Université de Montréal

Analyse bayésienne et classification pour modèles continus modifiés à zéro

par
Félix Labrecque-Synnott

Département de mathématiques et de statistique
Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)
en statistique

16 août, 2010

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée:

Analyse bayésienne et classification pour modèles continus modifiés à zéro

présentée par:

Félix Labrecque-Synnott

a été évaluée par un jury composé des personnes suivantes:

Alejandro Murua,	président-rapporteur
Jean-François Angers,	directeur de recherche
Mylène Bédard,	membre du jury
Jean-Philippe Boucher,	examinateur externe
William J. McCausland,	représentant du doyen de la FES

Thèse acceptée le: ....

RÉSUMÉ

Les modèles à sur-représentation de zéros discrets et continus ont une large gamme d'applications et leurs propriétés sont bien connues. Bien qu'il existe des travaux portant sur les modèles discrets à sous-représentation de zéro et modifiés à zéro, la formulation usuelle des modèles continus à sur-représentation – un mélange entre une densité continue et une masse de Dirac – empêche de les généraliser afin de couvrir le cas de la sous-représentation de zéros. Une formulation alternative des modèles continus à sur-représentation de zéros, pouvant aisément être généralisée au cas de la sous-représentation, est présentée ici. L'estimation est d'abord abordée sous le paradigme classique, et plusieurs méthodes d'obtention des estimateurs du maximum de vraisemblance sont proposées. Le problème de l'estimation ponctuelle est également considéré du point de vue bayésien. Des tests d'hypothèses classiques et bayésiens visant à déterminer si des données sont à sur- ou sous-représentation de zéros sont présentées. Les méthodes d'estimation et de tests sont aussi évaluées au moyen d'études de simulation et appliquées à des données de précipitation agrégées. Les diverses méthodes s'accordent sur la sous-représentation de zéros des données, démontrant la pertinence du modèle proposé.

Nous considérons ensuite la classification d'échantillons de données à sous-représentation de zéros. De telles données étant fortement non normales, il est possible de croire que les méthodes courantes de détermination du nombre de grappes s'avèrent peu performantes. Nous affirmons que la classification bayésienne, basée sur la distribution marginale des observations, tiendrait compte des particularités du modèle, ce qui se traduirait par une meilleure performance. Plusieurs méthodes de classification sont comparées au moyen d'une étude de simulation, et la méthode proposée est appliquée à des données de précipitation agrégées provenant de 28 stations de mesure en Colombie-Britannique.

Mots clés : sous-représentation de zéros, déflation à zéro, méthode d'agrégation bayésienne, précipitations agrégées, distribution de Laplace tronquée, algorithme EM, modèles de mélanges.

ABSTRACT

Zero-inflated models, both discrete and continuous, have a large variety of applications and fairly well-known properties. Some work has been done on zero-deflated and zero-modified discrete models. The usual formulation of continuous zero-inflated models – a mixture between a continuous density and a Dirac mass at zero – precludes their extension to cover the zero-deflated case. We introduce an alternative formulation of zero-inflated continuous models, along with a natural extension to the zero-deflated case. Parameter estimation is first studied within the classical frequentist framework. Several methods for obtaining the maximum likelihood estimators are proposed. The problem of point estimation is considered from a Bayesian point of view. Hypothesis testing, aiming at determining whether data are zero-inflated, zero-deflated or not zero-modified, is also considered under both the classical and Bayesian paradigms. The proposed estimation and testing methods are assessed through simulation studies and applied to aggregated rainfall data. The data is shown to be zero-deflated, demonstrating the relevance of the proposed model.

We next consider the clustering of samples of zero-deflated data. Such data present strong non-normality. Therefore, the usual methods for determining the number of clusters are expected to perform poorly. We argue that Bayesian clustering based on the marginal distribution of the observations would take into account the particularities of the model and exhibit better performance. Several clustering methods are compared using a simulation study. The proposed method is applied to aggregated rainfall data sampled from 28 measuring stations in British Columbia.

Keywords: zero-deflation, aggregate rainfall, truncated Laplace, Bayesian aggregation, EM algorithm, mixture models

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES

cdf	fonction de répartition (*cumulative distribution function*)
EMV	estimateur du maximum de vraisemblance
MAP	maximum *a posteriori*
MLE	estimateur du maximum de vraisemblance (*maximum likelihood estimator*)
MMZ	modèle modifié à zéro (*zero-modified model*)
pmf	fonction de masse (*probability mass function*)
pdf	fonction de densité (*probability density function*)
ZDM	modèle dégonflé à zéro (*zero-deflated model*)
ZIM	modèle gonflé à zéro (*zero-inflated model*)
ZMM	modèle modifié à zéro (*zero-modified model*)

NOTATION

$f_1(x\|\theta)$	la fonction de densité de base
$f_0(x)$	la fonction de densité de modification
$\mathbb{R}^+$	les nombres réels non négatifs
$n_{samples}$	le nombre d'échantillons considérés en classification
θ	les paramètres de la densité de base
Θ	l'espace paramétrique de la densité de base
ω	les paramètres d'intérêt (θ, p)
Ω	l'espace paramétrique sous le MMZ
$\|\mathcal{A}\|$	la cardinalité de l'ensemble $\mathcal{A}$.

À la mémoire d’Eugénie et d’Onias.

REMERCIEMENTS

J'aimerais tout d'abord remercier mon directeur de thèse, Jean-François Angers, pour son soutient et ses conseils tout au long de mes études doctorales.

Je remercie également Yves Lepage, m'ayant donné une première appréciation de la théorie statistique, et dont le cours me convainquit finalement d'opter pour cette spécialisation au baccalauréat, pour ensuite continuer aux cycles supérieurs. Je voudrais également souligner l'excellent travail du personnel administratif du département. Merci également à mes parents et ma soeur Delphine pour leurs appuis et encouragements au long de ces quatre années, ainsi qu'à Éloïse et André.

Merci au FQRNT, au CRSNG, au Département de mathématiques et de statistique et à la Faculté des études supérieures et postdoctorales pour les bourses m'ayant été octroyées lors de mes études doctorales. Je souhaite aussi remercier Jean-François et Véronique Hussin pour m'avoir donné l'opportunité d'une première expérience d'enseignement universitaire, et André Montpetit pour ses conseils relatifs à l'utilisation de LaTeX et pour m'avoir donné une belle expérience de travail en édition.

INTRODUCTION

Nous traitons ici de modèles continus modifiés à zéro, utilisés lorsque la proportion de zéros dans les observations diffère de ce qui serait prévu par un certain modèle.. La plupart des modèles modifiés à zéro rencontrés dans la littérature sont des modèles discrets : en effet, les modèles continus comportent certaines particularités rendant moins aisée l'utilisation de modèles modifiés à zéro. En particulier, la formulation usuelle des modèles continus gonflés à zéro ne permet pas de traiter également le cas d'un plus faible taux de zéros dans les observations. Nous proposons une nouvelle formulation, sous laquelle un modèle de base sera modifié sur un intervalle autour de zéro. À la limite, lorsque la taille de l'intervalle approche de zéro, la formulation classique des modèles continus gonflés à zéro est retrouvée. La nouvelle formulation permet le traitement de la sous-représentation de zéros, et sert également à la modélisation lorsque a proportion d'observations dans un intervalle autour de zéro diffère de la proportion prévue par un modèle donné. Cette formulation proposée est donc plus générale que la formulation classique des modèles continus gonflés à zéro.

Le développement de ce modèle est motivé par l'analyse de précipitations. La proportion de zéros dans ces données varie énormément selon le pas de temps considéré (des données journalières ayant une forte proportion de zéros, et des données mensuelles en ayant très peu). Ce problème de la proportion de zéros dans les données de précipitation est complexifié par le fait que de très faibles précipitations peuvent ne pas être détectées par les pluviomètres, ou n'être enregistrées que comme « traces » de précipitations. La documentation d'Environnement Canada (`http://www.climate.weatheroffice.gc.ca/prods\_servs/index\_e.html\#cdcd`) mentionne qu'une valeur de 0 mm de précipitation peut correspondre à une absence de précipitations (le cas le plus fréquent), mais aussi à des traces de précipitations, des précipitations de quantité incertaine ou encore à une possibilité de précipitation (sans qu'il soit possible de trancher dans un sens ou dans l'autre). Des données de ce type sont analysées afin d'illustrer la classe de modèles proposée.

Cette thèse traite de modèles continus modifiés à zéro (MMZ, en anglais *zero-modified models,* ou *ZMM*). La classe des MMZ comprend les modèles à sur-représentation de zéros (ou modèles gonflés à zéro, en anglais *zero-inflated models* ou *ZIM*) et les modèles à sous-représentation de zéros (ou dégonflés à zéro, en anglais *zero-deflated models* ou *ZDM*), correspondant respectivement à une plus forte et une plus faible proportion de zéros dans les données que ce que nous donnerait un modèle de base.

Considérons d'abord le cas discret, plus fréquemment traité dans la littérature. Si X dénote une variable aléatoire distribuée d'après un modèle (discret) modifié à zéro, alors elle sera de fonction de masse

$$f_X(x|p) = \begin{cases} p + (1-p)f_1(0|\theta) & \text{au point } x = 0 \\ (1-p)f_1(x|\theta) & \text{partout ailleurs,} \end{cases} \tag{1}$$

où $f_1(x|\theta)$ dénote la fonction de masse de base, dépendant d'un paramètre θ habituellement inconnu. Puisque p sera généralement aussi un paramètre inconnu, nous dénotons, tout au long de cette thèse et tant pour le cas discret que continu, l'ensemble des paramètres d'intérêt (p, θ) par ω. Similairement, alors que Θ représente l'espace paramétrique associé au modèle de base f_1, nous dénotons Ω l'espace paramétrique du modèle modifié à zéro (le produit cartésien de Θ et de l'ensemble des valeurs possibles pour p).

Si p est positif, et ce tant pour le cas discret que continu, l'équation (1) est un mélange entre une masse de Dirac au point 0 et la fonction de masse de base. Ce paramètre peut aussi être vu comme une quantité ajoutée à la probabilité que X soit identiquement égale à zéro. Fréquent dans la littérature scientifique, ce type de modèle fut d'abord présenté dans Singh (1963). Il s'agissait alors d'une loi de Poisson avec sur-représentation de zéros. L'estimation par maximum de vraisemblance (EMV) et ses propriétés asymptotiques pour ce modèle sont présentées dans El-Shaarawi (1985). De là, l'usage de ces modèles s'est répandu, et des modèles plus complexes furent construits sur cette première fondation (Lambert, 1992, Hall, 2000, Ridout *et al.*, 2001, Dalrymple *et al.*, 2003, Ghosh *et al.*, 2006, Rodrigues, 2003, Van den Broek, 1995, Jansakul et Hinde, 2002, Gupta *et al.*, 2005, Hasan et Sneddon, 2009).

La fonction de masse (1) demeure valide si p prend des valeurs négatives, mais supérieures à $-f_1(0|\theta)/[1-f_1(0|\theta)], f_1(0|\theta) \neq 1$. La littérature scientifique est très peu abondante à ce sujet ; en fait, bien que certain articles traitant de modèles discrets modifiés à zéros mentionnent le cas de la sous-représentation en plus de celui de la sur-représentation, aucun n'y est entièrement dédié. Des distributions de Poisson modifiées à zéro sont toutefois abordées dans Angers et Biswas (2003) et Dietz et Böhning (2000). Il faut noter qu'il ne s'agit plus d'un modèle de mélanges, et il est uniquement possible d'interpréter p comme une quantité retranchée à la probabilité que X prenne la valeur 0.

Dans le cas de modèles continus ($f_1(x|\theta)$ est alors une fonction de densité), seul le cas de la sur-représentation de zéros est mentionné dans la littérature. L'approche utilisée dans ce cas est de prendre comme distribution pour X une masse de Dirac au point 0 avec probabilité p et la densité de base $f_1(x|\theta)$ avec probabilité $(1-p)$. Le modèle (1) n'est alors plus continu, mais mixte. Il est aussi possible d'interpréter p comme la probabilité que X soit identiquement égale à 0. Bien que moins fréquemment rencontrés dans la littérature que les modèles discrets gonflés à zéro, les modèles continus gonflés à zéro, présentées dans Aitchison (1955), ont toutefois diverses applications, principalement en sciences biologiques (Lo *et al.*, 1992, Stefansson, 1996) et en modélisation de précipitations (Fernandes *et al.*, 2009, Feuerverger, 1979). Notons qu'aucune des deux interprétations possibles pour p dans ce cas ne permet de considérer des valeurs négatives pour ce paramètre.

Le but premier de cette thèse est de développer une formulation alternative pour les modèles continus gonflés à zéro pouvant aussi s'adapter au cas de la sous-représentation. La formulation proposée, à la limite, devient équivalente à la formulation usuelle. Les modèles continus modifiés à zéro constituent donc le lien unissant ces trois articles. Le premier article, après un relevé de la littérature existante au sujet des modèles discrets à sur- et sous-représentation de zéros et des modèles continus à sur-représentation de zéros, présente le nouveau modèle proposé. Celui-ci est continu et permet de traiter le cas de la sous-représentation de zéros. L'estimation par maximum de vraisemblance y est aussi abordée. Le second article en constitue la prolongation directe, traitant d'estimation ponctuelle bayésienne et de tests d'hypothèses permettant de tester pour la sur- ($p > 0$), sous- ($p < 0$) représentation de zéros, ou encore la non modification ($p = 0$). Le troisième article est centré sur une application plus complexe de cette classe de modèles. Il s'agit de la création de régions homogènes à partir de données continues modifiées à zéro, c'est-à-dire de la classification d'échantillons de tels données, sous une contrainte de contiguïté. En effet, nous souhaitons que les régions ainsi soient contigües (c'est-à-dire non disjointes). Ce dernier article explore donc différents critères d'arrêt pour la classification hiérarchique agglomérative dans ce contexte. La classe des modèles continus modifiés à zéro constitue donc le principal fil conducteur de la thèse, les deux premiers articles la développant et en explorant l'estimation, alors que le troisième article en présente une application plus complexe.

Les applications considérées sont le second fil conducteur de la thèse. Le développement de la classe des modèles continus modifiés à zéro fut principalement motivé par certains types de données – les précipitations agrégées sur 7, 14 ou 30 jours – et les deux premiers articles présentent une application des méthodes d'estimation proposées à un tel jeu de données. Cette application motive aussi en partie le choix de la densité de base $f_1(x|\theta)$ proposée dans les trois articles, et utilisée lors des applications présentées : les observations sont continues, positives, comportent des zéros, et semblent être dégonflées à zéro. De telles applications nécessitent donc une distribution continue, définie sur les réels positifs, et dont la fonction de densité n'est pas nulle au point 0, ce qui exclut potentiellement les distributions gamma et lognormale. La distribution de Laplace (ou double exponentielle) tronquée sur les réels positifs satisfait à ces critères.

Les deux premiers articles se terminent par l'application des méthodes d'estimation ponctuelles à un jeu de données de précipitations montréalaises bimensuelles. Outre la cohérence entre les résultats donnés par les différentes méthodes lorsqu'appliquées à ces données, soulignons que l'utilisation de critères d'information (critère d'information d'Akaike (Akaike, 1974), critère d'information bayésien (Schwarz, 1978), tous deux fréquemment utilisés en sélection de modèles (Kuha, 2004, Yang, 2005, Burnham et Anderson, 2004)) et de tests d'hypothèses (test du rapport de vraisemblance, test de score, et les tests bayésiens présentés au second article) permet de conclure que ces données sont bel et bien modifiées à zéro. Nous entendons par là que l'augmentation de la vraisemblance lorsque nous passons du modèle de base (Laplace tronquée sur les réels positifs) au modèle modifié à zéro est assez grande pour justifier la plus grande complexité du modèle (telle que définie par l'AIC et le BIC). En outre, l'hypothèse $p = 0$, correspondant à un modèle non modifié à zéro, est rejetée indépendamment du test considéré.

Les deux premiers articles ayant confirmé la pertinence de la classe de modèles proposée et son applicabilité à la modélisation de données de précipitations agrégées, le troisième propose de s'attaquer à un problème plus complexe (la classification d'échantillons de données continues modifiées à zéro). Nous considérons plusieurs échantillons, correspondant à des point géographiques donnés, et nous nous intéressons à la classification de ces échantillons en grappes homogènes et contigües (ne contenant pas de sites géographiquement disjoints). Dans ce contexte, nous privilégions l'utilisation de méthodes de classification hiérarchiques agglomératives. Si nous avons des mesures provenant de $n_{samples}$ sites, nous considérerons d'abord chacun de ces $n_{samples}$ échantillons comme appartenant à une grappe différente. Les méthodes de classification hiérarchiques nécessitent alors, itérativement, de construire une matrice de distance (ou de dissimilarité) entre les grappes et de tester si les deux grappes les moins distantes peuvent être combinées. Il faut noter ici que la matrice de distance ne fait pas référence à la distance géographique mais bien à une façon donnée de calculer des distances entre groupes d'observations. Par exemple, on pourrait prendre pour distance entre deux grappes la distance euclidienne entre les moyennes des observations comprises dans ces grappes. Cette approche est privilégiée dans ce contexte, puisqu'il sera plus intuitif de vérifier que les grappes à combiner sont bel et bien adjacentes (classification hiérarchique agglomérative) que de vérifier que le partitionnement d'une grappe en deux ne crée pas de « trous » dans l'une des deux grappes résultantes (classification hiérarchique divisive).

Cette problématique est motivée par l'analyse de données de précipitations agrégées sur un territoire géographique donné (plutôt qu'en un seul point donné, comme c'était le cas dans l'application présentée dans les deux premiers articles). Les précipitations agrégées sur une base hebdomadaire, bimensuelle ou mensuelle étant utilisées en planification agronome (Azhar *et al.*, 1992, Sharda et Das, 2005), en prévision d'écoulement des rivières (Dibike et Solomatine, 2001), en gestion de bassin-versants (Raghuwanshi *et al.*, 2006) et en estimation d'abondance d'espèces (Eklundh, 1998, Peco *et al.*, 1998), la classification de différents sites sur un territoire donné en régions homogènes en termes de précipitations agrégées serait d'intérêt pour plusieurs domaines d'application. Plusieurs stations de mesures en Colombie-Britannique sont considérées, ce territoire étant choisi pour ses précipitations abondantes et son relief intéressant.

RÉFÉRENCES

Aitchison, J. (1955). On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of the American Statistical Association*, **50**(271):901–908.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, **19**(6):716–723.

Angers, J. et Biswas, A. (2003). A Bayesian analysis of zero-inflated generalized Poisson model. *Computational statistics & data analysis*, **42**(1-2):37–46.

Azhar, A., Murty, V. et Phien, H. (1992). Modeling irrigation schedules for lowland rice with stochastic rainfall. *Journal of Irrigation and Drainage Engineering*, **118**:36.

Burnham, K. et Anderson, D. (2004). Multimodel inference : understanding AIC and BIC in model selection. *Sociological Methods & Research*, **33**(2):261.

Dalrymple, M. L., Hudson, I. L. et Ford, R. P. K. (2003). Finite mixture, zero-inflated Poisson and hurdle models with application to SIDS. *Computational Statistics & Data Analysis*, **41**(3-4):491–504.

Dibike, Y. et Solomatine, D. (2001). River flow forecasting using artificial neural networks. *Physics and Chemistry of the Earth, Part B : Hydrology, Oceans and Atmosphere*, **26**(1):1–7.

Dietz, E. et Böhning, D. (2000). On estimation of the Poisson parameter in zero-modified Poisson models. *Computational statistics & data analysis*, **34**(4):441–459.

Eklundh, L. (1998). Estimating relations between AVHRR NDVI and rainfall in East Africa at 10-day and monthly time scales. *International Journal of Remote Sensing*, **19**(3):563–570.

El-Shaarawi, A. (1985). Some goodness-of-fit methods for the Poisson plus added zeros distribution. *Applied and environmental microbiology*, **49**(5):1304.

Fernandes, M., Schmidt, A. et Migon, H. (2009). Modelling zero-inflated spatio-temporal processes. *Statistical Modelling*, **9**(1):3.

Feuerverger, A. (1979). On some methods of analysis for weather experiments. *Biometrika*, **66**(3):655–658.

Ghosh, S. K., Mukhopadhyay, P. et Lu, J.-C. (2006). Bayesian analysis of zero-inflated regression models. *Journal of Statistical Planning and Inference*, **136**(4): 1360–1375.

Gupta, P., Gupta, R. et Tripathi, R. (2005). Score test for zero inflated generalized Poisson regression model. *Communications in Statistics-Theory and Methods*, **33**(1):47–64.

Hall, D. (2000). Zero-inflated Poisson and binomial regression with random effects : a case study. *Biometrics*, **56**(4):1030–1039.

Hasan, M. et Sneddon, G. (2009). Zero-inflated Poisson regression for longitudinal data. *Communications in Statistics-Simulation and Computation*, **38**(3):638–653.

Jansakul, N. et Hinde, J. (2002). Score tests for zero-inflated Poisson models. *Computational statistics & data analysis*, **40**(1):75–96.

Kuha, J. (2004). AIC and BIC : Comparisons of assumptions and performance. *Sociological Methods & Research*, **33**(2):188.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**(1):1–14.

Lo, N., Jacobson, L. et Squire, J. (1992). Indices of relative abundance from fish spotter data based on delta-lognormal models. *Canadian Journal of Fisheries and Aquatic Sciences*, **49**(12):2515–2526.

Peco, B., Espigares, T. et Levassor, C. (1998). Trends and fluctuations in species abundance and richness in Mediterranean annual pastures. *Applied Vegetation Science*, :21–28.

Raghuwanshi, N., Singh, R. et Reddy, L. (2006). Runoff and sediment yield modeling using artificial neural networks : Upper Siwane River, India. *Journal of Hydrologic Engineering*, **11**:71.

Ridout, M., Hinde, J. et Demétrio, C. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, **57**(1):219–223.

Rodrigues, J. (2003). Bayesian analysis of zero-inflated distributions. *Communications in Statistics-Theory and Methods*, **32**(2):281–289.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2):461–464.

Sharda, V. et Das, P. (2005). Modelling weekly rainfall data for crop planning in a sub-humid climate of India. *Agricultural water management*, **76**(2):120–138.

Singh, S. (1963). A note on inflated Poisson distribution. *Journal of the Indian Statistical Association*, **1**(3):140–144.

Stefansson, G. (1996). Analysis of groundfish survey abundance data : combining the GLM and delta approaches. *ICES Journal of Marine Science*, **53**(3):577.

