainement. Ta bienveillance m'a permis de m'y sentir bien. Tes emails lapidaires d'une ligne me manquent un peu moins que ton clafoutis aux cerises! Jean-Michel et Christian, notre épopée mexicaine reste un joyeux souvenir.

Un grand merci à Marc Lavielle et Jean-Baptiste Poline par qui cette histoire a commencé.

Parmi mes collaborateurs, j'ai une pensée toute particulière pour Adeline Samson. Depuis 15 ans maintenant, nous poursuivons nos collaborations, de façon plus ou moins intensive mais avec un plaisir toujours renouvelé. Nos réunions épisodiques autour d'un problème statistique et d'une tasse de thé sont des moments précieux pour moi. Je suis très fière de cette amitié personnelle et professionnelle. Pour tout ça et tout le reste, merci.

L'autre collaboration qui me tient beaucoup à coeur est celle que je poursuis avec Avner Bar-Hen et Pierre Barbillon. Pas tout à fait aussi ancienne que ma collaboration avec Adeline, nous affichons tout de même un joli nombre d'heures au compteur ainsi qu'un nombre indécent d'éclats de rire. J'espère que nous continuerons longtemps à travailler et voyager ensemble.

Durant mes années au CEREMADE, j'ai rencontré des collègues et collaborateurs formidablement brillants et chaleureux en la personne de Judith Rousseau et Vincent Rivoirard. Judith, tu restes ma meilleure camarade de chambre d'hôtel! Vincent, merci pour tes encouragements, ton enthousiasme et ta gentillesse. J'ai adoré nos collaborations et voyages à vos côtés! Merci aussi à Bénédicte Haas, Robin Ryder et François Bolley pour nos "bières CEREMADE". Le CEREMADE ne serait pas totalement le CEREMADE sans Isabelle Bélier. Merci pour ton

travail d'abord, je mesure combien ton efficacité et ta disponibilité sont précieuses, mais aussi merci pour ton amitié. Enfin, last but not least, merci au Docteur Alessandra Iaccobucci!!!!! Ces années dans ton bureau ont été de très très belles années. Nous nous voyons moins mais avec toujours autant de plaisir pour ma part. Merci enfin à l'ensemble des collègues que j'ai pu croiser à l'université Paris-Dauphine. J'ai toujours pris beaucoup de plaisir à y travailler, à la fois en recherche et en enseignement.

Mes deux années au Mexique m'ont permis de prendre du recul sur mon métier, sur ce que je voulais faire et pour cela je remercie Andres Christen. Gracias por tu entusiasmo, tus palabras siempre positivas. Cuando el humor parisino me toca, trato de recordarme el sol del CIMAT.

Mon arrivée à l'INRA doit beaucoup au workshop "Statistiques au sommet de Rochebrune" organisé par Eric Parent. J'y ai découvert des applications toutes plus motivantes les unes que les autres, le tout dans une ambiance unique. Pour ça mais aussi pour nos collaborations en cours, j'adresse mes sincères remerciements à Eric. Parmi les personnes présentes que j'ai appris à connaître à Rochebrune, Marie-Pierre Etienne est l'une de celles qui m'a le plus marquée. J'admire ta façon d'appréhender les problèmes, la façon que tu as de t'approprier une question appliquée et d'en faire un problème de mathématiques intéressant. Pour tout ça merci!

Je remercie chaleureusement mes collaborateurs de MIA Paris: Stéphane Robin, Julien Chiquet pour les nombreux échanges passionnants et enrichissants, Tristan Mary-Huard et Pierre Gloaguen pour notre magnifique voyage, Gabriel Lang entre autre pour m'avoir encouragée à soutenir cette habilitation. Merci à Liliane Bel notre directrice d'unité et à Céline Lévy-Leduc sans qui l'unité ne pourrait pas fonctionner, je mesure le sacrifice que cela représente. Merci à Irène Desecures pour la gestion toujours efficace de mes missions. Un merci particulier à mes collègues de bureau (Laure Delattre, Jessica Tressou, Séverine Bord et Isabelle Albert) pour leur présence et leur soutien au jour le jour. Enfin à tous les autres qui font de MIA Paris un endroit si particulier dans le paysage des laboratoires de mathématiques: merci!

À ma famille,
À Alexis, Ana, Violette et Coline

Liste des publications

Articles scientifiques

Submitted papers

[S1] S. Donnet, S. Robin. *Shortened Bridge Sampler: Using deterministic approximations to accelerate SMC for posterior sampling*

Published papers

[A16] E. Lazega, A. Bar-Hen, P. Barbillon, S. Donnet. *Effects of competition on collective learning in advice networks.* Social Networks, 2016; 47:1–14

[A15] P. Barbillon, S. Donnet, E. Lazega, A. Bar-Hen. *Multiplex stochastic block model for social networks.*Journal of the Royal Statistical Society: Series A (Statistics in Society), 2017; 180(1)

[A14] S. Donnet, V. Rivoirard, J. Rousseau, C. Scricciolo. *Posterior concentration rates for empirical Bayes procedures, with applications to Dirichlet Process mixtures.* Bernoulli 2018, Vol. 24, No. 1, 231-256

[A13] S. Donnet, V. Rivoirard, J. Rousseau, C. Scricciolo. *Posterior concentration rates for counting processes with Aalen multiplicative intensities.* Bayesian Analysis, march 2017; 12(1),53–87

[A12] M. Capistran, A. Christen, S. Donnet. *Bayesian Analysis of ODE's: solver optimal accuracy and Bayes factors.* SIAM/ASA Journal on Uncertainty Quantification, 2016; 4(1), 829-849

[A11] S. Donnet, J. Rousseau, *Bayesian Inference for Partially Observed Branching Processes.* Bayesian analysis, march 2016; 11(1), 151–190

[A10] S. Donnet, R. Bartolo., J.M. Fernandes , J.P. Cunha, L. Prado and H. Merchant *Monkeys time their movement pauses and not their movement kinematics during a synchronization-continuation rhythmic task.* Journal of Neurophysiology, May 2014; 111(10), 2138

[A9] S. Donnet, A. Samson. *Using PMCMC in EM algorithm for stochastic mixed models: theoretical and practical issues.* Journal de la Société Française de Statistique, 155, 49-72, 2014.

[A8] S. Donnet, A. Samson. *A review on estimation of stochastic differential equations for pharmacokinetic/pharmacodynamic models.* Advanced Drug Delivery Reviews, 2013 Jun 30;65(7):929-39

[A7] I. Albert, S. Donnet, C. Guihenneuc, S. Low-Choy, K. Mengersen, J. Rousseau, *Combining expert opinions in prior elicitation (with discussion).* Bayesian Analysis, 7(3), 503-546 (2012)

[A6] S. Donnet, J.-M. Marin . *An empirical Bayes procedure for the selection of Gaussian graphical models.* Statistics and Computing, 22(5), 1113-1123 (2012)

[A5] S. Donnet , J-L Foulley, A. Samson. *Bayesian analysis of growth curves using mixed models defined by stochastic differential equations* . Biometrics, 66(3) :733-741, (2010)

[A4] P. Ciuciu, T. Vincent, L. Risser, S. Donnet. *A joint detection-estimation framework for analysing within-subject fMRI data.* Journal de la Société Française de Statistiques, Vol. 151, No 1 (2010)

[A3] S. Donnet, A. Samson. *Parametric inference for mixed models defined by stochastic differential equations.* ESAIM P&S, 12:196-218, (2008)

[A2] S. Donnet, A. Samson. *Estimation of parameters in missing data models defined by differential equations.* J. Statist. Plann. Inference (2007), 137(9), 2815–2831

[A1] S. Donnet, M. Lavielle, and J.-B. Poline. *Are fMRI event related reponses constant across events?.* Neuroimage,Volume 31, Issue 3, 1 (July 2006), 1169 – 1176

Book reviews and discussion

[B3] S. Donnet. Book review of *"Stochastic Modelling for Systems Biology (second edition)"* by Darren J. Wilkinson. CHANCE 25-4 (Décembre 2012)

[B2] S. Donnet. Book review of *"Statistical Thinking in Epidemiology"* by Yu-Kang Tu and Mark S. Gilthorpe. CHANCE 25-4 (Décembre 2012)

[B1] S. Donnet. Book review of *"Monte Carlo Simulation for the Pharmaceutical Industry: Concepts, Algorithms, and Case Studies"* by Mark Chang. International Statistical Review (Avril 2012)

[D1] S. Donnet, A. Samson. Discussion on *"Parameter estimation for differential equations: a generalized smoothing approach"* (by Ramsay JO, Hooker G, Campbell D and Cao J), Journal of the Royal Statistical Society: Series B, 69(5):741-796, (2007)

Proceedings

[P6] A. Bar-Hen , P. Barbillon, S. Donnet, *Modèles à blocs latents pour graphe multipartite. Application aux interactions entre espèces animales et plantes* Paru dans les actes de conférences des 49émes Journées de Statistique, SFdS, Avignon, France (juin 2017)

[P5] Chiquet, J., Donnet, S., Etienne, M. P., Samson, A. *Méthode conjointe de segmentation-classification pour des modèles d'écologie du déplacement* Paru dans les actes de conférences des 49èmes Journées de Statistique, SFdS, Avignon, France (juin 2017)

[P4] S. Donnet , J-L Foulley, A. Samson. *Analyse bayésienne de courbes de croissance par des modéles à effets mixtes définis par équations différentielles stochastiques.* Paru dans les actes de conférences des 41èmes Journées de Statistique, SFdS, Bordeaux, Bordeaux, France (juin 2009)

[P3] S. Donnet et A. Samson. *Estimation paramétrique d'un processus de diffusion à partir d'observations bruitées et à temps discrets.* Paru dans les actes de conférence des 38èmes journées de Statistique de la SFDS (juin 2006).

[P2] S. Donnet. *Inversion de données IRMf. Estimation et sélection de modéles.* Paru dans les actes de conférence des 37èmes journées de Statistique de la SFDS en (juin 2005).

[P1] S. Donnet, M. Lavielle, P. Ciuciu, and J.-B. Poline. *Selection of temporal models for event-related fMRI.* In Proc. 2th Proc. IEEE ISBI, Arlington, VA, pages 992–995, (Apr. 2004)

Contents

Introduction

This document is an overview of my research work. My PhD thesis took place in the field of statistics with an application to neurosciences. More precisely, I developed statistical model for hemodynamic response from functional magnetic resonance imaging (fMRI). In order to consider physiological dynamical models (as opposed to a non-parametric modeling of the hemodynamic response), I focused on statistical models relying on a system of differential equations with no explicit solution. To estimate the parameters of such models, I resorted to a stochastic version of the well-known Expectation Maximization (EM) algorithm, coupled to Monte Carlo Markov Chain (MCMC) algorithms (SAEM-MCMC algorithm).

After my PhD, as a Maître de Conférences at University Paris-Dauphine, I expanded my experience and knowledge in stochastic algorithms –essentially in a Bayesian inference context– to various fields of applications. Thus far, my research production mainly divides into three topics:

- Statistical inference for models defined by (stochastic) differential equations,
- Parametric and Non-parametric Bayesian inference for counting processes,
- Statistical inference for network data with application to sociology and ecology.

My various contributions to these three domains will be described hereafter, in dedicated chapters 1, 2 and 3. Note that other works, not directly linked to these three topics, are not evoked here but appear in the publications list.

Chapter 1 deals with models defined by ordinary or stochastic differential equations systems (ODEs and SDEs). Complex biological phenomena such as brain activity, drug action on bodies or glucose regulation system, etc., are usually modeled by complex systems of differential equations, dynamically linking the various observed or non-observed biological entities at stake. Except in very rare cases, it is impossible to exhibit an explicit solution of the differential system. Consequently, in practice, the exact solution of the system is approximated by a numerical scheme (Euler or Runge-Kutta being the most famous) and the statistical inference (estimation or model selection...) is performed using this approximation. In general, the replacement of the true solution by its approximation is ignored when studying the properties of the parameters estimation.

In a first work with Adeline Samson during my PhD thesis [**A2**], we consider the maximum likelihood estimation of models whose regression function is the non-explicit solution of an ODE, depending on non-observed latent variables. In this context, we maximize the likelihood using the SAEM-MCMC algorithm. Besides, we bound the error on the estimations introduced by the use of an approximate solution to the dynamical system. The obtained bound on the error is a function of the precision of the ODE numerical approximation scheme.

In a second work with Marcos Capistran and Andres J. Christen [**A**12] (during my two years stay in Mexico), we consider a similar problem but in a Bayesian context. In this work, we propose to control the error –still introduced by the numerical approximation scheme to solve the ODE– on the posterior distributions of the parameters. We also study the gain of precision when using a better approximation scheme in terms of Bayes Factor.

Ordinary differential system often fails to model completely the biological phenomena at stake. In order to take into account more variability in the process, random components may be introduced in the differential system, thus resulting into a system of stochastic differential equations (SDEs). These SDEs generally have no explicit solutions, and once again, a numerical approximation scheme is used to approximate the solution. In a second work with Adeline Samson during my PhD thesis [**A**3], we extend our work on models defined by ODEs to those defined by SDEs. More precisely, approximating the SDE solution but the Euler-Maruyama scheme, we develop an adapted SAEM-MCMC algorithm to maximize the likelihood. We prove the convergence of this algorithm and bound the error introduced by the Euler-Maruyama scheme on the parameters estimations.

In a joint work with Adeline Samson and Jean-Louis Foulley [**A**5], we consider the Bayesian analysis of growth curves using mixed models defined by stochastic differential equations. In this particular case, the solution of the SDE is explicit, and we illustrate the practical interest of SDEs over ODEs.

When the SDE has an explicit solution but impossible to simulate conditionally to the observations, we propose to take advantage of the new MCMC algorithms adapted to the filtering context (namely the Particle MCMC algorithm) and to combine them with the SAEM algorithm. This new method (a joint work with Adeline Samson [**A**9]) is presented in section 2.2.2 of Chapter 1.

Finally, I present my last project on models defined by SDEs. In a joint work with Julien Chiquet, Marie-Pierre Etienne and Adeline Samson [**P**5], we focus on SDE models for movement ecology (i.e. modeling a animals trajectories). We aim at identifying –from the observed trajectories– several types of activities (move from one point to an other, hunt...). In this particular context, we resort to change-point detection tools, associated to regularization methods.

Chapter 2 introduces my various works on counting processes.

My first work [**A**11] (joint work with Judith Rousseau) on counting processes is motivated by an application in reliability. More precisely, the aim is to model the failure and replacement of components of an electrical network. The model we propose is a counting process with an endogenous evolution of its intensity process. Inference methods for such processes exist when the trajectories are continuously fully observed. In this work, we deal with the case of a partially observed process: we assume that we observe the breakdowns instants but not the types of the breakdowns. Moreover, we consider the case where the initial state of the process at the beginning of the observation period is unknown. The Bayesian inference being strongly influenced by this quantity, we propose a sensible prior distribution on the initial state, using the probabilistic properties of the process. We prove the parameters identifiability and we illustrate the performances of our methodology on a large simulation study.

The process studied in this first paper belongs to the family of counting processes with multiplicative intensity. In this particular case, the intensity has a parametric form. In two other

papers (joint work with Judith Rousseau, Vincent Rivoirard and Catia Scricciolo [**A**14] [**A**13]) we consider the Bayesian non-parametric estimation of intensity function for Aalen Multiplicative Intensities. These two papers include two main parts, a theoretical one and a computational one. From a theoretical point of view, we focus on the concentration properties of the posterior distributions in the non-parametric framework. [**A**11] is dedicated to the multiplicatives Aalen processes. In [**A**14], we concentrate our efforts on the posterior concentration in the case of Empirical Byes procedure, namely when the prior distribution is data-dependent. Aalen processes are a special model from the ones studied in this paper. From a computational point of view, we enhance the papers with simulations studies. The non-parametric context requires the design of special sampling algorithms.

Finally, the last section Chapter 2 presents our last collaboration with Judith Rousseau and Vincent Rivoirard on Hawkes processes. Multivariate Hawkes processes are dependent counting processes, where the probability of occurrence of an event on a given process depends on the past of all the processes. They are used in neurosciences, to model the dependence between neurones. The dependence is written thanks to a linear filter. In our work, we consider the non-parametric estimation of the intensity functions, both from a theoretical and algorithmic point of view.

Chapter 3 gathers my works on the modeling of (social) networks. This topic is the object of a long time collaboration with Avner Bar-Hen and Pierre Barbillon. The two first papers [**A**15] [**A**16] are motivated by social sciences and are written with Emmanuel Lazega. Our last and our future works take place in the ecology (or eco-sociology) field. The inference of the probabilistic models involved in these papers are challenging from a computational point of view. This statement led to a methodological paper with Stéphane Robin [**S**1].

Modeling relationships between individuals is a classical question in social sciences and clustering individuals according to the observed patterns of interactions allows us to uncover a latent structure in the data. The stochastic block model is a popular approach for grouping individuals with respect to their social behavior. When several relationships of various types can occur jointly between individuals, the data are represented by multiplex networks where more than one edge can exist between the nodes. In [**A**15] and [**A**16], we extend stochastic block models to multiplex networks to obtain a clustering based on more than one kind of relation-ships. We propose to estimate the parameters – such as the marginal probabilities of assignment to groups (blocks) and the matrix of probabilities of connections between groups– through a variational Expectation Maximization procedure. The number of groups is chosen by maximizing a penalized likelihood criterion. This methodology is applied to a network of French cancer researchers.

Stochastic block models include latent random variables, making their likelihood intractable. In particular, processing large networks is computationally challenging. When talking about Bayesian inference, standard stochastic algorithms (such as MCMC or population Monte Carlo algorithms) reach their limit. To tackle this issue, deterministic approximations of the posterior distribution have been proposed in the literature (among them Variational Bayes is well suited to SBM and LBM). However, there is no theoretical guaranty. Moreover, it can be illustrated on examples, that such algorithms can underestimate the posterior variance. In a joint work with S. Robin [**S**1], we propose an algorithm taking advantage of the last development in Sequential Monte Carlo algorithms and deterministic approximations of the posterior distribution.

My most recent works are issued from ecological problematics. Aiming at inscribing my work in an ecological framework, I started working with W. Dattilo (ecologist at INECOL, Mexique)

on multipartite ecological networks. In a few words, a high number of interaction types between plants and animal species co-exist within the natural environment. Among them, we can think about mutualistic plants/animals interactions such as herbivory, protection of plants by ants, pollination, or seed dispersal by birds. These various interactions play a key role in structuring biodiversity. In the recent years, network tools have been intensively used to understand the structure of the ecological interaction networks. However, in most of the published papers, each type of interaction is considered individually, ignoring the other interactions. Few works have considered the joint study of several interactions. Among the various tools dedicated to the study of bipartite graph, Latent Blocks Models (LBM) provide a probabilistic framework for the simultaneous clustering of rows and columns of a matrix. In our last joint work with Pierre Barbillon and Avner Bar-Hen [**A**9], we consider the extension of LBM to the case of multipartite graphes. We develop an adapted estimation algorithm and provide a model selection criterion.

From a general point of view, I am now interested in the statistical inference of complex network datasets. In a few words, I aim at clustering entities from the observation of several non-independent networks. Such models derive from ecological or sociological concrete questions. These perspectives are described at the end of Chapter 3

My research perspectives with respect to the three topics of this manuscript are given at the end of each chapter.

Introduction (version française)

Ce document est un résumé de mes travaux de recherche. Ma thèse de doctorat s'inscrivait dans le champ de la statistique avec une application principale en neurosciences. Plus précisément, je me suis intéressée à la modélisation et à l'estimation de la réponse cérébrale hémodynamique à partir de données d'imagerie par résonance magnétique fonctionnelle (IRMf). Afin de prendre en compte un modèle physiologique de réponse hémodynamique (par opposition à une modélisation non-paramétrique), j'ai orienté mon travail vers les modèles statistiques définis à partir de systèmes d'équations différentielles sans solution explicite. Les paramètres de ces modèles sont estimés par une version stochastique de lalgorithme Expectation Maximization (EM) couplée avec des algorithmes de Monte Carlo par Chaîne de Markov (MCMC) (SAEM-MCMC).

Après mon doctorat, en tant que Maître de Conférences à l'Université Paris-Dauphine, j'ai enrichi mon expérience dans le domaine des algorithmes stochastiques (en particulier dans le contexte de l'inférence bayésienne) dans le cadre de domaines d'applications très variés. À ce stade, mes travaux se divisent principalement en trois thèmes :

- Inférence statistiques pour des modèles définis par équations différentielles (ordinaires ou stochastiques),
- Inférence bayésienne paramétrique et non paramétrique pour quelques processus de comptage.
- Inférence statistique de données de réseaux avec application en sociologie et en écologie.

Mes diverses contributions dans ces trois domaines sont décrites dans les Chapitres 1, 2 et 3. D'autres travaux ne se rattachant pas directement à ces trois thèmes ne sont pas décrits dans ce manuscrit mais sont référencés dans la liste des publications.

Le chapitre 1 est dédié aux modèles définis par équations différentielles ordinaires (EDO) et stochastiques (EDS). Les phénomènes biologiques complexes tels que l'activité cérébrale, l'action de médicaments ou le système de régulation du glucose sont très souvent décrit pas des systèmes d'équations différentielles, ces systèmes liant dynamiquement les différentes quantités biologiques observées et non observées. Exception faite de très rares cas, il est en général impossible d'exhiber une solution explicite du système et la solution est approchée par un schéma d'approximation numériques (Euler ou Runge-Kutta étant les plus connus). L'inférence statistique est alors faite sur ce modèle approché. En général, la substitution de la vraie solution par sa version approchée n'est pas prise en compte dans l'étude des propriétés des estimateurs des paramètres.

Dans mon premier travail de thèse avec Adeline Samson [**A**2], nous nous sommes intéressées à l'estimation par maximum de vraisemblance pour des modèles dans lesquels la fonction de régression est la solution non explicite d'un système d'EDO et dépend de variables aléatoires latentes. Nous maximisons la vraisemblance par un algorithme du type SAEM-MCMC. De plus,

nous bornons les erreurs d'estimation dues à l'utilisation d'une solution approchée de l'EDO. La borne obtenue est une fonction de la précision du schéma de résolution numériques de l'EDO.

Dans un second travail avec Marcos Capistran et Andres J. Christen (durant mon séjour de deux ans au Mexique) [**A**12], nous avons considéré un problème similaire, mais dans un cadre bayésien. Dans ce travail, nous contrôlons l'erreur (toujours introduite par l'utilisation d'une méthode numérique d'approximation) sur la loi a posteriori des paramètres. Nous nous intéressons aussi au gain de précision obtenu par l'utilisation d'un meilleur schéma de résolution en termes de facteurs de Bayes.

En général, les EDO ne permettent par de modéliser de façon satisfaisante les phénomènes biologiques à l'étude. Afin de prendre en compte une plus grande variabilité dans le processus, des composantes aléatoires peuvent être introduites dans le système différentiel, aboutissant ainsi à un système d'équations différentielles stochastiques. Comme précédemment, ces EDS n'ont, en général, pas de solution explicite et leurs solutions sont approchées par un schéma numérique du type Euler-Maruyama. Dans notre deuxième travail de thèse en commun avec Adeline Samson [**A**3], nous avons étendu nos résultats sur les modèles définis par EDO aux modèles définis par EDS. Plus précisément, après avoir approché la solution de l'EDS par un schéma d'Euler-Maruyama, nous maximisons la vraisemblance par un algorithme SAEM-MCMC. Nous étudions la convergence de l'algorithme et bornons l'erreur sur les paramètres d'estimations en fonction de l'erreur induite par l'utilisation du schéma d'Euler-Maruyama.

Dans un travail en collaboration avec Adeline Samson et Jean-Louis Foulley [**A**5], nous nous intéressons à l'analyse bayésienne de courbes de croissances, modélisées par un modèle d'EDS. Dans ce cas particulier, la solution de l'EDS est explicite et nous mettons en évidence l'intérêt de l'EDS par rapport à l'EDO.

Dans [**A**9], travail en collaboration avec Adeline Samson, nous nous intéressons au cas où l'EDS a une solution explicite qui ne peut être facilement simulée conditionnellement aux observations. Nous proposons alors d'avoir recours aux algorithmes MCMC spécialement construits pour des contextes de filtrage (l'algorithme MCMC particulaire) et de les combiner avec l'algorithme SAEM.

Le chapitre 1 se clôt sur mon dernier projet de recherche faisant intervenir des EDS. Dans ce projet, en collaboration avec Julien Chiquet, Marie-Pierre Etienne et Adeline Samson [**P**5], nous nous intéressons aux modèles spatio-temporels définis par EDS et utilisés en écologie pour modéliser les déplacements d'animaux. Dans ce cadre, nous cherchons à identifier différentes activités des animaux (chasse, déplacement d'un point à un autre...) à partir de l'observation des trajectoires d'animaux. Nous combinons des outils de détection de rupture avec des méthodes de régularisation.

Le chapitre 2 regroupe mes travaux sur les processus de comptage.

Mon premier travail dans ce domaine (co-écrit avec Judith Rousseau) a été motivé par une application en fiabilité. Nous modélisons l'évolution d'un réseau électrique, c'est-à-dire la succession des occurrences de pannes et réparations. Le modèle que nous proposons est un processus de comptage dont le processus d'intensité est à évolution endogène. L'estimation bayésienne d'un tel processus est immédiate lorsque le processus est observé de façon continue. Dans notre application, nous n'avons accès qu'à une observation partielle du phénomène : plus précisément, nous connaissons les instants de pannes mais pas le type de réparation. De plus, l'état du système

n'est pas connu au début de la période d'observation. L'inférence bayésienne étant très fortement influencée par ce paramètre, nous construisons une loi a priori informative, issue d'une étude théorique fine du processus. Nous montrons l'identifiabilité du modèle et illustrons l'efficacité de notre procédure sur des données simulées.

Le processus étudié dans ce premier article appartient à la famille des processus de comptage à intensité multiplicative. Dans ce premier article, nous avions considéré une forme paramétrique de l'intensité du processus. Dans les deux articles suivants [**A**14] [**A**13] (collaborations avec Judith Rousseau, Vincent Rivoirard and Catia Scricciolo) nous nous intéressons à l'estimation bayésienne non-paramétrique de fonctions d'intensité pour les processus de Aalen multiplicatifs. Ces articles comportent deux volets, un théorique et un numérique. D'un point de vue théorique, nous nous intéressons aux propriétés de concentration des lois a posteriori dans un cadre d'inférence bayésienne non-paramétrique. Dans [**A**13], nous traitons le cas des processus de Aalen. [**A**14] s'intéresse à la convergence de la loi a posteriori des estimateurs non-paramétriques dans le cas particulier où la loi a priori est dépendante des observations (bayésien empirique). Les processus de Aalen sont un des modèles étudiés dans ce article. D'un point de vue numérique, chaque article est doté d'une étude sur données simulées. L'inférence non-paramétrique bayésienne des processus que nous considérons dans ces articles requiert la construction d'algorithmes ad hoc d'échantillonnage de la loi a posteriori.

Enfin, la dernière partie du chapitre 2 décrit notre dernière collaboration avec Judith Rousseau et Vincent Rivoirard sur les processus de Hawkes multivariés. Ces processus sont classiquement utilisés en neurosciences pour modéliser la dépendance entre potentiels d'actions neuronaux. Ce sont des processus de comptage pour lesquels le processus d'intensité d'un processus donné dépend de la réalisation passée de tous les autres processus. Cette dépendance s'écrit au travers d'un filtre linéaire. Nous nous intéressons à l'estimation bayésienne non-paramétrique de fonctions d'interactions à la fois d'un point de vue théorique et algorithmique.

Mes travaux sur l'analyse de données de réseaux sont décrits dans le chapitre 3. Ce sujet est au cœur d'une collaboration de longue date avec Avner Bar-Hen and Pierre Barbillon. Les deux premiers articles [**A**15] [**A**16] ont été motivés par une application en sciences sociales et ont été écrits avec Emmanuel Lazega. Notre dernier travail [**P**6] et nos perspectives sont motivés par la modélisation de réseaux écologiques ou socio-écologiques. L'inférence des modèles probabilistes que nous utilisons est un défi computationnel, ce qui m'a conduit à un travail plus méthodologique avec Stéphane Robin [**S**1].

Modéliser les relations entre individus est une problématique classique en sciences sociales et regrouper les individus en fonction des motifs observés dans le réseau permet de comprendre la topologie du réseau. Le modèle à blocs stochastiques est une approche répandue pour regrouper les individus partageant le même comportement d'interaction.

Lorsque plusieurs types de relations peuvent coexister entre deux individus, on représente les observations par un réseau multiplexe. Dans [**A**15] and [**A**16], nous étendons le modèle à blocs stochastiques aux réseaux multiplexes de façon à obtenir une classification des individus tenant compte de leur comportements social vis-à-vis de plusieurs types de relations. Nous estimons les paramètres du modèle par une version variationnelle de l'algorithme EM, le nombre de blocs étant choisi par un critère de vraisemblance pénalisée. Ces deux articles sont motivés par la modélisation de relations de conseils entre chercheurs en cancérologie.

D'un point de vue méthodologique, les modèles à blocs stochastiques reposent sur l'introduction

des variables latentes, rendant leur vraisemblance incalculable de façon explicite dès que la taille des réseaux augmente. Lorsque l'on s'intéresse à l'inférence bayésienne, les algorithmes classiques d'échantillonnage de la loi a posteriori (MCMC ou population Monte Carlo) atteignent leurs limites sur ce type de modèle. Il leur est donc préféré des approximations déterministes de la loi a posteriori du type approximations variationnelles. Cependant, les propriétés théoriques de ces approximations variationnelles ne sont pas garanties; il est même possible en pratique de montrer que les variances a posteriori sont parfois sous-évaluées. Dans un travail avec S. Robin [**S**1], nous proposons une méthodologie permettant de combiner avantageusement les approximations déterministes de la loi a posteriori avec les derniers développements dans le domaine des algorithmes de Monte Carlo séquentiels.

Mes travaux récents découlent de problématiques écologiques. Cherchant à orienter mes travaux vers des thématiques environnementales, j'ai initié une collaboration avec Wesley Dattilo (écologue, INECOL, Mexique) sur le réseaux multipartites en écologie. En quelques mots, de multiples types d'interactions peuvent coexister dans la nature entre plantes et espèces animales. Parmi ces interactions, nous pouvons citer les interactions mutualistes entre plantes et pollinisateurs, plantes et fourmis protectrices ou oiseaux dispersant les graines. Ces divers interactions jouent un roÎe clé dans la structuration de la biodiversité. Au cours des dernières années, les outils d'analyse de réseaux ont été intensivement utilisés pour comprendre la structure de ces interactions écologiques. Cependant, en général, chaque type d'interaction est étudié de façon indépendante, peu de articles considèrent l'étude jointe des différents réseaux. Parmi les outils classiquement utilisés pour inférer la topologie des graphes bipartites, on trouve les modèles à blocs latents, qui permettent une classification simultanée des lignes et des colonnes (respectivement plantes et animaux dans notre cas). Dans notre travail en cours avec Pierre Barbillon, Avner Bar-Hen et Wesley Dattilo, nous étendons les outils pour graphes bipartites aux graphes multipartites. Ce travail est détaillé dans la section 4.1 du chapitre 3.

De façon plus large, je m'intéresse à l'inférence de données de réseaux ayant une structure complexe, en d'autres termes, je cherche à classifier les entités à partir d'une collection de réseaux dépendants et non à partir d'un unique réseau. L'écologie et la sociologie motivent en grande partie ces travaux.

Mes perspectives de recherche se rapportant aux trois thèmes traités ici sont décrites à la fin de chaque chapitre.

Chapter 1

Statistical inference for some models defined by differential systems

The modeling of dynamical physiological, ecological or biological processes often involves the use of complex systems of differential equations. In general, this system has no explicit solution and has to be approximated by a numerical scheme.

In my works with A. Samson [**A**2] and with A. J. Christen and M. Capistran [**A**12], we study the influence of the approximation of the solution when it comes out to estimate the parameters. The work with A. Samson assumes that the differential system depends on random latent parameters and we work in a frequentist paradigm whereas my work with A. J. Christen and M. Capistran takes place in a Bayesian framework. These two works are described in the following Section 1.

In order to take into account a possible lack of accuracy of the deterministic model, a stochastic volatility term can be introduced in the ODE, thus resulting into a Stochastic Differential Equation (SDE). My works dealing with the inference of statistical models involving SDEs [**A**3] [**A**5] [**A**9] [**P**5] are described in Section 2.

1 Statistical inference for models defined through ODE [A2] [A12]

Let $\mathbf{y} = (y_1, \dots, y_n)$ be n observations of a complex biological process at discrete times $t_1, \dots, t_n \in [t_0, T]$. $y_i \in \mathbb{R}^q, \forall i = 1 \dots n$. We consider the following observation model:

$$y_j = H(X_\phi(t_j)) + \varepsilon_j, \quad \varepsilon_i \sim_{i.i.d.} \mathcal{N}(0, \sigma^2 \mathbf{I}_q) \tag{1.1}$$

where X_ϕ is the solution of the following system of ordinary differential equations,

$$\frac{dX_\phi}{dt} = F(X_\phi, t, \phi); \quad X_\phi(t_0) = X_0. \tag{1.2}$$

$\phi \in A \subset \mathbb{R}^d$ is a vector of unknown parameters and $F : \mathbb{R}^r \times [0, T[\times A \mapsto \mathbb{R}^r$ is a known function defining the dynamical system. $H : \mathbb{R}^r \mapsto \mathbb{R}^q$ is the observation function, modeling the fact that, for instance, we only observe a restricted number of the r components of X_ϕ or some combinations of the r components. The observation variance $\sigma^2 \in S \subset \mathbb{R}^+$ is unknown.

Examples Several examples can be found. See for instance the ones we treated in [**A**12]:

1. Let $X(t)$ be the size of the tumor to time t. The growth is classically described by the following differential equation

$$\frac{dX}{dt} = \lambda X(t)(K - X(t)), \quad X(0) = X_0 \tag{1.3}$$

with λK being the growth rate and K the carrying capacity e.g. $\lim_{t\to\infty} X(t) = K$ and $\theta = (K, \lambda)$. The size of the tumor is observed with error at times $t_1, \ldots, t_n$:

$$y_j = X_\theta(t_j) + \varepsilon_j$$

and $H = id$.

2. An other example can be found when modeling the Oral Glucose Tolerance Test (OGTT). During that test, the individuals ingest a dose of glucose at time $t = 0$, and the blood glucose level is monitored at discrete times. The glucose is regulated by the insulin. Let $G(t)$ be the patient's blood glucose level at time t, in mg/dL. Let $I(t)$ be blood insulin level at time t and $L(t)$ "glucagon" level, to promote liver Glycogen glucose production, in arbitrary units. Let $D(t)$ be the digestive system 'glucose level'; we take it as a compartment in which glucose is first stored (eg. stomach and digestive tract) and in turn delivered into the blood stream. Let also G_b be the glucose base line, (=80 mg/dL, fixed). Andrés Christen et al. (2016) proposed the following system of ODE's

$$\begin{cases} \frac{dG}{dt} &= (L - I)\, G + \frac{D}{\theta_2}, \\ \frac{dI}{dt} &= \theta_0 \left(\frac{G}{G_b} - 1\right)^+ - \frac{I}{a}, \\ \frac{dL}{dt} &= \theta_1 \left(1 - \frac{G}{G_b}\right)^+ - \frac{L}{b}, \\ \frac{dD}{dt} &= -\frac{D}{\theta_2}. \end{cases} \tag{1.4}$$

We denote $X(t) = (G(t), I(t), L(t), D(t))$ and only observe the glucose $G(t)$. Let y_j be the blood glucose level at time t_j. We set the following observation model:

$$y_j = G(t_j) + \varepsilon_j = H(X_\phi(t_j)) + \varepsilon_j$$

where $H : \mathbb{R}^4 \mapsto \mathbb{R}$ with $H(x_1, x_2, x_3, x_4) = x_1$.

As soon as (1.2) has no explicit expression –see for instance system (1.4) in the examples– the regression function of model (1.1) has no explicit expression and a numerical scheme has to be applied to approximate it. We denote by X^h_ϕ the approximate solution of (1.2) derived from the numerical scheme where h is the discretization stepsize of the solver. Any statistical inference is then performed on an approximate model defined as :

$$y_i = H(X^h_\phi(t_i)) + \varepsilon_i, \quad \varepsilon_i \sim_{i.i.d.} \mathcal{N}(0, \sigma^2 \mathbf{I}_q) \tag{1.5}$$

My works [**A**2] and [**A**12] focus on the consequences of performing the statistical inference on the approximate model (1.5) rather than on the true model (1.1). The two works are developed in different inference frameworks and are described hereafter.

