UFAM

FEDERAL UNIVERSITY OF AMAZONAS

EXACT SCIENCE INSTITUTE

DOCTORAL PROGRAM IN MATHEMATICS PDM-UFPA/UFAM

DOCTORATE IN MATHEMATICS

ESSAYS ON CURE RATE MODEL

MÁRCIA BRANDÃO DE OLIVEIRA MARTINS

Doctoral thesis

MANAUS-AM

2024

Márcia Brandão de Oliveira Martins

ESSAYS ON CURE RATE MODELS

Doctoral thesis submitted to the doctoral program in mathematics jointly offered by the Federal University of Pará and the Federal University of Amazonas as a partial requirement for obtaining a Ph.D. in Mathematics.

Advisor: Prof. Dr. Jeremias da Silva Leão
Co-advisor: Prof. Dr. Marcelo Bourguignon

Concentration area: Applied Mathematics

MANAUS
2024

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

M386e

Martins, Márcia
Essays on cure rate models / Márcia Martins . 2024
95 f.: il. color; 31 cm.

Orientador: Jeremias da Silva Leão
Coorientador: Marcelo Bourguignon Pereira
Tese (Doutorado em Matemática) - Universidade Federal do Amazonas.

1. Cure-rate. 2. power-series. 3. birbaum-saunders. 4. EM-algorithm. 5. cancer. I. Leão, Jeremias da Silva. II. Universidade Federal do Amazonas III. Título

Ministério da Educação
Universidade Federal do Amazonas
Coordenação do Programa de Pós-Graduação em Matemática

FOLHA DE APROVAÇÃO

"*ESSAYS ON CURE RATE MODELS*"

MÁRCIA BRANDÃO DE OLIVEIRA MARTINS

Tese de Doutorado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Dr. Jeremias da Silva Leão - UFAM (Presidente)

Prof.ª Dr.ª Agatha Sacramento Rodrigues - UFES (Membro Externo)

Prof. Dr. Alex Leal Mota - (Membro)

Prof. Dr. Helton Saulo Bezerra dos Santos - UnB - (Membro Externo)

Prof. Dr. Manoel Ferreira dos Santos Neto - UFC - (Membro Externo)

Manaus, 06 de Maio de 2024

Documento assinado eletronicamente por **Jeremias da Silva Leão**, **Professor do Magistério Superior**, em 06/05/2024, às 12:17, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Agatha Sacramento Rodrigues**, **Usuário Externo**, em 06/05/2024, às 12:21, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Helton Saulo Bezerra dos Santos**, **Usuário Externo**, em 06/05/2024, às 12:57, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Alex Leal Mota**, **Professor do Magistério Superior**, em 06/05/2024, às 13:14, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Manoel Ferreira dos Santos Neto**, **Usuário Externo**, em 06/05/2024, às 16:50, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufam.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2036120** e o código CRC **3EE47F75**.

Av. General Rodrigo Octávio, 6200 - Bairro Coroado 1 Campus Universitário Senador Arthur Virgílio Filho, Setor Norte - Telefone: (92) 3305-1181 / Ramal 2405
CEP 69080-900, Manaus/AM, pos-matematica@ufam.edu.br

Referência: Processo nº 23105.019363/2024-21 SEI nº 2036120

To my parents.

Acknowledgements

I would like to express my deepest gratitude to God for giving me everything. To Him be all the honor, glory, and praise. Thanks to Our Lady of Aparecida, through her precious intercession.

To my advisors, Professors Jeremias Leão and Marcelo Bourguignon, thank you for your valuable guidance, respect, and support over the last four years. You always encouraged me and made me believe in my capabilities. Your efforts have truly paid off.

A sincere thank you to Professor Diego Gallardo for your invaluable support and generosity in sharing your academic knowledge. You were essential for this thesis to be carried out.

I want to express my heartfelt gratitude to Professor Celso Rômulo for absolutely everything. Your guidance has been instrumental in both my professional and personal development.

A sincere thank you to Professor Max Sousa for kindly dedicating your time to preparing me for the doctoral selection process. Your encouragement was fundamental to me.

I extend my heartfelt gratitude to all my teachers, from elementary school to higher education, who have contributed to this achievement. Especially to professors Amazoneida Sá Peixoto, Irapuan Pinheiro, James Dean, José Mir, José Raimundo and Themis Abensur.

I am profoundly grateful to my parents, Maria Esteva and Lindoval Francisco, for their endless love and sacrifices throughout all these years. To my brother, Márcio, thank you for all the support and friendship.

To my husband, Leonardo Martins, my love and my best friend. To our little cats: Arya, Pepê, Robb, Luna, Bolt, Ariel, Zangado, Tigrinha, Pingo, and Bolinha, thank you for the companionship.

To my friends Anderleuza, Camila, Carina, and Diego, thank you for your support, prayers, and companionship during the most difficult times.

Good does not spring from evil, any
more than figs grow from olive-trees:
the fruit corresponds to the seed.
(Sêneca)

Resumo

Em análise de sobrevivência, os modelos de fração de cura são ferramentas fundamentais em aplicações onde uma parcela significativa dos indivíduos estudados nunca experimentará o evento de interesse, mesmo se observados durante um longo período de tempo. Esses modelos assumem implicitamente que todos os indivíduos sob estudo pertencem a uma população homogênea e incluem a suposição da existência de uma variável aleatória não-observada, representando informações não diretamente disponíveis nos dados. Este trabalho está dividido em três capítulos, em que no primeiro apresentamos uma introdução aos modelos de fração de cura. Nos capítulos seguintes abordamos novas metodologias desenvolvidas neste trabalho no contexto de modelos de análise de sobrevivência com fração de cura, considerando a distribuição Weibull para o tempo de vida. No segundo capítulo, nossa proposta é estender o modelo de fração de cura com causas competitivas em séries de potência, assumindo uma mistura de duas causas competitivas provenientes dessa classe. A estimação dos parâmetros é discutida através do método da máxima verossimilhança, via o algoritmo do tipo EM (Expectation-Maximization). Estudos de Monte Carlo foram conduzidos para avaliação das propriedades assintóticas. Ilustramos nossa metodologia por meio de uma aplicação a um conjunto de dados reais de um estudo populacional de casos incidentes de melanoma cutâneo diagnosticados no estado de São Paulo, Brasil. No terceiro capítulo deste trabalho, apresentamos uma nova modelagem via fração de cura considerando que o número de causas competitivas para o evento de interesse segue um mistura das distribuições Poisson e Birnbaum-Saunders. Algumas propriedades estatísticas são apresentadas. A estimação dos parâmetros é conduzida através do método da máxima verossimilhança, utilizando um algoritmo do tipo EM (Expectation-Maximization). Ensaios de Monte Carlo são estudados para avaliar as propriedades assintóticas, bem como um estudo do poder do teste da razão de verossimilhanças. Uma aplicação é discutida utilizando dados reais de um estudo populacional de casos incidentes de câncer de mama no estado de São Paulo, Brasil.

palavras-chave: Misturas; Distribuição em série de potências, Poisson, Birnbaum-Saunders, causas concorrentes; Algoritmo EM; Melanoma; Câncer de Mama.

Abstract

In survival analysis, cure fraction models are fundamental in applications where a significant portion of the individuals studied will never experience the event of interest, even if observed over a long period of time. These models implicitly assume that all individuals under study belong to a homogeneous population and include the assumption of the existence of an unobserved random variable, representing information not directly available in the data. This work is divided into three chapters, in which in the first we present an introduction to the cure rate models. In the following chapters we address new methodologies developed in this work in the context of survival analysis models with cure fraction, considering the Weibull distribution for lifetime. In the second chapter our proposal is to extend the cure fraction model with competitive causes in Power Series assuming a mixture of two competitive causes belonging from this class. This mixture includes several well-known models as special cases. The estimation of parameters is discussed using the maximum likelihood method, with the proposition of an EM (Expectation-Maximization) type-algorithm. Monte Carlo studies were conducted to evaluate the asymptotic properties. We illustrate our methodology through an application to a set of medical data from a population study of incident cases of cutaneous melanoma diagnosed in the state of São Paulo, Brazil. In the third chapter, we present a new modeling via cure fraction considering that the number of competing causes for the event of interest follows a mixture of the Poisson and Birnbaum-Saunders distributions. Some statistical properties are presented, especially that the promotion time model appears as a limiting case. Parameter estimation is conducted using the maximum likelihood method, in which an EM (Expectation-Maximization) type-algorithm is proposed for this purpose. Monte Carlo experiment are studied to evaluate the asymptotic properties, as well as a study of the power of likelihood ratio test. An application is discussed using real data from a population study of incident cases of breast cancer in the state of São Paulo, Brazil.

keywords: Mixtures; Power series distribution, Poisson, Birnbaum-Saunders, competing causes; EM algorithm; Melanoma; Breast cancer.

List of Figures

2.1 Representation of the mixture model in a diagrammatic form for each subject in the population. 31
2.2 Variance of the number of competing causes as a function of the cure rate for some models with (left) and without (right) mixing. 33
2.3 Randomized quantile residuals (left) for mixture of concurrent causes for Bernoulli and Geometric model. Estimated SF (center) and $P(\text{Cured}|T \geq t)$ (right) for BER-GEO and BELL models to patients with (Profile 1) and without (Profile 2) chemotherapy, and both profiles considering that patients have made surgery, radiotherapy, full primary school, were women and have Stage 1 of disease. 48

3.1 Estimated SF obtained from the Kaplan-Meier (KM) estimator for overall patients diagnosed with breast cancer, by clinical stage, surgery, radiotherapy, chemotherapy and combinations of treatments. 78
3.2 Normalized randomized quantile residuals for PBS mixture applied to breast cancer dataset. Estimated SF for PBS Mixture, for patients with 20, 56 and 70 years old, who underwent radiotherapy and chemotherapy through stages of disease: Stage I (black), Stage II (red), Stage III (green) and Stage IV (blue). 83

List of Tables

2.1 Special cases for the distributions considered in the mixture of competing causes scheme. 32
2.2 AIC and BIC criteria for cure rate models with and without mixture of concurrent causes on PS applied to melanoma dataset. 45
2.3 ML estimates and SE obtained by fitting the four best combinations of PSMCC model and for BELL model to melanoma dataset. 46
2.4 Hypothesis test for $\gamma = 0$ and $\gamma = 1$ in different combinations for PSMCC model in melanoma data set. 47
2.5 Empirical bias, SE, RMSE and CP of the ML estimators for the Weibull distribution to time-to-event in the concurrent causes regression. 54
2.6 Empirical bias, SE, RMSE and CP of the ML estimators for the Weibull distribution to time-to-event in the concurrent causes regression. 55