Van den Broek, J. (1995). A score test for zero inflation in a Poisson distribution. *Biometrics*, **51**(2):738–743.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared ? A conflict between model indentification and regression estimation. *Biometrika*, **92**(4):937.

CHAPTER 1

AN EXTENSION OF ZERO-MODIFIED MODELS TO THE CONTINUOUS CASE

Cet article a été soumis pour publication à *Metron, International journal of statistics.* Le premier auteur est Félix Labrecque-Synnott et le coauteur est le directeur de recherche, Jean-François Angers.

De façon générale, le premier auteur était responsable de la majeure partie des travaux de recherche et de rédaction derrière cet article. Le seul coauteur est le directeur de recherche, et son rôle était principalement d'apporter un encadrement à la recherche et à la rédaction. Plus spécifiquement, l'élaboration de la problématique de recherche à la base de la thèse (les modèles continus modifiés à zéro) s'est faite au cours de multiples conversations. Par la suite, le premier auteur a travaillé à la formulation précise et formelle du modèle, ce qui constitua la base de cet article. Il a obtenu la borne inférieure sur p comme fonction de θ, et a considéré plusieurs choix possibles pour la densité de modification f_0. Cette dernière a une grande influence sur la complexité de la borne p_{min} en tant que fonction de θ, certains choix nous donnant une forme explicite pour la borne, alors que d'autres nécessitent des calculs numériques. Le premier auteur a également travaillé à l'estimation par maximum de vraisemblance, nécessitant généralement de passer par des méthodes numériques. Un estimateur analytique a toutefois été obtenu pour un certain choix de f_0. Il a également obtenu une façon de retrouver un modèle de mélange même dans le cas de la sous-représentation de zéros, permettant ainsi l'utilisation de méthodes classiques, fréquemment mentionnées dans la littérature, et utilisées pour le cas de modèles à sur-représentation de zéros de formulation habituelle. Finalement, s'étant chargé de l'algorithmique et de la programmation nécessaires aux différentes méthodes d'estimation présentées au premier article (l'une d'entre elles est toutefois fortement basée sur du code *Matlab* existant), il a élaboré les différents scénarios de simulation utilisés afin de comparer ces méthodes. À la suggestion du second auteur, ils se sont penchés sur les précipitations comme application pratique de la classe de modèles proposés.

La formulation proposée et ses propriétés, incluant la borne inférieure pour p, se retrouvent à la section 2 de cet article. La section 3 traite d'estimation par le maximum de vraisemblance; on y retrouve aussi l'EMV analytique pour p conditionnel à θ obtenu pour un choix particulier de f_0, ainsi qu'une méthode permettant de retrouver un modèle de mélanges même en cas de dégonflement à zéro. La section 4 porte sur les propriétés asymptotiques du modèles et la matrice d'information de Fisher. La section 5 comprend des résultats de simulation, et la section 6, une application à des données de précipitations bimensuelles.

ABSTRACT

Zero-inflated models, both discrete and continuous, have a large variety of applications and fairly well-known properties. Some work has been done on zero-deflated and zero-modified discrete models. The usual formulation of continuous zero-inflated models - a mixture between a continuous density and a Dirac mass at zero - precludes their extension to cover the zero-deflated case. An alternative formulation of zero-inflated continuous models is introduced, along with a natural extension to the zero-deflated case. Likelihood-based estimation is discussed. The model and estimation methods are illustrated with simulation results.
Keywords: zero-modified model, zero-deflated model, truncated Laplace, EM algorithm, aggregate precipitation data

1.1 Introduction

Zero-modified models (ZMMs) are used when the number of zeros observed is higher or lower than what can be explained by a given model. Let X be a discrete random variable. Under a zero-modified model, its probability mass function $f_X(x|p)$ will take the form:

$$f_X(x|p) = \begin{cases} p + (1-p)f_1(0) & \text{if } x = 0 \\ (1-p)f_1(x) & \text{otherwise,} \end{cases} \tag{1.1}$$

where $f_1(x)$ is a probability mass function defined on $\mathbb{N}$, and p takes values between $-f_1(0)/[1 - f_1(0)], f_1(0) \neq 1$ and 1. For p to take negative values, we must have that $f_1(0) > 0$. If the value of p is positive, the probability of observing zeroes is higher than $f_1(0)$ and the model is said to be zero-inflated. Similarly, negative values of p correspond to a lower probability mass function at zero and the model is then said to be zero-deflated. If p can take positive or negative values, the model is said to be zero-modified. Whether p is positive or negative, it is trivial to see that $\sum_{\mathbb{N}} f_X(x|p) = 1$.

It should also be noted that the zero-inflated model is a mixture model. Estimation methods developed for mixtures can therefore be used. The most common estimations methods used in this case, such as the EM algorithm (Muthén and Shedden, 1999) and the Gibbs sampler (Diebolt and Robert, 1994), are based on a latent variable approach. This type of model has a fairly widespread use. Introduced in Singh (1963), the zero-inflated Poisson distribution is especially common in the literature. El-Shaarawi (1985) obtained the maximum likelihood estimator for this model, as well as its asymptotic distribution. Lambert (1992) extended the family of zero-inflated models, using zero-inflated Poisson regression to model manufacturing defects. Hall (2000) considered zero-inflated Poisson and binomial regression models, and Ridout et al. (2001) discussed zero-inflated negative binomial regression models. Parameter estimation is often based on the interpretation of the model (1.1) as a mixture. Methods based on an expectation-maximization algorithm are used in Lambert (1992), Hall (2000), and Dalrymple et al. (2003) while Ghosh et al. (2006) and Rodrigues (2003) use the Gibbs sampler. Quasi-likelihood is also used in Hasan and Sneddon (2009). Score tests for zero-inflation have been developed for the Poisson (Van den Broek, 1995), Poisson regression (Jansakul and Hinde, 2002), and generalized Poisson regression (Gupta et al., 2005) models. Recent results comparing the performance of the score, likelihood ratio and Wald tests can be found in Min and Czado (2010).

The mass function (1.1) remains valid if p takes negative values larger than $-f_1(0)/[1 - f_1(0)], f_1(0) \neq 1$. Literature on this subject is less common than on zero-inflated models – no paper focuses solely on zero-deflated models – but zero-modified Poisson distributions are discussed in Angers and Biswas (2003) and Dietz and Böhning (2000). In this case, the interpretation of the probability mass function as a mixture is lost, as a negative mixture probability has no sense, and p simply represents the probability "removed" from the point 0 and "redistributed" amongst the other integers. The approaches most commonly used for positive values of p are thus impossible to use. Instead, Dietz and Böhning (2000) proposes a two-step method where an estimate for the Poisson parameter is first obtained by neglecting zeroes and maximizing the likelihood of the zero-truncated Poisson distribution, and p is then estimated by replacing the Poisson parameter by its estimate in the likelihood equation. Angers and Biswas (2003) consider the problem within the Bayesian framework, and use Monte-Carlo integration with importance sampling to estimate the posterior mean of the model parameters, iteratively applying a location-scale transformation to the random vector so that it is "more likely to be in the appropriate region of the parameter space."

In the continuous case, only positive values of p have been featured so far in the literature. The probability density function of X is a Dirac mass at 0 with probability p, and a density $f_1(x)$ defined on $\mathbb{R}^+$ with probability $1-p$. This type of model was introduced in Aitchison (1955), where the zero-inflated lognormal, exponential and Pearson Type III distributions were considered. Most often fitted with the gamma (Feuerverger, 1979, Stefansson, 1996) or lognormal (Fletcher, 2008, Tian, 2005) distributions, it is mostly applied to the life sciences (abundance data for fish or plankton, (Lo et al., 1992, Stefansson, 1996)). It is also used to model rainfall data (Fernandes et al., 2009, Feuerverger, 1979). Another possible way to model zero-inflated continuous data, within the context of geo-referenced data, is to use a compound Poisson process (Ancelet et al., 2009). It is impossible to take the zero-inflated continuous model and set $p < 0$ to obtain a zero-deflated model: in the continuous case, p can only be interpreted as the probability for X to be equal to 0, and a probability cannot be negative. Furthermore, it is also impossible to obtain a continuous zero-deflated model simply by lowering the value of $f_1(0)$ (much as it is impossible to obtain a zero-inflated model by raising $f_1(0)$), as changing the value of a probability density function at a single point has no effect on a continuous model.

In this paper, we propose an alternative formulation of zero-modified models in the continuous case, where inflation or deflation occurs on a small interval rather that at a single point. This can be seen as an extension of the classical formulation, which can be recovered in the limit as the interval length goes to zero. Different approaches to maximum likelihood estimation are proposed and compared using simulation studies. This work is mainly motivated by the analysis of rainfall data. While daily rainfall often presents an excess of zeroes and is therefore modelled by zero-inflated models (Fernandes et al., 2009), total rainfall over a longer period of time could conversely be expected to be zero-deflated, especially during periods of the year or in locations associated with higher than average rainfall. Aggregate rainfall data is used in agricultural planning (Azhar et al., 1992, Sharda and Das, 2005), in river flow forecasting (Dibike and Solomatine, 2001) and watershed management (Raghuwanshi et al., 2006), and to assess the abundance of species and vegetation (Eklundh, 1998, Peco et al., 1998).

The outline of this paper is as follows: the model and its features are introduced in Section 2. Section 3 discusses maximum likelihood estimation. Asymptotic properties of the model are discussed in Section 4. Section 5 presents simulation results to illustrate and compare estimation methods. In Section 6, the model is fitted to real data. Concluding remarks are given in the last section.

1.2 The model

Let X be a non-negative random variable which can be modelled by a probability density function $f_1(x|\theta)$ on $[0, \infty)$, and suppose that $X \in [0, x_0]$ is observed with a different frequency than indicated by $f_1(x|\theta)$. The probability density function of X could then be written as

$$f(x|\theta, p, x_0) = p \times f_0(x) + (1 - p) \times f_1(x|\theta), \tag{1.2}$$

where $f_0(x)$ is a probability density function on $[0, x_0]$ and p can take positive and negative values, corresponding respectively to higher and lower proportions of observations in $[0, x_0]$. We consider here x_0 to be known and $f_0(x)$ to be entirely specified. The parameters of interest are $\omega = (p, \theta)$. Positive values of p also correspond to a mixture of densities $f_0(x)$ and $f_1(x|\theta)$.

Since $f_0(x)$ is a probability density function on $[0, x_0]$, we must have $\int_0^{x_0} f_0(x)dx =$

1, for any positive value of x_0. Therefore, we have that

$$\lim_{x_0 \to 0} f_0(x) = \begin{cases} +\infty \text{ if } x = 0, \\ 0 \text{ otherwise.} \end{cases}$$

For positive values of p, the limit of the proposed model as x_0 goes to 0 is thus a mixture between a Dirac mass at 0 and the probability density function $f_1(x|\theta)$, which is the usual continuous zero-inflated model.

For $f(x|\theta, p)$ to be continuous at the point x_0, $f_0(x)$ must verify: $\lim_{x \to x_0} f_0(x) = 0$. Otherwise, a discontinuity appears at x_0, which could be hard to interpret or to justify in most applications.

As in the discrete zero-deflated model, p must be smaller or equal to 1 and greater or equal to a lower bound p_{min} for $f(x|\theta, p)$ to be non-negative, and thus be a valid probability density function. In particular, we must have:

$$\begin{aligned} pf_0(x) + (1-p)f_1(x|\theta) \geq 0 \; \forall \; x \in [0, x_0] \\ p(f_0(x) - f_1(x|\theta)) \geq -f_1(x|\theta) \; \forall \; x \in [0, x_0] \\ p \geq \frac{-f_1(x|\theta)}{f_0(x) - f_1(x|\theta)} \; \forall \; x \in [0, x_0] : f_0(x) > f_1(x|\theta) \\ p \geq \max_{x \in [0, x_0] : f_0(x) > f_1(x|\theta)} \frac{-f_1(x|\theta)}{f_0(x) - f_1(x|\theta)} = p_{min}(\theta). \end{aligned} \tag{1.3}$$

This bound depends on the unknown θ parameter, which makes the estimation of p more complex, especially if θ is high-dimensional. However, there are two special cases of $f_0(x)$ for which p_{min} will be easily obtainable. If $f_0(x)$ is non-increasing on $[0, x_0]$, and $f_1(x|\theta)$ is non-decreasing on $[0, x_0]$, then (1.2) takes its lowest value at 0, and it will be a valid probability density function as long as $p/x_0 + (1-p)f_1(0|\theta) \geq 0 \iff p \geq \frac{-f_1(0|\theta)}{1/x_0 - f_1(0|\theta)}$. If we choose $f_0(x) \propto f_1(x|\theta)\mathbb{I}_{[0,x_0]}(x)$, then the terms $f_1(x|\theta)$ in (1.3) cancel out, and $p_{min} = \frac{-1}{1/F_1(x_0|\theta) - 1}$, where $F_1(x|\theta)$ is the cumulative distribution function corresponding to the probability density function $f_1(x|\theta)$. This last choice has the disadvantage of creating a discontinuity at x_0, which may be hard to interpret.

1.3 Estimation

If p is known to be positive, then the problem of estimating $\omega = (p, \theta)$ is reduced to mixture model parameter estimations. A large body of literature exists on this subject, notably Titterington et al. (1985). In particular, the EM algorithm is a well-studied method for obtaining (with good choices of starting values) maximum likelihood estimators. Here we are interested in estimation when p is either completely unknown or known to be negative.

We consider maximum likelihood estimation for such models, and we propose different methods to obtain the maximum likelihood estimators for p and θ. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the proposed model (1.2). The log-likelihood of the model is given by

$$l(\omega|x) = \sum_{i=1}^{n} \log\left[p f_0(x_i) + (1-p) f_1(x_i|\theta)\right].$$

Its partial derivatives are given by:

$$\begin{aligned}\frac{\partial l(\omega|x))}{\partial p} &= \sum_{i=1}^{n} \frac{f_0(x_i) - f_1(x_i|\theta)}{p f_0(x_i) + (1-p) f_1(x_i|\theta)},\\ \frac{\partial l(\omega|x)}{\partial \theta} &= (1-p) \sum_{i=1}^{n} \frac{\frac{\partial f_1(x_i|\theta)}{\partial \theta}}{p f_0(x_i) + (1-p) f_1(x_i|\theta)}.\end{aligned}$$

However, since $f_0(x)$ is only non-zero on $[0, x_0]$, these can be rewritten in this way:

$$\frac{\partial l(\omega|x)}{\partial p} = \sum_{i:x_i<x_0} \frac{f_0(x_i) - f_1(x_i|\theta)}{p f_0(x_i) + (1-p) f_1(x_i|\theta)} - \frac{1}{(1-p)} \left|\{i : x_i > x_0\}\right|, \tag{1.4}$$

$$\frac{\partial l(\omega|x)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \log(f_1(x_i|\theta))}{\partial \theta} \left(1 - \frac{p f_0(x_i)}{p f_0(x_i) + (1-p) f_1(x_i|\theta)}\right), \tag{1.5}$$

where $|\mathcal{A}|$ is the cardinality of the set $\mathcal{A}$. Note that, in deriving equation (1.4), we implicitly assume $f_1(x|\theta)$ to be nonzero on $[x_0, \infty)$. It is possible to set these equations equal to 0 and solve numerically for p and θ with a Newton-Raphson algorithm to obtain the maximum likelihood estimator (Dennis and Schnabel, 1996). It is interesting to note that, for fixed θ,

$$\frac{\partial^2 l(p|x)}{\partial p^2} = -\sum_{i=1}^{n}\left[\frac{f_0(x_i) - f_1(x_i)}{pf_0(x_i) + (1-p)f_1(x_i)}\right]^2 \leq 0 \; \forall \; p.$$

Any zero of $\frac{\partial l(\omega|x)}{\partial p} = 0$ is therefore a maximum. An alternative to solving (1.4) and (1.5) is to use the Nelder-Mead simplex method (Lagarias et al., 1999).

Generally, it will not be possible to obtain an analytical expression for the maximum likelihood estimators. However, if we choose $f_0(x) \propto f_1(x)\mathbb{I}_{[0,x_0]}(x)$, and if θ is known, it is possible to obtain the maximum likelihood estimator for p, $\hat{p}_{MLE}$, in analytical form. It is given by:

$$\hat{p}_{MLE} = \frac{(n_0/n) - F_1(x_0|\theta)}{1 - F_1(x_0|\theta)},$$

where $n_0 = |\{i : x_i \leq x_0\}|$. It should be noted that this estimator depends on θ.

Another approach to obtaining maximum likelihood estimates is to slightly adapt estimation methods for mixtures to the ZMM model. Let ϵ be a positive quantity such that $p + \epsilon > 0$, and let $q = p + \epsilon$. We can re-write the ZMM model as:

$$\begin{aligned}
f(x|p,\theta,\epsilon) &= pf_0(x) + \epsilon f_1(x|\theta) + (1-p-\epsilon)f_1(x|\theta) \\
f(x|p,\theta,\epsilon) &= (p+\epsilon)\left[\frac{p}{p+\epsilon}f_0(x) + \frac{\epsilon}{p+\epsilon}f_1(x|\theta)\right] + (1-p-\epsilon)f_1(x|\theta) \\
f(x|q,\theta) &= q\tilde{f}_0(x|\theta) + (1-q)f_1(x|\theta), \qquad (1.6)
\end{aligned}$$

where

$$\tilde{f}_0(x|\theta) = \frac{p}{p+\epsilon}f_0(x) + \frac{\epsilon}{p+\epsilon}f_1(x|\theta).$$

It should be noted that $f(x|p,\theta,\epsilon)$ is a non-identifiable model since, given observations x and fixed parameters p and θ, the likelihood $L(p,\theta,\epsilon|x)$ will be the same regardless of the value of ϵ. To make the model identifiable, we only consider the smallest possible value of ϵ that will allow $\tilde{f}_0(x|\theta)$ to be a valid probability density function, that is, positive over $[0, x_0]$:

$$\epsilon = \begin{cases} 0 \text{ for } p \geq 0 \\ \max_{x\in[0,x_0]} |p| \frac{f_0(x)}{f_1(x|\theta)} \text{ for } p < 0. \end{cases}$$

This choice of ϵ also ensures that $q \in [0, 1]$. For positives values of p, this is trivial, as p is always smaller or equal to 1. If p is negative, we have that:

$$\begin{aligned} q = \epsilon + p &= \max_{x\in[0,x_0]} |p| f_0(x)/f_1(x) + p \\ &= p\left[1 - \max_{x\in[0,x_0]} f_0(x)/f_1(x)\right]. \end{aligned}$$

Since $\int_0^{x_0} f_0(x)dx = 1 \geq \int_0^{x_0} f_1(x)dx$, it follows that $\exists\, x \in [0, x_0] : f_0(x) \geq f_1(x)$, thus implying that $\max_{x\in[0,x_0]} f_0(x)/f_1(x) \geq 1$, and that $q \geq 0 \;\forall\; p < 0$.

Furthermore, for negative values of p

$$\begin{aligned} q \leq 1 &\iff p(1 - \max_{x\in[0,x_0]} f_0(x)/f_1(x)) \leq 1 \\ &\iff p \geq \frac{1}{1 - \max_{x\in[0,x_0]} f_0(x)/f_1(x)} \\ &= \frac{1}{\min_{x\in[0,x_0]} 1 - f_0(x)/f_1(x)} \\ &= \max_{x\in[0,x_0]} \frac{1}{1 - f_0(x)/f_1(x)} \\ &= \max_{x\in[0,x_0]} \frac{f_1(x)}{f_1(x) - f_0(x)} \\ &= \max_{x\in[0,x_0]} \frac{-f_1(x)}{f_0(x) - f_1(x)} = p_{min}. \end{aligned}$$

Therefore, for values of p in $[p_{min}, 1]$, q is in $[0, 1]$, and (1.6) is a mixture of the densities $\tilde{f}_0(x|\theta)$ and $f_1(x|\theta)$. It will thus be possible to use slightly adapted estimation methods for mixtures. A popular estimation method for mixture models is the expectation-maximization (EM) algorithm.

Since (1.6) is a mixture model, each observation X_i can be viewed as having the distribution $\tilde{f}_0(x|\theta)$ with probability q or $f_1(x|\theta)$ with probability $1-q$. Let the latent variables $Z = Z_1, \ldots, Z_n$ represent whether the observations X_i have the distribution $\tilde{f}_0(x_i|\theta)$ or $f_1(x_i|\theta)$ respectively. Each Z_i will take the value 1 with probability q and 0 with probability $1-q$.

Then the completed likelihood is

$$L(\theta, q|X, Z) = \prod_{i=1}^{n} [q\tilde{f}_0(x_i|\theta)]^{z_i}[(1-q)f_1(x_i|\theta)]^{1-z_i}.$$

The EM algorithm aims at maximizing the marginal likelihood $L(\theta, q|X)$ by iteratively applying two steps to the completed likelihood. First, the expectation (E) step consists of obtaining

$$Q^{(k)}(q, \theta) = \int_Z \log[L(q, \theta|X, Z)]f(Z|X, q^{(k-1)}, \theta^{(k-1)})dZ, \tag{1.7}$$

the expectation of the log-likelihood with respect to the conditional distribution of the latent variables Z given the observations X under the current estimates of the parameters.

Let

$$\begin{aligned} A_i &= P(Z_i = 1|X_i, q^{(k-1)}, \theta^{(k-1)}) \\ &= \frac{q^{(k-1)}\tilde{f}_0(x_i|\theta^{(k-1)})}{q^{(k-1)}\tilde{f}_0(x_i|\theta^{(k-1)}) + (1-q^{(k-1)})f_1(x_i|\theta^{(k-1)})}. \end{aligned}$$

Then, equation (1.7) can be written as

$$Q^{(k)}(q, \theta) = \sum_{i=1}^{n} A_i \log[q\tilde{f}_0(x_i|\theta)] + (1-A_i)\log[(1-q)f_1(x_i|\theta)].$$

The maximization (M) step then consists in finding values $\theta^{(k)}$ and $q^{(k)}$ which maximize $Q^{(k)}(q, \theta)$.

