Spatial Survival Models for Analysis of Exocytotic Events on Human beta-cells Recorded by TIRF Imaging

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof.ssa Giuliana Cortese

Dottorando/a: Thi Huong Phan

Abstract

Exocytosis on beta-cells is one of the fundamental cellular processes that releases insulin-containing secretory granules to blood through the plasma membrane due to stimulus. Studying survival of granules on the plasma membrane and their spatial correlation within cells during the exocytosis is of great interest to researchers in biological and medical area, as it is closely related to the regulation of insulin level in blood. Data are a collection of TIRF images recorded from 8 human beta-cells, containing granules and syntaxin information. One of the main objectives of this thesis is to investigate the relationship between the survival rates of granules and syntaxin levels, while adjusting for spatial correlation among granules within cells.

To answer our specific biological problem, we propose a semiparametric proportional hazard model, where the baseline hazard function is estimated nonparametrically and a multivariate normal distribution is assumed for individual frailties. Hence, the clustering structure, as well as the spatial correlation between granules are modeled via the variance-covariance matrix of frailties.

We firstly extend the penalized partial likelihood method and the Monte-Carlo EM method to estimate the parameters in the model. Then, we contribute a novel inferential approach based on pairwise likelihood, EM algorithm and quadrature approximation. We conduct simulations to validate and compare three approaches, hence the advantages and disadvantages for each approach are discussed. Finally, we apply our method to the exocytosis data and interpret the results.

Sommario

Nelle cellule beta, l'esocitosi è uno dei processi cellulari fondamentali che rilascia nel sangue granuli secretori contenenti insulina, i quali attraversano la membrana del plasma quando sono sotto stimolo. Lo studio del tempo di vita dei granuli fissati alla membrana, prima del loro distacco, il tasso di esocitosi dei granuli e di altri eventi correlati, e la loro correlazione spaziale all'interno delle cellule, sono aspetti di grande interesse per i ricercatori nel campo biomedico, poichè sono strettamente collegati alle disfunzioni del livello di insulina nel sangue. I dati consistono in un insieme di immagini di tipo TIRF registrate nel tempo su 8 cellule beta umane, le quali contengono molte informazioni sull'andamento e sulla posizione dei granuli, oltre che sui livelli di alcune proteine, come per esempio la sintassina. Uno degli scopi principali della tesi è quello di studiare la relazione tra il tasso degli eventi di scomparsa dei granuli dalla membrana e i livelli della sintassina, tenendo conto della correlazione spaziale tra i granuli all'interno di ciascuna cellula.

Per rispondere al problema biologico sotto studio, nella tesi è stato proposto un nuovo modello semiparametrico, un modello spaziale gerarchico di sopravvivenza ad effetti misti ("frailty") per dati raggruppati in clusters, dove la funzione hazard di riferimento è stimata nonparametricamente ed è assunta una distribuzione multivariata Normale per il vettore degli effetti casuali individuali. La struttura dei clusters e la correlazione spaziale tra le unità statistiche, sono modellati tramite la matrice di varianza e covarianza degli effetti casuali. Inizialmente, la tesi ha esteso il metodo della verosimiglianza parziale penalizzata ed il metodo EM Monte-Carlo, adattandoli all'inferenza per il modello spaziale di sopravvivenza proposto. In seguito, per tale modello, è stato presentato un nuovo approccio inferenziale, il quale si basa sulla verosimiglianza a coppie, l'algoritmo EM e l'approssimazione basata sull'integrazione numerica. Sono stati condotti studi di simulazione per confrontare il comportamento dei tre approcci inferenziali, e sono stati discussi i vantaggi e gli svantaggi di ciascun approccio. Infine, il modello ed i metodi proposti sono stati applicati ai dati sull'esocitosi ed è stata fornita una possibile interpretazione biologica del fenomeno.

Dedication to my family

Acknowledgements

Firstly, I would like to express my deepest gratitude to my supervisor, professor Giuliana Cortese for her patient guidance and continuous inspiration during my doctoral years. "Thank you very much for your friendly companionship that let me confidently share my expressions".

Next, I would like to thank my professor Duong Minh Duc in University of Science, Vietnam, for giving me first bricks of my research career. Specially, I hereby express my gratitude to my professor Tran Viet Chi in University of Lille 1, France, for giving me the first opportunity in my abroad study.

I thank to my colleagues in Department of Statistical Sciences, University of Padova for the great memories during the past three years. Specially, to Luca and Gioia, "I am very thankful for your helps and your cares in the time I was sick".

I also thank my Vietnamese friends in Padova for a lovely small Vietnamese community. Specially, thanks to Kim Hue for taking care me as a sister.

Warm hugs to my best friend Kim Phung and my best teacher Hoang Ha for constantly accompanying with me. Specially, many thanks to Hoang Ha for helping me, encouraging me from the first days I was in Europe. "Small things you did for me, great powers I got from you".

Finally, I would like to give all of my current success as a thankful gift for my family. "You are my endless love and pride".

Padova, October 1, 2018

Contents

List of Figures

List of Tables

Introduction

Overview

Data motivation Insulin is a main anabolic hormone of the body produced by beta cells of the pancreatic islets. It regulates various important metabolisms, especially glucose from the blood into liver, fat and skeletal muscle cells (Berg *et al.*, 2002). After being synthesized, insulins are packaged inside mature granules, called insulin-containing secretory granules, and wait to be released into blood through the plasma membrane due to stimulus (Joslin and Kahn (2005), Creighton (1993)). This releasing procedure is a particular case of a more general cellular process called exocytosis. Particularly, exocytosis of insulin-containing secretory granules on beta cells is triggered within milliseconds by a rise in cytosolic Ca^+ (Gandasi and Barg, 2014). The exocytosis contributes to managing insulin level in blood, that relates to various diseases, specially diabetes (Kahn, 2003). Consequently, there have been many biological studies on this exocytosis mechanism such as Olofsson *et al.* (2002), Knowles *et al.* (2010). Recent publications, Gandasi and Barg (2014) and Barg *et al.* (2010), claimed that clustering of syntaxin stabilizes granules on the plasma membrane and this contributes to increasing the exocytotic rate. However, there is no literature applying survival analysis techniques to study the interaction between syntaxin and single granules during the exocytosis process. Moreover, there is also an increasing interest in understanding if survival times of granules on the plasma membrane are spatially correlated across the cellular regions.

The dataset used in this thesis is part of some experimental results on human beta cells, obtained with the aim to study interaction between syntaxin, one of the most important proteins, and single secretory granules during exocytosis (Gandasi and Barg (2014), Barg *et al.* (2010)). The dataset contains sequences of images produced by high-resolution total-internal reflection microscopy on beta cells, which are typical raw imaging results in biological experiments. Presence of granules on the plasma membrane is followed from beginning of the experiment by the florescence intensity on the images until they disappeared or the experiment ends. Note that syntaxin is observed to be distributed hierarchically in beta cells (Gandasi and Barg (2014), Barg *et al.* (2010)), then this fact may cause spatial correlation of granules inside cells. This characteristic inspired us to introduce a spatial frailty survival model to describe the effects of syntaxin on the survival times of granules. The model also accounts

for independent right censored times, which are considered when granules end the experiment without experiencing any events.

The objective model   Let $\tilde{T}_{ik}$ be the event time for the ith observation in the kth cluster, where $i = 1, \ldots, n_k$, $k = 1, \ldots, m$ and $\sum_{k=1}^{m} n_k = n$. Here m is the number of clusters, n_k is the number of observations in cluster k, and n is the total number of observations. Let C_{ik} be the censoring time, the observed time is then $T_{ik} = \min(\tilde{T}_{ik}, C_{ik})$, and $\delta_{ik} = \mathbf{1}\{\tilde{T}_{ik} \leq C_{ik}\}$ is the event indicator. Let us also consider a vector of explanatory variables X_{ik}. The observed time-to-event data are denoted by $t = \{t_{ik}, \delta_{ik}, X_{ik}, i = 1 \ldots, n_k, \ k = 1, \ldots, m\}$.

The random effect, called frailties, provides a suitable way to introduce unobserved random factors in the model to account for association and unobserved heterogeneity. Given the frailties z_{ik}, the observed event times are typically assumed to be mutually independent with proportional hazard function for time t_{ik} of the ith observation in the kth cluster, written as

$$\lambda(t_{ik}|z_{ik}) = \lambda_0(t_{ik}) \exp(\beta^t X_{ik} + z_{ik}), \tag{1}$$

Here, $\lambda_0(t)$ is the baseline hazard function, and β is the vector of regression coefficients. Following standard spatial approaches, the frailty vector Z is assumed to follow a multivariate normal distribution with the block variance-covariance matrix $\Sigma(\theta_c) = \sigma^2 \times \Omega(d, \theta_\rho)$, where each block in $\Omega(d, \theta_\rho)$ presents a correlation matrix for units in the same cluster, while zeros are outside the blocks, meaning that units belonging to different clusters are independent. The correlation function depends on the unknown parameters θ_ρ and spatial distances d.

This model setting is similar to the well-known shared frailty model (Hougaard, 1995), where a frailty variable is presented for hidden aspects of clusters. However, in our case frailties are introduced for each individual observation to model the spatial variation inside clusters. Therefore, the model also aims at measuring the spatial correlation between observations and considering a clustering structure.

Inference approaches   Multivariate frailty models have been early used for modeling dependence in multivariate time-to-event data (Clayton (1978), Yashin and Iachine (1995)). They are characterized by challenges of high-dimensional integrals in the full likelihood, as well as complications in the maximization procedure. In the context of no spatial correlation between frailties, multivariate frailty models have been studied in many applications, and their inference can be based on two common estimation approaches: Monte-Carlo EM algorithm (MCEM) and penalized partial likelihood (PPL). Ripatti *et al.* (2002), Vaida and Xu (2000), Klein (1992), Guo and Rodriguez (1992) treated the unobserved random effects (frailties) as missing values, thus the EM algorithm is a natural choice. In the E-step, the intractable integrals in the conditional expectations of random effects, given the observed data, are usually approximated by Monte-Carlo methods. Ripatti and Palmgren (2000), Therneau and Grambsch (2013) proposed to use Laplace approximation for approximating the full likelihood

function. For likelihood maximization, they considered frailties as fixed effect parameters and the density of such frailties was considered as a penalty term in the likelihood. On the other hand, the history of spatial study on survival frailty models is relatively short and the majority of these studies are based on Bayesian inference approaches. See, for example, the study of Leukemia survival (Henderson *et al.*, 2002), infant mortality (Banerjee *et al.*, 2003) and breast cancer (Zhou *et al.*, 2015). On the other side, spatial frailty survival models have been rarely studied by likelihood-based inferential approaches. Two examples are the study of Loblolly pine based on the MCEM algorithm (Li *et al.*, 2015) and the East Boston Asthma study based on the Laplace approach (Li and Ryan, 2002). In the thesis, as an alternative choice, we suggest to use pairwise likelihood, i.e., a special case of a more general class of pseudo likelihoods called composite likelihoods (Lindsay, 1988). In this approach, the high-dimensional full likelihood is simplified to several two-dimensional pseudo-likelihoods for the benefit of computational time. In addition, the composite likelihood function surfaces are usually smoother than those of the full likelihood, and therefore, easier to be maximized (Katsikatsou *et al.*, 2012). Consequently, composite likelihood methods can reduce the computation cost substantially and also yield estimators with the desired asymptotic properties of unbiasedness, consistency, and normality (Varin *et al.*, 2011). Although pairwise likelihood approach applied on spatial random effects models have been well developed in context of noncensored data (Katsikatsou *et al.* (2012), Varin *et al.* (2005), Gao and Song (2011)), in the literature there has not been any work on modeling spatial survival data with presence of frailties.

In this thesis, we firstly have investigated how the common approaches MCEM and PPL can be extended on our objective model. Successively, a pairwise likelihood approach, conducted by EM algorithm and quadrature approximation, have been presented. To our knowledge, this study is the first attempt to develop such pairwise likelihood approach for spatial frailty survival models.

Main contributions of the thesis

Starting from raw experimental imaging data, we have described different cellular events of insulin-containing secretary granules during the exocytosis, recorded on human beta cells. We have implemented in Matlab a procedure to transform this imaging data into numerical datasets. We then applied a standard Cox model and compared results to a Gaussian frailty model, where frailties are introduced for each individual granule and they are assumed to be independent and normally distributed. The Matlab process does not only serve for the particular dataset used in thesis, but it is also useful for proceeding other raw biological experimental imaging data.

In order to investigate spatial dependence of granules in their survival rates, we have proposed a hierarchical spatial frailty model. Although the inferential methods, PPL and MCEM,

are both commonly used in frailty survival models, practically they are unable to be directly applied to our model setting. Therefore, a contribution of this thesis has been to extend the R package `coxme` to implement the PPL method with an Exponential variance-covariance matrix for accounting the spatial correlation between frailties. We also computed standard errors estimates for the variance-covariance parameters, which are missing in the `coxme` package. In addition, we have extended the MCEM inferential method Li *et al.* (2015) and the R code to be adapted to our model setting. Moreover, we contributed by proposing a stopping rule to study convergence of the algorithm. Furthermore, sandwich standard errors have been implemented and compared to the standard errors based on the Louis's formula to investigate effects of clustering on parameter estimation.

Another main contribution of the thesis focuses on the quadrature pairwise likelihood method based on EM algorithm (QPLH), that we introduced for the first time for inference in spatial frailty models in presence of independent right censored data. The estimators inherit asymptotic properties of the standard composite likelihood method and the cumulative baseline hazard function has been estimated following the Breslow-type estimator (Breslow, 1974). The quadrature approximation in the E-step, as well as the estimation procedure in the M-step, have been carefully shown in propositions and the mathematical proofs are also provided in the thesis. We have implemented the QPLH algorithm in R and C++ languages. The performance of the three inferential methods presented in the thesis have been investigated and compared in simulation studies. The QPLH approach shows off its stability in modeling spatial survival data because the complex structure of variance-covariance matrix does not complicate the algorithm. Specially, the QPLH method overcomes the problem of poor estimation in cases of few observation per cluster and heavy censored data. Finally, we contributed with an innovative application where the Exocytosis is the first time studied with a spatial correlation perspective. In detail, we have analyzed the hidden spatial correlation between granules by applying the QPLH algorithm to our data. This application explained statistically the biological properties studied by Gandasi and Barg (2014) and Barg *et al.* (2010), contributed to novel explorations and also can broaden further novel studies.

Chapter 1

Background

This chapter contains the main methodology of particle tracking procedure for TIRF images, described in Section 1.1. Two likelihood inference methods for general frailty models based on penalized partial likelihood and Monte-Carlo EM algorithm are, respectively, given in Section 1.2 and Section 1.3.

1.1   Particle tracking

TIRF image data include sequences of images produced by a total internal reflection fluorescence microscope (TIRF-M), which is a particular type of microscope. In this procedure, a thin region of a specimen, usually less than 200 nanometers, will be observed and imaged every few micro-seconds. Particles bounding to surface of the specimen will be detected due to increase of its fluorescence surrounding. Hence, TIRF-M produces high contrast images where bright fields may contain particles information. Due to the fact of sub-micron surface selectivity, TIRF-M has become a favor method of choice for single molecule detection in cell and molecular biology.

TIRF images contain a huge information of objects which may help us to understand underlying biological mechanisms. Particle tracking aims to identify target particles on the images and create their trajectories.

In the following chapters of this thesis, we will present a Matlab code to process TIRF image data where the particle tracking is an important step. In this section, we summarize the main methodology of the particle tracking which is constructed using a famous Matlab code authored by Blair and Dufresne (n.d.). In details, this section summarizes the main steps: Reducing noise, locating particles and creating particle trajectories.

1.1.1   Reducing noise

Fluorescence microscope has been widely used in biological experiments as a fundamental tool for the examination of cellular particles. However, it also produces a high signal-to-noise ratio because of light reflection through the specimen. Therefore, the first step in processing these image data is to reduce noise from the images using spatial filtering tools.

Generally, spatial filtering uses a defined square mask with diameter $2w + 1$, where w is an integer larger than the single sphere's image radius pixels, but smaller than the spacing between particles. The mask size is carefully defined to ensure that it smooths out the noise, but still preserves the features of interest. The mask is placed on the image and moved across all possible pixel position on the image. A filtered image is produced by replacing the intensity value at the center by a linear combination of the intensity values of the center pixel and all neighboring pixels covered by the mask. The filtering can perform many different functions but here, we mainly focus on two main functions that will be used in our Matlab code as well as commonly used for reducing image noise.

The average filter   If the target image is denoted by I, then to each pixel (x, y) of image I is assigned a new value

$$I_w(x, y) = \frac{1}{(2w + 1)^2} \sum_{i,j=-w}^{w} I(x + i, y + j),$$

which is the local average taken on a mask size $2w + 1$. Informally, the average filter consists of simply replacing the value at the center with an average of all grayscale intensity values in the mask.

The Gaussian filter   A Gaussian filter is one that has peak at the center of the mask, and has a Gaussian decay away from center by a variance ε. Each centroid is evaluated by a new value

$$I_\varepsilon(x, y) = \frac{\sum_{i,j=-w}^{w} I(x + i, y + j) \exp\left(-\frac{i^2 + j^2}{4\varepsilon^2}\right)}{\left[\sum_{i=-w}^{w} \exp(-i^2/4\varepsilon^2)\right]^2}.$$

Hence, the Gaussian filter also performs averaging, but performs a weighted average to give more emphasis on pixels near the center of the mask.

1.1.2   Locating particles

Determining granule locations is processed on a well cleared image which may be obtained from the previous step of reducing noise. This work is operated respectively on each image of video sequence following three steps: Identifying candidates, refining centroids, and computing exact centroids.

Identifying particle candidates The particles detection process firstly identifies candidates by using a threshold. Pixels whose brightness is greater than the threshold, are served as candidates of the particle locations. The authors recommended that a rough estimate of the threshold value is about 60% of the maximum grayscale intensity value of all video images.

Refining centroids A mask with diameter $2w + 1$ is again moved across all the locations of candidates to refine locally brightest pixels, which are defined as the target particle positions. The locally brightest pixels are those with the highest intensity values within the moving masks.

Computing exact centroids Note that, in cases one needs to get exact locations for particles in sub-pixel units, Blair and Dufresne (n.d.) proposed an additional refinement for the particle locations where sub-pixels are computed for each particle centroid location. Particularly, denoting centroids in the target image I that are obtained from the previous step by (x_0, y_0), then precise estimates of particle locations are updated by $(x_i, y_i) = (x_0 + \epsilon_x, y_0 + \epsilon_y)$ where sub-pixel corrections ϵ_x and ϵ_y are computed by

$$
\begin{pmatrix} \epsilon_x \\ \epsilon_y \end{pmatrix} = \frac{1}{m_0} \sum_{i^2 + j^2 \leq w^2} \begin{pmatrix} i \\ j \end{pmatrix} I(x_0 + i, y_0 + j)
$$

where the multiplication in the formula is the element-by-element vector multiplication and $m_0 = \sum_{i^2 + j^2 \leq w^2} I(x_0 + i, y_0 + j)$ is the integrated brightness of the particle image.

1.1.3 Creating particle trajectories

After completing the locating particle step, we get the lists of particle locations for each image in the video sequence. Given this information, now we aim to get particle trajectories as functions of time by linking the same particle in adjoining frames. The detection of particle trajectories requires two main restrictions: a maximum distance noted by r and a maximum missing time. The former defines the maximum distance that a particle can move from one frame to next frame. The later accepts that particles can be temporally missing on the focal plane. This has the advantage of generating longer trajectories, which not only avoids noise in the fluorescence microscopy, but also improves the statistics for longer lag times.

The central idea of this procedure is coding particle locations of each image in a binary matrix where the 1s indicate the locations. Then the couple of matrices related to two adjoining frames are consecutively overlapped and, at each particle location, an area with radius r is scanned to identify particle status (i.e., particle moved to new position, particle disappeared or a new particle recorded). If a particle is not found in its adjoining frame, then the algorithm continues to scan in the next frames within the accepted maximum missing time. The procedure is proceeded until the last frame.

1.2   Penalized partial likelihood for multivariate frailty models.

1.2.1   The multivariate frailty model

Let $\tilde{T}_i$ denote the survival time and C_i the censoring time for unit i, $i = 1, \ldots, n$. Define with $T_i = \min\{\tilde{T}_i, C_i\}$ and $\delta_i = I(\tilde{T}_i \leq C_i)$, respectively, the observed survival time and the corresponding event indicator. Ripatti and Palmgren (2000) proposed a general multivariate frailty model using the proportional hazard function

$$\lambda_i(t|Z) = \lambda_0(t) \exp(\beta^t X_i + I_i Z), \tag{1.1}$$

where X_i and I_i are vectors of explanatory variables. Here, the frailty vector Z is assumed to follow a multivariate distribution $p(z; \Sigma)$, with mean 0 and covariance matrix $\Sigma = \Sigma(\theta_c)$, with θ_c denoting a vector of unknown parameters.

Let assume that the censoring is independent and non-informative of Z. The marginal likelihood function of the parameter vector $(\Lambda_0(.), \beta, \theta_c)$ for model (1.1) is

$$
\begin{aligned}
L(\Lambda_0(.), \beta, \theta_c) &= \int \prod_{i=1}^{n} \lambda_i(t|z)^{\delta_i} S_i(t|z) p(z; \Sigma(\theta_c)) dz \\
&= \int \prod_{i=1}^{n} \left[\lambda_0(t) \exp\left(\beta^t X_i + I_i z\right)\right]^{\delta_i} \exp\left[-\Lambda_0(t) \exp\left(\beta^t X_i + I_i z\right)\right] \times p(z, \Sigma(\theta_c)) dz.
\end{aligned}
\tag{1.2}
$$

This is a class of semi-parametric models where $\Lambda_0(.)$ is treated non-parametrically, and $p(z, \Sigma(\theta_c))$ is a multivariate distribution, including as a special case the Gamma-Cox shared frailty model (Cox and Oakes, 1984).

Ripatti and Palmgren (2000) consider the case where the frailty Z follows a multivariate normal distribution. In this case, the formula (1.2) can be rewritten as

$$
\begin{aligned}
L(\Lambda_0(.), \beta, \theta_c) &= \frac{1}{|2\pi\Sigma(\theta_c)|^{1/2}} \int \prod_{i=1}^{n} \left[\lambda_0(t_i) \exp\left(\beta^t X_i + I_i z\right)\right]^{\delta_i} \\
&\quad \times \exp\left[-\Lambda_0(t_i) \exp\left(\beta^t X_i + I_i z\right)\right] \times e^{-\frac{1}{2} z^t \Sigma^{-1}(\theta_c) z} dz.
\end{aligned}
\tag{1.3}
$$

1.2.2   Laplace approximation

Presence of normally distributed frailties leads to complications in the optimization of the likelihood, since the high-dimensional integral in (1.3) does not have a closed form expression. To overcome the intractable integral, Ripatti and Palmgren (2000) proposed to use the second-order Laplace approximation, following the lines of Breslow and Clayton (1993), who applied

this approach to generalized linear mixed models. Breslow and Clayton (1993) claimed that when the likelihood function can be written in the form $c|\Sigma^{-1/2}(\theta_c)| \int e^{-\kappa(z)} dz$, where $\kappa(z)$ is a function of the frailty z, then the approximated marginal log-likelihood becomes

$$l(\lambda_0(.), \beta, \theta_c) \approx -\frac{1}{2} \log |\Sigma(\theta_c)| - \frac{1}{2} \log |\kappa''(\tilde{z})| - \kappa(\tilde{z}). \tag{1.4}$$

Therefore, in our case, we have

$$\kappa(\tilde{z}) = \sum_{i=1}^{n} \left\{ -\delta_i \left[\log (\lambda_0(t_i)) + \beta^t X_i + I_i \tilde{z} \right] + \Lambda_0(t_i) \exp \left(\beta^t X_i + I_i \tilde{z} \right) \right\} + \frac{1}{2} \tilde{z}^t \Sigma(\theta_c)^{-1} \tilde{z},$$

$$\kappa''(\tilde{z}) = \sum_{i=1}^{n} \Lambda_0(t_i) \exp \left(\beta^t X_i + I_i \tilde{z} \right) I_i I_i^t + \Sigma(\theta_c)^{-1}, \tag{1.5}$$

and $\tilde{z}$ denotes the solution to the first partial derivative of $\kappa(z)$ with respect to z, i.e

$$\kappa'(\tilde{z}) = \sum_{i=1}^{n} \left[-\delta_i I_i + \Lambda_0(t_i) \exp \left(\beta^t X_i + I_i \tilde{z} \right) I_i + \Sigma(\theta_c)^{-1} \tilde{z} \right] = 0.$$

Now, the approximated marginal log-likelihood can be fully written as

$$l(\Lambda_0(.), \beta, \theta_c) \approx -\frac{1}{2} \log |\Sigma(\theta_c)| - \frac{1}{2} \log |\sum_{i=1}^{n} \Lambda_0(t_i) \exp \left(\beta^t X_i + I_i \tilde{z} \right) I_i I_i^t + \Sigma(\theta_c)^{-1}|$$

$$+ \sum_{i=1}^{n} \delta_i \left[\log (\lambda_0(t_i)) + \beta^t X_i + I_i \tilde{z} \right] - \Lambda_0(t_i) \exp \left(\beta^t X_i + I_i \tilde{z} \right) - \frac{1}{2} \tilde{z}^t D(\theta)^{-1} \tilde{z} \tag{1.6}$$

1.2.3   Penalized partial likelihood

For estimating parameters β and $\Lambda_0(.)$, θ_c is assumed to be known and the frailty vector z is considered as a parameter of fixed effects. According to Ripatti and Palmgren (2000) the equation (1.6) is again simplified by using an additional approximation where terms $-\frac{1}{2} \log |\Sigma(\theta_c)|$ and $-\frac{1}{2} \log |\kappa''(\tilde{z})|$ which are in the first line of (1.6) are ignored. Hence, the marginal log-likelihood becomes the penalized likelihood (PL) for a Cox model with frailty Z, denoted by

$$PL = \sum_{i=1}^{n} \left\{ \delta_i \left[\log (\lambda_0(t_i)) + \beta^t X_i + I_i z \right] - \Lambda_0(t_i) \exp \left(\beta^t X_i + I_i z \right) \right\} - \frac{1}{2} z^t \Omega(\theta_c)^{-1} z \tag{1.7}$$

where term $-\frac{1}{2} z^t \Sigma(\theta_c)^{-1} z$ is considered as a penalized term for extreme values of Z. For maximization purpose, parameter $\Lambda_0(.)$ is eliminated by profiling out as typically done for Cox models (Johansen, 1983). Let $\mathcal{R}(t_i) = \{j = 1, \ldots, n \text{ such that } t_j \geq t_i\}$ be the risk set at time t_i, then the PL function in (1.7) becomes the PPL function for (β, z) defined as

$$l_1(\beta, z) = \sum_{i=1}^{n} \delta_i \left((\beta^t X_i + I_i z) - \log \sum_{j \in \mathcal{R}(t_i)} \exp\left(\beta^t X_j + I_j z\right) \right) - \frac{1}{2} z^t \Sigma^{-1}(\theta_c) z. \qquad (1.8)$$

For fixed θ_c, a solution $(\hat{\beta}(\theta_c), \hat{z}(\theta_c))$ can be iteratively obtained by solving estimation equations based on the first partial derivatives of the PPL function, as follows

$$\frac{\partial l_1(\beta, z)}{\partial \beta} = \sum_{i=1}^{n} \delta_i \left[X_i - \frac{\sum_{j \in R(t_i)} X_j \exp\left(\beta^t X_j + I_j z\right)}{\sum_{j \in R(t_i)} \exp(\beta^t X_j + I_j z)} \right] = 0 \qquad (1.9)$$

$$\frac{\partial l_1(\beta, z)}{\partial z} = \sum_{i=1}^{n} \delta_i \left[I_i - \frac{\sum_{j \in R(t_i)} I_i \exp\left(\sum_j \beta^t X_j + I_j z\right)}{\sum_{j \in R(t_i)} \exp(\beta^t X_j + I_j z)} \right] - \Sigma(\theta_c)^{-1} z = 0. \qquad (1.10)$$

Note that formulas (5) and (6) in Ripatti and Palmgren (2000) are reported erroneously in their paper since the summation $\sum_{j \in R(t_i)}$ is missing at the numerators.