3.1 Empirical standard deviation (SD), Bias, Root of MSE and CP of the ML estimators for PBS mixture model using the Weibull distribution to time-to-event in the concurrent causes regression. 73

3.2 Empirical Bias, Root of MSE, standard deviation (SD) and CP of the ML estimators for PBS mixture model using the Weibull distribution to time-to-event in the concurrent causes regression for variations of ϕ parameter. 75
3.3 Power (%) of LR Test for different values of ϕ and sample sizes. 76
3.4 AIC, BIC and BF values obtained by fitting the PBS mixture, NB, POI and BER (standard mixture) models to the breast cancer dataset. 80
3.5 ML estimates, standard error (SE) and respective p-value obtained by fitting of cure rate models for PBS mixture, NB, POI and BER (standard mixture) applied to breast cancer. 81
3.6 ML estimates of cure rate and 95% Confidence Interval (IC) obtained by Delta Method for PBS mixture cure rate model applied to breast cancer dataset through Stage of disease and treatments. 82
3.7 Empirical, Bias, Root of MSE, standard error (SE) and CP of the ML estimators for PBS mixture model using the Weibull distribution to time-to-event in the concurrent causes regression. 88
3.8 Empirical, Bias, Root of MSE, standard error (SE) and CP of the ML estimators for PBS mixture model using the Weibull distribution to time-to-event in the concurrent causes regression. 89

Contents

1 Introduction **15**
1.1 Objectives of the thesis 18
1.2 Organization of the chapters 19
1.3 Products of the thesis 19
References 21

2 Cure rate models for heterogeneous competing causes **24**
2.1 Introduction 26
2.2 Model based on the mixture competing causes 29
2.2.1 Model formulation 29
2.2.2 Special sub-models of the proposed model 34
2.3 Estimation of model parameters 36
2.4 Numerical applications 42
2.4.1 Simulation study 42
2.4.2 Application with a melanoma dataset 44
2.5 Concluding remarks 48
References 56

3 Poisson-Birnbaum-Saunders mixture cure rate model **59**
3.1 Introduction 61
3.2 The proposed model 63
3.2.1 Birnbaum-Saunders model (BS) 63
3.2.2 Poisson-Birnbaum-Saunders mixture model 65
3.3 Estimation 67
3.3.1 EM algorithm 68
3.4 Monte Carlo simulation studies 71
3.4.1 Asymptotic properties 72
3.4.2 Hypothesis testing 74
3.5 Application with breast cancer data 76
3.6 Concluding Remarks 84
References 90

CHAPTER 1

Introduction

In recent decades, studies on survival analysis have emerged as one of the areas with the most significant growth within statistics. This area of research is dedicated to analyzing data involving the time until the occurrence of a certain event of interest, such as death, cure or recurrence of diseases, during a certain period, that is, from a starting point to a predetermined end point. In addition to conveniently dealing with situations where the data is completely observed, due to the occurrence of censorship. However, in many situations, a significant portion of individuals will not experience the event of interest during the observation period and this fact results in the emergence of censorship in the data. In this case, the application of classical statistical techniques is negatively affected, as they require complete information about the failure time.

The incorporation of information from censored data in statistical data analysis makes survival analysis methods fundamental in the development of several areas of knowledge, with applications in different areas of science, including medicine, epidemiology, biology and public health studies. In medicine, for example, these models are often used to analyze the effectiveness of certain treatments, the progression of diseases and the patients survival, especially in the oncology field.

In several situations studied, in databases produced in the context of survival, a portion of individuals will never experience the event of interest, even after a long period of follow-up. In medical oncology studies, for example, a patient considered “cured” of cancer may not experience a recurrence of the tumor. In these cases, such individuals are considered “immune” to the event of interest, and the existence of a cure rate is assumed in this survival dataset to which these patients belong.

In this sense, models that deal with the existence of immune individuals are known as cure rate models, play an important role in survival analysis studies due to significant progress and advancements in treatment therapies, especially in scenarios where the presence of sampling units immune to the event studied. In clinical investigations, this event of interest may encompass several latent factors, causing the patient’s mortality under follow-up or the recurrence of malignant tumors, resulting from tumor cells capable of producing a metastatic tumor by remaining active after initial treatment, to these latent factors we call it concurrent causes or competing causes to the event of interest.

In the field of oncology, cure rate models are fundamental to modeling patients who are potentially cured of a certain type of cancer and those who are still at risk of disease recurrence. These models are particularly useful for understanding the long-term impact of treatments on different types of cancer, such as cutaneous melanoma and breast cancer. Various statistical models for studying the number of cancer cells have been used to analyze the lifetime of patients with an oncological diagnosis with the aim of estimating the fraction of patients considered cured after a certain period without recurrence of the disease, because it can analyze how specific variables for each type of cancer influence the probability of cure. Among multimodal interventions, it is possible to mention the combined use of surgery, radiotherapy, chemotherapy and target-directed therapies or immunotherapies. All of these advances allow these models to become more accurate and adaptive, in reflecting improvements in treatment and variations in patient responses. By providing a more accurate estimate of the fraction of patients who are potentially cured, these models help physicians personalize treatments and provide prognoses based on solid evidence, improving clinical decisions and optimizing healthcare resources.

Many studies have contributed to the theory of cure rate models, among which

Boag (1949) is the pioneer, in which the maximum likelihood method was used to estimate the proportion of survivors in a population of 121 women with breast cancer followed during 14 years. Based on Boag's idea, Berkson and Gage (1952) proposed a mixture model with the objective of estimate the proportion of people cured in a population undergoing stomach cancer treatment. The latter emerging as perhaps the best-known type of cure rate model. However, this model operates under the assumption that only a single cause is responsible for the event of interest. Despite this, in clinical studies, the event of interest, such as patient death, often arises from several competing latent causes, complicating identification of the precise causal factor. Furthermore, tumor recurrence, another crucial event in clinical research, can be attributed to the persistence of tumor cells capable of metastasis after initial treatment. In this sense, more complex long-term models, such as Yakovlev and Tsodikov (1996), Chen et al. (1999) and Ibrahim et al. (2001) among others, emerged with the aim of better explaining the biological effects involved, in which its structure is based on the assumption that the cumulative hazard function is bounded because of the existence of cured individuals.

This complexity underscores the need for models that can accommodate multiple latent competing causes. Their methodology operates under the premise that each individual harbors an unobservable (latent) quantity, denoted as M, of cells, with each possessing the capability to initiate the event of interest. Referred to in the literature as the promotion time cure rate model, this approach has garnered significant attention in research circles.

Rodrigues et al. (2009) proposed a comprehensive unification approach to long-term survival modeling, which has since inspired numerous extensions and applications in cure modeling, these models share a common underlying assumption: the initial cells are responsible for triggering the event of interest. The literature addressing modeling techniques that can effectively handle multiple latent competing causes is vast and continuously evolving. Notable contributions in this domain include significant insights from Castro et at. (2009) which delved into a cure rate model within a Bayesian framework, employing a negative binomial distribution to account for competing causes of the event of interest. Building upon this, Castro et at. (2010) utilized the Generalized Additive Models for Location, Scale, and Shape (GAMLSS) framework to fit long-term

survival models. They further expanded on this with a flexible Bayesian approach in Cancho et at. (2011). Cancho et at. (2012) offered another perspective by modeling cure rate survival, assuming competing causes follow a geometric distribution and event times adhere to a Birnbaum-Saunders distribution. Cancho et al. (2013) proposed The Power Series Cure Rate Model, with estimation via direct maximization of likelihood function. Ortega et al. (2015) presented A power series beta weibull regression model for predicting breast carcinoma. Rodrigues et al. (2016) presented a relaxed cure rate model, extending the Poisson cure rate model with an additional parameter for enhanced flexibility. Gallardo et al. (2017) gives to the Cancho et al. (2013) a simplified estimation procedure based on the em algorithm for the power series cure rate model, using maximum likelihood estimation in closed form via Expectation-Maximization (EM) algorithm. Among these advancements are notable works by various researchers. Yule-Simon (Gallardo et al., 2017); Polylogarithm (Gallardo et al., 2018); Zero-modified Geometric (ZMG) (Leão et al., 2020), compound Poisson (Gómez et al., 2023).

In this present work, we consider dataset from retrospective studies in oncological field provided by the Oncology Foundation of São Paulo (FOSP) which is a public institution associated with the State Health Secretariat that is responsible for the coordination of the state's Hospital Cancer Registry. As cited in (De andrade et al., 2012), the FOSP assists in the preparation and implementation of healthcare policies in the field of oncology, and serves as an instrument so that oncology hospitals can prepare protocols and improve care practices. With the aim of applying the new methodologies developed in this work, two datasets from FOSP are considered. In the second chapter we study a dataset related to cutaneous melanoma cancer and in the third chapter a dataset on breast cancer is studied.

1.1 Objectives of the thesis

In recent years, a variety of models addressing competing causes have emerged to improve estimation techniques in cure rate models. The primary aim of this thesis is to propose diverse statistical models for survival data, focusing on cure rates and based on a mixture of competing causes. The specific objectives include:

- To study concurrent causes supposing a mixture of competing causes in the class of power series distribution;
- To study competitive causes assuming that they come from a mixture of Poisson and Birnbaum-Baunders distributions.

1.2 Organization of the chapters

This thesis is organized is organized as follows. In Chapter 2, we presented the cure rate models for heterogeneous competing causes, a new methodology that extends the proposal of Cancho et al. (2013) to a mixture of competing causes and its properties and several important characteristics, dealing with estimation via EM-algorithm with close form for E-step, using the same idea propose by Gallardo et al. (2017), applying it to a melanoma cutaneous data.

In Chapter 3, we introduced a novel modeling approach for cure rate models, named the Poisson-Birnbaum-Saunders mixture cure rate model. This model builds upon the Poisson-Birnbaum-Saunders mixture model proposed by Gonçalves et al. (2022) within the context of cure rate models. Our proposition is inspired by the work of Barreto-Souza (2015), where it was assumed that one of the components of the mixture belongs to the Exponential Family (EF). We extend this assumption by replacing the EF with the Birnbaum-Saunders (BS) distribution, Birnbaum and Saunders (1969). We discuss an application to a breast cancer dataset.

1.3 Products of the thesis

This thesis allowed the following products to be obtained:

- Brandão, M., Leão, J., Gallardo, D., and Bourguignon, M. (2023). Cure rate models for heterogeneous competing causes. *Statistical Methods in Medical Research.* 32:9, 1823-1841.