The partial derivatives of Q are:

$$\begin{aligned}
\frac{\partial Q}{\partial q} &= \frac{\sum_{i=1}^n A_i}{q} - \frac{\sum_{i=1}^n (1-A_i)}{1-q}, \\
\frac{\partial Q}{\partial \theta} &= \sum_{i=1}^n A_i \frac{\partial \log \tilde{f}_0(x|\theta)}{\partial \theta} + \sum_{i=1}^n (1-A_i)\frac{\partial \log[f_1(x_i|\theta)]}{\partial \theta} \\
&= \sum_{i=1}^n A_i \frac{\partial \log[p f_0(x) + \epsilon f_1(x|\theta)]}{\partial \theta} + \sum_{i=1}^n (1-A_i)\frac{\partial \log[f_1(x_i|\theta)]}{\partial \theta}.
\end{aligned}$$

The k^{th} iteration estimate for q can be easily obtained by setting the partial derivative of Q with respect to q equal to zero and solving:

$$\hat{q}^{(k)} = \frac{\sum_{i=1}^n A_i}{n}.$$

Note that we have considered above $\tilde{f}_0(x|\theta)$ to be independent of q while p and ϵ (which are linked to q) appear it its expression. If we do not consider p and ϵ to be independent of q during the M-step, the partial derivative of Q with respect to q will be positive everywhere, which means that Q is maximized when q takes its largest possible value –1. We will still be able to obtain an estimate for p in this case, but convergence will be slower and p will be consistently overestimated (as can be seen in Section 5).

Estimates for θ are not so easy to obtain, and they will not usually result in a closed form expression. Estimates must therefore be obtained by numerical methods; either directly maximizing Q (for example, with the Nelder-Mead simplex method (Lagarias et al., 1999)) or numerically solving $\frac{\partial Q}{\partial \theta} = 0$ (for example, with a Newton-Raphson algorithm(Dennis and Schnabel, 1996)).

Another possibility is to simplify M-step calculations by considering θ in $\tilde{f}_0(x|\theta)$ to be known and equal to $\theta^{(k-1)}$. For some choices of $f_1(x|\theta)$ (notably, exponential families), this will allow us to obtain a closed form expression for $\theta^{(k)}$. However, as it is obtained by maximizing $\sum_{i=1}^n (1-A_i)\log[(1-q)f_1(x_i|\theta)]$, rather than $Q^{(k)}$, it would be more appropriate to speak of a generalized EM algorithm, which should nevertheless have good convergence properties (Wu, 1983).

It is also necessary to obtain, at each iteration, an estimate of ϵ (to deduce an estimate for p, and to update $\tilde{f}_0(x|\theta)$). For given values of q and θ, we have two candidates for the value of ϵ : 0 (corresponding to $p \geq 0$) and $\max_{x\in[0,x_0]} -p\frac{f_0(x)}{f_1(x|\theta)}$ (corresponding to negative values of p). It follows that we have two candidates for the value of $\hat{p}^{(k)}$ given $\hat{q}^{(k)}$ and $\hat{\theta}^{(k)}$, and we choose the one with the highest likelihood:

$$\hat{p}^{(k)} = \begin{cases} \hat{q}^{(k)} \\ \dfrac{\hat{q}^{(k)}}{1 - \max_{x\in[0,x_0]} \frac{f_0(x)}{f_1(x|\theta^{(k)})}} \end{cases}.$$

Initial values for θ can be obtained using the method of moments or by basing the estimation on a truncated distribution and considering only observations greater than x_0. Exploratory tests, not fully reported here for brevity, have shown that, for a given choice of f_0 and f_1, this estimation method is robust to the choice of a starting value for p. For example, if f_1 is the zero-truncated Laplace probability density function, then positive initial values of p will have good convergence properties. If f_1 is the gamma pdf, then negative starting values for p should be preferred. In both cases, the initial value for p needs only to be of the right sign to converge. An initial value for q can be directly obtained from the initial values of p, θ and the definition given above for ϵ.

1.4 Asymptotic properties

For the MLE (or, strictly speaking, a sequence of roots of the likelihood equation) to converge in distribution to a normal random variable, several conditions must be satisfied:

- the parameter space Ω must be an open subset of $\mathbb{R}^k$,
- the second partial derivatives of $f(x|\omega)$ with respect to ω must exist and be continuous for every $\omega \in \Omega$, and we must be able to pass the derivative under the integral sign in $\int f(x|\omega)dx$,

- there must be a function $g(x)$ such that $E(g(X))$ exists and that each component of the Fisher information matrix must be uniformly bounded in absolute value in some neighbourhood of the real value of ω,
- the Fisher information matrix must be positive definite.

In many applications, $f_1(x|\theta)$ will satisfy these regularity conditions and it will be possible to choose $f_0(x)$ so that $f(x|\omega)$ also satisfies the above conditions. However, it will be necessary to restrict p to values strictly greater than p_{min} (and strictly smaller than 1) for Ω to be an open subset of $\mathbb{R}^k$. If the above conditions are met, then $\sqrt{n}\hat{\omega}_{MLE} \xrightarrow{\mathcal{D}} \mathcal{N}\left(\omega, I(\omega)^{-1}\right)$, where $I(\omega)$ is the Fisher information matrix.

Generally, we will have to use numerical or Monte-Carlo methods to compute the elements of this matrix. Resampling-based Monte-Carlo methods for computing the Fisher information matrix have been proposed and discussed in Spall (2005) and Das et al. (2007), while Behboodian (1972) uses numerical quadrature to obtain the information matrix. Simply generating $x^{(k)} \sim f(x|\omega)$ allows us to approximate the elements $I(\omega)_{ij}$ by $-\dfrac{1}{n}\sum_{k=1}^{n} \frac{\partial^2 l(\omega|x^{(k)})}{\partial\omega_i \partial\omega_j}$, but this is not very efficient.

Using the fact that $f_0(x)$ is positive only on $[0, x_0]$, we can obtain the following expression for the i, j^{th} entry of the Fisher information matrix for $\omega = (p, \theta)$:

$$\begin{aligned} I(\omega)_{ij} &= -E_{f(x|\omega)}\left(\frac{\partial^2 \log\left[f(x|\omega)\right]}{\partial\omega_i \partial\omega_j}\right) \\ &= -\left(\int_0^{x_0} \frac{\partial^2 \log\left[f(x|\omega)\right]}{\partial\omega_i \partial\omega_j}[p f_0(x) + (1-p) f_1(x|\theta)]dx \right. \\ &\quad \left. + \int_{x_0}^{\infty} \frac{\partial^2 \log\left[(1-p) f_1(x|\theta)\right]}{\partial\omega_i \partial\omega_j}(1-p) f_1(x|\theta) dx\right), \\ &= -\left(\int_0^{x_0} \frac{\partial^2 \log\left[f(x|\omega)\right]}{\partial\omega_i \partial\omega_j}[p f_0(x) + (1-p) f_1(x|\theta)]dx \right. \\ &\quad + \int_0^{\infty} \frac{\partial^2 \log\left[(1-p) f_1(x|\theta)\right]}{\partial\omega_i \partial\omega_j}(1-p) f_1(x|\theta) dx \\ &\quad \left. - \int_0^{x_0} \frac{\partial^2 \log\left[(1-p) f_1(x|\theta)\right]}{\partial\omega_i \partial\omega_j}(1-p) f_1(x|\theta) dx\right). \end{aligned} \tag{1.8}$$

Note that

$$\begin{aligned}\frac{\partial^2 \log[(1-p)f_1(x|\theta)]}{\partial\omega_1\partial\omega_{i+1}} &= \frac{\partial^2 \log[(1-p)f_1(x|\theta)]}{\partial p\partial\theta_i}\\ &= \frac{-\partial f_1(x|\theta)/\partial\theta_i}{(1-p)f_1(x|\theta)} + \frac{(1-p)f_1(x|\theta)\partial f_1(x|\theta)/\partial\theta_i}{[(1-p)f_1(x|\theta)]^2} = 0\end{aligned}$$

and that

$$\begin{aligned}&\frac{\partial^2 \log[(1-p)f_1(x|\theta)]}{\partial\omega_{i+1}\partial\omega_{j+1}} = \frac{\partial^2 \log[(1-p)f_1(x|\theta)]}{\partial\theta_i\partial\theta_j}\\ &= \frac{(1-p)\partial^2 f_1(x|\theta)/\partial\theta_i\partial\theta_j}{[(1-p)f_1(x|\theta)]} - \left(\frac{(1-p)\partial f_1(x|\theta)/\partial\theta_i}{[(1-p)f_1(x|\theta)]}\right)\left(\frac{(1-p)\partial f_1(x|\theta)/\partial\theta_j}{[(1-p)f_1(x|\theta)]}\right)\\ &= \frac{\partial^2 f_1(x|\theta)/\partial\theta_i\partial\theta_j}{f_1(x|\theta)} - \left(\frac{\partial f_1(x|\theta)/\partial\theta_i}{f_1(x|\theta)}\right)\frac{\partial f_1(x|\theta)/\partial\theta_j}{f_1(x|\theta)} = \frac{\partial^2 \log f_1(x|\theta)}{\partial\theta_i\partial\theta_j}.\end{aligned}$$

This leads to the following expressions:

$$\begin{aligned}I(p) = I(\omega)_{11} &= \int_0^{x_0} \frac{\left(f_0(x) - f_1(x|\theta)\right)^2}{[pf_0(x) + (1-p)f_1(x|\theta)]}\, dx\\ &\quad + \frac{1}{1-p}[1 - F_1(x_0|\theta)], \qquad (1.9)\\ I(\omega)_{1,i+1} &= \int_0^{x_0} \frac{\partial f_1(x|\theta)/\partial\theta_i}{pf_0(x) + (1-p)f_1(x|\theta)}\\ &\quad \times [(1+p)f_0(x) - pf_1(x|\theta))]\, dx, \qquad (1.10)\\ I(\theta)_{ij} = I(\omega)_{i+1,j+1} &= \int_0^{x_0} \left[\frac{(1-p)^2}{(pf_0(x) + (1-p)f_1(x|\theta))}\frac{\partial f_1(x|\theta)}{\partial\theta_i}\frac{\partial f_1(x|\theta)}{\partial\theta_j}\right.\\ &\quad \left. -(1-p)\frac{\partial^2 f_1(x|\theta)}{\partial\theta_i\partial\theta_j}\right]\, dx\\ &\quad + (1-p)\int_0^{x_0} \frac{\partial^2 \log f_1(x|\theta)}{\partial\theta_i\partial\theta_j} f_1(x|\theta)\, dx\\ &\quad + (1-p)I_{f_1(x|\theta)}(\theta)_{ij}, \qquad (1.11)\end{aligned}$$

where $I_{f_1(x|\theta)}$ is the Fisher information matrix of a random variable with probability density function $f_1(x|\theta)$.

The Fisher information for $\hat{p}_{MLE}$ can be decomposed into a function of the cumulative distribution function $F_1(x|\theta)$ and an integral to be evaluated numerically on $[0, x_0]$. Similarly, the information on θ can be decomposed into two parts: (a) an integral to be evaluated numerically on $[0, x_0]$, and (b) a term which is proportional to the information on θ under $f_1(x|\theta)$. In most practical applications, $f_1(x|\theta)$ will be chosen amongst distributions with known asymptotic properties, and $I_{f_1(x|\theta)}(\theta)$ will thus be easily obtained. Finally, the information on p and θ is reduced to an integral to be evaluated numerically on $[0, x_0]$. Seeing that the integrals to be evaluated are unidimensional (as long as X is univariate), that the region of integration is compact, and that, given the choices of $f_1(x|\theta)$ and $f_0(x)$, the integrands are piecewise continuous, numerical quadrature methods (for example, Gauss or Chebyshev-Gauss quadrature) might be preferable to Monte-Carlo computations.

These expressions depend on unknown parameters which, in practical applications, will be unavailable. The empirical Fisher information matrix can be obtained by replacing these unknown parameters by their maximum likelihood estimators where necessary.

1.5 Simulation results

The performance of the proposed model and estimation methods can be assessed by a simulation study. It is thus necessary to be able to generate observations from the ZMM. We assume that it is possible to generate observations from the densities $f_0(x)$ and $f_1(x|\theta)$. If p is positive, then $f(x|\theta, p)$ is a standard mixture of two densities, and methods used to generate from this type of distribution are well-known. If p is negative, then we have that $f(x|\theta, p) = pf_0(x) + (1-p)f_1(x|\theta) \leq (1-p)f_1(x|\theta)$, and we can use rejection sampling to generate observations from $f(x|\theta, p)$ with unconditonnal acceptance probability $1/(1+|p|)$.

As an example, we choose $f_0(x) \propto (x_0 - x)^\tau$ and

$$f_1(x|\mu, \lambda) = \frac{1}{\lambda(2 - e^{-\mu/\lambda})} e^{-\frac{|x-\mu|}{\lambda}}, x \geq 0$$

the Laplace distribution with location parameter μ truncated on $[0,\infty]$. The exact form of $f_0(x)$ is $f_0(x) = c(x_0 - x)^\tau$, where $c = (\tau+1)/x_0^{\tau+1}$is a normalizing constant and τ is known. This choice of model ensures a probability density function that is continuous everywhere and easy to evaluate. Also, the truncated Laplace pdf is nonzero at zero (unlike, say, the lognormal), making it suitable to illustrate zero-deflation.

The lower bound p_{min} is strongly influenced by the values of the parameters μ and λ. Small values of μ concentrate a lot of probability mass around x_0, allowing p to take smaller values. Inversely, large values of μ will pull probability mass away from x_0, imposing a stricter lower bound on p. The parameter λ also has a strong influence on p_{min}, but this influence depends on the value of μ. Larger values of λ correspond to a higher variance and heavier tails. Heavier tails, in turn, correspond to lower values of p_{min} when μ is far from x_0. In that case, $[0, x_0]$ is in the left tail of the distribution. When μ is small, $[0, x_0]$ is near the peak of the distribution, and heavier tails correspond to a stricter bound.

Obviously, the values of τ and x_0 will also affect the bound. However, we suppose here that these parameters are known. The bound p_{min} as a function of μ and λ when $\tau = x_0 = 1/2$ is illustrated in Figures 1.1 and 1.2.

Let $p = -0.1$, $x_0 = 1$, $\mu = 1$, $\lambda = 2$, and $\tau = 1/2$. The densities $f_1(x|\mu,\lambda)$ and $f(x|p,\mu,\lambda)$ corresponding to this choice of parameters are illustrated in Figure 1.3. For different samples sizes ($n = 10, 30, 60, 200, 500$), 1000 random samples are generated. In Table 1.1, we report the empirical mean and variance of the maximum likelihood estimators for each sample size. Estimates are obtained using the Nelder-Mead simplex method to maximize the log-likelihood, and the EM algorithm. Unless otherwise noted, all numerical algorithms used are iterated until a relative tolerance of 10^{-3} is reached. By numerical integration of equations (1.9)

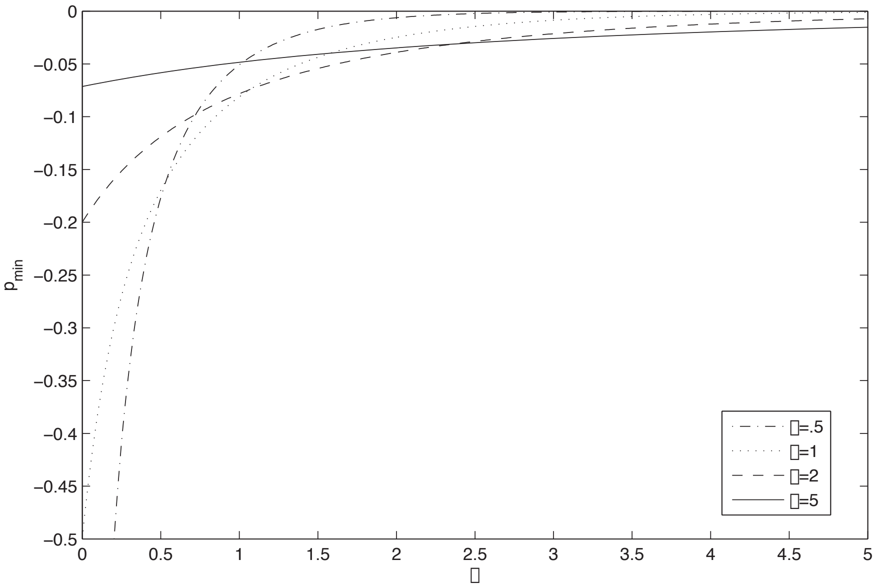

Figure 1.1: Bound p_{min} as a function of μ, λ fixed

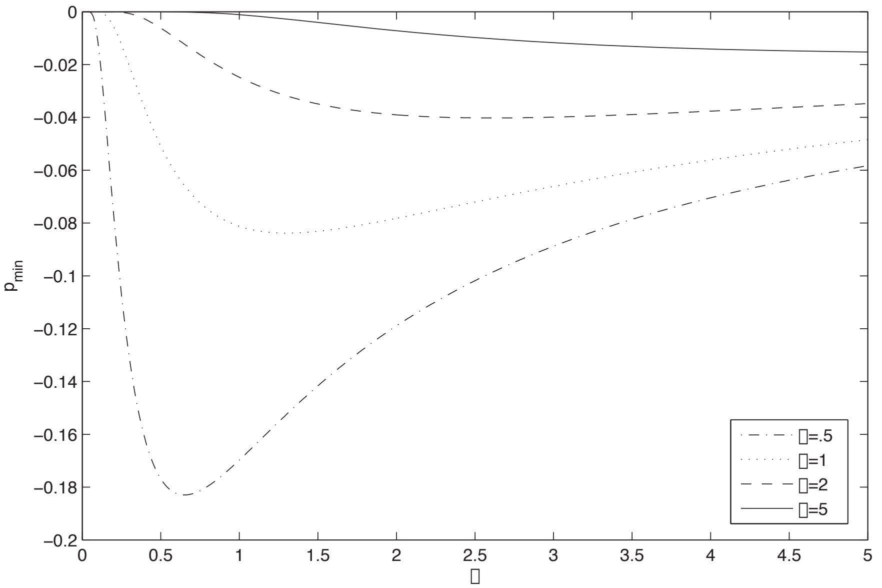

Figure 1.2: Bound p_{min} as a function of λ, μ fixed

to (1.11) using Gaussian quadrature, the asymptotic covariance matrix is:

$$I^{-1}(\omega) = \begin{array}{c} \begin{array}{ccc} p & \mu & \lambda \end{array} \\ \begin{pmatrix} 0.794 & 3 & -0.23 \\ 3 & 15.19 & -2.69 \\ -0.23 & -2.69 & 5.07 \end{pmatrix} \end{array}$$

The moderate bias present in small sample sizes becomes negligible as n gets larger. The empirical covariance matrix, however, seems unusually slow in converging to the asymptotic covariance matrix $I^{-1}(\omega)$. This might be caused by the parameter values chosen; a value of λ larger than μ corresponds to a very dispersed model, where the mode does not necessarily dominate the likelihood until larger samples sizes are reached. In addition, a large value of x_0 means that small variations of p would have large effects on the model as a whole. Estimates obtained with the EM algorithm here often have smaller variance than those obtained by direct optimization, but a much larger bias. This is most likely caused by convergence to a local mode or saddlepoint, perhaps due to the use of numerical optimization during the M-step. The EM algorithm took on average between 82 ($n = 30$) and 146 ($n = 10$) iterations to converge. Direct maximization of the likelihood should be preferred in this context.

We then set $(x_0, \tau, \mu, \lambda) = (1, 0.5, 0, 2)$, let p vary between $p_{min} = -0.5$ and 0.3, and consider μ to be known. Then, $f_1(x|\lambda)$ is the exponential distribution. When using the EM algorithm, if we replace $\tilde{f}_0(x|\lambda)$ by $\tilde{f}_0(x|\lambda^{(k-1)})$ during the M-step of each iteration, we can obtain the following closed-form expression: $\lambda^{(k)} = \frac{\sum_{i=1}^{n}(1-A_i)x_i}{\sum_{i=1}^{n}(1-A_i)}$. For each value of p considered, 500 samples of size $n = 60$ are generated. The mean and variance of estimators obtained by maximization of the log-likelihood with the simplex method and using the GEM algorithm are given in Table 1.2.

We see a larger bias when $p = p_{min}$, as well as higher variances for $\hat{\lambda}$ when p is positive. This is not unexpected: large values of p indicate that a large proportion

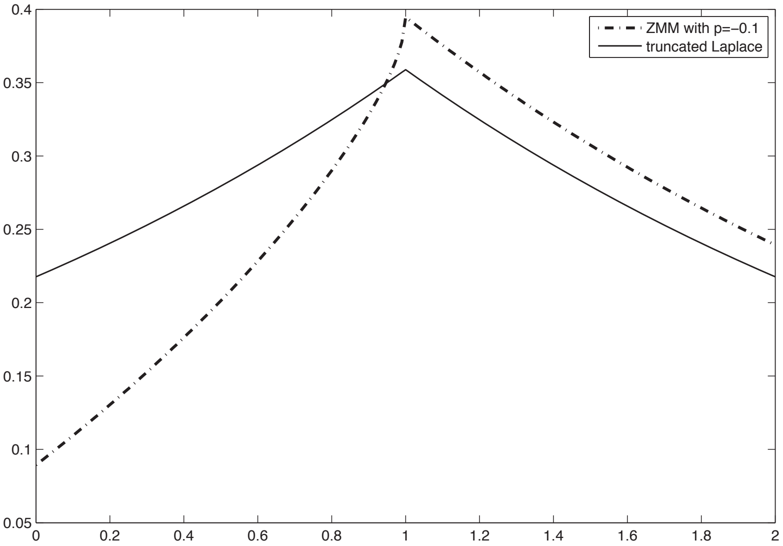

Figure 1.3: f_1 and f for $(p, \tau, \mu, \lambda) = (-0.1, 1/2, 1, 2)$

of observations are drawn from the distribution $f_0(x)$, which is entirely specified. This leaves us with only a few observations to estimate λ, whose variance steadily increases with p. Direct maximization of the likelihood performs better than the GEM for negative values of p close to p_{min} or large values of p. For intermediate values, the GEM estimates have slightly smaller variances and comparative bias. Unsurprisingly, the number of iterations required for the GEM to converge are lower when p is either very small or very large. The GEM, in this context, is not noticeably slower than the simplex optimization, since the M-step estimates are obtained as a closed-form expression. This, along with a less complex two-parameter estimation, explains why better results are obtained compared to the previous simulation.