1.1 Maximum likelihood estimation for incomplete data models defined by ODE [A2]

In my first joint work with A. Samson [**A2**], we consider the case where we observe the temporal trajectories of not one but I individuals. Let y_{ij} be the observation of individual i ($i = 1 \dots I$) at time t_{ij}. Each observation y_{ij} is the noisy observation of X_{ϕ_i} solution of the same dynamical system but depending on individual parameters ϕ_i. More precisely,

$$y_{ij} = H(X_{\phi_i}(t_{ij})) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim_{i.i.d.} \mathcal{N}(0, \sigma^2 \mathbf{I}_q), \tag{1.6}$$

where $X_{\phi_i} : \mathbb{R}^+ \mapsto \mathbb{R}^q$ is solution of

$$\frac{dX_{\phi_i}}{dt} = F(X_{\phi_i}, t, \phi_i); \quad X_{\phi_i}(t_0) = X_{0,i}. \tag{1.7}$$

For such population data, mixed effects models have proved their efficiency to distinguish the intra and the inter individual variability (Pinheiro and Bates, 2000). In mixed effects models, the individual parameters (ϕ_i) are latent (non-observed) variables such that

$$\phi_i \sim_{i.i.d.} \mathcal{N}_d(\mu, \Omega) \qquad \forall i = 1, \dots, I \tag{1.8}$$

where (μ, Ω) are the population parameters. *In this particular context,* $\theta = (\mu, \Omega, \sigma^2)$ *are the parameters of interest.* In what follows, $(\mathcal{M})$ refers to the "exact" model defined by equations (1.6), (1.7) and (1.8).

Remark 1.1. Note that if ϕ_i is a positive a parameter or constrained to a bounded interval, the Gaussian distribution can still be used provided a reparametrisation of ϕ_i.

Once again, (1.7) having no explicit expression, $X_{\phi_i}(t)$ has to be approximated by a numerical solver. Denoting $X^h_{\phi_i}$ the approximated solution, we perform the inference on the following approximate model:

$$\begin{array}{rcl} y_{ij} & = & H(X^h_{\phi_i}(t_{ij})) + \varepsilon_{ij} \\ \varepsilon_{ij} & \sim & _{i.i.d.}\mathcal{N}(0, \sigma^2 \mathbf{I}_q), \\ \phi_i & \sim & _{i.i.d}\mathcal{N}(\mu, \Omega) \end{array} \tag{1.9}$$

(1.9) defines the approximate model $(\mathcal{M}_h)$.

1.1.1 Maximization of the likelihood by a stochastic version of the EM algorithm

For the sake of simplicity in this manuscript, we consider hereafter that the observations y_{ij} are unidimensional ($q = 1$) and that H is the identity function. However, all the results have been proved for a general H and q.

EM algorithm and stochastic version The likelihood of the observations $\mathbf{y} = (y_{ij})_{i=1\dots I, j=1\dots n_i}$ under the approximate model $\mathcal{M}_h$ expresses as

$$\mathcal{L}_h(\mathbf{y}; \theta) = \int \ell_h(\mathbf{y}|\boldsymbol{\phi}, \sigma^2) p(\boldsymbol{\phi}|\mu, \Omega) d\phi, \tag{1.10}$$

where the latent variables (individual parameters) $\boldsymbol{\phi} = (\phi_1 \dots, \phi_I)$ have been integrated out and

$$\begin{array}{rcl} \ell_h(\mathbf{y}|\boldsymbol{\phi}, \sigma^2) & = & \displaystyle\prod_{i=1}^{I} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\left(y_{ij} - X^h_{\phi_i}(t_{ij})\right)^2} \\ p(\boldsymbol{\phi}|\mu, \Omega) & = & \displaystyle\prod_{i=1}^{I} \frac{1}{\sqrt{(2\pi)^d \det(\Omega)}} e^{-\frac{1}{2}\phi_i^t \Omega^{-1} \phi_i}. \end{array}$$

Due to the non-linearity of $X_{\phi_i}^h(t_{ij})$ as a function of ϕ_i, the "marginal" likelihood $\mathcal{L}_h(\mathbf{y};\theta)$ has no explicit expression. When the observations can be enhanced by latent variables, the Expectation-Maximisation (Dempster et al., 1977) algorithm is a well-known and powerful algorithm to maximise the likelihood. Taking advantage of the explicit expression of the complete likelihood (denoted $\mathcal{L}_h^{\text{compl}}$):

$$\log \mathcal{L}_h^{\text{compl}}(\mathbf{y}, \phi; \theta) = \log \ell_h(\mathbf{y}|\boldsymbol{\phi}, \sigma^2) + \log p(\boldsymbol{\phi}; \mu, \Omega),$$

the EM algorithm is an iterative maximisation method whose (m)-th iteration is decomposed as follows:

Algorithm 1 (EM)**.**

At iteration (m)*,*

- (Step E) *Compute* $Q(\theta|\theta^{(m)}) = \mathbb{E}[\log \mathcal{L}_h^{compl}(\mathbf{y}, \boldsymbol{\phi}; \theta)|\mathbf{y}, \theta^{(m)}]$
- (Step M) *Maximise* $\theta^{(m+1)} = \arg\max_{\theta\in\Theta} Q(\theta|\theta^{(m)})$

In model $(\mathcal{M}_h)$, step (E) has no closed form. To deal with such situations, Delyon et al. (1999) introduce a stochastic version (SAEM) of the EM algorithm, evaluating the $Q(\theta|\theta^{(m)})$ integral by a Robbins-Monroe stochastic approximation procedure. More precisely, the E-step is divided into a simulation step (S-step) of the non-observed data $\boldsymbol{\phi}^{(m)}$ with the conditional distribution $p_h(\boldsymbol{\phi}|\mathbf{y}, \theta^{(m)})$ and a stochastic approximation step (SA-step):

$$Q(\theta|\theta^{(m+1)}) = Q(\theta|\theta^{(m)}) + \gamma_m \left(\log \mathcal{L}_h^{\text{compl}}(\mathbf{y}, \boldsymbol{\phi}^{(m)}; \theta^{(m)}) - Q(\theta|\theta^{(m)}) \right)$$

where $(\gamma_m)_{m\geq 0}$ is a sequence of positive numbers decreasing to 0 such that $\sum_m \gamma_m = \infty$ and $\sum_m \gamma_m^2 < \infty$. They prove the convergence of this algorithm under general conditions in the case where $\mathcal{L}_h^{\text{compl}}(\mathbf{y}, \boldsymbol{\phi}; \theta)$ belongs to a regular curved exponential family.

Again, in model $(\mathcal{M}_h)$ (1.9), the conditional distribution of $\boldsymbol{\phi}$ given the observations $\mathbf{y}$ has no explicit expression and can not be simulated easily. In such cases, Kuhn and Lavielle (2004) suggest using a Monte Carlo Markov Chain (MCMC) algorithm which consists in generating a Markov chain with $p_h(\boldsymbol{\phi}|\mathbf{y}; \theta^{(m)})$ as unique stationary distribution at the m^{th} iteration. The principle of MCMC algorithm is described in Algorithm 3. Assume that $\mathcal{L}_h^{\text{compl}}(\mathbf{y}, \boldsymbol{\phi}; \theta)$ belongs to the exponential family of distributions, i.e. there exist ψ_h and ν_h two functions of θ such that:

$$\mathcal{L}_h^{\text{compl}}(\mathbf{y}, \boldsymbol{\phi}; \theta) = \exp\left\{ -\Psi_h(\theta) + \langle S_h(\mathbf{y}, \boldsymbol{\phi}), \nu_h(\theta) \rangle \right\},$$

where $\langle \cdot\cdot \rangle$ is the scalar product and $S_h(\mathbf{y}, \boldsymbol{\phi})$ is known as the minimal sufficient statistics of the complete model, taking its value in a subset $\mathcal{S}$ of $\mathbb{R}^m$. The SAEM-MCMC algorithm is as follows:

Algorithm 2 (SAEM-MCMC)**.**

At iteration (m)*,*

- (Step S): *Generate* $\boldsymbol{\phi}^{(m)}$ *through a few iterations of MCMC algorithm with stationary distribution* $p_h(\boldsymbol{\phi}|\mathbf{y}, \theta^{(m)})$

- (Step SA): *Stochastic approximation of* $\mathbb{E}[S_h(\mathbf{y},\boldsymbol{\phi})|\mathbf{y},\theta^{(m)}]$ *by*

$$s_{m+1} = s_m + \gamma_m(S_h(\mathbf{y},\boldsymbol{\phi}^{(m)}) - s_m)$$

- (Step M): *Maximisation*

$$\theta^{(m+1)} = \arg\max_{\theta\in\Theta} -\Psi_h(\theta) + < s_{m+1}, \nu_h(\theta) >$$

Remark 1.2. Note that our model $\mathcal{M}_h$ (1.9) is such that its complete likelihood belongs to the exponential family.

MCMC algorithms MCMC algorithms generate Markov chains whose stationary distribution is the distribution of interest. In our case, the distribution of interest is $p_h(\boldsymbol{\phi}|\mathbf{y},\theta^{(m)})$. In this particular context we resort to the most well-known MCMC, namely the Metropolis-Hastings algorithm (Robert and Casella, 2005). Assume that we target $p_h(\boldsymbol{\phi}|\mathbf{y};\theta)$, the iterative Metropolis-Hastings algorithm requires an easily simulable proposal density $\mathcal{K}$ and is written as follows:

Algorithm 3 (Metropolis-Hastings algorithm)**.**

At step $r+1$ *of the Metropolis-Hastings algorithm, given* $\boldsymbol{\phi}^{(r)}$*:*

- *Generate a candidate* $\boldsymbol{\phi}^c$ *from the proposal density* $\mathcal{K}(.|\boldsymbol{\phi}^{(r)})$*.*
- *Generate* $U \sim \mathcal{U}([0,1])$*. Then,*

$$\boldsymbol{\phi}^{(r+1)} = \begin{cases} \boldsymbol{\phi}^c & \text{if} \quad U < \alpha(\boldsymbol{\phi}^{(r)},\boldsymbol{\phi}^c), \\ \boldsymbol{\phi}^{(r)} & \text{if} \quad U > \alpha(\boldsymbol{\phi}^{(r)},\boldsymbol{\phi}^c), \end{cases}$$

where

$$\begin{aligned} \alpha(\boldsymbol{\phi}^{(r)},\boldsymbol{\phi}^c) &= \min\left\{1, \frac{p_h(\boldsymbol{\phi}^c|\mathbf{y};\theta)}{p_h(\boldsymbol{\phi}^{(r)}|\mathbf{y};\theta)}\frac{\mathcal{K}(\boldsymbol{\phi}^{(r)}|\boldsymbol{\phi}^c)}{\mathcal{K}(\boldsymbol{\phi}^c|\boldsymbol{\phi}^{(r)})}\right\} \\ &= \min\left\{1, \frac{\mathcal{L}_h^{compl}(\mathbf{y},\boldsymbol{\phi}^c;\theta)p(\boldsymbol{\phi}^c;\mu,\Omega)}{\mathcal{L}_h^{compl}(\mathbf{y},\boldsymbol{\phi}^{(r)};\theta)p(\boldsymbol{\phi}^{(r)};\mu,\Omega)}\frac{\mathcal{K}(\boldsymbol{\phi}^{(r)}|\boldsymbol{\phi}^c)}{\mathcal{K}(\boldsymbol{\phi}^c|\boldsymbol{\phi}^{(r)})}\right\} \end{aligned} \tag{1.11}$$

is the acceptance probability.

The choice of the proposal density $\mathcal{K}$ is theoretically arbitrary, although in practice a careful choice will help the algorithm to move quickly inside the parameters space. The theoretical and practical properties of such algorithms have been widely studied (see for instance Robert and Casella, 2005).

1.1.2 Contributions in [A2]

- From a *practical point of view*, we propose to reduce the computational cost of our procedure SAEM-MCMC. Indeed, the computation of the acceptance probability $\alpha(\boldsymbol{\phi}^{(r)},\boldsymbol{\phi}^c)$ (1.11) requires the evaluation of $X_{h,\boldsymbol{\phi}_i^c}(t_{ij})$ by the numerical scheme (at each iteration of the Metropolis-Hastings,

within each iteration of the SAEM), which can be burdensome. For a ϕ^c close enough to $\phi^{(r)}$, we propose an extension of the local linearization scheme proposed by Biscay et al. (1996), allowing to solve only partially the differential equation. A simulation study illustrates the fact that our new scheme decreases the computational time. The theoretical convergence of the algorithm is discussed in [**A**2].

- The convergence of SAEM is proved on the approximate model $\mathcal{M}_h$ (1.9) and the distance between the likelihoods of the two models $\mathcal{M}_h$ (1.9) and $\mathcal{M}$ (1.6) is quantified in the following theorem, leading to a bound on the estimates themselves.

Theorem 1. *Assume that equation* (1.7) *is approximated by a numerical scheme of step size h such that for a $p \in \mathbb{N}$ and $C > 0$:*

$$\sup_{t\in[t_0,T]} \|X_\phi(t) - X_\phi^h(t)\| \leq Ch^p$$

Let (γ_m) be a sequence of positive numbers decreasing to 0 such that for any m in $\mathbb{N}$, $\gamma_m \in [0,1]$, $\sum_{m=1}^{\infty} \gamma_m = \infty$ and $\sum_{m=1}^{\infty} \gamma_m^2 < \infty$.

1. *Assuming the sequence $(s_m)_{m\geq 0}$ takes its values in a compact set of $\mathcal{S}$, the sequence $(\theta^{(m)})_{m\geq 0}$ obtained by the SAEM-MCMC algorithm described in Algorithm 2 converges almost surely towards $\theta_{h\infty}$ a (local) maximum of the likelihood $\mathcal{L}_h(\mathbf{y};\theta)$ (equation 1.10).*

2. *Moreover, under regularity assumptions on H and F, for any $\sigma_0^2 > 0$, there exists a constant θ-independent $\mathcal{C}$ such that*

$$\sup_{\theta=(\beta,\sigma^2)\ |\sigma^2>\sigma_0^2} |\mathcal{L}_h(\mathbf{y};\theta) - \mathcal{L}(\mathbf{y};\theta)| \leq \mathcal{C}h^p.$$

3. *Finally, regularity assumptions on the likelihood $\theta \mapsto \mathcal{L}(\mathbf{y};\theta)$ and the pseudo likelihood $\theta \mapsto \mathcal{L}_h(\mathbf{y};\theta)$ imply results on the maximum likelihood estimates themselves. Let θ_∞ be the argmax of $\theta \mapsto \mathcal{L}(\mathbf{y};\theta)$, there exists $\mathcal{C}'$ such that*

$$\|\theta_\infty - \theta_{h,\infty}\| \leq \mathcal{C}'h^p.$$

[**A**2] also presents an application of the algorithm to a dataset simulated using a pharmacokinetic model defined by ODEs. The SAEM estimates are compared with those obtained by the standard estimation software NONMEM, the only available software providing estimates by maximum likelihood in nonlinear mixed models defined by ODEs at that time. SAEM provides satisfying estimates and standard errors of the parameters, while NONMEM does not converge on this simulated example and fails to evaluate the standard errors. The estimation algorithm implemented in NONMEM is based on the linearization of the regression function. The simulation results presented in [**A**2]point out the poor ability of this software to estimate the parameters in nonlinear mixed model defined through ODEs. The algorithm for Non Linear Mixed Effects Models with ODE is now implemented in MONOLIX.

1.2 Bayesian inference for models defined by ODE [A12]

In [**A**12], we consider the initial model

$$y_j = H(X_\phi(t_j)) + \varepsilon_j, \quad \varepsilon_j \sim_{i.i.d.} \mathcal{N}(0, \sigma^2 \mathbf{I}_q) \tag{1.12}$$

where we observe a unique individual temporal trajectory and the parameters of interest are (ϕ, σ). We work in a Bayesian framework and set a prior distribution on (ϕ, σ): $(\phi, \sigma) \sim \pi(\cdot)$ and are interested in the posterior distribution of (ϕ, σ):

$$p(\phi, \sigma|\mathbf{y}) = \frac{\ell(\mathbf{y}|\phi, \sigma)\pi(\phi, \sigma)}{m(\mathbf{y})} \propto \ell(\mathbf{y}|\phi, \sigma)\pi(\phi, \sigma) \tag{1.13}$$

where $m(\mathbf{y})$ is the marginal likelihood (where the unknown parameters have been integrated out). $\ell(\mathbf{y}|\phi)$ depends on $X_\phi(t)$ solution of the differential system, which has in general no explicit expression: the inference can not be performed on this "true" model $\mathcal{M}$. As in [**A2**], we approximate $X_\phi(t)$ by $X_\phi^h(t_i)$ thanks to a numerical scheme and get an explicit likelihood $\ell_h(\mathbf{y}|\phi)$ of the approximate model $\mathcal{M}_h$:

$$\begin{array}{rcl} y_j & = & H(X_\phi^h(t_j)) + \varepsilon_j \\ \varepsilon_j & \sim & _{i.i.d.}\mathcal{N}(0, \sigma^2 \mathbf{I}_q) \\ (\phi, \sigma) & \sim & \pi(\cdot) \end{array} \tag{1.14}$$

A sample from the posterior distribution $p_h(\phi|\mathbf{y}) \propto \ell_h(\mathbf{y}|\phi)\pi(\phi)$ is obtained with a Metropolis-Hastings algorithm (see Algorithm 3).

In general, the replacement of the theoretical (non-available) solution of the differential system by a numerical approximation is ignored, the solver being used as a black box. However, recently, research has been directed at trying to quantify the consequences of such an approximation, commonly by comparing expected values of the resulting posterior distributions, like the exact vs the numerical Posterior means. In [**A12**] we adopt a different approach, basing our comparison on the use of Bayes Factors (BFs), which is the natural tool for comparing models in a Bayesian context.

The results we obtain in [**A12**] are the following ones.

- From the results demonstrated in [**A2**] we derive the following results.

Theorem 2. *Assume that (ϕ, σ) remains in a compact set $A \times S$, and that the numerical scheme of step size h is such that $\{t_1, \ldots, t_n\} \subset h\mathbb{N}$ and*

$$\max_{t \in \{t_1, \ldots, t_n\}} \|X_\phi(t) - X_\phi^h(t)\|_{\mathbb{R}^p} \leq C_\phi h^p.$$

Also assume that the observation function H is differentiable with a bounded derivative. Then there exists a constant $C_\mathbf{y}$ such that for every (ϕ, σ) and h small enough

$$|p(\phi, \sigma|\mathbf{y}) - p_h(\phi, \sigma|\mathbf{y})| \leq C_\mathbf{y} \pi(\phi, \sigma) h^p.$$

As a consequence,

$$D_{T.V}\left(p(\theta, \sigma|\mathbf{y}), p_h(\theta, \sigma|\mathbf{y})\right) \leq C_\mathbf{y} h^p$$

where $D_{T.V}$ is the total variation distance. Moreover,

$$\|(\hat{\phi}^{L^2}, \hat{\sigma}^{L^2}) - (\hat{\phi}^{h,L^2}, \hat{\sigma}^{h,L^2})\| \leq h^p C'_\mathbf{y}$$

where $\hat{\phi}^{L^2}$ is the posterior expectation.

- In the Bayesian paradigm, model selection is performed using the Bayes Factor defined as follows. Let $\mathbf{y}$ be the observed data and let $\mathcal{M}$ and $\mathcal{M}_h$ be our two models in competition (defined in (1.12) and (1.14)).

Consider a prior distribution on the set of the models $\{\mathcal{M}, \mathcal{M}_h\}$, the decision between the competing models $\mathcal{M}$ and $\mathcal{M}_h$ is based on the ratio of the posterior probability for each model, namely the Bayes Factor (BF)

$$B_{\mathcal{M},\mathcal{M}_h} = \frac{P(\mathcal{M}|\mathbf{y})}{P(\mathcal{M}_h|\mathbf{y})} = \frac{m(\mathbf{y})}{m_h(\mathbf{y})} \frac{P(\mathcal{M})}{P(\mathcal{M}_h)},$$

where $m_h(\mathbf{y})$ and $m\mathbf{y}(\mathbf{y})$ are the marginal likelihoods of $\mathbf{y}$ from model $\mathcal{M}_h$ and $\mathcal{M}$, respectively, defined by

$$m_h(\mathbf{y}) = \int \ell_h(\mathbf{y}|\phi,\sigma)\pi(\phi,\sigma)d\phi d\sigma \quad \text{and} \quad m(\mathbf{y}) = \int \ell(\mathbf{y}|\phi,\sigma)\pi(\phi,\sigma)d\phi d\sigma$$

Theorem 3. *Under the same assumptions as Theorem 2, we get*

$$m(\mathbf{y}) = m_h(\mathbf{y}) + O(h^p).$$

That is, there exists a constant $B(\mathbf{y}) \in \mathbb{R}$ (which does not depend on h) such that

$$\frac{m(\mathbf{y})}{m_h(\mathbf{y})} \simeq 1 + B(\mathbf{y})h^p.$$

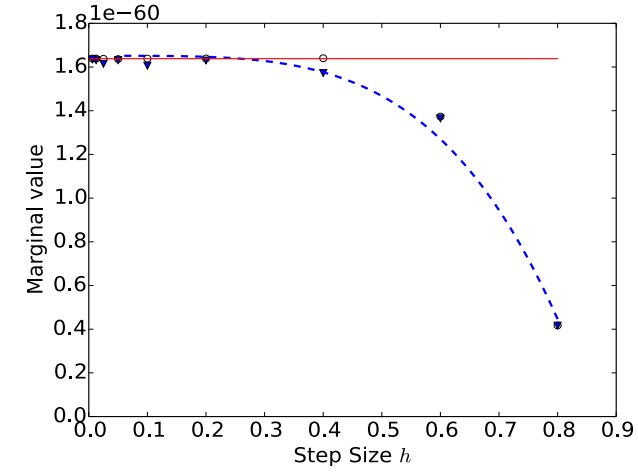
We comment the following regarding the above results. First, as can be noticed in the demonstration of Theorem 2 in [**A**12], we contribute to the intuitive idea that the ODE solver approximation error should be put in the perspective of the observational error σ. As a consequence, the numerical solver used may be viewed in this perspective and not solely as a black box number crunching routine. As far as the main aim is to make inference on parameters, there is no need to use to highest precision if the data are contaminated by a non-neglectable quantity of noise. In a domain where the computational time is important, we prove in [**A**12] that considerable CPU time savings may be obtained only by using a reasonable step size in the solver.

Secondly, from Theorem 3, we deduce that the marginal likelihood $m(\mathbf{y})$ of the unavailable theoretical model $\mathcal{M}$ now may be estimated. Indeed, an obvious method is to compute $m_{h_k}(\mathbf{y})$ for various step sizes $\{h_k, k = 1 \dots K\}$ and fit the simple linear regression $m_{h_k}(\mathbf{y}) = a + bh_k^p$; a then provides an estimation of $m(\mathbf{y})$. This means that by using a multi-resolution computation of $m_{h_k}(\mathbf{y})$ on various approximate models, we are able to estimate the marginal likelihood of the true model.

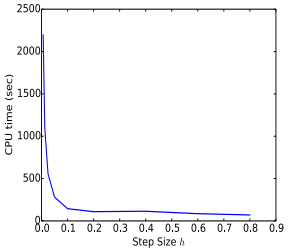
As an example, let us have a look at Figure 1.1 illustrating a study on synthetic data for the logistic growth (equation 1.3) with $\sigma = 30$. The marginal likelihood $m_h(\mathbf{y})$ has been computed for various step sizes, both exact (circles, using numerical integration) and estimated using the MCMC sample (triangles). The ODE is approached with a Runge-Kutta solver of order 4 (classical RK4, blue). In this particular case, the ODE has an explicit solution thus we can compare the estimates with the true $m(\mathbf{y})$ (horizontal red line). Following Theorem 3, we also indicate (dashed line) the regression for estimated values for $m_h(\mathbf{y}) = a + bh^p$ ($p = 4$). As expected, the polynomial curve provides a good approximation of the true marginal likelihood (plot (a) of 1.1). Moreover, we compare the CPU times relative to 10,000 iterations of the MCMC for the various step sizes (plot (b) of Figure 1.1). The algorithm requires 36 min for $h = 0.00625$ and and 2.5 minutes for $h = 0.1$. However, the posterior distributions for parameter λ can not be distinguished (plot (c) of Figure 1.1)

There are still some particular issues to be solved when applying our results to more realistic inverse problems like estimating the marginals in a multidimensional parameter problem and analyzing stiff problems where a multistep method would need to be used.

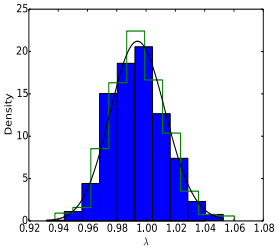
Note that between my two publications [**A**2] and [**A**12] several papers on this topic have been published. The introduction of Conrad et al. (2017) contains several interesting references.



(a)



(b)



(c)

Figure 1.1 – Study on synthetic data for the Logistic growth with $\sigma = 30$. (a) Marginal $m_h(\mathbf{y})$ for various step sizes, both exact (circles, using numerical integration) and estimated using the MCMC sample (triangles). We use a Runge-Kutta solver of order 4 (classical RK4, blue). The horizontal lines (red) is the true marginal $m(\mathbf{y})$ calculated using numerical integration on the analytic solution. Dashed lines indicate the regression for estimated values for $m_h(\mathbf{y}) = a + bh^p$ for the order $p = 4$. (b) Corresponding CPU time, relative to 10,000 iterations of the MCMC. (c) Posterior distribution of λ the for RK4 solver, $p = 4$, for step sizes $h = 0.00625$ and $h = 0.1$ (histograms; and exact posterior, black density). The former takes 36 min and the latter 2.5 min.

2 Inference for statistical models defined by SDEs [A3] [A5] [A8] [A9] [P5]

As underlined in the previous section, biological or physiological phenomena are often described by differential systems derived from physiology. In general, the proposed models are deterministic, that is, the observed dynamic is assumed to be driven exclusively by deterministic mechanisms. However, real biological processes are always exposed to influences that are not completely understood or not feasible to model explicitly. Ignoring these phenomena in the modeling may affect the estimation of the parameters and the derived conclusions. Therefore there is an increasing need to extend the deterministic models to models including a stochastic component. A natural extension of deterministic differential equations model is a system of stochastic differential equations (SDEs), where relevant parameters have been modeled as suitable stochastic processes, or stochastic processes have been added to the driving system equations. Note that with Adeline Samson, we wrote a review paper about the use of Stochastic Differential Equation in the particular field of Pharmacokinetics-Pharmacodynimics [**A**8].

In what follows, I will present three papers ([**A**3], [**A**5] and [**A**9]) dealing with non-linear mixed effects models defined with SDEs. My last work in progress [**P**5] deals with a different context where the SDE is used to represent ecological trajectories and we aim at performing a segmentation and classification of the trajectories.

Non linear mixed effects models with SDE Let y_{ij} be the observation of individual i $(i = 1 \dots I)$ at time t_{ij}. Each observation y_{ij} is the noisy observation of $Z(t_{ij}, \phi_i)$

$$y_{ij} = Z(t_{ij}, \phi_i) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim_{i.i.d.} \mathcal{N}(0, \sigma^2), \tag{1.15}$$

where $Z(t, \phi_i)$ is solution of a stochastic differential equation (SDE) depending on individual parameters ϕ_i:

$$dZ(t, \phi_i) = F(Z(t, \phi_i), t, \phi_i)dt + \Gamma(Z(t, \phi_i), t, \phi_i, \gamma^2)dB_t, \quad Z(t_0, \phi_i) = Z_0(\phi_i). \tag{1.16}$$

Adopting the non-linear mixed effects model approach, the individual parameters are distributed as:

$$\phi_i \sim_{i.i.d.} \mathcal{N}_d(\mu, \Omega) \qquad \forall i = 1, \dots, I. \tag{1.17}$$

The combination of equations (1.15), (1.16) *and* (1.17) *will be referred as the true model* $\mathcal{M}$ *in what follows.* The parameters of interest are $\theta = (\mu, \Omega, \sigma^2, \gamma^2)$. Taking into account the stochasticity of the underlying process $\mathbf{Z} = (Z(t_{ij}, \phi_i))_{i=1,\dots,I, j=1,\dots,n_i}$, the likelihood of the observations $\mathbf{y}$ with respect to model $\mathcal{M}$ is:

$$\mathcal{L}(\mathbf{y}; \theta) = \int \ell(\mathbf{y}|\mathbf{Z}, \sigma^2) p(\mathbf{Z}|\boldsymbol{\phi}; \gamma^2) p(\boldsymbol{\phi}; \theta) d\mathbf{Z} d\boldsymbol{\phi}$$

where:

$$\begin{aligned}
\ell(\mathbf{y}|\mathbf{Z}, \sigma^2) &= \prod_{i=1}^{I} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_{ij} - Z(t_{ij}, \phi_i))^2} \\
p(\mathbf{Z}|\boldsymbol{\phi}, \gamma^2) &= \prod_{i=1}^{I} \prod_{j=1}^{J} p(Z(t_{ij}, \phi_i)|Z(t_{ij-1}, \phi_i); \gamma^2) \\
p(\boldsymbol{\phi}|\mu, \Omega) &= \prod_{i=1}^{I} \frac{1}{\sqrt{(2\pi)^d \det(\Omega)}} e^{-\frac{1}{2}\phi_i^t \Omega^{-1} \phi_i}
\end{aligned}$$

For this model, we propose a Bayesian inference in a joint work with A. Samson and J.-L. Foulley with application to agronomy [**A**5]. With A. Samson, we propose a maximum likelihood estimation ([**A**3] and [**A**9]) thanks to the SAEM-MCMC algorithm. Whatever the framework is, two main difficulties arise.

1. First of all, in general the SDE (1.16) has no explicit solution i.e. their is no closed form for the transition densities $p(Z(t_{ij},\phi_i)|Z(t_{ij-1},\phi_i);\gamma^2)$. In this case, the solution has to be approximated with a numerical scheme. In [**A**3], we use the Euler-Maruyama scheme.

2. Secondly, the Bayesian inference and the SAEM-MCMC algorithm both require simulating the latent variables $(\mathbf{Z},\boldsymbol{\phi})$ under the conditional distribution $p(\mathbf{Z},\boldsymbol{\phi}|\mathbf{y},\theta)$. However, the dimension of the latent variables becomes problematic and powerful kernels have to be specially designed.

2.1 Bayesian inference for NLME models defined by SDE [A5]

In [**A**5], we illustrate the advantage of SDE over ODE on a particular dataset. We focus on the modeling of chicken growth. Data $\mathbf{y}$ are noisy weight measurements of $n = 50$ chickens at weeks t =0, 4, 6, 8, 12, 16, 20, 24, 28, 32, 36, 40 after birth.

Statistical model and Bayesian inference Several growth curve function can be found. Following the previous study of this dataset by Jaffrézic et al. (2006), we chose the Gompertz nonlinear mixed model with an additive measurement error on the logarithm of the observations:

$$\left\{\begin{array}{rcl}\log y_{ij} & = & \log A_i - B_i e^{-C_i t_{ij}} + \varepsilon_{ij},\ \varepsilon_{ij} \sim_{i.i.d.} \mathcal{N}(0,\sigma^2),\ \forall (i,j)\\ \phi_i & = & (\log A_i, B_i, \log C_i) \sim_{i.i.d.} \mathcal{N}(\mu,\Omega),\ \forall i = 1,\ldots,n\end{array}\right. \tag{1.18}$$

However, this model prescribes purely increasing curves and fails to capture unexpected variations of growth rate for some individuals (see Figure 1.2). As a consequence, we modify the ODE satisfied by the Gompertz function

$$f('t) = BCe^{-Cf(t)}, \quad f(0) = Ae^B$$

into a SDE. Given the heteroscedasticity of the process, the volatility function is set to be equal to $\Gamma(Z_t,\phi,\gamma^2) = \gamma Z_t$, resulting into:

$$dZ_t = BCe^{-Ct}Z_t dt + \gamma Z_t dW_t,\ Z_0 = Ae^{-B} \tag{1.19}$$

Equation (1.19) implies that the standard error of the random perturbations of the growth rate is proportional to the weight. This choice of volatility has two main advantages. First, SDE (1.19) has an explicit solution Z_t which is a multiplicative random perturbation of the solution of the Gompertz model $f(t)$. Secondly, due to the assumption of the non-negativity of A, Z_t is almost surely non-negative, which is a natural constraint to model weight records. Once the solution of the SDE has been discretized, we get to the following model:

$$\left\{\begin{array}{l}(\log y_{i0}, \log y_{i1},\ldots,\log y_{in_i})' = \left(\log(A_i) - B_i, Z_{t_{i1}},\ldots,Z_{t_{in_i}}\right)' + \varepsilon_i,\\ \varepsilon_i \sim_{i.i.d.} \mathcal{N}\left(0,\sigma^2\mathbf{I}_{n_i+1}\right)\\ \left(Z_{t_{i1}},\ldots,Z_{t_{in_i}}\right)' = \left(\log(A_i) - B_i e^{-C_i t_{i1}},\ldots,\log(A_i) - B_i e^{-C_i t_{in_i}}\right)' - \gamma^2\left(t_{i1},\ldots,t_{in_i}\right)' + \eta_i\\ \eta_i \sim_{i.i.d} \mathcal{N}\left(0_J,\gamma^2 T_i\right),\ T_i = (\min(t_{ij},t_{ij'}))_{1\leq j,j'\leq n_i}\\ (\log A_i, B_i, \log C_i) \sim_{i.i.d.} \mathcal{N}(\mu,\Omega)\end{array}\right. \tag{1.20}$$

Setting standard prior distributions on $(\mu, \Omega, \gamma, \sigma)$, we sample the posterior distribution using a Metropolis-Hastings within Gibbs algorithm. Note that in that particular model the conditional distribution of $\mathbf{Z}_i = (Z_{ij})_{1\leq j\leq n_i}$ given $(\phi_i, \theta, \mathbf{y}_i)$ is Gaussian with explicit conditional mean and variance. As detailed in [**A**5], some individual and population parameters also have explicit conditional distribution, thus allowing to design an efficient MCMC algorithm.

Numerical experiment [**A**5] presents a large simulation study to compare the models defined by ODE and SDE. In particular, we illustrate the fact that using the ODE model for data simulated with a SDE model leads to biases estimates, not only on the variance terms but also on the parameters of fixed effects. As a consequence, considering a ODE model instead of a SDE model can lead to inaccuracy in the estimation of the population parameters. On the contrary, if the data come from the ODE model and are estimated with SDE model, no lack of accuracy is detected.

The analysis of the real data set also supplies interesting results. The proposed models (ODE and SDE) are applied on real data of chicken growth. The estimate of γ^2 is strictly positive and its credibility interval puts the parameter of long way from zero. This means that the dynamical process that most likely represents the growth is a stochastic process with non-negligible noise. The diagnostic tools also show a clear improvement from the ODE model to SDE model for the whole population, both at early and late ages. The reduction in DIC from the ODE to the SDE models is equal to 393, which clearly indicates the better predictive ability of the SDE model.

Figure 1.2 reports, for four subjects, the observed weights, the ODE prediction, the empirical mean of the last 1000 simulated trajectories of the SDE (1.20) generated during the Gibbs algorithm, their empirical 95 % confidence limits (from the 2.5th percentile to the 97.5th percentile) and one simulated trajectory. Subjects 4 and 13 are examples of subjects with no growth slow down. Both ODE and SDE models satisfactorily fit the observations. Subject 14 has a small observed weight decrease. For subject 1, the weight decrease is more important. For both subjects, the ODE model fails to capture this phenomenon while the SDE model does. Furthermore, the SDE model provides different estimates for the individual parameters. For example for subject 1, the individual parameter A_1 (adult weight) is estimated at 3.922 kg and 3.484 kg by the ODE and SDE models, respectively. The use of Bayesian model validation tools validate the superiority of the SDE model over the ODE model on this dataset (see [**A**5] for more details).

2.2 Maximum likelihood inference via SAEM-MCMC algorithm

I now present two joint works with A. Samson, dealing with the maximum likelihood estimation of model defined by equations (1.15), (1.16) and (1.17).