- Gallardo, D., Brandão, M., Leão, J., Bourguignon, M., Calsavara, V. (2024) A new mixture model with cure rate applied to breast cancer data. *Biometrical Journal.* (Accepted for publication, in press.)

References

De Andrade, C. T., Magedanz, A., Escobosa, D. M. et al. (2012). The importance of a database in the management of healthcare services. *Einstein (São Paulo)*, 10:360–365.

Barreto-Souza, W. (2015). Long-term survival models with overdispersed number of competing causes. *Computational Statistics and Data Analysis*, 91(1):51–63.

Berkson, J. and Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47(259):501–515.

Birnbaum, Z. W. and Saunders, S. C. (1969). A new family of life distributions. *Journal of Applied Probability*, 6:319–327.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society B*, 11:15–53,

Cancho, V. G., Louzada, F., Barriga, G. D. C. (2012). The geometric Birnbaum-Saunders regression model with cure rate. *Journal of Statistical Planning and Inference*, 142:993–1000.

Cancho, V. G.; Rodrigues, J.; Castro, M. de. (2011). A flexible model for survival data with a cure rate: a Bayesian approach. *Journal of Applied Statistics*, 38:57–70.

Cancho, V. G., Louzada, F. and Ortega, E. M. The Power Series Cure Rate Model: An Application to a Cutaneous Melanoma Data. (2013). *Communications in Statistics-Simulation and Computation*, 42(3):586–602.

CASTRO, M. de; Cancho, V. G.; Rodrigues, J. (2009). A Bayesian long-term survival model parametrized in the cured fraction. *Biometrical Journal*, 51:443–455,

Castro, M. de, Cancho, V. G., Rodrigues, J. (2010). A hands-on approach for fitting long-term survival models under the GAMLSS framework. *Comput Methods Programs Biomed*, 97(2):168–77.

Chen, M. H., Ibrahim, J., and Sinha, D. (1999). A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, 94(447):909–919.

Gallardo, D. I., Gómez, H. W., and Bolfarine, H. (2017a). A new cure rate model based on the Yule–Simon distribution with application to a melanoma data set. *Journal of Applied Statistics*, 44(7):1153–1164.

Gallardo, D. I., Romeo, J. S., and Meyer, R. (2017b). A simplified estimation procedure based on the em algorithm for the power series cure rate model. *Communications in Statistics-Simulation and Computation*, 46(8):6342–6359.

Gallardo, D. I., Gómez, Y. M., and de Castro, M. (2018). A flexible cure rate model based on the polylogarithm distribution. *Journal of Statistical Computation and Simulation*, 88(11):2137–2149.

Gonçalves, J., Barreto-Souza, W., and Ombao, H. (2022). Poisson-Birnbaum-Saunders regression model for clustered count data. *https://doi.org/10.48550/arXiv.2202.10162*.

Gómez, Y., Gallardo, D., Bourguignon, M., Bertolli, E., and Calsavara, V. (2023). A general class of promotion time cure rate models with a new biological interpretation. *Lifetime Data Analysis*, 29:66–86.

Ibrahim, J. G., Chen, M. H. and Sinha, D. (2001). *Bayesian survival analysis*, volume 2. Springer.

Leão, J., Bourguignon, M., Gallardo, D. I., Rocha, R., and Tomazella, V. (2020). A new cure rate model with flexible competing causes with applications to melanoma and transplantation data. *Statistics in Medicine*, 39(24):3272–3284.

Ortega, E. M., Cordeiro, G. M., Campelo, A. K., Kattan, M. W., Cancho, V. G. (2015). A power series beta weibull regression model for predicting breast carcinoma. *Statistics in medicine*, 34(8):1366–1388.

Rodrigues, R., Cancho, V., De Castro, M., and Louzada-Neto, F. (2009c). On the unification of long-term survival models. *Statistics and Probability Letters*, 79(6):753–759.

Rodrigues, J., Cordeiro, Gauss. M., Cancho, V. G. and Balakrishnan, N. (2016). Relaxed Poisson cure rate models *Biometrical Journal (2016)* 58:397–4157.

Yakovlev, A. Y. and Tsodikov, A. (1996). *StochasticModels of Tumor Latency and their Biostatistical Applications*. World Scientific, New Jersey,US.

CHAPTER 2

Cure rate models for heterogeneous competing causes

Resumo

Os modelos de fração de cura têm sido amplamente estudados para analisar dados de tempo de vida com uma fração curada de pacientes, sob a abordagem de causas competitivas. Neste tipo de modelo, o número de causas competitivas (uma variável latente) é assumido como uma variável aleatória. Contudo, na prática, é natural assumir que esta é diferente de indivíduo para indivíduo. Nossa proposta é assumir que o número de causas concorrentes pertence a uma classe de mistura finita de distribuições de causas competitivas. Em particular, assumimos que o número de células malignas (causas concorrentes do evento de interesse) segue uma mistura de duas distribuições de séries de potências e assumimos que o tempo de falha é proveniente da distribuição Weibull. Consideramos a proporção do número de curados dependendo das covariáveis, permitindo uma modelagem direta da fração de cura. O modelo proposto inclui vários modelos bem conhecidos como casos especiais e define muitos novos modelos especiais (pelo menos dez novos casos). Alguns modelos especiais da classe proposta são discutidos detalhadamente. A estimação

dos parâmetros do modelo proposto é discutida por meio do método de estimação de máxima verossimilhança. Um algoritmo do tipo EM (Expectation-Maximization) é proposto para estimar os parâmetros, onde a etapa de esperança envolve o cálculo do número esperado de causas concorrentes para cada indivíduo. Um estudo de simulação foi proposto com a finalidade de examinar a recuperação dos valores dos parâmetros originais e a probabilidade de cobertura dos intervalos de confiança, com uma discussão dos resultados obtidos. A fim de mostrar o potencial do nosso modelo na prática, nós o aplicamos ao conjunto real de dados médicos de um estudo populacional de casos incidentes de melanoma cutâneo diagnosticados no estado de São Paulo, Brasil, ilustrando o fato de que o modelo proposto pode superar os modelos alternativos tradicionais em termos de ajuste.

palavras-chave: Causas concorrentes; Algoritmo EM; Melanoma; Misturas; Distribuição em série de potências.

Abstract

Cure rate models have been widely studied to analyze time-to-event data with a cured fraction of patients under competitive causes approach. In this type of model, the number of concurrent causes (a latent variable) is assumed to be a random variable. However, in practice, it is natural to assume that the distribution of the number of competing causes is different from individual to individual. Our proposal is to assume that the number of competing causes belongs to a class of a finite mixture of competing causes distributions. In particular, we assume the number of malignant cells (competing causes of the event of interest) follows a mixture of two power series (PS) distributions and assume that the time to the event of interest follows a Weibull distribution. We consider the proportion of the cured number of competing causes depending on covariates, allowing direct modeling of the cure rate through covariates. The proposed model includes several well-known models as special cases and defines many new special models (at least ten new special cases). Some special models of the proposed class are discussed in detail. The parameter estimation of the proposed model is discussed

through the maximum likelihood method. An EM algorithm is proposed for parameter estimation, where the expectation step involves the computation of the expected number of concurrent causes for each individual. A simulation study has been carried out to examine the parameter recovery and coverage probabilities of the confidence intervals with a discussion of the obtained results. In order to show the potential for the practice of our model, we apply it to the real medical data set from a population-based study of incident cases of cutaneous melanoma diagnosed in the state of São Paulo, Brazil, illustrating the fact that the model proposed can outperform traditional alternative models in terms of model fitting.
keywords: Concurrent causes; EM algorithm; Melanoma data set; Mixtures; Power series distribution.

2.1 Introduction

With the development of medical and health sciences and new treatments in recent years, it is expected that a proportion of patients responds favorably to a treatment, thus improving overall survival. This proportion of patients is commonly known as cure fraction. The models that can accommodate this feature are known in the literature as long-term survival or cure rate models. It should be noted that these models do not apply to overall survival because if a patient is cured of a disease, she/he remains at risk of death from other diseases, being never possible to cure her/him from all diseases.

In these models, it can be assumed that the occurrence of the event of interest might be a result of many competing causes, with the number of causes as well as survival times associated with each cause being unknown, which leads to the so called latent competing causes, and is assumed to be a random variable (not observable) following some discrete distribution. The latent competing causes can be assigned to metastasis-competent tumor cells left active after initial treatment, such as radiotherapy, chemotherapy, surgery, among others; see, for example, Ibrahim et al. (2001). According to Ortega et al. (2015), in a biological context, the idea behind these assumptions lies within a latent competing cause structure, in the sense that the event of interest can be

a tumor recurrence or the death of a patient, which can happen because of unknown competing causes. From this, several authors investigated the implications in the study of competitive causes. Two formulations of cure rate models have received attention in the literature, namely the mixture cure model by (Berkson and Gage, 1952), and the promotion time cure model by (Yakovlev and Tsodikov, 1996). The well-known mixture cure model presumes the number of latent causes to follow a Bernoulli (BER) distribution with at most one latent cause, while in the promotion time cure modeling this number follows a Poisson (POI) distribution.

Rodrigues et al. (2009) introduced a unified approach for long-term survival models by assuming that the number of competing causes associated with the event of interest follows any positive discrete distribution that possesses a probability generating function (Feller, 2008). This was significant in expanding the range of distributions that can be assumed for the number of competing causes. For example, the negative binomial distribution includes the Bernoulli (BER), binomial (Bin), Poisson (POI), and Geometric (GEO) distributions as particular cases. Rodrigues et al. (2009a) used the Conway-Maxwell Poisson (COMP-Poisson) distribution. Cancho et al. (2013) proposed the PS cure rate model as flexible model for modeling survival data with cure fraction. Ortega et al. (2015) employed the PS beta Weibull distribution, Gallardo et al. (2017) considered the Yule-Simon distribution, Gallardo et al. (2018) considered the polylogarithm distribution, Leão et al. (2020) studied the zero-modified geometric distribution and Gallardo et al. (2021) employed the Bell (which we will denoted by BELL) distribution. However, all the models cited above considered that the distribution for the competing causes is the same for all the individuals in the study and then, it is assumed the same biological process for all the observations. However, given the heterogeneity of the individuals and mainly the absence of measurement of relevant information, in practice such assumption might not be satisfied. Despite this, to the best of our knowledge, a specific cure rate model assuming a different biological process for all the observations has never been considered in the literature.