Replacing the maximized values $(\hat{\beta}(\theta_c), \hat{z}(\theta_c)$ of PPL into equation (1.4), we get an approximated profile likelihood function for θ_c, equal to

$$l_2(\theta_c) = -\frac{1}{2} \log |\Sigma(\theta_c)| - \frac{1}{2} \log |\kappa''(\theta_c)| - \frac{1}{2} \hat{z}^t \Sigma(\theta_c) \hat{z}. \qquad (1.11)$$

Note that function $\kappa(\theta_c)$ have been approximated by minus the PPL function (1.8) at the estimated parameters $(\hat{\beta}, \hat{z})$. In practice, the second derivative function $\kappa''(\theta_c)$ can be conveniently approximated by the Hessian matrix of PPL function obtained from the estimation procedure of parameter (β, z), which is here denoted by K_{22}. Then, the estimating equation of variance components θ_c can be simplified and expressed through differential of the profile likelihood (1.11) as follows

$$-\frac{1}{2} \left[tr\left(\Sigma^{-1}(\theta_c) \frac{\partial \Sigma(\theta_c)}{\partial \theta_c} \right) + tr\left(K_{22}^{-1} \frac{\partial \Sigma^{-1}(\theta_c)}{\partial \theta_c} \right) - \hat{z}^t \Sigma^{-1}(\theta_c) \frac{\partial \Sigma(\theta_c)}{\partial \theta_c} \Sigma^{-1}(\theta_c) \hat{z} \right] = 0, \quad (1.12)$$

where $tr()$ is the trace value of the matrix.

In shared frailty model with i.i.d. frailty terms, θ_c contains only the variance parameter σ^2, then the explicit formulation of the solution to equation (1.12) is

$$\hat{\sigma}^2 = \frac{\hat{z}^t \hat{z} + tr(K_{22}^{-1})}{n}$$

and the asymptotic variance of $\hat{\sigma}^2$ is

$$var(\hat{\sigma}^2) = 2\hat{\sigma}^4 \left[n + \frac{1}{\hat{\sigma}^4} tr\left(K_{22}^{-1} K_{22}^{-1} \right) - \frac{2}{\hat{\sigma}^2} tr\left(K_{22}^{-1} \right) \right]^{-1}.$$

In conclusion, the whole algorithm procedure can be summarized as follows. At the first step, the algorithm is started by initialing vector θ_c. Secondly, $(\hat{\beta}(\theta_c), \hat{z}(\theta_c))$ are found by alternating between solving equations (1.9) and (1.10) for fixed variance components θ_c. Thirdly, at maximized values of β and z, a new value of θ_c is obtained by solving the estimating equation (1.12). Lastly, the second and third step are iterated until convergence. After the algorithm converges, the cumulative baseline hazard can be estimated by using Breslow's estimator (Breslow (1974)),

$$\hat{\Lambda}_0(t) = \sum_{t_l \leq t} \frac{d_l}{\sum_{j \in R(t_i)} \exp\left[\hat{\beta}^t X_j + I_j \hat{z}\right]}$$

where $t_l = t_1, \ldots, t_L$ are L distinct ordered event time points, and d_l is the number of events at time t_l, for $l = 1, \ldots, L$.

1.3 Monte-Carlo EM (MCEM) algorithm for shared frailty models

1.3.1 A shared frailty model

Let $\tilde{T}_{ik}$ be the survival time and C_{ik} the censoring time for observation i in cluster k defined for $i = 1, \ldots, n_k$, $k = 1, \ldots, m$ and $\sum_{k=1}^{m} = n$. Denote with $T_{ik} = \min\{\tilde{T}_{ik}, C_{ik}\}$ and $\delta_{ik} = I(\tilde{T}_{ik} \leq C_{ik})$ as the observed survival time and the corresponding event indicator, respectively. Let also X_{ik} be a vector of exploratory variables. Then, all time-to-event data are identified by $t = \{t_{ik}, \delta_{ik}, X_{ik}, i = 1, \ldots, n_k, k = 1, \ldots, m\}$.

The hazard function of the ith observation in cluster k is modeled by

$$\lambda_{ik}(t) = \lambda_0(t) \exp(\beta^t X_{ik} + z_k), \tag{1.13}$$

where $\lambda_0(t)$ is the baseline hazard, β is the regression coefficient vector, and z_k is the shared random effect for all individuals in cluster k. The frailty vector $Z = (Z_1, Z_2, \ldots, Z_m)$ is assumed to follow a multivariate normal distribution, $Z \sim MVN(0, \Sigma)$, where $\Sigma = \sigma^2 \Omega$. Here Ω is a $m \times m$ correlation matrix and its (k, k') element measures the correlation between frailties z_k and $z_{k'}$. In this setting, the correlation is modeled as $corr(z_k, z'_k) = \rho(d(k, k'), \theta_\rho)$, where $\rho(\cdot)$ is a parametric correlation function of the spatial distance $d(\cdot)$ between two cluster k and k', which depends on parameter θ_ρ. The vector of unknown parameters in Σ is denoted by $\theta_c = (\sigma, \theta_\rho)$, and the vector of unknown parameters in the time-to-event model is denoted by $\theta_t = \{\beta, \Lambda_0(.)\}$. Let $\theta = \{\theta_t, \theta_c\}$ be a general notation for all parameters envolved in the model.

1.3.2   The EM algorithm for parameter estimation

The likelihood function for θ_t, conditional on the data t and random effects z, is

$$L_1(\theta_t|t, z) = \prod_{i,k} \lambda(t_{ik})^{\delta_{ik}} S(t_{ik})$$

$$= \prod_{i,k} \left[\lambda_0(t_{ik}) \exp\left(\beta^t X_{ik} + z_k\right)\right]^{\delta_{ik}} \exp\left[-\Lambda_0(t_{ik}) \exp(\beta^t X_{ik} + z_k)\right].$$

The corresponding log-likelihood function is

$$l_1(\theta_t|t, z) = \sum_{i,k} \delta_{ik} \log\left[\lambda_0(t_{ik})\right] + \sum_{i,k} \delta_{ik}\left(\beta^t X_{ik} + z_k\right) - \sum_{i,k} \Lambda_0(t_{ik}) \exp\left(\beta^t X_{ik} + z_k\right). \quad (1.14)$$

The likelihood and log-likelihood functions for θ_c, conditional on the random effect z, are respectively

$$L_2(\theta_c|z) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}z^t\Sigma^{-1}z\right)$$

and

$$l_2(\theta_c|z) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma|) - \frac{1}{2}z^t\Sigma^{-1}z. \quad (1.15)$$

If we consider the join density of the event times and the random effects, which is equal to

$$f(t, z) = L(\theta|t, z) = L_1(\theta_t|t, z) \times L_2(\theta_c|z),$$

as a function only of the frailty vector z, then $f(t, z)$ is proportional to a function $g(z)$, which in our case has the following expression:

$$g(z) = \exp\left\{\sum_{i,k}\left[\delta_{ik}z_k - \Lambda_0(t_{ik})\exp\left(\beta^t X_{ik} + z_k\right)\right] - \frac{1}{2}z^t\Sigma^{-1}z\right\}.$$

Then, the distribution of the random effect z, conditional on the data t, is

$$f(z|t) = \frac{f(t, z)}{\int f(t, z)dz} = \frac{g(z)}{\int g(z)dz} \quad (1.16)$$

Expectation step (E-step)   Given the estimate $\hat{\theta}^{(d)}$ obtained from the iteration (d), the algorithm requires to evaluate the expectation of the log-likelihood function with respect to the conditional distribution $f(z|t; \hat{\theta}^{(d)})$, that is

$$Q(\theta|\hat{\theta}^{(d)}) = \mathbb{E}_{z|t}^{(d)}\left\{\log[L(\theta|t, z)]\right\} = \int \log\left[L(\theta|t, z)\right] f(z|t; \hat{\theta}^{(d)})dz.$$

By simple factorization, $Q(\theta|\hat{\theta}^{(d)})$ can be written by $Q(\theta|\hat{\theta}^{(d)}) = Q_1(\theta_t|\hat{\theta}^{(d)}) + Q_2(\theta_c|\hat{\theta}^{(d)})$ where

$$Q_1(\theta_t|\hat{\theta}^{(d)}) = \sum_{i,k} \mathbb{E}_{z|t}^{(d)}(z_k)\delta_{ik} - \sum_{i,k} \mathbb{E}_{z|t}^{(d)}[\exp(z_k)]\Lambda_0(t_{ik})\exp(\beta^t X_{ik})$$
$$+ \sum_{i,k} \delta_{ik} \log[\lambda_0(t_{ik})\exp(\beta^t X_{ik})], \tag{1.17}$$

and

$$Q_2(\theta_c|\hat{\theta}^{(d)}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\log(|\Omega|) - \frac{1}{2\sigma^2}tr\left[\Omega^{-1}\mathbb{E}_{z|t}^{(d)}(z^t z)\right]. \tag{1.18}$$

In practice, samples of frailty vector z are generated from the posterior distribution $g(z;\hat{\theta}^{(d)})$ by using the Metropolis-Hastings algorithm within Gibbs sampling. Hence, the conditional expectations in (1.17) and (1.18) are approximated by sample means using such MCMC frailty samples.

Maximization step (M-step) Given MCMC frailty samples in the E-step, the M-step maximizes separately the conditional expected log-likelihoods $Q_1(\theta_t|\hat{\theta}^{(d)})$ and $Q_2(\theta_c|\hat{\theta}^{(d)})$.

Particularly, $Q_1(\theta_t|\hat{\theta}^{(d)})$ is maximized at $\hat{\theta}_t^{(d+1)} = \left\{\hat{\beta}^{(d+1)}, \hat{\Lambda}_0^{(d+1)}(.)\right\}$ in which $\hat{\beta}^{(d+1)}$ is obtained by maximizing the profile likelihood function

$$pl(\beta) = \sum_{i,k} \delta_{ik} \left[\mathbb{E}_{z|t}^{(d)}(z_k) + \beta^t X_{ik} - \log\left(\sum_{i',k'\in\mathcal{R}(t_{ik})} \mathbb{E}_{z|t}^{(d)}[\exp(z_{k'})]\exp(\beta^t X_{i'k'})\right)\right],$$

where $\mathcal{R}(t)$ is the risk set at time t. The cumulative hazard function $\Lambda_0(t)$ is estimated by a step function which is

$$\hat{\Lambda}_0^{(d+1)}(t) = \sum_{t_l \le t} \frac{d_l}{\sum_{i',k'\in\mathcal{R}(t_{ik})} \mathbb{E}_{z|t}^{(d)}[\exp(z_{k'})]\exp(\beta^t X_{i'k'})}.$$

Here t_l, for $l = 1,\ldots,L$, are L distinct event time points and d_l is the number of events at time t_l.

On the other side, by taking the first derivative of $Q_2(\theta_c|\hat{\theta}^{(d)})$ with respect to θ_c, the ML estimates $\hat{\theta}_c^{(d+1)}$ are obtained by computing

$$(\hat{\sigma}^2)^{(d+1)} = \frac{tr\left[\Omega^{-1}\mathbb{E}_{z|t}^{(d)}(z^t z)\right]}{n}$$

and maximizing the profile function

$$pl(\theta_\rho) = -\frac{n}{2}\log\left\{tr\left[\Omega^{-1}\mathbb{E}_{z|t}^{(d)}(z^t z)\right]\right\} - \frac{1}{2}\log(|\Omega|).$$

1.3.3   The observed information matrix

Since the EM algorithm does not directly provide the observed information matrix for the likelihood of the observed data, Louis's formula (Louis, 1982) is a well-known method used to obtain it. The observed information matrix under this formula is given by

$$
I_{\hat{\theta}} = \mathbb{E}_{z|t}\left[-\frac{\partial^2 l(\theta)}{\partial^2 \theta}\bigg|_{\theta=\hat{\theta}}\right] - \mathbb{E}_{z|t}\left[\left(\frac{\partial l(\theta)}{\partial \theta}\right)\left(\frac{\partial l(\theta)}{\partial \theta}\right)^t\bigg|_{\theta=\hat{\theta}}\right],
$$

where $l(\theta) = l_1(\theta_t) + l_2(\theta_c)$. The two contributions $l_1(\theta_t)$ and $l_2(\theta_t)$ are provided, respectively, in (1.14) and (1.15). The information matrix can be estimated by using the MCMC samples at the last iteration of the algorithm. Through discretization of the baseline hazard function $\Lambda_0(t)$, it is possible to estimate the variance of the jumps of this function at the event times, $\lambda_{01}, \ldots, \lambda_{0L}$, jointly with the other parameters in the model. The consistency of such estimators has been tested in Parner *et al.* (1998) and Andersen *et al.* (1997).

Chapter 2

Insulin-containing secretory granules data.

In this chapter we introduce a novel biological dataset. These data contain a vast of unexplored information about Exocytosis on human pancreatic beta-cells compressed in high-resolution total-internal reflection (TIRF) images. In details, we give a data overview in Section 2.1, we propose a Matlab procedure to process these TIRF images to numerical data in Section 2.2, some primary information on the numerical data is summarized in Section 2.3.1 and finally, in Section 2.3.2, we fit a standard Cox model and an independence Gaussian model to investigate presence of spatial random effects which motivates us to introduce spatial frailties in modeling survival rates.

2.1 Data overview

Insulin-containing secretory granules data is a collection of image sequences produced by a high-resolution total-internal reflection microscopy (TIRF-M). Data are a part of experimental results studying exocytosis on human pancreatic beta-cell conducted by biologists Nikhil R. Gandasi and Sebastian Barg, in the Department of Medical Cell Biologist, University of Uppsala, Sweden. Their experimental results were published in both Gandasi and Barg (2014) and Barg *et al.* (2010).

Briefly, exocytosis is a fundamental cellular process that fuses intracellular vesicles with the plasma membrane and is involved in secretion, protein trafficking and membrane repair. In beta cells, exocytosis of insulin-containing secretory granules is triggered within milliseconds by a rise in cytosolic Ca^+ (Gandasi and Barg, 2014).

Between proteins, one of the most important is syntaxin. To test how syntaxin in plasma membrane has influence on exocytosis, the expert biologists imaged several beta-cells where granules are marked by the granule marker neuropeptide-Y (NPY)-mCherry displayed in red

fluorescence and syntaxin clusters are expressed by green fluorescent protein (syx-GFP). Therefore, presence of a granule on the plasma membrane is recorded from the moment it approaches the plasma membrane, by the florescence intensity at its location on the images until it disappears or the experiment ends. Similarly, trajectories of syntaxin concentration levels are computed for each single granules by overlapping consecutive pairs of granule images and syntaxin images.

In details, Gandasi and Barg (2014) concluded that a successful exocytosis process requires three stages: Tethering, Docking, and Exocytotic event (Figure 2.1). Docking is considered as a prerequisite stage: a granule is considered to experience an exocytotic event only if previously it has been docked stably in the plasma membrane. A granule is defined to have an exocytotic event if it is successfully released to blood due to a stimulation by exposing to elevated K^+. By observing from TIRF images, the florescence of exocytotic events will be disappeared suddenly within few micro seconds (usually in $100 - 200$ms). There is another group of docked granules which are also docked successfully on the plasma membrane but they return to the cellular fluid. From TIRF images, these disappearing events are distinguished from exocytotic events because of their gradually loosing (usually in few seconds). These events are labeled by "undocking events". All docked granules, including "exocytotic events" and "undocking events", absorb proteins during the conversion from the tethering stage to the docking stage. On the other hand, granules who return to the intracellular fluid at the tethering stage or before being docked successfully at the docking stage, are labeled by "visitors". Furthermore, "visitors" fail to recruit proteins during their presence at the plasma membrane. Barg *et al.* (2010) mainly discovered that higher syntaxin levels are associated to an increased probability of exocytosis. Gandasi and Barg (2014) favored to focus on the docking stage. They concluded that docked granules recruit the syntaxin immediately at the moment of their contact with the plasma membrane.
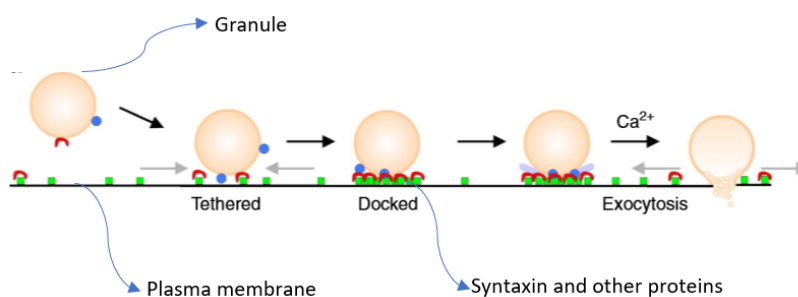
FIGURE 2.1: A successful exocytosis process.

Hence, in studying these secretory granules by TIRF imaging, all events including "visitors", "undocking events" and "exocytotic events", contain a huge unexplored information to investigate exocytosis as well as relative intracellular mechanisms. There has been also interest in studying the more general biological process, where granules disappear from the plasma membrane with different rate over time and the "disappearing event" (either an "undocking

event", an "exocytotic event" or a "vistor") is observed. Therefore, the effect of proteins, such as syntaxin, on the general disappearing event is also of great interest and will be considered in this thesis.

The data shows a wealth of granules and syntaxin information on 8 human pancreatic beta-cells. In details, each cell was imaged consecutively every 0.1 seconds in a certain time. The first row in Table 2.2 shows different experimental time for the 8 cells. Each cell image, called 'frame', is showed by a matrix 256×512 where the left 256×256 matrix is the image of syntaxin channel and the right 256×256 matrix is the image of granule channel. An example of such image is reported in Figure 2.2.
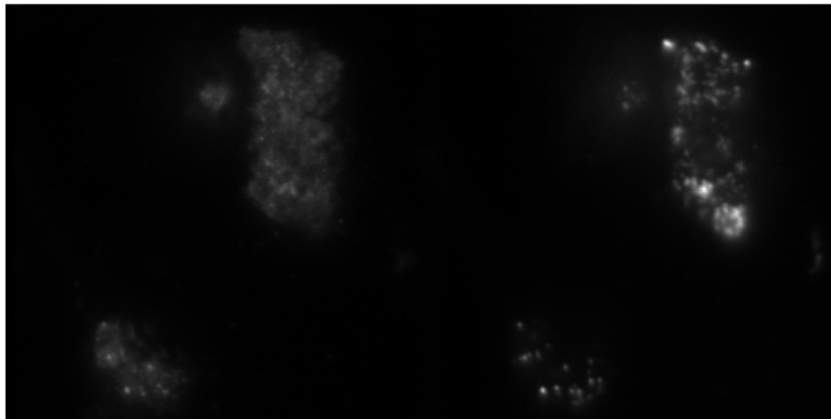
FIGURE 2.2: The first frame of cell 2466. The image size is 256×512 where the left 256×256 image is the syntaxin channel and the right one is the image of granule channel.

2.2 Data processing

To work with the Insulin-containing secretory granules data, firstly images are proceeded to detect granules, following the three main steps of reducing noise, locating granule, and creating granule trajectories. Finally, to complete the data processing, syntaxin levels are computed for each granule. We have implemented in Mathlab all the procedures explained in this section of the thesis, adapting the Matlab code on particle tracking by Blair and Dufresne (n.d.) (see Section 1.1) to our data setting. This processing procedure does not only produce a numerical dataset containing granule locations, survival times, censoring indicators and syntaxin levels which are used for spatial survival analysis in this thesis, but it is also potentially useful for many other kind of datasets in further analysis.

2.2.1 Granule detection

Reducing noise

To detect granules, noise in the images of granule channel are firstly reduced by using the average filter and the Gaussian filter, which have been described in Section 1.1. Generally, a filter moves a mask of size $2w + 1$ across the target image, where w is an integer lager than the single sphere's image radius pixels, but smaller than the spacing between particles. The filtered image is produced by replacing the intensity value at the center of the mask with a linear combination of all intensity values of its neighboring pixels. For convenience, in the computational procedure used later on, we suggest to choose the mask size to be equal to the granule size. In practice, a magnified image of a single granule, as shown in Figure 2.3, suggests us that a possible choice for the granule size is $2w + 1 = 5$. A representative granule image is shown in Figure 2.4 where the filtered image is much clearer.
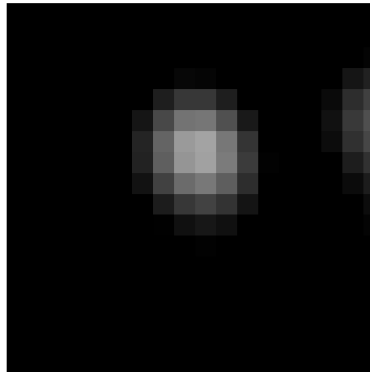
FIGURE 2.3: A magnified image of a single granule in the first frame of cell 2466 recommends a possible granule size to be $2w + 1 = 5$.

Locating granule

From the filtered granule images, the next step aims at identifying centroids of granules on the studied cell area. Applying the theory of particle tracking described in Section 1.1, a locating particle procedure includes three steps: Identifying centroid candidates by thresholds of images, refining centroids by local maxima and computing precise sub-pixel locations. However, sub-pixels are not much meaningful in our study, so we suggest to eliminate the last step about computing exact centroids in particle locating procedure. Hence, on each cell, determining granule locations is operated consecutively in each image of video sequence following two steps: Identifying centroid of granules by using thresholds and local maxima, and secondly,
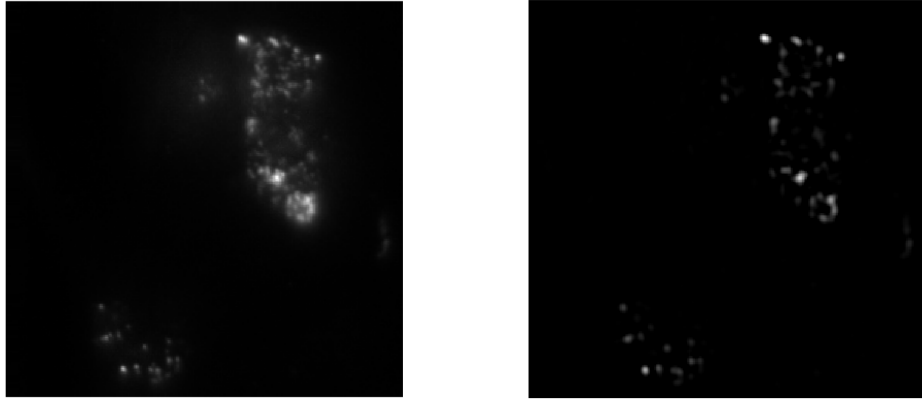
FIGURE 2.4: The granule channel on the first frame of cell 2466 where the original image is shown in the left side and the filtered image is in the right hand.

removing granules who are out of the cell area.

A rough estimate of the threshold value can be determined by 60% of the maximum pixel value in filtered images. However, this is infeasible in our data setting since the contrast quality of the images varied a lot from one image to another, due to the decay of microscopy light intensity during the experiment. Hence, it is essential to define different thresholds for every image in the sequence. In this work, thresholds are automatically computed using the gray scale of Otsu's method (Otsu, 1979). Generally, Otsu's method assumes that the target image contains two classes of pixels following bi-modal histogram (foreground pixels and background pixels), then it calculates the optimum threshold that separates the two classes, in such a way that their combined spread is minimal. The local maximum centroids are still defined on a mask size $2w + 1 = 5$. Figure 2.5 illustrates a representative result of granule detection where granule locations are identified and marked by red circles in Figure 2.5(c).

The correct cell area is found based on the first frame of syntaxin channel. The syntaxin image is firstly transformed into a binary image using the gray scale of Otsu's method (Otsu, 1979). Then connected areas are labeled by using 8-connected criterion that each pixel connects to its neighbors if the neighbors touch one of its edges or corners. The final cell area is defined as the largest connected one. Figure 2.5(a) is a binary image of a representative syntaxin channel in which three connected areas are labeled and the final cell area is shown in Figure 2.5(b). Finally, candidate granules who are out of the cell area are thrown away (Figure 2.5(d)).
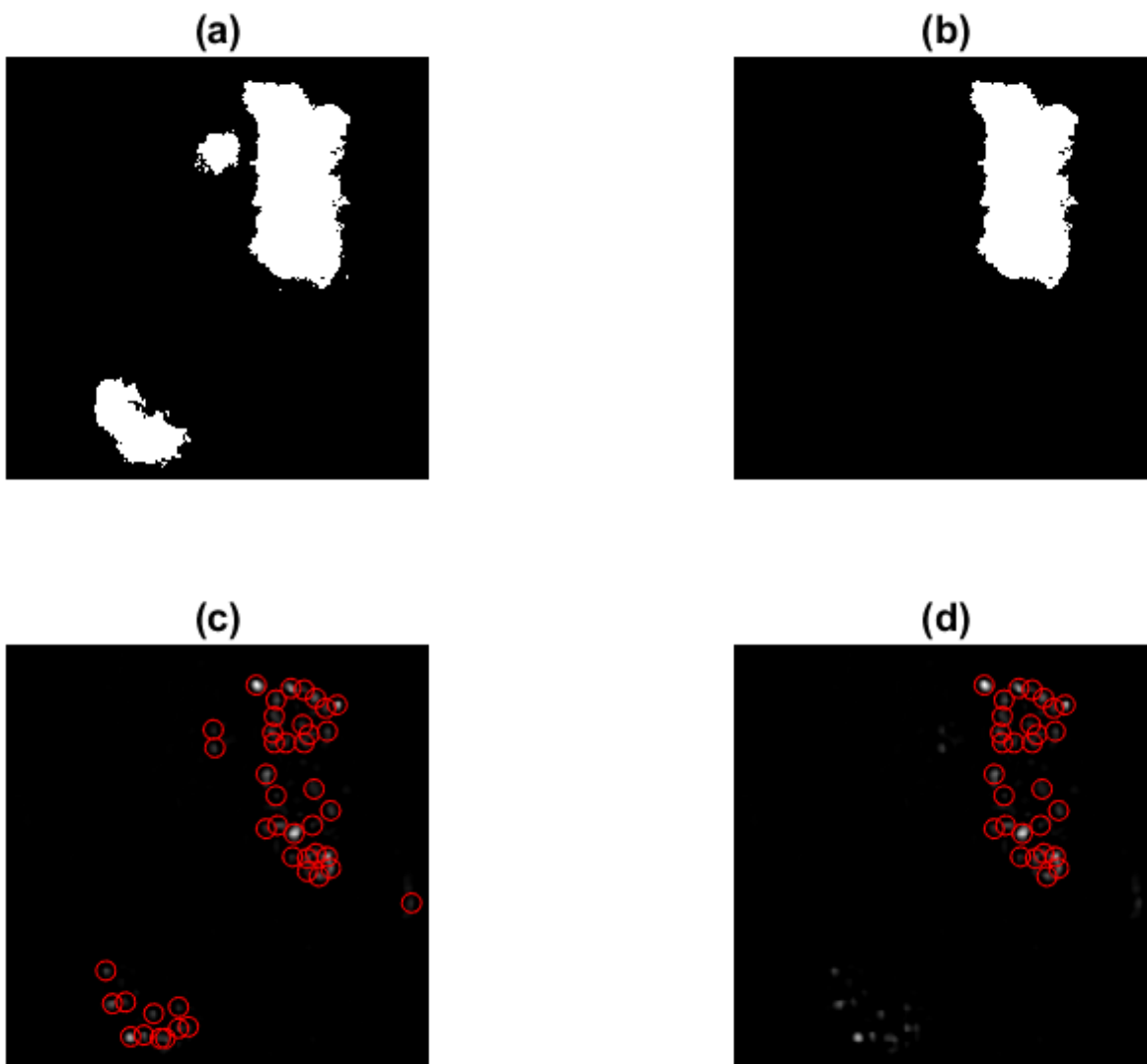
FIGURE 2.5: Detection granules on cell 2466. The entire binary of syntaxin channel (a); the cell area (b); granules are found out on the entire space (c); granules on the cell area (d).