Finally, if we choose $f_0(x) \propto f_1(x|\theta)\mathbb{I}_{[0,x_0]}(x)$, it is possible to obtain an analytical estimator for p. Simulation results (500 samples, $n = 200$) for this case are presented in Table 1.3. Figure 1.4 shows the densities $f_1(x|\mu, \lambda)$ and $f(x|p, \mu, \lambda)$ for different values of p. The discontinuity at x_0 in Figure 1.4 is due to the choice of $f_0(x) \propto f_1(x|\theta)\mathbb{I}_{[0,x_0]}(x)$, whose limit at x_0 is not 0. This discontinuity can be hard

Table 1.1: Simulation results for $(p, x_0, \tau, \mu, \lambda) = (-.1, 1, .5, 1, 2)$

MLE		EM	
$\bar{\hat{\omega}} = (\bar{\hat{p}}, \bar{\hat{\mu}}, \bar{\hat{\lambda}})$	$n \times \widehat{var}(\hat{\omega})$	$\bar{\hat{\omega}} = (\bar{\hat{p}}, \bar{\hat{\mu}}, \bar{\hat{\lambda}})$	$n \times \widehat{var}(\hat{\omega})$
$n = 10$			
$\begin{pmatrix} -0.089 \\ 1.15 \\ 1.82 \end{pmatrix}$	$\begin{pmatrix} 0.159 & 0.272 & 0.11 \\ 0.272 & 4.39 & -1.58 \\ 0.11 & -1.58 & 5.32 \end{pmatrix}$	$\begin{pmatrix} -0.049 \\ 1.51 \\ 1.71 \end{pmatrix}$	$\begin{pmatrix} 0.138 & 0.282 & -0.042 \\ 0.282 & 4.56 & -1.45 \\ -0.042 & -1.45 & 5.59 \end{pmatrix}$
$n = 30$			
$\begin{pmatrix} -0.104 \\ 1.07 \\ 1.91 \end{pmatrix}$	$\begin{pmatrix} 0.1967 & 0.658 & 0.054 \\ 0.658 & 6.55 & -1.98 \\ 0.054 & -1.98 & 5.45 \end{pmatrix}$	$\begin{pmatrix} -0.053 \\ 1.38 \\ 1.86 \end{pmatrix}$	$\begin{pmatrix} 0.182 & 0.542 & -0.069 \\ 0.542 & 5.82 & -1.93 \\ -0.069 & -1.93 & 5.53 \end{pmatrix}$
$n = 60$			
$\begin{pmatrix} -0.105 \\ 1.02 \\ 1.93 \end{pmatrix}$	$\begin{pmatrix} 0.27 & 0.93 & -0.005 \\ 0.93 & 8.21 & -2.3 \\ -0.005 & -2.3 & 5.07 \end{pmatrix}$	$\begin{pmatrix} -0.059 \\ 1.29 \\ 1.92 \end{pmatrix}$	$\begin{pmatrix} 0.23 & 0.75 & -0.044 \\ 0.75 & 7.08 & -1.85 \\ -0.044 & -1.85 & 5.52 \end{pmatrix}$
$n = 200$			
$\begin{pmatrix} -0.107 \\ 0.99 \\ 1.98 \end{pmatrix}$	$\begin{pmatrix} 0.61 & 2.28 & -0.14 \\ 2.28 & 12.9 & -2.54 \\ -0.14 & -2.54 & 4.81 \end{pmatrix}$	$\begin{pmatrix} -0.0725 \\ 1.22 \\ 1.95 \end{pmatrix}$	$\begin{pmatrix} 0.276 & 1.03 & -0.08 \\ 1.03 & 8.41 & -2.21 \\ -0.08 & -2.21 & 5.55 \end{pmatrix}$
$n = 500$			
$\begin{pmatrix} -0.105 \\ 0.99 \\ 1.99 \end{pmatrix}$	$\begin{pmatrix} 0.8 & 3.09 & -0.22 \\ 3.09 & 16.67 & -2.71 \\ -0.22 & -2.71 & 5.07 \end{pmatrix}$	$\begin{pmatrix} -0.076 \\ 1.17 \\ 1.98 \end{pmatrix}$	$\begin{pmatrix} 0.32 & 1.16 & -0.152 \\ 1.16 & 8.58 & -2.13 \\ -0.152 & -2.13 & 5.12 \end{pmatrix}$

to interpret, and it makes this choice of f_0 hard to justify in practical applications. As can be seen in Table 1.3, both the bias and variance of $\hat{p}$ remain small as p varies. The smallest bias and variances are attained when p is close to one, or smaller or equal to zero. The maximum likelihood estimates for p are practically the same whether obtained numerically or analytically, the only notable difference being when p is identically equal to 1. In that case, the analytical maximum likelihood estimate was 1 for all samples, while the numerical maximum likelihood has a very small variance. The EM algorithm was initialized with a starting value of 0.9, explaining its much faster convergence for (larger) positive values of p. It nevertheless has good convergence properties even for negative values of p, although it is often outperformed by the other two methods. This makes it a less than ideal

Table 1.2: Simulation results as p varies, with $\mu = 0$, $x_0 = 1$, $\tau = 0.5$, and $\lambda = 2$

	MLE				GEM				
p	$\bar{\hat{p}}$	$\widehat{Var}(\hat{p})$	$\bar{\hat{\lambda}}$	$\widehat{Var}(\hat{\lambda})$	$\bar{\hat{p}}$	$\widehat{Var}(\hat{p})$	$\bar{\hat{\lambda}}$	$\widehat{Var}(\hat{\lambda})$	$\overline{iter}$
- 0.5	-0.481	0.003	2.096	0.037	-0.416	0.009	2.289	0.151	72.82
- 0.4	-0.402	0.007	2.015	0.059	-0.409	0.012	2.034	0.088	66.13
- 0.3	-0.308	0.012	2.016	0.069	-0.312	0.015	2.001	0.084	64.31
- 0.2	-0.214	0.015	2.007	0.088	-0.217	0.014	2.004	0.088	66.93
- 0.1	-0.09	0.018	1.972	0.124	-0.103	0.016	1.976	0.089	111.52
- 0.05	-0.059	0.019	2.022	0.105	-0.058	0.014	1.976	0.090	135.41
0	-0.024	0.018	1.982	0.103	-0.023	0.017	1.975	0.098	130.48
0.05	0.043	0.018	2.002	0.116	0.042	0.018	1.996	0.116	132.64
0.1	0.098	0.019	1.973	0.132	0.097	0.018	1.978	0.119	99.43
0.2	0.189	0.018	1.998	0.133	0.192	0.016	2.037	0.133	66.24
0.3	0.291	0.017	2.020	0.146	0.275	0.016	1.992	0.154	50.26

choice in this context.

1.6 Application to real data

An application to a real dataset is presented in this section. The data are millimetric bimonthly precipitation data during the months of May and June (4 data points per year) in Montreal, from 1943 to 1992. A histogram of the data, and plots of a fitted truncated Laplace distribution, a fitted zero-modified model with $f_0 \propto (10 - x)^{0.05}$ and a fitted ZMM with $f_0(x) \propto f_1(x|\theta)$ are given in Figure 1.5. Ideally, we would want f_0 to be as close to f_1 as possible while still preserving continuity, so that x_0-truncated data could be accurately modelled. While $f_0(x) \propto f_1(x|\theta)$ satisfies this requirement, it creates a discontinuity. Choosing $f_0 \propto (x_0 - x)^{\tau}$, with a small value for τ, seems like a good compromise, as continuity is preserved, and $f_0(x)$ is nearly flat on most of the interval $[0, x_0]$.

The maximum likelihood estimators and their estimated standard errors obtained when fitting the truncated Laplace, the ZMM-tau and the ZMM-proportional models are given in Table 1.4. We can see that the peak of the fitted ZMM is closer to the actual mode of the data, and the modelling is improved on $[0, 10]$. The probability density functions for the two ZMM models are practically identical outside

Table 1.3: Simulation results for $\hat{p}_{mle}$ when $f_0 \propto f_1$, $(x_0, \mu, \lambda) = (0.5, 1, 1)$

	Closed form MLE		Numerical MLE		EM		
p	$\hat{p}$	$10^3\widehat{Var}(\hat{p})$	$\hat{p}$	$10^3\widehat{Var}(\hat{p})$	$\hat{p}$	$10^3\widehat{Var}(\hat{p})$	$\overline{iter}$
-0.15	-0.1502	0.1128	-0.1502	0.1128	-0.1514	0.1626	116.63
-0.10	-0.1000	0.3208	-0.1000	0.3208	-0.1008	0.4663	88.78
-0.05	-0.0510	0.6975	-0.0510	0.6975	-0.0522	0.7016	78.49
0	0.0014	0.8926	0.0014	0.8926	0.0003	0.9712	145.23
0.05	0.0504	0.1097	0.0504	0.1097	0.0456	0.1157	99.44
0.10	0.1008	0.1184	0.1008	0.1184	0.0833	0.0022	25.56
0.25	0.2513	0.1341	0.2513	0.1341	0.2516	0.1341	8.51
0.50	0.5023	0.1603	0.5023	0.1603	0.5025	0.1604	6.14
0.75	0.7527	0.1140	0.7527	0.1140	0.7529	0.1148	4.98
0.85	0.8494	0.0778	0.8494	0.0778	0.8495	0.0777	4.11
0.95	0.9499	0.0273	0.9499	0.0273	0.9497	0.0271	3.99
0.99	0.9901	0.0059	0.9901	0.0059	0.9899	0.0061	4.20
1	1	0	1.000	<0.0001	1.000	<0.0001	5.00

of $[0, x_0]$.

Table 1.4: MLEs and standard errors for the fitted models

model	$\hat{p}$	$SE_{\hat{p}}$	$\hat{\mu}$	$SE_{\hat{\mu}}$	$\hat{\lambda}$	$SE_{\hat{\lambda}}$
truncated Laplace	n/a	n/a	24.1	0.17	21.88	1.76
ZMM-tau	-0.066	0.018	22.3	0.16	22.75	1.75
ZMM-proportional	-0.072	0.018	22.3	0.12	22.64	1.74

Table 1.5 gives the Akaike information criterion and Bayesian information criterion for the fitted models. For both ZMMs, the improvement in fit compensates the penalty term for the extra model parameter. Furthermore, the ZMM with $f_0(x) \propto f_1(x|\theta)\mathbb{I}_{[0,x_0]}(x)$ has slightly lower AIC and BIC values than the ZMM with $f_0 \propto (10 - x)^{.05}$.

We also performed a sensitivity analysis, to see what impact did the choice of x_0 and τ had when fitting the ZMM. Results for $f_0 \propto (x_0 - x)^{\tau}$ and $f_0(x) \propto f_1(x|\theta)\mathbb{I}_{[0,x_0]}(x)$ can be found in Figures 1.6 and 1.7, respectively.

It should be noted that while a positive estimate for p is straightforward to interpret, negative estimates have no direct intuitive meaning. Negative values of p which seem, at first glance, almost zero, can nevertheless have a strong effect on

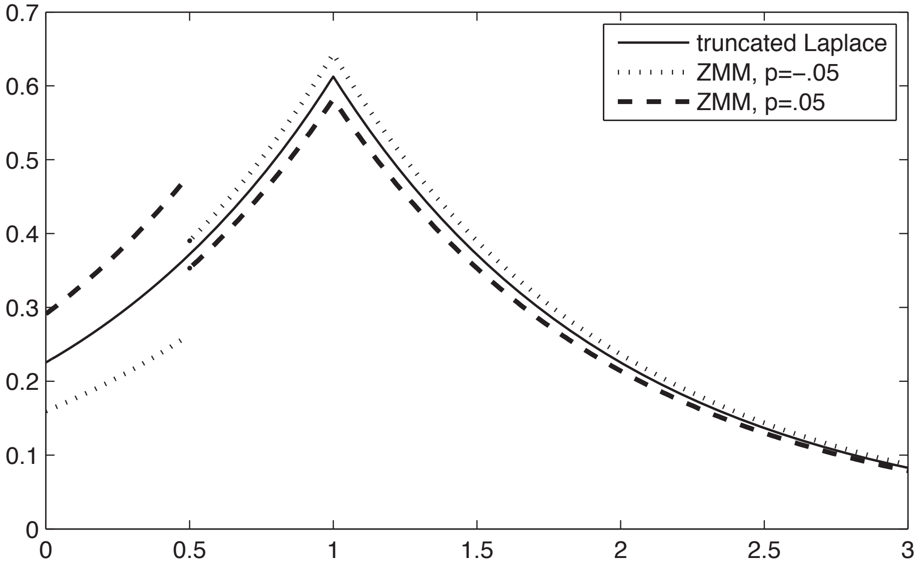

Figure 1.4: Densities f and f_1 for $f_0 \propto f_1$, $p = -.05$ and $p = .05$

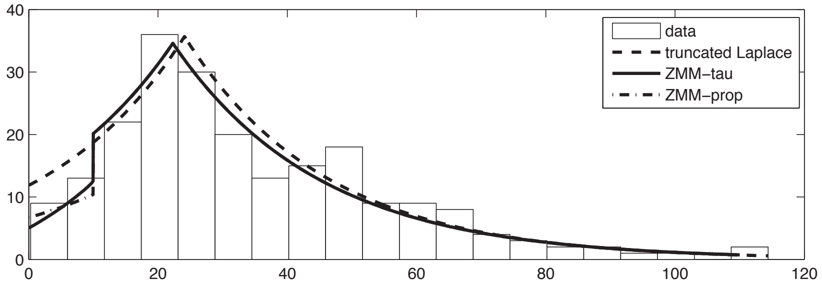

Figure 1.5: Data and fitted models for Montreal precipitations

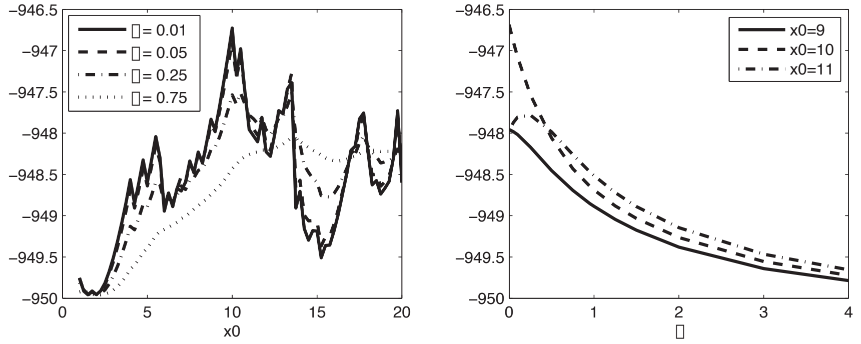

Figure 1.6: Log-likelihood as a function of x_0 and τ for $f_0 \propto (x_0 - x)^\tau$

Table 1.5: AIC and BIC for the fitted models

	Truncated Laplace	ZMM-tau	ZMM-proportionnal
AIC	1903.9	1899.5	1898.4
BIC	1910.7	1909.6	1908.6

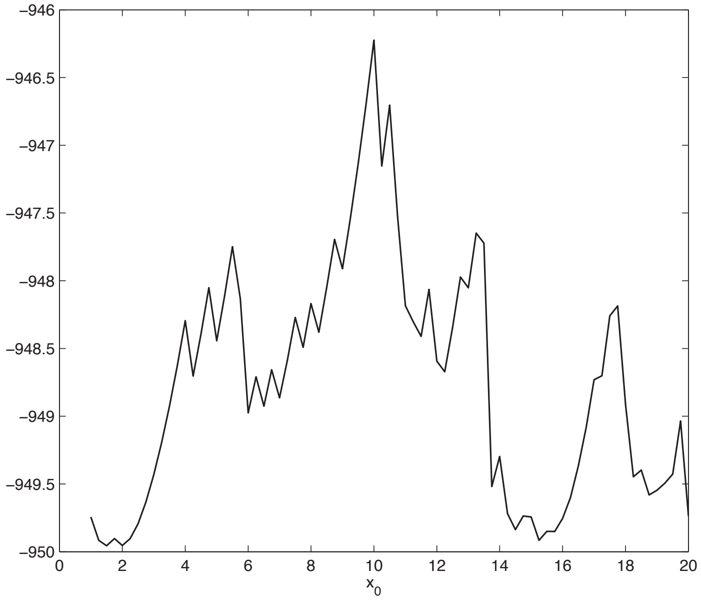

Figure 1.7: Log-likelihood as a function of x_0 for $f_0(x) \propto f_1(x|\theta)\mathbb{I}_{[0,x_0]}(x)$

the pdf on $[0, x_0]$. This may happen if p_{min} is very close to zero, which can occur, for example, if $f_0(x)$ if very concentrated, $f_1(x|\theta)$ is diffuse and x_0 is large. For ease of interpretation, a scaled estimate $\tilde{p} = -\hat{p}/p_{min}(\hat{\theta})$ should be reported along with the estimate $\hat{p}$. In particular, a $\tilde{p}$ value of -1 corresponds to $\hat{p} = p_{min}(\hat{\theta})$, i.e., as much probability mass has been removed in $[0, x_0]$ as possible. Although this removal may be small relative to $f_1(x|\theta)$. For the Montreal bimonthly precipitation example, with $f_0 \propto (10 - x)^{.05}$ we have $p_{min}(\hat{\theta}) = -0.1117$ and $\hat{p} = -0.0657$. The value of $\tilde{p}$ is 0.5884, indicating that roughly half of the probability mass that could possibly be removed in the interval $[0, x_0]$ has been removed. This highlights that the negative value of $\hat{p}$ cannot simply be ignored, even if it appears at first glance to be close to zero.

1.7 Concluding remarks

In the present paper we have introduced an extension of the zero-modified models to the continuous case. The proposed model can be viewed as a mixture of two continuous densities f_0 and f_1, but the "mixture" parameter can take positive and negative values, corresponding respectively to the zero-inflated and the zero-deflated cases. The particular case of $f_0 \propto f_1$ has some interesting properties. For this choice of f_0, it is possible to obtain a closed form maximum likelihood estimator for p. It is also possible to obtain an x_0-truncated distribution by taking $p = p_{min}$ under this choice of f_0, which is not generally possible for other choice of x_0. However, this creates a discontinuity in the model which may be hard to interpret or justify.

Likelihood-based estimation has been discussed. We can reduce the estimation of a ZMM to a simple mixture model and use an EM or GEM algorithm for parameter estimation. In the general case, using this approach is slower and leads to larger mean squared errors than numerical maximization of the likelihood function. The asymptotic properties of the model and maximum likelihood estimators have been briefly discussed. While the bias decreases appropriately as the sample size grows, the covariance matrix of the maximum likelihood estimates seems to be slow to converge to the Fisher information matrix. This might occur when the value of x_0 is too high, relatively to the bulk of the probability density function $f_1(x|\theta)$; caution is advised when fitting ZMMs with a value of x_0 approaching the mode of the distribution. The dependency of the lower bound p_{min} on unknown parameters is the main difficulty associated with the use of the proposed model.

Simulations studies have been used to evaluate the usability of the proposed model, and an example of an application to real-life precipitation data has been presented, demonstrating its usefulness. Despite its hard-to-justify discontinuity, choosing $f_0(x) \propto f_1(x)$ leads to a slightly smaller values for the AIC and BIC.

We have here considered f_0 and x_0 to be known. A Bayesian framework could be useful here, as a natural approach to the modelling of expert nowledge regarding

x_0 would be to use a prior concentrating the probability density around a certain value given by the experts.

Secondly, we could consider models of the form $f(x|p, x_0, \theta_0, \theta_1) = pf_0(x|\theta_0) + (1-p)f_1(x|\theta_1)$, where f_0 is no longer considered to be known. A larger number of observations would then be required for reliable parameter estimation, especially for very large and very small values of p.

Finally, tests for zero-inflation are fairly common in the literature. Developing a test for zero-deflation, or for zero-inflation using our model formulation, should be investigated in future research. A natural approach would be to try to adapt existing score tests to our model formulation.

REFERENCES

Aitchison, J. (1955). On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of the American Statistical Association*, **50**(271):901–908.

Ancelet, S., Etienne, M., Benoît, H., and Parent, E. (2009). Modelling spatial zero-inflated continuous data with an exponentially compound Poisson process. *Environmental and Ecological Statistics*, :1–30.

Angers, J. and Biswas, A. (2003). A Bayesian analysis of zero-inflated generalized Poisson model. *Computational statistics & data analysis*, **42**(1-2):37–46.

Azhar, A., Murty, V., and Phien, H. (1992). Modeling irrigation schedules for lowland rice with stochastic rainfall. *Journal of Irrigation and Drainage Engineering*, **118**:36.

Behboodian, J. (1972). Information matrix for a mixture of two normal distributions. *Journal of statistical computation and simulation*, **1**(4):295–314.

Dalrymple, M. L., Hudson, I. L., and Ford, R. P. K. (2003). Finite mixture, zero-inflated Poisson and hurdle models with application to SIDS. *Computational Statistics & Data Analysis*, **41**(3-4):491–504.

Das, S., Spall, J., and Ghanem, R. (2007). Efficient calculation of Fisher information matrix: Monte Carlo approach using prior information. *Master's thesis, Department of Applied Mathematics and Statistics, The Johns Hopkns University, Baltimore, Maryland, USA, May*, .

Dennis, J. and Schnabel, R. (1996). *Numerical methods for unconstrained optimization and nonlinear equations.* Society for Industrial Mathematics.

Dibike, Y. and Solomatine, D. (2001). River flow forecasting using artificial neural networks. *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere*, **26**(1):1–7.

Diebolt, J. and Robert, C. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, **56**(2):363–375.

Dietz, E. and Böhning, D. (2000). On estimation of the Poisson parameter in zero-modified Poisson models. *Computational statistics & data analysis*, **34**(4):441–459.

Eklundh, L. (1998). Estimating relations between AVHRR NDVI and rainfall in East Africa at 10-day and monthly time scales. *International Journal of Remote Sensing*, **19**(3):563–570.

El-Shaarawi, A. (1985). Some goodness-of-fit methods for the Poisson plus added zeros distribution. *Applied and environmental microbiology*, **49**(5):1304.

Fernandes, M., Schmidt, A., and Migon, H. (2009). Modelling zero-inflated spatio-temporal processes. *Statistical Modelling*, **9**(1):3.

Feuerverger, A. (1979). On some methods of analysis for weather experiments. *Biometrika*, **66**(3):655–658.

Fletcher, D. (2008). Confidence intervals for the mean of the delta-lognormal distribution. *Environmental and Ecological Statistics*, **15**(2):175–189.

Ghosh, S. K., Mukhopadhyay, P., and Lu, J.-C. (2006). Bayesian analysis of zero-inflated regression models. *Journal of Statistical Planning and Inference*, **136**(4):1360–1375.

Gupta, P., Gupta, R., and Tripathi, R. (2005). Score test for zero inflated generalized Poisson regression model. *Communications in Statistics-Theory and Methods*, **33**(1):47–64.

Hall, D. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, **56**(4):1030–1039.

Hasan, M. and Sneddon, G. (2009). Zero-inflated Poisson regression for longitudinal data. *Communications in Statistics-Simulation and Computation*, **38**(3):638–653.