Based on the above discussion, the main aim of this chapter is to propose a cure rate regression model that is tailored for situations where the number of competing causes is different from individual to individual based on discrete PS distribution. In

particular, we assume that the number of concurrent causes of the event of interest follows a mixture of two PS distributions by including a mixing additional parameter. The advantage of the proposed cure rate model is that the distribution of the number of competing is different from individual to individual, i.e., different biological process for all the observations. All the models cited above are not suitable for capturing this.

We note five motivations for the proposed model:

- We assume the number of malignant cells (competing causes of the event of interest) to follow mixture of two PS distributions. From a practical point of view, this generalization is based on the search for models that are more flexible in such a way that they fit better to the lifetime data. Furthermore, the wide usage of PS and the fact that the current generalization provides means of its continuous extension to still more complex situations. The PS distribution is very flexible, including several particular cases, such that, Bernoulli (the mixture model), Bell, geometric, logarithmic (LOG), and Poisson (the promotion time cure model), among others, and the probability generating function (PGF) can be expressed in a simple form;

- We consider the proportion of cured of competing causes depending on covariates, so allowing to a direct modeling of the cure rate through covariates. Thus, we obtain a straightforward interpretation of the regression coefficients in terms of the long-term survivors;

- The proposed model includes several well-known models as special cases, such as the mixture cure models by (Berkson and Gage, 1952) and promotion time cure rate model by (Yakovlev and Tsodikov, 1996) and defines many new special models (at least ten new special cases);

- The estimation and inference for the new model are possibly based on the likelihood paradigm (parametric approach), which can be easily computed using the R software (R Core Team, 2024). In particular, we provide a simple EM-algorithm which is more robust with regards to the estimation procedure, especially in situations with many covariates. Furthermore, the EM algorithm yields estimates of the number of latent causes for each individual and the coefficients of covariates

values, see (Gallardo et al., 2017). It has the big advantage that the maximization step can be decomposed into separate maximizations of two lower-dimensional functions of the regression and survival distribution parameters;

- The Monte Carlo simulations and empirical application show the good performance of the proposed model (see Subsection 2.4.2). In fact, the new models fit the data set well.

This chapter is organized as follows. Section 2.2 defines the proposed model and discusses how to obtain the class of power series mixture competing causes (PSMCC). In addition, some special cases are provided. The model parameters estimated via the EM algorithm and inference are both supplied in Section 2.3. In Section 2.4, we yield a numerical evaluation of the studied model where we evaluate the performance of the EM estimators by Monte Carlo (MC) simulations. In addition, we illustrate the proposed model and its diagnostics with a medical real-world data, comparing it to model studied by Gallardo et al. (2017). Some concluding remarks, and possible future studies are discussed in Section 2.5.

2.2 Model based on the mixture competing causes

In this section, we introduce the new cure rate model, its main properties and some special cases.

2.2.1 Model formulation

Let M_i be the number of competing causes related to the occurrence of an event of interest for the i-th individual in the population (in a cancer context M_i represents the carcinogenic cells of the individual), where $i = 1, \ldots, n$, and n denotes the sample size. In a competing causes scenario, the number of competing causes M_i is unobserved (latent) variables. It is natural to assume that the distribution of the number of concurrent

causes is different from individual to individual. In this context, a finite mixture of competing causes distributions can be used in order to describe situations in which one (or more) of the latent variables separates the population in study into two (or more) sub-populations. Here, we assume the distribution of M_{ji} in the sub-population j, with $j = 1, 2$, for each individual i is $P(M_{ji} = m_{ij})$. These distributions do not have the same functional form and different parameters, i.e., we assume that the number of competing causes M_i follows a mixture of two random variables M_{1i} and M_{2i} in $\mathbb{N}$, i.e., $M_i = \gamma M_{1i} + (1-\gamma)M_{2i}$, where $\gamma \in [0, 1]$ is the mixing probability. Let W_k, $k = 1, \ldots, M_i$ be the time to the k-th cell produces a detectable cancer. The W_k's are supposed to be independent and identically distributed with common (proper) survival function (SF) $S(t_i; \boldsymbol{\eta})$ and probability density function (PDF) $f(t_i; \boldsymbol{\eta})$, where $\boldsymbol{\eta}$ denotes a vector of unknown parameters. Moreover, we assume that M_{ji}, is independent of $W_1, \ldots, W_{m_{ij}}$. The observable time-to-event is defined as $T_i = \min\{W_1, \ldots, W_{M_i}\}$, for $M_i \geq 1$, and $T_i = \infty \text{ for } M_i = 0$, leading to a cured fraction denoted by p_i. Figure 2.1 illustrates this interpretation

Consider two random variables M_{1i} and M_{2i} in $\mathbb{Z}_+$ non-negative integers, with $P(M_{1i} = m_{1i}; p_i)$ and $P(M_{2i} = m_{2i}; p_i)$ probability mass functions, respectively, such that $P(M_i = 0; p_i) = p_i$. The probability function of the mixture components M_i has $\mathbb{Z}_+$, a subset of non-negative integers, as support. Then the random variable M_i has probability mass function given by

$$P(M_i = m_i; p_i) = \gamma\, P(M_{1i} = m_i; p_i) + (1-\gamma)\, P(M_{2i} = m_i; p_i), \quad m_i = 0, 1, 2, \ldots,$$

where $0 \leq \gamma \leq 1$ is the mixing parameter. Using the mixing parameter, it is possible to obtain the population proportion of the number of cancerous cells that follows the M_1 or M_2 distributions and the two specific models are particular cases for $\gamma = 0$ and $\gamma = 1$, respectively. By construction, it is immediate that $P(M_i = 0; p_i) = p_i$. In a cure rate models context, this is very important because $P(M_i = 0; p_i)$ is interpreted as the cure rate and then, if the model is parameterized directly in terms of this expression, then covariates can be introduced directly through it. The PGF of M_i, defined as

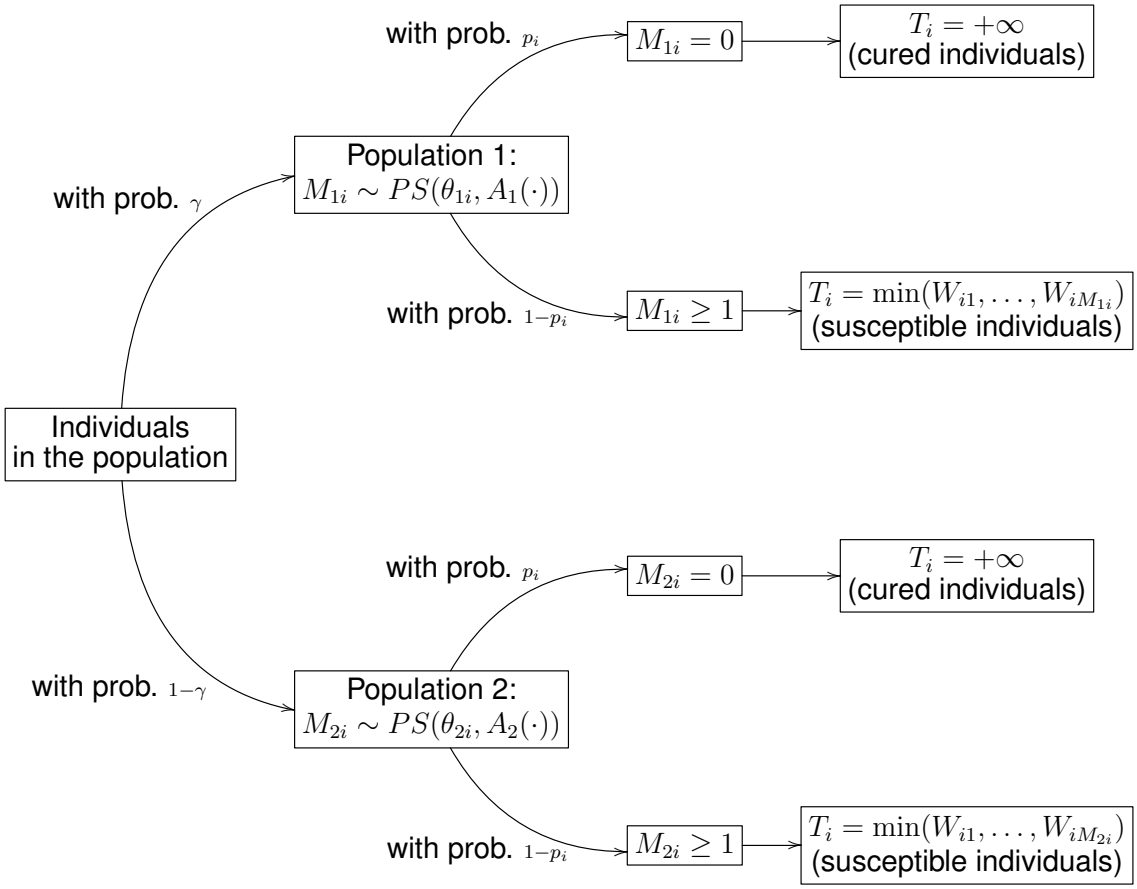

Figure 2.1: Representation of the mixture model in a diagrammatic form for each subject in the population.

$\varphi_{M_i}(s_i; p_i, \gamma) = \mathbb{E}[s_i^{M_i}; p_i]$, is given by

$$\varphi_M(s_i; p_i, \gamma) = \gamma\, \varphi_{M_{1i}}(s_i; p_i) + (1-\gamma)\varphi_{M_{2i}}(s_i; p_i).$$

In this work, we focused in the PS distributions for M_{ji}, for $j = 1, 2$ and $i = 1, 2, \ldots, n$; because (i) it is a class including many well known distributions in the literature; (ii) many of its particular cases were parameterized directly in the cure rate in the literature and; (iii) a nice EM algorithm can be applied to the model to perform parameter estimation. Specifically, we say that M_{1i} and M_{2i} have a PS$(\theta_{1i}, A_1(\cdot))$ and PS$(\theta_{2i}, A_2(\cdot))$ distribution, respectively, such that

$$P(M_{ji} = m_{ji}; \theta_{ji}) = \frac{a_j(m_{ji})\, \theta_{ji}^{m_{ji}}}{A_j(\theta_{ji})}, \quad m_{ji} \in \mathcal{S},\ \theta_{ji} \in (0, s). \tag{2.1}$$

Table 2.1: Special cases for the distributions considered in the mixture of competing causes scheme.