2.2.1 When the SDE has no explicit solution [A3]

In [**A**3], we consider the case where the transition density $p(Z(t,\phi_i)|Z(s,\phi_i), s<t;\gamma^2)$ has no closed form. In this case, we propose to approximate the SDE solution with a Euler-Maruyama scheme, which resorts to approximating the distribution $p(Z_{t+h}|Z_t,\phi,\theta)$ by a Gaussian distribution. This approach is widely used in finance for high-frequency datasets. However, in the applications we are interested in (pharmacokinetics, agronomics...) the laps time between two observations $t_{ij}-t_{ij-1}$ is too long for the gaussian approximation of $p(Z(t_{ij},\phi_i)|Z(t_{ij-1},\phi_i),\gamma^2)$ to be relevant. As a consequence, we have to introduce a finer time grid where the gaussian approximation will be performed. More precisely, $\forall i = 1\dots I$, let $(\tau_{im})_{m=0\dots M}$ be an increasing sequence such that: $\tau_{i0}=t_0$, $\tau_{iM}=t_J$, $\tau_{im}-\tau_{im-1}=h$ and $\forall j=1\dots n_i$, there exists m_{ij} such

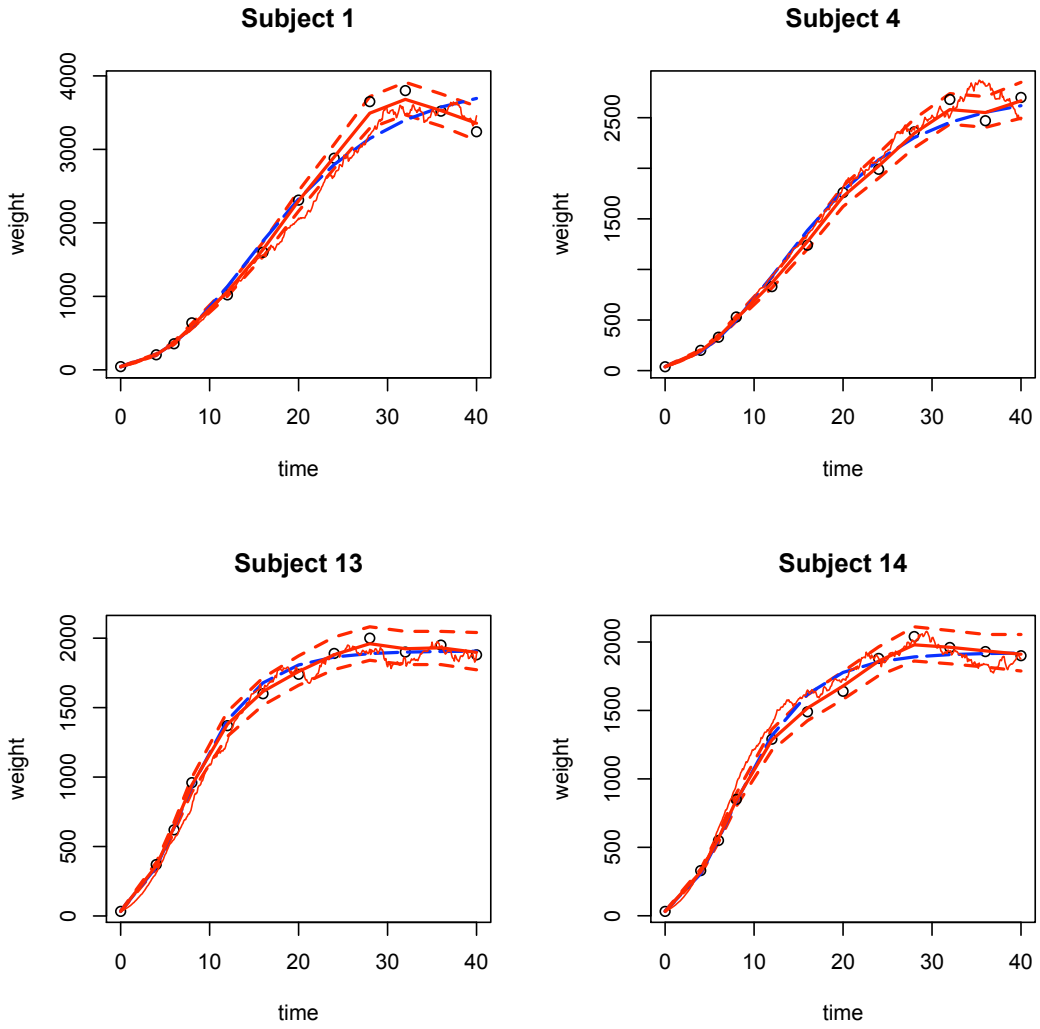


Figure 1.2 – Observations (circles), predictions obtained with the ODE mixed model (long dashed line), mean SDE prediction (smooth solid line), 95% credibility interval obtained with the SDE mixed model (dotted line) and one SDE realization (solid line), for subjects 1, 4 13 and 14.

that $t_{ij} = \tau_{m_{ij}}$. Then, the approximate model $(\mathcal{M}_h)$ is written as follows:

$$(\mathcal{M}_h) \left\{ \begin{array}{rcl} y_{ij} & = & \tilde{Z}^h_{m_{ij}}(\phi_i) + \varepsilon_{ij}, \\ \varepsilon_{ij} & \sim_{i.i.d.} & \mathcal{N}(0, \sigma^2) \\ \tilde{Z}^h_m(\phi_i) & = & \tilde{Z}^h_{m-1}(\phi_i) + h\, F(\tilde{Z}^h_{m-1}(\phi_i), \tau_{im-1}, \phi_i) + \Gamma(\tilde{Z}^h_{m-1}(\phi_i), \phi_i, \gamma^2)\sqrt{h}\xi_{im} \\ \xi_{im} & \sim & {}_{i.i.d}\mathcal{N}(0,1) \\ \phi_i & \sim & \mathcal{N}(\mu, \Omega) \end{array} \right.$$

The likelihood of the observations with respect to the approximate model $(\mathcal{M}_h)$ $\mathcal{L}_h(\mathbf{y};\theta)$ –which can be seen as a pseudo-likelihood now has the expression:

$$\mathcal{L}_h(\mathbf{y};\theta) = \int \ell(\mathbf{y}|\tilde{\mathbf{Z}}^h, \sigma^2) p_h(\tilde{\mathbf{Z}}^h|\boldsymbol{\phi};\gamma^2) p(\boldsymbol{\phi};\theta) d\tilde{\mathbf{Z}}^h d\boldsymbol{\phi}$$

where all the terms in the integral have explicit expression.

We propose to maximize the pseudo likelihood $\mathcal{L}_h$ by the SAEM-MCMC algorithm where $(\mathbf{y}, \tilde{\mathbf{Z}}^h, \boldsymbol{\phi})$ are the complete data. Assuming that $\mathcal{L}_h$ belongs to the exponential family (see section 1.1), we implement Algorithm 2 on the approximate model $\mathcal{M}_h$, the latent variables to be simulated being $\tilde{\mathbf{Z}}^h$ and $\boldsymbol{\phi}$ conditionally to $\mathbf{y}$ for a given θ.

Designing a MCMC algorithm to sample $p_h(\tilde{\mathbf{Z}}^h, \boldsymbol{\phi}|\mathbf{y}, \theta)$ can be tricky since the volume of latent variables is huge. At each iteration ℓ of the MCMC algorithm the task is divided into two steps:

1. Propose $(\tilde{\mathbf{Z}}^h)^c$ with a kernel $\mathcal{K}_1$ and accept $(\tilde{\mathbf{Z}}^h)^{(r)} = (\tilde{\mathbf{Z}}^h)^c$ with probability such that $p_h(\cdot|\boldsymbol{\phi}^{(r-1)}, \mathbf{y}, \theta)$ is stationary.
2. Propose $\boldsymbol{\phi}^c$ with a kernel $\mathcal{K}_2$ and accept $\boldsymbol{\phi}^{(r)} = \boldsymbol{\phi}^c$ with probability such that $p_h(\cdot|(\tilde{\mathbf{Z}}^h)^{(r)}, \mathbf{y}, \theta)$ is stationary.

The choice of the transition kernel $\mathcal{K}_1$ is crucial to guaranty the theoretical and practical convergence properties of the MCMC algorithm and consequently of the SAEM-MCMC algorithm. In [**A**3] we propose two kernels. The first naive kernel is $p_h((\tilde{\mathbf{Z}}^h)^c|\boldsymbol{\phi}^{(r-1)}, \theta)$; this kernel does not use the observations $\mathbf{y}$ thus leading a low acceptance rate and a slow exploration of the space. In order to take into account the observations, we propose to sample $(\tilde{\mathbf{Z}}^h)^c$ as a Brownian bridge between the observations. This second kernel allows to reach better convergence properties.

From a theoretical point of view, in the frequentist context [**A**3] we control the likelihood function of the approximate model as a function of the step-size h of the Euler-Maruyama scheme.

Theorem 4. *Under mild regularity assumptions on F and form assumptions on the volatility function $\Gamma(Z_t, \phi, \gamma^2)$ there exists $C_{\mathbf{y}}$ such that, for h small enough:*

$$\sup_{\theta=(\mu,\Omega,\sigma^2,\gamma^2)\ |\gamma_0^2<\gamma^2<\Gamma_0^2} |\mathcal{L}(\mathbf{y};\theta) - \mathcal{L}_h(\mathbf{y};\theta)| \leq C_{\mathbf{y}} h$$

Moreover, the sequence of the estimators $(\theta^{(m)})_{m\geq 1}$ supplied by the SAEM-MCMC algorithm implemented on the approximate model $\mathcal{M}_h$ converges almost surely toward a (local) maximum of $\theta \mapsto \mathcal{L}_h(\mathbf{y};\theta)$, denoted $\theta_{h,\infty}$. Besides, under regularity assumptions $\mathcal{L}_h(\mathbf{y};\theta)$ and $\mathcal{L}(\mathbf{y};\theta)$ there exist a constant C', independent of θ such that $\|\theta_{h,\infty} - \theta_\infty\|^2 \leq C' h^p$ where θ_∞ is the maximum likelihood of the true model ($\mathcal{M}$).

The proof of the theorem relies on the convergence rate of the approximation of the transition density by the Euler-Maruyama scheme (Bally and Talay, 1996). Note that a similar result is demonstrated on the posterior distributions in the Bayesian framework in [**A**5].

In [**A**3] a simulation study in the context of pharmacokinetics illustrates the convergence properties of the SAEM-MCMC algorithm. An application on a real dataset also illustrates the superiority of the SDE over the ODE model.

Remark 2.1. Note that as h decreases, on the one hand the approximation of the SDE's solution gets more accurate but on the other hand the volume of non-observed data (the individual parameters ϕ_i and the latent stochastic process $Z^h_{m_{ij}}$ at each instant of the grid) increases dramatically. As a consequence, in practice a compromise has to be found.

2.2.2 Combining SAEM with PMCMC [A9]

In a second work with A. Samson [**A**9] we propose to improve the simulation of the latent process $\mathbf{Z}$ conditionally to $(\boldsymbol{\phi}, \mathbf{y})$ using the recent particular methods proposed by Andrieu et al. (2010). In [**A**9] we restrict our study to the case where the transition density $p(Z(t_{ij}, \phi_i)|Z(t_{ij-1}, \phi_i); \gamma^2)$

has an explicit expression. As seen before, simulating $(\mathbf{Z}, \boldsymbol{\phi})$ conditionally to the observations $\mathbf{y}$ is a challenging algorithmic task: MCMC algorithm with naive kernels will lead to poor convergence properties due to the dimension of the space to be explored. Simulating $\mathbf{Z}|\boldsymbol{\phi}, \mathbf{y}; \theta$ is a filtering task. When this distribution is not explicit, Sequential Monte Carlo methods have been proposed (Doucet et al., 2001). However, the particle filtering is difficult to combine with the sampling of the parameters $\boldsymbol{\phi}$. Andrieu et al. (2010) proposed a powerful algorithm namely the Particle Markov Chain Monte Carlo (PMCMC) combining the strength of MCMC and SMC algorithms. The PMCMC has the great property to have $p(\mathbf{Z}|\boldsymbol{\phi}, \mathbf{y}; \theta)$ as invariant distribution, whatever the number of particles used at the SMC step is. In [**A**9], we propose to combine SAEM with PMCMC to get a more efficient algorithm for the estimation of the parameters of the model defined by (1.15), (1.16) and (1.17).

Using the convergence properties of PMCMC, we are able to prove the convergence of the SAEM-PMCMC algorithm. A simulation study performed on the toy example Ornstein-Uhlenbeck processus highlights the fact that the number of particles has no influence on the quality of the estimation. Moreover, a simulation study on a Gompertz diffusion (1.19) with heteroscedastic error proves that SAEM-PMCMC can help to improve the estimation of the volatility parameter γ^2.

3 Conclusion and perspectives

3.1 My contributions in a few words

I have been working on statistical models defined through ODE and SDE since my PhD and until recently.

On statistical models defined by ODE, my main contribution is the theoretical and practical study of the influence of the numerical approximation of the dynamical system solution on the parameters inference (frequentist or Bayesian). Indeed, whereas numerical solvers were classically used as black box in inference procedures, their influence was not studied. In our work, we monitored the error induced by their utilization. My work took place in frequentist and Bayesian frameworks where specific tools can be used to quantify the influence.

SDE are attractive and elegant tools to model biological processes (see for instance our review paper [**A**8] in the pharmacology field). As illustrated in [**A**5] their use can be clearly fruitful in a practical context. However, they imply methodological and theoretical significant difficulties. Estimating the parameters of such processes has been widely tackled in financial datasets with high-frequency data. They have been much less addressed in biology where the data are clearly not of that type. In my works, I endeavored to develop efficient estimation algorithms and supply control results when the SDE has no explicit solution and is replaced by a numerical scheme, in a biological framework.

This field of research is obviously still very active. Recently, Approximate Bayesian Computation has provided new perspectives and very flexible computational tools have been developed (see for instance Liepe et al., 2014, to name but a few).

As far as I am concerned, I slowly withdrew from this thematic to focus on the ones described in the following chapters. However, I still have a research project relying on SDEs and described here after.

3.2 Future work: SDE and rupture detection in ecology

I started recently a collaboration with J. Chiquet, M.P. Etienne and A. Samson around a problematic related to SDEs. Processes driven by stochastic differential equations have recently been seen as powerful tools to model ecological movements (such as birds, aquatic animals, fishes...). Movement ecology data typically deals with positions in space over a sequence of discrete points in time. The position is recorded thanks to Global Position System (GPS) and the acquisition frequency might vary from one position per day to one position per second depending on the species and the specificity of the studied movement.

Let $(Y_0, \dots, Y_n) \in (\mathbb{R}^2)^{n+1}$ denotes the sequence of relocations times $0 = t_0, \dots, t_n$. Among the standard movement models, we aim at considering the following ones:

- the Random walk with drift: $Y_{k+1} = Y_k + (t_{k+1} - t_k)\nu^{(RWD)} + E_{k+1}$ with $E_k \overset{i.i.d}{\sim} \mathcal{N}(0, \Psi^{(RWD)})$,
- the bi-dimensional Ornstein Ulhenbeck, defined as follows : $dY_t = B(X_t - \mu)dt + \Lambda dW_t$
- or the Continuous time Correlated random walk is defined as an Ornstein Uhlenbeck process on the velocity in each direction (no correlation between dimension) and the position is defined as the integral of the velocity. Note that this system has an explicit solution.

Many interesting question in ecology or in fisheries science are addressed through the study of such GPS data and one of the first step in the analysis of such data requires to identify homogeneous portion in the trajectory.

Two main class of approaches are used for such an identification: methods based on Hidden Markov Model (HMM) and methods based on segmentation. HMM based methods assume that the length of such homogeneous regions exhibits a geometric distribution and uses some Expectation Maximization approaches that might converge to a local maximum. Those methods are used for identifying underlying activities associated with those homogeneous regions. On the other hand, segmentation based approaches don't request anything regarding the length of the homogeneous regions but they are not meant to identify underlying activities as every region exhibits its own specificities. Such segmentation based methods might be associated with mixture models for the identification of underlying activities Picard et al. (2007) except that estimations methods for such model might not be tractable in practice. We propose a new approach for identifying underlying activities based on the identification of homogeneous region in a trajectory. This approach is built using a dynamic programing algorithm coupled with regularization technics.

A segmentation - classification strategy In general, assume that $(Y_k)_{k\in 1,\dots,K}$ are the discrete time observations of a stochastic process (Y_t) whose distribution (described by a time series or a SDE) depends on unknown parameters θ. We denote by $\mathbb{F}_\theta$ this distribution. We propose a stragety in two steps : first segmentation of the trajectory, secondly use of regularization tools to gather the segments with similar parameters.

- *Segmentation* We assume that the parameters are not constant along time but can take a finite number of values, i.e. there exist times instants $\tau_1 \leq \cdots \leq \tau_R$ such that the distribution of Y over $[\tau_r, \tau_{r+1}]$ is $\mathbb{F}_{\theta_r}$ where the $(\tau_r)_{r=1\dots R}$ and the $(\theta_r)_{r=1\dots R}$ are unknown. The change points (estimation of $(\tau_r)_{r=1\dots R}$ and $(\theta_r)_{r=1\dots R}$) are chosen to be of minimal cost where the cost is some well chosen criteria, such as minus the log-likelihood or the quadratic loss (depending on the model). Solving this problem is a hard computational task, however it can be enhanced by dynamic programming strategies.

- *Regularization* Once the change points have been identified, we want to gather the segments in order to be able to identify groups of segments to a specific activity. Concretely, we want to find non-consecutive segments $[\tau_r, \tau_{r+1}]$ sharing the same parameters θ_r. To that purpose, we propose to use regularization methods such as Group-Lasso. Dealing with the dependency between the sequential observations is a complex issue, we are working on it.

A first successful attempt on the Variational Auto Regressive (VAR) model has been presented in [**P**5]. The method has to be tested on other models and with non-regularly observed trajectories.

Chapter 2

Bayesian Inference for some multiplicative counting processes

Contributions In this chapter, I present four contributions written with J. Rousseau, V. Rivoirard (for three of them) and C. Scricciolo (for two of them). The first work [**A**11] is motivated by the analysis of an electrical network through time and we propose the Bayesian inference of a special counting process in the particular case where the process $N(t)$ is partially observed (Section 1). Contributions [**A**13] and [**A**14] are dedicated to the study of frequentist concentration properties of the posterior distribution in a Bayesian non-parametric context: [**A**13] focuses on the posterior concentration in multiplicative Aalen process, whereas the second one [**A**14] considers the concentration in the case of a data-dependent prior (Empirical Bayes). In both papers, I propose numerical illustrations requiring the design of adapted algorithmic tools. These works are described in Sections 2. Finally, Section 3 is dedicated to the Bayesian non-parametric inference of multivariate Hawkes processes [**S**2].

1 Bayesian inference for partially observed multiplicative intensity processes [A11]

This joint work with J. Rousseau [**A**11] is motivated by the analysis of an electrical network through time.

1.1 Context and model

Context Assume that the electrical network is composed of a cable (of constant length d) and accessories (such as joints, etc). We observe the evolution of the network and more precisely the sequences of incidents (failures) taking place either on the cable itself or on the accessories. When an incident takes place on the cable, it is repaired by exchanging the damaged part (very small) of the cable by a new piece of cable, connected to the remaining network by two accessories. When an incident takes place on an accessory, a small part of the network containing the damaged accessory is removed and replaced by a new piece of cable connected to the network by two accessories. (see Figure 2.1 for a graphical illustration of the reparation process).

Let $X(t)$ be the number of accessories on the network. The cable incident rate is assumed to be proportional to the length of the cable ($\nu_1(t) = d\nu_c$) whereas the accessory incident rate is proportional to the number of accessories ($\nu_a X(t^-)$), leading to the following exposure process:

$$Y(t) = \nu_a X(t^-) + \nu_c d.$$

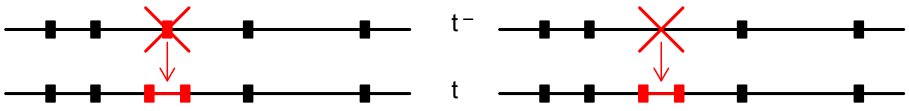


Figure 2.1 – Electrical network: the horizontal line represents a cable, each node on the line represents an accessory. On the left, **failure on an accessory** : the accessory is replaced by two of them. On the right, **failure on the cable** : a new cable is connected to the remaining network by two accessories.

ν_a and ν_c are the parameters of interest since they will allow to predict the evolution of the network in the future.

Partial observation We observe the process on a time interval $[\tau_0, \tau_0 + \tau]$. Meanwhile this time interval, we have two sources of partial observation.

1. First, in this practical context, we have access to the instants of breakdown but only partially to the types of breakdown. Indeed, the type of interventions are badly or not reported. As a consequence, we are in the situation described where the observations are restricted to the jump instants denoted T_i, the cause of the incidents (cable or accessories) are unobserved or partially observed.

2. Secondly, the number of accessories is known at the installation of the electrical network (a long time ago), but the systematic collection of the incident time starts a long time after this instant. Consequently, the number of accessories at the beginning of the observation period $X(\tau_0)$ is unknown and this quantity has to be inferred.

As a consequence, $X(t)$ is a multiplicative intensity process whose parameters of interest have to be estimated from the observation of the jumps instants collected from a truncated period excluding the initial state of the process.

Birth and death point of view and generalization Note that, calling "particle" the accessory, an incident on an accessory can be seen as the birth of one particle ($j_0 = 1$) whereas a breakdown on the cable corresponds to the immigration of two particles ($K = 1, j_1 = 2$). In [**A**11], we consider a more general model : $X(t)$ is a pure-birth processes with multi-size immigrations. More precisely, we consider a population of particles such that the particles give birth (randomly) to j_0 particles (or equivalently divides into $j_0 + 1$ particles) with rate ν_0 and immigration groups of sizes $j_1, \ldots, j_K$ arrive with respective rates $\nu_1, \ldots, \nu_K$. Let $X(t)$ be the number of particles at time t: $X(t)$ is a counting process where the exposure process $Y(t)$ is

$$Y(t) = X(t^-)\nu_0 + \sum_{k=1}^{K} \nu_k \quad \text{where} \quad X(t^-) = \lim_{s \to t, s<t} X(t)$$

$X(t^-)$ is predictable and $\nu_k(t)$ $(k = 0, \cdots, K)$ are positive constants. *The aim is then to estimate the different rate parameters $\theta = (\nu_0, \ldots, \nu_K)$ from the partial observation of the counting process over a finite period $[\tau_0, \tau_0 + \tau]$.*

1.2 Bayesian inference

Likelihood and prior distributions Let $N(t)$ be the total number of events occurring in $[\tau_0, t]$. In the following, we use the following notation: $N^\star := N(\tau_0 + \tau)$, i.e. $N^\star$ is the total number of events occurring in the observation period. For every $k = 1, \dots K$, we denote by $N_k(t)$ the number of immigration events of size j_k occurred in $[\tau_0, t]$ and $N_0(t)$ is the number of birth events occurred in $[\tau_0, t]$. Obviously $N(t) = N_0(t) + N_1(t) + \cdots + N_K(t)$. $\{N_{0:K}(t), \tau_0 \leq t \leq \tau_0 + \tau\}$ is a multivariate counting process with multiplicative intensity $(\nu_0 X(t^-), \nu_1, \dots, \nu_K)$ where $X(t^-) = \lim_{s \to t, s<t} X(t)$, $X(t)$ is the number of particles at time t and

$$X(t) = X(\tau_0) + \sum_{k=0}^{K} j_k N_k(t) \tag{2.1}$$

Let $T_1, \dots, T_{N^\star}$ be the occurrence times of the events during the observation period $[\tau_0, \tau_0 + \tau]$. Let Z_i be a discrete variable representing the type of the i-th event : $Z_i \in \{0, \dots, K\}$ is equal to k if the i-th event is of type k, then we have $X(T_i) = X(T_{i-1}) + j_{Z_i}$.

With these notations, the process is said to be *fully observed* if $N_{0:K}(\cdot)$ is continuously observed on $[\tau_0, \tau_0 + \tau]$, or equivalently if the total number of events $N^\star$, the time events $\{T_i\}_{i=1\dots N^\star}$ and the nature of the events $\{Z_i\}_{i=1\dots N^\star}$ are observed.

In the fully observed setup, the likelihood is (see Andersen et al., 1993):

$$\mathcal{L}(\mathbf{D}; \theta, X(\tau_0)) = \prod_{k=0}^{K} \nu_k^{N_k(\tau_0+\tau)} \prod_{i=1}^{N^\star} X(T_{i-1})^{\mathbb{I}_{Z_i=0}} \times \exp\left[-\nu_0 \sum_{i=1}^{N^\star} (T_i - T_{i-1}) X(T_{i-1}) - \nu_\bullet \tau\right] \tag{2.2}$$

here $T_0 = \tau_0$, $T_{N^\star+1} = \tau + \tau_0$ and $\nu_\bullet = \sum_{k=1}^{K} \nu_k$.

We set prior distributions on the ν_k:

$$\nu_k \sim \Gamma(\alpha_k, \beta_k), \quad \forall k = 0, \dots K.$$

Remark 1.1. It is easy to notice that in this case of a complete observation of the process, the model is conjugate and the Gamma posterior distributions of (ν_0, ν_1, ν_K) are easy to calculate.

Estimation from the partial observation of the process We now consider the case where we partially observe the process: more precisely, we observe all the instants of occurrences $T_{1:N^\star}$ and partially the types of the events $(Z_j)_{j=1\dots N^\star}$. Let $\mathbf{Z}$ denote $(Z_1, \dots, Z_{N^\star})$. We introduce n_{nobs} and n_{obs} the numbers of non-observed and observed event types respectively. Let $\mathbf{Z}_{nobs}$ be the vector composed of the non-observed Z_i's and $\mathbf{Z}_{obs} = \mathbf{Z} \setminus \mathbf{Z}_{nobs}$.

We also assume that $X(\tau_0)$ is unknown and has to be estimated. We first prove that the parameters can be identified.

Identifiability of the model The likelihood of the observations $\mathbf{D} = (N^\star, T_{1:N^\star}, \mathbf{Z}_{obs})$ is

$$\mathcal{L}(\mathbf{D}; \theta, X(\tau_0)) \quad = \quad \sum_{\mathbf{z} \in \{0,\dots,K\}^{n_{nobs}}} \mathcal{L}(N^\star, T_{1:N^\star}, \mathbf{Z}_{obs}, \mathbf{z}; \theta, X(\tau_0)) \tag{2.3}$$

Proposition 1. *Let $(\theta, X(\tau_0))$ and $(\theta', X'(\tau_0))$ be two sets of parameters such that for any partial dataset $\mathbf{D} = (N^\star, T_{1:N^\star}, \mathbf{Z}_{obs})$,*

$$\mathcal{L}\left(\mathbf{D}; \theta, X(\tau_0)\right) = \mathcal{L}\left(\mathbf{D}; \theta', X'(\tau_0)\right)$$

Then $\theta = \theta'$ and $X(\tau_0) = X'(\tau_0)$.

Interestingly, even if $n_{obs} = 0$, i.e. if none of the types of events are observed, the parameter θ can still be identified. The result is proved in [**A**11] for the least favorable case when $n_{obs} = 0$.

Prior derivation on $X(\tau_0)$ Since $X(\tau_0)$ has a strong influence on the inference, the choice of its prior π is a key issue. A first solution is to propose a uniform distribution on $\{x(\tau_0)^-, \dots, x(\tau_0)^+\} \subset \mathbb{N}$: $X(\tau_0) \sim \mathcal{U}_{\{x(\tau_0)^-,\dots,x(\tau_0)^+\}}$ where $x(\tau_0)^-$ and $x(\tau_0)^+$ are elicited.
An alternative is to use the probabilistic structure of the counting process $N_{0:K}$ to construct a coherent prior distribution on $X(\tau_0)$. It is often the case (see for instance linear assets, as in our motivating example based on the electrical network) that although $X(\tau_0)$ is not known, the state of the network at its installation –several decades prior to the beginning of the study at time τ_0– is known. When the observation period starts, the system has evolved until a certain number $X(\tau_0)$ of particles. *As a consequence we propose to derive the prior distribution on $X(\tau_0)$ from the asymptotic distribution of the number of particles.* This asymptotic distribution is given in Theorem 5.

Theorem 5. *Let $X(t)$ be the number of particles issued from the pure birth multi-immigration process described previously. We assume that $X(0) = x_0$ and the following two conditions:*

- *(i) $\forall k = 1, \dots, K$, $j_k/j_0 = r_k \in \mathbb{N}^*$.*
- *(ii) For all $k \geq 1$ $\nu_k(t) = \nu_k$ and there exists $t_1 > 0$ such that $\nu_0(t) = \nu_{0,1}\mathbb{I}_{t \leq t_1} + \nu_{0,2}\mathbb{I}_{t > t_1}$ with $0 < \nu_{0,1} \leq \nu_{0,2}$.*

Then setting $V_0(t) = \int_0^t \nu_0(u)du$ and $\nu_\bullet = \sum_{k=1}^K \nu_k$,

$$e^{-j_0 V_0(t)} X(t) \xrightarrow[t \to \infty]{\mathcal{L}} \Gamma\left(\frac{x_0}{j_0}, \frac{1}{j_0}\right) + \sum_{l=0}^{r_K - 1} G_l$$

where the G_l's are independent random variables with $G_0 \sim \Gamma\left(\frac{\nu_\bullet}{\nu_{0,2} j_0}, \frac{1}{j_0}\right)$ and for $l = 1, \dots, r_K - 1$,

$$G_l \sim \sum_{j=1}^{\infty} \omega_{j,l} \Gamma\left(jl, \frac{1}{j_0}\right)$$

with $\omega_{j,l} = e^{\lambda_l} \frac{\lambda_l^j}{j!}$, $\lambda_l = \frac{\alpha_l}{l \nu_{0,2} j_0}$, $\alpha_l = \nu_\bullet$, $\forall l \in \{1, \dots r_1 - 1\}$ and $\alpha_l = \nu_l + \dots + \nu_K$, $\forall l \in \{r_{k-1}, \dots, r_k - 1\}$, $\forall k = 2 \dots K$

Theorem 5 shows that as τ_0 increases, conditionally to the ν_j's and x_0, $X(\tau_0)$'s distribution can be approximated by the product of $e^{j_0 V_0(\tau_0)}$ and the sum of infinite mixtures of Gamma random variables. Neglecting the modification of the system through time may lead to strongly biased estimation, as soon as $V_0(\tau_0) j_0$ is not negligible. For intermediate value of τ_0 it is possible to improve the approximation by re-centering the distribution using the true mean of $X(\tau_0)$ which can be deduced from the Laplace transform given in [**A**11]. We denote by π_∞^R the re-centered asymptotic distribution.

Posterior sampling for Bayesian inference Once the prior distribution has been derived from the asymptotic properties of the process, the model is not fully conjugate anymore (see equation (2.2)). As a consequence, we have to resort to a Metropolis-Hastings algorithm. The proposal distributions on $X(\tau_0)$ and $(\nu_0, \dots, \nu_K)$ can be found in [**A**11] and have proved their efficiency on the simulation study.

1.3 Numerical studies

We give here an insight of the large simulation study performed in [**A**11]. All the numerical experiments take place in the reliability framework described in Section 1.1.

Influence of the non-observation of Z First, assuming that the initial state of the process $X(\tau_0)$ is known, we illustrate the influence of the non-observation of $\mathbf{Z}$ on the quality of estimation of the parameters. We considered 4 scenarios with a varying rate of non-observed $\mathbf{Z}$ (0%, 33%, 66% and 100 %) Denoting by $\widehat{\nu}_a^{(m)}$ and $\widehat{\nu}_c^{(m)}$ the posterior means of ν_a and ν_c respectively associated to dataset m, we compute the relative bias and relative root mean square error. and report them in Table 2.1 in percentage. As expected, the quality of estimation decreases when the number of observations decreases but remains at a reasonable level.

%	of non-observed $\mathbf{Z}$	0%	33%	66%	100%
ν_a	Relative Bias (%)	-0.85	-0.99	-1.46	-3.36
	RMSE (%)	6.58	7.14	8.31	8.66
ν_c	Relative Bias (%)	-2.12	-3.09	-1.47	4.76
	RMSE (%)	12.34	14.06	18.48	11.48

Table 2.1 – Simulation study 1 ($X(\tau_0)$ known and fixed): relative bias and RMSE (in percentage) for $\widehat{\nu}_a$ and $\widehat{\nu}_c$ in the 4 scenarios

Estimation of $X(\tau_0)$ and θ We also focus on the inference of $X(\tau_0)$, either fixing or estimating it using either a uniform prior distribution $\mathcal{U}_{\{100\cdots1000\}}$ or using the recentered asymptotic distribution as a prior π_∞^R. We study how the strategy influences the estimation of ν_a and ν_c.

In Figure 2.2, we plot the posterior densities of ν_a (upper) and ν_c (bottom) for one arbitrarily chosen dataset. As expected, $X(\tau_0)$ does not influence the posterior distribution of ν_c and the posterior densities corresponding to the 4 scenarios nearly overlap. On the contrary the posterior density for ν_a clearly depends on $X(\tau_0)$. If $X(\tau_0)$ is under-evaluated (scenario 1), the posterior density of ν_a (dashed line) is shifted to the right. When a prior on $X(\tau_0)$ is considered, the re-centered asymptotic prior distribution clearly outperforms the uniform prior distribution.

[**A**11] presents additional simulation studies highlighting the fact that our ad-hoc prior distribution performs well. Many extension of the model can be considered and are discussed in [**A**11]. In the next work, we are interested in the non-parametric intensity estimation of Aalen multiplicative intensity models.

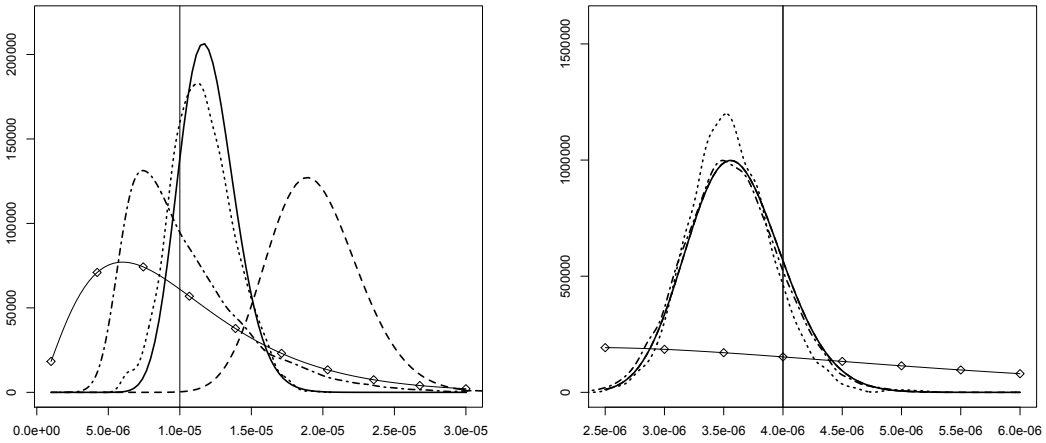


Figure 2.2 – Influence of the non-observation of $X(\tau_0)$ on the posterior distributions of ν_a (left) and ν_c (right) for one dataset: prior distribution (plain line with diamonds), posterior distribution with the true $X(\tau_0)$ (Scenario 0) (plain line), posterior distribution with under-evaluated $X(\tau_0)$ (Scenario 1) (dashed line), posterior distribution with a uniform prior distribution on $X(\tau_0)$ (Scenario 2) ($\cdot - \cdot$) and posterior distribution with asymptotic prior distribution on $X(\tau_0)$ (dotted line).

2 Bayesian non-parametric inference for counting processes with Aalen multiplicative intensities [A13] [A14]

Two of my papers [**A**13] and [**A**14] were written with V. Rivoirard, J. Rousseau and C. Scricciolo in non-parametric Bayesian framework. These papers include theoretical results on asymptotic results of the posterior distribution and numerical experiments requiring the development of adapted and complex algorithms. These works are presented here in a unified framework.

[**A**13] studies the concentration properties of the posterior distribution around the true parameter for Aalen multiplicative intensities models. Several families of non-parametric prior distributions are studied (Dirichlet Process Mixtures, Log-spline and log-linear priors). All the prior distributions considered in [**A**13] are assumed to be data-free. However, it is of common usage to choose hyperparameters depending on the data, thus resulting into empirical Bayes. Paper [**A**14] studies the properties on concentration of the posterior distribution in several non-parametric models with Dirichlet Process Mixtures (DPM) prior distributions, when the hyperparameters are data-dependent. The numerical experiments were conducted on two particular Aalen models, namely the inhomogeneous Poisson process and the right-censoring model. In both cases, we aimed at estimating the intensity function. While, the density estimation with DPM has been widely studied in the literature, the estimation of intensity function is less addressed.

In the following, we first present the Aalen multiplicative intensity model and the two particular models we treat hereafter (Section 2.1). Section 2.2 supplies a short introduction to Dirichlet process mixtures and sets the prior distribution used in the numerical experiments. In Section 2.3, I give some concentration results derived in papers [**A**13] and [**A**14]. For the sake of clarity, I only give results adapted to the numerical experiments. However, much more general theorems can be found in the original papers [**A**13] [**A**14]. Posterior sampling algorithms adapted to DPM prior are evoked in Section 2.4. Finally, Section 2.5 provides some insights on the various numerical results obtained in the papers.