Jansakul, N. and Hinde, J. (2002). Score tests for zero-inflated Poisson models. *Computational statistics & data analysis*, **40**(1):75–96.

Lagarias, J., Reeds, J., Wright, M., and Wright, P. (1999). Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal on Optimization*, **9**(1):112–147.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**(1):1–14.

Lo, N., Jacobson, L., and Squire, J. (1992). Indices of relative abundance from fish spotter data based on delta-lognormal models. *Canadian Journal of Fisheries and Aquatic Sciences*, **49**(12):2515–2526.

Min, A. and Czado, C. (2010). Testing for zero-modification in count regression models. *Statistica Sinica*, **20**:323–341.

Muthén, B. and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, **55**(2):463–469.

Peco, B., Espigares, T., and Levassor, C. (1998). Trends and fluctuations in species abundance and richness in Mediterranean annual pastures. *Applied Vegetation Science*, :21–28.

Raghuwanshi, N., Singh, R., and Reddy, L. (2006). Runoff and sediment yield modeling using artificial neural networks: Upper Siwane River, India. *Journal of Hydrologic Engineering*, **11**:71.

Ridout, M., Hinde, J., and Demétrio, C. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, **57**(1):219–223.

Rodrigues, J. (2003). Bayesian analysis of zero-inflated distributions. *Communications in Statistics-Theory and Methods*, **32**(2):281–289.

Sharda, V. and Das, P. (2005). Modelling weekly rainfall data for crop planning in a sub-humid climate of India. *Agricultural water management*, **76**(2):120–138.

Singh, S. (1963). A note on inflated Poisson distribution. *Journal of the Indian Statistical Association*, **1**(3):140–144.

Spall, J. (2005). Monte Carlo computation of the Fisher information matrix in non-standard settings. *Journal of Computational and Graphical Statistics*, **14**(4):889–909.

Stefansson, G. (1996). Analysis of groundfish survey abundance data: combining the GLM and delta approaches. *ICES Journal of Marine Science*, **53**(3):577.

Tian, L. (2005). Inferences on the mean of zero-inflated lognormal data: the generalized variable approach. *Statistics in medicine*, **24**(20):3223–3232.

Titterington, D., Smith, A., and Makov, U. (1985). *Statistical analysis of finite mixture distributions*. John Wiley & Sons.

Van den Broek, J. (1995). A score test for zero inflation in a Poisson distribution. *Biometrics*, **51**(2):738–743.

Wu, C. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, **11**(1):95–103.

CHAPTER 2

BAYESIAN ESTIMATION AND TESTING FOR CONTINUOUS ZERO-MODIFIED MODELS

Cet article sera prochainement soumis a *Statistics & Probability Letters.* Le premier auteur est Félix Labrecque-Synnott et le coauteur est le directeur de recherches, Jean-François Angers.

La problématique derrière le deuxième article – un traitement bayésien de la classe des modèles continus modifiés à zéro, ainsi qu'une approche aux tests d'hypothèses – découlait naturellement du premier article. Les tests d'hypothèses font souvent suite à l'estimation ponctuelle en statistique, il était naturel d'aller dans cette direction après avoir démontré qu'il était possible de travailler avec la classe de modèles proposée, et que celle-ci était pertinente et avait au moins une application pratique. Différentes lois *a priori* non informatives furent considérées par le premier auteur. Après avoir mis en oeuvre l'estimation basée sur le maximum de la loi *a posteriori* (basée sur une des méthodes d'estimation par maximum de vraisemblance, elle-même basée sur du code *Matlab* existant), quelques difficultés furent rencontrées au moment de passer à l'estimation par espérance *a posteriori* des paramètres. À la suggestion du deuxième auteur, des méthodes d'estimation adaptatives furent alors considérées. Le premier auteur s'est chargé de la programmation de deux méthodes de ce type tirées de la littérature scientifique (Angers and Biswas, 2003, Naylor and Smith, 1982), et de l'élaboration des tests d'hypothèses présentés dans l'article. Face aux longs temps de calculs demandés par les méthodes adaptatives, il a identifié la tâche demandant le gros des calculs dans l'évaluation de la vraisemblance en un point donné: le calcul de $p_{min}(\theta)$. Le premier auteur a donc élaboré une méthode d'obtention de l'espérance *a posteriori* basée sur une quadrature gaussienne et une simple transformation des paramètres. Cette méthode n'est pas adaptative, mais les bornes $p_{min}(\theta_i)$ sont gardées en mémoire et réutilisées si plusieurs échantillons doivent être comparés ou si une analyse doit être refaite.

On retrouve dans cet article une proposition de lois *a priori* non informatives à la section 2. L'estimation ponctuelle bayésienne est abordée à la section 3, où trois méthodes sont proposées afin d'obtenir les espérances *a posteriori*. La section 4 traite de tests d'hypothèses sur p, tant sous le paradigme bayésien que classique. La section 5 contient des résultats de simulation comparant les différentes méthodes d'estimation et de tests, ainsi qu'une ré-analyse des données présentées à la section 6 du premier article.

ABSTRACT

While discrete zero-inflated and zero-deflated models are fairly well known, continuous zero-modified models are mostly limited to the zero-inflated case in the literature, where they are often represented as a mixture between a probability density function and a Dirac mass at zero. An alternative formulation of zero-inflation for continuous models which can be extended to the zero-deflated case was introduced in Labrecque-Synnott and Angers (2009), where likelihood-based estimation was discussed. We consider here estimation and testing in a Bayesian framework. In particular, we discuss the choice of priors, the posterior maximum estimators and the posterior expectation estimators. Some simulation results and an application to a real data set illustrate the estimation and testing procedures in the context of continuous zero-deflated models.

2.1 Introduction

In the present paper, we consider continuous zero-modified models within a Bayesian framework. The parameter p controlling whether the model is zero-inflated, zero-deflated, or unmodified, is considered unknown; we are therefore interested in estimation or testing procedures providing good results in any of those cases.

Discrete zero-modified models, introduced in Singh (1963), are frequently encountered in the literature. Over the years, various and more complex models have been proposed, notably in Lambert (1992), Hall (2000), Ridout et al. (2001). For this class of models, the score test is the preferred method to test hypotheses on p; difficulties in testing within the more common zero-inflated models are often related to the fact that the value of p under the null hypothesis lies at the boundary of the parameter space (Van den Broek, 1995, Jansakul and Hinde, 2002, Gupta et al., 2005). Discrete zero-deflated and zero-modified models where the sign of p is unknown are discussed in Angers and Biswas (2003) and Dietz and Böhning

(2000).

In the continuous case, until recently, only zero-inflated models were considered. Such models, and their estimation, are introduced in Aitchison (1955), and further discussed in Feuerverger (1979), Stefansson (1996), Fernandes et al. (2009), Fletcher (2008), Tian (2005). Zero-deflated continuous models were introduced in Labrecque-Synnott and Angers (2009), where maximum likelihood estimation was considered. Here, we consider estimation within a Bayesian context, as well as hypothesis testing under the classical and Bayesian paradigms.

In Section 2, we review the zero-modified continuous model and present appropriate priors. In Section 3, we discuss parameter estimation and their asymptotic properties. Section 4 focuses on testing hypotheses on p. We present simulation results and an application to a real data set in Section 5. Section 6 contains some concluding remarks.

2.2 The model

Let X be a non-negative random variable such that it can be modelled by a probability density function of the form

$$f(x|\theta, p, x_0) = p \times f_0(x) + (1-p) \times f_1(x|\theta), \tag{2.1}$$

where $f_0(x)$ is a probability density function on $[0, x_0]$, $f_1(x|\theta)$ is a probability density function on $\mathbb{R}_+$, and p is the zero-modification parameter, which can take both positive and negative values. For positive values of p, this probability density function can therefore be viewed as a mixture of f_0 and f_1. We cannot generally suppose that p or θ are known; we therefore note $\omega = (p, \theta)$, the set of parameters of interest.

As discussed in Labrecque-Synnott and Angers (2009), some constraints on $f_0(x)$ and p must be respected in order for $f(x|\omega)$ to be a continuous probability density function. In particular, we must have $\lim_{x \to x_0} f_0(x) = 0$ for continuity, and

$$1 \geq p \geq \max_{x \in [0,x_0]: f_0(x) > f_1(x|\theta)} \frac{-f_1(x|\theta)}{f_0(x) - f_1(x|\theta)} = p_{min}(\theta) \tag{2.2}$$

for $f(x|\omega)$ to be non-negative everywhere. The lower bound $p_{min}(\theta)$ depends on the parameter θ, creating a dependance between the parameters. It might sometimes be useful to work with

$$\tilde{p} = \begin{cases} p/|p_{min}(\theta)| & \text{if } p < 0, \\ p & \text{otherwise.} \end{cases} \tag{2.3}$$

In the absence of information on p, we consider non-informative priors on p and θ. Jeffreys prior is a poor choice here, since obtaining parameter estimates will usually require the use of numerical integration or Monte-Carlo methods. The prior must therefore be evaluated at a large number of points. Since the Jeffreys prior is based on the Fisher information matrix, which must also be obtained numerically, its use leads to nested numerical procedures, which is much more computationally intensive than other non informative priors. Instead, we can use a suitable non-informative prior on θ, and either

$$\begin{aligned} \pi_1(p|\theta) &\sim \text{uniformly on } [p_{min}(\theta), 1], \\ \pi_2(p|\theta) &\sim \text{uniformly on } [p_{min}(\theta), 0] \text{ with probability } 0.5, \\ &\quad \text{and uniformly on } [0,1]\text{with probability } 0.5, \end{aligned}$$

where choosing $\pi_2(p|\theta)$ is equivalent to choosing a uniform distribution on $[-1, 1]$ for $\tilde{p}$.

More particularly, we are interested in the zero-modified location-scale Laplace family of distributions truncated on $\mathbb{R}^+$. This is a mathematically simple distribution that we can easily work with to assess the applicability of the class of continuous zero-modified models; unlike more common distributions on $\mathbb{R}^+$, such as the gamma and lognormal distributions, the truncated Laplace pdf is nonzero at 0, which allows us to consider zero-deflation. While the simulation results presented in Section 5 might not be applicable to other distributions, the underlying theoretical work is not hinged on any property specific to the truncated Laplace distribution, and should therefore be straightforward to apply to other distributions on $\mathbb{R}^+$.

Let $f_0(x) \propto (x_0 - x)^\tau \mathbb{I}_{[0,x_0]}(x)$, where $\mathbb{I}$ is the indicator function, and

$$f_1(x|\mu,\lambda) = \frac{1}{\lambda\left[1 + \text{sign}(\mu)(1 - e^{-|\mu|/\lambda})\right]} \exp\left\{\frac{-|x-\mu|}{\lambda}\right\}, x \geq 0. \tag{2.4}$$

For negative values of μ, we can write

$$\begin{aligned} f_1(x|\mu,\lambda) &= \frac{1}{\lambda\left[1 - (1 - e^{\mu/\lambda})\right]} \exp\left\{\frac{\mu - x}{\lambda}\right\} \\ &= \frac{1}{\lambda} \exp\left\{\frac{-x}{\lambda}\right\}, \end{aligned}$$

which is the exponential distribution and corresponds to the case $\mu = 0$. We will therefore restrict μ to non-negative values in order to have an identifiable model. We can then write (2.4) as:

$$f_1(x|\mu,\lambda) = \frac{1}{\lambda\left[2 - e^{-\mu/\lambda})\right]} \exp\left\{\frac{-|x-\mu|}{\lambda}\right\}.$$

In this case, as μ and λ were originally a location and a scale parameter, respectively, improper non informative priors could be

$$\begin{aligned} \pi_3(\mu) &\propto 1, \forall \mu \geq 0, \qquad (2.5) \\ \pi_4(\lambda) &\propto 1/\lambda, \forall \lambda > 0. \end{aligned}$$

It is possible to show that, even with this choice of improper priors, the posterior will be proper.

2.3 Estimation

Let $x = (x_1, \dots, x_n)$ be a sample of n i.i.d. random variables from a continuous zero-modified model $f(x_i|\omega)$. Within a Bayesian framework, inference on the parameters $(p, \theta) = \omega$ will be based on the posterior distribution $\pi(\omega|x)$, which is proportional to the product of the prior $\pi(\omega)$ and the likelihood $L(\omega|x) = \Pi_{i=1}^{n} f(x_i|\omega) = \Pi_{i=1}^{n} \left[p f_0(x) + (1-p) f_1(x|\theta)\right]$. Due to the form of the likelihood function, neither the posterior maximum nor the posterior expectations can generally be obtained analytically.

It is nonetheless rather easy to maximize the posterior, as most statistical software packages include routines to quickly maximize the likelihood of an arbitrary distribution. Such routines can easily be adapted to maximize the product of the likelihood and the prior. As in the frequentist framework, the main issue with maximum posterior estimates is that of multimodality: if the function to be maximized is not unimodal, there is a possibility that the algorithm used to compute the estimate converges to a local, rather than global, mode. In a Bayesian framework, possible prior-data conflicts can introduce multimodality in the posterior distribution. While non-informative priors are likely to be dominated by the likelihood function, and therefore carry a lesser risk of multimodality than informative priors, using the posterior expectation of the parameters as estimators completely sidesteps the multimodality issue.

However, the posterior expectation also cannot be explicitly obtained, and must be approximated by numerical quadrature or a Monte-Carlo estimate. The dependence between p and θ will complicate matters; in particular, since the support of p depends on the value of θ, direct numerical integration is not trivial. When using numerical quadrature, it is therefore useful to work with $\tilde{p}$ instead, as its support is always $[-1, 1]$. Another problem is the curse of dimensionality: our simulations (not shown here) showed us that, as the dimension of θ increases, the space where the posterior is nearly 0 grows much faster than the space where most of the posterior probability is concentrated. This is worsened by the parameter p, as typically a narrow range of values of θ will be compatible with zero-deflation, while another will be compatible with zero-inflation.

A first approach to obtaining posterior expectation estimators could be to use importance sampling. Under this approach, to estimate a function $g(\omega)$ of the parameters ω, we would first generate a Monte-Carlo sample $Z = Z_1, \ldots, Z_k$, where each Z_i has probability density function $h(z_i)$, called the importance function. The resulting estimator would be $\widehat{t_h}(\omega) = \sum_{i=1}^{k} g(z_i)\pi(\omega)L(\omega|z_i)/h(z_i)$. The difficulty here lies in the fact that to have a quick convergence, we must choose h as close as possible to the posterior distribution (Rubinstein and Kroese, 2008). The complex relationship between p and θ makes this difficult. It is therefore possible that Monte-Carlo estimation with importance sampling require a very large random sample to converge, making it a slow and unattractive choice, and more complex numerical methods might be needed to obtain quality estimators.

2.4 Numerical methods

We consider three numerical methods to obtain posterior expectation estimators. The first two methods are adaptive methods designed so that most of the computational work is done in the region where the bulk of the posterior density is concentrated. The last method considered, on the other hand, aims at reducing the computational load required to obtain the estimators.

2.4.1 Adaptive Monte-Carlo

We consider here the adaptive Monte-Carlo estimation with importance sampling proposed in Angers and Biswas (2003). The idea is to first apply a transformation $t : \Omega \to \mathbb{R}^k$ to the parameter space, so that a convenient importance function (from which large samples can be generated easily) can be used. The Monte-Carlo sample $z_1, \ldots, z_k$ is iteratively transformed so that most z_i's end up in the region where the posterior density is concentrated. This approach avoids the need for tweaking the importance function in order to obtain satisfactory results.

Here, we work with the transformed parameters

$$\omega' = t(\omega) = \left(\log \mu, \log \lambda, \log \left[-\log \left(\frac{\tilde{p}+1}{2}\right)\right]\right).$$

Having transformed the parameter space to the whole $\mathbb{R}^3$, we can use the standard multivariate Student's t distribution as an importance function to generate a random sample $z^{(0)}$. We iteratively apply the following steps to obtain the estimators for ω' :

1. compute the $l^{th}-$ iteration importance weights $h(z_i^{(l)})$, the estimator of ω', $\hat{\omega'}$ and the estimator of the covariance matrix of ω', $\hat{S_{\omega'}}$;

2. for $i = 1, \ldots, k$, compute $z_i^{(l+1)} = \widehat{S_{\omega'}}^{1/2} z_i^{(0)} + \widehat{\omega'}$, where $S^{1/2}$ is the square root matrix of S;

3. repeat steps 1 and 2 until the differences between the estimators $\widehat{\omega'}$ and $\widehat{S_{\omega'}}$ of successive iterations are small enough;

4. once the method has converged, the estimators for ω are obtained by taking the inverse transformation $t^{-1}(\omega')$.

2.4.2 Adaptive Gaussian quadrature

Monte-Carlo methods are not the only tools available to evaluate non-tractable integrals. Another methods with widespread use is Gaussian numerical quadrature. In this case the integral $\int_a^b f(x)dx$ is approximated by the weighted $\sum_{1=1}^{k} w(x_i)f(x_i)$, where the x_i's are the sample points (or nodes) of the quadrature and the w_i's are the quadrature weights. Weights and nodes are determined by the domain of integration ($[-1,1]$, $[0,\infty)$, or $(-\infty,\infty)$), and can be found either in tables or in most mathematical software packages. Multiple integrals can also be approximated by Gaussian quadratures. The approximation will then take the form of an iterated sum.

In the context of zero-modified models, we are interested in obtaining the posterior expectation of a given function of the parameters, say $g(\omega)$,

$$E(g(\omega)|x) = \int_{\Theta} \int_{p_{min}}^{1} g(\omega) L(\omega|x) \pi(\omega) dp d\theta,$$

which could be approximated by a quadrature of the form

$$\sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{ij} g(\omega_{ij}) L(\omega_{ij}|x) \pi(\omega_{ij}).$$

Gaussian quadrature does not require us to find a suitable importance function. However, due to the dependency between p and θ, most quadrature nodes ω_{ij} used in the computation will give us little information about the integral. A possible solution would be to augment the number of quadrature nodes, but obviously this increases the computational load required to obtain estimators.

An adaptive quadrature method, similar to the one introduced in Naylor and Smith (1982), is a possible solution. Just like the adaptive Monte-Carlo method described above, its principle is to transform the parameter space and quadrature nodes so that most of the computations are made where most of the posterior density is concentrated. The first step is to transform the parameters so that the transformed parameter space is $\mathbb{R}^3$. We will once again use the transformed parameters $\omega' = t(\omega) = \left(\log \mu, \log \lambda, \log \left[- \log \left(\frac{\tilde{p} + 1}{2} \right) \right] \right)$. To estimate a given function $g(\omega')$ of the parameters, we iteratively apply the following steps:

1. obtain the initial quadrature nodes ω'_{ijk} and quadrature weights w_{ijk}, and specify initial values for ν and S, the posterior expectation and covariance matrix, respectively;

2. compute $z_{ijk}^{(l)} = \nu + \sqrt{2} S^{1/2} \omega'_{ijk}$;

3. obtain the $l^{th}-$step estimates of ν and S using the approximation $\int_{\mathbb{R}^3} g(\omega') pi(\omega') L(\omega'|x) d\omega \approx \sum_{ijk} w_{ijk} \, g(z_{ijk})$;

4. repeat Steps 2 and 3 until the differences between the estimators of ν and S from successive iterations are small enough.

2.4.3 Gauss-Legendre-Laguerre quadrature

Another option available to us is to try to reduce the computational work needed to evaluate the posterior at a given point. If we choose to work with $\tilde{p}$, for a function of ω, $g(\omega)$, the integral

$$\int_{\Omega} g(\omega)\pi(\omega)L(\omega|x)d\omega$$

can be expressed as the iterated integral

$$\int_0^{\infty}\left[\int_0^{\infty}\left(\int_{-1}^{1} g(\omega)\pi(\omega)L(\omega|x)dp\right)d\lambda\right]d\mu.$$

The Gauss-Legendre quadrature can then be used to evaluate the innermost integral, while the Gauss-Laguerre quadrature can be used to integrate with respect to μ and λ. It is interesting to note that, in this case, the quadrature nodes ω_{ijk} where the integrand is to be evaluated depend only on the number of quadrature nodes used. They are independent of the data data x, the function of interest g, and the prior π. Just as Gaussian quadrature weights and nodes are usually obtained from tables (or software packages), the bound p_{min} associated to each quadrature node does not depend on the data x and can also be tabulated (or written to disk). Any given node in the Gauss-Legendre-Laguerre iterated quadrature will always be associated to the same value of p_{min}. Since the numerical computation of p_{min} is the task with the highest computational cost in the evaluation of the integrand, this can save quite a bit of time, especially when a large number of integrals have to be computed (e.g., when analyzing multiple samples, or in simulation studies). This approach, computing once and then storing the bound $p_{min}(\mu, \lambda)$ for different values of μ and λ, cannot be used when using adaptive quadrature methods (as the nodes will then be adjusted to better fit the data) or Monte-Carlo methods.

2.5 Hypothesis testing

Testing the hypothesis that p lies in a certain set P_i, formally written as "$H_i : p \in P_i''$, within a Bayesian framework requires computing the posterior probability

$$\int_0^\infty \left[\int_0^\infty \left(\int_{P_i} \pi(\omega) L(\omega|x) dp \right) d\lambda \right] d\mu. \tag{2.6}$$

If multiple hypotheses are to be confronted, the hypothesis with the highest posterior probability is selected. When comparing two hypotheses, Bayesian testing can also be carried out through the use of Bayes factors, which can be obtained from the prior and posterior probabilities. When considering a point hypothesis, such as "$H_0 : p = 0''$, it is necessary to add a point mass to our prior to have a nonzero posterior probability for the point hypothesis.

Equation (2.6) can be rewritten as $\int_\Omega g(\omega)\pi(\omega)L(\omega|x)d\omega$, with $g(\omega) = \mathbb{I}_{P_i}(p)$, where $\mathbb{I}$ is the indicator function. The problem of Bayesian hypothesis testing can therefore be reduced to the computation of the posterior expectation of a certain function $g(\omega)$, and will therefore face the same issues as described in the previous sections. The solutions considered are also the same, namely adaptive Monte-Carlo (as in Angers and Biswas (2003)), adaptive quadrature (as in Naylor and Smith (1982)), and Gauss-Legendre-Laguerre quadrature with tabulated bounds $p_{min}(\mu, \lambda)$.