Distribution	$a_j(m)$	$A_j(\theta)$	$A_j^{'}(\theta)$	$A_j^{-1}(\theta)$	$(0,s)$	$\mathbb{E}_\theta[M^d]$ for $d=1,2$
Bernoulli	$I_{\{0,1\}}^{(m)}$	$1+\theta$	1	$\theta-1$	$(0,\infty)$	$\frac{\theta}{(1+\theta)}$
Poisson	$(m!)^{-1}$	$\exp(\theta)$	$\exp(\theta)$	$\log(\theta)$	$(0,\infty)$	$\theta+(1-d)\theta^2$
Geometric	1	$(1-\theta)^{-1}$	$(1-\theta)^{-2}$	$1-1/\theta$	$(0,1)$	$\frac{\theta}{1-\theta}\left(\frac{1+\theta}{1-\theta}\right)^{d-1}$
Logarithmic	$(m+1)^{-1}$	$-\frac{\log(1-\theta)}{\theta}$	$\frac{\log(1-\theta)}{\theta^2}-\frac{1}{\theta(1-\theta)}$	$1+\frac{W\left(-\theta\,\mathrm{e}^{-\theta}\right)}{\theta}$	$(0,1)$	$\frac{(-1)^d\theta(1-2\theta)^{d-1}}{(1-\theta)^d\log(1-\theta)}+(-1)^d$
Bell	$D_m/m!$	$\exp(\mathrm{e}^\theta-1)$	$\exp(\mathrm{e}^\theta-1+\theta)$	$\log(1+\log(\theta))$	$(0,\infty)$	$\theta\,\mathrm{e}^\theta(1+\theta(1+\mathrm{e}^\theta))^{d-1}$

NOTE: $D_m=e^{-1}\sum_{k=0}^{\infty}k^m/k!$ denotes the Bell numbers. $W(\cdot)$ denotes the Lambert function.

The support $\mathcal{S}$ of M_{ji} in (2.1) is a subset of $\mathbb{Z}_+$, $a_j(m_{ji})\geq 0$ depends only on m_{ji}, and there is $s>0$ such that the normalizing constant $A_j(\theta_{ji})=\sum_{m_{ji}=0}^{\infty}a_j(m_{ji})\,\theta_{ji}^{m_{ji}}$ is finite for all $\theta_{ji}\in(0,s)$ (s can be ∞). Although we will always consider θ_{ji} as a value in $(0,s)$, we will also assume that the PS for $A_j(\theta_{ji})$ converges, in fact, to a finite value for $\theta_{ji}\in(-s,s)$. If this is the case, then, $A_j(\theta_{ji})$ has derivatives of all orders in $(-s,s)$ and those derivatives can be obtained by differentiating the PS term to term. Since $a_j(m_{ji})\geq 0$ for all m_{ji}, $A_j(\theta_{ji})$ and all its derivatives will be positive in $(0,s)$. When $0\in\mathcal{S}$ and, from (2.1), $p_i=\Pr(M_{ji}=0;\theta_{ji})=a_j(0)/A_j(\theta_{ji})$. Thus, $\theta_{ji}=\theta_{ji}(p_i)=A_j^{-1}(a_j(0)/p_i)$. For more detail on the PS class of distributions, one can refer to Noak (1950).

The corresponding mean and variance of M_{ji} are, respectively, represented as

$$\mathbb{E}_\theta(M_{ji})=\theta_{ji}\frac{A_j^{'}(\theta_{ji})}{A_j(\theta_{ji})}\quad\text{and}\quad\mathrm{Var}(M_{ji})=\theta_{ji}^2\left(\frac{A_j^{''}(\theta_{ji})}{A_j(\theta_{ji})}-\frac{A_j^{'}(\theta_{ji})}{A_j(\theta_{ji})}\right).$$

The PGF of M_{1i} and M_{2i} are given by $\varphi_{M_{1i}}(s_i;p_i)=A_1(\theta_{1i}s_i)/A_1(\theta_{1i})$ and $\varphi_{M_{2i}}(s_i;p_i)=A_2(\theta_{2i}s_i)/A_2(\theta_{2i})$, respectively, and then

$$\varphi_{M_i}(s_i;p_i,\gamma)=\gamma\,\frac{A_1(\theta_{1i}s_i)}{A_1(\theta_{1i})}+(1-\gamma)\frac{A_2(\theta_{2i}s_i)}{A_2(\theta_{2i})}. \tag{2.2}$$

Remark 2.2.1. *When $\gamma=0$ or $\gamma=0$, the boundaries of γ's parameter space, the proposed model is reduced to the PS cure rate model (Ortega et al.,2015).*

Table 2.1 provides some results for special cases to distributions considered on

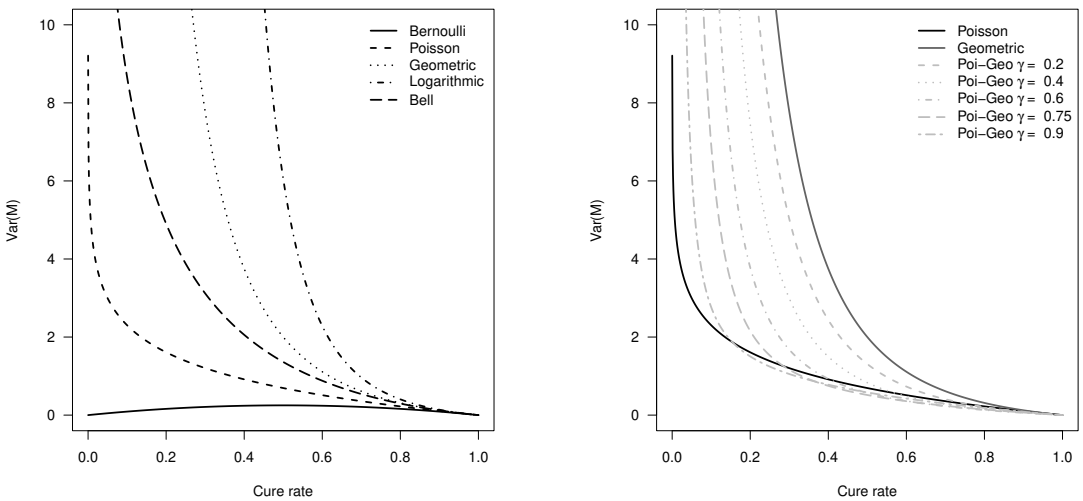

Figure 2.2: Variance of the number of competing causes as a function of the cure rate for some models with (left) and without (right) mixing.

the mixing. Observe that, as shown earlier, θ can be denoted in function of the cure rate through. Figure 2.2 presents the variance's behavior of the number of competing causes as a function of the cure rate for particular cases (left) and for two of these distributions by using Poisson and Geometric distribution as particular case and mixing both for different values of γ (right). Can be seen different behaviors for variances of particular cases. Observe that given two models with one component, the mixture of them presents variance's behavior limited almost everywhere by the ones for the particular case. Therefore, the variance of model for mixing concurrent causes is influenced by the fixed value of γ. If the value of this parameter is low, it can be seen that the variance of mixing model is next to the first component considered in the model, in this case the Geometric one. If the value for of γ is high, the variance of second component, in this case Poisson, has more weight on the estimation.

Following the relationship with PGF as in Rodrigues et al. (2009) (2009), the population SF and PDF are given by

$$S_{pop}(t_i; p_i, \boldsymbol{\eta}, \gamma) = \gamma \, \frac{A_1\left(\theta_{1i} S(t_i; \boldsymbol{\eta})\right)}{A_1\left(\theta_{1i}\right)} + (1-\gamma) \frac{A_2\left(\theta_{2i} S(t_i; \boldsymbol{\eta})\right)}{A_2(\theta_{2i})}, \tag{2.3}$$

and

$$f_{pop}(t_i; p_i, \boldsymbol{\eta}, \gamma) = f(t_i; \boldsymbol{\eta}) \left[\gamma\, \theta_{1i} \frac{A_1^{'}\left(\theta_{1i} S(t_i; \boldsymbol{\eta})\right)}{A_1\left(\theta_{1i}\right)} + (1-\gamma)\theta_{2i} \frac{A_2^{'}\left(\theta_{2i} S(t_i; \boldsymbol{\eta})\right)}{A_2(\theta_{2i})} \right]. \tag{2.4}$$

Remark 2.2.2. *Note that the inversion of coefficients $a_1(m_{1i})$ and $a_2(m_{2i})$ and the functions $A_1(\cdot)$ and $A_2(\cdot)$ in Equation (2.1) will produce the same PGF in* (2.2)*, with mixing probability $\gamma^* = 1 - \gamma$. Therefore, based on the examples in Table 2.1, we are introducing ten new cure rate models.*

2.2.2 Special sub-models of the proposed model

Following the results presented in Table 2.1, we can obtain ten new cure rate models. Below we present four of these sub-models. These cases were chosen according to the best results obtained in the application to real data presented in Subsection 2.4.2.

Bernoulli and Poisson: Consider $A_1(\theta_{1i}) = 1 + \theta_{1i}$ with $\theta_{1i} = 1/p_i - 1$ and $A_2(\theta_{2i}) = \exp(\theta_{2i})$ with $\theta_{2i} = \log(1/p_i)$ in (2.3) and (2.4). The population SF and PDF are given by

$$S_{pop}(t_i; p_i, \boldsymbol{\eta}, \gamma) = \gamma\left(p_i + (1-p_i)S(t_i; \boldsymbol{\eta})\right) + (1-\gamma)p_i^{(1-S(t_i;\boldsymbol{\eta}))},$$

and

$$f_{pop}(t_i; p_i, \boldsymbol{\eta}, \gamma) = f(t_i, \boldsymbol{\eta}) \left[\gamma(1-p_i) + (1-\gamma)\log(1/p_i)p_i^{(1-S(t_i;\boldsymbol{\eta}))}\right].$$

The BER-POI cure rate model includes the long-term survival models (Berkson and Gage, 1952) when $\gamma = 1$ and the promotion time cure rate model (Yakovlev and Tsodikov, 1996) when $\gamma = 0$.