Creating granule trajectories

The final step is to identify the same granule in subsequent frames (Section 1.1). In our application, the detection of granule trajectories is based on two assumptions: A granule can not move further than its diameter (5 pixels) and a granule can be observed to disappear and then appear again in a maximum time of 5 frames. The first assumption is commonly assumed in particle tracking and enables granules to be detected from one frame to another. Moreover, it is also feasible because granules are known not to move in a strong convection (Gandasi and Barg, 2014). The second assumption implies that, if a granule disappears and it is found again in one of the subsequent frames, then it is recorded as the same granule. This procedure

aims at avoiding microscopy noise and at obtaining longer trajectories (see Section 1.1). The second row in Table 2.2 shows the number of trajectories (the total number of granules) in each studied cell. Figure 2.6 gives us a panorama of trajectory length. Unsurprisingly, the shortest trajectories (less than 5 seconds) accounts the highest proportion in every histogram and most of the granules disappeared before 25 seconds.
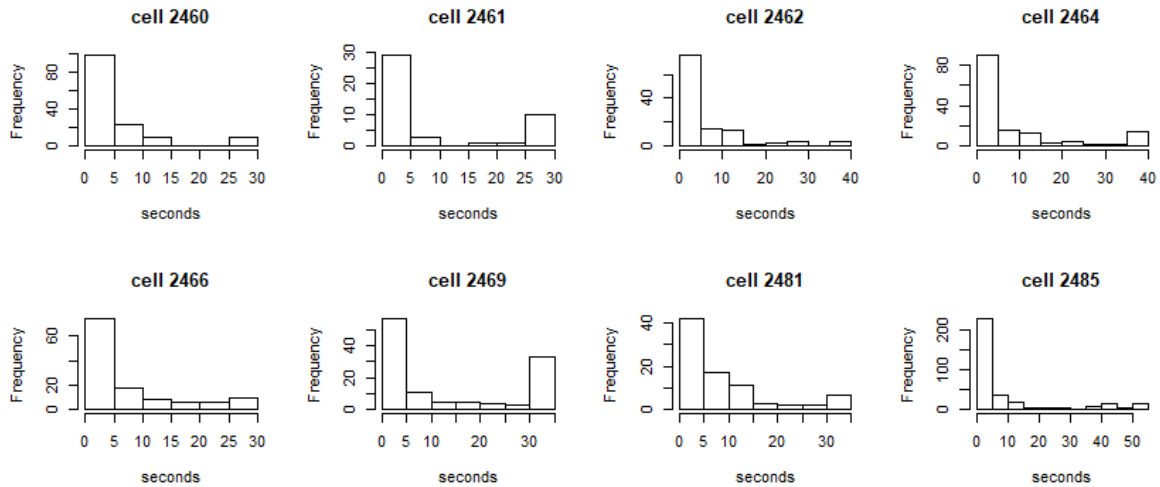
FIGURE 2.6: The histograms of trajectory length (survival time) in seconds for each cell.

2.2.2 Computing syntaxin

To study the concentration of syntaxin on granules in the plasma membrane, we overlap granule and syntaxin frames to quantify syntaxin level around each granule. Barg *et al.* (2010) observed that the most visible syntaxin clusters did not coincide with granule positions. Syntaxin still congregated in clusters, but the clusters were rarely centered on granules (see Figure 2.7). To document this point, the biologists defined on-granule and off-granule clusters and the concentration of syntaxin level at each granule centroid is then computed by the formula $\Delta F/S = \frac{c-a}{a-bg}$ (Gandasi and Barg (2014), Barg *et al.* (2010)). In details, the notation has the following meaning:

- c: The average pixel fluorescence in the granule area, which is a central circle of 5 pixels in our application;

- a: The average pixel fluorescence in a surrounding annulus with an outer diameter of 2 pixels lager than the granule diameter (in our application it is 7 pixels);

- bg: A background fluorescence that is not including the cell area.
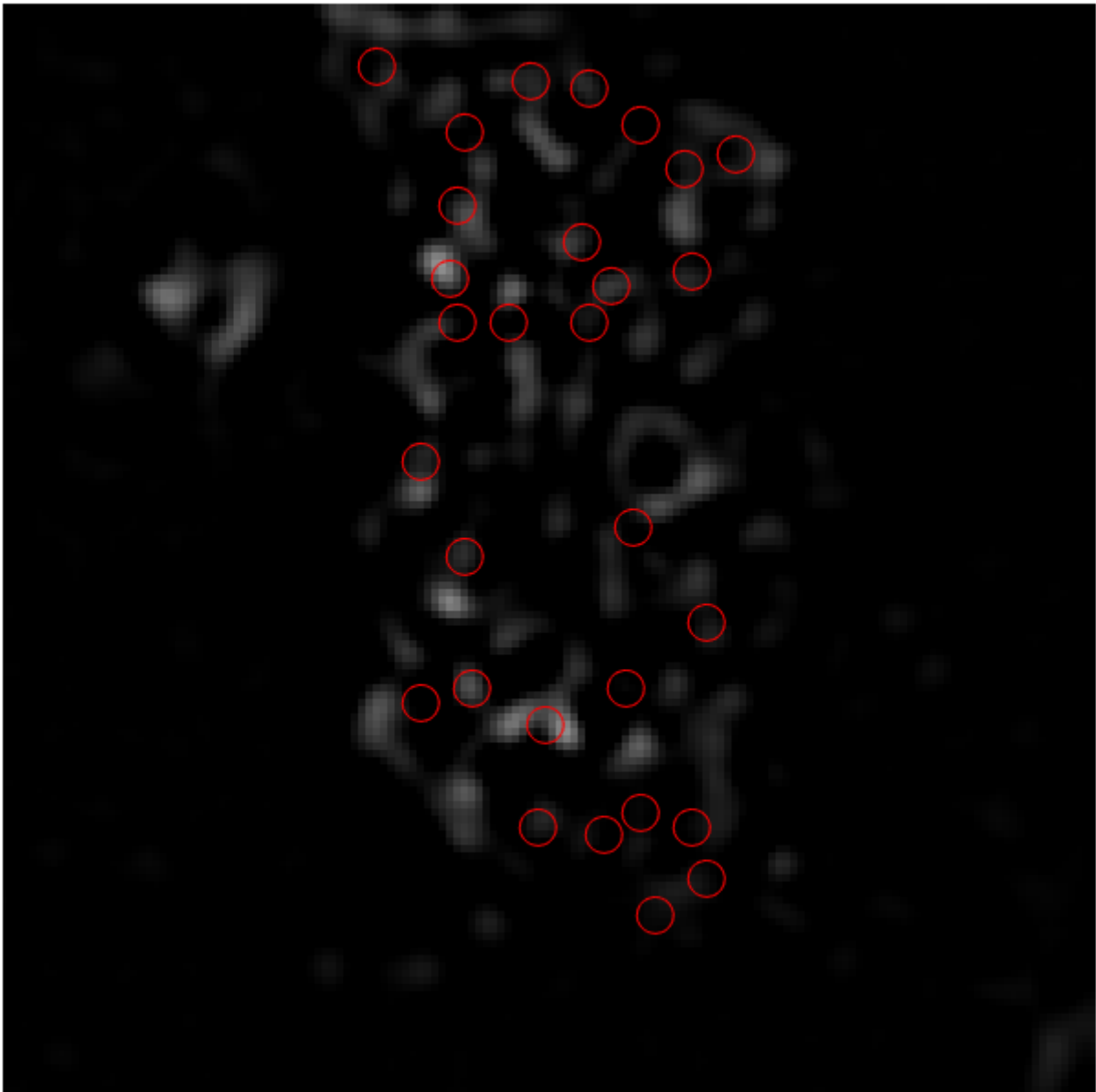
FIGURE 2.7: An enlarged image of syntaxin channel in the first frame of cell 2466 where red circles are granule positions and fluorescence areas are syntaxin clusters. The figure shows that most sysntaxin clusters are not centered at granule positions.

Therefore, $\Delta F = c - a$ quantifies the specific on-granule fluorescence which implies that syntaxin is concentrated at the granule for positive values ΔF, while negative values indicate exclusion. The result $S = a - bg$ is proportional to the concentration of free syntaxin molecules unaffected by any granule. The syntaxin level of each granule is recorded following its trajectory, so it is showed as a function of time. Figure 2.8 shows representative syntaxin trajectories on cell 2466. It appears that syntaxin level varies during the experiment and this pattern is also visible in the other cells.
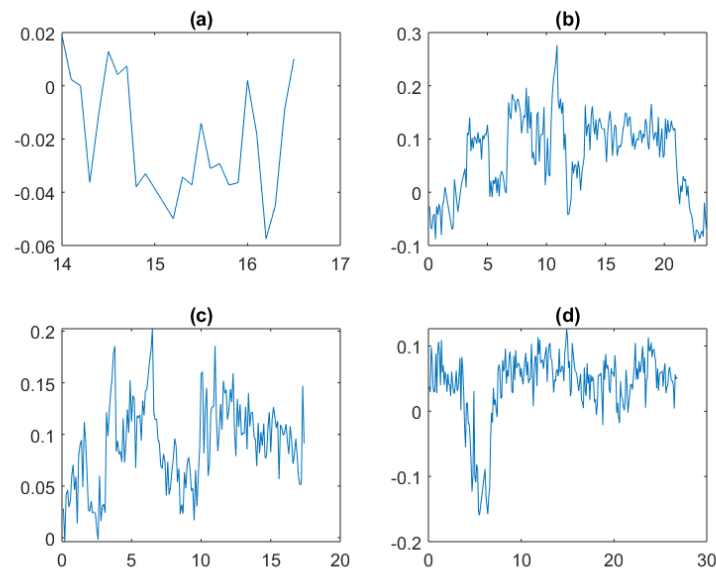
FIGURE 2.8: Four syntaxin trajectories of cell 2466 includes a granule disappeared in a short time (2.6 seconds) (fig.(a)), a granule disappeared with a long survival time (23.6 seconds) (fig.(b)), a granule that ends the experiment with a short survival time (17.4 seconds) (fig.(c)) and a granule that ends the experiment with a full experimental time (26.8 seconds) (fig.(c)).

2.3 Survival data

2.3.1 Primary exploratory analysis

The numerical data is obtained from the data processing and image analysis described in Section 2.2. The data includes 1117 granules clustered in 8 cells. The event of interest is disappearance of granule from the plasma membrane. Survival time is recored from the moment a granule appears on the plasma membrane until it disappears or ends of the experiment. The data includes two explanatory variables, namely **app** and **syn**. The former is a categorical variable indicating whether a granule is present (and visible) at the first frame or not. The **syn** is the average syntaxin level for each granule computed during its survival time. The dataset contains also the spatial coordinates of each granule in pixel unit. Figure 2.9 shows the histogram of the pairwise Euclidean distances of granules within cells. With an average of 91.1% over all cells, pairwise distances are less than 100 pixels and there are very few granules (about 8.35%) who are close in a distance of 10 pixels. Table 2.1 shows a description of the variables in the dataset and Table 2.2 summarizes the primary information classified by cells.
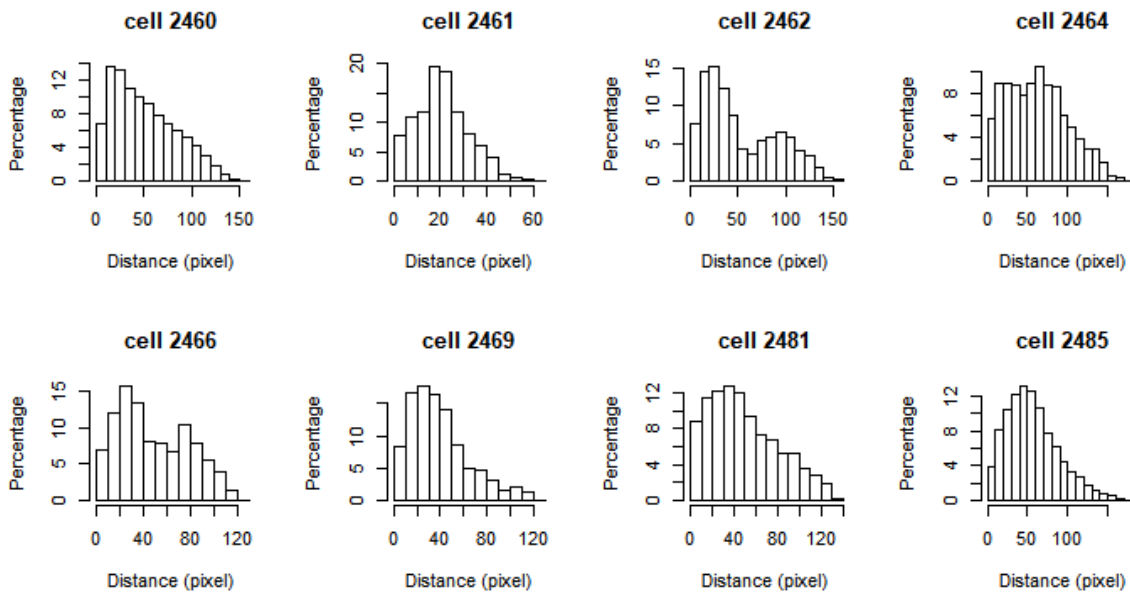
FIGURE 2.9: Histogram of the pairwise Euclidean distances of granules within cells.

Variable	Description
cell_id	A serial number to denote each cell
id	A serial number to denote each granule
x_coord, y_coord	Spatial coordinates of each granule in pixel unit
sur_time	Survival time of each granule in seconds
event	Event indicator (1 = disappear, 0 = censored)
syn	Average synaxin over granule survival time
app	First frame indicator (1 = presence at the first frame, 0 = otherwise)

TABLE 2.1: Description of variable in the dataset.

2.3.2 Initial survival analysis

First, we analyzed data with a standard survival analysis ignoring any spatial variation. Let us define with t_{ik} the observed time for the ith granule in cell k, with δ_{ik} the event/censoring indicator and with X_{ik} the covariate vector. Results from fitting the standard Cox proportional hazards model,

$$\lambda(t_{ik}) = \lambda_0(t_{ik}) \exp(\beta X_{ik})$$

are summarized in the left part of Table 2.3. In this model, we have included the interaction term between syntaxin level and the indicator of granule presence at the first frame. A p-value lower than 0.001 for the regression coefficient of the variable **app** indicates that presence of the granule at the first frame of the experiment is very significant. Moreover, its negative value (-1.167) means that, if a granule is observed from beginning of the experiment, it shows a lower rate of disappearing. The estimate of coefficient for **syn_0**, equal to $= 0.734$, is

	2460	2461	2462	2464	2466	2469	2481	2485
Time (seconds)	25.7	29.2	37.6	35.1	26.8	34.7	32	52
No.granule	143	44	117	146	121	118	84	344
No.event	122	30	103	112	94	73	60	291
E.time(seconds)	3.789	3.227	4.626	4.124	4.35	5.718	6.53	5.136
C.time(seconds)	11.3	21.521	20.564	22.212	15.915	26.267	13.158	32.192
E.syn	0.026	0.147	0.077	0.057	0.038	0.089	0.029	0.022
C.syn	0.029	0. 138	0. 055	0.075	0.035	0. 127	0.028	0. 028

TABLE 2.2: Primary information from the data classified by cells: experimental time in seconds (Time), number of granule (No.granule), number of event (No.event), the average survival time for events and censored granules (E.time & C.time), the average syntaxin level for events and censored granules (E.syn & C.syn).

interpreted as the single effect of syntaxin level on the event rate when the indicator variable **app** is equal to zero. We observe a p-value of 0.13, indicating that the on-granule syntaxin level is not significant in disappearance of those granules that enter later during the experiment. In contrast, the coefficient for **syn_1** is estimated by a value -2.647, and indicates that the syntaxin level decreases the disappearing rate for those granules present at the first frame of the experiment. These conclusions are also visualized in Figure 2.10 where the estimated survival curves are plotted at different syntaxin levels for the two groups of granules being present or absence at the first frame. This figure shows clearly that at any moment, the survival probability of granules present at the first frame of the experiment is always higher than the others. Solid lines of group 0 ($app = 0$) almost coincided and this is coherent with the conclusion of non-influence of syntaxin level on disappearance of these granules. On the other hand, dashed lines in group 1 ($app = 1$) are different and showing higher survival probabilities for increased on-granule syntaxin levels.

We considered also a Cox model containing an individual frailty term z_{ik}, that is

$$\lambda(t_{ik}|z_{ik}) = \lambda_0(t_{ik})\exp(\beta^t X_{ik} + z_{ik}),$$

where frailties are commonly assumed to be independent and follow a Normal distribution $\mathcal{N}(0, \sigma^2)$. Results from fitting this model are reported in Table 2.3 (right side). As expected, the estimated coefficient for **syn_0** has increased and the other coefficients have decreased when the frailty term is included, moreover the standard errors are all increased while p-values are almost the same. In addition, the estimated frailty variance $\hat{\sigma}^2 = 0.506$, as well as the higher log-likelihood (-5456.180 in the standard Cox model and -5270.934 in the independent frailty model), show strong evidence that there exists unexplained variation that should be accounted in modeling this data. Finally, Figure 2.11 shows potential spatial variation within cells. Hence, these results motivate us to introduce spatial frailties in modeling survival rates

	Cox			Cox/Gaussian		
	β	standard error	p.value	β	standard error	p.value
syn_0	0.734	0.468	0.117	0.845	0.552	0.130
app	-1.167	0.117	0.000	-1.532	0.14	0.000
syn_1	-2.647	1.006	0.008	-3.048	1.146	0.008
Log-likelihood		-5456.180		Frailty variance $\hat{\sigma}^2 = 0.506$ -5270.934		

TABLE 2.3: Standard Cox and independence Gaussian survival results. The variables are coded by **syn_0**, **syn_1** are respectively variables modeling syntaxin levels for groups of granules who appear within the experiment and the first frame of the experiment; **app** is an indicator variable that **app** = 1 if granule is present at the first frame of the experiment otherwise **app** = 0.
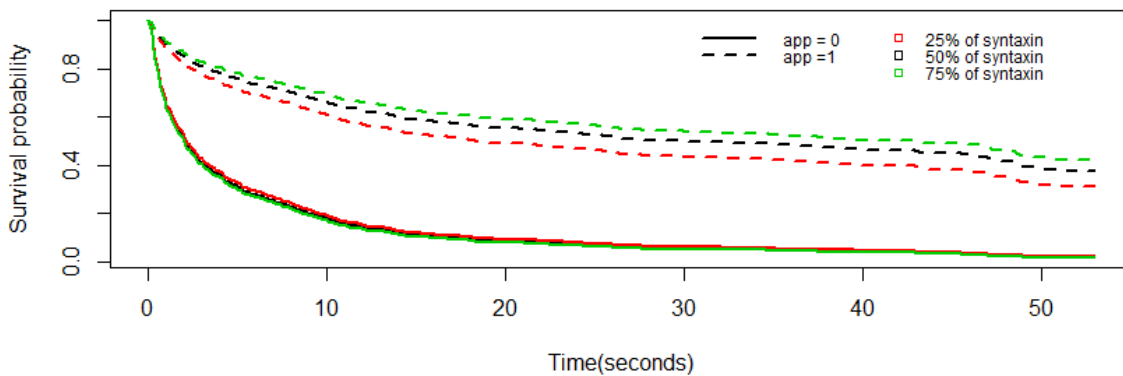
FIGURE 2.10: Estimated survival curves at different level of syntaxin for each group of granule: presence/absence at that first frame.

for these types of events.
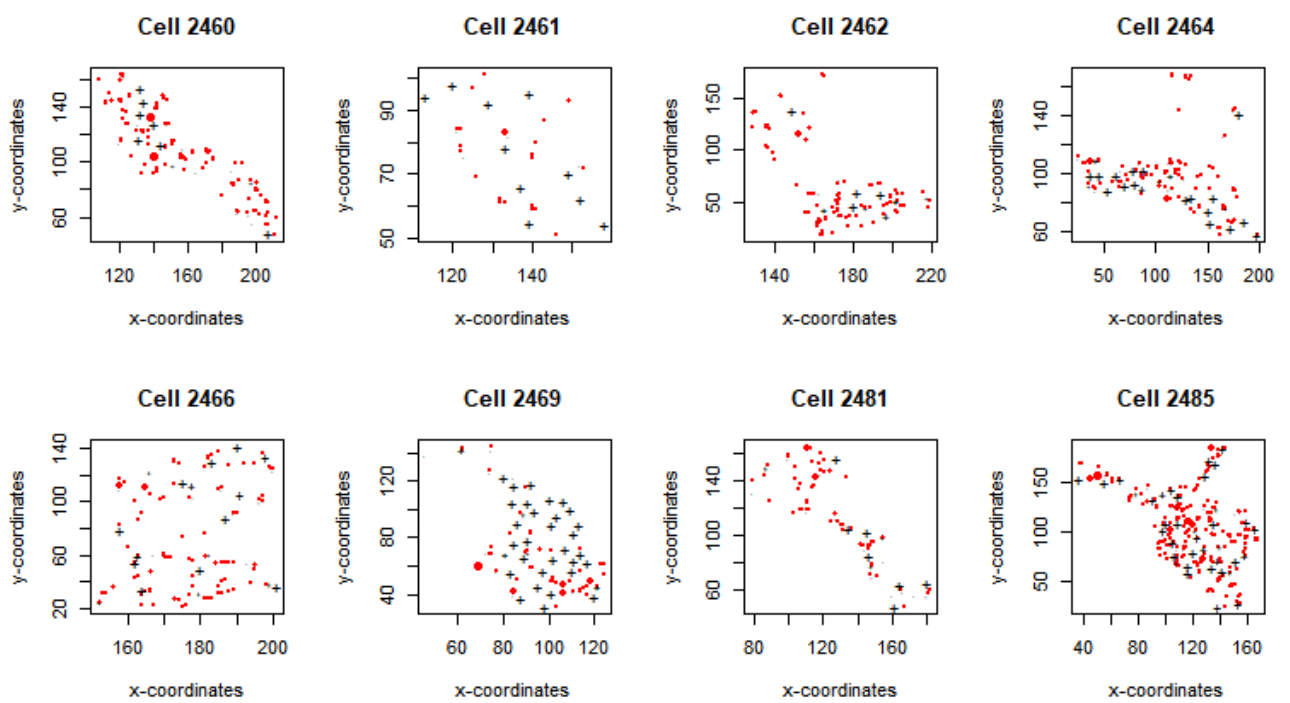
FIGURE 2.11: Survival time of granules in each cell, where black pluses and red dots indicate, respectively, events and censored granules. The size of the dot or plus is proportional to the observed survival time.

Chapter 3

Likelihood-based methods extended to the hierarchical spatial frailty model.

We propose a semiparametric proportional hazard model where frailties are addressed as random effects in the hazard function. The baseline hazard is estimated non-parametrically while frailties are assumed to be clustered and normally distributed. Moreover, the spatial correlation between frailties is incorporated in the variance-covariance matrix. The inference for this proposed hierarchical spatial frailty model is challenged by the computation of high-dimensional integrals in the marginal likelihood. Furthermore, the burden of computation is much considerable since frailties are introduced for individual observations, and the model also accounts for hierarchical structure via independent clusters.

We firstly investigate the penalized partial likelihood (PPL) method that was proposed by Ripatti and Palmgren (2000) for inference in frailty survival models where the frailty term follows a multivariate distribution. The intractable integral in the marginal likelihood term is approximated by Laplace approximation following the lines of Breslow and Clayton (Breslow and Clayton, 1993). Frailties are considered as fixed effect parameters and the baseline hazard function is estimated non-parametrically by the Breslow estimator (Breslow, 1974). The PPL approach for frailty survival models has been implemented in the R package `coxme` (Therneau, 2015). In practice, shared frailty models with i.i.d. frailty terms, where the variance-covariance matrix is diagonal, have been well analyzed and discussed in both Ripatti and Palmgren (2000), and Therneau and Grambsch (2013). However, it is not clear how the method can be extended to allow the presence of spatially correlated frailties, although the methodology is quite general for any kind of semi-parametric frailty model. Moreover, the R package `coxme` does not allow straightforward implementation of frailty survival models for spatially correlated data and no standard error estimates of frailty covariance are given. Hence, we propose to investigate PPL methods applied on the hierarchical spatial frailty model. Practically, parameters

are estimated from `coxme` package where the package is extended to handle an Exponential variance-covariance matrix for spatial frailties. Moreover, we present the standard errors associated to the estimated variance-covariance parameters by numerical derivative methods.

An alternative common approach based on likelihood maximization is based on the Monte-Carlo EM (MCEM) algorithm. The EM algorithm, formalized by Dempster *et al.* (1977), is a well-known procedure for finding maximum likelihood estimates in contexts of incomplete data, such as frailty models. The algorithm is alternated between two steps: finding the expectation of the unobserved part of the data given the observed data (E-step) and maximizing the conditional expectation obtained in the E-step (M-step). In the MCEM algorithm, the conditional expectation in the E-step is approximated by the random effects (frailties) samples which are drawn from the posterior distribution of frailties using Gibbs sampling. In the M-step, given frailties, parameters are conveniently estimated in two separated maximization procedures for the variance-covariance parameters and the time-to-event parameters, including the regression parameters and the baseline hazard function. Particularly, time-to-event parameters can be maximized by using a standard Cox partial likelihood maximization procedure and the variance-covariance parameters can be solved by simple least square tools. The MCEM approach was applied on shared frailty survival models in Ripatti and Palmgren (2000), Vaida and Xu (2000) and Li *et al.* (2015), but only the last one studied the spatial correlation for frailties. However, in Li *et al.* (2015) frailties are introduced as clustered random effects while our proposed hierarchical model requires frailties for each individual observation and also handles the hierarchical structure via independence clusters. Hence, based on the studies of Li *et al.* (2015), we present a MCEM procedure for modeling our proposed hierarchical spatial frailty model. Furthermore, Li *et al.* (2015) proposed stopping the iterations after a fixed amount of time, but for investigating the algorithm convergence speed, we apply a stopping rule using a defined tolerance value. In addition, besides the Louis method, which is the common approach for finding standard errors of estimates in the EM algorithm, we also implemented sandwich standard errors to investigate the influence of clustering as well as the independence assumption in the semi-parametric model.

In details, the hierarchical spatial frailty model is reintroduced in Section 3.1, then the two approaches based on PPL and MCEM are respectively given in Section 3.2 and Section 3.3. Simulation studies will be presented later in the thesis in Chapter 5.

3.1   The hierarchical spatial frailty model

Let $\tilde{T}_{ik}$ be the event time for the ith observation in cluster k, where $i = 1, \ldots, n_k$, $k = 1, \ldots, m$ and $\sum_{k=1}^{m} n_k = n$. Here m is the number of clusters, n_k is the number of observations in cluster k, and n is the total number of observations. Let C_{ik} be the censoring time, the observed time is then $T_{ik} = \min(\tilde{T}_{ik}, C_{ik})$, and $\delta_{ik} = \mathbf{1}\{\tilde{T}_{ik} \leq C_{ik}\}$ is the event indicator. Let us also consider a vector of explanatory variables X_{ik}. The observed time-to-event data are

denoted by $t = \{t_{ik}, \delta_{ik}, X_{ik}, i = 1 \ldots, n_k, \; k = 1, \ldots, m\}$.

The random effect, called frailty, provides a suitable way to introduce unobserved random factors in the model to account for association and unobserved heterogeneity. Given the frailties z_{ik}, the observed event times are typically assumed to be mutually independent with proportional hazard function for time t_{ik} of the ith observation in cluster k, written as

$$\lambda(t_{ik}|z_{ik}) = \lambda_0(t_{ik}) \exp(\beta^t X_{ik} + z_{ik}),. \tag{3.1}$$

Here, $\lambda_0(t)$ is the baseline hazard function, and β is the vector of regression coefficients. The conditional density function is written as a product of the conditional hazard function and the conditional survival function, that is

$$
\begin{aligned}
f(t_{ik}|z_{ik}) &= \lambda(t_{ik}|z_{ik})^{\delta_{ik}} \; S(t_{ik}|z_{ik}) \\
&= \left[\lambda_0(t_{ik}) \exp(\beta^t X_{ik} + z_{ik})\right]^{\delta_{ik}} \; \exp\left[-\Lambda_0(t_{ik}) \exp(\beta^t X_{ik} + z_{ik})\right],
\end{aligned}
$$

where $\Lambda_0(t)$ denotes the cumulative baseline hazard function.

Our approach consists of investigating the spatial variation of observations within clusters. This is translated in studying the spatial association between the underlying random effects, in our case between the individual frailties z_{ik} in cluster k. Following standard spatial approaches, we propose the normal distribution for individual frailties, i.e $z_{ik} \sim \mathcal{N}(0, \sigma^2)$. The spatial structure is addressed by assuming that observations who belong to different clusters are independent, while observations from the same cluster are spatially correlated. We introduce $\rho(d_{ii'}^k; \theta_\rho)$, which is a correlation function containing a parameter θ_ρ that depends solely on the distance $d_{ii'}^k$ between observation i and observation i' in the kth cluster. The frailty vector z is distributed as a multivariate normal, $z \sim \text{MVN}(0, \Sigma)$, where the variance-covariance matrix is $\Sigma = \sigma^2 \Omega$. Here, Ω is a $n \times n$ diagonal block correlation matrix, where correlations are denoted by $\text{corr}(z_{ik}, z_{i'k'}) = 0$ if $k \neq k'$ and $\text{corr}(z_{ik}, z_{i'k'}) = \rho(d_{ii'}^k; \theta_\rho) > 0$ if $k = k'$. The variance parameter σ^2 denotes common variation for all frailties.

Let $\theta_t = (\Lambda_0(.), \beta)$ be the time-to-event parameters and $\theta_c = (\sigma, \theta_\rho)$ be the unknown parameters in the variance-covariance matrix, then, we denote by $\theta = (\theta_t, \theta_c)$ the vector of all model parameters.