2.1 Introduction to Aalen multiplicative processes

Counting processes (say $N(t)$) are commonly used in various fields of applications such as medicine –see Gusto and Schbath (2005) for instance– public health biology or reliability –see Chen (2011) for instance– or more generally in risk theory, see Ogata (1999) for instance. These processes are driven by their intensity process $\lambda(t)$ such that $\mathbb{P}(\text{occurrence} \in [t, t+dt]) = \lambda(t)dt$. The most simple counting processes are homogeneous Poisson processes, whose intensity process is a constant deterministic positive number $\lambda(t) = \lambda$. A classical generalization of the homogeneous Poisson process is the inhomogeneous Poisson process whose intensity process is a positive deterministic function. Although widely used in practice and flexible, these processes are limited by the fact they do not allow for endogenous evolution of the intensity function.

Aalen multiplicative intensity models allow for such an evolution. They are point processes such that the stochastic intensity is written as $\tilde{\lambda}(t) = \lambda(t)Y_t$, where λ is a non-negative deterministic function called *intensity function* and $(Y_t)_t$ is a non-negative predictable process. Informally,

$$\mathbb{E}[N([t,\, t+\ \mathrm{d}t]) \mid \mathcal{G}_{t^-}] = \mathbb{P}[N([t,\, t+\ \mathrm{d}t]) = 1 \mid \mathcal{G}_{t^-}] = \mathbb{P}[N([t,\, t+\ \mathrm{d}t]) > 0 \mid \mathcal{G}_{t^-}] = Y_t\lambda(t)\ \mathrm{d}t. \tag{2.4}$$

where $(\mathcal{G}_t)_t$ its adapted filtration. Note that, almost surely, we have no jump of N on sets where λ or Y vanishes.

The log-likelihood at λ with respect to the filtration $(\mathcal{G}_t)_{t\geq 0}$ can be expressed as

$$\log \mathcal{L}_n(\lambda) = \int_0^T \log(\lambda(t))\ \mathrm{d}N_t - \int_0^T \lambda(t)Y_t\ \mathrm{d}t, \tag{2.5}$$

see Daley and Vere-Jones (2003) or Karr (1986).

Inhomogeneous Poisson processes and *Right-censoring models* are two examples of Aalen multiplicative intensity models.

Inhomogeneous Poisson processes Poisson processes correspond to the case where the process $(Y_t)_{t\in[0,T]}$ is equal to 1. Assume that we observe n independent Poisson processes with common intensity λ on $[0, T]$. This model is equivalent to the model where we observe a Poisson process with intensity $n \times \lambda$, so it corresponds to the case $Y_t = n$ for all $t \in [0, T]$. In this case, if $T_1, \ldots, T_{N_T}$ are the jump times of N over $[0,\, T]$, we have

$$\log \mathcal{L}_n(\lambda) = \sum_{i=1}^{N_T} \log(\lambda(T_i)) - n \int_0^T \lambda(t)\ \mathrm{d}t. \tag{2.6}$$

In this example the observations are $\mathbf{D} = (N_t)_{t\leq T}$.

Right-censoring models Right-censoring models are very popular in biomedical problems, (see, for instance, Example I.3.9 of Andersen et al. (1993) concerning the survival analysis with right-censoring of patients with malignant melanoma). We consider n patients and, for each patient i, T_i (a non-negative random variable) is the lifetime with density f that may be censored. C_i is the censoring time assumed to be independent of T_i. We face with censoring when, for instance, the patient drops out of a hospital study: the time of death is not observed, but we know that the patient was still alive when he left the study. In right-censoring models, we observe $(Z_i,\, \delta_i)$ on $[0,\, T]$, with $Z_i = \min\{T_i,\, C_i\}$ and $\delta_i = \mathbb{I}_{T_i\leq C_i}$. In this case, the processes to be considered are

$$N_t^i = \delta_i \times \mathbb{I}_{Z_i\leq t} \quad \text{and} \quad Y_t^i = \mathbb{I}_{Z_i\geq t}.$$

We assume that the vectors $(T_i, C_i)_{1\leq i\leq n}$ are i.i.d. and we denote by λ the common hazard rate of the T_i's assumed to be finite at least on $[0, T]$:

$$\lambda(t) = \frac{f(t)}{\mathbb{P}(T_1 > t)}, \quad t \in [0, T]. \tag{2.7}$$

Note that we do not force the Z_i's to be supported in $[0, T]$. Finally, consider N (respectively Y) by aggregating the n independent processes N^i's (respectively the Y^i's), so

$$N_t = \sum_{i=1}^{n} N_t^i \quad \text{and} \quad Y_t = \sum_{i=1}^{n} Y_t^i$$

and straightforward computations show that the compensator of N is $\Lambda_t = \int_0^t Y_s\lambda(s)\ \mathrm{d}s, \quad t \in [0, T]$, thus right-censoring models obey the Aalen multiplicative model.

Expressing the log-likelihood, we obtain

$$\log \mathcal{L}_n(\lambda) \quad = \quad \sum_{i=1}^{n} \delta_i \log(\lambda(Z_i)) - \sum_{i=1}^{n} \int_0^{Z_i} \lambda(t)\ \mathrm{d}t. \tag{2.8}$$

2.2 Dirichlet process mixture priori distribution

Non parametric Bayesian estimation has gained popularity in a large number of applications in statistics and machine learning (image segmentation, clustering, density estimation, etc.). For a large introduction to non-parametric Bayesian inference, we refer the reader to Hjort et al. (2010). I quickly introduce here the tools required herafter. Nonparametric statistics arises when the parameter of interest (here λ) belongs to a space $\mathcal{F}$ of infinite dimension. λ being an intensity function, we set

$$\mathcal{F} = \left\{ \lambda : \Omega \to \mathbb{R}_+ \;\middle|\; \int_\Omega \lambda(t)\ \mathrm{d}t < \infty \right\}.$$

Defining a nonparametric Bayesian model requires defining a prior probability distribution on that infinite-dimensional space. A distribution on an infinite-dimensional space $\mathcal{F}$ is a stochastic process with paths in $\mathcal{F}$. Such distributions are typically harder to define than distributions on $\mathbb{R}^d$, but stochastic process theory and applied probability supply a large arsenal of tools.

Non-parametric Bayesian estimation has been widely developed for density estimation. Although Aalen processes do not lead to independent and identically distributed observations and estimating λ is not the same as estimating a density, there are strong connections between the two problems. To emphasize these connections, for any $\lambda \in \mathcal{F}$, we introduce the following parametrization

$$\lambda = M_\lambda \times \bar{\lambda}, \qquad \text{with} \qquad M_\lambda = \int_\Omega \lambda(t)\ \mathrm{d}t, \quad \bar{\lambda} \in \mathcal{F}_1, \quad \text{and} \quad \mathcal{F}_1 = \{\lambda \in \mathcal{F} : \int_\Omega \lambda(t)\ \mathrm{d}t = 1\}. \tag{2.9}$$

Estimating λ is equivalent to estimating M_λ and $\bar{\lambda}$ where $\bar{\lambda}$ is a density function.

M_λ being a positive constant, we will set a Gamma prior distribution on this parameter:

$$M_\lambda \sim \Gamma(a_M, b_M). \tag{2.10}$$

For the definition of a prior distribution on $\mathcal{F}_1$, Dirichlet process mixture models have become ubiquitous in Bayesian nonparametric modeling (see Müller and Mitra, 2013). Adopting a similar

strategy, we set mixture of a parametric family with a discrete random probability as a prior distribution for $\bar{\lambda}$:

$$\bar{\lambda}(\cdot) = \int_0^\infty f_\theta(\cdot)\, \mathrm{d}P(\theta) \tag{2.11}$$

with $\theta \in \Theta$. In DPM, the prior over the mixing probability P is the realization of a Dirichlet Process (DP) (Ferguson, 1973), which is a probability measure on probability measures:

$$P \mid A, G \sim \mathrm{DP}(AG) \tag{2.12}$$

where $A \in \mathbb{R}^{*+}$ and G is a probability distribution on Θ. The combination of equations (2.11) and (2.12) defines the Dirichlet Process Mixture model (DPM). One representation of the DP is the stick-breaking scheme, where a realization P is introduced explicitly as an infinite sum of atomic measures:

$$P = \sum_{k=1}^{\infty} \omega_k \delta_{\theta_k^*} \tag{2.13}$$

where $\forall k \geq 1$,

$$\theta_k^* \sim_{i.i.d.} G, \quad \text{and} \quad \omega_k = r_k \prod_{j=1}^{k-1}(1 - r_k), \quad r_k \sim_{i.i.d.} \mathcal{B}(1, A),$$

$\mathcal{B}$ denoting the Beta distribution and $\delta_{\theta_k^*}$ the Dirac delta measure located in θ_k^*. The underlying random measure P is then discrete with probability one. Using this representation, it comes that the following flexible prior model is adopted for the unknown function $\bar{\lambda}$:

$$\bar{\lambda}(t) = \sum_{k=1}^{\infty} w_k f_{\theta_k^*}(\cdot) \tag{2.14}$$

By using DPM, it is important to mention that our family of functions will be able to capture the right shape and hence statistical inference for $\bar{\lambda}$ will be improved and reliable. A drawback of using a model based on the Dirichlet process is that it is infinite dimensional and therefore inference will be complicated. However, recent innovations in sampling algorithms within infinite dimensional frameworks have lead to considerable progress in recent years to such an extent that it is now possible to perform exact inference without the need to set up arbitrary approximations in the context of density estimation. These algorithms and their adaptation to the Aalen context will be exposed hereafter.

Particular case of monotone functions In [**A**13] and [**A**14], we perform the numerical experiments on monotone non-increasing intensity for the inhomogeneous Poisson process and non-decreasing intensity for the right censoring model. To construct a prior on the set of monotone non-decreasing densities over $[0, T]$, we use their representation as mixtures of uniform densities as in Williamson (1956)

$$\bar{\lambda}(\cdot) = \int_0^\infty \frac{\mathbb{I}_{(T-\theta, T)}(\cdot)}{\theta}\, \mathrm{d}P(\theta), \qquad P \mid A, G \sim \mathrm{DP}(AG)., \tag{2.15}$$

where $\theta \in [0, T]$. Equivalently, to construct a prior on the set of monotone non-increasing densities over $[0, T]$, we use the following representation:

$$\bar{\lambda}(\cdot) = \int_0^\infty \frac{\mathbb{I}_{(0, \theta)}(\cdot)}{\theta}\, \mathrm{d}P(\theta), \qquad P \mid A, G \sim \mathrm{DP}(AG)., \tag{2.16}$$

In both models, G is chosen to be a translated inverse Gamma distribution:

$$G(\cdot) = G_{a,\gamma}(\cdot) = \left(\frac{1}{T} + \frac{1}{\Gamma(a,\gamma)}\right)^{-1} \tag{2.17}$$

Note that other distributions G can be considered, but this one verifies the assumptions required to guaranty the asymptotic concentration ([**A**13],[**A**14]). Besides, it presents computational advantages due to conjugacy.

2.3 Theoretical results

Bayesian analysis starts with a prior distribution on the parameter space and this prior distribution is updated into the posterior distribution given the data **D**. As a consequence, it is of utmost importance to ensure that the updated knowledge on the parameter is more and more accurate as the number of collected data becomes infinite. Such consistency results have been the subject of a huge literature, first in the parametric context and secondly in the semi-parametric and non-parametric fields. A large overview of this research field is provided in Ghosal and van der Vaart (2017)

Papers [**A**13] and [**A**14] provide sufficient conditions for assessing contraction rates of posterior distributions in various contexts. Judging that it would have been too burdensome to present the results in their general form in this manuscript, I chose to give a glimpse of the proved theoretical results on the particular models treated in the numerical sections e.g. right censoring model and inhomogeneous Poisson process with monotone intensity and DPM prior distributions.

2.3.1 Contraction rates of posterior distributions of intensities for Aalen models and DPM prior

[**A**13] provide sufficient conditions for assessing contraction rates of posterior distributions of intensities in general Aalen models on a compact observation time interval $[0, T]$. The derivation of asymptotic results requires conditions on the predictable process Y_t and on the prior distribution.

Assumptions on Y_t Let the true intensity λ_0 to be estimated be such that $\int_0^T \lambda_0(t)\mathrm{d}t < \infty$. We denote by $\mathbb{P}_{\lambda_0}^{(n)}$ (resp. $\mathbb{E}_{\lambda_0}^{(n)}$) the probability measure (resp. the expectation) associated with λ_0. Define

$$\mu_n(t) := \mathbb{E}_{\lambda_0}^{(n)}[Y_t] \quad \text{and} \quad \tilde{\mu}_n(t) := \frac{1}{n}\mu_n(t). \tag{2.18}$$

[C1] We assume the existence of a non-random set $\Omega \subseteq [0, T]$ such that there are positive constants m_1 and m_2 satisfying for any n,

$$m_1 \leq \inf_{t\in\Omega} \tilde{\mu}_n(t) \leq \sup_{t\in\Omega} \tilde{\mu}_n(t) \leq m_2, \tag{2.19}$$

[C2] There exists $\alpha \in (0, 1)$ such that, if

$$\Gamma_n := \left\{\sup_{t\in\Omega} |n^{-1}Y_t - \tilde{\mu}_n(t)| \leq \alpha m_1\right\} \cap \left\{\sup_{t\in[0,T]\setminus\Omega} Y_t = 0\right\},$$

then

$$\lim_{n\to\infty} \mathbb{P}_{\lambda_0}^{(n)}(\Gamma_n) = 1. \tag{2.20}$$

[C3] For some $k \geq 1$ there exists a constant $C_{1k} > 0$ such that

$$\mathbb{E}_{\lambda_0}^{(n)} \left[\left(\int_\Omega [Y_t - \mu_n(t)]^2 \, \mathrm{d}t \right)^k \right] \leq C_{1k} n^k. \tag{2.21}$$

[C1], **[C2]** and **[C3]** are the general conditions required for the demonstration. **[C1]** and **[C2]** allow to control quite precisely the number of jumps of the process N on subsets of Ω. In particular, the number of jumps of N is bounded by the number of jumps of a Poisson process with intensity $n\lambda(\cdot)$. This trick allows us to use the classical machinery for density estimation developed by Ghosal and van der Vaart (2007) in the density setting. For inhomogeneous Poisson processes, conditions **[C1]** and **[C2]** are obviously satisfied with $m_1 = m_2 = 1$ and $\Omega = [0, T]$ since for any $t \in [0, T]$, $Y_t = \mu_n(t) = n$. It is also verified for right-censoring models with different values for Ω depending on the support of the Z_i's (see page 39).

Assumptions on the prior distribution We now set the following assumptions on the prior distribution, adapted to the DPM context.

[C4a] The prior on λ is the one defined in (2.15): $\bar{\lambda}(\cdot) = \int_0^\infty \frac{\mathbb{I}_{(T-\theta, T)}(\cdot)}{\theta} \, \mathrm{d}P(\theta)$ with $P \mid A, G_\gamma \sim \mathrm{DP}(AG_\gamma)$ where G_γ is a distribution on $[0, T]$ having density g_γ with respect to Lebesgue measure (γ being a fixed hyperparameter).

[C4b] The prior on λ is the one defined in (2.16): $\bar{\lambda}(\cdot) = \int_0^\infty \frac{\mathbb{I}_{(0, \theta)}(\cdot)}{\theta} \, \mathrm{d}P(\theta)$ with $P \mid A, G_\gamma \sim \mathrm{DP}(AG_\gamma)$ where G_γ is a distribution on $[0, T]$ having density g_γ with respect to Lebesgue measure (γ being a fixed hyperparameter).

[C5] As in Salomond (2014), assume that there exist a_1, $a_2 > 0$ such that

$$\theta^{a_1} \lesssim g_\gamma(\theta) \lesssim \theta^{a_2} \quad \text{for all } \theta \text{ in a neighbourhood of } 0. \tag{2.22}$$

where g_γ is the density function of G_γ. Note that this condition is checked by the inverse Gamma distribution defined in (2.17).

Concentration result The following result holds.

Theorem 6. *Assume that the counting process N verifies conditions **[C1]**, **[C2]** and **[C3]**. Consider a prior π_1 on $\bar{\lambda}$ satisfying conditions **[C4a]** (resp. **[C4b]**) and **[C5]**. Assume that the prior π_M on M_λ is absolutely continuous with respect to Lebesgue measure with positive and continuous density on $\mathbb{R}_+$, independent of $\bar{\lambda}$. Suppose that λ_0 is monotone non-increasing (resp. non-decreasing) and bounded on $[0, T]$. Let $\bar{\epsilon}_n = (n/\log n)^{-1/3}$. Then, there exists a constant $J_1 > 0$ such that*

$$\mathbb{E}_{\lambda_0}^{(n)}[\pi(\lambda : \ \|\lambda - \lambda_0\|_1 > J_1 \bar{\epsilon}_n \mid \mathbf{D})] = 1 - \mathbb{P}_{\lambda_0}^{(n)}(\Gamma_n) + O((\log n)^k (n\bar{\epsilon}_n^2)^{-k})$$

where $\|\cdot\|_1$ is the $\mathbb{L}_1$-norm.

Note that $1 - \mathbb{P}_{\lambda_0}^{(n)}(\Gamma_n) = 0$ for the inhomogeneous Poisson model and decreases at exponential rate to 0 for the right censoring context.

Comments on the general results [**A**13] states the same type of concentration inequalities for other prior distributions such as log-spline and log-linear priors. The general conditions on the prior distributions can be found in [**A**13]. The same kinds of results have been first obtained by Belitser et al. (2015) for inhomogeneous Poisson processes. The theorem in [**A**13] is proved for Aalen multiplicative counting processes *in general*.

2.3.2 Contraction rates of posterior distributions for data dependent prior

Consider a statistical model $(\mathbb{P}_\theta^{(n)} : \theta \in \Theta)$ on a sample space $\mathcal{X}^{(n)}$, together with a family of prior distributions $(\pi(\cdot \mid \gamma) : \gamma \in \Gamma)$ on a parameter space Θ. A Bayesian statistician would either set the hyper-parameter γ to a specific value γ_0 or integrate it out using a probability distribution for it in a hierarchical specification of the prior for θ. Both approaches would lead to prior distributions for θ that do not depend on the data. However, it is often the case that knowledge is not *a priori* available to either fix a value for γ or elicit a prior distribution for it, so that a value for γ may be more easily chosen using the data.The prior is then data-dependent and the approach falls under the umbrella of empirical Bayes methods, as opposed to fully Bayes methods.

Throughout the paper, we will denote by $\hat{\gamma}_n$ a data-driven choice for γ. There are many instances in the literature where an empirical Bayes choice for the prior hyper-parameters is performed, sometimes without explicitly mentioning it. Some examples concerning the parametric case can be found in Ahmed and Reid (2001), Berger (1985) and Casella (1985). In [**A**14], contraction rates of posterior distributions for data dependent prior are proved. As before, I do not expose here the general theorem obtained in [**A**14] but give an insight of the one obtained for the Aalen multiplicative processus with DPM prior on the normalized intensity function $\bar{\lambda}$.

We define the following conditions.

[C6] Assume that G_γ is one on these two distributions:

$$g_\gamma(\theta) \propto \gamma^a \theta^{a-1} e^{-\theta\gamma} \mathbb{I}_{0\leq\theta\leq T} \quad \text{or} \quad \left(\frac{1}{\theta} - \frac{1}{T}\right)^{-1} \sim \text{Gamma}(a, \gamma)$$

[C7] Assume that $\hat{\gamma}_n$ (the data-driven choice for γ) is a measurable function of the observations satisfying $\mathbb{P}_{\lambda_0}^{(n)}(\hat{\gamma}_n \in \mathcal{K}) = 1 + o(1)$ for some fixed compact subset $\mathcal{K} \subset (0, \infty)$.

Thus, we have the following result.

Theorem 7. *Assume that the counting process N verifies conditions **[C1]**, **[C2]** and **[C3]**. Consider a prior π_1 on $\bar{\lambda}$ satisfying conditions **[C4a]** (resp. **[C4b]**) and **[C6]** with $\hat{\gamma}_n$ as hyperparameter. Assume that $\hat{\gamma}_n$ verifies **[C7]**. Assume that the prior π_M for the mass M_λ is absolutely continuous with respect to Lebesgue measure, with positive and continuous density on $\mathbb{R}^+$, and has finite Laplace transform in a neighborhood of 0. Suppose that λ_0 is monotone non-increasing (resp. non-decreasing) and bounded on $[0, T]$. Let $\bar{\epsilon}_n = (n/\log n)^{-1/3}$. Then, there exists a sufficiently large constant $J_1 > 0$ such that:*

$$\mathbb{E}_{\lambda_0}^{(n)}[\pi(\lambda : \ \|\lambda - \lambda_0\|_1 > J_1\bar{\epsilon}_n \mid \hat{\gamma}_n, \mathbf{D})] = o(1)$$

and

$$\sup_{\gamma\in\mathcal{K}} \mathbb{E}_{\lambda_0}^{(n)}[\pi(\lambda : \ \|\lambda - \lambda_0\|_1 > J_1\bar{\epsilon}_n \mid \gamma, \mathbf{D})] = o(1).$$

where $\|\cdot\|_1$ is the $\mathbb{L}_1$-norm.

Comments As observed in [**A**13], condition **[C3]**, is quite mild and is satisfied for inhomogeneous Poisson processes, censored data and Markov processes. Notice that the concentration rate $\bar{\epsilon}_n$ of the empirical Bayes posterior distribution is the same as that obtained by Salomond (2014) for the fully Bayes posterior. Up to a $(\log n)$-factor, this is the minimax-optimal convergence rate over the class of bounded monotone non-increasing (decreasing) intensities.

In the simulation study of Section 2.5.1, a moment type estimator γ_n has been considered which converges almost surely to a fixed value, so that $\mathcal{K}$ is a fixed interval around such value.

2.4 Algorithmic developments

In this section, we give insights of the algorithmic tools required to sample from the posterior distribution in case of a DPM prior on the normalized intensity. The difficulty comes from the countably infinite representation of P in (2.13). For inhomogeneous Poisson process, we highlight the fact that, conditionally to $N(T)$, estimating the intensity is equivalent to the estimation of a density function of the $T_1, \dots, T_{N(T)}$. As a consequence, we can use standard algorithms specially designed for this case. We adopted the algorithm proposed by Fall and Barat (2014) and described in section 2.4.1. For the right censoring model, we can not resort to existing algorithms and have to develop an ad-hoc one. Our algorithm is described in Section 2.4.2.

Hyperparameters The prior distribution previously defined depends on hyperparameters a_M, b_M, A, a and γ. Depending on the numerical experiments, we chose to set them (at a data-dependent value or not) or to adopt a hierarchical strategy, setting an hyperprior distribution on A or γ.

2.4.1 MCMC algorithm for inhomogeneous Poisson process with DPM prior

Using the expression of the likelihood (2.6), the posterior distribution on $(M_\lambda, \bar{\lambda})$ is

$$\begin{aligned} p(M_\lambda, \bar{\lambda}|\mathbf{D}) &\propto \mathcal{L}_n(\lambda)\pi(M_\lambda, \bar{\lambda}) \\ &\propto M_\lambda^{a_M+N(T)-1} e^{-(b_M+N(T))M_\lambda} \pi_{DPM}(\bar{\lambda}) \prod_{i=1}^{n} \bar{\lambda}(T_i). \end{aligned} \tag{2.23}$$

where $N(T)$ is the number of jumps in $[0, T]$. As a consequence of (2.23), the estimation of M_λ and $\bar{\lambda}$ can be done separately with

$$M_\lambda \mid \mathbf{D} \sim \Gamma(a_M + N(T),\, b_M + N(T)).$$

and

$$p(\bar{\lambda}|\mathbf{D}) \propto \pi_{DPM}(\bar{\lambda}) \prod_{i=1}^{N(T)} \bar{\lambda}(T_i) \tag{2.24}$$

Due to the Dirichlet Process Mixture prior, nonparametric Bayesian estimation of $\bar{\lambda}$ is more involved. However, having a look at (2.24), we notice that, in this particular case, estimating $\bar{\lambda}$ is equivalent to estimating it as the density function of the $T_1, \dots, T_{N(T)}$.

Handling with the countably infinite representation of P in (2.13) is not an easy task. Ishwaran and James (2004) resorted to an approximation, truncating the infinite sum at a deterministic value. The idea of a random truncation has been first introduced by Muliere and Tardella (1998) and introduced in MCMC sampler by Papaspiliopoulos and Roberts (2008) and in Walker (2007) using the slice sampler strategy. Kalli et al. (2011) and Fall and Barat (2014) improved this algorithm. The idea of the slice sampler for DPM is to introduce auxiliary variables making the mixture model (2.14) conditionally finite. More precisely, we consider the stick breaking representation of $\bar{\lambda}$. Let c_i be the affectation variable of data T_i. The DPM model is written as:

$$T_i|c_i, \boldsymbol{\theta}^* \sim f_{\theta^*_{c_i}}, \quad P(c_i = k) = w_k, \forall k \in \mathbb{N}^* \quad (w_k)_{k\in\mathbb{N}^\star} \sim Stick(A), \quad (\theta^*_k)_{k\in\mathbb{N}^\star} \sim_{i.i.d} G_{a,\gamma}.$$

The slice sampler strategy consists in introducing a latent variable u_i such that the joint distribution of (T_i, u_i) is $p(T_i, u_i|\bar{\lambda}) = \sum_{k=1}^{\infty} w_k f_{\theta^*_k}(T_i) \frac{1}{\xi_k} \mathbb{I}_{[0,\xi_k]}(u_i)$ with $\xi_k = \min(w_k, \zeta)$, which can be reformulated as:

$$p(T_i, u_i|\bar{\lambda}) = \frac{1}{\zeta}\mathbb{I}_{[0,\zeta]}(u_i) \sum_{k|w_k>\zeta} w_k f_{\theta^*_k}(T_i) + \sum_{k|u_i\le w_k\le\zeta} f_{\theta^*_k}(T_i)\mathbb{I}_{[0,w_k]}(u_i) \tag{2.25}$$

Noticing the fact that $(w_k)_{k\geq 1}$ verifies $\sum_{k\geq 1} w_k = 1$ (implying $\lim_{k\to\infty} w_k = 0$), the cardinal of $\{k, w_k > \varepsilon\}$ is finite for every $\varepsilon > 0$, and the sum in (2.25) is finite.

Posterior sampling with slice strategy The Gibbs algorithm we propose is decomposed into three blocks. At each iteration, we sample

$$[1.] \quad [\bar{\lambda}, \mathbf{u}|\mathbf{D}; A, \gamma] \qquad [2.] \quad [A|\bar{\lambda}, \mathbf{u}, \mathbf{D}, \gamma] \qquad [3.] \quad [\gamma|\bar{\lambda}, \mathbf{u}, \mathbf{D}, A]$$

Remark 2.1. In case where γ or A are fixed to a given value, then the corresponding part in the algorithm is removed.

[1.] Sampling $[\bar{\lambda}, \mathbf{u}|\mathbf{D}; A, \gamma]$ is the most challenging but the strategy proposed by Fall and Barat (2014) can be easily adapted. Details are given in the arxiv version of our paper Donnet et al. (2014).

[2.] For $[A|\bar{\lambda}, \mathbf{u}, \mathbf{D}, \gamma]$, we use West (1992) to deduce the conditional distribution of A given $\bar{\lambda}, \mathbf{u}, \mathbf{D}, \gamma$. Assume that $A \sim \Gamma(a_A, b_A)$, then:

$$A|x, K_{N(T)} \sim \pi_x \Gamma(a_A + K_{N(T)}, b_A - \log(x)) + (1 - \pi_x)\Gamma(a_A + K_{N(T)} - 1, b_A - \log(x)) \quad (2.26)$$

where $K_{N(T)}$ is the current number of non-empty classes obtained after step [1.] and

$$\begin{aligned} x|A, K_{N(T)} &\sim \mathcal{B}(A+1, N(T)) \\ \frac{\pi_x}{1-\pi_x} &= \frac{a_A + K_{N(T)} - 1}{n(b_A - \log(x))}. \end{aligned}$$

[3.] Finally, if $\gamma \sim \Gamma(a_\gamma, b_\gamma)$, we have to sample from $[\gamma|\bar{\lambda}, \mathbf{u}, \mathbf{D}, A]$. We can prove that:

$$\gamma|\bar{\lambda}, \mathbf{u}, \mathbf{D}, \sim \Gamma\left(a_\gamma + aK^\star, b_\gamma + \sum_{k=1}^{K^\star} \frac{1}{(\frac{1}{\theta_k^\star} - \frac{1}{T})}\right) \quad (2.27)$$

where $K^\star$ is the total number of classes used to represent $\bar{\lambda}$ ($K > K_n$).

2.4.2 Posterior sampling algorithm for the right censoring model

Recall that for $i = 1, \ldots, n$, we observe $Z_i = \min\{T_i, C_i\}$, where $T_i \sim f(\cdot)$, T_i and C_i are independent, $C_i \in [0, 1]$ and the likelihood function is:

$$\mathcal{L}_n(\mathbf{D}; \bar{\lambda}, M_\lambda) = M_\lambda^{n^\star} \left(\prod_{i\in\mathcal{O}} \bar{\lambda}(Z_i)\right) \exp\left[-M_\lambda \sum_{i=1}^{n} \bar{\Lambda}(Z_i)\right]. \quad (2.28)$$

where $\bar{\Lambda}(t) = \int_0^t \bar{\lambda}(u)du$, $\mathcal{O} = \{i \in \{1, \ldots, n\} |\ \delta_i = 1\}$ and $n^\star = \#\mathcal{O}$. Equation (2.28) induces the fact that in this model, we can not plunge the problem into the density estimation framework anymore. The previous algorithm has to be adapted. Under the assumption of a DPM prior on $\bar{\lambda}$,

$$\bar{\lambda}(t) = \sum_{k=1}^{\infty} w_k \frac{\mathbb{I}_{(1-\theta_k, 1)}(t)}{\theta_k},$$

where $w_1 = v_1, \quad w_k = v_k \prod_{j=1}^{k-1}(1 - v_j)$, $v_k \overset{\text{i.i.d}}{\sim} \mathcal{B}(1, A)$, $\theta_k \overset{\text{i.i.d}}{\sim} G(\cdot)$, we can write:

$$\bar{\Lambda}(t) = \sum_{k=1}^{\infty} w_k F_{\mathcal{U}(1-\theta_k, 1)}(t),$$

where $F_{\mathcal{U}(1-\theta_k,1)}$ is the cumulative distribution function of a uniform distribution over $(1-\theta_k,\,1)$, leading to:

$$\mathcal{L}_n(\mathbf{D};\bar{\lambda},\,M_\lambda) = M_\lambda^{n^\star}\left(\prod_{i\in\mathcal{O}}\sum_{k=1}^{\infty} w_k \frac{\mathbb{I}_{(1-\theta_k,1)}}{\theta_k}(Z_i)\right)\exp\left[-M_\lambda\sum_{k=1}^{\infty} w_k H(\theta_k)\right], \tag{2.29}$$

where

$$H(\theta_k) = \sum_{i=1}^{n} F_{\mathcal{U}(1-\theta_k,1)}(Z_i).$$

In (2.29), two infinite sums have to be handled. In [**A**13], we propose the following strategy.

- On the one hand, we suggest to introduce a deterministic truncation R to approximate $\sum_{k=1}^{\infty} w_k H(\theta_k)$, leading to the following pseudo-likelihood:

$$\bar{\mathcal{L}}_{n,R}(\mathbf{D};\bar{\lambda},M_\lambda) \;=\; M_\lambda^{n^\star}\left(\prod_{i\in\mathcal{O}}\sum_{k=1}^{\infty} w_k \frac{\mathbb{I}_{(1-\theta_k,1)}}{\theta_k}(Z_i)\right)\exp\left[-M_\lambda\sum_{k=1}^{R} w_k H(\theta_k)\right]. \tag{2.30}$$

  The effect of the deterministic truncation R is studied in the numerical illustration.

- On the other hand, we use the slice sampling strategy proposed by Walker (2007) based on the auxiliary variables $\mathbf{u} = (u_i)_{i\in\mathcal{O}}$ to deal with $\prod_{i\in\mathcal{O}}\sum_{k=1}^{\infty} w_k \frac{\mathbb{I}_{(1-\theta_k,1)}}{\theta_k}(Z_i)$ in (2.30):

$$\begin{aligned}\bar{\mathcal{L}}_{n,R}(\mathbf{u},\mathbf{D};\bar{\lambda},M_\lambda) \;=\; & M_\lambda^{n^\star}\left(\prod_{i\in\mathcal{O}}\sum_{k=1}^{\infty} w_k \frac{\mathbb{I}_{(1-\theta_k,1)}}{\theta_k}(Z_i)\frac{\mathbb{I}_{(0,\,w_k)}(u_i)}{w_k}\right)\\ & \times\exp\left[-M_\lambda\sum_{k=1}^{R} w_k H(\theta_k)\right].\end{aligned} \tag{2.31}$$

  $\bar{\mathcal{L}}_{n,R}(\mathbf{u},\mathbf{D};\bar{\lambda},M_\lambda)$ is such that its marginal expression after having integrated $\mathbf{u}$ is $\bar{\mathcal{L}}_{n,R}(\mathbf{D};\bar{\lambda},M_\lambda)$. The sequence $(w_k)_{k\geq 1}$ being stochastically decreasing the infinite sum in (2.32) only has (a.s.) a finite number of positive terms. We denote by $K_i^\star = \min\{k\in\mathbb{N}^*|\forall\, l\geq k, w_l\leq u_i\}$, $K^\star = \max\{R,(K_i^\star)_{i\in\mathcal{O}}\}$, $c_i\in\mathbb{N}^*$ the allocation variable of individual $i\in\mathcal{O}$ and $\mathbf{c} = (c_i)_{i\in\mathcal{O}}$. The augmented likelihood can then be written as

$$\begin{aligned}\widetilde{\mathcal{L}}_{n,R}(\mathbf{c},\,\mathbf{u},\,\mathbf{Z};\boldsymbol{\omega},\,\boldsymbol{\theta},\,M_\lambda) \;=\; & M_\lambda^{n^\star}\left(\prod_{i\in\mathcal{O}}\frac{\mathbb{I}_{(1-\theta_{c_i},1)}}{\theta_{c_i}}(Z_i)\frac{\mathbb{I}_{(0,\,w_{c_i})}(u_i)}{w_{c_i}}\right)\\ & \times\exp\left[-M_\lambda\sum_{k=1}^{R} w_k H(\theta_k)\right]\times\prod_k w_k^{n_k}\end{aligned} \tag{2.32}$$

  where $n_k = \#\{i\in\mathcal{O}|c_i = k\}$.

Following (2.32), the MCMC will sequentially sample

$$[1.]\quad [M_\lambda|\mathbf{u},\mathbf{c},\boldsymbol{\theta},\boldsymbol{\omega},\mathbf{D}] \qquad [2.]\quad [\boldsymbol{\theta},\mathbf{c},\mathbf{u},\boldsymbol{\omega},|M_\lambda\mathbf{D}]$$

From (2.32) and the prior distribution, we have: [1.] $M_\lambda\sim\Gamma\left(a_M+n^\star, b_M+\sum_{k=1}^{R} w_k H(\theta_k)\right)$, where H has been defined in equation (2.4.2). Step [2.] is detailed in [**A**14].

2.5 Numerical experiments

We present here a glimpse of the numerical experiments included in [**A**13] and [**A**14].

2.5.1 Inhomogeneous Poisson process : the benefit of the empirical Bayes strategy

In the context of [**A**14], the main goal of the numerical experiment is to highlight the impact of an empirical Bayes prior distribution for finite sample sizes in the case of an inhomogeneous Poisson process. We simulate datasets with the following intensity function:

$$\lambda(t) \quad = \quad \left[\cos^{-1}\Phi(t)\mathbb{I}_{[0,\,3[}(t) - \left(\frac{1}{6}\cos^{-1}\Phi(3)t - \frac{3}{2}\cos^{-1}\Phi(3)\right)\mathbb{I}_{[3,\,8]}(t)\right],$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution. We aim at estimating the λ from 3 datasets corresponding to $n = 500$, 1000, 2000, respectively. In what follows, we denote by $\mathbf{D}_n$ the dataset associated with n.