Poisson and geometric: As a second example, consider $A_1(\theta_{1i}) = \exp(\theta_{1i})$ with $\theta_{1i} = \log(1/p_i)$ and $A_2(\theta_{2i}) = (1-\theta_{2i})^{-1}$ with $\theta_{2i} = 1 - p_i$. In this case, the population SF and PDF are given by

$$S_{pop}(t_i; p_i, \boldsymbol{\eta}, \gamma) = \gamma \times p_i^{1-S(t_i;\boldsymbol{\eta})} + (1-\gamma)\frac{p_i}{1 - S(t_i;\boldsymbol{\eta}) + p_i S(t_i;\boldsymbol{\eta})},$$

and

$$f_{pop}(t_i; p_i, \boldsymbol{\eta}, \gamma) = f(t_i, \boldsymbol{\eta}) \left[\gamma p_i \left(\log(1/p_i) \log\left[\log(1/p_i) S(t_i;\boldsymbol{\eta})\right]\right) - p_i^2\right]$$

Geometric and Bell: Now, let us consider $A_1(\theta_{1i}) = (1-\theta_{1i})^{-1}$ with $\theta_{1i} = 1 - p_i$ and $A_2(\theta_{2i}) = \exp(\exp(\theta_{2i}) - 1)$ with $\theta_{2i} = \log(1 - \log p_i)$. Then, we have that SF and PDF are given by

$$S_{pop}(t_i; p_i, \boldsymbol{\eta}, \gamma) = \gamma \frac{p_i}{1 - S(t_i;\boldsymbol{\eta}) + p_i S(t_i;\boldsymbol{\eta})} + (1-\gamma) p_i \exp\left[(1 - \log p_i)^{S(t_i;\boldsymbol{\eta})} - 1\right],$$

and

$$f_{pop}(t_i; p_i, \boldsymbol{\eta}, \gamma) = f(t_i, \boldsymbol{\eta}) \left\{\gamma \frac{p_i(1-p_i)}{(1 - S(t_i;\boldsymbol{\eta}) + p_i S(t_i;\boldsymbol{\eta}))^2} + (1-\gamma) p_i (1 - \log p_i) \right.$$
$$\left. \times \log(1 - \log p_i) \exp((1 - \log p_i)^{S(t_i;\boldsymbol{\eta})} - 1)\right\}.$$

Bernoulli and Bell: In the last example, we consider $A_1(\theta_{1i}) = 1 + \theta_{1i}$ with $\theta_{1i} = 1/p_i - 1$ and $A_2(\theta_{2i}) = \exp(e^{\theta_{2i}} - 1)$ with $\theta_{2i} = \log(1 - \log p_i)$. Therefore, the survival population function and probability are given by

$$S_{pop}(t_i; p_i, \boldsymbol{\eta}, \gamma) = \gamma \left(p_i + (1-p_i) S(t_i;\boldsymbol{\eta})\right) + (1-\gamma) p_i \exp\left[(1 - \log p_i)^{S(t_i;\boldsymbol{\eta})} - 1\right],$$

and

$$f_{pop}(t_i; p_i, \boldsymbol{\eta}, \gamma) = f(t_i, \boldsymbol{\eta}) \left\{ \gamma p_i(1-p_i) + (1-\gamma)p_i(1-\log p_i) \right.$$
$$\left. \times \log(1-\log p_i) \exp((1-\log p_i)^{S(t_i;\boldsymbol{\eta})} - 1) \right\}.$$

For heterogeneous populations we can incorporate explanatory variables into the parametric cure rate model through the cure parameter p. When these variables are incorporated, we have a different cure rate parameter for each patient, which is denoted by p_i, with $i = 1, \ldots, n$. In order to model the effect of the explanatory variables on the cure rate, we can use different link functions. Let $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_r)^\top$ be the vector of regression coefficients to be estimated of dimension $q = (r+1)$. Note that $\boldsymbol{\beta}$ is related to explanatory variables with observed values for the patient i denoted by $\mathbf{x}_i = (1, x_{1i}, \ldots, x_{ri})^\top$, which are associated with the cured fraction. Then, considering the link logit function, we obtain

$$\log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i^\top \boldsymbol{\beta} \quad \Leftrightarrow \quad p_i = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}, \quad i = 1, \ldots, n. \tag{2.5}$$

2.3 Estimation of model parameters

We assume that the data are obtained under a right censoring scheme. Thus, the observed data for the ith individual can be represented by $T_i = \min(T_i^*, C_i)$ and $\delta_i = I(T_i^* \leq C_i)$, for $i = 1, \ldots, n$, with T_i^* and C_i denoting failure and censoring time, respectively. Denote the observed data as $\mathrm{D}_{obs} = (\mathbf{t}, \boldsymbol{\delta}, \mathbf{x})$, with $\mathbf{t} = (t_1, \ldots, t_n)^\top$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^\top$, and $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$, where $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ir})^\top$ is the covariate vector of dimension $(r+1) \times 1$ related to the cure of the ith individual. For each individual, let the latent number of causes $M_i = \gamma M_{1i} + (1-\gamma) M_{2i}$, with $M_{ji} \sim$ PS$(\theta_{ji}, A(\theta_{ji}))$ and W_{ki} be independent and identical distribution non-negative random variables with SF S$(\cdot; \boldsymbol{\eta})$ for $k = 1, \ldots, M_i$, and $T_i = \min W_k$ for, $k = 0, \ldots, M_i$ and $i = 1, \ldots, n$. Also, let consider the vector of latent variables $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ indicating the group to which belongs each individual. The influence of the covariates $\mathbf{x}_i$ related to the cure

of the individuals p_i as in (2.5). On the other hand, $\mathbf{M}_j = (M_{j1}, \ldots, M_{jn})^\top$, for $j = 1, 2$, and $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ is non-observable and thus the complete data are denoted by $\mathsf{D}_{comp} = (\mathbf{t}, \boldsymbol{\delta}, \mathbf{x}, \mathbf{M}_1, \mathbf{M}_2, \mathbf{Y})$.

In order to obtain the estimates for $\boldsymbol{\psi} = (\boldsymbol{\eta}^\top, \boldsymbol{\beta}^\top, \gamma)$, with $\boldsymbol{\eta}^\top$, we can use the ML method by maximizing the observed likelihood function given by

$$\begin{aligned}
\ell(\boldsymbol{\psi}|\mathsf{D}_{obs}) &= \sum_{i=1}^{n} \{\delta_i \log[f_{pop}(t_i, \boldsymbol{\eta})] + (1-\delta_i)\log[S_{pop}(t_i, \boldsymbol{\eta})]\} \\
&= \sum_{i=1}^{n} \left\{ \delta_i \left[\log[f(t_i; \boldsymbol{\eta})] + \log\left(\gamma\, \theta_{1i} \frac{A_1'\left(\theta_{1i} S(t_i; \boldsymbol{\eta})\right)}{A_1\left(\theta_{1i}\right)} + (1-\gamma)\theta_{2i} \frac{A_2'\left(\theta_{2i} S(t_i; \boldsymbol{\eta})\right)}{A_2(\theta_{2i})} \right) \right] \right. \\
&\quad \left. + (1-\delta_i) \log\left(\gamma\, \frac{A_1\left(\theta_{1i} S(t_i; \boldsymbol{\eta})\right)}{A_1\left(\theta_{1i}\right)} + (1-\gamma) \frac{A_2\left(\theta_{2i} S(t_i; \boldsymbol{\eta})\right)}{A_2(\theta_{2i})} \right) \right\}. \qquad (2.6)
\end{aligned}$$

However, the maximization of (2.6) cannot be simple because the maximization procedure need to be performed for $\boldsymbol{\psi}$, i.e., for $q+3$ parameters. To obtain a simplified and more robust estimation procedure, we use a similar idea that Gallardo et al. (2017) based on the EM algorithm for the PS cure rate model. According to the results in the latter paper, even though the observed (marginal) likelihood is available in closed form and all-purpose maximization algorithms such as the Newton-Raphson algorithm exists that could be applied, convergence depends on the choice of starting values and is guaranteed only to local maxima. Adapting the results presenting in Appendix of Gallardo et al. (2017), we obtain

$$\begin{aligned}
f(t_i, \delta_i, m_{1i}, m_{2i}, y_i; p_i, \boldsymbol{\eta}, \gamma) &= \left\{ S(t_i; \boldsymbol{\eta})^{m_{1i}-\delta_i} [m_{1i} f(t_i; \boldsymbol{\eta})]^{\delta_i} \right\}^{y_i} \left\{ S(t_i; \boldsymbol{\eta})^{m_{2i}-\delta_i} [m_{2i} f(t_i; \boldsymbol{\eta})]^{\delta_i} \right\}^{1-y_i} \\
&\quad \times \frac{a_1(m_{1i})\theta_{1i}^{m_{1i}}}{A_1(\theta_{1i})} \frac{a_2(m_{2i})\theta_{2i}^{m_{2i}}}{A_2(\theta_{2i})} \gamma^{y_i} (1-\gamma)^{1-y_i},
\end{aligned}$$

where $\theta_{1i} = \theta_{1i}(p_i) = A_1^{-1}(a_0/p_i)$ and $\theta_{2i} = \theta_{2i}(p_i) = A_2^{-1}(b_0/p_i)$. Therefore, up to a constant that does not depend on $\boldsymbol{\psi} = (\boldsymbol{\eta}\boldsymbol{\beta}, \gamma)^\top$, the vector of parameters, the complete log-likelihood is given by

$$\ell_c(\boldsymbol{\psi} \mid \mathsf{D}_{comp}) = \ell_{1c}(\boldsymbol{\eta} \mid \mathsf{D}_{comp}) + \ell_{2c}(\boldsymbol{\beta} \mid \mathsf{D}_{comp}) + \ell_{3c}(\gamma \mid \mathsf{D}_{comp}), \qquad (2.7)$$

where

$$\ell_{1c}(\boldsymbol{\eta} \mid \mathsf{D}_{comp}) = \sum_{i=1}^{n} \left\{ [Y_i M_{1i} + (1 - Y_i) M_{2i} - \delta_i] \log[S(t_i; \boldsymbol{\eta})] + \delta_i \log[f(t_i; \boldsymbol{\eta})] \right\},$$

$$\ell_{2c}(\boldsymbol{\beta} \mid \mathsf{D}_{comp}) = \sum_{i=1}^{n} \left\{ M_{1i} \log(\theta_{1i}) - \log[A_1(\theta_{1i})] + M_{2i} \log(\theta_{2i}) - \log[A_2(\theta_{2i})] \right\}, \quad \text{and}$$

$$\ell_{3c}(\gamma \mid \mathsf{D}_{comp}) = \sum_{i=1}^{n} \left\{ Y_i \log \gamma + (1 - Y_i) \log(1 - \gamma) \right\}.$$

When the estimation involves latent variables or missing data, the EM algorithm is typically used to deal with the ML estimates of the parameters of interest by using incomplete data to facilitate the process. The EM algorithm was proposed by Dempster et al. (1977) and it uses the conditional distribution of the latent variables given the data and current estimates of the parameters in the E-step to find the ML estimates of the parameters interactively. Then, this conditional expectation is maximized on the M-step to find ML of the unknown parameters.

The following proposition and corollary can be used to derive the formula for the E-step. The proofs of the latter are given in the Appendix A.