Assume that the censoring is independent and non-informative of frailty Z. The marginal log-likelihood function of parameter θ for model (3.1) is

$$
\begin{aligned}
L(\Lambda_0(.), \beta, \theta_c) = \frac{1}{|2\pi\Sigma(\theta_c)|^{1/2}} \int \prod_{i,k} & \left[\lambda_0(t_{ik}) \exp\left(\beta^t X_{ik} + z_{ik}\right)\right]^{\delta_{ik}} \\
& \times \exp\left[-\Lambda_0(t_{ik}) \exp\left(\beta^t X_{ik} + z_{ik}\right)\right] \times e^{-\frac{1}{2}z^t \Sigma^{-1}(\theta_c)z} dz
\end{aligned} \tag{3.2}
$$

The integral in this likelihood function is intractable because its dimension corresponds to the number of individual frailties. Dealing with this intractable integral is a prerequisite step for

the model inference, hence leading to different statistical approaches.

3.2 Penalized partial likelihood (PPL)

In a study of penalized partial likelihood method on estimation of multivariate frailty models, Ripatti and Palmgren (2000) proposed to approximate the likelihood (3.2) by Laplace approximation following the technique in Breslow and Clayton (1993). Hence, an approximated marginal log-likelihood for the likelihood (3.2) is given by

$$l(\Lambda_0(.),\beta,\theta_c) \approx -\frac{1}{2}\log|\Sigma(\theta_c)| - \frac{1}{2}\log|\sum_{i,k}\Lambda_0(t_{ik})\exp\left(\beta^t X_{ik} + \tilde{z}_{ik}\right)I_{ik}I_{ik}^t + \Sigma(\theta_c)^{-1}|$$

$$+ \sum_{i,k}\delta_{ik}\left[\log\left(\lambda_0(t_{ik})\right) + \beta^t X_{ik} + \tilde{z}_{ik}\right] - \Lambda_0(t_{ik})\exp\left(\beta^t X_{ik} + \tilde{z}_{ik}\right) - \frac{1}{2}\tilde{z}^t\Sigma(\theta_c)^{-1}\tilde{z}$$

$$(3.3)$$

Here, the frailties $\tilde{z} = (\tilde{z}_{11}, \tilde{z}_{12}, \ldots, \tilde{z}_{n_m n_m})$ are solution to the first derivative with respect to z of the last two terms in the approximation. In addition, I_{ik} is a binary vector expressed by $I_{ik} = (i_{11}, i_{12}, \ldots, i_{ik}, \ldots, i_{n_m n_m})$ with value 1 at the i_{ik} position and 0 elsewhere.

However, it is still not possible to maximize the approximated log-likelihood (3.3) since it depends on $\tilde{z}$. Ripatti and Palmgren (2000) observed that, if the variance-covariance parameters θ_c were known and the frailties z were considered as fixed effect parameters, then by ignoring the first two terms in the log-likelihood (3.3), the remaining parts would be a penalized log-likelihood function (Green, 1987) for parameters $(\beta, z, \Lambda_0(.))$. Hence, given θ_c, the estimation for the time-to-event parameters θ_t and the frailties z can be simply obtained by using a Cox-partial likelihood maximization routine for (β, z) and the Breslow estimator for the cumulative baseline hazard function $\Lambda_0(.)$ (Klein and Moeschberger, 2006). Here, the penalized partial log-likelihood (PPL) function for (β, z) is given by

$$l_{ppl}(\beta, z) = \sum_{i=1}^{n_k}\sum_{k=1}^{m}\delta_{ik}\left(\left(\beta^t X_{ik} + z_{ik}\right) - \log\sum_{i'k'\in\mathcal{R}(t_{ik})}\exp\left(\beta^t X_{i'k'} + z_{i'k'}\right)\right) - \frac{1}{2}z^t\Sigma^{-1}(\theta_c)z$$

$$(3.4)$$

where $\mathcal{R}(t_{ik}) = \{i', k' : t_{i'k'} \geq t_{ik}\}$ is the risk set at time t_{ik}.

Obviously, by ignoring the first two terms in the log-likelihood (3.3), the maximization is much more simple, but consequently some information may be lost. The characteristics of the information loss will be discussed in the simulation studies in Chapter 5.

Note that the Laplace approximation in (3.3) can be originally written in the following formula

$$l(\Lambda_0(.),\beta,\theta_c) \approx -\frac{1}{2}\log|\Sigma(\theta_c)| - \log|\kappa''(\tilde{z})| - \kappa(\tilde{z}), \tag{3.5}$$

where $\kappa(\tilde{z})$ is equal to minus the last two terms in the approximation (3.3), which has been approximated by the penalized partial log-likelihood in (3.4). Now, if we consider the approximation (3.5) as a function of parameter θ_c given the estimated parameters $(\hat{\beta}, \hat{z}, \hat{\Lambda}_0)$, then (3.5) can be seen as an approximated profile likelihood function for the variance-covariance parameters θ_c, that is

$$l(\theta_c) = -\frac{1}{2}\log|\Sigma(\theta_c)| - \frac{1}{2}\log|\kappa''(\theta_c)| - \frac{1}{2}z^t\Sigma^{-1}(\theta_c)z. \tag{3.6}$$

For computational efficiency, we follow Ripatti and Palmgren (2000) to approximate the second derivative function $\kappa''(\theta_c)$ by minus the second derivative of PPL function (3.4) with respect to z, that has the following analytic formula

$$
\begin{aligned}
K''(\hat{z}, \hat{\beta}, \theta_c) &= -\frac{\partial^2 PPL}{\partial z \partial z^t} \\
&= -\sum_{i=1}^{n_k}\sum_{k=1}^{m} \delta_{ik} \left\{ \left[\sum_{i'k'\in\mathcal{R}(t_{ik})} \exp\left(\hat{\beta}^t X_{i'k'} + \hat{z}_{ik}\right) I_{i'k'} \right] \right. \\
&\quad \times \left[\sum_{i'k'\in R(t_{ik})} \exp\left(\hat{\beta}^t X_{i'k'} + \hat{z}_{ik}\right) I_{i'k'} \right]^t \times \left[\sum_{i'k'\in\mathcal{R}(t_{ik})} \exp\left(\hat{\beta}^t X_{i'k'} + \hat{z}_{i'k'}\right) \right]^{-2} \right\} \\
&\quad + \sum_{i=1}^{n_k}\sum_{k=1}^{m} \delta_i \left[\frac{\sum_{i'k'\in\mathcal{R}(t_{ik})} \exp\left(\hat{\beta}^t X_{i'k'} + \hat{z}_{ik}\right) I_{i'k'} I_{i'k'}^t}{\sum_{i'k'\in\mathcal{R}(t_{ik})} \exp\left(\hat{\beta}^t X_{i'k'} + \hat{z}_{ik}\right)} \right] + \Sigma^{-1}(\theta_c).
\end{aligned}
$$

In practice, this formula is conveniently approximated by the Hessian matrix of PPL function obtained from the estimation procedure of parameter (β, z), which is here denoted by K_{22}. Then, it is straightforward to show that the estimator of the component σ of the variance-covariance parameter vector θ_c has the explicit formula

$$\hat{\sigma}^2 = \frac{\hat{z}^t\Omega(\hat{\theta}_\rho)^{-1}\hat{z} + tr\left[K_{22}^{-1}\Omega(\hat{\theta}_\rho)^{-1}\right]}{n}, \tag{3.7}$$

and the correlation parameters θ_ρ is obtained by solving the following estimating equation

$$tr\left(\Omega^{-1}\frac{\partial\Omega}{\partial\theta_\rho}\right) + \frac{n}{\sigma^2(\theta_\rho)}tr\left(K_{22}^{-1}\frac{\partial\Omega^{-1}}{\partial\theta_\rho}\right) - \frac{1}{\sigma^2(\theta_\rho)}z^t\Omega^{-1}\frac{\partial\Omega}{\partial\theta_\rho}\Omega^{-1}z = 0, \tag{3.8}$$

with $tr()$ being the trace value of the matrix.

The standard error of $\hat{\theta}_c$ is derived from the second derivative of (3.6). The standard error of $\hat{\sigma}^2$ is derived from the following explicit form

$$\frac{\partial^2 l(\theta_c)}{\partial\sigma^2} = -\frac{1}{2}\left\{ -\frac{n}{\sigma^4} + \frac{2}{\sigma^6}tr\left[K_{22}^{-1}\Omega^{-1}(\theta_\rho)\right] - \frac{1}{\sigma^8}tr\left[K_{22}^{-1}\Omega(\theta_\rho)^{-1}K_{22}^{-1}\Omega(\theta_\rho)^{-1}\right] + \frac{2}{\sigma^6}z^t\Omega^{-1}(\theta_\rho)z \right\},$$

whereas computation is complicated for the standard error of $\hat{\theta}_\rho$. Hence, practically these

standard error terms will be obtained by numerical tools.

For more details of the PPL technique applied on multivariate frailty models without accounting the spatial correlation between frailties (Ripatti and Palmgren, 2000), see Section 1.2.

3.3   Monte-Carlo Expectation Maximization (MCEM)

The MCEM algorithm

Following the approach of Li *et al.* (2015), the frailties are now treated as missing values. Hence, instead of maximizing the marginal log-likelihood in (3.2), in the EM algorithm the maximization is focusing on an expectation of the full log-likelihood given the observed data. Moreover, unlike Ripatti and Palmgren (2000), in this approach in order to maximize the likelihood, frailties will be randomly generated rather than being estimated.

The full log-likelihood function for parameter θ is

$$
\begin{aligned}
l(\theta; t, z) = \log(f(t, z; \theta)) &= \log\left(f(t|z; \theta_t)\right) + \log\left(f(z; \theta_c)\right) \\
&= \sum_{i,k} \delta_{ik} \log\left[\lambda_0(t_{ik})\right] + \sum_{i,k} \delta_{ik}\left(\beta^t X_{ik} + z_k\right) - \sum_{i,k} \Lambda_0(t_{ik}) \exp\left(\beta^t X_{ik} + z_k\right) \\
&\quad - \frac{1}{2}\log(|\Sigma|) - \frac{1}{2}z^t \Sigma^{-1} z.
\end{aligned}
\tag{3.9}
$$

Given the value of $\hat{\theta}^{(d)}$ obtained from the iteration (d), the **E-step** involves the evaluation of

$$
Q(\theta|\hat{\theta}^{(d)}) = \int \log\left[f\left(t, z; \theta\right)\right] f(z|t; \hat{\theta}^{(d)}) dz.
$$

Using the factorization in (3.9), $Q(\theta|\theta^{(d)})$ can be written as a sum of the components $Q_1(\theta_t|\theta^{(d)})$ and $Q_2(\theta_c|\theta^{(d)})$, which separately contain information on the time-to-event parameters $\theta_t = (\beta, \Lambda_0(.))$ and the variance-covariance parameters $\theta_c = (\sigma, \theta_\rho)$, respectively. They are given by

$$
\begin{aligned}
Q_1(\theta_t|\hat{\theta}^{(d)}) = \sum_{i,k} \mathbb{E}_{z|t}^{(d)}(z_{ik})\delta_{ik} &- \sum_{i,k} \mathbb{E}_{z|t}^{(d)}[\exp(z_{ik})]\Lambda_0(t_{ik})\exp(\beta^t X_{ik}) \\
&+ \sum_{i,k} \delta_{ik}\log[\lambda_0(t_{ik})\exp(\beta^t X_{ik})]
\end{aligned}
$$

and

$$
Q_2(\theta_c|\hat{\theta}^{(d)}) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\log(|\Omega|) - \frac{1}{2\sigma^2} tr\left[\Omega^{-1}\mathbb{E}_{z|t}^{(d)}(z^t z)\right].
$$

Then, the conditional expectations in $Q_1(\theta_t|\hat{\theta}^{(d)})$ and $Q_2(\theta_c|\hat{\theta}^{(d)})$ are approximated by sample means using MCMC frailty samples drawn from the Gibbs sampling technique with the

posterior distribution $f(z|t; \hat{\theta}^{(d)}) \approx f(t|z, \hat{\theta}_t^{(d)}) f(z|\hat{\theta}_c^{(d)})$. In particular, this distribution is proportional to the function

$$g(z) = \exp \left\{ \sum_{i,k} \left[\delta_{ik} z_{ik} - \hat{\Lambda}_0^{(d)}(t_{ik}) \exp \left((\hat{\beta}^{(d)})^t X_{ik} + z_{ik} \right) \right] - \frac{1}{2} z^t \Sigma^{-1} (\hat{\theta}_c^{(d)}) z \right\}.$$

In the **M-step**, given frailty samples obtained from the E-step, the estimates of the parameter vectors θ_t and θ_c are respectively obtained by maximizing $Q_1(\theta_t|\hat{\theta}^{(d)})$ and $Q_2(\theta_c|\hat{\theta}^{(d)})$. Particularly, $\theta_t = (\beta, \lambda_0(.))$ is conveniently estimated as in standard Cox models using the partial likelihood technique for parameter β and the Breslow estimator for $\Lambda_0(.)$. On the other side, $Q_2(\theta_c|\theta^{(d)})$ is straightforward to be maximized by the least square tool.

For the **convergence criterion**, Li *et al.* (2015) suggested to stop iterations after a specific amount of time and to check convergence efficiency by graphical methods. However, for the aim of investigating the algorithm convergence as well as comparing convergence efficiency between inference approaches, we propose an alternative criterion based on a fixed tolerance. In details, we define the relative difference of the iteration (d) by $\max_i |\hat{\theta}_i^{(d)} - \hat{\theta}_i^{(d-1)}|/|\hat{\theta}_i^{(d-1)}|$, where the maximum is taken over parameter components. The algorithm is defined to be successfully converged if the relative difference is smaller than the defined tolerance in three consecutive iterations. This replication of the criterion in three iterations is proposed to ensure that the algorithm will not stop because of stochastically small changes in parameter estimates, i.e, unlike the standard EM algorithm, the MCEM algorithm does not guarantee monotonic changes in parameters over iterations. In practice, the algorithm will be stopped if it is successfully convergent or it finishes a defined maximum number of iterations. For more details of MCEM technique (Li *et al.*, 2015), see Section 1.3.

Sandwich standard error estimates

The well-known technique for finding the standard error estimates in the EM algorithm is Louis's formula (Louis, 1982) that was given in Section 1.3. However, note that in our proposed hierarchical spatial model, we assumed that observations are independent between clusters and correlated within clusters. While the Louis's formula does not account for the clustering structure and therefore may lead to underestimate standard errors, we propose to use additionally the sandwich standard errors, also referred as the cluster robust standard errors developed by Huber *et al.* (1967) and White (1980). Given the estimated variance-covariance matrix $\hat{V}_L$ obtained by Louis's formula, the sandwich variance-covariance estimate is given by

$$\hat{V}_S = \hat{V}_L{}^t \left(\sum_{k=1}^m u_k u_k^t \right) \hat{V}_L, \tag{3.10}$$

where u_k is the contribution of the kth cluster to the total score function $\frac{\partial l(\theta)}{\partial \theta}$. More specifically, Rogers (1994) noted that, if the log-likelihood is additive at the observation level, then the cluster contribution can be written as

$$u_k = \sum_{i=1}^{n_k} \frac{\partial l_i(\theta)}{\partial \theta},$$

where $l(\theta) = \log\left(f(t, z; \theta)\right)$ is the full log-likelihood function given in (3.9) and $l_i(\theta) = \log\left(f(t_i, z_i; \theta)\right)$ is the log-likelihood of the ith unit. Consequently, we derive the sandwich variance-covariance estimate by the formula below

$$\hat{V}_S = \hat{V}_L{}^t \sum_{k=1}^{m} \left\{ \left(\sum_{i=1}^{n_k} \frac{\partial l_{ik}(\theta)}{\partial \theta} \right) \left(\sum_{i=1}^{n_k} \frac{\partial l_{ik}(\theta)}{\partial \theta} \right)^t \right\} \hat{V}_L,$$

so the standard error terms are computed by taking square root of the diagonal elements of $\hat{V}_S$.

Chapter 4

Quadrature pairwise likelihood

A frailty survival model is a common choice to account for unexplained aspects that helps to describe dependence and heterogeneity of observations. In survival analysis, frailties are addressed as random effects in the proportional hazard function. However, inference in this model requires computation of high dimensional integral in the marginal likelihood. Moreover, the problem becomes more complicated when spatial correlation is introduced between frailties and the baseline hazard function is estimated non-parametrically. Penalized likelihood or Monte-Carlo EM methods may be computationally very slow and also poor in parameter estimation. Alternatively, composite likelihood has been widely used for dealing with estimation in parametric models where the full likelihood approach is computationally intractable (Varin *et al.* (2011), Gao and Song (2011)). There have been many applications of composite likelihood in different statistical areas, including geostatistics, spatial extreme, time series, as well as survival analysis (Varin *et al.*, 2011). However, there is no study on spatial frailty survival models for censored data.

In particular, we propose a pairwise likelihood approach, that is, a composite likelihood based on pairs of observations is adopted, for inference in the spatial frailty survival model. Here, the frailties are assumed to be normally distributed and spatially correlated. For parameter estimation, we proceed with an EM algorithm where a Gauss-Hermite approximation is used in the M-step.

In details, a general theory of pairwise likelihood approach applied on spatial frailty survival models is introduced in Section 4.1. The EM algorithm with Gauss-Hermite approximation is provided in Section 4.2. A procedure to get variance of parameter estimates and its difficulties are discussed in Section 4.3 and finally, proofs of propositions are given in Section 4.4.

4.1   Inference based on the pairwise likelihood

Let $\tilde{T}_{ik}$ be the event time for the observation i^{th} in cluster k where $i = 1, \ldots, n_k$, $k = 1, \ldots, m$ and $\sum_{k=1}^{m} n_k = n$. Let C_{ik} be the censoring time, the observed time is then $T_{ik} = \min(\tilde{T}_{ik}, C_{ik})$,

and $\delta_{ik} = \mathbf{1}\{\tilde{T}_{ik} \leq c_{ik}\}$ is the event indicator. Let us also consider a vector of explanatory variables X_{ik}. The observed time-to-event data are denoted by $t = \{t_{ik}, \delta_{ik}, X_{ik}, i = 1 \ldots, n_k, \ k = 1, \ldots, m\}$. The hierarchical frailty survival model is defined by the conditional hazard function, that is

$$\lambda(t_{ik}|z_{ik}) = \lambda_0(t_{ik}) \exp\left(\beta^t X_{ik} + z_{ik}\right).$$

Here, the frailty vector $Z = (Z_{11}, Z_{12}, \ldots, Z_{n_m n_m})$ is assumed to follow a multivariate normal distribution with the block variance-covariance matrix $\Sigma(\theta_c) = \sigma^2 \times \Omega(d, \theta_\rho)$, where each block in $\Omega(d, \theta_\rho)$ presents a correlation matrix for units in the same cluster, while zeros are outside the blocks, meaning that units belonging to different clusters are independent. The correlation function depends on the unknown parameters θ_ρ and spatial distances d.

For simplicity in writing, we generally denote subscript i, $i = 1, \ldots, n$ for index units. With the observed time-to-event data $t = \{t_i, \delta_i, X_i, i = 1, \ldots, n\}$, the conditional distribution function for unit i at time t_i becomes

$$\begin{aligned}
f(t_i|z_i; \theta) &= \lambda(t_i|z_i)^{\delta_i} S(t_i|z_i) \\
&= \left[\lambda_0(t_i) \exp(\beta^t X_i + z_i)\right]^{\delta_i} \exp\left[-\Lambda_0(t_i) \exp(\beta^t X_i + z_i)\right].
\end{aligned} \tag{4.1}$$

We define again $\theta_t = (\Lambda_0(.), \beta)$ and $\theta_c = (\sigma^2, \theta_\rho)$. Then, the likelihood for $\theta = (\theta_t, \theta_c)$ is

$$L(\theta; t) = \int \cdots \int \prod_{i=1}^{n} f(t_i|z_i; \theta) f(z; \theta_c) dz,$$

which requires the computation of a high-dimensional integral in order to be optimized. As an alternative, we propose to use a pairwise likelihood approach (Lindsay (1988)). The pairwise likelihood is defined by

$$\begin{aligned}
PL(\theta; t) &= \prod_{i=1}^{n} \prod_{j \in \mathcal{A}_i} L(\theta; t_i, t_j) \\
&= \prod_{i=1}^{n} \prod_{j \in \mathcal{A}_i} \int \int f(t_i|z_i; \theta) f(t_j|z_j; \theta) f(z_i, z_j; \theta_c) dz_i dz_j,
\end{aligned} \tag{4.2}$$

where A_i is a subset of all considered pairwise neighbors of unit i indexed by j. For example, in the application of our proposed hierarchical spatial frailty model, we proposed that A_i is equivalent to the cluster containing unit i. This proposal aims to eliminate pairs having no spatial correlation.

The pairwise maximum likelihood estimator (PMLE) $\hat{\theta}$ is obtained by maximizing the pairwise likelihood $PL(\theta; t)$ or equivalently the pairwise log-likelihood

$$pl(\theta; t) = \log(PL(\theta; t)) = \sum_{i=1}^{n} \sum_{j \in \mathcal{A}_i} pl(\theta; t_i, t_j)$$

over the parameter θ, where $pl(\theta; t_i, t_j) = \log(f(t_i, t_j; \theta))$.

For n independent and identically distributed observations, a central limit theorem for the PMLE has been available from Lindsay (1988), Kent (1982) and reviewed by Varin *et al.* (2011). It is claimed that under regularity conditions on the component log-densities, the PMLE $\hat{\theta}$ is asymptotically normally distributed:

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, G^{-1}(\theta)\right),$$

where G is the Godambe information matrix, also referred to as the sandwich information matrix (Godambe, 1960). In particular,

$$G(\theta) = H(\theta)J^{-1}(\theta)H(\theta),$$

where $H(\theta)$ is the sensitivity matrix, $H(\theta) = \mathbb{E}_t\left\{-\nabla^2 pl(\theta; t)\right\}$ and $J(\theta)$ is the variability matrix, $J(\theta) = Var_t\left\{\nabla pl(\theta; t)\right\}$. The expectation and the variance are computed with respect to the unknown density of the observed time-to-event data t.

Other properties of composite likelihood, or pairwise likelihood, are discussed well in Varin *et al.* (2011).

4.2 The Quadrature pairwise EM algorithm

The main subjective is to estimate the parameter vector $\theta = (\Lambda_0, \beta, \sigma, \theta_\rho)$ by maximizing the pairwise likelihood function. The EM algorithm is a natural choice since frailties are unobserved. Moreover, the EM algorithm allows us to estimate the time-to-event parameters $\theta_t = (\Lambda_0, \beta)$ and correlation parameters $\theta_c = (\sigma, \theta_\rho)$ separately. The first use of EM algorithm for pairwise likelihood was proposed by Liang and Yu (2003) in network tomography. The pairwise EM algorithm inherits the main properties of standard EM algorithm for full likelihood (Gao and Song (2011), Varin *et al.* (2005)).

Now, instead of maximizing the pairwise likelihood (4.2) directly, the EM algorithm proceeds by using an initial estimate $\theta^{(0)}$ and solving iteratively the pseudo-complete data problem: $\max_\theta \sum_{i,j} \mathbb{E}_{z_i z_j | t_i t_j} \left\{\log f(t_i, t_j, z_i, z_j; \theta)\right\}$ (Liang and Yu, 2003). In details, starting from an initial value $\theta^{(0)}$ such that $PL(\theta^{(0)}; t) > 0$ and setting the first iteration to $d = 0$, the algorithm alternates an expectation step (E-step) and a maximization step (M-step). The E-step and M-step are described as follows.

The approximate E-step

In the E-step, we aim to evaluate the sum of the conditional expectations with respect to the conditional distribution of $z|t$ given $\theta^{(d)}$:

$$
\begin{aligned}
Q(\theta|\theta^{(d)}) &= \sum_{(i,j)\in\mathcal{A}} \int\int \log\{f(z_i,z_j,t_i,t_j;\theta)\} \times f(z_i,z_j|t_i,t_j;\theta^{(d)})dz_idz_j \\
&= \sum_{(i,j)\in\mathcal{A}} \int\int \log\{f(t_i,t_j|z_i,z_j;\theta_t)\} \times f(z_i,z_j|t_i,t_j,;\theta^{(d)})dz_idz_j \\
&\quad + \sum_{(i,j)\in\mathcal{A}} \int\int \log\{f(z_i,z_j;\theta_c)\} \times f(z_i,z_j|t_i,t_j;\theta^{(d)})dz_idz_j \\
&= Q_1(\theta_t|\theta^{(d)}) + Q_2(\theta_c|\theta^{(d)}),
\end{aligned}
\tag{4.3}
$$

For ease of reading from now on, the notation $\sum_{(i,j)\in\mathcal{A}}$ indicates the double sums $\sum_{i=1}^n \sum_{j\in\mathcal{A}_i}$. The decomposition in (4.3) has the advantage that parameters θ_c and θ_t can be maximized separately. However, double integrals cannot be expressed in close form, and the expectation must be estimated by numerical approximation tools.

Gauss-Hermite quadrature is a form of Gaussian quadrature for approximating the value of integrals of form $\int_{-\infty}^{\infty} e^{-x^2}g(x)dx$. In this case, the integral will be approximated by a weighted sum of function g evaluated at some predetermined quadrature nodes $\int_{-\infty}^{\infty} e^{-x^2}g(x)dx \cong \sum_{i=1}^M w_i g(x_i)$. The nodes x_i are the zeros of the Mth order Hermite polynomial and the w_i are suitably corresponding weights (Liu and Pierce (1994)). Couples (x_i, w_i) can be obtained from tables in Abramowitz and Stegun (1964) or by many R tools, for example the `statmod` package. For double integrals, Jäckel (2005) showed that they can be effectively approximated by $M \times M$ function evaluations using one-dimension Gaussian-Hermite quadrature nodes and weights (x_i, w_i), that is expressed by the summation $\sum_{i,j=1}^M w_i w_j g(x_i, x_j)$. In addition, Varin *et al.* (2005) claimed that the adaptive Gauss-Hermite quadrature could give acceptable accuracy with a low order M. Hence, we propose to use Gauss-Hermite quadrature to approximate conditional expectations in the E-step and the approximation procedure is given below.

Given the estimated standard deviation $\hat{\sigma}^{(d)}$ and the estimated correlation $\hat{\rho}_{ij}^{(d)}$, we define binary nodes $(\tilde{z}^{m_1}, \tilde{z}_{ij}^{m_1 m_2})$ by

$$
\tilde{z}^{m_1} = \hat{\sigma}^{(d)} h_{m_1}
\tag{4.4}
$$

$$
\tilde{z}_{ij}^{m_1 m_2} = \hat{\sigma}^{(d)}(1 - \hat{\rho}_{ij}^{2(d)})^{1/2} h_{m_2} + \hat{\sigma}^{(d)}\hat{\rho}_{ij}^{(d)} h_{m_1}
\tag{4.5}
$$

and weights by

$$
\hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}) = \frac{f\left(t_i|\tilde{z}^{m_1};\theta^{(d)}\right) f\left(t_j|\tilde{z}_{ij}^{m_1 m_2};\theta^{(d)}\right) k_{m_1} k_{m_2}}{\sum_{m_1,m_2} f\left(t_i|\tilde{z}^{m_1};\theta^{(d)}\right) f\left(t_j|\tilde{z}_{ij}^{m_1 m_2};\theta^{(d)}\right) k_{m_1} k_{m_2}}
\tag{4.6}
$$

where couples (h_{m_1}, h_{m_2}) and (k_{m_1}, k_{m_2}) for $m_1, m_2 = 1 \ldots, M$ are respectively Gauss-Hermite quadrature nodes and weights. Then, the Gauss-Hermite quadrature approximation of $Q_1(\theta_t|\theta^{(d)})$ and $Q_2(\theta_c|\theta^{(d)})$ in (4.3) are respectively given in Proposition 4.1.

Proposition 4.1.

For the bivariate nodes $(\tilde{z}^{m_1}, \tilde{z}^{m_1, m_2})$ and weights $\hat{w}_{ij}^{m_1 m_2}(\theta^{(d)})$ given in equations (4.4), (4.5) and (4.6), the final Gaussian-Hermite quadrature approximations for $Q_1(\theta_t|\theta^{(d)})$ and $Q_2(\theta_c|\theta^{(d)})$ are

$$\hat{Q}_1(\theta_t|\theta^{(d)}) = \sum_{(i,j)\in\mathcal{A}} \sum_{m_1,m_2} \log\left\{ f(t_i|\tilde{z}^{m_1};\theta_t) f(t_j|\tilde{z}_{ij}^{m_1 m_2};\theta_t) \right\} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}) \qquad (4.7)$$

and

$$\hat{Q}_2(\theta_c|\theta^{(d)}) = \sum_{(i,j)\in\mathcal{A}} \sum_{m_1,m_2} \log\left\{ f\left(\tilde{z}^{m_1}, \tilde{z}_{ij}^{m_1 m_2}; \theta_c \right) \right\} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}). \qquad (4.8)$$

The proof for this proposition is given in section 4.4.1.