As stressed at page 45, three hyper-parameters are involved in this prior, namely, the mass A of the Dirichlet process, a and γ. The hyper-parameter A strongly influences the number of classes in the posterior distribution of $\bar{\lambda}$. In order to mitigate its influence on the posterior distribution, we propose to consider a hierarchical approach by putting a gamma prior distribution on A, thus $A \sim \text{Gamma}(a_A,\, b_A)$. In absence of additional information, we set $a_A = b_A = 1/10$, which corresponds to a weakly informative prior. We arbitrarily set $a = 2$; the influence of a is not studied in this work. We compare three strategies for determining γ in our simulation study.

Strategy 1: Empirical Bayes - We propose the following estimator:

$$\hat{\gamma}_n = \Psi^{-1}\left[\overline{W}_{N(T)}\right], \quad \overline{T}_{N(T)} = \frac{1}{N(T)}\sum_{i=1}^{N(T)} T_i, \tag{2.33}$$

where the (T_i)'s are the jump instants. Moreover,

$$\Psi(\gamma) := \mathbb{E}\left[\overline{T}_{N(T)}\right] = \frac{\gamma^a}{2\Gamma(a)}\int_{1/T}^{\infty} \frac{e^{-\gamma/(\nu-\frac{1}{T})}}{(\nu-\frac{1}{T})^{(a+1)}}\frac{1}{\nu}\,\mathrm{d}\nu,$$

$\mathbb{E}[\cdot]$ denoting expectation under the marginal distribution of N. Hence, $\hat{\gamma}_n$ converges to $\Psi^{-1}(\mathbb{E}[\overline{T}_{N(T)}])$ as n goes to infinity, thus verifying condition **[C7]** where $\mathcal{K}$ can be chosen as any small but fixed compact neighborhood of $\Psi^{-1}(\mathbb{E}[\overline{T}_{N(T)}]) > 0$.

Strategy 2: Fixed γ - In order to avoid an empirical Bayes prior, one can fix $\gamma = \gamma_0$. To study the impact of a bad choice of γ_0 on the behaviour of the posterior distribution, we choose γ_0 different from the calibrated value $\gamma^* = \Psi^{-1}(\mathbb{E}_{theo})$, with $\mathbb{E}_{theo} = \int_0^T t\bar{\lambda}_0(t)\,\mathrm{d}t$. We thus consider

$$\gamma_0 = \rho \cdot \Psi^{-1}(\mathbb{E}_{theo}), \qquad \rho \in \{0.01, 30, 100\}.$$

Strategy 3: Hierarchical Bayes - We consider a prior on γ, that is, $\gamma \sim \text{Gamma}(a_\gamma,\, b_\gamma)$. In order to make a fair comparison with the empirical Bayes posterior distribution, we center the prior distribution at $\hat{\gamma}_n$. Besides, in the simulation study, we consider two different hierarchical hyper-parameters $(a_\gamma,\, b_\gamma)$ corresponding to two prior variances. More precisely, $(a_\gamma,\, b_\gamma)$ are such that the prior expectation is equal to $\hat{\gamma}_n$ and the prior variance is equal to σ_γ^2, the values of σ_γ being specified in Table 2.2. The posterior sample is obtained with the algorithm previously described.

To compare the three different strategies used to calibrate γ, several criteria are taken into account: tuning of the hyper-parameters, quality of the estimation, convergence of the MCMC and computational time.

		Empirical		Fixed		Hierarchical		Hierarchical 2	
	n	$\hat{\gamma}_n$	CpT	$\rho\Psi^{-1}(E_{theo})$	CpT	σ_γ	CpT	σ_γ	CpT
D_{500}	483	0.4094	782.19		822.12		788.14		788.00
D_{1000}	1058	0.4398	1610.47	30×0.4302	2012.96	0.1	1559.17	0.01	1494.75
D_{2000}	2055	0.4677	3546.57		9256.71		3179.96		2770.83

Table 2.2 – Computational Time (CpT in seconds), hyper-parameters for the different strategies and datasets

		$\mathbf{D}_{500}$	$\mathbf{D}_{1000}$	$\mathbf{D}_{2000}$
d_{L_1}	Empir	0.1382	0.0596	0.0606
	Fixed	0.3114	0.2852	0.2885
	Hierar	0.2154	0.1378	0.1405
	Hiera 2	0.1383	0.0607	0.0724

Table 2.3 – $\mathbb{L}_1$-distances between the estimates and the true densities for all datasets and strategies

- In terms of tuning effort on γ, the empirical Bayes and the fixed γ approaches are comparable and significantly simpler than the hierarchical one, which increases the space to be explored by the MCMC algorithm and consequently slows down its convergence. Moreover, setting an hyper-prior distribution on γ requires to choose the parameters of this additional distribution, that is, a_γ and b_γ, and, thus, to postpone the problem, even though these second-order hyper-parameters are presumably less influential.

- In our simulation study, the computational time, for a fixed number of iterations, here equal to $N_{iter} = 30000$, turned out to be also a key point. Indeed, the simulation of $\bar{\lambda}$, conditionally on the other variables, involves an accept-reject (AR) step (see equation (B3) in Donnet et al., 2014), whose acceptance rate may drastically drop for some values of γ as illustrated in Table 2.2.

- Talking about the quality of estimation a bad choice of γ - here γ too large in strategy 2 - or a not enough informative prior on γ, namely, a hierarchical prior with large variance, has a significant negative impact on the behavior of the posterior distribution. Contrariwise, the medians of the empirical and informative hierarchical posterior distributions of λ have similar losses, as seen in Table 2.3.

As a conclusion, the empirical Bayes strategy is, as expected, efficient from a computational and accuracy point of view. Its theoretical concentration asymptotic properties reinforce its importance.

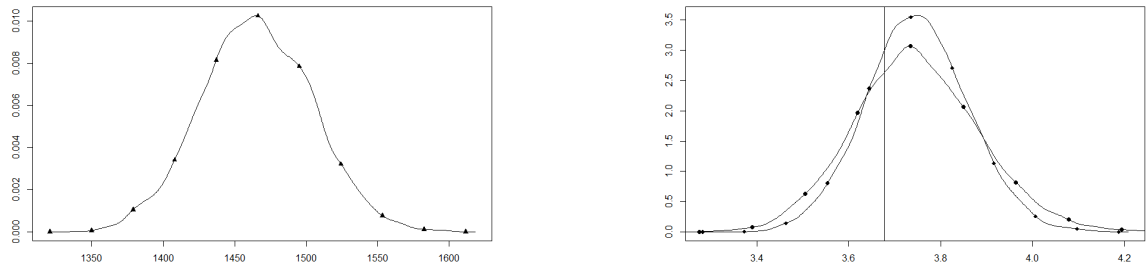
2.5.2 Numerical results for the right censoring model

We conduct a simulation study to illustrate the performances of the MCMC algorithm based on the truncation presented in Section 2.4.2.

Simulation parameters We consider the following common hazard function:

$$\lambda(t) = 2.5\left[\arctan\left(20\,t - 10\right) - \arctan\left(-10\right)\right].$$

The censoring times C_i are distributed as $C_i \overset{\text{i.i.d}}{\sim} \frac{1}{3}\,\mathcal{U}_{(0,1)} + \frac{2}{3}\,\delta_{\{1\}}$. The chosen λ and censoring time distribution ensure a censoring rate equal to 21.46 %.



(a) (b)

Figure 2.3 – *Posterior distributions of* M_λ for different values of R. (a) $R = 100$. (b) $R = 500$ and 1000.

Hyperparameters Going back to the prior distribution described in (2.10) and (2.16), we set the hyper-parameters (A, a_M, b_M, a) in the following way: $A = 15$ and $(a_M, b_M) = (4, 1)$. The choice of (a, γ) may influence a lot the inference. To avoid this problem, we propose a hierarchical strategy on a, setting $a \sim \Gamma(1, 1)$ and $\gamma = 3$.

Effect of the truncation R To study the effect of truncating with R, we have simulated one dataset with $n = 2000$, and run the MCMC algorithm with $R = 20, 80, 100, 500, 1000$. From the output in terms of the (approximation) of the posterior distribution of M_λ we observe that for $R = 500$ and $R = 1000$, the results are equivalent and the posterior distribution concentrates around the true value. Not surprisingly, for the small values of R, the estimation degenerates and the posterior distributions concentrate around aberrant values. This is shown in Figure 2.3. A finer study of this phenomena is proposed in [**A**13]. We also propose a strategy to calibrate R at a low cost.

Results With each simulated dataset, we concatenate the 5 chains to obtain a sample from the posterior distribution. For 4 of the datasets arbitrarily chosen, we plot 100 realizations of the posterior distribution of λ (Figure 2.4, on the left). Using the formula $S(t) = \exp(-\int_0^t \lambda(t)\, \mathrm{d}u)$, we also plot 100 posterior realizations of F and compare it with the true cumulative distribution function (Figure 2.4, on the right). The estimation of λ is of good quality over $[0, 0.7]$, the estimation is less accurate at the end of the interval, due to the increasing proportion of censored data. However, it corresponds to the tail of the distribution, and so this phenomenon is less noticeable on the F plots.

As a consequence, we have proposed an algorithm partially handling the countably infinite representation of the DPM and exhibiting good performances in practice.

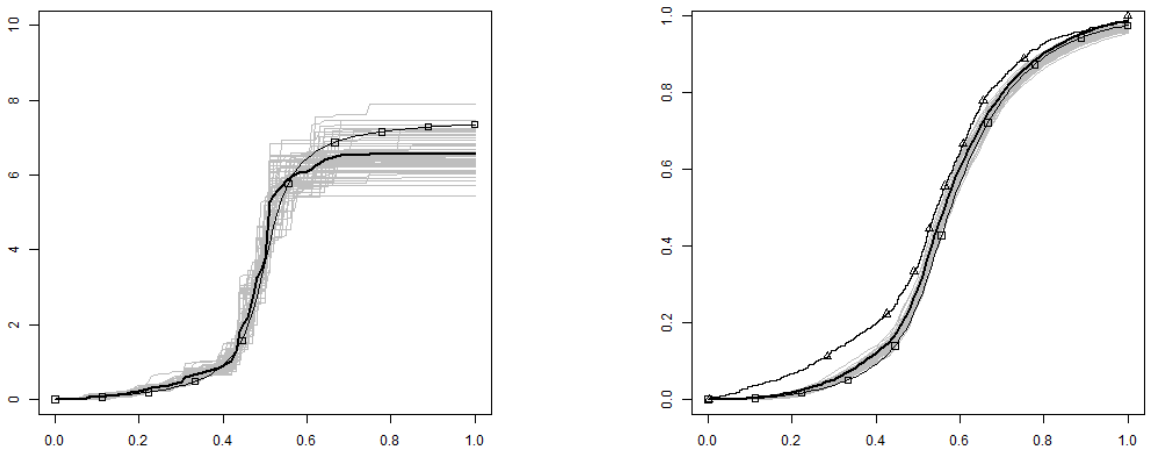


Figure 2.4 – *Posterior distributions.* For one arbitrarily chosen dataset, on the left 100 realizations (gray lines) of λ under the posterior distribution issued from the last iterations of the 5 MCMC chains; The posterior mean is plotted in plain line, the true λ is the line with squares. On the right, the corresponding curves for F: posterior simulation in gray, estimated in plain line, true F in line with square; the empirical probability function of the Z_i is the line with triangles

3 Bayesian non-parametric inference for multivariate Hawkes processes

3.1 Introduction

Hawkes processes, introduced by Hawkes (1971), are specific point processes which are extensively used to model data whose occurrences depend on previous occurrences of the same process. To describe them, we consider N a point process on $\mathbb{R}$. For any Borel set A of $\mathbb{R}$, we denote by $N(A)$ the number of occurrences of N in A. For short, for any $t \geq 0$, N_t denotes the number of occurrences in $[0, t]$. We assume that for any $t \geq 0$, $N_t < \infty$ almost surely. If $\mathcal{G}_t$ is the history of N until t:

$$\mathcal{G}_t = \sigma\left\{N(C) : \ C \in \mathcal{B}(\mathbb{R}), \ C \subset (-\infty, t]\right\},$$

then, λ, the predictable intensity of N at time t, which represents the probability to observe a new occurrence at time t given previous occurrences, is defined by

$$\lambda_t dt = \mathbb{P}(dN_t = 1 \,|\, \mathcal{G}_{t^-}),$$

where dt denotes an arbitrary small increment of t and $dN_t = N([t, t + dt])$. For the case of *univariate Hawkes processes*, we have

$$\lambda_t = \phi\left(\int_{-\infty}^{t^-} h(t - s) dN_s\right),$$

for $\phi : \mathbb{R} \mapsto \mathbb{R}_+$ left-continuous and $h : \mathbb{R} \mapsto \mathbb{R}$. We recall that the last integral means

$$\int_{-\infty}^{t^-} h(t - s) dN_s = \sum_{T \in N : T < t} h(t - T).$$

The case of *linear Hawkes processes* corresponds to $\phi(x) = \nu + x$ and $h(x) \geq 0$ for any x. The parameter ν is referred as the *spontaneous rate* and h as the *self-exciting function*. We now assume that N is a *marked* point process, meaning that each occurrence T_i of N is associated to a mark $m_i \in \{1, \ldots, K\}$ (see Daley and Vere-Jones, 2003). In this case, we can identify N with a *multivariate* point process and for any $k \in \{1, \ldots, K\}$, $N^k(A)$ denotes the number of occurrences of N in A with mark k. In the sequel, we only consider linear multivariate Hawkes processes, so we assume that λ^k, the intensity of N^k, is

$$\lambda_t^k = \nu_k + \sum_{\ell=1}^{K} \int_{-\infty}^{t-} h_{\ell,k}(t - u) dN_u^\ell,$$

where $\nu_k > 0$ and $h_{\ell,k} \geq 0$ is the interaction function of N^ℓ on N^k.

Hawkes processes have extensively been used in a large number of applications and in particular to model earthquakes, interactions in social networks, financial data, violence rates or to analyze genomes, to name but a few (see [**S2**] and references therein). In this work, we pay specific attention on the use of Hawkes processes to model neuronal activities in the same spirit as Brillinger (1988); Chornoboy et al. (1988); Okatan et al. (2005); Paninski et al. (2007); Pillow et al. (2008); Hansen et al. (2015); Reynaud-Bouret et al. (2013).

Parametric inference for Hawkes models based on the likelihood is the most classical in the literature and we refer the reader to Ogata (1988); Carstensen et al. (2010) for instance. Non-parametric estimation has first been considered by Reynaud-Bouret and Schbath (2010) who proposed a procedure based on minimization of an ℓ_2-criterion penalized by an ℓ_0-penalty for

univariate Hawkes processes. Their results have been extended to the multivariate setting by Hansen et al. (2015) where the ℓ_0-penalty is replaced with an ℓ_1-penalty. The resulting Lasso-type estimate leads to an easily implementable procedure providing sparse estimation of the structure of the underlying connectivity graph. To generalize this procedure to the high-dimensional setting, Chen et al. (2017) proposed a simple and computationally inexpensive edge screening approach, whereas Bacry et al. (2015) combine ℓ_1 and trace norm penalizations to take into account the low rank property of their *self-excitement matrix*. Other alternatives based on spectral methods (Bacry et al., 2012) or estimation through the resolution of a Wiener-Hopf system (Bacry and Muzy, 2016) can also been found in the literature. While frequentist inference for Hawkes processes has extensively been studied, the Bayes approach has received less attention. To the best of our knowledge, the only contribution for the Bayesian inference is due to Rasmussen (2013) who explored and compared two parametric approaches and used MCMC to approximate the posterior distribution of the parameters.

Our working paper is dedicated on the one hand to the study of the concentration properties of the posterior distribution for some particular non-parametric prior on the interaction functions, and on the other hand to numerical illustrations for the posterior sampling using a reversible jump algorithm. The theoretical part being still in progress, in this manuscript, I only present the second part of the work.

3.2 Non-parametric Bayesian inference

We aim at estimating the parameters of the multivariate Hawkes models, namely $((\nu_k, h_{\ell,k})_{l,k\in\{1,\dots,K\}})$. In a neurosciences context, the focus is on the interaction functions, since they drive the interactions between the neurones at stake. The general objective is to identify the groups of neurons which are independent from the others, i.e. we want as an output of the method an interaction network like the one in Figure 2.5. However, for any pair on neurones, we are interested in the duration of the influence, i.e. we want to estimate the support of the interactions functions $h_{\ell,k}$. The prior distribution we set on the $(h_{\ell,k})$ takes into account these two specificities.

Prior distribution on $(\nu_k, h_{\ell,k})_{l,k\in\{1,\dots,K\}}$ First of all, ν_k being strictly positive quantities, we set a log-normal prior distribution:

$$\log \nu_k \sim_{i.i.d} \mathcal{N}(\mu_\nu, \sigma_\nu^2), \quad \forall k = 1, \dots, K \tag{2.34}$$

with $\mu_\nu = 3$ and $\sigma_\nu^2 = 1$. About the interaction functions $(h_{\ell,k})_{\ell,k\in\{1,\dots,K\}}$, we assume that we know an upper bound of $h_{\ell,k}$'s support, denoted $[0, 0.04]$. The prior distribution is defined on the set of piecewise constant functions, $h_{\ell,k}$ being written as follows:

$$h_{\ell,k}(t) = \delta_{\ell,k} \sum_{m=1}^{M_{\ell,k}} \alpha_{\ell,k}^{(m)} \mathbb{I}_{[s_{\ell,k}^{(m-1)}, s_{\ell,k}^{(m)}]}(t) \tag{2.35}$$

with $s_{\ell,k}^{(0)} = 0$ and $s_{\ell,k}^{(M_{\ell,k})} = A$

- $\delta_{\ell,k}$ is a global parameter of nullity for $h_{\ell,k}$. For all $(\ell, k) \in \{1, \dots, K\}^2$,

$$\delta_{\ell,k} \sim_{i.i.d} \mathcal{B}ern(\pi_\delta). \tag{2.36}$$

The idea of this parameter is to encourage complete nullity for the interaction functions and so a sparse interaction network between neurons.

- For non-null intensities functions, we work in a non-parametric framework. For all $(\ell, k) \in \{1, \dots, K\}^2$, the number of steps $(M_{\ell,k})$ follows a translated Poisson prior distribution:

$$M_{\ell,k} | \{\delta_{\ell,k} = 1\} \sim_{i.i.d.} 1 + \mathcal{P}(\lambda). \tag{2.37}$$

  To minimize the influence of λ on the posterior distribution, we consider an hyperprior distribution on the hyperparameter λ:

$$\lambda \sim \Gamma(a_\lambda, b_\lambda). \tag{2.38}$$

- Given $M_{\ell,k}$, we consider a spike and slab prior distribution on $(\alpha_{\ell,k}^{(m)})_{m=1\dots M_{\ell,k}}$. Let $Z_{\ell,k}^{(m)} \in \{0, 1\}$ denote a sign indicator for each step, we set: $\forall m \in \{1, \dots, M_{\ell,k}\}$:

$$\begin{array}{rcl} \mathbb{P}(Z_{\ell,k}^{(m)} = z | \delta_{\ell,k} = 1) & = & \pi_z, \quad \forall z \in \{0, 1\} \\ \alpha_{\ell,k}^{(m)} | \delta_{\ell,k} = 1 & \sim & Z_{\ell,k}^{(m)} \times \log \mathcal{N}(\mu_\alpha, s_\alpha^2) \end{array} \tag{2.39}$$

  These spike and slab prior on the $(\alpha_{\ell,k}^{(m)})_{m=1\dots M_{\ell,k}}$ will help to identify the time intervals where the activity on neuron ℓ really generates activity on neuron k with possibly latency periods. Note that $Z_{\ell,k}^{(m)} \alpha_{\ell,k}^{(m)} \geq 0$, thus resulting into pure mutual exciting behavior. If $\mathbb{P}(Z_{\ell,k}^{(m)} = -1 | \delta_{\ell,k} = 1) \neq 0$, then we are able to handle inhibition behavior. However, it leads to computational difficulties, see the discussion on page 60.

- On $(s_{\ell,k}^{(m)})_{m=0\dots M_{\ell,k}}$, we consider two possible prior distributions. In the first one, referred as *the regular prior* in the following, $(s_{\ell,k}^{(m)})_{m=0\dots M_{\ell,k}}$ is set equal to a regular partition of $[0, A]$:

$$s_{\ell,k}^{(m)} = \frac{m}{M_{\ell,k}} A \qquad \forall m = 0, \dots, M_{\ell,k}. \tag{2.40}$$

  This prior verifies the assumptions required in the theoretical results. However, in practice, we also tried a *continuous prior* distribution on $(s_{\ell,k}^{(m)})_{m=0\dots M_{\ell,k}}$, setting :

$$\begin{array}{rcl} (u_1, \dots, u_{M_{\ell,k}}) & \sim & \mathcal{D}ir(a_s, \dots a_s) \\ s_{\ell,k}^{(0)} & = & 0 \\ s_{\ell,k}^{(m)} & = & A \sum_{r=1}^{m} u_r, \quad \forall m = 1, \dots, M_{\ell,k} \end{array} \tag{2.41}$$

  where $\mathcal{D}ir(\cdot)$ is the Dirichlet probability distribution.

Posterior sampling The posterior distribution is sampled using a standard Reversible-jump Markov chain Monte Carlo. Considering the current parameter $(\boldsymbol{\nu}, \boldsymbol{h})$, $\boldsymbol{\nu}^{(c)}$ is proposed using a Metropolis-adjusted Langevin proposal: $\boldsymbol{\nu}^{(c)} := \boldsymbol{\nu} + \tau \left[\nabla \log \pi(\boldsymbol{\nu}) + \nabla \log L(\boldsymbol{\nu}, \boldsymbol{h})\right] + \sqrt{2\tau}\xi^{(i)}$, where $\xi^{(i)} \sim \mathcal{N}(\mathbf{0}_K, \boldsymbol{I}_K)$. For a fixed $M_{l,k}$, the heights $\alpha_{l,k}^{(m)}$ are proposed using a random walk proposing null or non-null candidates. Changes in the number of steps $M_{\ell,k}$ are proposed by standard birth and death moves (Green, 1995). In this simulation study, we generate chains of length 30000 removing the first 10000 burn-in iterations.

3.3 Numerical results

In the paper, we consider three simulation scenarios involving respectively $K = 2$ and $K = 8$ neurons. In this document, I only present the $K = 8$ scenario.

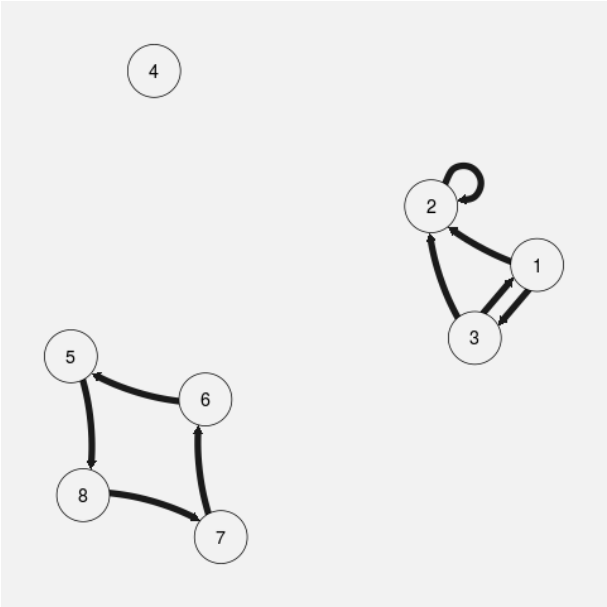


Figure 2.5 – **Scenario 2**. True interaction graph between the $K = 8$ neurones. A directed edge is plotted from vertex ℓ to vertex k if the interaction functions $h_{\ell,k}$ is non-null.

In this scenario, we mimic $K = 8$ neurons belonging to three independent groups. The non-null interactions are the piecewise constant functions defined as:

$$h_{2,1} = h_{3,1} = h_{2,2} = h_{1,3} = h_{2,3} = h_{8,5} = h_{5,6} = h_{6,7} = h_{7,8} = 30 \cdot \mathbb{I}_{(0,0.02]}.$$

In Figure 2.5, we plot the subsequent interactions directed graph between the 8 neurons: the vertices represent the K neurons and an oriented edge is plotted from vertex ℓ to vertex k if the interaction function $h_{\ell,k}$ is non-null.

Moreover, $\nu_k = 20, \forall k = 1 \dots K$. 25 datasets are simulated on the time interval $[0, 22]$ seconds. The Bayesian inference is performed considering recordings on two possible periods of length $T = 10$ seconds and $T = 20$ seconds. For any dataset, we remove the initial period of 2 seconds –corresponding to 25 times the length of the support of the $h_{\ell,k}$– assuming that, after this period, the Hawkes processes have reached their stationary distribution.

In the simulations studies, we set the following hyperparameters:

$$\begin{array}{rclrcl} \mu_\alpha & = & 3.5, & s_\alpha & = & 1 \\ \mu_\nu & = & 3.5, & s_\nu & = & 1 \\ \mathbb{P}(Z = 1) & = & 1/2, & P(\delta = 1) & = & 1/2 \\ a_s & = & 2 & & & \end{array}$$

Let us first have a look at the L_1 distances on λ^k and $h_{\ell,k}$ for all 3 the scenarios, all length observation time. In Table 2.4, we compile the estimated L_1 distances on λ^k and $h_{\ell,k}$. More precisely, we compute:

$$\frac{1}{25 * K^2} \sum_{sim,\ell,k} \widehat{\mathbb{E}}[d_{L_1}(h_{\ell,k}, h^0_{\ell,k}) | (N^{sim}_t)_{t\ in[0,T]}]$$

and

$$\frac{1}{25 * K} \sum_{sim,k} \widehat{\mathbb{E}}[d_{L_1}(\lambda^k, \lambda^{0k}) | (N^{sim}_t)_{t\ in[0,T]}]$$

where $h^0_{\ell,k}$ is the true interaction function and λ^{0k} is the "true" conditional intensity function, dependent $h^0_{\ell,k}$ and the observations $(N^{sim}_t)_{t\ in[0,T]}$. The posterior expectation is estimated by Monte Carlo using the output of the MCMC algorithms. In a few words, as expected, the error decreases as T increases. As we will detail later, the continuous prior on s gives better results that the regular prior. In **[S2]**, we also simulated data with smooth interaction functions. As expected, we perform better when the true interaction function $(h_{\ell,k})$ are step functions (with respect to smooth functions), due to the form of the prior distribution.

		K=8
L1 distances on λ^k	T=10	5.65
	T=20	3.17
L1 distances on $h_{\ell,k}$	T=10	0.1199
	T=20	0.0616

Table 2.4 – L1 distances on $h_{\ell,k}$ and λ^k

The posterior distribution of the $(\nu_k)_{k=1...K}$ for a randomly chosen dataset is plotted in Figure 2.6. The prior distribution is in dotted line and is clearly uninformative. The posterior distribution concentrate around the true value (here 20) with a smaller variance when T increases.

In a neurosciences context, we are especially interested in recovering the interaction graph of the $K = 8$ neurons. In Figure 2.7, we consider the same dataset as the one used in Figure 2.6 and plot the posterior estimation of the interaction graph, for respectively $T = 10$ on the left and $T = 20$ on the right. The width and the gray level of the edges are proportional to the estimated posterior probability $\widehat{\mathbb{P}}(\delta_{\ell,k} = 1|(N_t)_{t\ in[0,T]})$. The global structure of the graph is recovered (to be compared to the true graph plotted in Figure 2.5). We observe that the false positive edges appearing when $T = 10$ disappear when $T = 20$. In Figure 2.8, we consider the mean of the estimates of the graph over the 25 datasets. The resulting graph for $T = 10$ is on the left and for $T = 20$ on the right.

Note that, in this example, for any (ℓ, k) such that the true $\delta_{\ell,k} = 1$, the estimated posterior probability $\widehat{\mathbb{P}}(\delta_{\ell,k} = 1|(N^{sim}_t)_{t\ in[0,T]})$ is equal to 1, for any dataset and any length of observation interval. In other words, the non-null interactions are perfectly recovered. In a simulation scenario with other interaction functions, the results could have been different.

In Figure 2.9, we plot the posterior estimation (with credible regions) of the non-null interaction functions for the simulated dataset used in Figure 2.7. The time intervals where the interaction functions are null are exactly recovered. The posterior incertitude of the non-null steps $\alpha^{(m)}_{\ell,k}$ decreases when T increases.

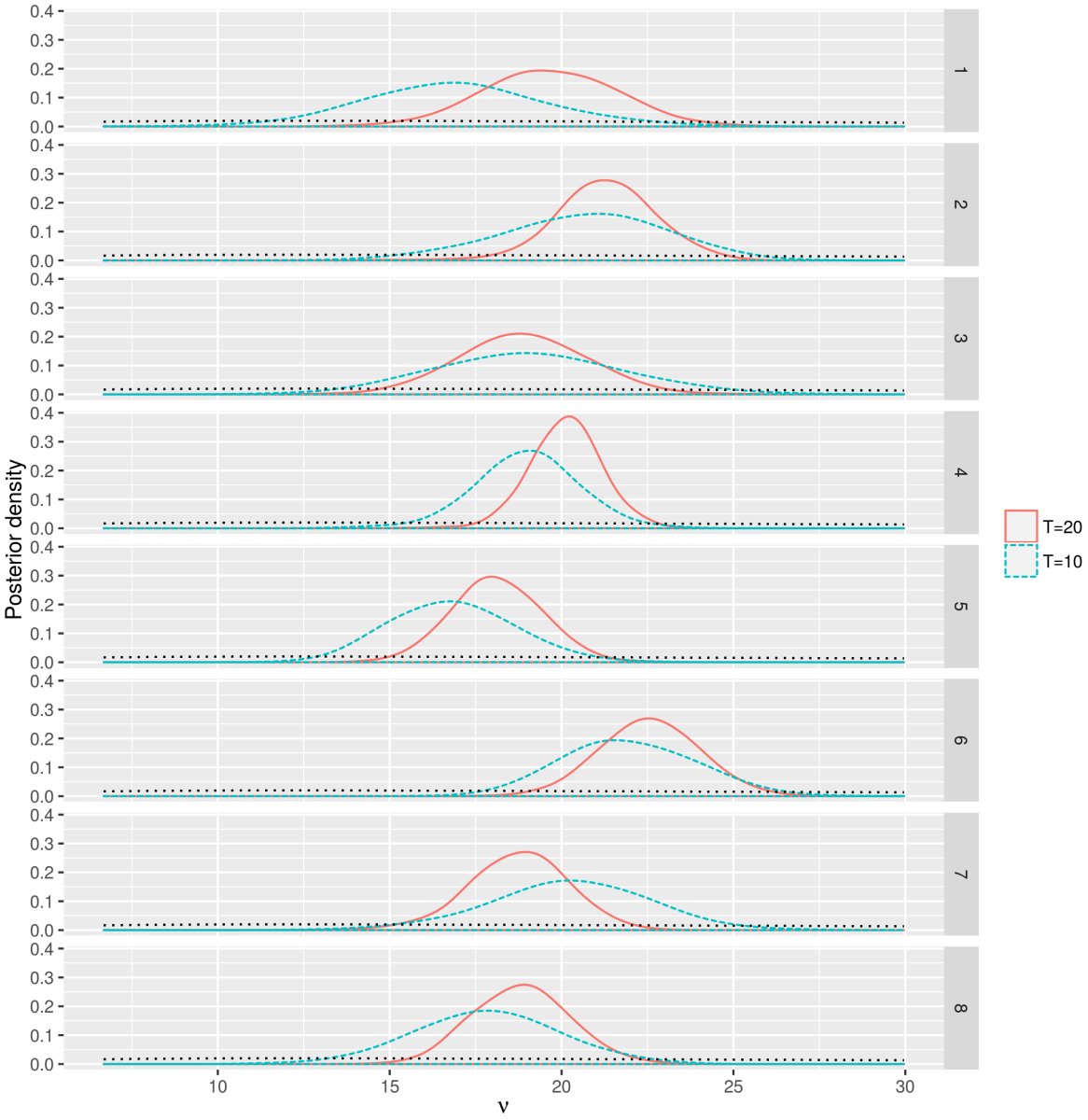


Figure 2.6 – **Scenario 2**. Results on $(\nu_k)_{k=1\ldots K}$ for a particular dataset: Prior distribution (dotted line), Posterior distributions for $T = 10$ (dashed line) and $T = 20$ (plain line).

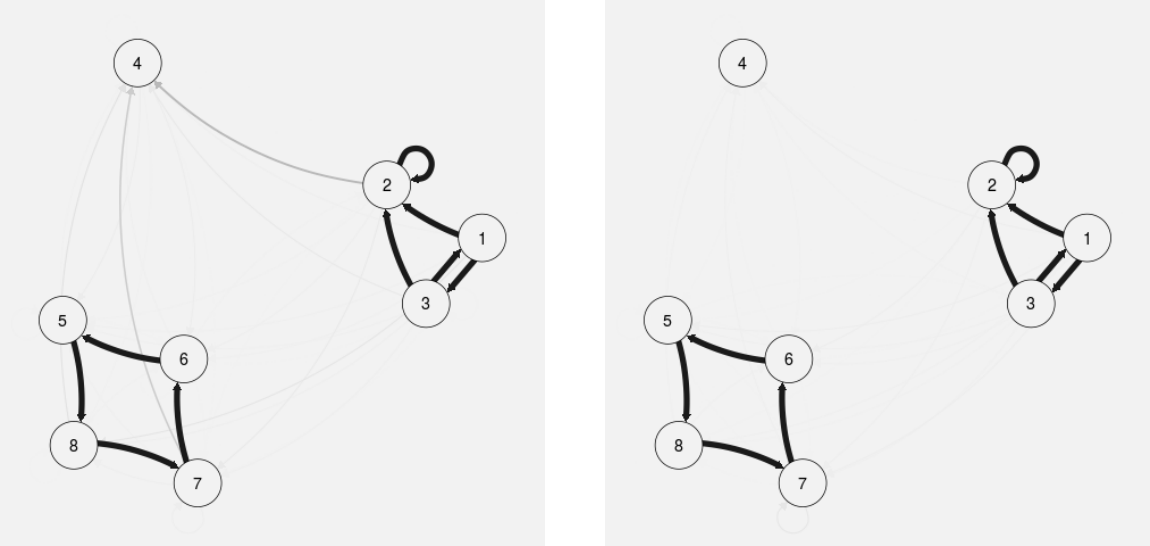


Figure 2.7 – **Results for scenario 2 for one given dataset**. Posterior estimation of the interaction graph for $T = 10$ *on the left* and $T = 20$ *on the right*, for one randomly chosen dataset. Level of grey and width of the edges proportional to the posterior estimated probability of $\widehat{\mathbb{P}}(\delta_{\ell,k} = 1|(N_t^{sim})_{t\ in[0,T]})$.

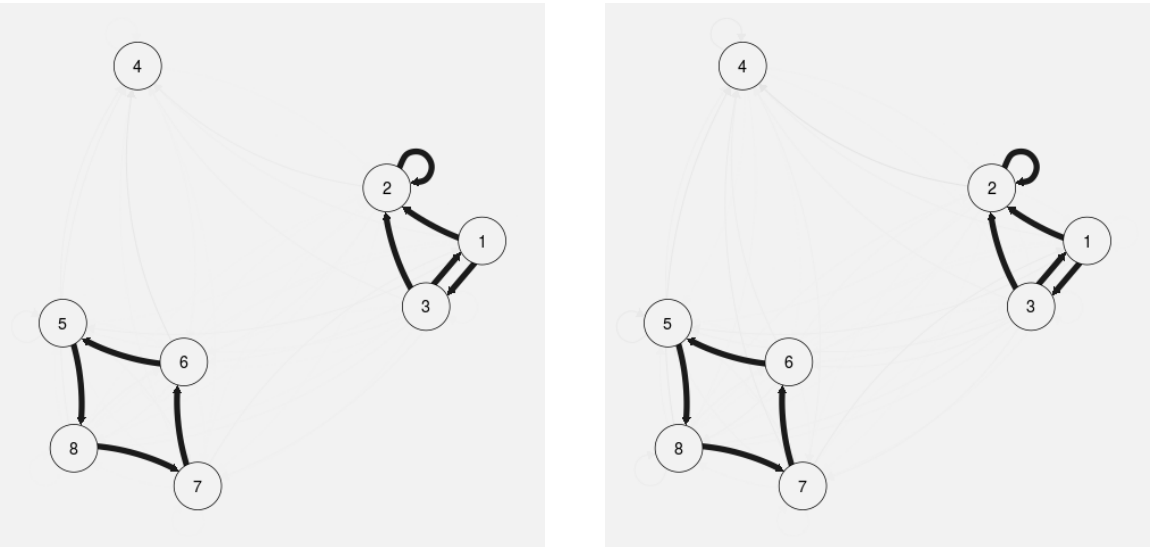


Figure 2.8 – **Results for scenario 2 over the** 25 **simulated datasets**. Posterior estimation of the interaction graph for $T = 10$ *on the left* and $T = 20$ *on the right*. Level of grey and width of the edges are proportional to the posterior estimated probability of $\frac{1}{25}\sum_{sim=1}^{25}\widehat{\mathbb{P}}(\delta_{\ell,k} = 1|(N_t^{sim})_{t\ in[0,T]})$.