2.6 Simulation and application results

The performance of Bayesian tests and point estimators can be assessed through a simulation study. Let $\tau = 0.05$, and $x_0 = 10$. For samples sizes $n = 10, 20, 30, 60,$ and 100, 500 random samples are generated with $p = -0.1, \mu = 20, \lambda = 20$. Table 2.1 lists the mean and variance of parameter estimates as a function of sample size. We consider both maximum posterior and posterior mean estimates; the latter are computed using an adaptive Monte-Carlo method, adaptive Gauss-Hermite quadrature, and Gauss-Legendre-Laguerre (GLL) iterated quadrature. The convergence criterion used for the Monte-Carlo and numerical quadrature methods is a maximal relative error of at most 1×10^{-2} while the absolute error is less than 1×10^{-3}. GLL quadrature fulfills this criterion with $20 \times 20 \times 20$ or fewer nodes for more than half the random samples. Adaptive Gauss-Hermite quadrature requires less nodes ($12 \times 12 \times 12$) for the same precision, but is slower since more complex computations are required (as the grid of nodes is iteratively translated and scaled). Furthermore, the values of p_{min} computed for a given sample cannot be used for other samples, increasing the computational load of this method.

To obtain the adaptive Monte-Carlo estimates (as in Angers and Biswas (2003)), samples of 750 triplets (x, y, z) are generated from the multinormal$(0, I_3)$ distribution and are used to estimate the transformed parameters

$$\left(\log(\mu), \log(\lambda), \log\left[-\log\left(\frac{\tilde{p}+1}{2}\right)\right]\right).$$

These triplets are iteratively translated and scaled. Convergence is usually attained after 20 to 80 iterations. Simple importance sampling was also considered with a variety of importance functions (gamma, truncated normal, lognormal, Laplace and mixtures) and was mostly outperformed by the aforementioned estimation methods. These results are not reported here for the sake of brevity.

Table 2.1: Mean and variance of parameter estimates as a function of n, with $p = -0.1$, $\mu = \lambda = 20$

n		ML			MAP			Monte-Carlo			Gauss-Hermite			GLL		
		p	μ	λ	p	μ	λ	p	μ	λ	p	μ	λ	p	μ	λ
10	mean	-0.052	23.06	18.69	-0.046	24.54	16.01	-0.087	24.99	15.73	-0.111	23.28	21.78	-0.056	23.57	20.16
	var	0.008	54.23	58.93	0.0074	55.91	45.20	0.0059	52.72	42.05	0.002	70.96	35.16	0.0029	55.52	37.1
20	mean	-0.082	21.1	19.28	-0.079	21.38	17.98	-0.11	22.81	17.86	-0.126	21.25	19.64	-0.071	22.03	17.67
	var	0.0048	27.5	24.6	0.0058	30.4	23.39	0.0033	23.22	29.45	0.0017	25.76	28.31	0.0024	24.78	30.32
30	mean	-0.092	20.61	18.52	-0.086	21.08	17.58	-0.112	20.53	16.41	-0.135	23.57	19.91	-0.084	20.48	18.61
	var	0.0035	23.45	21.81	0.0033	23.15	20.2	0.0043	21.17	23.33	0.0023	27.43	24.61	0.0028	22.92	26.13
60	mean	-0.102	19.88	19.76	-0.1	20.01	19.33	-0.093	20.2	18.88	-0.123	22.23	17.37	-0.106	21.11	19.18
	var	0.0026	15.82	11.72	0.0025	15.33	11.94	0.0018	12.2	15.75	0.0019	13.3	18.62	0.0026	12.97	22.82
100	mean	-0.095	20.09	19.7	-0.094	20.23	19.4	-0.095	20.08	19.91	-0.112	20.93	19.85	-0.096	20.51	19.62
	var	0.0016	9.537	7.38	0.0016	9.14	7.103	0.0011	7.576	7.911	0.0013	8.21	9.42	0.0018	7.54	8.94

Table 2.2: Mean and variance of parameter estimates as a function of p, $\mu = \lambda = 20$, $n = 50$

p		ML			MAP			Monte-Carlo			Gauss-Hermite			GLL		
		p	μ	λ	p	μ	λ	p	μ	λ	p	μ	λ	p	μ	λ
-0.075	mean	-0.068	21.08	18.59	-0.065	21.35	17.67	-0.072	20.25	18.23	-0.069	20.23	18.25	-0.063	20.48	18.1
	var	0.0057	31.28	24.61	0.0067	34.01	22.94	0.0066	20.66	41.3	0.0061	20.72	42.26	0.0038	26.13	35.49
-0.05	mean	-0.047	20.51	19.23	-0.042	20.96	18.3	-0.056	20.75	17.18	-0.062	20.82	17.08	-0.056	21.18	18.62
	var	0.0067	22.48	19.77	0.0069	24.09	17.87	0.0088	19.78	37.8	0.0078	19.71	37.81	0.0022	18.85	26.46
-0.025	mean	-0.037	19.1	19.69	-0.034	19.38	18.75	-0.031	19.29	18.51	-0.032	19.23	18.47	-0.032	21.54	18.02
	var	0.0084	24.61	19.28	0.0086	24.34	18.42	0.0083	13.9	38.44	0.011	16.94	38.48	0.0054	19.28	22.88
-0.01	mean	-0.028	19.09	20.15	-0.024	19.43	19.17	-0.011	20.17	19.05	-0.0092	20.32	19.03	-0.013	20.92	18.32
	var	0.0061	33.65	30.55	0.0061	32.72	28.34	0.0073	16.24	32.87	0.0093	16.19	32.76	0.0046	28.88	30.29
-0.005	mean	-0.015	20.07	18.38	-0.015	20.54	17.44	-0.009	19.6	18.98	-0.011	19.66	19.03	-0.007	20.15	18.31
	var	0.0058	23.65	18.98	0.0048	24.14	17.53	0.012	15.52	46.65	0.009	15.57	44.24	0.003	18.47	23.32
0	mean	-0.0015	19.81	18.86	0.0046	20.34	17.83	-0.0018	19.9	18.22	-0.0039	19.79	18.52	-0.0036	20.64	17.82
	var	0.006	25.72	19.9	0.006	23.44	17.8	0.0105	15.39	43.79	0.0135	16.44	41.75	0.0035	18.82	27.28
0.05	mean	0.0091	16.87	20.64	0.0322	18.77	19.16	0.0621	19.68	18.15	0.0351	18.75	18.17	0.0057	21.86	17.51
	var	0.0077	57.98	25.62	0.0089	53.25	25.57	0.0122	13.74	36.84	0.0179	13.7	37.45	0.0075	26.15	31.02
0.1	mean	0.0191	15.66	20.27	0.019	16.41	19.03	0.0958	20.15	18.78	0.1292	20.13	18.75	0.092	21.24	18.44
	var	0.0114	65.34	23.66	0.0118	75.01	21.79	0.0084	17.19	34.09	0.0103	17.2	34.18	0.0091	19.98	33.42
0.2	mean	0.0601	9.97	21.41	0.0567	8.53	20.28	0.166	20.83	17.84	0.162	20.76	17.91	0.18	21.72	17.9
	var	0.0126	73.80	22.09	0.0122	78.38	20.78	0.0116	16.51	40.85	0.0103	16.54	39.77	0.0151	19.13	22.77
0.3	mean	0.109	5.258	22.12	0.114	3.186	20.57	0.271	20.31	18.30	0.286	20.28	18.34	0.281	21.12	18.02
	var	0.025	36.70	20.00	0.0235	41.88	19.52	0.011	17.68	36.2	0.0145	17.76	36.26	0.0197	21.01	28.27

While the estimates of μ and λ appear to be highly variable regardless of the estimation method, these variances do not seem out of place when considering the inverse of the Fisher information matrix:

$$I^{-1}(\omega) = \begin{bmatrix} 0.1302 & 6.8851 & -2.8064 \\ 6.8851 & 660.6605 & -256.2556 \\ -2.8064 & -256.2556 & 598.5605 \end{bmatrix}.$$

We next consider a fixed sample size $n = 50$ and let p take values between -0.1 and 1. In this scenario, the lower bound on p is $p_{min} = -0.1202$. For each value of p, 500 random samples have been generated; the mean and variance of the parameter estimates are reported in Table 2.2.

An especially interesting point is that simple maximization of the likelihood or posterior density simply breaks down for positive values of p, as both the bias and variance of the estimates of p and μ become large as p grows. This is most likely due to multimodality; for positive values of p, the zero-modified model is essentially a mixture model, and there is a possibility that the maximization algorithms converge to a local mode. Estimates obtained with these methods should be considered carefully when the estimate of p is positive; re-analyzing the data using a method well-suited to mixture models (e.g., based on a latent variable approach, as the EM algorithm or Gibbs sampling) is a possible solution. Posterior expectation estimates do not suffer from this disadvantage and remain reliable as p varies.

Also of interest is the fact that the adaptive Monte-Carlo, adaptive Gauss-Hermite quadrature and GLL quadrature all produce different estimates, and those differences can be important in some cases. This can be surprising, as all three methods give an estimate of the same quantities - the posterior expectation of p, μ, and λ. This difference can be partly explained by the use of iterative methods, which can sometimes converge to local rather than global modes. We also note that all three methods integrate transformed parameters, and then inversely transform the resulting estimates to obtain estimates of (p, μ, λ). In the case of the GLL, this transformation is simple –working with $\tilde{p}$ instead of p to ensure a constant domain of integration– but the adaptive methods take the logarithm of the parameters so that the transformed parameter space is $\mathbb{R}^3$.

We next consider testing hypotheses on p. As mentioned in section 4, posterior probabilities can be viewed as expectations of indicator functions – the testing methods considered will therefore be the three methods for obtaining posterior expectation estimates. Let $p = -0.1$. Figure 2.1 plots the posterior probabilities, obtained with the three estimation methods considered, of $H_- : \ p < 0, \ H_0 : \ p = 0$, and $H_+ : p > 0$ as a function of the sample size. The numerical quadratures lead to a more conservative test than adaptive Monte-Carlo. To test if $p = 0$, we can easily obtain $P(p \neq 0) = P(p < 0) + P(p > 0)$.

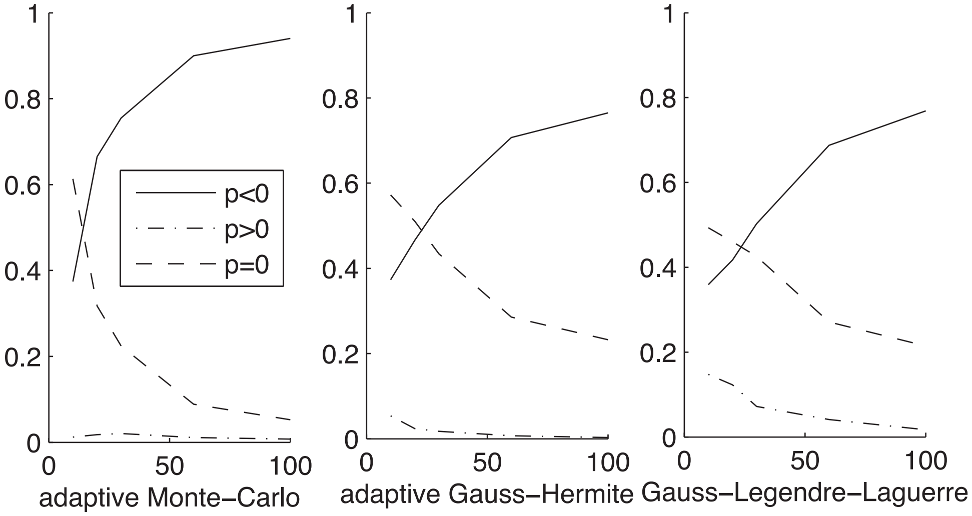

Figure 2.1: Posterior probabilities for tests as a function of n.

Under the classical paradigm, we consider two methods for testing $H_0 : \ p = 0$ against $H_1 : \ \ p \neq 0$: the likelihood ratio test, and Rao's score test. Tables 2.3 and 2.4 compare the power of all testing procedures discussed in this paper (likelihood ratio, score, and posterior probabilities computed using adaptive Monte-Carlo, adaptive Gauss-Hermite quadrature and Gauss-Legendre-Laguerre quadrature) as a function of n and p, respectively. While there is no testing procedure uniformly better (or worse) than the others, we can note that the score test performs poorly when p is negative. The Bayesian test with adaptive Gauss-Hermite quadrature seems the most powerful for positive values of p, but it is outperformed by both other Bayesian tests and by the likelihood ratio test when p is negative. Both classical tests are much faster than all three Bayesian tests; however, should Bayesian testing be preferable, the GLL approach is the fastest. This, combined with its generally good performance, can make it an attractive choice when working in a Bayesian context. Interestingly, a recent comparison of testing procedures in the discrete case (Min and Czado, 2010) has found that while score tests had been previously recommended in the literature to test for zero-modification, they were outperformed by the likelihood ratio test and the Wald test.

Finally, we apply these estimation methods to the data set analyzed in Labrecque-Synnott and Angers (2009). The data are bimonthly millimetric rainfall in Montreal for the months of May and June, from 1943 to 1992. Parameter estimates, as can be seen in Table 2.5, are mostly consistent amongst the different methods. In particular, the MAP and ML estimates are nearly identical. This is not surprising, given our use of a non-informative prior and the large sample size ($n = 217$). The three Bayesian and two classical tests (using $\alpha = 0.05$), as seen in Table 2.6, are in agreement that p is negative, although the iterated GLL quadrature is still the most conservative.

2.7 Concluding remarks

In this paper, we have considered the use of continuous zero-modified models within a Bayesian framework. The choice of non-informative priors was discussed, and several methods for obtaining posterior expectation estimates have been proposed. The performance of these estimators has been assessed through a simulation study. While the three proposed methods of estimation of the posterior expectation are generally in agreement, they also outperform maximum likelihood and maximum posterior estimators when p is positive. Testing hypotheses on p in this

Table 2.3: Power as a function of n, with $p = -0.1$, $\mu = \lambda = 20$

n	LK	Score	Monte-Carlo	Gauss-Hermite	GLL
10	0.00	0.02	0.00	0.00	0.02
20	0.18	0.03	0.14	0.08	0.26
30	0.30	0.04	0.20	0.18	0.35
60	0.52	0.05	0.51	0.46	0.63
100	0.61	0.06	0.64	0.64	0.73

Table 2.4: Power as a function of p, $\mu = \lambda = 20$, $n = 50$

p	LK	Score	Monte-Carlo	Gauss-Hermite	GLL
-0.1	0.45	0.10	0.36	0.28	0.34
-0.075	0.32	0.08	0.27	0.19	0.21
-0.05	0.19	0.07	0.24	0.18	0.18
-0.025	0.07	0.07	0.12	0.13	0.15
-0.01	0.05	0.06	0.06	0.11	0.13
-0.005	0.03	0.05	0.09	0.09	0.08
0	0.05	0.06	0.06	0.06	0.05
0.05	0.01	0.09	0.13	0.17	0.09
0.1	0.07	0.19	0.24	0.29	0.16
0.2	0.15	0.37	0.51	0.65	0.41
0.3	0.22	0.43	0.70	0.85	0.78

Table 2.5: Estimation of a real data set of aggregated rainfall data

	p	μ	λ
ML	-0.0569	22.7	22.31
MAP	-0.0569	22.7	22.43
Monte-Carlo	-0.0557	22.61	22.45
Gauss-Hermite	-0.0551	22.61	22.48
GLL	-0.0562	22.58	22.74

Table 2.6: Testing a real data set of aggregated rainfall data

	χ_1^2	p-val	$P(p < 0 \| x)$	$P(p = 0 \| x)$	$P(p > 0 \| x)$
LR	6.131	0.013	n/a	n/a	n/a
Score	8.696	0.003	n/a	n/a	n/a
Monte-Carlo	n/a	n/a	0.992	0.0063	0.0015
Gauss-Hermite	n/a	n/a	0.962	0.0321	0.0059
GLL	n/a	n/a	0.88	0.116	0.004

kind of model was also considered, and can be viewed as the posterior expectation of an indicator function. Re-analysis of the aggregate rainfall data presented in Labrecque-Synnott and Angers (2009) confirms that this data set is best modeled by a zero-deflated model.

REFERENCES

Aitchison, J. (1955). On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of the American Statistical Association*, **50**(271):901–908.

Angers, J. and Biswas, A. (2003). A Bayesian analysis of zero-inflated generalized Poisson model. *Computational statistics & data analysis*, **42**(1-2):37–46.

Dietz, E. and Böhning, D. (2000). On estimation of the Poisson parameter in zero-modified Poisson models. *Computational statistics & data analysis*, **34**(4):441–459.

Fernandes, M., Schmidt, A., and Migon, H. (2009). Modelling zero-inflated spatio-temporal processes. *Statistical Modelling*, **9**(1):3.

Feuerverger, A. (1979). On some methods of analysis for weather experiments. *Biometrika*, **66**(3):655–658.

Fletcher, D. (2008). Confidence intervals for the mean of the delta-lognormal distribution. *Environmental and Ecological Statistics*, **15**(2):175–189.

Gupta, P., Gupta, R., and Tripathi, R. (2005). Score test for zero inflated generalized Poisson regression model. *Communications in Statistics-Theory and Methods*, **33**(1):47–64.

Hall, D. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, **56**(4):1030–1039.

Jansakul, N. and Hinde, J. (2002). Score tests for zero-inflated Poisson models. *Computational statistics & data analysis*, **40**(1):75–96.

Labrecque-Synnott, F. and Angers, J. (2009). An Extension of zero-modified models to the continuous case. Technical Report CRM-3288, Université de Montréal.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**(1):1–14.

Min, A. and Czado, C. (2010). Testing for zero-modification in count regression models. *Statistica Sinica*, **20**:323–341.

Naylor, J. and Smith, A. (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics*, **31**(3):214–225.

Ridout, M., Hinde, J., and Demétrio, C. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, **57**(1):219–223.

Rubinstein, R. and Kroese, D. (2008). *Simulation and the Monte Carlo method*. Wiley-interscience.

Singh, S. (1963). A note on inflated Poisson distribution. *Journal of the Indian Statistical Association*, **1**(3):140–144.

Stefansson, G. (1996). Analysis of groundfish survey abundance data: combining the GLM and delta approaches. *ICES Journal of Marine Science*, **53**(3):577.

Tian, L. (2005). Inferences on the mean of zero-inflated lognormal data: the generalized variable approach. *Statistics in medicine*, **24**(20):3223–3232.

Van den Broek, J. (1995). A score test for zero inflation in a Poisson distribution. *Biometrics*, **51**(2):738–743.

CHAPTER 3

BAYESIAN MODEL-BASED CLUSTERING OF CONTINUOUS ZERO-MODIFIED DATA

Cet article sera prochainement soumis a *Statistics & Probability Letters.* Le premier auteur est Félix Labrecque-Synnott et le coauteur est le directeur de recherche, Jean-François Angers.

La question de recherche derrière cet article provient de discussions entre les coauteurs, et vient rejoindre certains travaux antérieurs du premier auteur, qui a élaboré et programmé la méthode proposée basée sur les distributions marginales des observations. Il a également programmé les différentes méthodes classiques auxquelles fut comparée la méthode bayésienne proposée. Bien qu'il ait initialement été prévu d'appliquer cette méthode à différentes stations de mesures des précipitations au Québec, la Colombie-Britannique fut plutôt retenue à la suggestion du deuxième auteur, le relief de cette province étant propice à un découpage plus intéressant. Finalement, la récupération et la préparation des données (validation et passage d'un pas de temps journalier à un pas de temps hebdomadaire) fut effectuée par le premier auteur.

Un algorithme basé sur la distribution marginale des observations conditionnellement à une classification donnée est présenté en détail à la section 3. La section 4 contient une description de certains critères d'arrêts classiques, ainsi que des résultats de simulation comparant les diverses méthodes. La section 5 contient l'application à des données de précipitations hebdomadaires en Colombie-Britannique motivant cette approche.

ABSTRACT

We consider model-based clustering of continuous zero-modified data under contiguity constraints, where there is a spatial relationship between the different samples. In this context, clusters should be contiguous, and very similar clusters should not be joined together unless they are adjacent. As each sample is assumed to follow a continuous zero-modified model, stopping rules and clustering procedures based on the normality of the data seem less appealing. A Bayesian approach based on posterior probabilities is proposed. The performance of the outlined methodology is compared to that of existing stopping rules using a simulation study. The proposed approach is then applied to the analysis of aggregate rainfall data at different measurement stations in the province of British Columbia. A sample is associated to each measuring station, and those samples are re-grouped into clusters when allowed by adjacency and statistical similarity constraints.

3.1 Introduction

This paper focuses on model-based clustering within the context of continuous zero-modified data. Let $X_i = X_{i1}, \ldots, X_{ini}$, for $i = 1, \ldots, k$ be a collection ok k samples of sizes $n_1, \ldots, n_k$ respectively, and corresponding to k points in a given space, with a spatial relationship. For example, each sample could come from a given geographical location. For a given sample i, the X_{ij} are assumed to be independent identically distributed random variables with probability density function $f_{xi}(x|p, \theta) = pf_0(x) + (1-p)f_1(x|\theta)$, where p and θ are unknown parameters, f_0 is a probability density function on $[0, x_0]$, and f_1 is a probability density function on $\mathbb{R}^+$. This class of continuous zero-modified models was introduced in Labrecque-Synnott and Angers (2009), where maximum likelihood estimation was considered, and was studied within the Bayesian framework in Labrecque-Synnott and Angers (2010), which also discussed hypothesis testing for this class of models. We are interested in clustering such data, where similar (and adjacent) samples can be combined to create homogenous and contiguous regions. Two samples should not be in the same cluster unless they are adjacent, or can be linked by pairwise-adjacent samples within the same cluster. Such clustering procedures would then be used if the data point or samples to be clustered correspond to measurements taken at different locations.

We begin with a brief overview of the model, its properties and suitable priors in Section 2. In Section 3, we describe the proposed algorithm of Bayesian model-based clustering. A simulation study and its results are presented in Section 4, while a real dataset is analyzed in Section 5. Finally, Section 6 contains some concluding remarks.

3.2 The model

Let X be a random variable with probability density function given by

$$f_X(x|p,\theta,x_0) = pf_0(x)\mathbb{I}_{[0,x_0]}(x) + (1-p)f_1(x|\theta), \tag{3.1}$$

where $f_0(x)$ is a probability density function on $[0, x_0]$ and $f_1(x|\theta)$ is a probability density function on $\mathbb{R}^+$. We then say that equation (3.1) is a zero-modified model, $f_0(x)$ is the zero-modifying distribution and $f_1(x|\theta)$ is the base distribution. More precisely, if $p \in [0, 1]$ it is said to be a zero-inflated model, with a higher proportion of observation in $[0, x_0]$ than under the probability density function $f_1(x|\theta)$. Conversely, if p is smaller than 0 (but larger than a certain bound p_{min} depending on both θ and x_0), equation (3.1) remains a valid probability density function and is said to be zero-deflated. The parameters of interest here are p and θ; it is assumed that x_0 is either known or that its value can be adequately assessed by expert opinion. We denote the set of parameters of interest (p, θ) by ω; similarly, while Θ refers to the parameter space under the base model, Ω represents the parameter space of the zero-modified model.