Proposition 2.3.1. *Define*

$$\mathbb{E}_{\theta_{ji} S(t_i; \boldsymbol{\eta})}[M_{ji}^d] = \sum_{m_{ji}=0}^{\infty} m_{ji}^d \frac{a_j(m_{ji})(\theta_{ji} S(t_i; \boldsymbol{\eta}))^{m_{ji}}}{A_j(\theta_{ji} S(t_i; \boldsymbol{\eta}))} \quad \textit{with} \quad d, j = 1, 2.$$

For the PSMCC, the conditional distribution of the number of latent initial causes $M_{ji} \mid y_i, t_i, \delta_i$*, for* $j = 1, 2$*, is given by*

$$P(M_{1i} = m_{1i} \mid y_i, t_i, \delta_i) = \frac{a_1(m_{1i}) \left(\theta_{1i} \left[S(t_i; \boldsymbol{\eta})\right]^{y_i}\right)^{m_{1i}}}{A_1\left(\theta_{1i} \left[S(t_i; \boldsymbol{\eta})\right]^{y_i}\right)} \left(\frac{m_{1i}}{\mathbb{E}_{\theta_{1i} S(t_i; \boldsymbol{\eta})}[M_i]} \right)^{y_i \delta_i}, \quad \textit{and}$$

$$P(M_{2i} = m_{2i} \mid y_i, t_i, \delta_i) = \frac{a_2(m_{2i}) \left(\theta_{2i} \left[S(t_i; \boldsymbol{\eta})\right]^{1-y_i}\right)^{m_{2i}}}{A_2\left(\theta_{2i} \left[S(y_i; \boldsymbol{\eta})\right]^{1-t_i}\right)} \left(\frac{m_{2i}}{\mathbb{E}_{\theta_{2i} S(t_i; \boldsymbol{\eta})}[M_i]} \right)^{(1-y_i)\delta_i}.$$

In addition, the distribution for $Y_i \mid t_i, \delta_i$ *is Bernoulli with success probability* $\omega_i/(1 + \omega_i)$*,*

where

$$\omega_i = \left(\frac{\mathbb{E}_{\theta_{1i}S[t_i;\boldsymbol{\eta}]}[M_{1i}]}{\mathbb{E}_{\theta_{2i}S[t_i;\boldsymbol{\eta}]}[M_{2i}]}\right)^{\delta_i} \frac{A_1(\theta_{1i}S[t_i;\boldsymbol{\eta}])A_2(\theta_{2i})}{A_1(\theta_{1i})A_2(\theta_{2i}S[t_i;\boldsymbol{\eta}])}\frac{\gamma}{(1-\gamma)}. \tag{2.8}$$

The following Corollary 2.3.1 presents the expectation of the conditional distribution of M for the case of censored and uncensored data.

Corollary 2.3.1. *The expected value for the number of initial causes M_{ji} given (y_i, t_i, δ_i), for $j = 1, 2$, can be written as*

$$\mathbb{E}[M_{1i} \mid y_i, t_i, \delta_i] = \left(\frac{\mathbb{E}_{\theta_{1i}S(t_i;\boldsymbol{\eta})}[M_{1i}^{1+\delta_i}]}{\mathbb{E}_{\theta_{1i}S(t_i;\boldsymbol{\eta})}[M_{1i}]^{\delta_i}}\right)^{y_i} (\mathbb{E}_{\theta_{1i}}[M_{1i}])^{1-y_i}, \quad \textit{for } y_i, \delta_i = 0, 1, \quad \textit{and}$$

$$\mathbb{E}[M_{2i} \mid y_i, t_i, \delta_i] = (\mathbb{E}_{\theta_{2i}}[M_{2i}])^{y_i} \left[\frac{\mathbb{E}_{\theta_{2i}S(t_i;\boldsymbol{\eta})}[M_{2i}^{1+\delta_i}]}{\mathbb{E}_{\theta_{2i}S(t_i;\boldsymbol{\eta})}[M_{2i}]^{\delta_i}}\right]^{1-y_i}, \quad \textit{for } y_i, \delta_i = 0, 1.$$

Let $\boldsymbol{\psi}^{(k)}$ be the estimate of $\boldsymbol{\psi}$ at the k-th iteration and $Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}^{(k)})$ denotes the conditional expectation of $\ell_c(\boldsymbol{\psi} \mid \mathrm{D}_{comp})$ on Equation (2.7) given the observed data and $\boldsymbol{\psi}^{(k)}$. With these notations, we obtain

$$Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}^{(k)}) = Q_1(\boldsymbol{\eta} \mid \boldsymbol{\psi}^{(k)}) + Q_2(\boldsymbol{\beta} \mid \boldsymbol{\psi}^{(k)}) + Q_3(\gamma \mid \boldsymbol{\psi}^{(k)}),$$

where

$$Q_1(\boldsymbol{\eta} \mid \boldsymbol{\psi}^{(k)}) = \sum_{i=1}^{n} \left\{ \left[\widetilde{Y_i M_{1i}}^{(k)} + \left[\widetilde{M_{2i}}^{(k)} - \widetilde{Y_i M_{2i}}^{(k)}\right] - \delta_i\right] \log[S(t_i;\boldsymbol{\eta})] \right.$$
$$\left. + \delta_i \log[f(t_i;\boldsymbol{\eta})] \right\}, \tag{2.9}$$

$$Q_2(\boldsymbol{\beta} \mid \boldsymbol{\psi}^{(k)}) = \sum_{i=1}^{n} \left\{ \widetilde{M_{1i}}^{(k)} \log(\theta_{1i}) - \log[A_1(\theta_{1i})] + \widetilde{M_{2i}}^{(k)} \log(\theta_{2i}) \right.$$
$$\left. - \log[A_2(\theta_{2i})] \right\}, \tag{2.10}$$

and

$$Q_3(\gamma \mid \boldsymbol{\psi}^{(k)}) = \sum_{i=1}^{n} \left\{ \widetilde{Y}_i^{(k)} \log(\gamma) + \left(1 - \widetilde{Y}_i^{(k)}\right) \log(1-\gamma) \right\}, \tag{2.11}$$

where, for the components $j = 1, 2$, the expressions for $\widetilde{M_{ji}}^{(k)} = \mathbb{E}\left[M_{ji} \mid \mathsf{D}_{obs}; \boldsymbol{\psi}^{(k)}\right]$, $\widetilde{Y}_i^{(k)} = \mathbb{E}\left[Y_i \mid \mathsf{D}_{obs}; \boldsymbol{\psi}^{(k)}\right]$ and $\widetilde{Y_i M_{ji}}^{(k)} = \mathbb{E}\left[Y_i M_{ji} \mid \mathsf{D}_{obs}; \boldsymbol{\psi}^{(k)}\right]$.

Therefore, the k-th iteration of the EM algorithm consists of the following steps:

- **E-step:** Given the observed data and the estimate for the vector of parameters at the $k-1$ iteration $\boldsymbol{\psi}^{(k-1)}$, for $i = 1, \ldots, n$, compute

$$\begin{aligned}
\widetilde{M_{1i}}^{(k)} &= \mathbb{E}_{\theta_{1i}^{(k)}}[M_{1i}] \left(1 - \widetilde{Y}_i^{(k)}\right) + \left(\frac{\mathbb{E}_{\theta_{1i}^{(k)} S(t_i;\boldsymbol{\eta}^{(k)})}[M_{1i}^{1+\delta_i}]}{\mathbb{E}_{\theta_{1i}^{(k)} S(t_i;\boldsymbol{\eta}^{(k)})}[M_{1i}]^{\delta_i}} \right) \widetilde{Y}_i^{(k)}, \\
\widetilde{M_{2i}}^{(k)} &= \left[\frac{\mathbb{E}_{\theta_{2i}^{(k)} S(t_i;\boldsymbol{\eta}^{(k)})}[M_{2i}^{1+\delta_i}]}{\mathbb{E}_{\theta_{2i}^{(k)} S(t_i;\boldsymbol{\eta}^{(k)})}[M_{2i}]^{\delta_i}} \right] \left(1 - \widetilde{Y}_i^{(k)}\right) + \mathbb{E}_{\theta_{2i}^{(k)}}[M_{2i}] \widetilde{Y}_i^{(k)} \\
\widetilde{Y_i M_{1i}}^{(k)} &= \left(\frac{\mathbb{E}_{\theta_{1i}^{(k)} S(t_i;\boldsymbol{\eta}^{(k)})}[M_{1i}^{1+\delta_i}]}{\mathbb{E}_{\theta_{1i}^{(k)} S(t_i;\boldsymbol{\eta}^{(k)})}[M_{1i}]^{\delta_i}} \right) \widetilde{Y}_i^{(k)}, \\
\widetilde{Y_i M_{2i}}^{(k)} &= \mathbb{E}_{\theta_{2i}^{(k)}}[M_{2i}] \widetilde{Y}_i^{(k)}, \qquad \text{and} \\
\widetilde{Y}_i^{(k)} &= \left(\frac{\omega_i^{(k)}}{1 + \omega_i^{(k)}} \right).
\end{aligned} \tag{2.12}$$

- **M-step:** Given $\boldsymbol{M}_j^{(k)} = (\widetilde{M_{j1}}^{(k)}, \ldots, \widetilde{M_{jn}}^{(k)})$, $\boldsymbol{Y}\boldsymbol{M}_j^{(k)} = (\widetilde{Y_1 M_{j1}}^{(k)}, \ldots, \widetilde{Y_n M_{jn}}^{(k)})$, for $j = 1, 2$ and $\boldsymbol{Y}^{(k)} = (\widetilde{Y_1}^{(k)}, \ldots, \widetilde{Y_n}^{(k)})$. Find $\boldsymbol{\psi}^{(k)} = (\boldsymbol{\eta}^{(k)}, \boldsymbol{\beta}^{(k)}, \gamma^{(k)})$, that maximize (2.9), (2.10) and (2.11), respectively. Moreover, for γ such maximization provides

$$\gamma^{(k)} = \frac{1}{n} \sum_{i=1}^{n} \widetilde{Y}_i^{(k)}.$$

The E- and M-steps are repeatedly alternated until a suitable convergence rule is satisfied, that is, the difference in successive values of the estimates is less than a pre-specified tolerance. On the other hand, SE corresponding to the estimator $\widehat{\boldsymbol{\psi}}$ can

be obtained from the Hessian matrix of the observed log-likelihood function in (2.6), i.e.,

$$\boldsymbol{\Sigma}(\widehat{\boldsymbol{\psi}}) = -\left.\frac{\partial^2 \ell(\boldsymbol{\psi}|\mathbf{D}_{\text{obs}})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^\top}\right|_{\boldsymbol{\psi}=\widehat{\boldsymbol{\psi}}}.$$

Such matrix can be obtained using the `hessian` function included in the `pracma` (Borchers, 2022) package of R Core Team (2024). Under suitable regularity conditions, it can be shown that in (Kalbfleisch and Prentice, 2002)

$$\sqrt{n}\left[\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\psi})\right]^{-1/2}\left(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right) \xrightarrow{\mathcal{D}} N_{q+3}(\mathbf{0}_{q+3}, \mathbf{I}_{q+3}), \qquad \text{as } n \to \infty. \tag{2.13}$$

It converges in distribution to the Normal distribution, where $\mathbf{0}_{q+3}$ represents a vector of zeros with dimension $q+3$, and $\mathbf{I}_{q+3}$ denotes the identity matrix of order $(q+3)$, representing respectively the vector of means and the covariance matrix.