The M-step

In the M-step, we aim to maximize $\hat{Q}_1(\theta_t|\theta^{(d)})$ and $\hat{Q}_2(\theta_c|\theta^{(d)})$ to get the estimators for the next iteration, namely $\hat{\theta}_t^{(d+1)}$ and $\hat{\theta}_c^{(d+1)}$. However, when the baseline hazard function is assumed to be a nonparametric function, then the model functions can not be directly solved by common numerical tools such as Newton-Raphson or Nelder-Mead. By solving $\hat{Q}_1(\theta_t|\theta^{(d)})$ analytically with the profile likelihood method, Proposition 4.2 shows that the maximum likelihood estimator for the baseline hazard function is a step function defined at distinct event time points, where its analytic form is known. The variance-covariance parameters are estimated for step $(d+1)$ by maximizing $\hat{Q}_2(\theta_c|\theta^{(d)})$. The analytic formula for the estimated variance $\hat{\sigma}^2$ and the profile likelihood function for the correlation parameter θ_ρ are given in Proposition 4.3. The proofs are respectively given in sections 4.4.2 and 4.4.3.

Proposition 4.2.

For the bivariate nodes $(\tilde{z}^{m_1}, \tilde{z}^{m_1, m_2})$ and weights $\hat{w}_{ij}^{m_1 m_2}(\theta^{(d)})$ given in equations (4.4), (4.5) and (4.6), the approximated expectation $\hat{Q}_1(\theta_t|\theta^{(d)})$ in (4.7) is maximized at $\hat{\theta}_t = \left(\hat{\beta}^{(d+1)}, \hat{\Lambda}_0^{(d+1)(.)} \right)$, where $\hat{\Lambda}_0^{(d+1)}(.)$ is a step function which has jumps at L distinct event time points, t_l for $l = 1, \ldots, L$, computed as

$$\hat{\Lambda}_0^{(d+1)}(t) = \sum_{t_l \leq t} \frac{w_l}{a_1^l(\hat{\beta}^{(d+1)}) + a_2^l(\hat{\beta}^{(d+1)})} \qquad (4.9)$$

where

$$w_l = \sum_i |\mathcal{A}_i|\delta_i \mathbf{1}_{\{t_i=t_l\}} + \sum_i \sum_{j\in\mathcal{A}_i} \delta_j \mathbf{1}_{\{t_j=t_l\}}$$

$$a_1^l(\beta) = \sum_{i\in\mathcal{R}(t_l)} \sum_j \sum_{m_1,m_2} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}) \exp\left(\beta^t X_i + \tilde{z}^{m_1}\right)$$

$$a_2^l(\beta) = \sum_i \sum_{j\in\mathcal{R}(t_l)\cap\mathcal{A}_i} \sum_{m_1,m_2} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}) \exp\left(\beta^t X_j + \tilde{z}_{ij}^{m_1 m_2}\right),$$

$\mathcal{R}(t)$ is the risk set at time t, $\mathcal{A}_i$ is the neighbor set of unit i and $|\mathcal{A}_i|$ is the number of units in set $\mathcal{A}_i$.

The regression coefficient $\hat{\beta}^{(d+1)}$ is obtained by maximizing the profile likelihood function for β

$$\begin{aligned}
pl(\beta) = &\sum_{l=1}^{L} \left\{ \beta\left(\sum_{j\in D_l} X_j |\mathcal{A}_j|\right) - \log\left(a_1^l(\beta) + a_2^l(\beta)\right) \sum_{j\in D_l} |\mathcal{A}_j| \right\} \\
&+ \sum_i \sum_{l=1}^{L} \left\{ \beta\left(\sum_{j\in D_l\cap A_i} X_j\right) - d_l^i \log\left(a_1^l(\beta) + a_2^l(\beta)\right) \right\},
\end{aligned} \qquad (4.10)$$

where D_l is the set of subjects failing at event time point t_l, d_l^i is the number of subjects failing at event time t_l on subset $\mathcal{A}_i$ and $a_2^l(\beta)$, $a_2^l(\beta)$ are given above.

Proposition 4.3.

For the bivariate nodes $(\tilde{z}^{m_1}, \tilde{z}^{m_1,m_2})$ and weights $\hat{w}_{ij}^{m_1 m_2}(\theta^{(d)})$ given in equations (4.4), (4.5) and (4.6), the approximated expectation $\hat{Q}_2(\theta_c|\theta^{(d)})$ in (4.8) is maximized at $\hat{\theta}_c^{(d+1)} = \left(\hat{\sigma}^{2\ (d+1)}, \hat{\theta}_\rho^{(d+1)}\right)$ where

$$(\hat{\sigma}^2)^{(d+1)} = \frac{1}{|\mathcal{A}|} \sum_{(i,j)\in\mathcal{A}} \sum_{m_1,m_2} \frac{\left[(\tilde{z}^{m_1})^2 + (\tilde{z}_{ij}^{m_1 m_2})^2 - 2\hat{\rho}_{ij}\tilde{z}^{m_1}\tilde{z}_{ij}^{m_1 m_2}\right]}{2\left(1 - \hat{\rho}_{ij}^2\right)} \, w_{ij}^{m_1 m_2}(\theta^{(d)}). \quad (4.11)$$

Here, $|\mathcal{A}|$ is the total pairwise units, and $\hat{\rho}_{ij} = \rho_{ij}(\hat{\theta}_\rho^{(d+1)})$ is the correlation between individuals i and j computed at the estimate $\hat{\theta}_\rho^{(d+1)}$ of θ_ρ.

The correlation parameters $\hat{\theta}_\rho^{(d+1)}$ are obtained by maximizing the profile function

$$pl(\theta_\rho) = \sum_{(i,j)\in\mathcal{A}} \sum_{m_1,m_2} \left\{ -\log((\hat{\sigma}^2(\theta_\rho))^{(d+1)}) - \frac{1}{2}\log\left(1 - \rho_{ij}^2(\theta_\rho)\right) \right\} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}). \quad (4.12)$$

Conclusion

Starting from initialized values $\theta^{(0)}$ such that $PL(\theta^{(0)}; t) > 0$, the algorithm updates the estimates by applying Propositions 4.2 and 4.3. In details, the regression coefficient β is maximized by the profile likelihood (4.10) and an estimator of the baseline hazard function is computed by formula (4.9). Similarly, the MLE of the correlation parameter θ_ρ is produced by maximizing its profile likelihood function (4.12) and substituting this value into equations (4.11) to get the MLE of variance σ^2. The algorithm converges when the relative difference defined by $\max_i |\theta_i^{(d+1)} - \theta_i^{(d)}|/|\theta_i^{(d)}|$ is less than a specific tolerance value.

For the choice of initial values, the random effects are firstly neglected and the regression parameters β and the baseline hazard function are estimated by fitting a standard Cox model. Variance and correlation parameters are guessed, then local maximum should be checked by changing the initial starting values and computing the corresponding results of the algorithm. Different correlation functions should be tested for model validation.

An opportune choice of neighbors for each unit i (defined as $\mathcal{A}_i$) does not only produce precise estimators but also speeds the algorithm convergence. The reason of this is that, between all possible pairs of units, some of them do not give significant contribution to the pairwise likelihood. Nott and Rydén (1999) claimed that only distinct pairs showing significant spatial dependence need to be included in the product. Varin *et al.* (2005) showed that pairs that are far apart and have little spatial correlation will give negligible contribution in the pairwise likelihood, so the authors used a moving window to exclude these pairs. Hence, based on our model setting where we ignore spatial correlation between units that belongs to different clusters, we propose the neighbor sets to be equivalent to the clusters. Consequently, in $\mathcal{A}_i$, all pairs (i, j) where i and j belongs to different clusters are ignored.

4.3   Variance of parameter estimates

As discussed in section 4.1, in order to estimate the standard error of the pairwise likelihood estimators, we need to estimate the Godambe information matrix,

$$
\begin{aligned}
G(\theta) &= H(\theta) J^{-1}(\theta) H(\theta) \\
&= \mathbb{E}_t \left\{ -\nabla^2 pl(\theta; t) \right\} Var_t^{-1} \left\{ \nabla pl(\theta; t) \right\} \mathbb{E}_t \left\{ -\nabla^2 pl(\theta; t) \right\},
\end{aligned}
$$

where the expectation and the variance are taken with respect to the unknown true density of the observed data t.

For $H(\theta)$, we use a common approximation of the negative observed Hessian matrix, which is the matrix of partial second derivatives of pairwise log-likelihood evaluated at the pairwise MLE $\hat{\theta}$ using the original data, i.e, $\hat{H}(\theta) = -\sum_{i,j} \nabla^2 pl(\hat{\theta}; t_i, t_j)$ (Efron and Hinkley, 1978). While the observed variability $\nabla pl(\hat{\theta}; t) \nabla pl(\hat{\theta}, t)^t$ is zero, so it does not approximate the expected variability matrix $J(\theta)$. To approximate $J(\theta)$, we bootstrap the data and calculate an estimated variability matrix $\hat{J}(\theta)$ by

$$
\hat{J} = \frac{1}{K} \sum_{k=1}^{K} \left\{ \sum_{i,j} \nabla pl(\hat{\theta}; t_i^k, t_j^k) \right\} \left\{ \sum_{i,j} \nabla pl(\hat{\theta}; t_i^k, t_j^k) \right\}^t,
$$

where the gradients $\nabla pl(\hat{\theta}; t_i^k, t_j^k)$ are evaluated at the pairwise MLE $\hat{\theta}$ and $t^1, \ldots, t^K$ are K bootstrap data sets.

Hence, to estimate $G(\theta)$, we need the gradient and the Hessian of each bivariate log-likelihood contribution $pl(\hat{\theta}; t_i, t_j)$. These quantities are not directly obtainable because they contain the frailties (z_i, z_j) which are considered as missing values. To solve this problem, we follow the approach of Louis (1982), who proposed a method to compute the observed information matrix in the EM algorithm. Then by straightforward differentiation, for every couple (i, j), we have

$$
\nabla pl(\theta; t_i, t_j) = \mathbb{E}_{z_i, z_j | t_i, t_j} \left\{ \nabla \log \left[f(t_i, t_j, z_i, z_j; \theta) \right] \right\} \tag{4.13}
$$

and

$$
\nabla^2 pl(\theta; t_i, t_j) = \mathbb{E}_{z_i, z_j | t_i, t_j} \left\{ \nabla^2 \log \left[f(t_i, t_j, z_i, z_j; \theta) \right] \right\} + Var_{z_i, z_j | t_i, t_j} \left\{ \nabla \log \left[f(t_i, t_j, z_i, z_j; \theta) \right] \right\}. \tag{4.14}
$$

Moreover, by taking the first derivative of $Q(\theta | \theta^{(d)})$ computed in $\theta = \hat{\theta}$, we have

$$
\nabla Q(\theta | \theta^{(d)}) |_{\theta = \hat{\theta}} = \sum_{i,j} \int \int \nabla \log \left\{ f(t_i, t_j, z_i, z_j; \theta) \right\} |_{\theta = \hat{\theta}} \ f(z_i, z_j | t_i, t_j; \theta^{(d)}) dz_i dz_j.
$$

Hence, by evaluating the above quantity at the last iteration of the EM algorithm, we get the approximation

$$\nabla Q(\theta|\theta^{(d)})|_{\theta=\hat{\theta}} \approx \sum_{i,j} \nabla pl(\theta; t_i, t_j)|_{\theta=\hat{\theta}}.$$

Similarly, by taking the second derivative of $Q(\theta|\theta^{(d)})$ and evaluating it at the last iteration, we get an expression for the first term of equation (4.14)

$$\nabla^2 Q(\theta|\theta^{(d)})|_{\theta=\hat{\theta}} \approx \sum_{i,j} \mathbb{E}_{z_i,z_j|t_i,t_j} \left\{ \nabla^2 \log \left[f(t_i, t_j, z_i, z_j; \theta) \right] \right\}$$

The last term in (4.14) is complicated to be expressed in terms of $Q(\theta|\theta^{(d)})$, so it is not provided in this thesis. Proofs of equations (4.13) and (4.14) are provided in the Appendix A.1.

4.4   Proofs of propositions

4.4.1   Proof of Proposition 4.1

Proof. In this proposition, we aim to approximate $Q_1(\theta_t|\theta^{(d)})$ and $Q_2(\theta_c|\theta^{(d)})$ by using Gaussian-Hermite quadrature approximation. Given parameter $\theta^{(d)}$, we note that the conditional distribution $f(z_i, z_j|t_i, t_j; \theta^{(d)})$ does not depend on θ, so its contribution is only to change the weights of the quadrature approximation. Here, we firstly denote it by $w_{ij}(\theta^{(d)})$, then the decompositions in (4.3) can be rewritten in terms of $w_{ij}(\theta^{(d)})$

$$\begin{aligned} Q_1(\theta_t|\theta^{(d)}) &= \sum_{(i,j)\in\mathcal{A}} \int\int \log\left\{ f(t_i, t_j|z_i, z_j; \theta_t) \right\} \times f(z_i, z_j|t_i, t_j, ; \theta^{(d)}) dz_i dz_j \\ &= \sum_{(i,j)\in\mathcal{A}} \int\int \log\left[f(t_i|z_i; \theta_t) f(t_j|z_j; \theta_t) \right] w_{ij}(\theta^{(d)}) dz_i dz_j \end{aligned}$$

and

$$\begin{aligned} Q_2(\theta_c|\theta^{(d)}) &= \sum_{(i,j)\in\mathcal{A}} \int\int \log\left\{ f(z_i, z_j; \theta_c) \right\} \times f(z_i, z_j|t_i, t_j, ; \theta^{(d)}) dz_i dz_j \\ &= \sum_{(i,j)\in\mathcal{A}} \int\int \log\left\{ f(z_i, z_j; \theta_c) \right\} w_{ij}(\theta^{(d)}) dz_i dz_j. \end{aligned}$$

Since z_i, z_j are generally correlated normal variables with standard deviation $\sigma^{(d)}$ and correlation $\rho_{ij}^{(d)}$, to use Gaussian-Hermite quadrature approximation, we transform these random variables into an independent scale by taking

$$v_i = \frac{z_i}{\sigma^{(d)}}, \qquad v_j = \frac{z_j - \rho_{ij}^{(d)} z_i}{\sigma^{(d)} \left(1 - [\rho_{ij}^{(d)}]^2 \right)^{1/2}}$$

or

$$z_i(v_i) = v_i \sigma^{(d)}, \qquad \text{and} \qquad z_j(v_i, v_j) = \sigma^{(d)} \left(1 - [\rho_{ij}^{(d)}]^2\right)^{1/2} v_j + \sigma^{(d)} \rho_{ij}^{(d)} v_i.$$

Note that

$$w_{ij}(\theta^{(d)}) = f(z_i, z_j | t_i, t_j; \theta^{(d)}) = \frac{f(t_i, t_j | z_i, z_j; \theta^{(d)}) f(z_i, z_j; \theta^{(d)})}{f(t_i, t_j; \theta^{(d)})}$$

$$= \frac{f(t_i | z_i; \theta^{(d)}) f(t_j | z_j; \theta^{(d)}) f(z_i, z_j; \theta^{(d)})}{\int \int f(t_i | z_i; \theta^{(d)}) f(t_j | z_j; \theta^{(d)}) f(z_i, z_j; \theta^{(d)}) dz_i dz_j}.$$

Here $f(z_i, z_j; \theta^{(d)})$ is a two-dimensional normal distribution. Hence, by solving $z_i(v_i)$ and $z_j(v_i, v_j)$ with some simple calculations, the denominator in $w_{ij}(\theta^{(d)})$ becomes

$$\frac{1}{2\pi} \int \int f(t_i | z_i(v_i); \theta^{(d)}) f(t_j | z_j(v_i, v_j); \theta^{(d)}) e^{-v_i^2/2} e^{-v_j^2/2} dv_i dv_j.$$

Then, it can be estimated conveniently by $M \times M$ function evaluations from a one-dimensional Gauss-Hermite quadrature with notes h_m and weights k_m, for $m = 1, \ldots, M$ (Jäckel, 2005), by using a double summation

$$\frac{1}{2\pi} \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} f\left(t_i | \tilde{z}^{m_1}; \theta^{(d)}\right) f\left(t_j | \tilde{z}_{ij}^{m_1 m_2}; \theta^{(d)}\right) k_{m_1} k_{m_2},$$

where frailty pairs (z_i, z_j) are respectively replaced by transformed nodes, $\tilde{z}^{m_1} = z^{(d)}(h_{m_1}) = \hat{\sigma}^{(d)} h_{m_1}$ and $\tilde{z}_{ij}^{m_1 m_2} = z_{ij}^{(d)}(h_{m_1}, h_{m_2}) = \hat{\sigma}^{(d)} (1 - [\hat{\rho}_{ij}^{(d)}]^2)^{1/2} h_{m_2} + \hat{\sigma}^{(d)} \hat{\rho}_{ij}^{(d)} h_{m_1}$.

The same procedure is carried out for the integral in the numerator of $Q_1(\theta_t | \theta^{(d)})$ and $Q_2(\theta_c | \theta^{(d)})$, then we can get the weights

$$\hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}) = \frac{f\left\{t_i | \tilde{z}^{m_1}; \theta^{(d)}\right\} f\left\{t_j | \tilde{z}_{ij}^{m_1 m_2}; \theta^{(d)}\right\} k_{m_1} k_{m_2}}{\sum_{m_1, m_2} f\left\{t_i | \tilde{z}^{m_1}; \theta^{(d)}\right\} f\left\{t_j | \tilde{z}_{ij}^{m_1 m_2}; \theta^{(d)}\right\} k_{m_1} k_{m_2}}.$$

Therefore, the final Gaussian-Hermite quadrature approximations for $Q_1(\theta_t | \theta^{(d)})$ and $Q_2(\theta_c | \theta^{(d)})$ become

$$\hat{Q}_1(\theta_t | \theta^{(d)}) = \sum_{(i,j) \in \mathcal{A}} \sum_{m_1, m_2} \log \left\{ f(t_i | \tilde{z}^{m_1}; \theta_t) f(t_j | \tilde{z}_{ij}^{m_1 m_2}; \theta_t) \right\} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)})$$

and

$$\hat{Q}_2(\theta_c | \theta^{(d)}) = \sum_{(i,j) \in \mathcal{A}} \sum_{m_1, m_2} \log \left\{ f\left(\tilde{z}^{m_1}, \tilde{z}_{ij}^{m_1 m_2}; \theta_c\right) \right\} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}),$$

where $f\left(\tilde{z}^{m_1}, \tilde{z}_{ij}^{m_1 m_2}; \theta_c\right)$ is the bivariate normal distribution with variance-covariance matrix:

$$\Sigma_{ij}(\theta_c) = \sigma^2 \begin{pmatrix} 1 & \rho_{ij}(\theta_\rho) \\ \rho_{ij}(\theta_\rho) & 1 \end{pmatrix}. \qquad \Box$$

4.4.2 Proof of Proposition 4.2

Proof. In this proposition, we aim to maximize $\hat{Q}_1(\beta, \Lambda_0(.)|\theta^{(d)})$. For simplicity, we decompose (4.7) into two parts:

$$
\begin{aligned}
\hat{Q}_1(\beta, \Lambda_0(.)|\theta^{(d)}) &= \sum_{(i,j)\in\mathcal{A}} \sum_{m_1,m_2} \log\left\{ f(t_i|\tilde{z}^{m_1};\theta_t) f(t_j|\tilde{z}_{ij}^{m_1 m_2};\theta_t) \right\} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}) \\
&= \sum_{(i,j)\in\mathcal{A}} \sum_{m_1,m_2} \log\left\{ f(t_i|\tilde{z}^{m_1};\theta_t) \right\} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}) \\
&\quad + \sum_{(i,j)\in\mathcal{A}} \sum_{m_1,m_2} \log\left\{ f(t_j|\tilde{z}_{ij}^{m_1 m_2};\theta_t) \right\} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}) \\
&= A_1(\beta, \Lambda_0(.)|\theta^{(d)}) + A_2(\beta, \Lambda_0(.)|\theta^{(d)}).
\end{aligned}
$$

Using the conditional distribution in (4.1), $A_1(\beta, \Lambda_0(.)|\theta^{(d)})$ becomes

$$
\begin{aligned}
A_1(\beta, \Lambda_0(.)|\theta^{(d)}) &= \sum_{(i,j)\in\mathcal{A}} \sum_{m_1,m_2} \left\{ \delta_i \log[\lambda_0(t_i)] + \delta_i\left(\beta^t X_i + \tilde{z}^{m_1}\right) \right\} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}) \\
&\quad - \sum_{(i,j)\in\mathcal{A}} \sum_{m_1,m_2} \left\{ \Lambda_0(t_i) \exp\left(\beta^t X_i + \tilde{z}^{m_1}\right) \right\} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)})
\end{aligned}
$$

Let $V = \{i : \delta_i = 1\}$ be the set of all observed event time points. Since $\sum_{m_1 m_2} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}) = 1, \forall(i,j)$, then $A_1(\beta, \Lambda_0(.)|\theta^{(d)})$ and $A_2(\beta, \Lambda_0(.)|\theta^{(d)})$ can be rewritten as follows:

$$
\begin{aligned}
A_1(\beta, \Lambda_0(.)|\theta^{(d)}) &= \sum_{i\in V} \sum_{j\in\mathcal{A}_i} \log(\lambda_0(t_i)) + \sum_{i\in V} \sum_{j\in\mathcal{A}_i} \sum_{m_1,m_2} \left(\beta^t X_i + \tilde{z}^{m_1}\right) \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}) \\
&\quad - \sum_{(i,j)\in\mathcal{A}} \sum_{m_1,m_2} \left\{ \Lambda_0(t_i) \exp\left(\beta^t X_i + \tilde{z}^{m_1}\right) \right\} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}) \qquad (4.15)
\end{aligned}
$$

$$
\begin{aligned}
A_2(\beta, \Lambda_0(.)|\theta^{(d)}) &= \sum_{i} \sum_{j\in\mathcal{A}_i\cap V} \log(\lambda_0(t_j)) + \sum_{i} \sum_{j\in\mathcal{A}_i\cap V} \sum_{m_1,m_2} \left(\beta^t X_j + \tilde{z}_{ij}^{m_1 m_2}\right) \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}) \\
&\quad - \sum_{(i,j)\in\mathcal{A}} \sum_{m_1,m_2} \left\{ \Lambda_0(t_j) \exp\left(\beta^t X_j + \tilde{z}_{ij}^{m_1 m_2}\right) \right\} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}). \qquad (4.16)
\end{aligned}
$$

Now, using the profile likelihood approach, we fix β and maximize $\hat{Q}_1(\beta, \Lambda_0(.)|\theta^{(d)})$, considered as a function of $\Lambda_0(.)$ only. Note that all weights $\hat{w}_{ij}^{m_1 m_2}(\theta^{(d)})$ are positive for all pairs (i,j) and (m_1, m_2), therefore this function is maximized when $\lambda_0(t) = 0$ except for times at which an event occurs.

Let

- t_l, $l = 1, \ldots, L$ be L distinct event time points,

- D_l be the set of subjects failing at event time point t_l and $|D_l|$ be the number of subjects in D_l,

- $|\mathcal{A}_j|$ be the number of subjects in neighbor set $\mathcal{A}_j$,

- $\mathcal{R}(t)$ be the risk set at time t

Denote by $\lambda_{0l} = \lambda_0(t_l)$ the baseline hazard function computed at event time point t_l. Then the cumulative baseline hazard function can be written as $\Lambda_0(t) = \sum_{l=1}^{L} \lambda_{0l}|D_l|\mathbf{1}_{\{t_l \leq t\}}$. Rewriting terms A_1, A_2 as functions of the parameter vector $(\lambda_{01}, \ldots, \lambda_{0L})$, one gets

$$
\begin{aligned}
A_1(\lambda_{01}, \ldots, \lambda_{0L}; \beta) &= \sum_{l=1}^{L} \log(\lambda_{0l}) \sum_{j \in D_l} |\mathcal{A}_j| \\
&\quad - \sum_{(i,j) \in \mathcal{A}} \sum_{m_1, m_2} \left(\sum_{l=1}^{L} \lambda_{0l}|D_l|\mathbf{1}_{\{t_l \leq t_i\}} \right) \exp\left(\beta^t X_i + \tilde{z}^{m_1}\right) \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}) \\
&= \sum_{l=1}^{L} \log(\lambda_{0l}) \sum_{j \in D_l} |\mathcal{A}_j| \\
&\quad - \sum_{l=1}^{L} \lambda_{0l}|D_l| \sum_{i \in \mathcal{R}(t_l)} \sum_{j \in \mathcal{A}_i} \sum_{m_1, m_2} \exp\left(\beta^t X_i + \tilde{z}^{m_1}\right) \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}). \quad (4.17)
\end{aligned}
$$

Denote by $d_l^i = |D_l \cap \mathcal{A}_i|$ the number of subjects belonging to the subset $\mathcal{A}_i$ and failing at event time t_l. Then we can write

$$
\begin{aligned}
A_2(\lambda_{01}, \ldots, \lambda_{0L}; \beta) &= \sum_i \sum_{l}^{L} \log(\lambda_{0l}) d_l^i \\
&\quad - \sum_{l=1}^{L} \lambda_{0l}|D_l| \sum_i \sum_{j \in \mathcal{R}(t_l) \cap \mathcal{A}_i} \sum_{m_1, m_2} \exp\left(\beta^t X_j + \tilde{z}_{ij}^{m_1 m_2}\right) \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}). \quad (4.18)
\end{aligned}
$$

At fixed event times t_l, for $l = 1, \ldots, L$, the partial derivatives with respects to λ_{0l} are

$$
\frac{\partial A_1(\lambda_{01}, \ldots, \lambda_{0L}; \beta)}{\partial \lambda_{0l}} = \sum_{j \in D_l} |\mathcal{A}_j| \frac{1}{\lambda_{0l}} - |D_l| \sum_{i \in \mathcal{R}(t_l)} \sum_j \sum_{m_1, m_2} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}) \exp\left(\beta^t X_i + \tilde{z}^{m_1}\right)
$$

$$(4.19)$$

$$
\frac{\partial A_2(\lambda_{01}, \ldots, \lambda_{0L}; \beta)}{\partial \lambda_{0l}} = \sum_i d_l^i \frac{1}{\lambda_{0l}} - |D_l| \sum_i \sum_{j \in \mathcal{R}(t_l) \cap \mathcal{A}_i} \sum_{m_1, m_2} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}) \exp\left(\beta^t X_j + \tilde{z}_{ij}^{m_1 m_2}\right)
$$

$$(4.20)$$

By summing equations (4.19) and (4.20), we get

$$\frac{\partial \hat{Q}_1(\lambda_{01}, \dots, \lambda_{0L} | \theta^{(d)})}{\partial \lambda_{0l}} = \frac{w_l}{\lambda_{0l}} - |D_l| \left(a_1^l(\beta) + a_2^l(\beta) \right),$$

where

$$w_l = \sum_i |\mathcal{A}_i| \delta_i \mathbf{1}_{\{t_i = t_l\}} + \sum_i \sum_{j \in \mathcal{A}_i} \delta_j \mathbf{1}_{\{t_j = t_l\}}$$

$$a_1^l(\beta) = \sum_{i \in \mathcal{R}(t_l)} \sum_j \sum_{m_1, m_2} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}) \exp\left(\beta^t X_i + \tilde{z}^{m_1} \right)$$

$$a_2^l(\beta) = \sum_i \sum_{j \in \mathcal{R}(t_l) \cap \mathcal{A}_i} \sum_{m_1, m_2} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}) \exp\left(\beta^t X_j + \tilde{z}_{ij}^{m_1 m_2} \right),$$

By setting $\frac{\partial \hat{Q}_1(\lambda_{01}, \dots, \lambda_{0L} | \theta^{(d)})}{\partial \lambda_{0l}} = 0$, for all $l = 1, \dots, L$, we get the estimators of $\lambda_{01}, \dots, \lambda_{0L}$ as a function of β

$$\hat{\lambda}_{0l}(\beta) = \frac{w_l}{|D_l| \left(a_1^l(\beta) + a_2^l(\beta) \right)},$$

and the cumulative baseline hazard at time t is

$$\hat{\Lambda}_0(\beta) = \sum_{t_l \leq t} \frac{w_l}{a_1^l(\beta) + a_2^l(\beta)}. \tag{4.21}$$

The profile likelihood function for β is

$$\hat{Q}(\beta | \theta^{(d)}) = A_1(\beta | \theta^{(d)}) + A_2(\beta | \theta^{(d)}),$$

which is obtained by substituting function $\hat{\lambda}_{0l}(\beta) = \frac{w_l}{|D_l|(a_1^l(\beta) + a_2^l(\beta))}$ into $A_1(\beta, \lambda_0(.) | \theta^{(d)})$ and $A_2(\beta, \lambda_0(.) | \theta^{(d)})$. Using (4.15) and (4.17), and omitting terms that do not contain β, $A_1(\beta | \theta^{(d)})$ is written as a function of β by

$$A_1(\beta | \theta^{(d)}) = \sum_{l=1}^L \left\{ \beta \left(\sum_{j \in D_l} X_j |\mathcal{A}_j| \right) - \log\left(a_1^l(\beta) + a_2^l(\beta) \right) \sum_{j \in D_l} |\mathcal{A}_j| \right\} - W_1$$

where

$$W_1 = \sum_{l=1}^L \frac{w_l}{a_1^l(\beta) + a_2^l(\beta)} \sum_{i \in \mathcal{R}(t_l)} \sum_{j \in \mathcal{A}_i} \sum_{m_1, m_2} \left\{ \exp\left(\beta^t X_i + \tilde{z}^{m_1} \right) \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}) \right\}$$

$$= \sum_{l=1}^L \frac{w_l}{a_1^l(\beta) + a_2^l(\beta)} a_1^l(\beta)$$

Similarly,

$$A_2(\beta|\theta^{(d)}) = \sum_i \sum_{l=1}^{L} \left\{ \beta \left(\sum_{j \in D_l \cap A_i} X_j \right) - d_l^i \log \left(a_1^l(\beta) + a_2^l(\beta) \right) \right\} - W_2$$

where

$$W_2 = \sum_{l=1}^{L} \frac{w_l}{a_1^l(\beta) + a_2^l(\beta)} a_2^l(\beta)$$

Since $W_1 + W_2 = \sum_{l=1}^{L} w_l$ is a constant, it can be disregarded for optimization purpose. Therefore, sum of $A_1(\beta|\theta^{(d)})$ and $A_2(\beta|\theta^{(d)})$ produces the profile likelihood function of β, that is

$$pl(\beta) = \sum_{l=1}^{L} \left\{ \beta \left(\sum_{j \in D_l} X_j |\mathcal{A}_j| \right) - \log \left(a_1^l(\beta) + a_2^l(\beta) \right) \sum_{j \in D_l} |\mathcal{A}_j| \right\}$$
$$+ \sum_i \sum_{l=1}^{L} \left\{ \beta \left(\sum_{j \in D_l \cap A_i} X_j \right) - d_l^i \log \left(a_1^l(\beta) + a_2^l(\beta) \right) \right\},$$

given in Proposition 4.2.