The main difficulty in working with this class of models is the dependance between p and θ, induced partly by the bound p_{min}. While it is generally impossible to find a closed-form expression for the maximum likelihood estimators (MLE) of the model, various numerical methods can be used to obtain estimates, usually with good performance (Labrecque-Synnott and Angers, 2009). Bayesian estimation is a bit more tricky, as it is often based on the posterior expectation or MAP estimator of the model parameters. Once again, there is usually no closed-form expression for this quantity. Furthermore, the dependency between p and θ will mean that, for large regions of the parameter space, the likelihood function $L(\omega|x)$ will be negligible, as the complex relationship between θ and p means that only a certain range of values of θ will be "compatible" (have non-negligible likelihood) with a given value of p. Adaptive Monte-Carlo methods or adaptive numerical quadrature can solve this problem, but they usually have a heavy computational load, as seen in Labrecque-Synnott and Angers (2010). It should be noted that, in evaluating $L(\omega|x)$ at a certain point ω, the most computationally-intensive task is determining the bound $p_{min}(\theta)$. A possible alternative to obtain posterior expectations (or the marginal $m(x)$) when working with multiple samples would therefore be to consider non-adaptive Gaussian quadrature, and to compute the bound $p_{min}(\theta)$ at each quadrature node only once, storing these bounds to compute the posterior expectation for further samples. This method is slightly less accurate than adaptive Monte-Carlo/quadrature, but much faster.

As in Labrecque-Synnott and Angers (2009, 2010), we consider

$$f_0(x) \propto (x_0 - x)^{\tau} \mathbb{I}_{[0,x_0]}(x)$$
$$f_1(x|\mu, \lambda) \propto \exp\{\frac{-|x-\mu|}{\lambda}\} \mathbb{I}_{\mathbb{R}^+}(x).$$

Note that this choice of f_1 corresponds to the Laplace distribution with a location parameter truncated on the positive real numbers. This choice of f_0 is simple, easy to work with, flexible and preserves continuity of the zero-modified model (Labrecque-Synnott and Angers, 2009). The choice of f_1 is mainly motivated by the necessity of having an easily-computed probability density function defined on $\mathbb{R}^+$ whose support includes 0, and which is strictly positive at 0, so as to allow for zero-deflation.

3.3 Contiguous model-based clustering

Since the different samples that we wish to cluster correspond to different measuring points with known geographical locations, we would like our clusters to be made of contiguous samples – that is, not to contain geographically disjoint samples. For simple cases, with a low number of measuring points arranged along a line or in a regular grid, this constraint is fairly intuitive. We propose to use a $k \times k$ connectivity matrix, whose $(i, j)^{\text{th}}$ entry is 1 if the samples i and j are adjacent, and 0 otherwise. Such a matrix can easily accommodate both simpler (lines, grids) and more complex relationships (such as the proposed application in Section 6).The proposed model-based clustering algorithm can easily make use of the information contained in the connectivity matrix to produce contiguous clusters. Classical agglomerative hierarchical clustering methods can also easily be adapted to only merge adjacent clusters using the connectivity matrix. We consider agglomerative hierarchical clustering, as it is easier to check if two clusters have any adjacencies and can be regrouped than to check if removing a sample from a cluster makes that cluster non-homogenous. Checking if this contiguity constraint is satisfied is also more complex when using relocation-based partitional clustering methods, such as k-means clustering. Divisive hierarchical clustering is also conceptually more complex, as it must determine at each step which cluster to split and how to split it. Since it is also less common in the literature (Fraley and Raftery, 1998), it was not considered here.

Starting with $n_{samples}$ clusters, containing each the observations from a measuring point, we propose to iteratively use Bayesian hypothesis testing to determine whether two clusters should be combined or not. Under the Bayesian paradigm, a way of testing the hypothesis that ω lies in certain subset Ω_A of the parameter space is to obtain its posterior probability. Let us suppose that at a given point in the clustering procedure, where there are k clusters, we want to test whether or not to combine two clusters. Our hypotheses are "H_0 : the pair $\{i,j\}$ should be combined" (which implies that there are $k-1$ clusters in the dataset) and "H_1 : the pair $\{i,j\}$ should not be combined" (and there are k clusters in the dataset). Under both of these hypotheses, the parameters ω can take any value within Ω, as long as we combine – or not – the two clusters in question. Under H_0, we must have that $p_i = p_j$ and $\theta_i = \theta_j$, but the actual values of the parameters are unimportant, and

$$P(H_0|x) \propto = \int_{\Theta} \int_{p_{min}(\theta_i)}^{1} L(p_i, \theta_i|x_i) L(p_i, \theta_i|x_j) \pi(p_i|\theta_i) \pi(\theta_i) dp_i d\theta_i.$$

Conversely, under H_1, the specific values taken by p_i, p_j, θ_i and θ_j are also unimportant; we are only concerned with the fact that $(p_i, \theta_i) \neq (p_j, \theta_j)$, and

$$\begin{aligned} P(H_1|x) \propto & \int_{\Theta} \int_{p_{min}(\theta_i)}^{1} L(p_i, \theta_i|x_i) \pi(p_i|\theta_i) \pi(\theta_i) dp_i d\theta_i \\ & \times \int_{\Theta} \int_{p_{min}(\theta_j)}^{1} L(p_j, \theta_j|x_j) \pi(p_j|\theta_j) \pi(\theta_j) dp_j d\theta_j. \end{aligned}$$

In both cases, we integrate on the entire parameter space – p_i is integrated on $[p_{min}, 1]$, and θ_i is integrated on Θ. In the case of H_1, p_j and θ_j are also integrated on $[p_{min}, 1]$ and Θ, respectively. The only difference between the hypotheses is the dimension of the parameter space considered. Essentially, we have to compute the marginal distribution of the observations when combining the clusters, and when keeping them separate. Our decision to merge or not the clusters is based on these marginal distributions, which correspond to the posterior probabilities of H_0 and H_1.

The proposed Bayesian model-based clustering algorithm is as follows:

1. For each sample, compute the maximum likelihood parameter estimators $\hat{p_i}$ and $\hat{\theta}_i$, and the marginal density $m(x_i)$;

2. Compute a distance (or dissimilarity) matrix between the samples, based on either the observations or the parameter estimates;

3. Set the distance between non-adjacent samples and the diagonal of the distance matrix to ∞;

4. Find the samples couple X_i, X_j with the smallest distance;

5. Test whether or not to combine those two samples. The samples are combined if $m(x_{ij \text{ combined}}) > m(x_i) \times m(x_j) / \exp(\dim(\Theta)+1)$, where $\dim(\Theta)$ denotes the dimension of the parameter space associated to a single probability density function $f_1(x|\theta)$. This is a penalty term similar to the one found in the AIC, as can be seen below;

6. If the two samples are combined, recompute the distance matrix; otherwise, set the distance between X_i and X_j to ∞;

7. Steps 3 to 6 are re-iterated until all entries of the distance matrix are ∞ (indicating that, for every remaining pair of clusters, either combining the clusters has already been tested and rejected, or the pair is non-adjacent) or until all samples are combined in the same cluster.

The parameter estimators and the marginal distribution in Step 1 will usually have to be obtained numerically. Here, we have used the iterated Gauss-Legendre-Laguerre quadrature described in Labrecque-Synnott and Angers (2010). In Step 6, if the samples are merged, the dimension of the distance matrix is adjusted (to account for the fact that there is now one less sample to consider) before being recomputed. A variety of distance metrics can be used to compute the distance matrices. Here, we have used the Euclidean distances between the sample means.

Note that the distances computed in Step 2 do not refer to spatial or geographical distances, but to a given statistical distance or dissimilarity between samples. Setting the distance between non-adjacent samples to ∞ and searching for the minimum distance (in Steps 3 and 4) accommodates the constraint that clusters must be homogenous, in the sense that they must be composed of adjacent samples. In Step 5, rather than simply comparing the marginal densities corresponding to combined and separate clusters, we have divided the marginal density corresponding to distinct clusters by a penalty term. This approach is very similar to the Akaike and Bayes information criteria (AIC and BIC, respectively), widely used in variable selection problems (Burnham and Anderson, 2004, Yang, 2005, Kuha, 2004). In fact, the penalty used here – dividing the marginal density by the exponential of the number of additional parameters – is equivalent to the penalty in the AIC – subtracting, from the logarithm of the likelihood, the number of parameters (Akaike, 1974).

3.4 Simulation results

Performance of the proposed clustering method is best assessed through simulation studies. We are interested, on the first hand, in assessing the reliability of the proposed algorithm – that is, its capacity to detect the "true" clusters – both when different clusters are present and when all samples belong to the same cluster. On the other hand, we want to see whether the proposed model-based algorithm provides better performance than usual clustering methods. To this end, in addition to the Bayesian algorithm based on the marginals given above, we also apply hierarchical agglomerative clustering to our simulations.

This type of clustering is conceptually simple (and commonly found in the literature): starting with a cluster per object, a distance matrix between clusters is computed. While there are several ways to compute the distance matrix, we consider the Euclidean distance between cluster centroids. Since our data are univariate, this is equivalent to take the Euclidean distance between cluster means. Once the distance matrix is computed, the following steps will be iterated:

1. merge the two closest clusters;

2. compute the new distance matrix.

The main challenge in hierarchical agglomerative clustering lies in determining the number of clusters. In certain applications, expert opinion will be used. In other cases, statistical procedures or stopping rules will serve to determine the number of clusters. Next, we briefly describe four known clustering procedures that will be compared to the proposed method.

3.4.1 Je(2)/Je(1)

The Je(2)/Je(1) , introduced in Duda et al. (2001), is a statistical test meant to test "H_0 : the data considered belong to a single cluster" against "H_1 : the data belong to two different clusters." It is used as a stopping rule, before merging any two clusters in agglomerative clustering. The test uses only the data points contained in the two clusters to be merged (or not) and is independent of clusters not currently being considered for agglomeration. Clusters are agglomerated until there is a single cluster remaining, or until H_0 is rejected, at which point the procedure stops.

Let x_{1i}, $i = 1, \ldots, n_1$ and x_{2j}, $j = 1, \ldots, n_2$ be the data points of the two samples being considered, $\bar{x}_1$ and $\bar{x}_2$ be the means of the two clusters and $\bar{x}$ the overall mean of the two clusters. Then Je(2)/Je(1) is defined as:

$$\frac{Je(2)}{Je(1)} = \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)^2}{\sum_{i=1}^{n_1}(x_{1i} - \bar{x})^2 + \sum_{j=1}^{n_2}(x_{2j} - \bar{x})^2}.$$

Duda et al. (2001) have shown that, for large n, if the observations form a random sample from a $Normal(\nu, \sigma^2)$, then Je(1) is approximately $Normal(n\sigma^2, 2n\sigma^4)$ and Je(2) will be approximately $Normal(n(1/2/\pi), 2n(1-8/\pi^2))$. Using these approximate distributions, and estimating σ^2 by $Je(1)/n$, the following stopping rule is found: reject H_0 if Je(2)/Je(1) $\leq 1 - 2/\pi - Z_{1-\alpha}\sqrt{\frac{2(1-8/\pi^2)}{n_1+n_2}}$, where α is the significance level of the test.

3.4.2 Beale test

The Beale test, introduced in Beale (1969), is also a stopping rule used in hierarchical agglomerative clustering. It also tests "H_0 : the data considered belong to a single cluster" against "H_1 : the data belong to two different clusters.". Like the Je(2)/Je(1), it is also based only on the two clusters currently considered. The test is based on mean square deviation from cluster centroids, which is simply the sum of squared deviations from the mean when the data are unidimensional. Using the same notation as above, and defining $W_1 = \sum_{i=1}^{n_1}(x_{1i} - \bar{x})^2 + \sum_{j=1}^{n_2}(x_{2j} - \bar{x})^2$ and $W_2 = \sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)^2$, the test statistic is:

$$B = \frac{\frac{W_1 - W_2}{W2}}{\frac{n_1+n_2-1}{n_1+n_2-2}2^2 - 1}.$$

Under the null hypothesis, B is approximately F_{1,n_1+n_2-2}. The hierarchical agglomerative clustering stops as soon as the hypothesis that two given clusters should be merged is rejected.

3.4.3 pseudo-F

Unlike the Je(2)/Je(1) and the Beale test, the pseudo-F, introduced in Caliński and Harabasz (1974) is not a stopping rule, but a statistic to be computed for $k = 2, \ldots, n_{samples}$ clusters, where $n_{samples}$ is the initial number of clusters before starting the agglomerative clustering process. There is no statistical test associated with the pseudo-F, we simply select the number of clusters maximizing it.

For k clusters, the pseudo-F is given by

$$\frac{\sum_{r=1}^{k}(\bar{x}_r - \bar{x})^2/(r-1)}{\sum_{r=1}^{k}\sum_{i=1}^{n_r}(x_{r_i} - \bar{x}_r)^2/(n-r)}.$$

It should be noted that, contrarily to the Je(2)/Je(1) and Beale stopping rules, the pseudo-F index can only point out the best division of the data into two or more clusters, as the index is not defined for $k = 1$. Using the pseudo-F to determine the number of clusters in this case should therefore be preceded (or followed) by a test comparing H_0 : all data belongs to a single cluster against H_1 : two or more clusters are present in the data.

3.4.4 AIC

We also consider using the AIC to determine the correct number of clusters. A similar approach has been used before in Chen and Gopalakrishnan (1998) and Fraley and Raftery (1998), using the BIC instead of the AIC. Clusters are determined using the centroid method; for $k = 1, \ldots, n_{samples}$, the maximum likelihood estimators of the model parameters are computed for each cluster. Within the context of model-based clustering, this is an intuitive approach, as each additional cluster corresponds to $\dim(\Omega) = \dim(\Theta) + 1$ extra parameters, but will likely produce a higher likelihood. The problem of determining the number of clusters, in this particular context, can therefore be seen as determining the number of model parameters that strike the best balance between a high likelihood and low model complexity. This approach is also quite similar to the marginal-based clustering presented above. Indeed, the marginal $m(x)$ can be seen as the expectation of the likelihood $L(\omega|x)$ with respect to the prior distribution $\pi(\omega)$. The main difference between minimizing the AIC and our method is therefore that the former is based on the mode of the likelihood function, while the latter is based on its expectation.

Finally, we have also considered using the likelihood ratio test as a stopping rule, combining clusters until the null hypothesis that the two closest clusters should be combined is rejected.

We consider simulation scenarios where measures are taken at nine distinct points, arranged in a grid as follows:

$$\begin{array}{ccc}
1 & 2 & 3 \\
4 & 5 & 6 \\
7 & 8 & 9
\end{array}$$

We consider a sample to be adjacent to any sample directly to its side, above or below it. For example, sample 5 is adjacent to samples 2, 4, 6 and 8, but not to samples 1, 3, 7 and 9. This gives us the following connectivity matrix, where the

$(i,j)^{th}$ entry is 1 if i and j are adjacent and 0 otherwise:

$$\begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

In the first simulation scenario that we consider, all samples belong to the same cluster, with parameter values $(p, \mu, \lambda, \tau, x_0) = (-0.05, 20, 10, 0.05, 10)$. In the second scenario considered, samples 3, 6, and 9 are regrouped in a first cluster with parameters $(p_1, \mu_1, \lambda_1) = (0, 10, 8)$, while the remaining samples belong to a second, larger cluster with $(p_2, \mu_2, \lambda_2) = (-0.05, 20, 10)$. In this last scenario, the values of $x_0 = 10$ and $\tau = 0.05$ are the same for both clusters. The two densities used in scenario 2 are plotted in Figure 3.1. Under each scenario, we generate 500 sets of nine samples of size $n = 60$ with the parameter values given above. For each generated set of samples, we apply the proposed algorithm, as well as centroid hierarchical clustering using the Je(2)/Je(1), Beale test, pseudo-F, LRT and AIC to determine the number of clusters. The empirical probability of correctly assigning all samples to the correct cluster is given in Table 3.1. These are computed as the proportion of simulation runs in which all samples were correctly assigned. The average number of misclassified sites is given in Table 3.2.

While at first glance, these numbers do not seem to draw a very clear picture (other than the AIC and LRT being significantly outperformed), some precisions must be made for a better interpretation of these results. First, the proposed method – using the marginals combined with an AIC-style penalty term to deter-

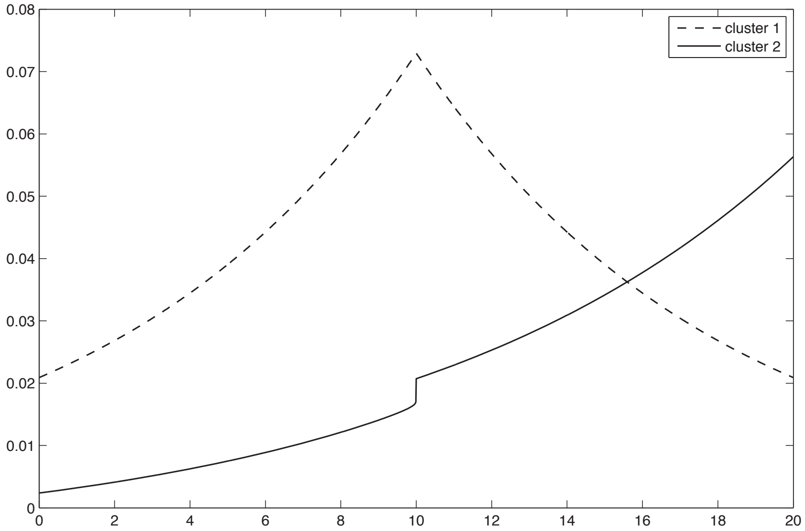

Figure 3.1: The two probability density functions used in scenario 2

Table 3.1: Probability of identifying the correct clusters

Method	Scen. 1	Scen. 2
Bayesian marginal	0.972	0.752
Je(2)/Je(1)	1.000	0.724
Beale	1.000	0.682
Likelihood ratio	0.522	0.474
pseudo-F	n/a	0.720
AIC	0.424	0.458

Table 3.2: Average number of misclassified sites

Method	Scen.1	Scen. 2
Bayesian marginal	0.102	0.600
Je(2)/Je(1)	0.000	0.676
Beale	0.000	0.936
Likelihood ratio	1.918	1.996
pseudo-F	n/a	0.730
AIC	2.146	1.822

mine whether clusters should be regrouped – performs well when all samples belong to the same cluster, in effect having a probability of type I error or the order of 2.5%. It is also the method that is most likely to correctly assign all samples when different clusters exists, and it has the lowest misclassification rate in that case. Both the Beale and Je(2)/Je(1) tests are perfectly accurate when all samples belong to the same cluster, and the Je(2)/Je(1) test is only slightly outperformed by the marginal-based method in the second scenario. However, the question of the test levels must be mentioned at this point: the literature recommends specific levels for using the Beale and Je(2)/Je(1) tests in clustering, different from the more common levels $\alpha = 0.05, 0.01$ or 0.1. (Milligan and Cooper (1985) recommends using the formula appearing in Duda et al. (2001) to determine the cutoff point, with a normal score of 3.2, and a significance level of $\alpha = 0.005$ for the Beale test, while Gordon (1998) and Hardy and Andre (1998) recommend a normal score of 4.0 to compute the cutoff point for the Je(2)/Je(1) test) Using either the "classical" test levels or those proposed in the literature results in the systematic failure of both the Beale and Je(2)/Je(1) tests to identify multiple clusters; both methods systematically agglomerate all samples in a single cluster.

To get the better performance reflected in Tables 3.1 and 3.2, the cutoff point at which we reject the hypothesis that clusters must be combined has to be finely tuned throughout simulations, essentially choosing a cutoff point ensuring that we get acceptable results. This makes using these methods in practical applications without further and more extensive simulation studies a risky proposition, as there is no guarantee that the cutoff points which give good results in the two scenarios considered here will also give good results in other cases (for example, with different sample sizes, numbers of samples, different "true" repartition of the samples within clusters, different true parameter values for the samples, etc.).

While the pseudo-F does not suffer from this disadvantage, its probability to correctly assign clusters is slightly lower than the marginal-based method (and its average number of misclassified sites is slightly higher) under the second scenario. In addition, as mentioned above, the pseudo-F is not defined for a single cluster, meaning that it can only identify the “best” subdivision of data in 2 or more clusters. Its use should therefore be accompanied by testing if all samples can be agglomerated in a single cluster; such a test at a significance level $\alpha = 0.05$ would presumably give similar results than the marginal-based method under scenario 1. However, this would further increase the gap between the pseudo-F and the marginal-based method, as the pseudo-F correctly identifying the clusters (with probability 0.72) would then be conditional on the statistical test rejecting the hypothesis of a single cluster (with power unlikely to be 1). All these factors point to the proposed marginal-based method to be the most reliable clustering method for clustering samples of zero-modified data.

The poor performance of the AIC, relatively to the marginal-based method, could be explained by the fact that the AIC is based on the likelihood function evaluated at its mode. Since there is no closed-form expression for the MLEs under this model, there is no guarantee that the estimates obtained correspond to a global maximum of the likelihood function, which may make the likelihood evaluated at the MLE less stable than the marginal density. This could also explain the poor performance of the likelihood ratio test used as a stopping rule. Note that in this last case, we have used here a test of level $\alpha = 0.05$. While adjusting the test level could modify the performance of this approach, it is unlikely that it could become competitive with the proposed approach – taking a smaller α would lead to less rejections of H_0, improving performance under scenario 1 but worsening it under scenario 2, while taking a larger α would have the opposite effect.

3.5 Application to a real dataset

We have also applied the marginal-based clustering to weekly millimetric rainfall data in the month of January in the province of British Columbia, based on Environment Canada data. We used data from 21 sites, chosen as to include major population centres and at least one measuring station per district in Environment Canada’s CDCD database(available at `http://www.climate.weatheroffice.gc.`

`ca/prods\_servs/index\_e.html\#cdcd`), and focusing mostly on measuring stations active from the 1940s to the late 1990s/early 2000s. The measuring stations and maximum likelihood estimates for each station are given in Table 3.3. Figure 3.3 contains the map of the measuring stations considered, with combined samples contained within polygons. An example of a fitted sample (Prince Rupert) can be seen in Figure reffigfitted.