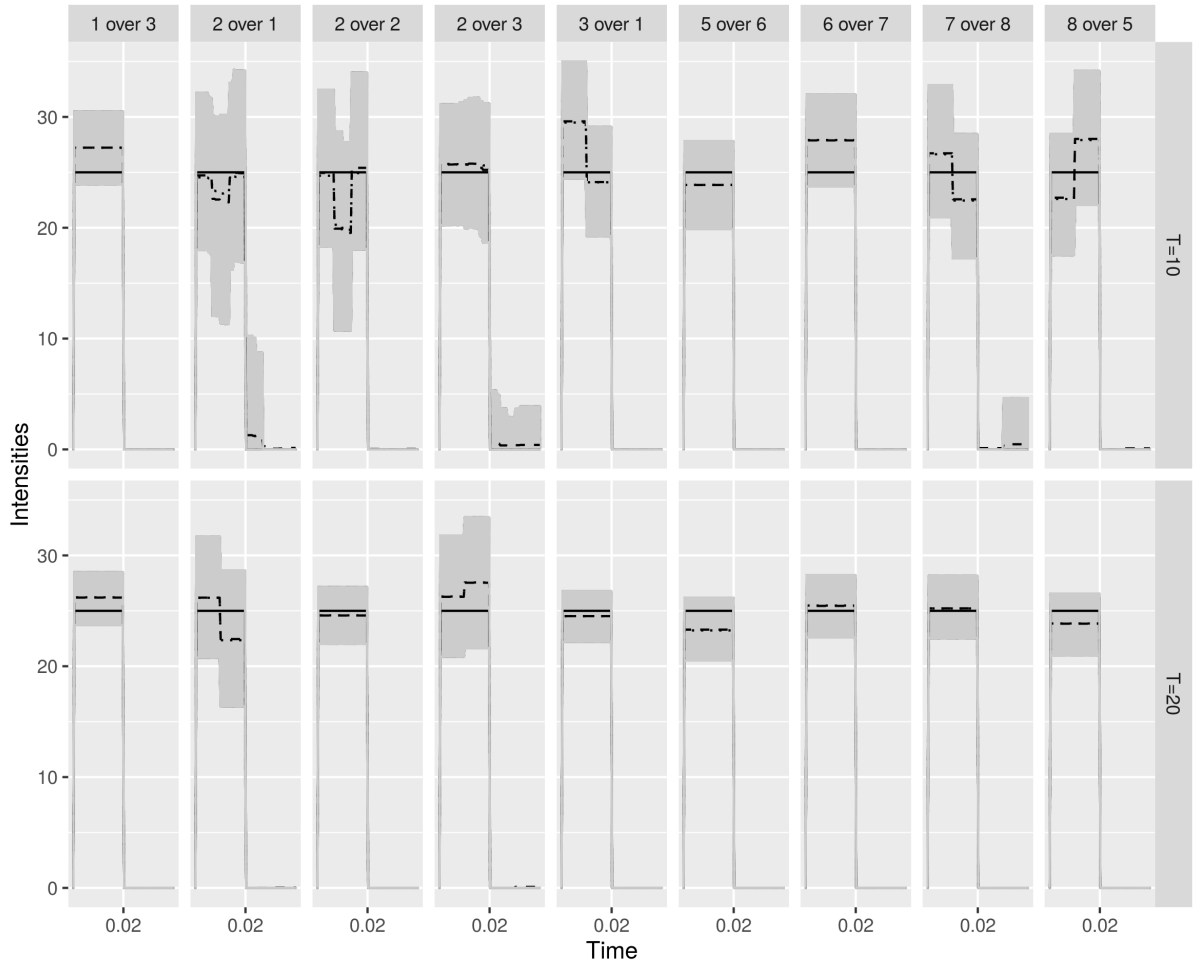


Figure 2.9 – **Results for scenario 2 for one given dataset**. Estimation of the non null interaction functions $(h_{\ell,k})_{\ell,k=1,\dots,8}$ using the regular prior for $T = 10$ (upper panel) and $T = 20$ (bottom). The gray region indicates the credible region for $h_{\ell,k}(t)$ (delimited by the 5% and 95% percentiles of the posterior distribution). The true $h_{\ell,k}$ is in plain line, the posterior expectation and posterior median for $h_{\ell,k}(t)$ are in dotted and dashed lines respectively (often undistinguishable).

4 Perspectives

This chapter presented my works on counting processes. I now give a few working perspectives.

About the modeling of electrical network through time [A11] First, in this model, we consider that the $Z_1, \dots Z_{N(\tau)}$ are partially observed. An other interesting scenario would be to consider a mis-reporting of the event types Z_j's. More precisely we observe types of events $Z_1^r \dots Z_{N(\tau)}^r$ which are reported with error, defined by a probabilistic model $P[Z^r|Z]$. Writing the new full likelihood

$$\tilde{\mathcal{L}}(N^*, (T_i, Z_i, Z_i^r)_{i=1,\cdots,N^*}, X(\tau_0)) = \mathcal{L}(N^*, (T_i, Z_i)_{i=1,\cdots,N^*}, X(\tau_0)) \times \prod_{i=1}^{N^*} P[Z_i^r|Z_i]$$

we obtain a tractable posterior distribution on $(\boldsymbol{Z}, X(\tau_0), \theta)$ which we can simulate using a Gibbs sampler.

In our description of the model and methodology, emphasize has been put in the case where the rates ν_j are constant. We explain in [**A**11] how the methodology can be extended to the case where they depend on time in a parametric way. The structure of the algorithms would remain the same, apart from possible loss in conjugacy so that Hasting-Metropolis steps within Gibbs might have to be considered in such situations, depending on the parametric form of the function $\nu_j(t;\theta)$. Depending of the form of $\nu_j(t;\theta)$, extensions of Theorem 5 to cases where the ν_k's are allowed to vary could be obtained.

An other direct extension from our model is to consider covariates which do not vary with time. In that case a hierarchical formulation of our Bayesian model can be stated as follows. Let C denote the covariate taking values in a set $\mathcal{C}$, typically $\mathcal{C}$ would be finite, then given C, define the same process $(N_C(t), X_C(t), t \in [0,T])$ with parameters $\nu_C = (\nu_{C,0}, \dots, \nu_{C,K})$, assume that the parameters ν_C are independent and identically distributed from the prior distribution proposed previously.

An easier way to consider an aging in the system is to say that after a given time $\tau^\star$, the accessories are replaced by a new type of material with their proper failure rate $\nu^\star$. In that context, we would have a multi-type counting process. Let $X^\star(t)$ denote the number of new type-accessories and $X(t)$ the number of old type accessories. After τ^*, at each event (immigration or birth) $X(t)$ decreases and $X^\star(t)$ increases conjointly. The study of that process and the estimation of the parameters would remain essentially the same as the one presented in the paper.

About the non-parametric estimation of Intensities for Aalen processes [A13] [A14]. From an algorithmic point of view, in [**A**13] [**A**14], we adapted algorithms able to handle (partially) with the countably infinite representation of P in (2.13). However, the result is not completely satisfactory since we have to introduce an arbitrary truncation for the right censoring model. Moreover the algorithms are specially designed for each model. Looking for an other slice sampler strategy adapted to the Aalen multiplicative intensities process would be a challenging but interesting perspective.

Perspectives for multivariate Hawkes processes In order to stick to the theoretical results, we restricted the simulation study to null or positive interaction functions, setting $Z_{\ell,k}^{(m)} \in \{0,1\}$ in (2.39). In practice, the methodology presented before could be easily extended to any type of interactions, i.e. $Z_{\ell,k}^{(m)} \in \{-1,0,1\}$. However, this extension could lead to additional non-negligible computational time.

Indeed, when the interaction functions are non-negative, the conditional intensity for neuron k defined as :

$$\lambda^\star(t,k) = \nu_k + \sum_{\ell=1}^{K} \int_{-\infty}^{t-} h_{\ell,k}(t-u) dN_u^{(\ell)}.$$

has to be integrated over $[0,T]$ to get the likelihood function. Since $t \mapsto \lambda^\star(t,k)$ is not regular, its integration can not be performed with a standard numerical solver. When $(h_{\ell,k})_{\ell,k\in\{1,\dots K\}}$ are piecewise constant functions, the integral can be computed in a close form with a complexity linear in the number of step sizes $(M_{\ell,k})$ and the number of occurrence times.

However, when considering non-positive interactions, the conditional intensity has to be modified to guarantee its positivity. A standard modification is the following one:

$$\lambda^\star(t,k) = \phi\left\{\nu_k + \sum_{\ell=1}^{K} \int_{-\infty}^{t-} h_{\ell,k}(t-u) dN_u^{(\ell)}.\right\}$$

where $\phi : \mathbb{R} \mapsto \mathbb{R}^+$ is a non linear function ($\phi(x) = x^2$ or $\{x\}_+$). In this context, the integration of $t \mapsto \lambda^\star(t,k)$ is an operation of larger complexity, thus implying, in practice, a possibly substantial increase of the computational time.

Moreover, at that time, for computational time reasons, we are not able to handle real neurosciences datasets. Indeed, with our non-parametric Bayesian strategy, we are able to obtain not only good estimates but also posterior variances, which is a progress with respect to Hansen et al. (2015). However, the computational time become unreasonable after 10 neurones. A solution could be sought among the deterministic approximations of the posterior distributions such as variational approximations. Linderman and Adams (2015) propose such a promising strategy after having discretized the process. It would be interesting to understand the practical and theoretical consequences of such a discretization. Also note that assessing in general the properties of variational approximations of the posterior is a hot topic at that time.

Chapter 3

Statistical inference of network datasets

As described before, my work on Hawkes processes aims at inferring a connection graph between neurons. The inference of graphs representing dependence (or independence) between stochastic variables or processes is a hot topic both from the methodological and applied points of views. Note that in a paper with Jean-Michel Marin [**A**3] we worked on the calibration of prior parameters for Gaussian Graphical Models (through an adapted Stochastic version of the EM algorithm).

In the recent years, I focused my research toward the study of network datasets. This topic is also linked to graph theory but the point of view is different. Whereas in graph/network inference, the graph is a tool to summarize the dependence between entities of interest, here I am interested in modeling network datasets (in sociology and later in ecology). This work lead to two published papers [**A**15] and [**A**16] with Avner Bar-Hen, Pierre Barbillon and Emmanuel Lazega, one submitted paper [**S**1] with Stéphane Robin and one on-going work [**P**6] with Avner Bar-Hen, Pierre Barbillon and Wesley Dattilo. This chapter is dedicated to the description of these various works.

Modeling relations between entities (individuals, plants, insects...) is a classical question in sociology or ecology. Clustering individuals according to the observed patterns of interactions allows to uncover a latent structure in the data. Stochastic block model (SBM) and Latent Block models (LBM) are popular approaches for grouping the individuals with respect to their interaction profile. These models are presented in Section 1.

Stochastic block models or latent block models include latent random variables, making their likelihood intractable. Moreover, processing large networks is computationally challenging. When talking about Bayesian inference, standard stochastic algorithms (such as MCMC or population Monte Carlo algorithms) quickly reach their limitations when the size of the network and/or the number of blocks increase. To tackle that point, deterministic approximations of the posterior distribution have been proposed (among them Variational Bayes is well suited to SBM and LBM). However, there is no theoretical guarantee for the accuracy of such algorithms. Moreover, as can be illustrated on examples, such algorithms may underestimate the posterior variance. In a joint work with Stéphane Robin [**S**1], we propose an algorithm taking advantage of the last development in Sequential Monte Carlo algorithms and deterministic approximations of the posterior distribution. This work is presented in Section 2.

From an application oriented point of view, I am interested in modeling complex networks. In [**A**15] and [**A**16], my colleagues and I are interested in the case where relationships of various types occur conjointly between the individuals. In this situation, the data are represented by

multiplex networks where more than one type edge may exist between the nodes. To handle such data, we extend the SBM to multiplex networks, thus obtaining a clustering based on more than one kind of relationship. Multiplex Stochastic Block Model arises in many situations but [**A**15] is motivated by a network of French cancer researchers. The two possible links (edges) between researchers are a direct advising connection or a connection through their labs. [**A**16] is the application of the same model and same inference method to a different dataset, where we observe at the same time relations of advice and competition between cancer researchers. This work is presented in Section 3.

Aiming at inscribing my work in an ecological framework, I started working with W. Dattilo, on multipartite ecological network. To that purpose, with my colleagues P. Barbillon and A. Bar-Hen we developed a model (and its inference methodology) for multipartite networks. From my involvement in the interdisciplinary group MIRES I identified several structures of complex networks for which no model nor inference methods have been proposed until now. These working perspectives are given in Section 4.

I now give a very short introduction to Stochastic Block Models (SBM) and Latent Block Models (LBM).

1 Stochastic and Latent block models in a few words and equations

1.1 Several application contexts but a unified mathematical framework

Networks are now standard tools in sociology and ecology. Mathematically, a network is composed of a set of nodes and a list of edges between pairs of nodes. The set of nodes and the significations of the edges are context-dependent. For the sake of clarity, I will first present in a few words the three contexts motivating my work. For each of them, I will identify the set(s) of nodes and the edges. The SBM and LBM will be introduced hereafter, using a general formulation.

- **Relations between researchers and between laboratories.** In [**A**15] and [**A**16], we study various relations between some researchers. In [**A**15], we are interested in the advice relations (who gives advices to whom, etc...) between researchers. In this case, nodes are researchers and there is an edge from node/researcher i to node/researcher i' if researcher i gives advices to researcher i'. We are also interested in the resources exchange network between their laboratories: the nodes are the laboratories and there is an edge from node j to node j' if there is a resource transfer (staff, scientific material...) from laboratory j to laboratory j'. Such relations are oriented. See page 74 for more details on this dataset.

- **Mutualistic relations between plants and animals.** In my recent work in ecology [**P**6], I am interested in the mutualistic relations between plants and animals (pollinators, seed dispersal birds or ants...). As an example, a plant specie is interacting with a given ant specie if it has been observed in interaction with the plant within a given period, in a given area. In this case, we have two sets of nodes, namely the plants on the one hand and the animals on the other hand, corresponding to two functional groups. Their is an edge between plant i and animal j if animal i has been observed on plant j. The resulting graph is a bipartite graph *i.e.* interactions only occur between plants and animals. See page 77 for more details and plots.

- **Social relations between farmers and agrobiodiversity.** Finally, at the frontier of

ecology and sociology, I study on the one hand the social relations between farmers or gardeners (for instance two farmers interact if they exchange seeds) and on the other hand I am interested in the diversity of species in their agricultural production. In this context, two sets of nodes arise, namely farmers and plants. An edge from farmer i to farmer i' represents an exchange of seeds and an edge between farmer i and plant j means that farmer i cultivates plant j. See also page 77 for more details on the dataset.

Adjacency matrices are used to represent interaction networks taking place *inside* a group of entities (see interactions between researchers, laboratories or farmers). Let $E = \{1, \dots, n\}$ be the set of the nodes and X be the corresponding adjacency matrix. $X \in \mathcal{M}_{n,n}(\{0,1\})$ is such that $X_{ij} = 1$ if their is an edge from i to j, 0 otherwise. If the relation is oriented (see examples in sociology) then X is not symmetric; otherwise, X is symmetric. Moreover, we assume that $X_{ii} \neq 0$ (no loop).

Incidence matrices are used to represent bipartite graphs, *i.e.* relations between individuals belonging to two distinct functional groups. Let $E_1 = \{1, \dots, n_1\}$ and $E_2 = \{1, \dots, n_2\}$ be the two sets nodes corresponding to functional groups 1 and 2. The *incidence matrix* $Y \in \mathcal{M}_{n_1,n_2}(\{0,1\})$ is such that $Y_{ij} = 1$ if entity i of functional group 1 interacts with entity j of functional group 2.

1.2 Stochastic block models (SBM) and Latent Block models (LBM) definitions

Stochastic block models (Nowicki and Snijders, 2001) and Latent Block models for random graphs have emerged as a natural tool to model heterogeneity in the connection patterns and to perform clustering of entities based on their interaction profile. SBM (respectively LBM) are adapted to adjacency (respectively incidence) matrices.

SBM Assume that we want to study the relations inside a group of n individuals. Let $E = \{1, \dots, n\}$ be the set of nodes and X be the adjacency matrix. The simplest stochastic model is the Erdös-Rényi model which sets that $\forall (i,j) \in \{1, \dots, n\}^2, i \neq j$,

$$X_{ij} \sim_{i.i.d} \mathcal{B}\text{ern}(\alpha).$$

As a consequence, individual connects independently with the same probability. Heterogeneity in the connection phenomenon can be introduced by considering latent variables $\mathbf{Z} = (Z_1, \dots, Z_n) \in \{1, \dots, K\}^n$ such that $(Z_i)_{i=1\dots n}$ are independent with distribution

$$\mathbb{P}(Z_i = k) \quad = \quad \pi_k \quad = \quad \pi_{Z_i}, \tag{3.1}$$

and conditionally to the latent variables $\mathbf{Z}$, the $(X_{ij})_{i,j}$ are independent with distribution:

$$\mathbb{P}(X_{ij} = 1 | Z_i = k, Z_j = l) \quad = \quad \alpha_{kl} \quad = \quad \alpha_{Z_i,Z_j}, \tag{3.2}$$

resulting into the so-called Stochastic Block Models. Note that other models with latent variables can be proposed to model heterogeneity (see Matias and Robin (2014) and references therein). The likelihood deriving from (3.1) and (3.2) is:

$$\begin{aligned} \ell(X;\theta) &= \sum_{\mathbf{Z} \in \{1,\dots K\}^n} \prod_{i,j,i \neq j} \alpha_{Z_i Z_j}^{X_{ij}} (1 - \alpha_{Z_i Z_j})^{1 - X_{ij}} \prod_{i=1}^{n} \pi_{Z_i} \\ &= \sum_{\mathbf{Z} \in \{1,\dots K\}^n} \ell_c(Y, \mathbf{Z}; \theta). \end{aligned} \tag{3.3}$$

where $\ell_c(X, \mathbf{Z}; \theta)$ is the so-called complete likelihood:

$$\ell_c(X, \mathbf{Z}; \theta) = \prod_{i,j} \prod_{i,j,i \neq j} \alpha_{Z_i Z_j}^{X_{ij}} (1 - \alpha_{Z_i Z_j})^{1 - X_{ij}} \prod_{i=1}^{n} \pi_{Z_i}, \tag{3.4}$$

LBM Now, consider two *functional groups* of respective sizes n_1 and n_2. Let Y be the incidence matrix corresponding to the bipartite graph of interest. Assuming that the interaction phenomenon is not homogeneous in the population, one can introduce connection heterogeneity through latent blocks, resulting into the so-called Latent Block Models (LBM). Let $\left(Z_1^{(1)}, \ldots, Z_{n_1}^{(1)}\right) \in \{1, \ldots, K_1\}^{n_1}$ and $\left(Z_1^{(2)}, \ldots, Z_{n_2}^{(2)}\right) \in \{1, \ldots, K_2\}^{n_2}$ be $n_1 + n_2$ independent random variables, such that:

$$\begin{array}{rcll} \mathbb{P}(Z_i^{(1)} = k) & = & \pi_k^{(1)}, & \forall k = 1 \ldots K_1, \forall i = 1 \ldots n_1 \\ \mathbb{P}(Z_j^{(2)} = l) & = & \pi_l^{(2)}, & \forall l = 1 \ldots K_2, \forall j = 1 \ldots n_2 \end{array} \tag{3.5}$$

with $\sum_{k=1}^{K^{(q)}} \pi_k^{(q)} = 1$ for $q = 1, 2$. Then conditionally to $\mathbf{Z} = \{Z_i^{(q)}, i = 1 \ldots n_q, q = 1, 2\}$, the (Y_{ij}) are independent with distribution:

$$\mathbb{P}(Y_{ij} = 1 | Z_i^{(1)} = k, Z_j^{(2)} = l) = \alpha_{kl}. \tag{3.6}$$

Equations (3.5) and (3.6) define the Latent Block Model (LBM) resulting into a co-clustering of rows and columns of Y, whose likelihood is:

$$\begin{aligned} \ell(Y; \theta) &= \sum_{\mathbf{Z} \in \boldsymbol{\mathcal{Z}}} \prod_{i,j} \alpha_{Z_i^{(1)} Z_j^{(2)}}^{Y_{ij}} (1 - \alpha_{Z_i^{(1)} Z_j^{(2)}})^{1 - Y_{ij}} \prod_{i=1}^{n_1} \pi_{Z_i^{(1)}}^{(1)} \prod_{j=1}^{n_2} \pi_{Z_j^{(2)}}^{(2)} && (3.7) \\ &= \sum_{\mathbf{Z} \in \boldsymbol{\mathcal{Z}}} \ell_c(Y, \mathbf{Z}; \theta) && (3.8) \end{aligned}$$

where $\boldsymbol{\mathcal{Z}} = \{1, \ldots, K_1\}^{n_1} \times \{1, \ldots, K_2\}^{n_2}$. $\ell_c(Y, \mathbf{Z}; \theta)$ is referred as the complete likelihood:

$$\ell_c(Y, \mathbf{Z}; \theta) = \prod_{i,j} \alpha_{Z_i^{(1)} Z_j^{(2)}}^{Y_{ij}} (1 - \alpha_{Z_i^{(1)} Z_j^{(2)}})^{1 - Y_{ij}} \prod_{i=1}^{n_1} \pi_{Z_i^{(1)}}^{(1)} \prod_{j=1}^{n_2} \pi_{Z_j^{(2)}}^{(2)}. \tag{3.9}$$

1.3 Estimation and model selection

Parameters estimation As soon as (n, K) or (n_1, n_2, K_1, K_2) increase, the observed likelihoods (3.3) or (3.7) become intractable (due to the summation over $\boldsymbol{\mathcal{Z}}$) and their maximization is a challenging task. Several approaches have been developed (for a review, see Matias and Robin, 2014), both in the frequentist and Bayesian frameworks, starting from Snijders and Nowicki (1997) and Nowicki and Snijders (2001). However, when the latent data space is really large, these techniques may be burdensome. Some other strategies have been proposed, such as Bickel and Chen (2009) relying on a profile-likelihood optimization or the moment estimation proposed by Ambroise and Matias (2012), to name but a few.

The variational EM (adapted to SBM context by Daudin et al., 2008) is a flexible tool to tackle the computational challenge in many types of graphs. In a few words, the variational EM aims at optimizing a lower bound of the log-likelihood, namely

$$\begin{array}{rcl} \mathcal{I}_\theta(\mathcal{R}_\mathbf{D}) & = & \log \ell(\mathbf{D}; \theta) - \mathbf{KL}[\mathcal{R}_\mathbf{D}, p(\cdot | \mathbf{D}; \theta)] \\ & = & \mathbb{E}_{\mathcal{R}_\mathbf{D}} \left[\log \ell_c(\mathbf{D}, \mathbf{Z}; \theta)\right] - \mathbb{E}_{\mathcal{R}_\mathbf{D}} \left[\mathcal{R}_\mathbf{D}(\mathbf{Z})\right], \end{array} \tag{3.10}$$

where $\mathbf{D}$ denotes the observations (X or Y), $\mathbf{KL}$ is the Kullback-Leibler divergence, $p(\cdot|\mathbf{D};\theta)$ is the true conditional distribution of the latent variables $\mathbf{Z}$ given the observed data $\mathbf{D}$ and $\mathcal{R}_\mathbf{D}$ is an approximation of this conditional distribution $p(\cdot|\mathbf{D};\theta)$. Note that $\mathcal{I}_\theta(\mathcal{R}_\mathbf{D}) = \log \ell(\mathbf{D};\theta)$ if and only if the $\mathcal{R}_\mathbf{D} = p(\cdot|\mathbf{D};\theta)$. Dealing with the exact distribution $p(\cdot|\mathbf{D};\theta)$ being impossible, the principle is to approximate it by $\mathcal{R}_\mathbf{D}$, where $\mathcal{R}_\mathbf{D}$ belongs to a certain class of "simple" distributions. The variational EM alternates between the maximization of $\mathcal{I}_\theta(\mathcal{R}_\mathbf{D})$ with respect to $\mathcal{R}_\mathbf{D}$ and its maximization with respect to θ using the two formulations of equation (3.10).

Simulation studies show its practical efficiency (Mariadassou et al., 2010). Moreover, its theoretical convergence towards the maximum likelihood estimates has been studied by Bickel et al. (2013) for binary graphs. Its application to LBM has also been proposed by Govaert and Nadif (2008). Ad-hoc extensions of the variational EM algorithm have to be designed for every new probabilistic model.

Model selection through ICL The selection of the most adequate number of blocks K in SBM (or K_1 and K_2 for LBM) is a challenging issue. In the recent years, the Integrated Completed Likelihood criterion (ICL) has become a standard criterion to select the most adequate number of blocks.

Let $\mathcal{M}$ be a model, the well-known BIC (Bayesian Information Criterion) is a penalized likelihood criterion, widely used for model choice. It is defined as $\text{BIC} = \log \ell(\mathbf{D};\widehat{\theta},\mathcal{M}) - \text{Pen}_\text{BIC}(\mathcal{M})$. The $\text{Pen}_\text{BIC}(\mathcal{M})$ (penalizing the complexity of the model) derives from a Laplace approximation of the marginal likelihood $\int \ell(\mathbf{D};\theta,\mathcal{M})p(\theta)d\theta$. The BIC criterion provides –under regularity conditions– a reliable approximation of this integrated likelihood. However, these regularity conditions on the likelihood function do not hold for mixture models or stochastic blocks models. Moreover, in the SBM/LBM context, the quantity $\log \ell(\mathbf{D};\widehat{\theta},\mathcal{M})$ has no explicit expression (due to the integration over the latent variables $\mathbf{Z}$).

ICL has been proposed as an alternative to the BIC in the model-based clustering context (Biernacki et al., 2000). Let $\ell_c(\mathbf{D},\mathbf{Z};\theta,\mathcal{M})$ be the complete likelihood [see equations (3.4) and (3.9)]. ICL is a penalized conditional complete likelihood. More precisely, first considering that $\mathbf{Z}$ is observed, a Laplace of approximation the completed marginal likelihood gives:

$$\log \int_\theta \ell_c(\mathbf{D},\mathbf{Z};\theta,\mathcal{M})p(\theta)d\theta \approx \log \ell_c(\mathbf{D},\mathbf{Z};\widehat{\theta},\mathcal{M}) - \text{Pen}_\text{ICL}(\mathcal{M}) \tag{3.11}$$

when the prior distribution on θ $p(\theta)$ is a well-chosen non-informative one. The penalty term $\text{Pen}_\text{ICL}(\mathcal{M})$ expresses as:

$$\text{Pen}_\text{ICL}(\mathcal{M}) = \frac{1}{2}\left\{K^2 \log(n(n-1)) + (K-1)\log n\right\} \quad \text{for non-symmetric SBM,} \tag{3.12}$$

and

$$\text{Pen}_\text{ICL}(\mathcal{M}) = \frac{1}{2}\left\{K_1 K_2 \log(n_1 n_2). + (K_1-1)\log n_1 + (K_2-1)\log n_2\right\} \quad \text{for LBM.} \tag{3.13}$$

As in the BIC criterion, the log refers to the number of data. Thus, in (3.12), the n nodes are used to estimate the $K-1$ probabilities $\pi_1,\ldots,\pi_{K-1}$. The $n(n-1)$ edges are used to estimate $\boldsymbol{\alpha}$.

$\mathbf{Z}$ being non-observed, Biernacki et al. (2000) propose to integrate the latent variables, thus defining:

$$\text{ICL} = \mathbb{E}_{\mathbf{Z}|\mathbf{D},\widehat{\theta},\mathcal{M}}\left[\log \ell_c(\mathbf{D},\mathbf{Z};\widehat{\theta},\mathcal{M})\right] - \text{Pen}_\text{ICL}(\mathcal{M}). \tag{3.14}$$

Note that the latent variables $\mathbf{Z}$ can also be imputed replacing them with their posterior mode, leading to this second version of the ICL :

$$\begin{aligned} \text{ICL} &= \log \ell_c(\mathbf{D}, \hat{\mathbf{Z}}; \widehat{\theta}, \mathcal{M}) - \text{Pen}_{\text{ICL}}(\mathcal{M}) \\ \hat{\mathbf{Z}} &= \arg\max_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{D}, \widehat{\theta}, \mathcal{M}). \end{aligned}$$

When the VEM is used to estimate the parameters, we obtain an approximation of the conditional distribution $p(\mathbf{Z}|\mathbf{D}, \widehat{\theta}, \mathcal{M})$ by $\widehat{\mathcal{R}}_{\mathbf{D},\mathcal{M}}$ (minimizing the Kullback distance in a particular class of distributions). We are then able to give an explicit approximation of $\mathbb{E}_{\mathbf{Z}|\mathbf{D},\widehat{\theta},\mathcal{M}}\left[\log \ell_c(\mathbf{D}, \mathbf{Z}; \widehat{\theta}, \mathcal{M})\right]$ by $\mathbb{E}_{\widehat{\mathcal{R}}_{\mathbf{D},\mathcal{M}}}\left[\log \ell_c(\mathbf{D}, \mathbf{Z}; \widehat{\theta}, \mathcal{M})\right] = \log \ell_c(\mathbf{D}, \mathbb{E}_{\widehat{\mathcal{R}}_{\mathbf{D},\mathcal{M}}}[\mathbf{Z}]; \widehat{\theta}, \mathcal{M})$ in the particular cases of SBM and LBM.

Here are a few comments on the behavior of the ICL criterion. We can decompose the marginal likelihood $\log \ell(\mathbf{D}; \widehat{\theta}, \mathcal{M})$ as follows:

$$\log \ell(\mathbf{D}; \widehat{\theta}, \mathcal{M}) = \mathbb{E}_{\mathbf{Z}|\mathbf{D},\widehat{\theta},\mathcal{M}}\left[\log \ell_c(\mathbf{D}, \mathbf{Z}; \widehat{\theta}, \mathcal{M})\right] + H(\mathbf{Z}|\mathbf{D}; \widehat{\theta}, \mathcal{M}) \tag{3.15}$$

where $H(\mathbf{Z}|\mathbf{D}, \widehat{\theta}, \mathcal{M}) = -\int \log p(\mathbf{Z}|\mathbf{D}; \widehat{\theta}, \mathcal{M}) p(\mathbf{Z}|\mathbf{D}; \widehat{\theta}, \mathcal{M}) d\mathbf{Z}$ is the entropy of the conditional distribution $p(\mathbf{Z}|\mathbf{D}, \widehat{\theta}, \mathcal{M})$. Consequently, ICL of equation (3.14) may be seen as a penalized maximum likelihood, where the penalty includes not only the complexity of the model $\text{Pen}_{\text{ICL}}(\mathcal{M})$ but also the entropy of $p(\mathbf{Z}|\mathbf{D}, \widehat{\theta}, \mathcal{M})$. Thus, using the ICL will automatically encourage clustering configurations with well separated groups. Its capacity to outline the clustering structure in the data has been tested, either in mixture models (Baudry et al., 2008), LBM (Keribin et al., 2014) or SBM (Mariadassou et al., 2010).

Moreover, note that, in the SBM context, under standard asymptotic assumptions, the posterior distribution $p(\mathbf{Z}|\mathbf{D}, \widehat{\theta}, \mathcal{M})$ concentrates on the true affections (see Mariadassou and Matias, 2015). As a consequence, for n large, the entropy vanishes and ICL becomes a standard penalized maximum likelihood. We expect the same type of results in the multiplex SBM, but additional theoretical work is required.

I now present a methodological joint work with S. Robin about the Bayesian Inference for SBM with covariates.

2 Bayesian inference for SBM with covariates [S1]

This joint work with S. Robin deals with the combination of variational methods and Sequential Monte Carlo methods for Bayesian inference. This work is motivated by the modeling of networks where any pair of individuals is described by a set of covariates, resulting into the so-called Stochastic Block model with covariates. I first describe shortly the model and then present the general methodology we propose in [**S**1].

2.1 SBM with covariates

We consider the combination of SBM and logistic regression (shortened as 'SBM-reg' in the sequel) considered in Latouche et al. (2015). This model aims at deciphering some residual structure in an observed network once accounted for the effect of some edge covariates. The model is an extension of the SBM defined in Section 1, equations (3.1) and (3.2), page 65. More

precisely, consider a set of n nodes; for each pair $(1 \leq i < j \leq n)$ of nodes, we observe a p-dimensional covariates vector $\boldsymbol{c}_{ij}$. Likewise in SBM, we further assume that each node belongs to one among K groups and we denote Z_i the (unobserved) group where node i is affected; $\boldsymbol{\pi} = (\pi_k)_{k=1,\dots,K}$ denotes the vector of group proportions. The model states that the edges of the observed binary undirected network $\mathbf{X} = (X_{ij})$ are drawn independently conditionally on the set of latent variables $\mathbf{Z} = (Z_i)$ as Bernoulli variables:

$$(X_{ij}|Z_i, Z_j, \boldsymbol{\alpha}, \boldsymbol{\beta}) \sim \mathcal{B}(p_{ij}), \quad \text{logit}(p_{ij}) = \boldsymbol{c}_{ij}^{\intercal}\boldsymbol{\beta} + \alpha_{Z_i,Z_j}$$

where $\boldsymbol{\alpha} = (\alpha_{kl})_{k,l=1,\dots,K}$ stands for the matrix of between-group effects (analogous to the between-group connection probabilities from SBM, in logit scale) and $\boldsymbol{\beta} = (\beta_\ell)_{\ell=1,\dots,p}$ for the vector of regression coefficients.

We are interested in understanding the influence of the covariates on the connection patterns. We work in a Bayesian framework, setting prior distributions on the parameters. As for the priors, π has a Dirichlet distribution, both α and β are Gaussian. When considering model selection or averaging, the number of groups K is supposed to be uniformly distributed among $\{1, \dots, K_{\max}\}$. When inferring $\boldsymbol{\beta}$, we need to take into account the uncertainty on the number of groups K. Rather than to choose the 'best' model (choosing K), Bayesian Model averaging (BMA) (Hoeting et al., 1999) is a general principle, which consists in combining the results obtained with several models. Among other interests, it allows to account for model uncertainty. More precisely, while model selection consists in choosing K as $\widehat{K} = \arg\max_K p(K|\mathbf{X})$ and considering the posterior $p(\beta_\ell|\mathbf{X}, K = \widehat{K})$, BMA directly considers the unconditional posterior

$$p(\beta_\ell|\mathbf{X}) = \sum_K p(K|\mathbf{X}) p(\beta_\ell|\mathbf{X}, K).$$

In terms of moments, it results into

$$\begin{aligned} \mathbb{E}(\beta_\ell|\mathbf{X}) &= \sum_K p(K|\mathbf{X}) \mathbb{E}(\beta_\ell|\mathbf{X}, K) \\ \mathbb{V}(\beta_\ell|\mathbf{X}) &= \mathbb{V}_{\text{within}}(\beta_\ell|\mathbf{X}) + \mathbb{V}_{\text{between}}(\beta_\ell|\mathbf{X}) \end{aligned}$$

where $\mathbb{V}_{\text{within}}$ measures the mean variance of the parameter conditionally on K and $\mathbb{V}_{\text{between}}$ is the variance of the parameter due to model uncertainty:

$$\begin{aligned} \mathbb{V}_{\text{within}}(\beta_\ell|\mathbf{X}) &= \sum_K p(K|\mathbf{X}) \mathbb{V}(\beta_\ell|\mathbf{X}, K), \\ \mathbb{V}_{\text{between}}(\beta_\ell|\mathbf{X}) &= \sum_K p(K|\mathbf{X}) \left(\mathbb{E}(\beta_\ell|\mathbf{X}, K) - \mathbb{E}(\beta_\ell|\mathbf{X})\right)^2. \end{aligned}$$

2.2 Bayesian inference for SBM with covariates

Existing methods SBM with covariates involves a large number of latent variables ($\mathbf{Z}$). In Bayesian statistics, this is a typical situation where standard algorithms for posterior sampling such as MCMC reach their limitations. Indeed, when the space of parameters to explore is of high dimension, MCMC algorithms will have difficulties in reaching their equilibrium distribution within a reasonable computational time.

Recently, population based Monte Carlo methods have proved their efficiency and robustness in front of high dimensional and multimodal spaces. Among population based Monte Carlo, Sequential Monte Carlo (SMC) (Del Moral et al., 2006) is a method combining parameters sampling and resampling. More precisely, a sequence of distributions of interest is designed, such that the first one is simple (*i.e.* easy to sample from) and the last one is the posterior distribution.