The estimators are obtained by firstly maximizing the profile likelihood function to get $\hat{\beta}^{(d+1)}$ and then the estimator of the cumulative hazard function in (4.21) is computed at $\hat{\beta}^{(d+1)}$. This has proved Proposition 4.2.   □

4.4.3   Proof of Proposition 4.3

Proof. This proposition aims at maximizing $\hat{Q}_2(\theta_c|\theta^{(d)})$. Firstly, rewriting $\hat{Q}_2(\theta_c|\theta^{(d)})$, where $f(\tilde{z}^{m_1}, \tilde{z}_{ij}^{m_1 m_2}; \theta_c)$ is the density function of a bivariate normal, we have

$$\hat{Q}_2(\theta_c|\theta^{(d)}) = \sum_{(i,j) \in \mathcal{A}} \sum_{m_1, m_2} \left\{ -\log(2\pi) - \log(\sigma^2) - \frac{1}{2} \log \left(1 - \rho_{ij}^2(\theta_\rho) \right) \right\} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)})$$
$$- \sum_{(i,j) \in \mathcal{A}} \sum_{m_1, m_2} \left\{ \frac{1}{2\sigma^2(1 - \rho_{ij}^2(\theta_\rho))} \left[(\tilde{z}^{m_1})^2 + (\tilde{z}_{ij}^{m_1 m_2})^2 - 2\rho_{ij}(\theta_\rho)\tilde{z}^{m_1} \tilde{z}_{ij}^{m_1 m_2} \right] \right\} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)})$$

The first derivative of $\hat{Q}_2(\theta_c|\theta^{(d)})$ with respect to σ^2 is

$$\frac{\partial \hat{Q}_2(\theta_c|\theta^{(d)})}{\partial \sigma^2} = \sum_{(i,j) \in \mathcal{A}} \sum_{m_1, m_2} \left\{ -\frac{1}{\sigma^2} + \frac{\left[(\tilde{z}^{m_1})^2 + (\tilde{z}_{ij}^{m_1 m_2})^2 - 2\rho_{ij}(\theta_\rho)\tilde{z}^{m_1} \tilde{z}_{ij}^{m_1 m_2} \right]}{2\sigma^4(1 - \rho_{ij}^2(\theta_\rho))} \right\} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)})$$

Evaluating equation $\frac{\partial \hat{Q}_2(\theta_c|\theta^{(d)})}{\partial \sigma^2} = 0$ at the estimate $\hat{\theta}_\rho^{(d+1)}$, we get the estimator $(\hat{\sigma}^2)^{(d+1)}$ of σ^2 shown in the expression (4.11) of Proposition 4.3, that is

$$(\hat{\sigma}^2)^{(d+1)} = \frac{1}{|\mathcal{A}|} \sum_{(i,j)\in\mathcal{A}} \sum_{m_1,m_2} \frac{\left[(\tilde{z}^{m_1})^2 + (\tilde{z}_{ij}^{m_1 m_2})^2 - 2\hat{\rho}_{ij}\tilde{z}^{m_1}\tilde{z}_{ij}^{m_1 m_2}\right]}{2\left(1 - \hat{\rho}_{ij}^2\right)} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}), \qquad (4.22)$$

where $|\mathcal{A}|$ is the total number of pairwise neighbours and $\hat{\rho}_{ij} = \rho_{ij}(\hat{\theta}_\rho^{(d+1)})$ is the correlation between individuals i and j computed at the estimate $\hat{\theta}_\rho^{(d+1)}$.

If we consider the expression (4.22) as a function of θ_ρ and substitute it into $\hat{Q}_2(\theta_c|\theta^{(d)})$, we get

$$\hat{Q}_2(\theta_\rho|\theta^{(d)}) = \sum_{(i,j)\in\mathcal{A}} \sum_{m_1,m_2} \left\{-\log(2\pi) - \log((\hat{\sigma}^2(\theta_\rho))^{(d+1)}) - \frac{1}{2}\log\left(1 - \rho_{ij}^2(\theta_\rho)\right)\right\} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)}) - B$$

where

$$B = \sum_{(i,j)\in\mathcal{A}} \sum_{m_1,m_2} \left\{\frac{\left[(\tilde{z}^{m_1})^2 + (\tilde{z}_{ij}^{m_1 m_2})^2 - 2\rho_{ij}(\theta_\rho)\tilde{z}^{m_1}\tilde{z}_{ij}^{m_1 m_2}\right]}{2(\hat{\sigma}^2(\theta_\rho))^{(d+1)}(1 - \rho_{ij}^2(\theta_\rho))}\right\} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)})$$

$$= \frac{1}{(\hat{\sigma}^2(\theta_\rho))^{(d+1)}} \sum_{(i,j)\in\mathcal{A}} \sum_{m_1,m_2} \left\{\frac{\left[(\tilde{z}^{m_1})^2 + (\tilde{z}_{ij}^{m_1 m_2})^2 - 2\rho_{ij}(\theta_\rho)\tilde{z}^{m_1}\tilde{z}_{ij}^{m_1 m_2}\right]}{2(1 - \rho_{ij}^2(\theta_\rho))}\right\} \hat{w}_{ij}^{m_1 m_2}(\theta^{(d)})$$

$$= |\mathcal{A}|$$

is a constant. Therefore, we get the profile likelihood function for θ_ρ which is shown in equation (4.12) of Proposition 4.3.

□

Chapter 5

Simulation studies

In this chapter, we present Monte Carlo simulation studies for the proposed hierarchical spatial frailty model presented in Section 3.1. The general procedure used for generating the simulated datasets, the settings and the assumptions for simulations are described in Section 5.1. The performance of inferential theory based on QPLH, which has been presented in Chapter 4, is studied by simulations and reported in Section 5.2. The performance of the MCEM and PPL methods, whose theory has been shown in Chapter 3, is discussed and compared to QPLH approach in Section 5.3. Finally, discussions and conclusions are given in Section 5.4.

5.1   Simulated datasets

In this section, we present MCMC datasets simulated from the hierarchical spatial model defined in Section 3.1 to test how our methodologies perform. We simulate survival times from the proportional hazard model

$$\lambda(t_{ik}|z_{ik}) = \lambda_0(t_{ik}) \exp(\beta^t X_{ik} + z_{ik})$$

for the ith individual in cluster k. Survival times T_{ik}, given z_{ik}, are assumed to follow a Weibull distribution with scale parameter $\lambda \exp(\beta^t X_{ik} + z_{ik})$ and shape parameter ρ. Then, the hazard and cumulative hazard functions are, respectively, $\lambda_0(t) = \lambda \rho t^{\rho-1}$ and $\Lambda_0(t) = \lambda t^\rho$. To simplify the setting but still gain insights, we consider the case where we have one explanatory variable X. For the spatial random effects, we assume that the vector z is generated from a multivariate normal distribution $MVN(0, \Sigma)$, with the variance-covariance matrix $\Sigma = \sigma^2 \Omega$, where σ^2 is the common variance and Ω describes the spatial correlation within clusters. For the units i and i' in cluster k, we assume an Exponential correlation function, $\rho(d_{ii'}^k; \theta_\rho) = \exp\left(-d_{ii'}^k/\nu\right)$, where $d_{ii'}^k$ is the Euclidean distance between them.

Following lines of Bender *et al.* (2005) who presented a general technique to generate survival times for simulation studies regarding Cox proportional hazards models, survival time T_{ik} can be obtained by transforming a uniform random variable, $U_{ik} \sim U[0, 1]$, via a cumulative hazard function as follow

$$T_{ik} = \Lambda_0^{-1} \left[\frac{-\log(U_{ik})}{\exp\left(\beta^t X_{ik} + Z_{ik}\right)} \right].$$

Censored times c_{ik} are independently generated from a uniform distribution $U[0, c_{max}]$ where c_{max} will be chosen to designate censoring rates, approximately 30%, 50% and 70%. The event indicator for observation i in cluster k is obtained by calculating $\delta_{ik} = I(t_{ik} \leq c_{ik})$.

All parameter values are chosen to mimic the real survival dataset. In details, we choose $\lambda = 14$, $\rho = 2$ for the Weibull hazard function. The exploratory variable X is generated from a uniform distribution, $X \sim U[-1, 1]$ with the regression coefficient $\beta = 0.6$. We set $\sigma = 0.5$, and the frailties are assumed to be correlated by an Exponential correlation function with parameter $\nu = 10$. In the implementation, the distances are rescaled by a factor 10, to make them numerically more stable. Thus the results will be obtained on $\tilde{\nu} = \nu/10$ instead of the original values.

Clusters are studied in regions of 36×36 squares where 3 and 5 clusters are considered, respectively. Spatial locations in each cluster are assumed to follow an equally spaced squared grid of points. We also investigate different cluster sizes by 7×7 and 9×9 units (see Figure 5.1). Thus, four settings are considered in the simulation studies.
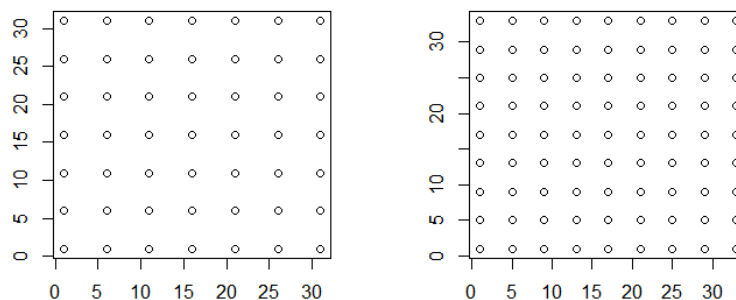
FIGURE 5.1: Illustration of site locations in a cluster of 49 units and a cluster of 81 units.

5.2   Results from quadrature pairwise likelihood method

In this section, we show results of QPLH method applied on 1000 simulated datasets. The theory has been shown in Chapter 4 and the coding is implemented in R and $C++$ languages. For

what concerns the quadrature approximation in the M-step, note that the algorithm works well also with a low number of nodes (Varin *et al.*, 2005). Then, for the practical setting, we decided to use $M = 7$ nodes after investigating a range of possible values. At the iteration (d), the algorithm is defined to successfully converge if the relative difference, $\max_i |\theta_i^{(d)} - \theta_i^{(d-1)}|/|\theta_i^{(d-1)}|$ is less than a specific tolerance value, $\tau = 5 \times 10^{-4}$, otherwise it is stopped when a maximum number of 1000 iterations has been completed.

Table 5.1 investigates parameter estimation for different number of clusters and different cluster sizes at level of 50% of censoring. The reported estimates are computed at sample means on the successfully convergent datasets. To investigate the algorithm convergence, we define the convergence rate by total successful convergent datasets over 1000 considered datasets (notation "Conv.rate" in the table) and the average number of iteration for a dataset to be successfully convergent (notation "Avg.iter" in the table). Studying the convergence rates and the average iterations, we concluded that the algorithm is successful and converges fast for all settings, even in case of many observations (5 clusters with 9×9 units). Unsurprisingly, we observe that the convergence rate increases and the average iteration number decreases gradually when more observations are considered. Particularly, a high convergence rate (96.1%) is obtained in the last setting (5 clusters with 9×9 units). Moreover, compared to the maximum of 1000 iterations fixed in the implementation, the algorithm is very fast with an average of 401.57 iterations to get a successful convergence.

Parameter β is always stably estimated with negligible bias, whereas covariance parameters, σ and $\tilde{\nu}$, are more difficult to estimate. Particularly, the algorithm shows some bias in estimating parameter $\tilde{\nu}$. However, at the same number of observations (3 clusters with 9×9 units and 5 clusters with 7×7 units), an increasing number of clusters leads to a slight decrease in bias for the estimates $\hat{\beta}$ and $\hat{\sigma}$, but noticeably, there is a significant bias decrease for $\hat{\tilde{\nu}}$ (from 0.231 to 0.093) and it tends to be more precise when more observations are considered (5 clusters with 9×9 units). It means that this algorithm is more effective in correcting the estimation of the correlation parameter $\tilde{\nu}$, as compared to the variance parameter. In the last column of Table 5.1 , low empirical standard deviations are shown for $\hat{\beta}$ and $\hat{\sigma}$. On the other hand, parameter $\tilde{\nu}$ is characterized by high standard deviation regardless of clustering setting. Furthermore, at the same number of observations (3 clusters with 9×9 units and 5 clusters with 7×7 units), clustering increases slightly the standard deviation terms. However, both $\tilde{\nu}$ and σ are well estimated with acceptable bias and a significant reduction in standard deviations was noted for a lager number of observation (5 cluster with 9×9 units). In case of 30% censoring, the same conclusions about the performance of the QPLH method are also reported. Results are given in Table 5.2.

From the literature (see, e.g., Xu *et al.* (2016)), it is known that when composite likelihood are information-biased, this is the case especially in complex dependence data, then the efficiency is not necessary increased by incorporating additional independent component likelihoods. However, our model setting is a special case, where the likelihood dimension and

No. of locations	Parameter	True value	Bias	Empirical SD
3 clusters	β	0.6	-0.012	0.227
	σ	0.5	-0.194	0.137
7×7	$\tilde{\nu}$	1.0	-0.289	0.851
Conv.rate = 93.4%, Avg.iter = 609.506				
3 clusters	β	0.6	-0.008	0.173
	σ	0.5	-0.145	0.129
9×9	$\tilde{\nu}$	1.0	-0.231	0.735
Conv.rate = 95.4%, Avg.iter = 513.647				
5 clusters	β	0.6	-0.006	0.183
	σ	0.5	-0.134	0.138
7×7	$\tilde{\nu}$	1.0	0.093	0.795
Conv.rate = 95.1%, Avg.iter = 498.079				
5 clusters	β	0.6	- 0.009	0.138
	σ	0.5	-0.096	0.121
9×9	$\tilde{\nu}$	1.0	-0.061	0.669
Conv.rate = 96.1%, Avg.iter = 401.57				

TABLE 5.1: Results from QPLH method with Exponential correlation frailty on 1000 simulated datasets for 3 and 5 clusters at level of 50% of censoring. "Conv.rate" is the convergent rate of datasets and "Avg.iter" is the average iteration number for a dataset to be successfully convergent.

.

the frailty vector dimension are equal to the number of observations. Hence, presence of more observations should give more precise estimates as discussed before about the results in Tables 5.2 and 5.1. Consequently, the algorithm also consumes more computing time. Figure 5.2 shows the average computing times of QPLH algorithm in case of 3 clusters when the number of observations increases, where the computed time is averaged on 100 simulated datasets for each studied case. In case of few observations (75 observations for 3 clusters), the algorithm takes about 9 minutes, but it grows up quadratically with the number of observations (with the dimension of the frailty vector), in line with the conclusions of Lindsay *et al.* (2011). Computational expense of composite likelihoods are discussed in details in Lindsay *et al.* (2011).

Table 5.3 describes results from QPLH method at different levels of censoring for the case of 5 clusters with 81 observations per cluster. Unsurprisingly, a high censoring proportion (70% censoring) produces a lower convergent rate (93.2%) of the algorithm, even though the algorithm is faster (with an average of 373.223 iterations for a successful convergence), because fewer observed events require less computational burden. Moreover, when the censoring proportion increases, the bias of $\hat{\tilde{\nu}}$ substantially raises about 15-fold, compared to the case of

No. of locations	Parameter	True value	Bias	Empirical SD
3 clusters	β	0.6	-0.028	0.207
	σ	0.5	-0.223	0.137
7×7	$\tilde{\nu}$	1.0	-0.206	0.881
Conv.rate =90.5%, Avg.iter = 662.368				
3 clusters	β	0.6	-0.017	0.153
	σ	0.5	-0.169	0.145
9×9	$\tilde{\nu}$	1.0	-0.161	0.798
Conv.rate = 94.3%, Avg.iter = 548.312				
5 clusters	β	0.6	-0.017	0.165
	σ	0.5	-0.164	0.14
7×7	$\tilde{\nu}$	1.0	-0.058	0.738
Conv.rate = 92.8%, Avg.iter = 534.155				
5 clusters	β	0.6	-0.017	0.12
	σ	0.5	-0.119	0.12
9×9	$\tilde{\nu}$	1.0	0.003	0.683
Conv.rate = 94.8%, Avg.iter = 429.255				

TABLE 5.2: Results from QPLH method with Exponential correlation frailty on 1000 simulated datasets for 3 and 5 clusters at level of 30% of censoring. "Conv.rate" is the convergent rate of datasets and "Avg.iter" is the average iteration number for a dataset to be successfully convergent.

.

low censoring level (30% censoring). On the other hand, the bias terms in estimating parameters β and $\tilde{\nu}$ are mostly stable across the all scenarios considered. The empirical standard deviations increase when more observations are censored, and we note that the most clearly raise is for parameter $\tilde{\nu}$. It implies that parameter $\tilde{\nu}$ is much sensitive to the censoring level. However, generally we observed that QPLH method is working well with high convergence rates, negligible bias and standard deviations in all censoring settings, even in case of highly censored data.

5.3   Comparison of results

After the discussion of QPLH method in the previous section, here we present simulation results for the PPL and MCEM approaches as well as comparison of all three methods. Particularly, adopting the theory given in Section 3.2, PPL method is practically implemented in the R package `coxme`. We have extended this package to handle an Exponential correlation matrix allowing to incorporate the spatial correlation as well as clustering between frailties. Standard
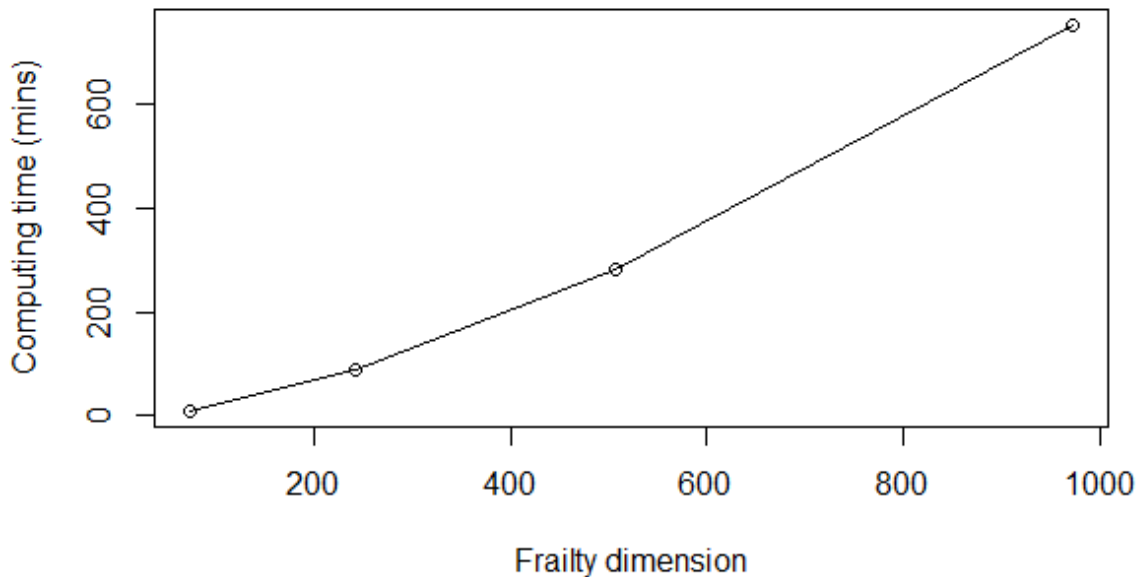
FIGURE 5.2: The average computing time of QPLH algorithm in case of 3 clusters with different numbers of observation.

errors of the variance-covariance parameters σ and $\tilde{\nu}$ are estimated by numerical tools. We have implemented the MCEM method (see Section 3.3) in the R software and the C language was used for the Gibbs sampling step, adapting the code in Li *et al.* (2015) to our proposed model. In this practical setting, thinned MCMC samples of 2000 frailty vectors are used for the approximation at each iteration in the E-step, and we also used a tolerance of $\tau = 5 \times 10^{-4}$ for the convergence criterion as described in Section 3.3. The results are given in Table 5.4.

First of all, we observe that PPL is the most effective method in estimating parameters β and σ, with the lowest bias for σ and bias similar to the other methods for β. Moreover, in the PPL method, negligible bias and low empirical standard deviations are stably observed in all settings. However, this method fails to estimate the correlation parameter $\tilde{\nu}$ because, in the simulations, we observed many extreme values (that is the reason why sample medians are reported for this parameter estimates, rather than sample means). The reason of this result could be probably attributed to non convergence of the optimization numerical procedure. Furthermore, even though the estimated standard errors of $\hat{\beta}$ are very similar to the empirical counterpart, the estimated standard errors for $\hat{\tilde{\nu}}$ and $\hat{\sigma}$ are either heavily underestimated or overestimated across the different settings, showing that the PPL-based method is not working well for estimating uncertainty of these two parameters. Consequently, PPL method failed to provide complete inferential good results for the proposed hierarchical spatial frailty model.

We compare now the QPLH and MCEM methods. With exception for the the first setting,

Censoring level	Parameter	True value	Bias	Empirical SD
30% censoring	β	0.600	-0.017	0.120
	σ	0.500	-0.119	0.120
	$\tilde{\nu}$	1.000	0.003	0.683
Conv.rate = 94.8%, Avg.iter= 429.255				
50% censoring	β	0.600	- 0.009	0.138
	σ	0.500	-0.096	0.121
	$\tilde{\nu}$	1.000	-0.061	0.669
Conv.rate = 96.1%, Avg.iter = 401.57				
70% censoring	β	0.600	-0.002	0.165
	σ	0.500	-0.081	0.070
	$\tilde{\nu}$	1.000	-0.043	0.755
Conv.rate = 93.2%, Avg.iter = 373.223				

TABLE 5.3: Results from QPLH method with Exponential correlation frailty on 1000 simulated datasets for 5 clusters with 9×9 units per cluster. The results are implemented at different levels of censoring. "Conv.rate" is the convergent rate of datasets and "Avg.iter" is the average iteration number for a dataset to be successfully convergent.

.

the regression parameter β is more accurately estimated with slightly lower bias under QPLH method, as compared to MCEM method. In contrast, the estimate $\hat{\sigma}$ produces lower bias under the MCEM approach. Interestingly, the estimate of parameter $\tilde{\nu}$ shows a very unstable behavior across settings. Particularly, in cases of few observation per cluster (3 clusters with 7×7 units and 5 clusters with 7×7 units), the MCEM method fails in estimating the correlation parameter $\tilde{\nu}$, because we observe a very high bias as well as very large standard deviations. QPLH method has instead acceptable bias and empirical standard deviation, in particular, $\tilde{\nu}$ is just slightly underestimated (bias 0.093) in the third setting (5 clusters with 7×7 units), compared to a heavy overestimation (bias 0.45) produced by MCEM method. However, when more observations are considered for each cluster, the MCEM approach produces slightly better results.

In all settings, $\hat{\beta}$ and $\hat{\sigma}$ are associated to similar empirical standard deviations, which assume low values, whereas the estimate of parameter $\tilde{\nu}$ is characterized by a larger empirical standard deviation. This results about a high standard deviation was observed in the simulation studies, regardless of the choice of the inferential method, and this implies challenges in estimating this parameter. Noticeably, compared to MCEM method, QPLH method always produces lower standard deviations for all parameters.

Both QPLH and MCEM methods show off particular limitations in computing the estimated standard errors. Since QPLH leads to complex calculations for estimating standard error terms, they have not been implemented in these simulation studies. However, challenges were also shown in the alternative methods. In details, standard errors computed by Louis information matrix in MCEM approach, are totally underestimated for the estimates of parameter σ and $\tilde{\nu}$. On the other side, sandwich standard errors of parameter $\hat{\sigma}$ are very similar to the empirical standard deviations, while sandwich standard errors for $\hat{\beta}$ are much more underestimated regardless of clustering setting. For parameter $\tilde{\nu}$, the sandwich standard error is acceptable in cases of having more observations per cluster (3 clusters with 9×9 units and 5 clusters with 9×9 units). These results implies that the assumption of independence between clusters has much more influence on the estimation of the covariance parameters σ and $\tilde{\nu}$, than on the estimation of β. Moreover, estimates of parameter $\tilde{\nu}$ is very sensitive to the number of observations per cluster. More discussions about Louis and sandwich standard are given in Appendix A.2.

The last point to be noted is the convergence speed. As discussed in the previous section, QPLH algorithm is very fast to converge, with more than 93% of datasets that converged before the maximum of 1000 iterations. Averagely, about 400 to 600 iterations are needed for a successful convergence. On the other hand, unsurprisingly, MCEM method is computationally very slow. For more than 98% of datasets, the algorithm needs to arrive to the maximum of 1000 iterations to obtain full converge. Therefore, when we performed our simulation studies, computational time to obtain results from the MCEM algorithm was about twice longer than the one for the QPLH algorithm.

5.4   Discussions and Conclusions

We structure the discussion by underlying: advantages and disadvantages for each method as given belows.

PPL method

   Advantages: This method may be effective in simple frailty models, such as the case of i.i.d shared frailty model that contains only the regression parameter β and the variance parameter σ. Moreover, the R package is well implemented and flexible to be extended to different kind of variance-covariance matrix.