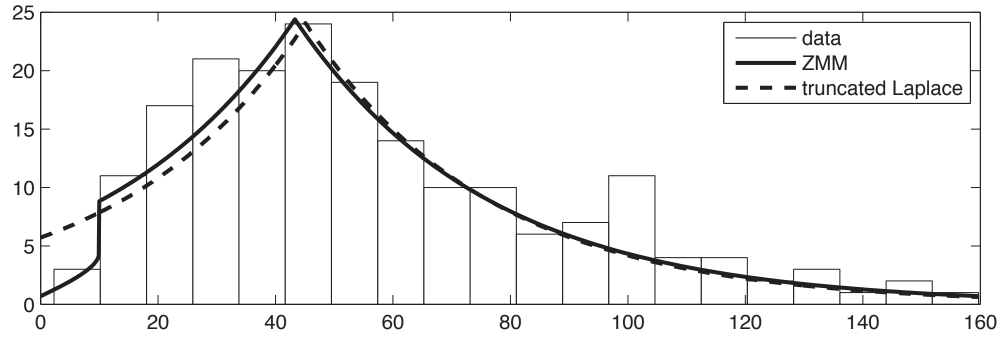

Figure 3.2: Prince Rupert data, fitted ZMM and fitted truncated Laplace

We can see quite a lot of variability in the MLEs, with rainfall being less abundant as we move to the North and away from the coasts. Of particular interest on the map are three measuring stations at roughly the same latitude and fairly close by, but not agglomerated: Prince Rupert, Terrace and Smithers. This is easily explained as these stations are separated by a mountain range, which undoubtedly has an influence on local rainfall. Further south, we have Nanaimo, Vancouver and Whistler in the same cluster, and belonging to the same coastal (hence, abundant in rainfall) region. it is also interesting to note that Prince George and Williams Lake, while separated by a couple hundred of kilometres, nevertheless can be identified as belonging to the same cluster – since the two measuring stations are along the same river, at roughly the same elevation, this is not surprising altogether.

Table 3.3: Measuring stations and their MLE, $(x_0, \tau) = (10, 0.05)$

ID	Cluster	Station	Neighbours	$\hat{\mu}$	$\hat{\lambda}$	$\hat{p}$
1	1	Victoria	3, 15	17.5	24.082	-0.066
2	2	Campbell	3, 4, 5, 11	36.5	26.285	-0.002
3	3	Nanaimo	1, 2, 5, 15	28.0	25.311	-0.019
4	2	Port Hardy	2, 7, 11	41.2	30.121	-0.015
5	4	Tofino	2, 3	84.2	49.953	-0.016
6	3	Whistler	2, 11, 15, 16	26.4	22.487	-0.024
7	5	Sandspit	4, 8	27.8	20.392	-0.059
8	6	Prince Rupert	7, 9	43.3	32.774	-0.042
9	7	Terrace	4, 8, 10, 12	22.6	27.100	-0.007
10	8	Smithers	9, 12	6.60	9.383	0.036
11	9	Tatlayoko Lake	2, 4, 6, 12, 14, 16	0.00	8.884	-0.238
12	10	Wistaria	9, 10, 11, 13, 14	4.80	8.992	0.1443
13	11	Prince George	12, 14	5.80	8.432	-0.323
14	11	Williams Lake	11, 12, 13, 16	4.40	7.959	0.173
15	3	Vancouver	1, 3, 6, 17, 18	28.4	19.139	-0.014
16	12	Shalalth	6, 11, 14, 20	4.40	16.711	0.030
17	13	Kelowna	15, 18, 20, 21	3.40	6.448	0.134
18	13	Penticton	15, 17	2.00	6.214	0.063
19	14	Grand Forks	21	5.20	8.208	-0.263
20	15	Kamloops	14, 16, 17, 21	0.00	5.922	-0.091
21	16	Revelstroke	17, 19, 20	21.00	13.726	0.074

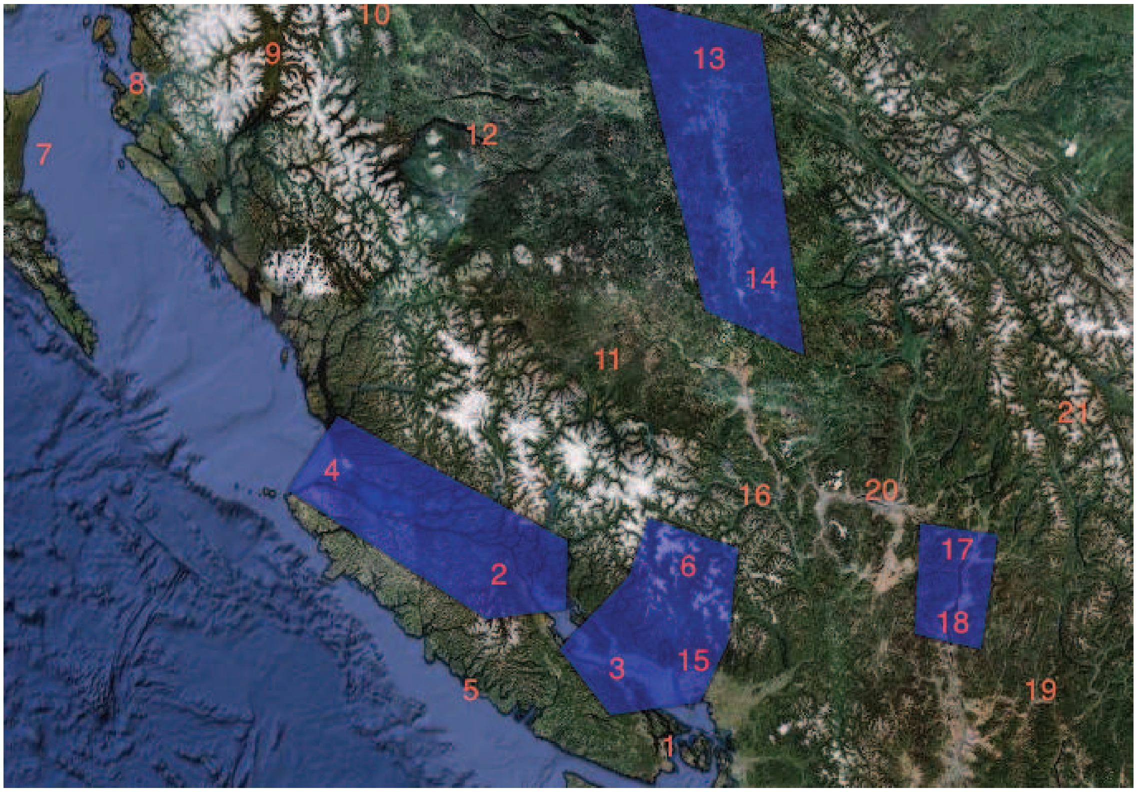

Figure 3.3: Measuring stations and clusters for British Columbia

3.6 Concluding remarks

The clustering of samples of continuous zero-modified data is the focal point of this paper. This problem is approached as an hypothesis testing problem within the Bayesian paradigm. In that case, we can see that, at a given iteration, the posterior probabilities that two clusters should or should not be combined correspond respectively to the marginal density $m(X_1, X_2)$ and to the product of the marginal densities $m(X_1) \times m(X_2)$. Our hypothesis was that, since these marginal densities incorporate information regarding the shape of the distribution of the observations, and since the zero-modified Laplace distribution is strongly non-normal (asymmetrical, heavy-tailed), this clustering approach would perform better than more common methods, based on sums of squares and (usually) supposing normality of the data. This was assessed throughout a simulation study; while classical methods remain competitive in terms of performance, they are burdened with complications (the necessity of testing for a single cluster when using the pseudo-F, and the necessity of fine-tuning the cutoff point for the Beale and Je(2)/Je(1) tests, which can be done easily for simulation studies, but not so much for practical applications). This makes the marginal-based method proposed in this paper a more attractive option for zero-modified data. An application to weekly rainfall data in British Columbia is also presented.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, **19**(6):716–723.

Beale, E. (1969). Euclidean cluster analysis. *Bulletin of the International Statistical Institute*, **43**(2):92–94.

Burnham, K. and Anderson, D. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*, **33**(2):261.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-Simulation and Computation*, **3**(1):1–27.

Chen, S. and Gopalakrishnan, P. (1998). Clustering via the Bayesian information criterion with applications in speech recognition. *IEEE International Conference On Acoustics Speech And Signal Processing*, **2**:645–648.

Duda, R., Hart, P., and Stork, D. (2001). *Pattern classification*. Citeseer.

Fraley, C. and Raftery, A. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, **41**(8):578.

Gordon, A. (1998). How many clusters? An investigation of five procedures for detecting nested cluster structure. In *Data science, classification, and related methods: proceedings of the fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan, March 27-30, 1996*, page 109. Springer.

Hardy, A. and Andre, P. (1998). An investigation of nine procedures for detecting the structure in a data set. In *Advances in data science and classification: proceedings of the 6th Conference of the International Federation of Classification Societies (IFCS-98), Università" La Sapienza", Rome, 21-24 July, 1998*, page 29. Springer Verlag.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research*, **33**(2):188.

Labrecque-Synnott, F. and Angers, J. (2009). An Extension of zero-modified models to the continuous case. Technical Report CRM-3288, Université de Montréal.

Labrecque-Synnott, F. and Angers, J. (2010). Bayesian estimation and testing for continuous zero-modified models. Technical Report CRM-3303, Université de Montréal.

Milligan, G. and Cooper, M. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50**(2):159–179.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, **92**(4):937.

CHAPITRE 4

CONCLUSION

Dans le cas d'un modèle discret à sur-représentation de zéros, le paramètre p peut être interprété comme une probabilité de mélange (le modèle est alors vu comme un mélange entre une loi donnée et une masse de Dirac à 0) ou comme un ajout direct à la fonction de masse au point 0 (et donc à la probabilité que X prenne la valeur 0). En utilisant cette seconde interprétation, il est facile de voir que p peut également prendre des valeurs négatives si la fonction de masse de base $f_1(x|\theta)$ est strictement positive au point 0 ; on parlera alors de sous-représentation de zéro. Ce modèle ne peut plus être interprété comme un mélange, et p représente alors une quantité retirée à la fonction de masse au point 0 (et redistribuée également entre les autres points du support de $f_1(x|\theta)$).

Dans le cas d'un modèle continu à sur-représentation de zéros, le modèle peut encore une fois être vu comme un mélange entre une densité de base $f_1(x|\theta)$ et une masse de Dirac en 0, où p représente un paramètre de mélange. Il est également possible de l'interpréter, tout simplement, comme la probabilité que la variable X soit identiquement égale à zéro. Aucune de ces deux interprétations ne nous permet d'utiliser des valeurs négatives de p; les modèles continus à sous-représentation de zéros sont d'ailleurs absents de la littérature scientifique (excluant les articles présentés dans cette thèse). Bien qu'il soit possible de diminuer (au d'augmenter) la valeur de $f_1(0|\theta)$, changer la valeur d'une fonction de densité sur un ensemble de mesure nulle (par exemple, en un seul point) n'affectera pas les probabilités découlant de cette fonction de densité. Un tel changement restera donc sans effet. Par contre, il est possible de diminuer (ou d'augmenter) la fonction de densité de base $f_1(x|\theta)$ par une densité de modification $f_0(x)$ définie sur un intervalle $[0, x_0]$, ce qui nous donne un modèle pouvant traiter les cas de la sur-représentation et de la sous-représentation de zéros. À la limite, lorsque $x_0 \to 0$, $f_0(x)$ devient une masse de Dirac en 0, et le modèle continu modifié à zéro proposé devient le modèle continu à sur-représentation de zéros usuel. Le paramètre p est alors nécessairement non négatif.

Deux principales difficultés surviennent lors de l'utilisation de modèles continus modifiés à zéro tels que présentés ici. Premièrement, il est généralement impossible d'obtenir une forme analytique pour les estimateurs du maximum de vraisemblance (pour un certain choix de $f_0(x)$, il est possible d'obtenir un estimateur du maximum de vraisemblance pour p conditionnel à θ), du maximum de la densité *a posteriori*, et de l'espérance *a posteriori*. Des méthodes numériques, dont la convergence n'est pas garantie et au temps de calcul parfois long, doivent alors être utilisées. Deuxièmement, la borne inférieure sur p, $p_{min}(\theta)$, crée une dépendance entre p et θ, nécessitant un plus grand nombre de noeuds de quadrature (ou un plus grand échantillon de Monte-Carlo) afin d'obtenir des résultats fiables.

En termes d'estimation ponctuelle, plusieurs estimateurs ont été considérés : maximum de vraisemblance, maximum *a posteriori* et espérance *a posteriori*. Dans le cas du maximum de vraisemblance, outre la maximisation numérique directe de la fonction de vraisemblance, il est aussi possible de reformuler le modèle afin de retrouver un mélange. Des méthodes d'estimation telles l'algorithme EM peuvent alors être utilisées. L'estimation par espérance *a posteriori*, nécessitant d'intégrer plutôt que de maximiser la densité *a posteriori*, est habituellement plus lourde en termes de calculs. La dépendance entre p et θ augmente le nombre de noeuds de quadrature (ou la taille de l'échantillon de Monte-Carlo) nécessaire, ce qui rend les calculs encore plus longs. Deux solutions peuvent être envisagées pour résoudre ce problème. Il est possible d'utiliser des méthodes adaptatives, déplaçant les noeuds de quadrature (ou l'échantillon de Monte-Carlo) là où est concentrée la densité *a posteriori*, et pouvant donc donner de bons résultats avec un plus faible nombre de noeuds. Lorsque plusieurs échantillons doivent être analysés (ou lors d'études de simulation), il est également possible d'utiliser une quadrature gaussienne non adaptative, et de garder en mémoire les valeurs de $p_{min}(\theta)$ associées à chaque noeud de quadrature. Le calcul de $p_{min}(\theta)$ étant de loin l'étape la plus coûteuse en termes de temps de calculs dans l'évaluation de la densité *a posteriori* en un point donné, cette méthode permet d'augmenter la nombre de noeuds sans que les calculs ne deviennent trop longs.

Les tests d'hypothèses relatifs au signe de p (et permettant donc de savoir si il y a sur- ou sous-représentation de zéros, ou si $f_1(x|\theta)$ modélise adéquatement les observations) ont également été considérés. D'un point de vue classique, le test du rapport de vraisemblance et le test du score, fréquemment utilisé pour tester la sur-représentation de zéros sous des modèles discrets (Van den Broek, 1995, Jansakul et Hinde, 2002, Gupta *et al.*, 2005), furent étudiés. D'un point de vue bayésien, les tests d'hypothèses sur p se ramènent au calcul de l'espérance de fonction indicatrices, et donc à un problème d'estimation ponctuelle. Cependant, les tests bayésiens utilisant des méthodes adaptatives, celles-ci donnent souvent de très faibles probabilités à l'hypothèse $p = 0$, nécessitant soit d'accorder un poids *a priori* plus grand à cette hypothèse, soit de ne la rejeter que lorsque sa probabilité *a posteriori* est au-dessous d'un certain seuil (plus petit que 0,5 ; et à déterminer par simulation).

La méthodologie développée fut appliquée à des données montréalaises de précipitations bimensuelles au printemps. Différents critères (AIC, BIC, tests d'hypothèses classiques et bayésiens) appuient l'idée que ces données soient mieux modélisées par un modèle modifié à zéro que par la densité de base considérée. La classification d'échantillons de données continues modifiées à zéro fut également étudiée. Des données de précipitations hebdomadaires provenant de différentes stations de mesure en Colombie-Britannique on été considérées, de façon à les regrouper en différentes régions à précipitations de distribution similaire. Lors de la classification d'échantillons correspondants à des points géographiques, il est intéressant que les grappes résultantes soient contigües. Intuitivement, des méthodes de classification hiérarchiques agglomératives ne joignant deux grappes que si elles sont adjacentes devraient pouvoir produire des classifications de ce type. Différentes façon de déterminer le nombre de grappes ont été étudiées. Celle que nous proposons, basée sur la distribution marginale des observations conditionnelle à la répartition des échantillons entre les différentes grappes (pénalisée selon le nombre de paramètres), se démarque par sa bonne performance et sa simplicité d'usage. Les tests de Beale et du Je(2)/Je(1) donnent aussi de très bons résultats, mais le seuil critique au-delà duquel deux grappes ne seront pas combinées doit être soigneusement ajusté d'après les simulations. Les seuils plus usuels (par exemple, prendre $\alpha = 0.05$, ou encore utiliser l'un des seuils mentionnés dans la littérature (Milligan et Cooper, 1985, Gordon, 1998, Hardy et Andre, 1998)) donnent des résultats catastrophiques, où tous les échantillons sont systématiquement agglomérés en une seule grappe. Ceci peut s'expliquer par le fait que ces tests soient approximatifs et fondés sur une hypothèse de normalité des données qui n'est pas respectée dans les scénarios de simulation considérés. Le pseudo-F donne de bons résultats, mais ne peut indiquer si tous les échantillons devraient être combinés en une seule grappe, ou si plusieurs grappes sont présentes. Son utilisation devrait donc être précédée d'un test d'hypothèse. Finalement, les autres méthodes considérées (test du rapport de vraisemblance et minimisation de l'AIC) donnent des résultats nettement moins bons, pouvant être expliqués par une plus grande volatilité du maximum de la fonction de vraisemblance par rapport à la marginale des observations (qui correspond à l'espérance de la fonction de vraisemblance sous la distribution *a priori*).

4.1 Travaux futurs

(1) Modèles continus modifiés à zéro plus complexes Le développement des premiers modèles discrets modifiés à zéro (Singh, 1963) fut rapidement suivi du développement de modèles plus complexes. Certains ont incorporé des variables explicatives (Lambert, 1992, Hall, 2000, Ridout *et al.*, 2001) ou des effets aléatoires (Hall, 2000). D'autres furent développés pour données manquantes, censurées ou longitudinales (Hasan et Snedonn, 2009). La question de l'adéquation de la fonction de masse modifiée fut également étudiée : des tests pour la sur-dispersion d'une variable aléatoire de Poisson en présence de sur-représentation de zéros ont été développés, de même que des tests confrontant une loi de Poisson avec sur-représentation à une loi binomiale négative avec sur-représentation (Ridout *et al.*, 2001). En prenant cette thèse comme point de départ, un développement similaire serait possible dans le cas de modèles continus modifiés à zéro. En particulier, l'utilisation de variables explicatives, de modèles censurés et l'étude du choix de la densité de base pourraient permettre de mieux modéliser certains types de données et d'utiliser les modèles continus modifiés à zéro de façon prédictive plutôt qu'uniquement descriptive.

(2) Meilleure modélisation des données de précipitation agrégées Tel que mentionné au point précédent, l'incorporation de variables explicatives ou de modèles censurés pourraient permettre une meilleure modélisation. Par exemple, si nous considérons les applications présentées dans les articles (données de précipitation hebdomadaires ou bimensuelles), certaines données sont en fait censurées, le pluviomètre n'ayant recueilli que de l'information partielle (par exemples, traces de précipitations pour une date donnée, sans mesure précise ; précipitations de plus de x millimètres, etc). Cette censure n'a pas été considérée explicitement par les modèles utilisés dans cette thèse, le but étant de développer d'abord une fondation fiable pour la modification à zéro de modèles continus. Un traitement explicite de la censure pouvant être présente dans les données en améliorerait la modélisation.

(3) Choix de f_0 Bien que diverses alternatives pour le choix de f_0 aient été proposées dans le premier article, il serait envisageable d'étudier les différentes options possibles pour cette densité. En particulier, il serait intéressant de considérer une famille paramétrique $f_0(x|\theta_0)$, indexée par un paramètre θ_0 inconnu, permettant ainsi une meilleure adéquation du modèle aux données.

(4) Modèles continus modifiés Bien que la méthodologie développée et appliquée dans cette thèse traite de modèles continus modifiés à zéro (ou, pour être exact, sur $[0, x_0]$), les résultats développés ne reposent pas sur le fait que le modèle soit modifié précisément en cet endroit. Il serait donc facilement envisageable de développer des modèles continus modifiés sur un intervalle arbitraire $[x_1, x_2]$, pouvant aider à modéliser des données dont l'histogramme comporte un pic ou un creux en un endroit donné. Dans le cas discret, une telle approche a été proposée dans Murat et Szynal (1998). Dans le même ordre d'idées, bien que les modèles considérés ici soient des densités définies sur $\mathbb{R}^+$, il serait également possible de travailler avec des densités définies sur $\mathbb{R}$. Il serait intéressant de reprendre la comparaison des différentes méthodes de classification présentée dans le troisième article avec une densité normale modifiée sur $[0, x_0]$ (ou encore sur $[x_1, x_2]$) de façon à déterminer si les différences de performance entre les méthodes considérées dans l'article étaient principalement dues à la modification à zéro, ou à la non normalité de la loi $f_1(x|\theta)$.

RÉFÉRENCES

Gordon, A. (1998). How many clusters ? An investigation of five procedures for detecting nested cluster structure. *In Data science, classification, and related methods : proceedings of the fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan, March 27-30, 1996*, page 109. Springer.

Gupta, P., Gupta, R. et Tripathi, R. (2005). Score test for zero inflated generalized Poisson regression model. *Communications in Statistics-Theory and Methods*, **33**(1):47–64.

Hall, D. (2000). Zero-inflated Poisson and binomial regression with random effects : a case study. *Biometrics*, **56**(4):1030–1039.

Hardy, A. et Andre, P. (1998). An investigation of nine procedures for detecting the structure in a data set. *In Advances in data science and classification : proceedings of the 6th Conference of the International Federation of Classification Societies (IFCS-98), Università" La Sapienza", Rome, 21-24 July, 1998*, page 29. Springer Verlag.

Hasan, M. et Snedonn, G. (2009). Zero-inflated Poisson regression for longitudinal data. *Communications in Statistics-Simulation and Computation*, **38**(3):638–653.

Jansakul, N. et Hinde, J. (2002). Score tests for zero-inflated Poisson models. *Computational statistics & data analysis*, **40**(1):75–96.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**(1):1–14.

Milligan, G. et Cooper, M. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50**(2):159–179.

Murat, M. et Szynal, D. (1998). Non–zero inflated modified power series distributions. *Communications in Statistics-Theory and Methods*, **27**(12):3047–3064.

Ridout, M., Hinde, J. et Demétrio, C. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, **57**(1):219–223.

Singh, S. (1963). A note on inflated Poisson distribution. *Journal of the Indian Statistical Association*, **1**(3):140–144.

Van den Broek, J. (1995). A score test for zero inflation in a Poisson distribution. *Biometrics*, **51**(2):738–743.