This sequence of distributions defines the iterations of the algorithm. At the first iteration, a sample of parameters is simulated with the first distribution. In the following iterations, the parameters are stochastically moved, weighted and resampled to follow the current distribution. The true posterior distribution is reached at the last iteration.

In the recent years, particular fields (such as genomics or network analysis to name but a few) brought news statistical problems involving an increasing amount of data or statistical models with a large number of parameters. In such cases, not only MCMC but also population Monte Carlo algorithms have reached their limitations, requiring unreasonable computational time to explore the posterior distribution. To deal with such difficulties, deterministic approximations of the posterior distribution through optimization mathematical tools – such as variational approximation (Wainwright and Jordan, 2008; Blei et al., 2016), Expectation-Propagation (Minka, 2001) or Integrated nested Laplace approximation (Rue et al., 2009) for instance– have been proposed. These methods have the great advantage to be computationally light and can handle large data. However, their theoretical properties and accuracy are still under study. In particular, we do know that variational approximations may supply underestimated posterior variances (see for instance Consonni and Marin, 2007, for a large illustration of this phenomena on the Probit model).

One may therefore be tempted to take advantage of the two approaches in a combined strategy. The idea of combining variational Bayes inference with SMC is actually not new. Rabinovich et al. (2015) split the data into block and compute the posterior distribution of θ given each block. They use a variational argument to propose the product of this partial posterior as a proxy for the true posterior. Focusing on Gaussian mixtures, McGrory et al. (2016) consider online-inference and propose a sequential sampling scheme where, for each new batch of data, the variational approximation is iteratively updated and used as a proposal. Naesseth et al. (2017) use a SMC approach to get an improved, but still biased, variational approximation.

From the approximate posterior distribution to the true posterior distribution Our approach is different from all these ones. Our main idea is to design a bridge sampling from the approximated posterior distribution to the true posterior distribution, the transfer from the approximate to the exact distribution being performed with an SMC algorithm (Del Moral et al. (2006)). The sampling method we propose may be considered from two points of view: either SMC is seen as a tool to correct the approximate distribution, or the approximate posterior distribution is seen as a mean to drastically accelerate the SMC procedure.

Let $\ell(\mathbf{X}|\boldsymbol{\theta})$ be the likelihood function with $\boldsymbol{\theta} \in \Theta$ the unknown parameters and possibly the latent variables ($\mathbf{Z}$). $\pi(\boldsymbol{\theta})$ is the prior distribution on $\boldsymbol{\theta}$. Let $p(\boldsymbol{\theta}|\mathbf{X}) = \frac{\ell(\mathbf{X}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\mathbf{X})}$ be the posterior distribution where $p(\mathbf{X})$ is the marginal likelihood.

In what follows, $\widetilde{p}_{\boldsymbol{Y}}$ is an approximate posterior distribution on $\boldsymbol{\theta}$. *We assume that $\widetilde{p}_{\boldsymbol{Y}}$ can be easily intensively simulated and that the density function of $\widetilde{p}_{\boldsymbol{Y}}$ has an explicit expression.*

Sequential Monte Carlo samplers generate samples from a sequence of intermediate distributions $(p_h)_{h=0\dots H}$ where the intermediate distributions $(p_h)_{h=0\dots H}$ are smooth transitions from a simple distribution p_0 to the distribution of interest $p_H = p(\cdot|\mathbf{X})$. A classical choice for $(p_h)_{h=0\dots H}$ (Neal, 2001) is to consider $p_h(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta})\ell(\mathbf{X}|\boldsymbol{\theta})^{\rho_h}$ where $\rho_0 = 0$, $\rho_H = 1$, thus slowly shrinking the prior distribution into the posterior by progressively integrating the data $\mathbf{X}$ through the likelihood function. In this paper, we propose an alternative scheme moving smoothly from the approximate posterior distribution $\widetilde{p}_{\boldsymbol{Y}}$ to the true $p(\cdot|\boldsymbol{Y})$. The path is thus defined by:

$$\begin{aligned} p_h(\boldsymbol{\theta}) &\propto \widetilde{p}_{\boldsymbol{Y}}(\boldsymbol{\theta})^{1-\rho_h}(p(\boldsymbol{\theta}|\mathbf{X}))^{\rho_h} \\ &\propto \widetilde{p}_{\boldsymbol{Y}}(\boldsymbol{\theta})^{1-\rho_h}(\pi(\boldsymbol{\theta})\ell(\mathbf{X}|\boldsymbol{\theta}))^{\rho_h}. \end{aligned} \tag{3.16}$$

where, $\rho_0 = 0$, $\rho_H = 1$. In a few words, we start from the easy-to-sample distribution $\widetilde{p}_{\boldsymbol{Y}}(\boldsymbol{\theta})$ and progressively replace it with the true posterior distribution, this strategy being known as annealed importance sampling procedure (Neal, 2001). We claim that this scheme significantly reduces the computational time and is robust with respect to $\widetilde{p}_{\boldsymbol{Y}}$. Note that, when thinking about $\widetilde{p}_{\boldsymbol{Y}}$ has a Bayesian variational approximation, the first distribution $\widetilde{p}_{\boldsymbol{Y}}$ is likely to be more spiked than the distribution of interest $p(\cdot|\mathbf{X})$ and will assume untrue dependencies between the parameters: the procedure will be used to get back to the true variance and the possibly ignored dependencies between the parameters.

To sample from the sequence of distributions $(p_h)_{h=1,\dots,H}$, we adopt the sequential sampler proposed by Del Moral et al. (2006) where the annealing coefficients ρ_h will be adjusted dynamically. With respect to MCMC strategies, Annealing Importance Sampling and SMC have the great advantage to supply good estimators of the marginal likelihood. Indeed, as proved by Del Moral et al. (2006), a non-biased estimator of the marginal likelihood derives as a by-product of SMC. Moreover, the path sampling identity also provides an estimate of the marginal likelihood. Details are in [**S**1], along with comparison with other existing strategies. Our resulting path sampler is referred as Shorten Bridge Sampler (SBS) in what follows.

In [**S**1], the robustness and the efficiency of our algorithm is illustrated on several models: logistic regression, latent class analysis (LCA) model and finally on SBM with covariates. The logistic regression serves as a toy example to illustrate the efficiency and the robustness of our methodology. In particular, we point out the fact that even if the approximated posterior distribution $\widetilde{p}$ has an underestimated variance (which is known to be the case for the Variational Bayes estimator in this case), our methodology will supply a sample from the true posterior distribution. Moreover, the algorithm is also robust when $\widetilde{p}_{\boldsymbol{Y}}$ is spiked around an absurd value.

2.3 Numerical experiments on SBM with covariates

A first comment can be made on the Variational Bayes approximation itself for SBM with covariates. As illustrated in [**S**1], the VB approximate posterior is quite accurate for logistic regression and Gazal et al. (2012) also proved its empirical accuracy for SBM. A first goal of this simulation study is to check if this accuracy still holds when the two models are combined into the SBM-reg model. To this aim, we focus on the posterior distribution of the regression parameters. Secondly, we want to check the accuracy of the VB posterior distribution of the number of groups, that can be used either to assess goodness-of-fit or for model averaging (Latouche et al., 2015).

Simulation design. We simulate networks with $n \in \{20, 50\}$ nodes according to an SBM-reg model with $K^* \in \{1, 2\}$ groups and $p = 3$ covariates. The parameters are sampled from the prior distribution. $S = 100$ replicates are simulated for each configuration and, for each of them, the SBM-reg models with $K \in \{1, \dots, K_{\max} = 5\}$ were fitted with the VB algorithm described in Latouche et al. (2015). The SBS algorithm is then run on each dataset.

Results for parameter estimation. We first consider the posterior distribution of β when the number of groups K is known. On Figure 3.1, we plot on the left the boxplots for the posterior means $(\widehat{\mathbb{E}}^{VB}(\beta_\ell|\mathbf{X}^s))_{s=1\dots100}$ and $(\widehat{\mathbb{E}}^{SBS}(\beta_\ell|\mathbf{X}^s))_{s=1\dots100}$. The boxplot (over the 100 simulated datasets) of the posterior standard deviations $(\widehat{\sigma}^{VB}(\beta_\ell|\mathbf{X}^s))_{s=1\dots100}$ and $(\widehat{\sigma}^{SBS}(\beta_\ell|\mathbf{X}^s))_{s=1\dots100}$ are on the top-right. We clearly observe that the posterior means provided by VB and SBS are both accurate and similar, but the VB's posterior standard deviations (sd) are smaller than SBS's posterior standard deviations.

We also point out the fact that, although the VB approximate posterior distribution is accurate

for logistic regression and SBM separately, it is biased for the SBM-reg model, and that the proposed SBS is a way to correct it. As a consequence of this phenomenon, the empirical level of VB's credibility intervals is equal to 84.75%, which is below the nominal level 95%, whereas SBS's credibility intervals almost reach the targeted level (93.75%).

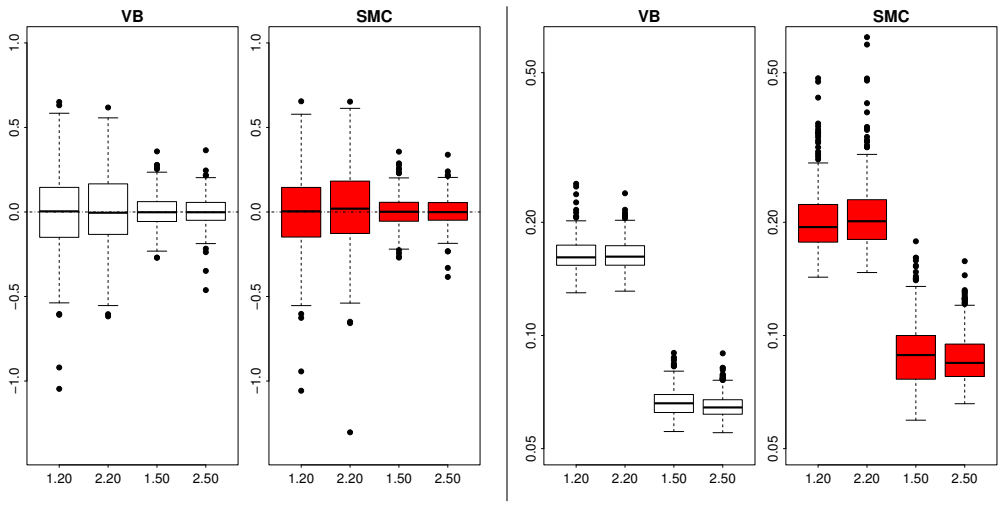


Figure 3.1 – Simulation results for the SBM-regmodel: VB (white) and SMC (red) posterior of the regression coefficients $\boldsymbol{\beta} = (\beta_\ell)$. Top: posterior mean (left), posterior standard deviation (right); x-axis label: $K^*.n$ (e.g. '1.20' means $K^* = 1$, $n = 20$). Left: $K = K^*$, right: with model averaging.

Results for model selection. We now consider the posterior distribution of the number of groups $p(K|\boldsymbol{Y})$ and its use for model selection. Figure 3.2 provides a comparison of the posterior provided by VB and SBS. We observe that the VB approximation always results in a more concentrated distribution than SBS. This behavior can be compared to the under-estimation of the posterior variance of the parameters that we already discussed. To compare the results in terms of model selection we computed the frequency at which the right model is selected (*i.e.* when $\widehat{K} = K^*$) and the mean posterior probability of the K^* (see Table 3.1). We observe that VB performs better than SBS for both criteria. This parallels Minka (2005), who shows that the minimization of the Kullback-Leibler (KL) divergence leads to an accurate estimate of the mode, which is convenient for model selection.

	frequency of $\widehat{K} = K^*$ (%)			
	$n = 20$		$n = 50$	
	$K^* = 1$	$K^* = 2$	$K^* = 1$	$K^* = 2$
VB	100	10	100	42
SBS	46	23	60	36

	mean value of $P(K = K^*\|\mathbf{X})$			
	$n = 20$		$n = 50$	
	$K^* = 1$	$K^* = 2$	$K^* = 1$	$K^* = 2$
VB	0.947	0.138	0.982	0.410
SBS	0.435	0.257	0.562	0.387

Table 3.1 – Simulation results for the SBM-regmodel: model selection

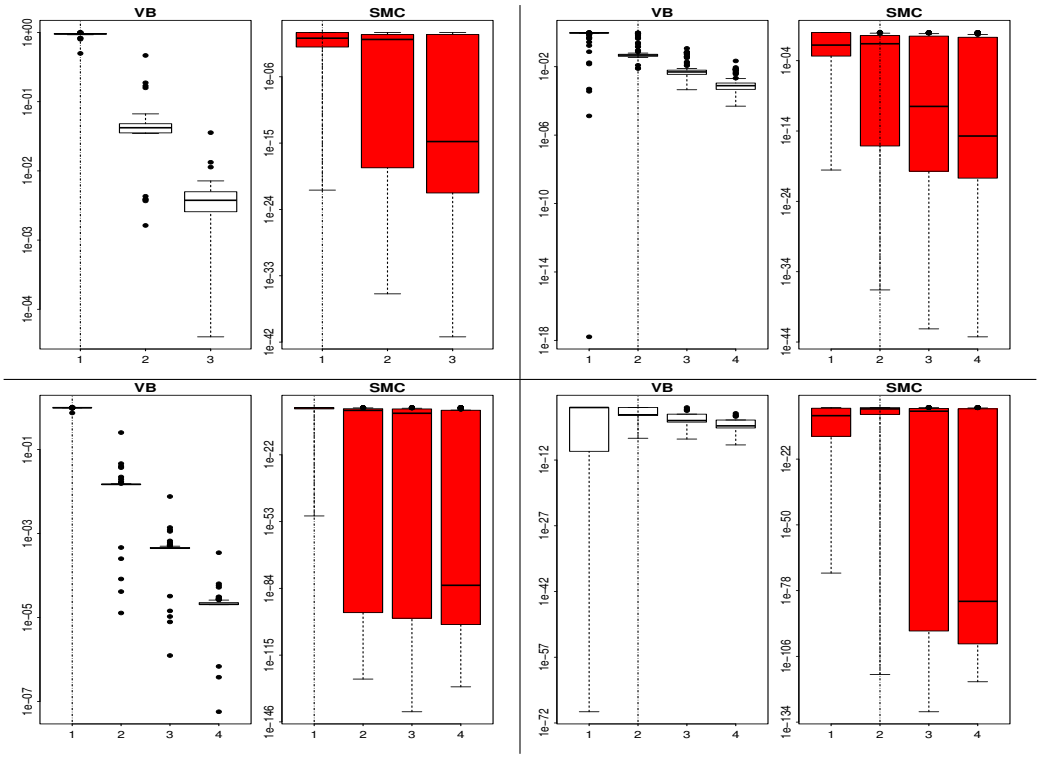


Figure 3.2 – Simulation results for the SBM-regmodel: box-plots for the posterior probability $p(g|\boldsymbol{Y})$ as a function of K. Top $n = 20$, bottom: $n = 50$. Left: $K^* = 1$, right: $K^* = 2$.

Although it does not seem to hamper model selection, the biased estimation of the posterior $p(K|\boldsymbol{Y})$ may have undesired consequences when used for model averaging. To illustrate this point, we simply computed the empirical coverage of credibility intervals for each β_ℓ after model averaging. The mean coverage across simulation condition and covariate index ℓ for VB (85.8%) is still below the nominal level, whereas this of SBS (93.25%) is close to 95%. Figure 3.1 (bottom right) also shows that the distribution of the ecdf after model averaging is almost confounded with the uniform for SBS, whereas it still displays a significant bias for VB.

2.4 Conclusion and comments

In this paper, we present a simple strategy to combine the strength of deterministic approximations of the posterior distribution with sequential Monte Carlo samplers. We illustrated the efficiency of our approach and its robustness with respect to the deterministic approximation on a large simulation study. Its application on network datasets stresses the fact that the well-known underestimation of the posterior variance by the variational approximation can be easily corrected, sometimes leading to different statistical conclusions. Besides, if dependencies between parameters have been neglected in the deterministic posterior approximation, they will be recovered by the sequential sampling.

Our approach is not restricted to the case where a standard deterministic posterior approximation can be derived (such as Variational Bayes, Laplace or Expectation Propagation estimate). Any point estimate can be used to design a rough posterior (using a Gaussian or a log-Gaussian seems to be the simplest solution) and serves as an accelerator of the sampling sequence. This strategy is different from an empirical Bayes strategy, the point estimate being only used to explore more efficiently the posterior distribution and not to elicit a prior distribution. The method is not as

sensitive as standard Importance Sampling to an eventual under-evaluation of the approximate posterior variance : even with a too narrow approximation of the posterior distribution, the algorithm is able to get back to the true posterior variance.

SMC directly supplies a final population of particles arising from the true posterior distribution, as opposed to MCMC strategies, whose convergence is difficult to assess. The proposed SBS algorithm is adaptive in the sense that the sequence $\widetilde{p}_{\boldsymbol{Y}}(\boldsymbol{\theta})^{1-\rho_h}(p(\boldsymbol{\theta}|\mathbf{X}))^{\rho_h}$ is determined on the fly in an automatic way. Furthermore, the algorithm path (summarized by the sequence ρ_h) is an indicator of the quality of the deterministic posterior distribution used to initiate the bridge sampling.

A natural extension of the present work is its adaptation to Approximate Bayesian Computation (ABC) context for models with no explicit likelihood, following Del Moral et al. (2012). The difficulty will arise from the specification of the distributions sequence.

As exposed in the introduction, in the last years, I focused my research on complex networks, "complex" referring to the fact that I am interested in the modeling of several co-occurring networks. I now present my first work on this topic, in collaboration with A. Bar-Hen, P. Barbillon and E. Lazega. Section 4 is dedicated to my on-going work with a discussion on perspectives.

3 Stochastic block model for multiplex networks [A15], [A16]

[**A**15] arose from a discussion with my colleague E. Lazega in University Paris Dauphine. He is a specialist of advice social networks.

Application context French scandals during the 1990s involving the voluntary sector around the cancer research dried up large donations that funded research laboratories. In the 2000s, the cancer research became politicized, with the launch of the Cancer Plan and the creation of a dedicated institution. The aim of this public agency is to coordinate the cancer research and to promote collaborations about top researchers. In this context, Lazega et al. (2008) studied the relations of advice between French cancer researchers identified as "Elite" conjointly with the relations of their respective laboratories. At the inter-individual level, two researchers are considered as linked if at least one kind of relationship exists among advice to deal with choices about the direction of projects, advice to find institutional support, advice to handle financial resources, advice for recruitment, and finally advice about manuscripts before submitting them to journals. Obviously the links are directed. An oriented link between laboratories is defined as an exchange of resources, as defined in the paper.

Objectives Our objective is to study the advice network between researchers conjointly with the network of resources exchanges of the laboratories they belong to. The modeling of multilevel networks is a hot topic in social sciences, aiming at understanding how individual interactions interact with institutional connections (Snijders and Lazega, 2016).

An individual-oriented strategy In this first work, we decide to adopt the following individual-oriented strategy (this point is discussed in the paper): the institutional network is used to define a new network on the individual level *i.e.* the set of nodes consists in the set of individuals and for a pair of individuals, two kinds of link are possible: a direct connection given by the individual network and a connection through their organizations given by the organizational network. As a

consequence, the individual and institutional levels are fused into a multiplex. We then develop a statistical model able to detect in multiplex substantial non-trivial topological features, with patterns of connection between their elements that are not purely regular.

A stochastic block model for multiplex networks Assume in general, that we observe not 2 but Q directed graphs $\mathbf{X}^1, \ldots, \mathbf{X}^Q$ relying on the same set of nodes $E = \{1, \ldots, n\}$. We assume that $\forall (i,j), i \neq j, \forall q \in \{1, \ldots, Q\}$, $X_{ij}^q \in \{0,1\}$ and $X_{ii}^q \neq 0$. We set $X_{ij}^{1:Q} = (X_{ij}^1, \ldots, X_{ij}^Q) \in \{0,1\}^Q$. Moreover, $\mathbf{X}^{1:Q} = \left(X_{ij}^{1:Q}\right)_{i \neq j}$. Let K be the number of blocks and $\mathbf{Z} = (Z_1, \ldots, Z_n)$ the latent variable such that $Z_i = k$ if the individual i belongs to block k (note that an individual only belongs to one block in this version).

The multiplex version of SBM is written as follows: $\forall (i,j) \in \{1, \ldots, n\}^2, i \neq j$, $\forall w \in \{0,1\}^Q$, $\forall (k,l) \in \{1, \ldots, K\}^2$,

$$\begin{aligned} \mathbb{P}(X_{ij}^{1:Q} = w | Z_i = k, Z_j = l) &= \alpha_{kl}^{(w)} \\ \mathbb{P}(Z_i = k) &= \pi_k, \end{aligned} \tag{3.17}$$

where the $(Z_i)_i$ are independent, and the $(X_{ij}^{1:Q})_{ij}$ are independent conditionally to $\mathbf{Z}$. Such a model involves $(2^Q - 1)K^2 + (K-1)$ parameters. Introducing the following notations:

$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K), \quad \boldsymbol{\alpha} = (\alpha_{kl}^{(w)})_{w \in \{0,1\}^Q, (k,l) \in \{1,\ldots,K\}^2}, \quad \theta = (\boldsymbol{\alpha}, \boldsymbol{\pi}),$$

the likelihood function is written as:

$$\begin{aligned} \ell(\mathbf{X}^{1:Q}; \theta) &= \int_{\mathbf{z} \in \{1,\ldots K\}^n} p(\mathbf{X}^{1:Q} | \mathbf{Z}; \boldsymbol{\alpha}) p(\mathbf{Z}; \boldsymbol{\pi}) d\mathbf{Z}, \\ &= \sum_{\mathbf{Z} \in \{1,\ldots K\}^n} \prod_{i,j,i \neq j} \alpha_{Z_i Z_j}^{(X_{ij}^{1:K})} \prod_{i=1}^n \pi_{Z_i}, \end{aligned} \tag{3.18}$$

Note that, conditionally to the groups, the $(X_{ij}^{1:Q})_{ij}$ are independent but the various levels $q = 1, \ldots, Q$ are not independent : no assumption is made on the structure of the $(\alpha_{kl}^{(w)})$ with respect to w. Once integrated out, the latent variables $\mathbf{Z}$ introduce dependence between the edges.

Statistical inference The likelihood is maximized by an adapted version of the Variational EM algorithm described in the appendix section of [**A**15]. This procedure has been added to the R package blockmodels (Leger, 2015).

The number of blocks K is selected with the ICL criterion, defined in this case as:

$$ICL(\mathcal{M}_K) = \max_\theta \log p(\mathbf{X}^{1:Q}, \widetilde{\mathbf{Z}}; \theta) - \frac{1}{2}\left\{K^2(2^Q - 1)\log(Qn(n-1)) + (K-1)\log n\right\}.$$

where $\widetilde{\mathbf{Z}}$ is the approximated conditional expectation or the conditional maximum (see Section 1.3). The identifiability of the model can be proved and the maximum likelihood estimators are consistent (see [**A**15] and theorems there in). No theoretical results are available for the ICL but numerical experiments performed on data mimicking the real ones prove that this criterion detects the true number of groups.

Applications On our dataset of interest in [**A**15], we were able to detect interesting patterns from a sociological point of view.

For the sake of clarity, we index by R and L (rather than X^1, X^2, where R stands for researchers and L for laboratories) the two adjacency matrices.

In Figure 3.3 we plot the marginal and conditional probabilities of the connections of researchers (respectively labs) between and within blocks. Note that the study of the estimated marginal distributions allows us to have results on the researchers without considering the laboratories. This gives a clear interpretation of the importance of the lab for the researcher network structure.

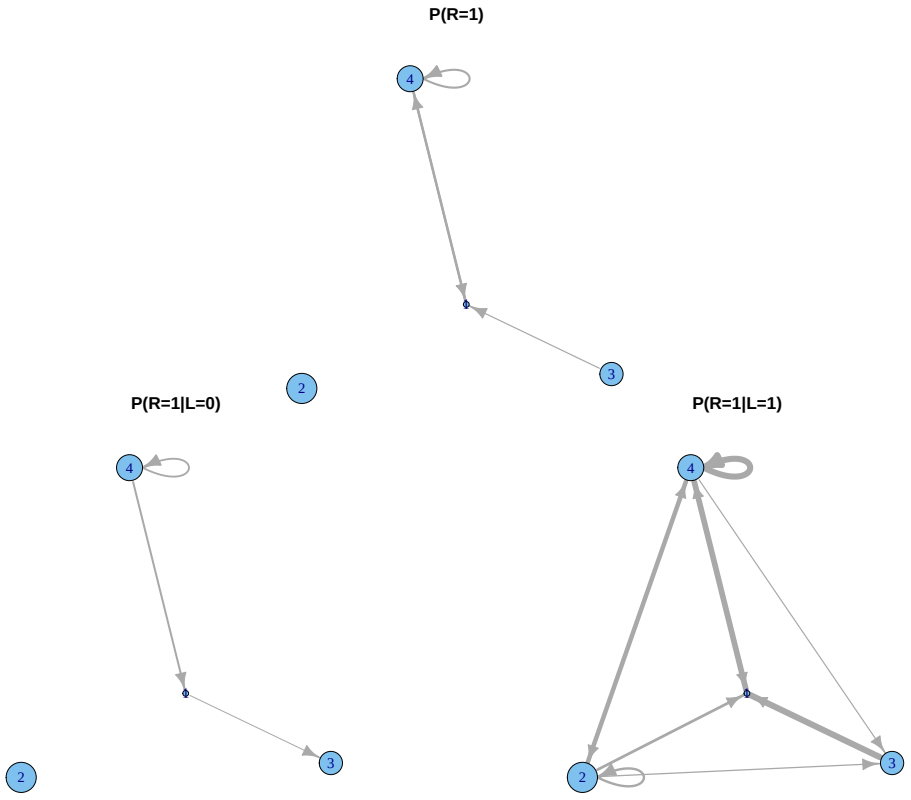


Figure 3.3 – Marginal probabilities of researcher connections between and within blocks (top) and probabilities of researcher connections between and within blocks conditionally on absence (bottom left-hand-side) or presence (bottom right-hand-side) of Lab connection. Vertex size is proportional to the block size. Edge width is proportional to the probabilities of connection; if this probability is smaller than 0.1, edges are not displayed.

Multiplex SBM reveals interesting structural features of the multiplex network. More precisely, collaboration takes place in a clustered manner for both researchers and laboratories; collaborating laboratories tend to have affiliated researchers seeking advice from one another. Indeed, Figure 3.3 shows that the existence of a connection (exchange of resources) between labs clearly increases the probability of connection (sharing advice) between researchers. The reinforcement of this probability of connection is clearly outstanding in block 2. In this block, the researcher connections are quite unlikely within the block or with other blocks. However, conditionally to the existence of a laboratory connection, the researcher connections become more important especially with block 4. In block 4, the links between researchers are strengthened given a connection between their laboratories. Researchers in block 3 seem to be the least affected by the connections provided by their laboratories. The case of block 1 is quite peculiar since it contains

two researchers only. This clustering demonstrates that not all researchers benefit on equal terms from the institutional level. Some researchers are more dependent on their laboratories in terms of connections.

Additional interpretation comments were made in [**A**15] about the identified groups, taking into account the additional informations (covariates) available on the researchers ("location", "director of not" and "specialty", "age", etc...). [**A**16] is the application of the same model and same inference method to a different dataset, where we observe at the same time relations of advice and competition between cancer researchers.

4 On going work and perspectives : towards more complex structures of networks

This section is dedicated to my most recent work and several perspectives.

4.1 Latent block models for multipartite networks [P6]

I first present my ongoing work on multipartite networks. Section 4.1.1 is dedicated to the presentation of two different motivating examples involving multipartite networks. Section 4.1.2 presents a probabilistic model able to handle both examples. The inference tools are described in the following section.

4.1.1 Two motivating contexts

Application context 1 Example 1 takes place in ecology. A high number of interaction types between plants and animal species co-exist within the natural environment. Among them, we may think about plant/animal interactions such as herbivory, protection of plants by ants, pollination, or seed dispersal. These various interactions play a key role in structuring biodiversity. In the recent years, network tools have been intensively used to understand the structure of these ecological interaction networks. However, in most of the works, each type of interaction is considered individually, ignoring the other interactions. A few recent works have considered the joint study of several interactions. See for instance the papers by Dáttilo et al. (2016), Fontaine et al. (2011) or Kéfi et al. (2016).

The data we consider is the one provided by Dáttilo et al. (2016) simultaneously studying the mutualistic interactions between plants and three functional groups, namely ants, pollinators and seed dispersal birds. The structure of the dataset is given in Table 3.2 in its matrix form and represented in plotted in Figure 3.4.

In this context, we aim at co-clustering the plants and the animals under the constraint that the clusters of "animals" must respect the functional classification. In other words, the clusters must be subsets of the functional groups.

Application context 2 The second example is motivated by the working group MIRES [1], gathering statisticians, geneticians, ethonologists and ecologists to name but a few. This group aims at understanding the influence of the social relations on the agrobiodiversity. More precisely, as an example, we study on the one hand the relations between individuals (farmers ou gardeners) and on the other hand the diversity of species in the agricultural production. The relations between individuals are of type "seed exchange" or social link.

[1] MIRES' website: https://sites.google.com/site/miresssna/home

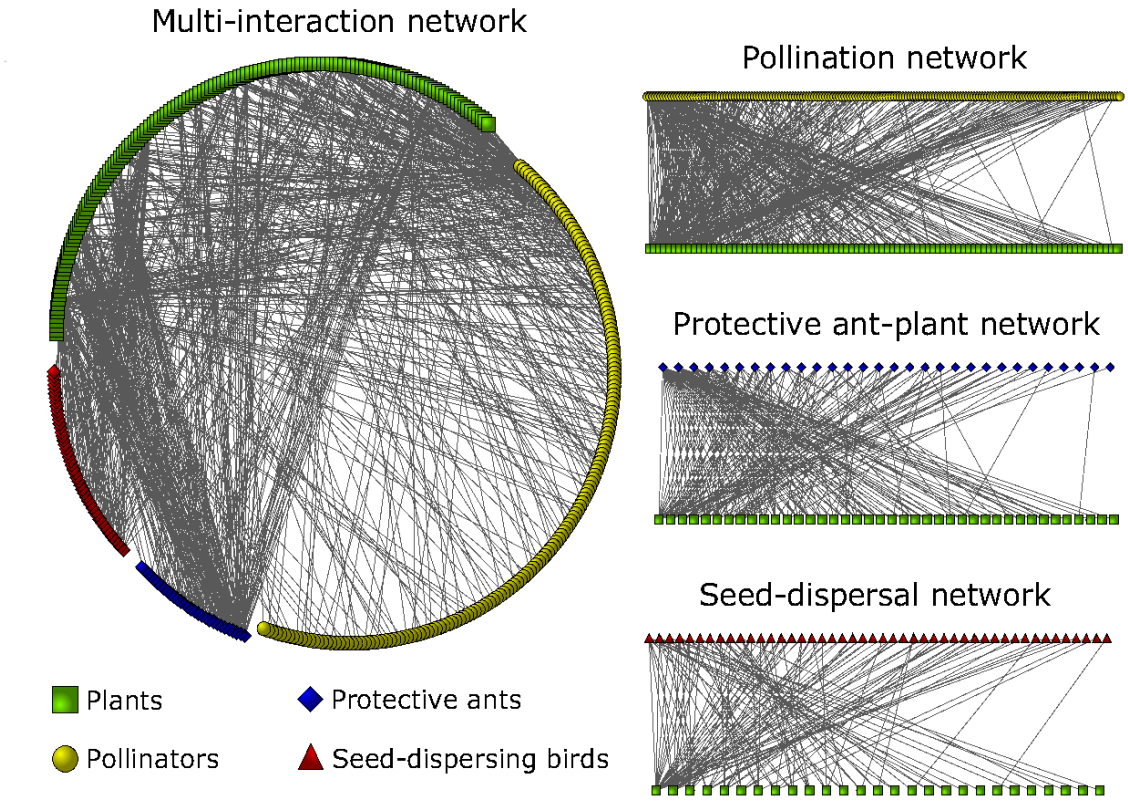


Figure 3.4 – *Application context 1.* Ecological multipartite network (extracted from Dáttilo et al. (2016))

Defining two functional groups, namely the farmers (or gardeners) and the cultivated species, the relations between farmers supply an adjacency matrix X^{11} , whereas the information of "who cultivates what" defines an incidence matrix X^{12}. The structure of the dataset is given in Table 3.3. Once again, we are interested in a co-clustering of individuals and plants with respect to these two matrices with clusters respecting the functional group structures.

Finally, there is a need for a probabilistic model supplying a co-clustering of the various functional groups, the clustering being based on the observation of several adjacency and/or incidence matrices. This unified model is presented in the following section.

4.1.2 A unified statistical model : the block model for complex multipartite networks

Let me consider Q functional groups (for instance plants, pollinators, ants... or farmers and plants) of respective sizes $n_1, \dots n_Q$. Assume that our dataset is a collection of matrices (adjacency or incidence) between or inside the functional groups. We denote by $\mathcal{E}$ the list of couples (q, q') for whom we observe an interaction matrix between functional groups q and q':

$$\mathbf{X} = \{X^{qq'}, (q, q') \in \mathcal{E}\}$$

where $X^{qq'}$ is a matrix of size $n_q \times n_{q'}$ with values in $\{0, 1\}$.

Note that, if $q = q'$, $X^{qq'}$ is an adjacency matrix (symmetric or not, depending on the context), wherea $X^{qq'}$ is an incidence matrix if $q \neq q'$.

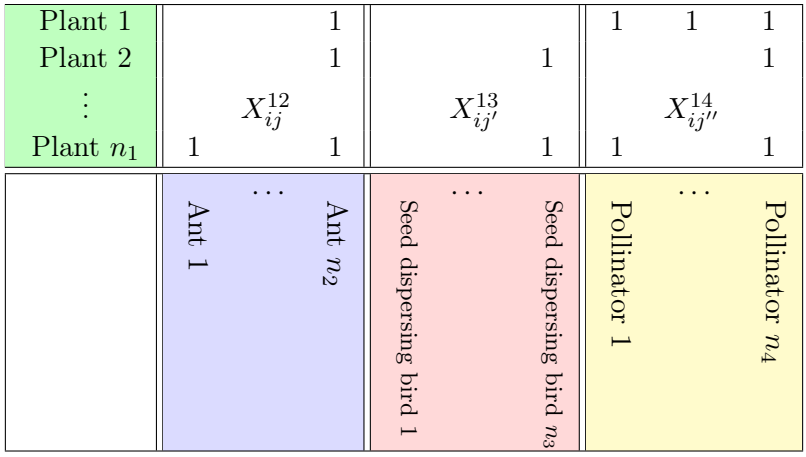


Table 3.2 – *Application context 1.* Ecological multipartite network : structure of the data.

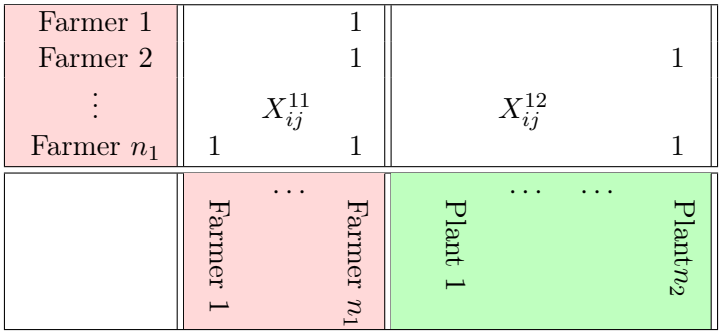


Table 3.3 – *Application context 2.* A social/ agrobiodiversity multipartite network : structure of the data

.