   Disadvantages: As shown in the previous section, this method failed to estimate the correlation parameters, showing many wrong extreme values as well as underestimates of the standard errors. It may perform poorly in modeling the complex dependence structure of multidimensional random effects. The reason may be explained in the following aspects. Firstly, to get the penalized partial likelihood from the Laplace approximation of the full likelihood, the method ignores some terms. In details, two terms $-\frac{1}{2}\log|\Sigma(\theta_c)|$

		3 clusters: 7×7								
		QPLH		MCEM				PPL		
Par	True.val	Bias	Sd	Bias	Sd	L.SE	S.SE	Bias	Sd	SE
β	0.600	-0.012	0.227	0.006	0.230	0.199	0.133	0.012	0.232	0.226
σ	0.500	-0.194	0.137	-0.074	0.243	0.157	0.231	0.001	0.258	0.108
$\tilde{\nu}$	1.000	-0.289	0.851	1.243	9.447	2.207	1.765	-0.088*	1.034*	0.476*

		3 clusters: 9×9								
		QPLH		MCEM				PPL		
Par	True.val	Bias	Sd	Bias	Sd	L.SE	S.SE	Bias	Sd	SE
β	0.600	-0.008	0.173	0.021	0.182	0.155	0.099	0.018	0.176	0.174
σ	0.500	-0.145	0.129	-0.022	0.223	0.130	0.230	0.006	0.084	0.217
$\tilde{\nu}$	1.000	-0.231	0.735	0.150	1.778	0.837	1.135	-0.163*	0.238*	0.776*

		5 clusters: 7×7								
		QPLH		MCEM				PPL		
Par	True.val	Bias	Sd	Bias	Sd	L.SE	S.SE	Bias	Sd	SE
β	0.600	-0.006	0.183	0.013	0.250	0.154	0.118	0.011	0.186	0.173
σ	0.500	-0.134	0.138	-0.035	0.215	0.118	0.134	-0.012	0.072	0.201
$\tilde{\nu}$	1.000	-0.093	0.795	0.450	2.300	1.037	0.912	0.014*	0.802*	0.272*

		5 clusters: 9×9								
		QPLH		MCEM				PPL		
Par	True.val	Bias	Sd	Bias	Sd	L.SE	S.SE	Bias	Sd	SE
β	0.600	-0.009	0.138	0.010	0.143	0.120	0.090	0.001	0.136	0.134
σ	0.500	-0.096	0.121	0.004	0.194	0.108	0.191	-0.002	0.169	0.057
$\tilde{\nu}$	1.000	-0.061	0.669	0.060	0.785	0.475	0.554	0.024*	0.742*	0.118*

TABLE 5.4: A simulation study compares results of three approaches: QPLH, MCEM, and PPL methods. It includes true values (True.val), bias (Bias), empirical standard errors (Sd), estimated standard errors (SE). In the MECM method, the estimated standard errors are reported by Louis method (L.SE) and sandwich standard errors (S.SE). The results are implemented by using an Exponential correlation frailty on 1000 simulated datasets at level of 50% of censoring. Notations "*" indicate sample medians.

and

$-\frac{1}{2} \log | \sum_{i,k} \Lambda_0(t_{ik}) \exp \left(\beta^t X_{ik} + \tilde{z}_{ik} \right) I_{ik} I_{ik}^t + \Sigma(\theta_c)^{-1} |$ in the expression (3.3) are omitted. It is clear that some information of parameters β and $\Lambda_0(.)$ is lost in the second term, while parameters $\theta_c = (\sigma, \theta_\rho)$ are present in both two terms. Hence, more information of θ_c is lost when the model handles complex correlation structures of frailties and this may explain imprecise estimates of θ_c. Secondly, PPL method considers frailties as fixed effect parameters and estimates them. The algorithm involves a repeated inversion of the design matrix and the variance-covariance matrix for the individual frailties.

Consequently, it should show off maximization issues when dealing with cases of large number of frailties. Lastly, another limitation of this approach is implied by the adjusted iteration mechanism of the maximization procedure. Particularly, the algorithm is iterated between estimation of parameters (β, z) and estimation of θ_c at given starting values of θ_c, using profile likelihood method. Consequently, some bias of $\hat{\beta}$ may be arisen from bias of $\hat{\theta}_c$.

MCEM method

   Advantages: Unlike PPL method, the original likelihood function is maximized in MCEM method. Moreover, MCEM method does not aim to maximize the unknown frailties as in the PPL approach, alternatively they are generated for the approximation of the full likelihood function. Hence, the number of frailties as well as their complex dependence structure has influence only on the computational expense of the Gibbs sampling, and it does not complicate the maximization of the algorithm. In addition, the M-step of the algorithm separately maximizes the time-to-event parameters and the variance-covariance parameters, and then all the uncertainty of parameters is accounted.
   Disadvantage: In the simulation studies, we have observed that the algorithm is very sensitive to the clustering structure, showing high bias of the correlation parameter estimates in cases of few observation per cluster. Moreover, its limitation in modeling the independence of clusters can also be observed in the standard error terms where both Louis's formula and sandwich formula are unable to produce good standard error estimates of all parameters. In addition, MCEM method is typically very slow to converge. This issue may come from a large number of frailties and their complex dependence structure, which reduce the efficiency of the Gibbs sampling in the E-step. Lastly, both MCEM method and PPL method commonly struggle to deal with the problem of high-dimensional matrix inversion produced by the variance-covariance matrix of individual frailties.

QPLH method

   Advantages: QPLH method shows off its advantages in both the parameter estimation precision and the convergence speed. The potential key is that QPLH method simplifies the high-dimensional full likelihood to several 2-dimensional pseudo-likelihoods. In this way of simplification, the algorithm avoids to deal with high-dimensional matrix inversions that might produced issues in the estimation procedure of PPL and MCEM methods. Moreover, the composite likelihood function surface is usually much smoother than the one of the full likelihood and therefore, easier to be maximized (Katsikatsou *et al.*, 2012). Note that if a cluster has m observations, then QPLH method considers $m \times (m-1)$ pair observations for model inference. That seems to explain why MCEM method is unsuccessful in cases of few observation per cluster whereas QPLH is still stably working well. The pairwise likelihood is also not sensitive to the percentage of

censored data, since it shows stable estimates across different censoring levels in simulation studies. Lastly, QPLH algorithm uses the quadrature approximation in the E-step, while on the contrary, MCMC samples are used in the MCEM algorithm. Therefore, it produces a deterministic algorithm of QPLH, consequently the assessment of convergence is simpler and empirical standard deviations are also lower compared to MCEM approach.

Disadvantages: Besides the advantages, QPLH has also some limitations. The precision of the estimates, as well as computation expense of QPLH, depend on the particular set of all pair observations that contributes to the pairwise likelihood, and also on the number of quadrature nodes used in the E-step. Both issues should be verified practically by implementing the algorithm in different settings. Moreover, an increased number of observations leads to raising quadratically the computational time. Practically, this algorithm still needs to be improved for the benefit of computational time, but importantly the simulation studies have concluded that the high number of independent component in the likelihood does not complicate the parameter maximization and the algorithm is still computable in high dimension. For example, 972 frailties (likelihood dimensions) were considered in the last case of Figure 5.2. Lastly, we are currently working for the standard error estimates under the QPLH method, which will be included in the future submission of a scientific paper.

In conclusion, for modeling i.i.d frailty models, PPL method is a convenient choice, but for handling the spatial correlation between frailties, QPLH and MCEM methods are preferable. QPLH method is flexible for modeling many types of complex correlated data and has shown a good performance in the parameter estimation. Specially it overcomes the problem of poor estimation in case of heavy censoring (i.e 70%) and few observations per cluster.

Chapter 6

Real data application

In this chapter, we fit the proposed hierarchical spatial frailty model in Section 3.1 to the insulin-containing secretory granules data described in Chapter 2 by using the QPLH method. We implemented the QPLH algorithm with Exponential and Gaussian correlation functions, then we compared results to the standard Cox models. The setting of the algorithm is the same as used in simulation studies (tolerance 5×10^{-4}, quadrature nodes $M = 7$).

6.1   The biological dataset

We analyzed our real dataset (see Chapter 2) by fitting the proposed hierarchical spatial frailty model and applying the QPLH inferential approach with an Exponential correlation function. The results are reported in Table 6.1 and compared with those from standard Cox models. The QPLH algorithm converged after 365 iterations. As discussed in Section 2.3.2, there exist two distinct groups of granules identified by their presence/absence at beginning of the experiment. A higher level of syntaxin contributes to obtain a higher survival rate for granules that are present at beginning of the experiment, whereas syntaxin level has no effect on survival rates of granules in the other group. This conclusion seems to be coherent with results obtained by fitting QPLH method in this chapter, where regression coefficient estimates are similar between models. The estimate of σ is equal to 1.060 and 0.711, respectively, in the model with Exponential spatially correlated frailties and in the model with independent frailties, indicating that there exist unexplored aspects in the dataset. The small estimated value of parameter ν, equal to 0.432, in the Exponential correlation function indicates no spatial correlation between survival rates of granules. However, in this fitting, parameters σ and ν describe common variation and correlation of granules on the entire data, but actually they may also be affected by the entrance time in the experiment. Therefore, we propose to perform the analysis separately in the two groups of granules being presence or absence of the beginning of the experiment, in order to have a precise view on spatial variation.

	syn_0	syn_1	app	sigma	nu
Exponential	0.300	-2.332	-1.888	1.060	0.432
Cox (i.i.d frailty)	0.845	-3.048	-1.532	0.711	-
Cox (no.frailty)	0.734	-2.647	-1.167	-	-
n.obs = 1117; n.event = 885					

TABLE 6.1: Estimates of parameters in the hierarchical spatial frailty model by the QPLH method with an Exponential correlation function, and by standard Cox models (Cox models with independent frailties and without frailties). Parameter "app" indicates if granule is present at beginning of the experiment ($app = 1$) or not ($app = 0$). "syn_0" and "syn_1" are respectively syntaxin levels in group 0 (when $app = 0$) and in group 1 (when $app = 1$). Here, "n.obs" is the total number of observations and "n.event" is the total number of events.

6.2   Analysis of the subsets in the dataset

In this section, we analyze data separately in the two subsets: group 1 containing granules that are present at beginning of the experiment and group 0 containing granules who enter later in the experiment.

Group 1: Granules that are present at beginning of the experiment

Table 6.2 shows parameter estimates for data containing only granules that are present at beginning of the experiment. The results are obtained under the QPLH method and the algorithm converged after 867 and 135 iterations respectively for using Exponential and Gaussian correlation function. The regression parameter "syn" is stably estimated in all inferential methods, while the variance parameter estimate $\hat{\sigma}$ is slightly lower in the case of independent Gaussian frailty model. The correlation parameter ν is respectively equal to 109.606 and 291.771 in the models of Exponential correlation and Gaussian correlation, indicating very strong correlation between event rates of granules within cells. Figure 6.1 is shown to visualize this conclusion. For granules with distances within 150 pixels, Gaussian correlation function produces higher correlation than the Exponential function and a faster decay. Interestingly, the histogram shows off that most granules have at least one neighbor that is at the distance less than 150 pixels to itself. It means that within cells, event rates of granules are strongly correlated and the correlations among those granules range from 0.2 to 1. Figure 6.2 illustrates the estimated survival curves at a representative frailty of $z = 0.5$ for different syntaxin levels. Within the first 5 seconds, concentration of syntaxin on single granules does not seem to affect significantly the granule survival rates, but for longer survival times, a higher level of syntaxin contributes to a higher survival probability. Figure 6.3 shows a strong effect of frailty on the survival rate of granules. Considering the 50th second as an example, when the frailty value increases from -0.5 to 0, the survival probability decreases from 60% to 40% and falls to under 20% if the frailty value climbs to 0.5. The same behavior has been recorded in using Gaussian

correlation function and the related figures are given in Appendix A.3 (Figure A.3, Figure A.4). Moreover, the algorithm has been run at different initial values and a different number of quadrature node ($M = 8$), showing very similar results (see Appendix A.3, in particular Table A.2 and Table A.3).

	syn	sigma	nu
Exponential	-2.282	0.53	109.606
Gaussian	-2.238	0.606	291.771
Cox (i.i.d.frailty)	-2.186	0.363	-
Cox (no.frailty)	-2.086	-	-
	n.obs = 232; n.event = 138		

TABLE 6.2: Estimates of parameters in the hierarchical spatial frailty model on group 1. QPLH method with Exponential and Gaussian correlations are implemented and compared to the other standard Cox models (Cox models with independent frailties and without frailties). Here, "n.obs" is the total number of observations and "n.event" is the total number of events.
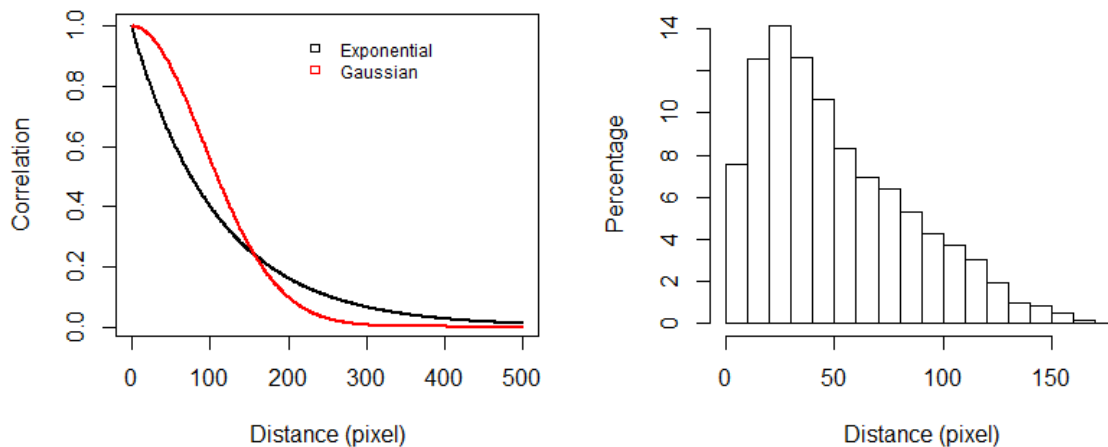
FIGURE 6.1: Estimated correlation functions of granules within cells and the histogram of granule distances in unit of pixel.

Group 0: Granules that enter later in the experiment

On the other hand, in the group of granules with later entrance, negligible estimates of parameter "syn" (about 0.1 in spatial models and about 0.6 in standard Cox models) and parameter ν (0.604 and 0.874), given in Table 6.3, indicate no interaction between syntaxin level and event rate, as well as no spatial correlation between granules in this group. However, random effects may still exist in this dataset because of variance parameter estimates ($\hat{\sigma} = 0.1, 0.562, 0.565$) in the different models. This behavior is coherent with the results discussed in Section 2.3.2.

exponential correlation

FIGURE 6.2: Estimated survival curves at a representative frailty $z = 0.5$.

The QPLH algorithm converged after 239 and 260 iterations respectively, when using an Exponential correlation function or a Gaussian correlation function. The issue of local maxima has also been checked by running the algorithm at different initial values, and similar results are also recorded with different number of quadrature node ($M = 8$). These additional results are provided in Appendix A.3 (Table A.4, Table A.5).

	syn	sigma	nu
Exponential	0.113	0.565	0.604
Gaussian	0.096	0.562	0.874
Cox (i.i.d.frailty)	0.694	0.1	-
Cox (no.frailty)	0.691	-	-
n.obs = 885; n.event = 747		

TABLE 6.3: Estimates of parameters in the hierarchical spatial frailty model on group 0. QPLH method with Exponential and Gaussian correlations are implemented and compared to the other standard Cox models (Cox models with independent frailties and without frailties). Here, "n.obs" is the total number of observations and "n.event" is the total number of events
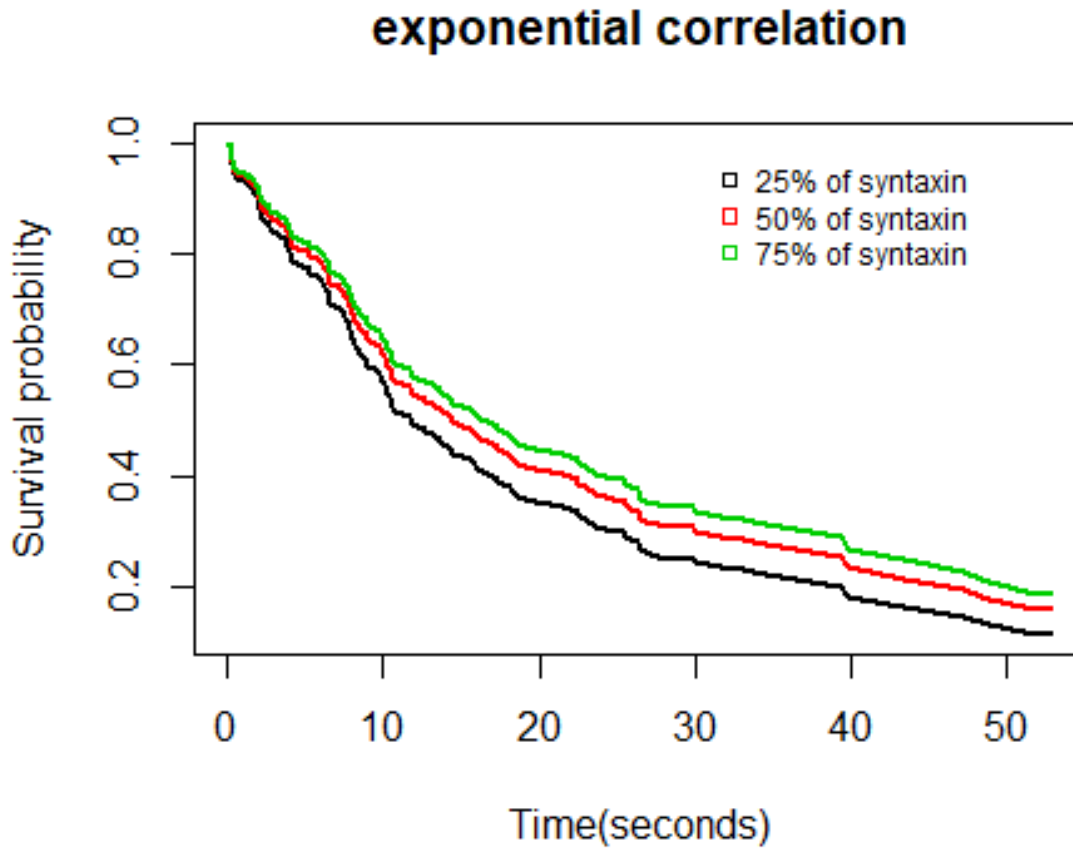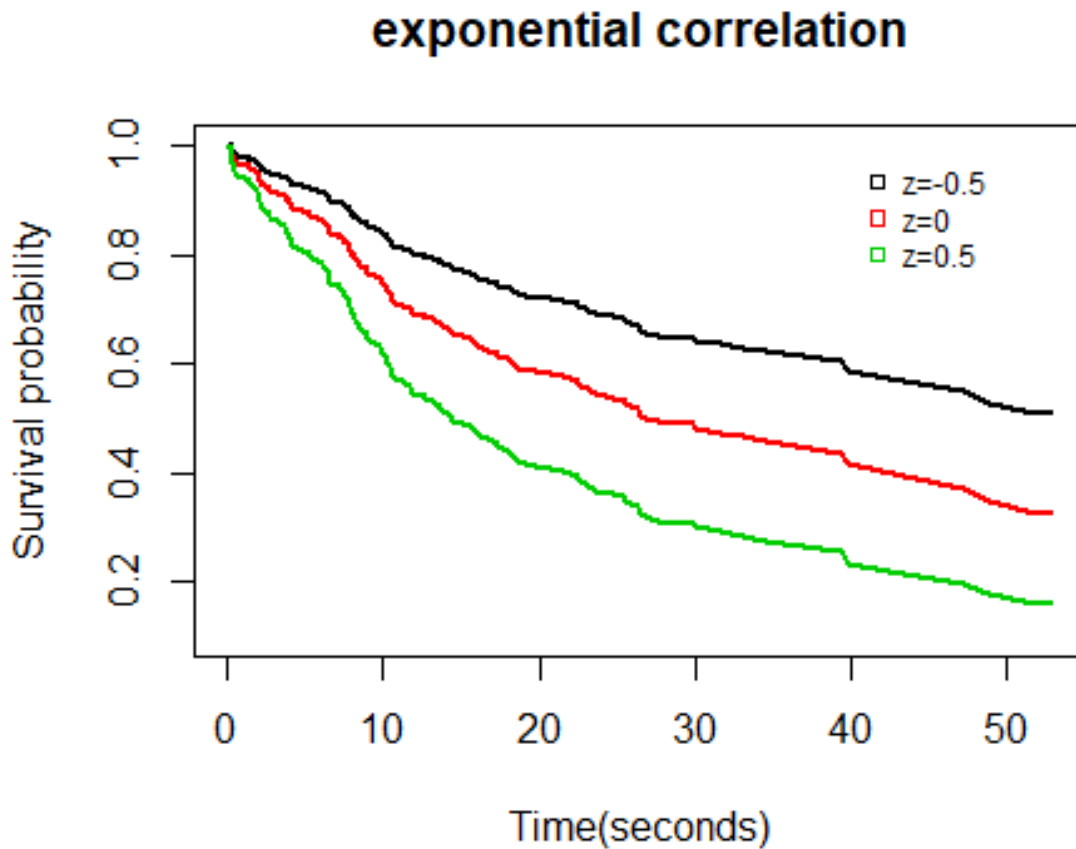
FIGURE 6.3: Estimated survival curves at a fixed syntaxin level for different values of frailties.

6.3 Conclusions and discussions

Starting from the entire dataset, we have discovered that there exist two groups of granules showing opposite behaviors in the interaction with syntaxin level as well as in the spatial correlation between granules. Linking this conclusion to the biological papers (Gandasi and Barg (2014), Barg *et al.* (2010)), our results indicate indirectly that with a high probability, exocytotic events (granules who are successfully released to blood) will be concentrated in group 1, because granules need syntaxin clusters to be docked before being released. On the other hand, for granules in group 0, our results showed off no interaction between syntaxin level and their event rates on the plasma membrane. Hence, this group might mainly contain "visitors" which are failed to recruit proteins during their presence at the plasma membrane (see Section 2.1). Furthermore, as mentioned in Section 2.1, there exist two event types in the dataset, which are very difficult to be distinguished visually from TIRF images. The first one is "exocytotic event" observed in granules that are successfully released to blood, and the other event is "undocking", which is defined for granules who have been stabilized on the

plasma membrane by syntaxin clusters but then they return to cellular fluid (Gandasi and Barg (2014), Barg *et al.* (2010)). Therefore, these results open further research related to granule classification by event type, which could be interestingly studied by competing risks survival models.

Conclusions

Discussion

Exocytosis is itself a complex biological mechanism that has never been studied in statistics. Moreover, starting from a raw biological imaging data required us much efforts to well understand the internal biological mechanism as well as to learn the imaging processing techniques. On working with this dataset, we started with the aim of modeling the exocytotic rates of granules (granules who are successfully released to blood), but we coped with two intractable issues. Firstly, it is very complicated to identify exocytotic events among other types. Secondly, the model need to deal with high censored data due to a low probability of exocytotic events. Consequently, we have alternatively worked on survival times of granules on the plasma membrane which disappear because of either an exocytotic event or a return to the cellular fluid. Interestingly, our first results explained indirectly some biological properties reported in the reference papers (Gandasi and Barg (2014), Barg *et al.* (2010)), contributed also to new explorations and broadened potential studied directions.

Inference on spatial frailty survival models based on maximum likelihood approaches is a challenging problem because of the high dimension of the full likelihood and complications of the maximization procedure. Common approaches, such as the PPL method, might be preferable to be used with uncorrelated data, while the MCEM method might be effective in shared frailty models without the frailty clustering structure. However, both of them are still intractable in dealing with a large number of frailties or with the complex dependence structure between frailties. Here, we have tackled such problem by using the pairwise likelihood technique, which in our simulation studies, has presented advantages compared to other common approaches. Specially, QPLH method overcome the problem of poor performance in the cases of few observation per cluster or heavy censored data. The first feature is frequently observed in cellular studies. For example, in our real data of group 1, there are few granules per cell. On the other side, the difficulty of heavy censored data in the study of exocytotic rate, that we have coped at beginning, is expected to be solved by our proposed QPLH method. Moreover, we have also implemented the QPLH algorithm in a R package. Therefore, our work on this method has contributed to an innovative inference approach in modeling spatial frailty survival models with an available R package for dissemination of the scientific results and application to other practical real data problems.

Future directions of research

First of all, we need to complete the work on standard error estimates for the QPLH method. Then, it is expected to investigate with simulation studies, as well as in the application, different types of correlation functions.

The proposed hierarchical spatial frailty model can be extended to incorporate time-dependent covariates, such as trajectories of syntaxin level, and multiple covariates. See, for example, our biological problem where there is great interest in studying the effects of multiple proteins on the behaviour of granules in β-cells.

Another interesting avenue for future work is to generalize the QPLH method to other types of spatial frailty survival models, such as shared frailty models, nested frailty models or to handle competing risks data.

Appendix

A.1 The first and second derivatives of the bivariate log-likelihood $pl(t_i, t_j; \theta)$

Proof of equation (4.13):

$$
\begin{aligned}
\nabla pl(\theta; t_i, t_j) &= \nabla \log\left[f(t_i, t_j; \theta)\right] = \nabla \log\left\{\int \int f(t_i, t_j, z_i, z_j; \theta) dz_i dz_j\right\} \\
&= \frac{\nabla \int \int f(t_i, t_j, z_i, z_j; \theta) dz_i dz_j}{\int \int f(t_i, t_j, z_i, z_j; \theta) dz_i dz_j} = \frac{\int \int \nabla f(t_i, t_j, z_i, z_j; \theta) dz_i dz_j}{f(t_i, t_j; \theta)} \\
&= \int \int \frac{\nabla f(t_i, t_j, z_i, z_j; \theta)}{f(t_i, t_j, z_i, z_j; \theta)} \frac{f(t_i, t_j, z_i, z_j; \theta)}{f(t_i, t_j; \theta)} dz_i dz_j \\
&= \int \int \nabla \log\left\{f(t_i, t_j, z_i, z_j; \theta)\right\} f(z_i, z_j | t_i, t_j; \theta) dz_i dz_j \\
&= \mathbb{E}_{z_i z_j | t_i t_j}\left\{\nabla \log\left[f(t_i, t_j, z_i, z_j; \theta)\right]\right\}
\end{aligned}
\tag{A.1}
$$

Proof of equation (4.14):

$$
\begin{aligned}
\nabla^2 pl(\theta; t_i, t_j) &= \nabla \left\{\int \int \nabla \log\left\{f(t_i, t_j, z_i, z_j; \theta)\right\} f(z_i, z_j | t_i, t_j; \theta) dz_i dz_j\right\} \\
&= \int \int \nabla^2 \log\left\{f(t_i, t_j, z_i, z_j; \theta)\right\} f(z_i, z_j | t_i, t_j; \theta) dz_i dz_j \\
&\quad + \int \int \nabla \log\left\{f(t_i, t_j, z_i, z_j; \theta)\right\} \nabla f(z_i, z_j | t_i, t_j; \theta) dz_i dz_j \\
&= \mathbb{E}_{z_i z_j | t_i t_j}\left\{\nabla^2 \log\left[f(t_i, t_j, z_i, z_j; \theta)\right]\right\} \\
&\quad + \int \int \nabla \log\left\{f(t_i, t_j, z_i, z_j; \theta)\right\} \nabla f(z_i, z_j | t_i, t_j; \theta) dz_i dz_j \\
&= \mathbb{E}_{z_i z_j | t_i t_j}\left\{\nabla^2 \log\left[f(t_i, t_j, z_i, z_j; \theta)\right]\right\} + A
\end{aligned}
\tag{A.2}
$$

Denote the second term in (A.2) by A and it can be analyzed as follows

$$A = \int \int \nabla \log \left\{ f(t_i, t_j, z_i, z_j; \theta) \right\} f(z_i z_j | t_i, t_j; \theta) \left[\frac{\nabla f(z_i, z_j | t_i, t_j; \theta)}{f(t_i, t_j, z_i, z_j; \theta)} \right] dz_i dz_j$$

$$= \int \int \nabla \log \left\{ f(t_i, t_j, z_i, z_j; \theta) \right\} f(z_i, z_j | t_i, t_j; \theta) \nabla \log \left\{ f(z_i, z_j | t_i, t_j; \theta) \right\} dz_i dz_j$$

$$= \int \int \left[\nabla \log \left\{ f(t_i, t_j, z_i, z_j; \theta) \right\} \right] \left[\nabla \log \left\{ f(t_i, t_j, z_i, z_j; \theta) \right\} \right]^t f(z_i, z_j | t_i, t_j; \theta) dz_i dz_j$$

$$- \int \int \nabla \log \left\{ f(t_i, t_j, z_i, z_j; \theta) \right\} \nabla \log \left\{ f(t_i, t_j; \theta) \right\} f(z_i, z_j | t_i, t_j; \theta) dz_i dz_j$$

$$= \mathbb{E}_{z_i z_j | t_i t_j} \left\{ \left[\nabla \log \left\{ f(t_i, t_j, z_i, z_j; \theta) \right\} \right] \left[\nabla \log \left\{ f(t_i, t_j, z_i, z_j; \theta) \right\} \right]^t \right\}$$

$$- \nabla \log \left\{ f(t_i, t_j; \theta) \right\} \int \int \nabla \log \left\{ f(t_i, t_j, z_i, z_j; \theta) \right\} f(z_i, z_j | t_i, t_j; \theta) dz_i dz_j$$

By using the result (A.1), we get

$$A = \mathbb{E}_{z_i z_j | t_i t_j} \left\{ \left[\nabla \log \left\{ f(t_i, t_j, z_i, z_j; \theta) \right\} \right] \left[\nabla \log \left\{ f(t_i, t_j, z_i, z_j; \theta) \right\} \right]^t \right\}$$

$$- \left[\mathbb{E}_{z_i z_j | t_i t_j} \left\{ \nabla \log f(t_i, t_j, z_i, z_j; \theta) \right\} \right] \left[\mathbb{E}_{z_i z_j | t_i t_j} \left\{ \nabla \log f(t_i, t_j, z_i, z_j; \theta) \right\} \right]^t$$

$$= \text{Var}_{z_i z_j | t_i t_j} \left\{ \nabla \log f(t_i, t_j, z_i, z_j; \theta) \right\} \qquad (A.3)$$

Combine equations (A.3) and (A.2), we have

$$\nabla^2 pl(\theta; t_i, t_j) = \mathbb{E}_{z_i z_j | t_i t_j} \left\{ \nabla^2 \log f(t_i, t_j, z_i, z_j; \theta) \right\} + \text{Var}_{z_i z_j | t_i t_j} \left\{ \nabla \log f(t_i, t_j, z_i, z_j; \theta) \right\}$$

that is equation (4.14).