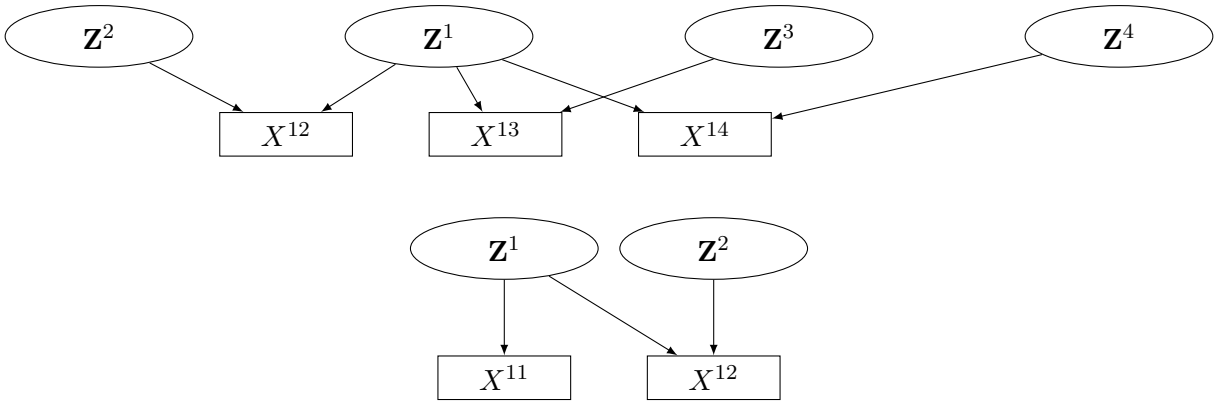


Figure 3.5 – DAG's corresponding to model (3.20) for Examples *1* (upper figure) and *2* (lower figure).

Remark 4.1. Given these notations, and adopting the notation 1 for plants, 2 for ants, 3 for pollinators and 4 for seed dispersal birds, the $\mathcal{E}$ corresponding to the first context is $\mathcal{E} = \{(1,2),(1,3),(1,4)\}$. $\mathcal{E}$ is equal to $\{(1,1),(1,2)\}$ for Context 2 if 1 refers to the farmers and 2 to the plants.

We propose a probabilistic model on $(X^{qq'})_{q,q'\in\mathcal{E}}$. Heterogeneity in the connections and dependence between the various matrices are handled through the following Latent Block structure. Assume that each functional group q is divided into K_q blocks/clusters. For any $q = 1 \dots Q$, we denote by $(Z_i^q)_{i=1\dots n_q}$ the independent random variables such that $Z_i^q = k$ if entity i of functional group q belongs to cluster k. We set the following model: $\forall k = 1 \dots K_q, \forall i = 1 \dots n_q, \forall q = 1 \dots Q$,

$$\mathbb{P}(Z_i^q = k) = \pi_k^q, \tag{3.19}$$

with $\sum_{k=1}^{K_q} \pi_k^q = 1$ for any $q = 1, \dots Q$.
Conditionally to the latent variables $\boldsymbol{Z} = \{Z_i^q, i = 1 \dots n_q, q = 1 \dots Q\}$, the observations $\boldsymbol{X} = \{X^{qq'}, (q,q') \in \mathcal{E}\}$ are distributed as follows:

$$X_{ij}^{qq'} | Z_i^q, Z_j^{q'} \sim_{i.i.d} \mathcal{B}ern(\alpha^{qq'}_{Z_i^q, Z_j^{q'}}) \,. \tag{3.20}$$

meaning that the probability of connection between i and j depends on the groups to which i and j belong.

Remark 4.2. Note that, if $q \neq q'$ then the distribution of $X_{ij}^{qq'}$ corresponds to a LBM, if $q = q'$ then we are dealing with a SBM-type model. As a consequence, this model is able to handle LBM and SBM at the same time. Moreover, talking about probabilistic dependences, note that conditionally to the affectation variables $\mathbf{Z}$, the adjacency/incidence matrices are independent. However, $\mathbf{Z}$ being non-observed (latent), the marginalization introduces dependence between the $\{X^{qq'}, (q,q') \in \mathcal{E}\}$ as can be visualized in the DAG corresponding to the models adapted to *Context 1* and *Context 2* in Figure 3.5.

4.1.3 Parameters inference and model selection

The parameters of interest are the connection probabilities $\boldsymbol{\alpha} = \{\alpha_{kk'}^{qq'}, k = 1 \dots K_q, k' = 1 \dots K_q, (q,q') \in \mathcal{E}\}$ and the clustering parameters $\boldsymbol{\pi} = \{\pi_k^q, k = 1 \dots K_q, q = 1 \dots Q\}$. In the following, we denote $\theta = (\boldsymbol{\alpha}, \boldsymbol{\pi})$.

Complete and marginal likelihoods Let $\ell(\mathbf{X};\theta)$ denote the likelihood of the observations $\mathbf{X}$ for parameter θ. Equations (3.20) and (3.19) allow us to write explicitly the joint distribution of $\mathbf{X}$ and $\mathbf{Z}$ (*complete likelihood*):

$$\begin{aligned}\ell_c(\mathbf{X},\mathbf{Z};\theta) &= p(\mathbf{X}|\mathbf{Z};\boldsymbol{\alpha})p(\mathbf{Z};\boldsymbol{\pi}) \\ &= \prod_{q,q'\in\mathcal{E}}\prod_{i=1}^{n_q}\prod_{j=1}^{n_{q'}}(\alpha^{qq'}_{Z_i^q,Z_j^{q'}})^{X_{ij}^{qq'}}(1-\alpha^{qq'}_{Z_i^q,Z_j^{q'}})^{1-X_{ij}^{qq'}}\prod_{q=1}^{Q}\prod_{i=1}^{n_q}\pi^q_{Z_i^q}\,. \end{aligned} \tag{3.21}$$

$\mathbf{Z}$ being latent, the log-likelihood of the observed data $\log\ell(\mathbf{X};\theta)$ is obtained by integrating (3.21) over all the possible values of $\mathbf{Z}$.

$$\log\ell(\mathbf{X};\theta) = \log\sum_{\mathbf{Z}\in\boldsymbol{\mathcal{Z}}}\ell_c(\mathbf{X},\mathbf{Z};\theta)\,. \tag{3.22}$$

However, $\boldsymbol{\mathcal{Z}} = \otimes_{q=1\ldots Q}\{1,\ldots,K_q\}^{n_q}$, which implies that when Q or K_q increase, this summation becomes impossible to perform in a close form. Following [**A**15], we are developing an adapted version of the variational Expectation Maximization algorithm to maximize the likelihood function. The estimated clusters $\widehat{\boldsymbol{Z}} = \{\widehat{Z}_i^q, i = 1\ldots n_q, q = 0\ldots Q\}$ will be a by-product of the inference method. Moreover, the ICL criterion can be easily adapted to this model.

4.1.4 Perspectives

The selection of the number of clusters in each functional group is a hard computational task. Indeed, from a practical point of view, like in any model selection problem, the difficulty comes from the huge number of models to scan. Let $K_q^{\max}$ be the maximum number of clusters authorized for functional group q. Then, $\prod_{q=1}^{Q}K_q^{\max}$ models have to be estimated trough the variational EM algorithm.

Moreover, as for any EM algorithm, the variational EM is very sensitive to its initialization. As a consequence, for each model, the algorithm will have to be run not once but several times from several initial parameters points, making the computational time unreasonable. An exhaustive strategy –where all the possible models would be estimated– cannot be considered in a general case. We have to propose a clever procedure to "travel" across the models space.

Several strategies has been proposed, see for instance Leger (2015) or the PhD thesis of Valérie Robert (Robert, 2017). Note that Valérie Robert developed a model similar to our multipartite latent block models but in a very different framework, namely pharmacovigilance.

The idea we consider implementing is the following iterative one.

Algorithm 4 (Model selection procedure)**.**

At step r of the iterative procedure, let $\mathcal{M}^r = \mathcal{M}(K_1^r,\ldots,K_q^r,\ldots K_Q^r)$ be the current model and ICL_r its corresponding ICL criterion.

(r.1) For any, $q = 1\ldots,Q$,

- *If $K_q + 1 \leq K_q^{(\max)}$, compute the ICL criterion of model $\mathcal{M}(K_1^r,\ldots,K_q^r+1,\ldots,K_Q^r)$ deriving from the VEM optimization using several initialization points. We denote by $ICL_{r,q,+}$ the obtained value.*
- *If $K_q \geq 2$, compute the ICL criterion of model $\mathcal{M}(K_1^r,\ldots,K_q^r-1,\ldots,K_Q^r)$ deriving from the VEM optimization using several initialization points. We denote by $ICL_{r,q,-}$ the obtained value.*

(r.2). Set

$$\mathcal{M}^{r+1} = \arg\max_{\mathcal{M}} \left\{ (ICL_{r,q,+})_{q|K_q+1\leq K_q^{(\max)}}, (ICL_{r,q,-})_{q|K_q\geq 2}, ICL_r, \right\}$$

(r.3) If $\mathcal{M}^{r+1} \neq \mathcal{M}^r$, go back to (r.1). Otherwise stop and set $\mathcal{M}^\star = \mathcal{M}^r$.

When talking about proposing several strategies for the initialization of the VEM at step $(r.1)$, we think about the strategy used by Leger (2015) in the R-package blockmodels, *i.e.* the likelihood maximization of model $\mathcal{M}(K_1^r, \ldots, K_q^r + 1, \ldots, K_Q^r)$ is initialized by dividing clusters in model $\mathcal{M}(K_1^r, \ldots, K_q^r, \ldots, K_Q^r)$. On the contrary, the likelihood maximization of model $\mathcal{M}(K_1^r, \ldots, K_q^r -, \ldots, K_Q^r)$ is initialized by merging clusters in model $\mathcal{M}(K_1^r, \ldots, K_q^r, \ldots, K_Q^r)$. Note that the tasks in (r.1) can easily be parallelized thus inducing a computational time reduction.

In order to promote the diffusion of our models in the ecology research community, we have to write the corresponding R-package. In particular, the visualization of the results is of prime importance. Moreover, the comparison with other clustering strategies standardly used in this community (such as modularity for community detection and nestedness) has to be discussed.

From an ecological point of view (*Application context 1*), the robustness of our model with respect to the definition of the functional groups has to be tested. Indeed, sometimes, the definition of the functional groups is not clearly established. We should study the robustness of our clustering method with respect to this misspecification. This work is highly linked to ecological aspects and so has to be discussed with my colleagues experts in the field.

4.2 Multilevel network : a new perspective?

Going back to the data described in Section 3, we had two functional groups, namely researchers (group number 1) and laboratories (group number 2). Whereas in [**A**15], we transformed the dataset into a multiplex network (using the fact that very few laboratories contained more than one researcher), we propose here to consider the original data. Let X^{11} be the advice matrix between researchers, X^{22} is the matrix of resources exchanges between laboratories and finally X^{12} is the matrix of affiliation : $X_{ij}^{12} = 1$ if researcher i belongs to laboratory j. The structure of the data is given in Table 3.4.

In this context too, we aim at co-clustering the laboratories and the researchers with respect to these three matrices. We adopt a block model strategy. Assume that researchers (resp. laboratories) are in K_1 (resp. K_2) classes . For any $q = 1, 2$, we denote by $(Z_i^q)_{i=1\ldots n_q}$ the independent random variables such that $Z_i^q = k$ if entity i of functional group q belongs to cluster k:

$$\mathbb{P}(Z_i^q = k) = \pi_k^q, \tag{3.23}$$

with $\sum_{k=1}^{K_q} \pi_k^q = 1$ for any $q = 1, \ldots Q$.
X^{11} and X^{22} being adjacency matrices, it is reasonable to assume that the $(X_{ij}^{11})_{ij}$ and $(X_{ij}^{22})_{ij}$ are independent conditionally to the latent variables, naturally leading to the SBM type model:

$$\begin{aligned} X_{ii'}^{11} | Z_i^1, Z_{i'}^1 &\sim_i \mathcal{B}ern\left(\alpha_{Z_i^1, Z_{i'}^1}^{11}\right) \\ X_{ij}^{22} | Z_j^2, Z_{j'}^2 &\sim_i \mathcal{B}ern\left(\alpha_{Z_j^2, Z_{j'}^2}^{22}\right) \end{aligned} \tag{3.24}$$

However, talking about the affiliation matrix X^{12}, the dataset we are interested in is such that the researchers can only belong to one laboratory. As a consequence, the independence assumption

	n_1	n_2
Researcher 1	1	
Researcher 2	1	1
⋮	$X^{11}_{ii'}$	X^{12}_{ij}
⋮		
Researcher n_1	1 1	1
Laboratory 1		1
⋮		$X^{22}_{jj'}$
Laboratory n_2		1
	⋯ ⋯	⋯
	Researcher 1 … Researcher n_1	Laboratory 1 … Laboratory n_2

Table 3.4 – Sociological multipartite network : structure of the data

.

of the $(X^{12}_{ij})_{i,j}$ conditionally to the $\mathbf{Z}$ is not realistic anymore. We have to define a different emission distribution on X^{12}. The one we consider is as follows. Let A_i be the affiliation of i: $A_i \in \{1, \dots n_2\}$ and $A_i = j$ if researcher i belongs to laboratory j. We set the following probability distribution :

$$\mathbb{P}(A_i = j | Z^1_i = k, Z^2_j = l, \mathbf{Z}^2) \quad = \quad \frac{\alpha^{12}_{kl}}{n_{2l}} \tag{3.25}$$

where $n_{2l} = \#\{j | Z^2_j = l\}$. In a few words, individual i (known to be in cluster k) has a probability α^{12}_{kl} to work in a laboratory of cluster l. Knowing that i works in a laboratory of cluster l, he may work in any of the n_{2l} laboratories of cluster l with equi-probability $\frac{1}{n_{2l}}$.
$\boldsymbol{\alpha} = (\alpha^{12}_{kl})_{k=1\dots K_1, l=1\dots K_2}$ is such that $\forall (k,l), \alpha^{12}_{kl} \in [0,1]$ and

$$\sum_{l=1}^{K_2} \alpha_{kl} = 1.$$

Finally, let $X^{12}_{i,1:n_2}$ be line i of X^{12}, we set:

$$X^{12}_{i,j} = \begin{cases} 1 & \text{if} \quad j = A_i \\ 0 & \text{if} \quad j \neq a_i \end{cases} \tag{3.26}$$

X^{12} is a deterministic transformation of $A = (A_1, \dots, A_{n_1})$: $X^{12} = \phi(A)$. As may be noticed in the DAG corresponding to this model – given in Figure 3.6 – the three matrices X^{11}, X^{12} and X^{22} will take part into the clustering.

Inference, model selection This model is a new way to handle multilevel networks, which are of high interest in social sciences (Snijders and Lazega, 2016). The block model structure is a standard tool. We propose here its extension to the context where we combine adjacency

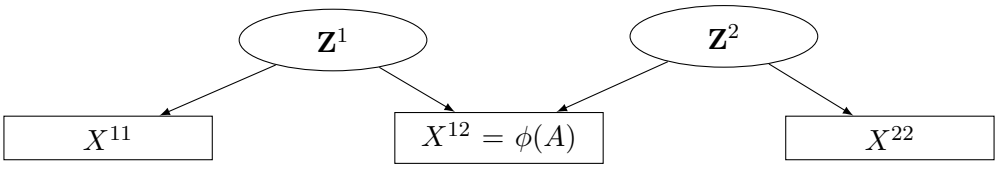


Figure 3.6 – DAG's corresponding to equations (3.23–3.26)

and affiliation matrices. Variational EM is a promising tool to maximize the likelihood. It has to be specially designed for this model. Similarly, the model selection criterion (to choose the number of clusters) has to be derived. The strategies to explore the models space developed for multipartite networks will be possibly reused in this context.

4.3 Other perspectives

Multiplex - multipartite networks The next natural extension is the multipartite networks where some adjacency matrices take into account several types of interactions. Such a model is expected in the MIRES group, where several types of interactions between farmers are studied (exchanges of seeds, social relations...) conjointly with agrodiversity in each farm property. The modeling step is straightforward, combining the models previously presented in Sections 3 and 4.1. Similarly, the variational EM algorithm can easily be written and the ICL derives naturally. The difficulties may arise from the practical implementation and from the interpretation of the results.

Temporal networks An other interesting issue deriving from the observation of (ecological or social) interactions is the evolution of their patterns along time. Two situations may occur.

- *Snapshot framework.* Either snapshots of the network are available at different times. For instance ecological networks such as plants/pollinators are sampled along months.
- *Continuous time framework.* Alternatively, the network may be observed continuously and the connections between entities i and j can occur at any time. For instance think about a network of subway stations and assume that there is a connection from station i to station j if a user starts its journey at station i and stops at station j.

The modeling in each case requires different probabilistic tools. I won't make here an exhaustive review of the existing literature on that subject. If needed, see the introductions of Matias and Miele (2017), Matias et al. (2017), Corneli et al. (2016b) or Corneli et al. (2016a) and references therein. However, I will suggest a few possible research perspectives.

1. From an ecological point of view, I am more interested in the *Snapshot framework*. Indeed, ecological interactions such as insect /plant interactions are (as far as I know) never observed continuously but are observed at different moments of the year. In this case, there is a need to understand the evolution of the connection patterns along time.

    - A first strategy is to infer a Latent Block Model at each time, thus giving rise to a co-clustering at each time and compare the different observed structures. The comparison in itself is a complex task. Comparisons can be helped by the alluvial flow diagrams, widely used in the ecologist community. Some statistics such as the Adjusted Rand Index (ARI) can give a clue on the proximity of the various clusterings but this is not sufficient to perform a precise interpretation. Some other tools have to be though of.

- A second option is to use a time-evolving block models, thus modeling the several networks conjointly. A time dependent SBM and its application to ecological datasets has been published recently by Matias and Miele (2017). A first interesting perspective would be its extension to LBM where we need a co-clustering (plants and insects) evolving through time. However, in ecology, the time is not important per-se but what matters is the modification of climate conditions through seasons. Among climate conditions we think about temperature, but also humidity rates... If considering a unidimensional variable (only temperature or only humidity), time-dependent methods can be directly transposed to this case (replacing the time by the humidity rate). Tools and models have to be developed if several variables have to be taken into account.
- Finally, comparing ecological through time leads to a problem of data sampling. Indeed, depending on the seasons, some plants are or are not present. As a consequence, this phenomena has to be taken into account in the model or at least kept in mind when interpreting the results.

2. From a statistical/methodological point of view, the modeling of continuously observed time-dependent network is an interesting and challenging problem requiring the use of counting processes. Considering the approached proposed by Matias et al. (2017), the sequence of the interaction occurrences between any two entities (i, j) along time can be modeled through counting processes. More precisely, the authors propose to use Poisson inhomogeneous processes whose intensity $\lambda_{ij}(t)$ only depends on a latent block structure, *i.e.* assuming that i et and j belong respectively to clusters k and l, then the occurrences sequence $T_{ij}^1 \dots T_{ij}^{N_{ij}}$ is the realization of a Poisson process of unknown intensity λ_{kl}. The occurrences sequences are independent processes conditionally to the clusters, the dependence is introduced when marginalizing over the latent blocks. Poisson processes seem well adapted to the modeling of transport network (city bikes or subway). However, when willing to consider emails networks, one may want to take into account the fact that sending en email may provoke a response, thus needing self-exciting processes. The Hawkes processes presented in Chapter 2 would be natural tools. Note that Hawkes processes have already been used in the network framework (see Blundell et al., 2012; Cho et al., 2013; Linderman and Adams, 2014, for instance). However, their combination with SBM structure is a new approach that could be thought of as a long-term working perspective.

Chapter 4

Autres perspectives et conclusion (en français)

Suite à mon recrutement à l'INRA dans l'équipe MORSE (MOdélisation et Risque en Statistique Environnementale) de l'Unité MIA Paris, j'ai choisi d'inscrire mes perspectives de recherche dans le domaine de la statistique pour l'écologie et l'environnement. Les outils méthodologiques que je considère sont ceux que j'ai pu développer ces dernières années (modèles définis par équations différentielles stochastiques, modèles pour données de réseaux, statistiques bayésiennes,etc.) mais je souhaite me concentrer dorénavant sur des projets essentiellement motivés par des collaborations avec des écologues ou des chercheurs en sciences de l'environnement. J'ai décrit à la fin de chaque chapitre mes perspectives de recherches se rapportant aux trois domaines évoqués. J'évoque ici quelques autres projets de recherche.

Je me suis impliquée dernièrement (avec des collègues de l'équipe MORSE) dans une collaboration avec le Museum d'Histoire Naturelle sur la modélisation de données de surveillance des dépôts métalliques atmosphériques par les mousses terrestres. Le dispositif BRAMM (Biosurveillance des Retombées Atmosphériques Métalliques par les Mousses) permet de cartographier et de suivre l'évolution, à l'échelle métropolitaine, des niveaux de concentrations en contaminants accumulés dans des mousses. Les mousses n'ont pas de système racinaire et absorbent directement les éléments présents dans l'air. Elles sont donc de bons capteurs des contaminants atmosphériques. Dans ce cadre, nous cherchons à modéliser la répartition spatio-temporelle des niveaux de concentrations en métaux accumulés dans les mousses. La difficulté (et l'originalité) de la modélisation à réaliser porte sur le caractère multi-éléments de la réponse ainsi que sur les localisations non concordantes des sources d'information.

Par ailleurs, je travaille avec des chercheurs de l'observatoire PELAGIS (UMS 3462, Université de La Rochelle / CNRS). Plus précisément, nous nous intéressons à des données collectées depuis 2004 sur les campagnes bateau MEGASCOPE dans l'Atlantique Nord-Est. Lors de ces campagnes opportunistes, les observateurs à bord de bateaux ou d'avions dénombrent les groupes de mammifères marins observés. De ces comptages seront déduites les estimations d'abondance des différentes espèces. Une étude rapide des données montre que les observateurs ne peuvent dénombrer exactement le nombre d'animaux dans les grands groupes; ils fournissent alors un comptage arrondi à la dizaine ou la centaine. Ne pas prendre en compte ces arrondis a pour effet de biaiser les estimations d'abondance. Notre premier travail consiste donc à modéliser les différents comportements d'arrondis pour les observateurs, puis étudier les propriétés statistiques des estimateurs d'abondance sous ces diverses façons d'arrondir les comptages de données groupées .

Enfin, j'ai profité de l'arrivée dans notre unité de Séverine Bord (issue de l'UMR INRA Épidémi-

ologie des Maladies Animales et Zoonotiques) pour m'intéresser à la répartition spatiale des tiques. En particulier, nous cherchons à comprendre la répartition des tiques en fonction des covariables du terrain afin de définir par la suite une méthode optimale d'échantillonnage.

Conclusion Mes travaux couvrent essentiellement trois domaines de compétence en statistiques (modèles définis par équations différentielles, processus de comptage et modélisation de réseaux) auxquels s'ajoutent des travaux non évoqués ici. J'ai donné en 2013 un tournant à ma carrière en étant recrutée comme chargée de recherches à l'INRA dans une équipe ayant pour thématique MOdélisation et Risque en Statistique Environnementale. Depuis, je cherche à donner une couleur "environnement" à mes travaux tout en mettant à profit mes compétences acquises jusque là en méthodologie statistique. Pour ce faire, je me suis impliquée dans diverses collaborations dans diverses domaines de l'environnement (écologie et biodiversité, animaux marins, etc.). Pour les années à venir, je souhaiterais poursuivre cette orientation, en me spécialisant sur certaines de ces applications.

Bibliography

S. Ahmed and N. Reid. *Empirical Bayes and Likelihood Inference*. Lecture Notes in Statistics, Springer, 2001. 44

C. Ambroise and C. Matias. New consistent and asymptotically normal parameter estimates for random-graph mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 74(1):3–35, 2012. 66

P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York, 1993. 35, 39

J. Andrés Christen, M. Capistrán, A. Monroy, S. Alavez, S. Quintana Vargas, H. A. Flores-Arguedas, and N. Kuschinski. A Diabetes minimal model for Oral Glucose Tolerance Tests. *ArXiv e-prints*, January 2016. 16

C. Andrieu, A. Doucet, and R. Holenstein. Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010. 28, 29

E. Bacry, K. Dayri, and J. F. Muzy. Non-parametric kernel estimation for symmetric hawkes processes. application to high frequency financial data. *The European Physical Journal B*, 85 (5):157, 2012. 53

E. Bacry, S. Gaïffas, and J.-F. Muzy. A generalization error bound for sparse and low-rank multivariate Hawkes processes. *ArXiv e-prints*, January 2015. 53

E. Bacry and J.-F. Muzy. First- and second-order statistics characterization of Hawkes processes and non-parametric estimation. *IEEE Trans. Inform. Theory*, 62(4):2184–2202, 2016. 53

V. Bally and D. Talay. The law of the Euler scheme for stochastic differential equations (I): convergence rate of the distribution function. *Probability Theory and Related Fields*, 104(1), 1996. 28

J.-P. Baudry, G. Celeux, and J.-M. Marin. Selecting models focussing on the modeller's purpose. In P. Brito, editor, *COMPSTAT 2008*, pages 337–348. Physica-Verlag HD, 2008. 68

E. Belitser, P. Serra, and H. van Zanten. Rate-optimal Bayesian intensity smoothing for inhomogeneous Poisson processes. *J. Statist. Plann. Inference*, 166:24–35, 2015. 43

J. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, second edition, 1985. 44

P. Bickel and A. Chen. A nonparametric view of network models and Newman-Girvan and other modularities. *JProc. Natl. Acad. Sci. U.S.A*, 106(50):21068–21073, 2009. 66

P. Bickel, D. Choi, X. Chang, and H. Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Statist.*, 41(4):1922–1943, 08 2013. 67

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(7):719–725, Jul 2000. 67

R. Biscay, J. C. Jimenez, J. J. Riera, and P. A. Valdes. Local linearization method for the numerical solution of stochastic differential equations. *Ann. Inst. Statist. Math.*, 48(4):631–644, 1996. 20

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational Inference: A Review for Statisticians. *arXiv*, pages 1–33, 2016. 70

C. Blundell, J. Beck, and K. A. Heller. Modelling reciprocating relationships with hawkes processes. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2600–2608. Curran Associates, Inc., 2012. 85

D. R. Brillinger. Maximum likelihood analysis of spike trains of interacting nerve cells. *Biological Cybernetics*, 59(3):189–200, 1988. 52

L. Carstensen, A. Sandelin, O. Winther, and N. Hansen. Multivariate hawkes process models of the occurrence of regulatory elements. *BMC Bioinformatics*, 2010. 52

G. Casella. An introduction to empirical Bayes data analysis. *The American Statistician*, 39: 83–87, 1985. 44

F. Chen. Maximum local partial likelihood estimators for the counting process intensity function and its derivatives. *Satistica Sinica*, 21:107–128, 2011. 39

S. Chen, D. Witten, and A. Shojaie. Nearly assumptionless screening for the mutually-exciting multivariate Hawkes process. *Electron. J. Stat.*, 11(1):1207–1234, 2017. 53

Y. Cho, A. Galstyan, P. J. Brantingham, and G. E. Tita. Latent point process models for spatial-temporal networks. *CoRR*, abs/1302.2671, 2013. 85

E. Chornoboy, L. Schramm, and A. Karr. Maximum likelihood identification of neural point process systems. *Biological cybernetics*, 59(4):265–275, 1988. 52

P. R. Conrad, M. Girolami, S. Särkkä, A. Stuart, and K. Zygalakis. Statistical analysis of differential equations: introducing probability measures on numerical solutions. *Statistics and Computing*, 27(4):1065–1082, 2017. 22

G. Consonni and J.-M. Marin. Mean-field variational approximate bayesian inference for latent variable models. *Comput. Stat. Data Anal.*, 52(2):790–798, October 2007. 70

M. Corneli, P. Latouche, and F. Rossi. Exact ICL maximization in a non-stationary temporal extension of the stochastic block model for dynamic networks. *Neurocomputing*, 192(Supplement C):81 – 91, 2016a. Advances in artificial neural networks, machine learning and computational intelligence. 84

M. Corneli, P. Latouche, and F. Rossi. Block modelling in dynamic networks with non-homogeneous poisson processes and exact ICL. *Social Network Analysis and Mining*, 6(1): 55, Aug 2016b. 84

D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes. Vol. I: Elementary Theory and Methods*. Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2003. 39, 52

W. Dáttilo, N. Lara-Rodríguez, P. Jordano, P. R. Guimarães, J. N. Thompson, R. J. Marquis, L. P. Medeiros, R. Ortiz-Pulido, M. A. Marcos-García, and V. Rico-Gray. Unravelling darwin's entangled bank: architecture and robustness of mutualistic networks with multiple interaction types. *Proceedings of the Royal Society of London B: Biological Sciences*, 283(1843), 2016. 77, 78

J. J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statistics and Computing*, 18(2):173–183, June 2008. 66

P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 68(3):411–436, 2006. 69, 70, 71

P. Del Moral, A. Doucet, and A. Jasra. An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, 22(5):1009–1020, sep 2012. 74

B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.*, 27:94–128, 1999. 18

A. P. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977. With discussion. 18

S. Donnet, V. Rivoirard, J. Rousseau, and C. Scricciolo. Posterior concentration rates for empirical Bayes procedures, with applications to Dirichlet Process mixtures. *arXiv:1406.4406v1*, June 2014. 46, 49

A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, 2001. 29

M. D. Fall and É. Barat. Gibbs sampling methods for Pitman-Yor mixture models. working paper or preprint, 2014. 45, 46

T. S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *Ann. Statist.*, 1(2): 209–230, 03 1973. 41

C. Fontaine, P. R. Guimarès, S. Kéfi, N. Loeuille, J. Memmott, W. H. van der Putten, F. J. F. van Veen, and E. Thébault. The ecological and evolutionary implications of merging different types of networks. *Ecology Letters*, 14(11):1170–1181, 2011. 77

S. Gazal, J.-J. Daudin, and S. Robin. Accuracy of variational estimates for random graph mixture models. *Journal of Statistical Computation and Simulation*, 82(6):849–862, 2012. 71

S. Ghosal and A. van der Vaart. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223, 2007. 43

S. Ghosal and A. van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017. 42

G. Govaert and M. Nadif. Block clustering with bernoulli mixture models: Comparison of different approaches. *Comput. Stat. Data Anal.*, 52(6):3233–3245, February 2008. 67

P. J. Green. Reversible jump Markov chain monte carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995. 54

G. Gusto and S. Schbath. FADO: a statistical method to detect favored or avoided distances between occurrences of motifs using the hawkes model. *Statistical Applications in Genetics and Molecular Biology*, 4:1–26, 2005. 39

N. R. Hansen, P. Reynaud-Bouret, and V. Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143, 2015. 52, 53, 61

N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker. *Bayesian Nonparametrics*. Cambridge University Press, Cambridge, UK, 2010. 40

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417, 1999. 69

H. Ishwaran and L. F. James. Computational methods for multiplicative intensity models using weighted gamma processes: proportional hazards, marked point processes, and panel count data. *Journal of the American Statistical Association*, 99:175–190, 2004. 45

F. Jaffrézic, C. Meza, M. Lavielle, and J. Foulley. Genetic analysis of growth curves using the SAEM algorithm. *Genetics Selection Evolution*, 38:583–600, 2006. 25

M. Kalli, J. E. Griffin, and S. G. Walker. Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105, 2011. 45

A. F. Karr. *Point processes and their statistical inference*, volume 2 of *Probability: Pure and Applied*. Marcel Dekker, Inc., New York, 1986. 39

S. Kéfi, V. Miele, A. Wieters, Evie, S. A. Navarrete, and E. L. Berlow. How structured is the entangled bank? the surprisingly simple organisation of multiplex ecological networks leads to increased persistence and resilience. *PLOS Biology*, 14(8):1–21, 08 2016. 77

C. Keribin, V. Brault, G. Celeux, and G. Govaert. Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, pages 1–16, 2014. 68

E. Kuhn and M. Lavielle. Coupling a stochastic approximation version of EM with a MCMC procedure. *ESAIM Probab. Stat.*, 8:115–131, 2004. 18

P. Latouche, S. Robin, and S. Ouadah. Goodness of fit of logistic models for random graphs. Technical report, arXiv:1508.00286, 2015. 68, 71

E. Lazega, M.-T. Jourda, L. Mounier, and R. Stofer. Catching up with big fish in the big pond? multi-level network analysis through linked design. *Social Networks*, 30(2):159 – 176, 2008. 74

J.-B. Leger. *blockmodels: Latent and Stochastic Block Model Estimation by a 'V-EM' Algorithm*, 2015. R package version 1.1.1. 75, 81, 82

J. Liepe, P. Kirk, S. Filippi, T. Toni, C. P. Barnes, and M. P. H. Stumpf. A Framework for Parameter Estimation and Model Selection from Experimental Data in Systems Biology Using Approximate Bayesian Computation. *Nature Protocols*, 2014. 29

S. W. Linderman and R. P. Adams. Scalable Bayesian Inference for Excitatory Point Process Networks. *ArXiv e-prints*, July 2015. 61

S. W. Linderman and R. P. Adams. Discovering latent network structure in point process data. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1413–II–1421. JMLR.org, 2014. 85

M. Mariadassou and C. Matias. Convergence of the groups posterior distribution in latent or stochastic block models. *Bernoulli*, 21(1):537–573, 02 2015. 68

M. Mariadassou, S. Robin, and C. Vacher. Uncovering latent structure in valued graphs: A variational approach. *The Annals of Applied Statistics*, 4(2):715–742, 06 2010. 67, 68

C. Matias and V. Miele. Statistical clustering of temporal networks through a dynamic stochastic block model. *JRSSB*, 79(4), 2017. 84, 85

C. Matias and S. Robin. Modeling heterogeneity in random graphs through latent space models: a selective review. *ESAIM: Proc.*, 47:55–74, 2014. 65, 66

C. Matias, T. Rebafka, and F. Villers. A semiparametric extension of the stochastic block model for longitudinal networks. working paper or preprint, July 2017. 84, 85

C. A. McGrory, A. N. Pettitt, D. M. Titterington, C. L. Alston, and M. Kelly. Transdimensional sequential Monte Carlo using variational Bayes – SMCVB. *Computational Statistics & Data Analysis*, 93:246–254, 2016. 70

T. P. Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, pages 362–369, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. 70

T. Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research Ltd, 2005. 72

P. Muliere and L. Tardella. Approximating distributions of random functionals of ferguson-dirichlet priors. *Canadian Journal of Statistics*, 26(2):283–297, 1998. 45

P. Müller and R. Mitra. Bayesian nonparametric inference : Why and how. *Bayesian Anal.*, 8 (2):269–302, 06 2013. 40

C. A. Naesseth, S. W. Linderman, R. Ranganath, and D. M. Blei. Variational Sequential Monte Carlo. *ArXiv e-prints*, May 2017. 70

R. M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001. 70, 71

K. Nowicki and T. A. B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001. 65, 66

Y. Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association.*, 83:9–27, 1988. 52

Y. Ogata. Seismicity analysis through point-process modelling: A review. *Pure and applied Geophysics*, 155:471–507, 1999. 39

M. Okatan, M. A. Wilson, and E. N. Brown. Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity. *Neural computation*, 17(9):1927–1961, 2005. 52

L. Paninski, J. Pillow, and J. Lewi. Statistical models for neural encoding, decoding, and optimal stimulus design. *Progress in brain research*, 165:493–507, 2007. 52

O. Papaspiliopoulos and G. O. Roberts. Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika*, pages 169–186, 2008. 45

F. Picard, S. Robin, E. Lebarbier, and J.-J. Daudin. A segmentation/clustering model for the analysis of array cgh data. *Biometrics*, 63(3):758–766, 2007. 30

J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. Chichilnisky, and E. P. Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, 2008. 52

J. Pinheiro and D. Bates. *Mixed-effect models in S and Splus*. Springer-Verlag, 2000. 17

M. Rabinovich, E. Angelino, and M. I. Jordan. Variational consensus Monte Carlo. *ArXiv e-prints*, June 2015. 70

J. G. Rasmussen. Bayesian inference for Hawkes processes. *Methodol. Comput. Appl. Probab.*, 15(3):623–642, 2013. 53

P. Reynaud-Bouret and S. Schbath. Adaptive estimation for Hawkes processes; application to genome analysis. *Ann. Statist.*, 38(5):2781–2822, 2010. 52

P. Reynaud-Bouret, V. Rivoirard, and C. Tuleau-Malot. Inference of functional connectivity in neurosciences via hawkes processes. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pages 317–320. IEEE, 2013. 52

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. 19

V. Robert. *Classification croisée pour l'analyse de bases de données de grandes dimensions de pharmacovigilance*. PhD thesis, Paris Saclay, 2017. 81

H. Rue, S. Martino, and N. Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009. 70

J.-B. Salomond. Concentration rate and consistency of the posterior distribution for selected priors under monotonicity constraints. *Electronic Journal of Statistics*, 8(1):1380–1404, 2014. 43, 44

T. Snijders and E. Lazega. *Multilevel Network Analysis for the Social Sciences: Theory, Methods and Applications*. Methodos Series. Springer, 2016. 74, 83

T. A. B. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *J. Classification*, 14(1):75–100, 1997. 66

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1–2):1–305, 2008. 70

S. G. Walker. Sampling the dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 36(1):45–54, 2007. 45, 47

M. West. Hyperparameter estimation in Dirichlet Process Mixture models. Technical report, Duke University, 1992. 46

R. E. Williamson. Multiply monotone functions and their Laplace transforms. *Duke Mathematical Journal*, 23(2):189–207, 1956. 41