A.2   Standard error discussion of MCEM method

Table A.1 shows a study on the coverage probabilities in the MCEM method. It is clear that sandwich standard errors give lower coverage probabilities than the Louis standard errors regardless of the clustering setting. The last column contains the percentages of cases where the Louis standard errors are greater than the sandwich standard errors. This percentages are observed very high for all settings and all parameters. Consequently, confidence intervals produced by sandwich standard errors are narrower in most of the simulations, leading to a lower coverage of the true value.

Coverage probabilities for σ and $\tilde{\nu}$ are systematically lower than those for β. One reason for that may be the substantial asymmetry of confidence intervals because they are strongly shifted to the left, leading to very asymmetric type-1 errors on the tails (two numbers in the parentheses). Moreover, when computing Louis standard errors for $\hat{\sigma}$ and $\hat{\tilde{\nu}}$, a percentage of samples is lost (about 20%) because of numerical problem, therefore the related coverage probabilities might be affected by a reduced sample size. Figure A.1 and A.2 give a visualization

of this result for the case of 5 clusters with 9×9 units per cluster. From boxplots, we observe that the poor estimates in standard errors of $\hat{\tilde{\nu}}$ and $\hat{\sigma}$ may be affected substantially by the long tails of outliers. On the other hand, from the histograms of figure A.1, we see that estimates of β are nearly following a normal distribution, explaining the reason why its standard error estimates have higher coverage probabilities.

Par	True.val	Bias	Sd	L.SE	S.SE	L.cov.prob (%)	S.cov.prob (%)	L.se > S.se (%)
3 clusters: 7×7								
β	0.600	0.006	0.230	0.199	0.133	91.1 (3.9, 5.0)	67.8 (15.2,17.0)	82.1
σ	0.500	-0.074	0.243	0.157	0.231	65.3 (27.8, 6.9)	45.0 (41.6, 13.4)	86.6
$\tilde{\nu}$	1.000	1.243	9.447	2.207	1.765	84.2(15.0, 0.8)	50.3(41.5, 8.2)	88.6
3 clusters: 9×9								
β	0.600	0.021	0.182	0.155	0.099	97.3(3.3, 6.0)	66.7(13.1, 20.2)	84.7
σ	0.500	-0.022	0.223	0.130	0.230	66.6(20.4, 13.0)	46.8(33.6, 15.9)	82.6
$\tilde{\nu}$	1.000	0.150	1.778	0.837	1.135	71.6(17.9, 0.5)	46.7(45.3, 8.0)	83.3
5 clusters: 7×7								
β	0.600	0.013	0.250	0.154	0.118	88.6(4.2, 7.2)	73.8(11.1, 13.1)	79.0
σ	0.500	-0.035	0.215	0.118	0.134	66.2(22.9, 10.9)	51.2(33.2, 15.2)	79.5
$\tilde{\nu}$	1.000	0.450	2.300	1.037	0.912	81.8(17.7, 0.5)	61.0 (33.2, 5.8)	80.5
5 clusters: 9×9								
β	0.600	0.010	0.143	0.120	0.090	90.1(4.0, 5.6)	75.1 (10.2, 14.7)	81.7
σ	0.500	0.004	0.194	0.108	0.191	67.0(17.4, 15.4)	54.7(25.8, 19.5)	72.9
$\tilde{\nu}$	1.000	0.060	0.785	0.475	0.554	69.7(28.6, 1.7)	61.6(33.4, 4.9)	74.4

TABLE A.1: Simulation results of MCEM method by using an Exponential correlation frailty on 1000 simulated datasets at level of 50% of censoring. The results include parameter names (par), true values of parameters (True.val), the bias of estimates (bias), the empirical standard deviations (Sd), the estimated standard errors and the corresponding coverage probabilities by Louis method (L.SE & L.cov.prob), the estimated sandwich standard errors and the corresponding coverage probabilities(S.SE & S.cov.prob), and lastly the percentages such that the estimated standard errors is greater than the estimated sandwich standard errors (L.se > S.se). Two numbers in the parentheses of the coverage probabilities are the upper rate (when the true value is greater than the estimated confidence interval) and lower rate (when the true value is lower than the estimated confidence interval) respectively.
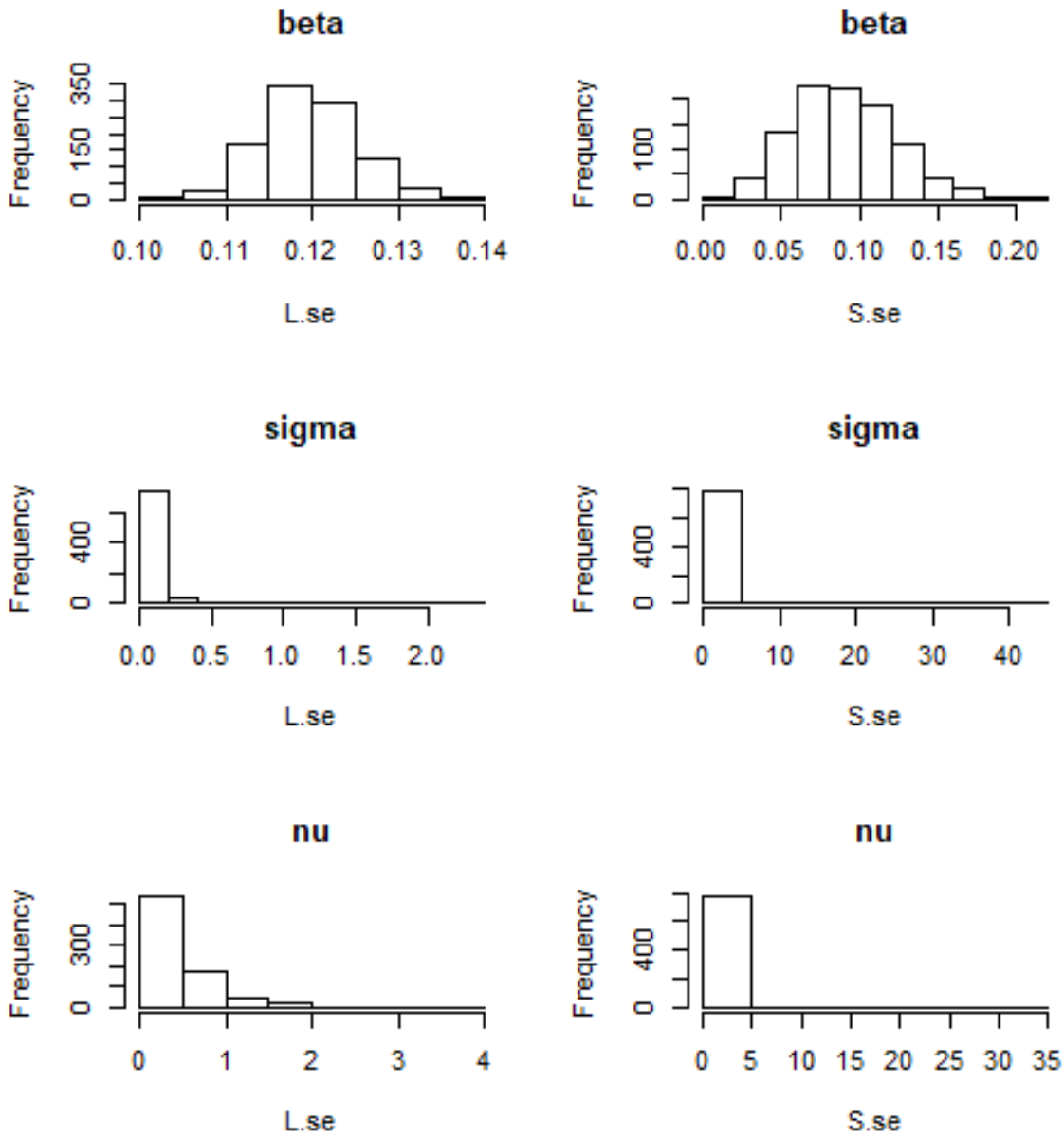
FIGURE A.1: Histograms of standard errors computed by Louis method and sandwich method in a simulated study of 5 clusters with 9×9 observations per cluster.

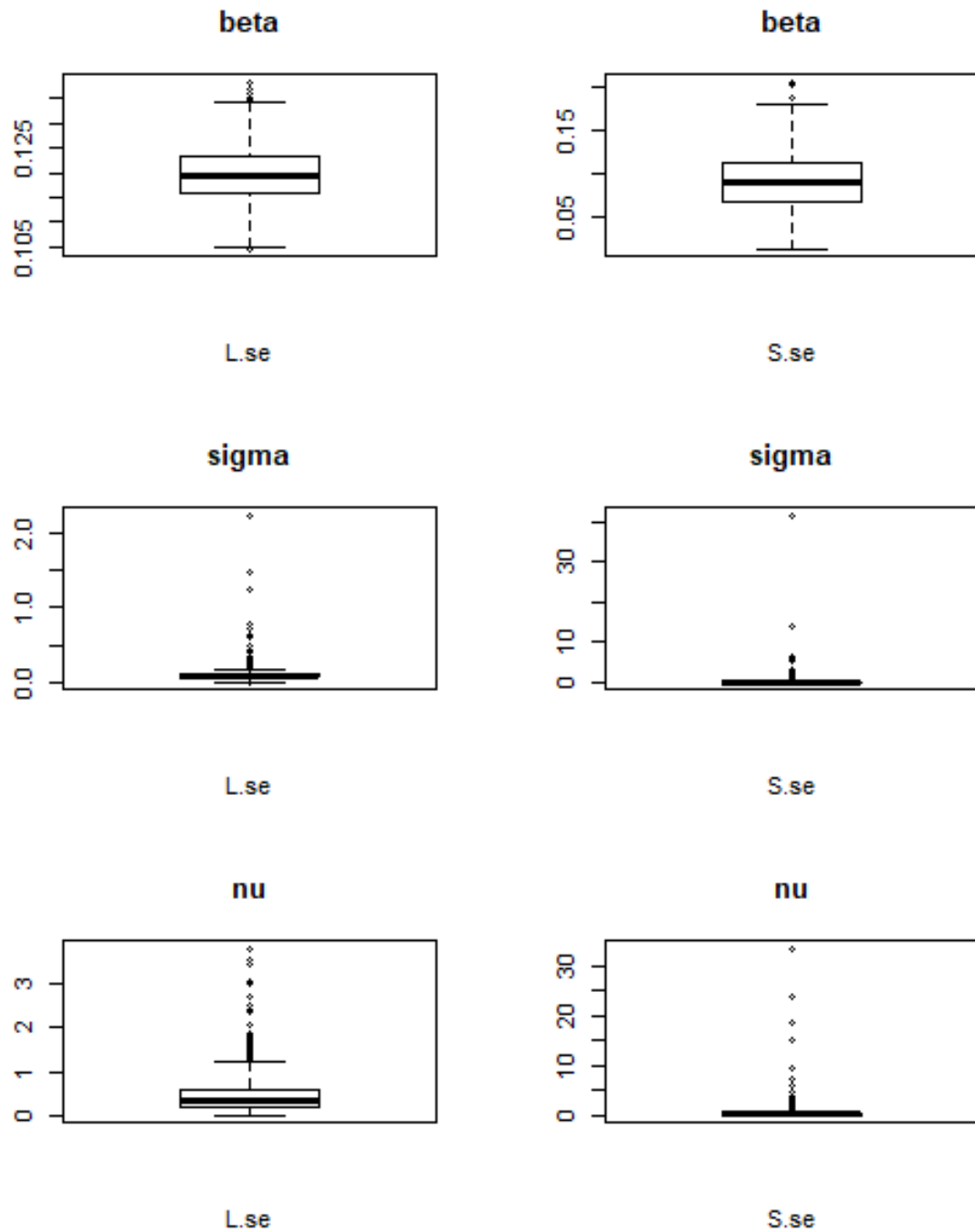## A.3   Additional results on the real data application

FIGURE A.2: Boxplots of standard errors computed by Louis method and sandwich method in a simulated study of 5 clusters with 9×9 observations per cluster.

Parameters	Exponential	Gaussian
syn	-2.287	-2.263
sigma	0.532	0.518
nu	107.206	78.338
no.iter	1000	1000

TABLE A.2: Estimates of parameters in the hierarchical spatial frailty model on group 1. QPLH method with Exponential and Gaussian correlations are implemented at initial values $\sigma = 0.7$, $\nu = 20$ compared to results in Table 6.2 at initial values $\sigma = 0.7$, $\nu = 200$.

Parameters	Exponential	Gaussian
syn	-2.284	-2.239
sigma	0.530	0.476
nu	110.071	131.414
no.iter	867	135

TABLE A.3: Estimates of parameters in the hierarchical spatial frailty model on group 1. The QPLH method with Exponential and Gaussian correlations are implemented at the number of quadrature node $M = 8$ and compared to results in Table 6.2 with $M = 7$.

Parameters	Exponential	Gaussian
syn	0.113	0.069
sigma	0.565	0.455
nu	0.604	1.078
no.iter	245	569

TABLE A.4: Estimates of parameters in the hierarchical spatial frailty model on group 0. The QPLH method with Exponential and Gaussian correlations are implemented at initial values $\sigma = 0.7$, $\nu = 20$ and compared to results in Table 6.2 at initial values $\sigma = 0.7$, $\nu = 2$.

Parameters	Exponential	Gaussian
syn	0.110	0.095
sigma	0.563	- 0.560
nu	0.586	- 0.870
no.iter	259	- 256

TABLE A.5: Estimates of parameters in the hierarchical spatial frailty model on group 0. The QPLH method with Exponential and Gaussian correlations are implemented at the number of quadrature node $M = 8$ and compared to results in Table 6.2 with $M = 7$.
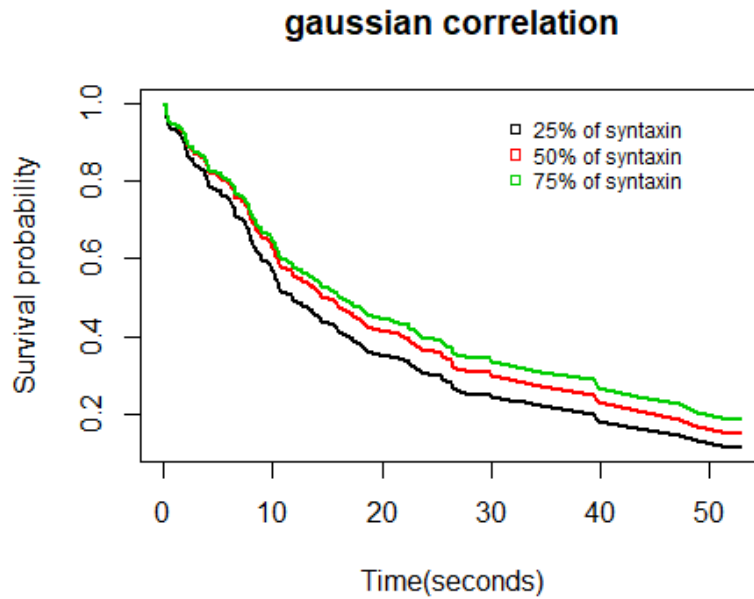
FIGURE A.3: Estimated survival curves at a representative frailty $z = 0.5$. The result is obtained from the implementation of the QPLH method with a Gaussian correlation function.
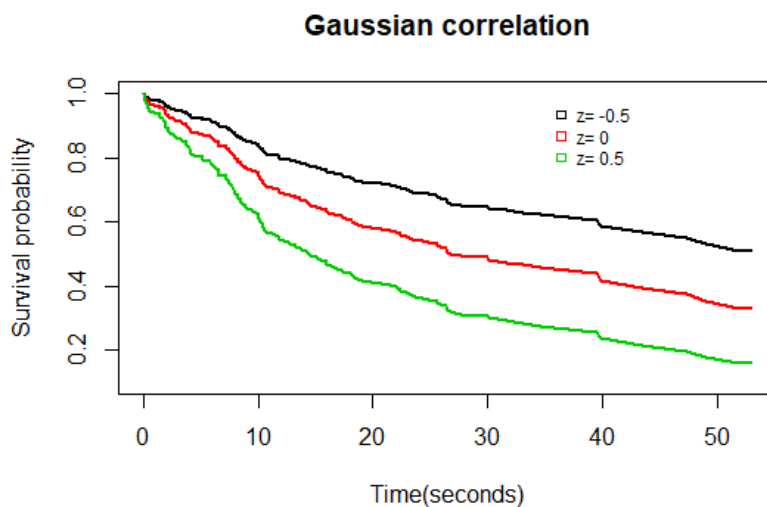

FIGURE A.4: Estimated survival curves at a fixed syntaxin level for different values of frailties. The result is obtained from the implementation of the QPLH method with a Gaussian correlation function.

Bibliography

Abramowitz, M. and Stegun, I. A. (1964) *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Volume 55. Courier Corporation.

Andersen, P. K., Klein, J. P., Knudsen, K. M. and y Palacios, R. T. (1997) Estimation of variance in cox's regression model with shared gamma frailties. *Biometrics* pp. 1475–1484.

Banerjee, S., Wall, M. and Carlin, P. (2003) Frailty modeling for spatially correlated survival data, with application to infant mortality in minnesota. *Biostatistics* **4**(1), 123–142.

Barg, S., Knowles, M., Chen, X., Midorikawa, M. and Almers, W. (2010) Syntaxin clusters assemble reversibly at sites of secretory granules in live cells. *Proceedings of the National Academy of Sciences* **107**(48), 20804–20809.

Bender, R., Augustin, T. and Blettner, M. (2005) Generating survival times to simulate cox proportional hazards models. *Statistics in medicine* **24**(11), 1713–1723.

Berg, J. M., Tymoczko, J. L., Stryer, L. and Stryer, L. (2002) *Biochemistry. 5th edit.* WH Freeman and Company New York.

Blair, D. and Dufresne, E. (n.d.) The matlab particle tracking code repository. `http://site.physics.georgetown.edu/matlab/`.

Breslow, N. (1974) Covariance analysis of censored survival data. *Biometrics* pp. 89–99.

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American statistical Association* **88**(421), 9–25.

Clayton, D. G. (1978) A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**(1), 141–151.

Cox, D. R. and Oakes, D. (1984) Analysis of survival data. 1984. *Chapman&Hall, London* .

Creighton, T. E. (1993) *Proteins: structures and molecular properties*. Macmillan.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)* pp. 1–38.

Efron, B. and Hinkley, D. V. (1978) Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika* **65**(3), 457–483.

Gandasi, N. R. and Barg, S. (2014) Contact-induced clustering of syntaxin and munc18 docks secretory granules at the exocytosis site. *Nature communications* **5**, 3914.

Gao, X. and Song, P. X.-K. (2011) Composite likelihood em algorithm with applications to multivariate hidden markov model. *Statistica Sinica* pp. 165–185.

Godambe, V. P. (1960) An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics* **31**(4), 1208–1211.

Green, P. J. (1987) Penalized likelihood for general semi-parametric regression models. *International Statistical Review/Revue Internationale de Statistique* pp. 245–259.

Guo, G. and Rodriguez, G. (1992) Estimating a multivariate proportional hazards model for clustered data using the em algorithm, with an application to child survival in guatemala. *Journal of the American Statistical Association* **87**(420), 969–976.

Henderson, R., Shimakura, S. and Gorst, D. (2002) Modeling spatial variation in leukemia survival data. *Journal of the American Statistical Association* **97**(460), 965–972.

Hougaard, P. (1995) Frailty models for survival data. *Lifetime data analysis* **1**(3), 255–273.

Huber, P. J. *et al.* (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pp. 221–233.

Jäckel, P. (2005) A note on multivariate gauss-hermite quadrature. *London: ABN-Amro. Re* .

Johansen, S. (1983) An extension of cox's regression model. *International Statistical Review/Revue Internationale de Statistique* pp. 165–174.

Joslin, E. P. and Kahn, C. R. (2005) *Joslin's Diabetes Mellitus: Edited by C. Ronald Kahn...[et Al.].* Lippincott Williams & Wilkins.

Kahn, S. (2003) The relative contributions of insulin resistance and beta-cell dysfunction to the pathophysiology of type 2 diabetes. *Diabetologia* **46**(1), 3–19.

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F. and Jöreskog, K. G. (2012) Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics & Data Analysis* **56**(12), 4243–4258.

Kent, J. T. (1982) Robust properties of likelihood ratio tests. *Biometrika* **69**(1), 19–27.

Klein, J. P. (1992) Semiparametric estimation of random effects using the cox model based on the em algorithm. *Biometrics* pp. 795–806.

Klein, J. P. and Moeschberger, M. L. (2006) *Survival analysis: techniques for censored and truncated data.* Springer Science & Business Media.

Knowles, M., Barg, S., Wan, L., Midorikawa, M., Chen, X. and Almers, W. (2010) Single secretory granules of live cells recruit syntaxin-1 and synaptosomal associated protein 25 (snap-25) in large copy numbers. *Proceedings of the National Academy of Sciences* **107**(48), 20810–20815.

Li, J., Hong, Y., Thapa, R. and Burkhart, E. (2015) Survival analysis of loblolly pine trees with spatially correlated random effects. *Journal of the American Statistical Association* **110**(510), 486–502.

Li, Y. and Ryan, L. (2002) Modeling spatial survival data using semiparametric frailty models. *Biometrics* **58**(2), 287–297.

Liang, G. and Yu, B. (2003) Maximum pseudo likelihood estimation in network tomography. *IEEE Transactions on Signal Processing* **51**(8), 2043–2053.

Lindsay, B. G. (1988) Composite likelihood methods. *Contemporary mathematics* **80**(1), 221–239.

Lindsay, B. G., Yi, G. Y. and Sun, J. (2011) Issues and strategies in the selection of composite likelihoods. *Statistica Sinica* pp. 71–105.

Liu, Q. and Pierce, D. A. (1994) A note on gausshermite quadrature. *Biometrika* **81**(3), 624–629.

Louis, T. A. (1982) Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 226–233.

Nott, D. J. and Rydén, T. (1999) Pairwise likelihood methods for inference in image models. *Biometrika* **86**(3), 661–676.

Olofsson, C. S., Göpel, S. O., Barg, S., Galvanovskis, J., Ma, X., Salehi, A., Rorsman, P. and Eliasson, L. (2002) Fast insulin secretion reflects exocytosis of docked granules in mouse pancreatic b-cells. *Pflügers Archiv* **444**(1-2), 43–51.

Otsu, N. (1979) A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics* **9**(1), 62–66.

Parner, E. *et al.* (1998) Asymptotic theory for the correlated gamma-frailty model. *The Annals of Statistics* **26**(1), 183–214.

Ripatti, S., Larsen, K. and Palmgren, J. (2002) Maximum likelihood inference for multivariate frailty models using an automated monte carlo em algorithm. *Lifetime Data Analysis* **8**(4), 349–360.

Ripatti, S. and Palmgren, J. (2000) Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics* **56**(4), 1016–1022.

Rogers, W. (1994) Regression standard errors in clustered samples. *Stata technical bulletin* **3**(13).

Therneau, T. (2015) Mixed effects cox models. *R package version* **2**(3).

Therneau, T. M. and Grambsch, P. M. (2013) *Modeling survival data: extending the Cox model.* Springer Science & Business Media.

Vaida, F. and Xu, R. (2000) Proportional hazards model with random effects. *Statistics in medicine* **19**(24), 3309–3324.

Varin, C., Høst, G. and Skare, Ø. (2005) Pairwise likelihood inference in spatial generalized linear mixed models. *Computational statistics & data analysis* **49**(4), 1173–1191.

Varin, C., Reid, N. and Firth, D. (2011) An overview of composite likelihood methods. *Statistica Sinica* pp. 5–42.

White, H. (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society* pp. 817–838.

Xu, X., Reid, N. and Xu, L. (2016) Note on information bias and efficiency of composite likelihood. *arXiv preprint arXiv:1612.06967* .

Yashin, A. I. and Iachine, I. A. (1995) Genetic analysis of durations: correlated frailty model app ed to survival of danish twins. *Genetic epidemiology* **12**(5), 529–538.

Zhou, H., Hanson, T., Jara, A. and Zhang, J. (2015) Modelling county level breast cancer survival data using a covariate-adjusted frailty proportional hazards model. *The annals of applied statistics* **9**(1), 43.

Thi Huong Phan

CURRICULUM VITAE

Contact Information

University of Padova
Department of Statistics
Via Cesare Battisti, 241-243
35121 Padova, Italy.

Tel. +39 347 353 3381
Email: thihuong.phan@studenti.unipd.it

Current Position

Since October 2015; (expected completion: October 2018)
PhD Student in Statistical Sciences, University of Padova.
Thesis title: Spatial Survival Models for Analysis of Exocytotic Events on Human β-cells Recorded by TIRF Imaging.
Supervisor: Prof.Giuliana Cortese

Research interests

- Survival analyis, dynamic survival regression models, competing risks models and prediction.
- Statistical inference based on the likelihood.

Education

September 2014 – July 2015
Master (*laurea specialistica/magistrale***) degree in Applied Mathematics** .
University of Lille 1, Faculty of Sciences and Technologies
Title of dissertation: Adaptive pointwise estimation for pure jump Lévy processes
Supervisor: Prof. Emeline Schmisser

September 2010 – March 2014
Bachelor degree (*laurea triennale***) in Sciences**.
University of Sciences-Vietnam National University, Faculty of Mathematics and Computer Science.

Awards and Scholarship

October 2015
PhD scholarship at University of Padova, Department of Statistics.
September 2014
Master scholarship in University of Lille 1.
September 2013
Scholarship of Vietnam Institute for Advance Study in Mathematics (VIASM).

Computer skills

- Programing languages: R, C/C++, Mathlab.

Language skills

Vietnamese: native;
English: good;

Conference presentations

Thi Huong Phan, Giuliana Cortese (09/2017). Spatial Survival Models for Analysis of Exocytotic Events on Human β-cells Recorded by TIRF Imaging. (Contributed oral presentation)
11th Scientific Meeting of the Classification and Data Analysis Group of the Italian Statistical Society - CLADAG 2017 , Milan, Italy, September 13rd to 15th 2017.

Thi Huong Phan, Giuliana Cortese (07/2018). Spatial Survival Models for Analysis of Exocytotic Events on Human β-cells Recorded by TIRF Imaging. (Contributed oral presentation)
XXIXth International Biometric Conference (29th IBC) , Barcelona, Italy, July 8th to 13rd 2018.

References

Prof. Giuliana Cortese
University of Padova
Department of Statistics
Via Cesare Battisti, 241-243, 35121 Padova. Italy

Web: `http://homes.stat.unipd.it/giulianacortese/`
Email: gcortese@stat.unipd.it
Phone:+39 049 8274159

Prof. TRAN Viet Chi
Université des Sciences et Technologies de Lille,
Laboratoire Paul Painlevé - UMR CNRS 8524
Bureau 316, Bat M3, Cité scientifique
59 655 Villeneuve d'Ascq Cedex

Web: `http://math.univ-lille1.fr/~tran/`
Email: chi.tran@math.univ-lille1.fr
Phone: 03 20 43 49 88

Prof. DUONG Minh Duc
University of Science, VNU-HCM
Deparment of Mathematics and Computer Science
227 Nguyen Van Cu, Dist. 5, HCM, Vietnam

Email: dmduc@hcmus.edu.vn
Phone: +84 83 89 96 484