Remark 2.3.1. *The expected values for the E-step have a simple form to the models considered in Table 2.1. For instance, for the BER-POI model is obtained*

$$\begin{aligned}
\mathbb{E}_{\theta_{1i}}[M_{1i}] &= \frac{\theta_{1i}}{1+\theta_{1i}}, & \mathbb{E}_{\theta_{2i}}[M_{2i}] &= \theta_{2i}, \\
\mathbb{E}_{\theta_{1i}S(t_i;\boldsymbol{\eta})}[M_{1i}^{1+\delta_i}] &= \frac{\theta_{1i}S(t_i;\boldsymbol{\eta})}{1+\theta_{1i}S(t_i;\boldsymbol{\eta})}, & \mathbb{E}_{\theta_{2i}S(t_i;\boldsymbol{\eta})}[M_{2i}^{1+\delta_i}] &= \theta_{2i}S(t_i;\boldsymbol{\eta}) + \delta_i\left(\theta_{2i}S(t_i;\boldsymbol{\eta})\right)^2, \\
\mathbb{E}_{\theta_{1i}S(t_i;\boldsymbol{\eta})}[M_{1i}]^{\delta_i} &= \left(\frac{\theta_{1i}S(t_i;\boldsymbol{\eta})}{1+\theta_{1i}S(t_i;\boldsymbol{\eta})}\right)^{\delta_i}, & \mathbb{E}_{\theta_{2i}S(t_i;\boldsymbol{\eta})}[M_{2i}]^{\delta_i} &= \left(\theta_{2i}S(t_i;\boldsymbol{\eta})\right)^{2\delta_i}.
\end{aligned}$$

With those values it is possible to compute all the expressions in the E-step.

Remark 2.3.2. *Up to this moment the procedure was developed without the specification of $S(t_i;\boldsymbol{\eta})$ and $f(t_i;\boldsymbol{\eta})$. In practice, the Weibull distribution is a very suitable model in this context. For this reason, we consider this model with parameterization $S(t_i;\boldsymbol{\eta}) = \exp(-e^{\alpha}t_i^{\nu})$ and $f(t_i;\boldsymbol{\eta}) = e^{\alpha}\nu t_i^{\nu-1}\exp(-e^{\alpha}t_i^{\nu})$, where $\alpha \in \mathbb{R}$ and $\nu \in \mathbb{R}^+$. In this case, $\boldsymbol{\eta} = (\alpha, \nu)$.*

Remark 2.3.3. *As mentioned previously, for $\gamma = 1$ is recovered the model assumed for M_{1i} (the first model in the mixture). This is a particular case of the PS cure rate model introduced by Cancho et al. (2013), where an EM-type algorithm was developed by*

(Gallardo et al., 2017). Initial values for $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ in our proposal can be obtained based on such model. In addition, we also can take an arbitrary value for γ. For instance, we take $\gamma^{(0)} = 0.5$. These initial values showed a satisfactory convergence in both, the simulation studies and the application.

2.4 Numerical applications

In this section, we carry out a simulation study to evaluate the performance of the ML based on EM estimators of the PSMCC model. The aim was to assess how accurately the true parameters can be recovered by the proposed EM algorithm procedure presented in Section 2.3. Then, we illustrate the proposed methodology by applying it to a real-world medical data set. We compare the proposed model with the model studied in Gallardo et al. (2017). Then, we find the one that provides an adequate fit to the data.

2.4.1 Simulation study

This study was conducted as follows. The time-to-event values for the simulation study were drawn from the Weibull distribution with fixed parameters $\alpha = -4$ and $\nu = 1.8$. Since the PSMCC proposed model considers a mixture of two different distributions for concurrent causes M_1 and M_2, we can have 10 possible combinations for this mixture with distributions in Table 2.1. The following 6 combinations of models were used to deal with the simulation study: Bernoulli and Geometric; Geometric and Logarithmic; Poisson an Logarithmic; Bernoulli and Poisson; Poisson and Geometric; Geometric and Bell. For each mixture, a MC study was conduced with 1,000 replications.

For each individual was considered two covariates. One of them was assumed to be a dichotomous variable, that is, x_{1i}, for $i = 1, \ldots, n$ where x_{1i} was drawn from a Bernoulli distribution with mean 0.5. The second one, x_{2i}, for $i = 1, \ldots, n$, was drawn from an Uniform distribution between 0 and 10. With the vector of regression parameters

fixed in $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top = (1.9, -1.5, -0.2)^\top$ we compute p_i in (2.5). For the distribution of time-to-event, was fixed the parameters $\alpha = -4$ and $\nu = 1.8$. For the parameter that set the mixture of concurrent causes was assumed that $\gamma = 0.5$. Finally, in all cases the censoring times were drawn from the uniform distribution between 0 and 20, yielding on average approximately 20% percentage of censoring. The values that were considered for the parameters comes from the approximation of the estimates obtained for the parameters of the proposed model in Gallardo et al. (2017).

For each value of the parameter, sample size and combinations of PSMCC, was reported the empirical values for the estimated bias and root of the estimated mean squared error (RMSE) and 95% coverage probability (CP) of the ML estimators based on the asymptotic distribution of the ML estimators in (2.13) in Tables 2.5 and 2.6 presented in the Appendix B. The mean of the SE and the RMSE were obtained through the Hessian matrix, considering the asymptotic distribution of maximum likelihood based on EM estimators.

Some comments concerning the behavior of the maximum likelihood estimates obtained in the simulation study are presented.

- From these tables, it is noted that for each model tested, as the sample size n increases, the results of SE and RMSE of the EM estimators were became smaller, showing us the efficiency of maximum likelihood estimates for each model tested in this study.

- The average bias of the coefficients regression parameters in $\boldsymbol{\beta}$ for all of 6 considered combinations and samples size were reduced, showing that the maximum likelihood estimates found for all combinations of models proposed were close from the values fixed for all the parameters that generate the samples used for each simulation study.

- The bias of the time-to-event distribution parameters α and ν became smaller as the sample size n increased. Furthermore, it can be noted that these estimated values are greater compared to the ones considered in the regression structure. This happened because the values chosen for the simulation study were larger, in absolute terms, than the ones for the $\boldsymbol{\beta}$ vector.

- The coverage probabilities (CP's) for all scenarios studied were next to the nominal value (95%).

In general, the presented simulation study shows that the EM algorithm proposed in this work presents consistent estimates asymptotically for a given set of parameters in situations where a sample is generated from a cure fraction model with competitive causes from a mixture of PS. Furthermore, the introduction of the regression structure directly into the cure fraction produced consistent estimates for its coefficients, which is interesting in practice, as there are many situations where covariates can better explain the failure time behavior of a particular individual or object of interest. The simulation study suggests good finite sample properties for all estimates of mixture models.

2.4.2 Application with a melanoma dataset

The observations provided by FOSP are from a retrospective survey of $7,166$ records of patients diagnosed with cutaneous melanoma in the State of São Paulo, Brazil. Where the patients were registered on between years 2000 and 2014, with follow-up conducted until 2018. Data from patients who did not die of melanoma during the follow-up period were right-censored. After the removal of data from 417 patients due to missing values for the observed covariates, the set contained data from $6,749$ patients. The death due to cancer was defined as the event of interest and the time to event was defined as the period from the date of melanoma diagnosis time of death due to cancer. Looking up for some aspects of the data, the observation that refers the maximum time found was 18.54 years, and the median of follow-up time was about 5.24 years. The level of censorship is up to $71,6\%$.

To have a regression structure on fitting, the explanatory variables measured at baseline for $i = 1, \ldots, 6,749$ were X_{i1} : surgery (0: no, $n_0 = 771$; 1: yes, $n_1 = 5,978$), X_{i2} : education level (0: illiterate, $n_0 = 349$; 1: elementary school, $n_1 = 2,649$; 2: high school, $n_2 = 841$; 3: college graduate, $n_3 = 667$; 4: not informed, $n_4 = 2,243$), X_{i3} : clinical cancer stage (0: Stage I, $n_0 = 3,011$; 1: Stage II, $n_1 = 1,541$; 2: Stage III, $n_2 = 1,229$; 3: Stage IV, $n_3 = 968$), X_{i4} : age at diagnosis (mean $\pm$ standard deviation,

58.04 ± 16.36 years), X_{i5} : sex (0: male, $n_0 = 3,334$; 1: female, $n_1 = 3,415$), X_{i6} : service category (0: service 1, $n_0 = 3,128$; 1: service 9, $n_1 = 3,621$), X_{i7} : radiotherapy (0: no, $n_0 = 6,162$; 1: yes, $n_1 = 587$) and X_{i8} : chemotherapy (0: no, $n_0 = 5,645$; 1: yes, $n_1 = 1,104$). All of these variables were introduced on the cure fraction of the proposed model.

The approach presented in Section 2.3 via EM algorithm was applied to obtain ML estimates. The authors had developed a computer program in `R` language (R Core Team, 2024) and it is available to community upon request. The estimates for the Melanoma data obtained in the best fit obtained from the work proposed by Gómez et al. (2021) were used in this one as initial values related to the parameters under estimation of the model under estimation.

Table 2.2: AIC and BIC criteria for cure rate models with and without mixture of concurrent causes on PS applied to melanoma dataset.

Model	Number of Parameters	Log-likelihood	AIC	BIC
BER-GEO	17	−5,482.76	10,999.53	11,115.42
GEO-BELL	17	−5,482.91	10,999.82	11,115.71
POI-GEO	17	−5,483.93	11,001.86	11,117.75
BER-BELL	17	−5,504.97	11,043.95	11,159.85
BELL	16	−5,510.79	11,055.58	11,164.66
POI-BELL	17	−5,519.60	11,073.21	11,189.10
BER-POI	17	−5,521.47	11,076.95	11,192.84
BER-LOG	17	−5,546.43	11,126.86	11,242.75
POI-LOG	17	−5,546.50	11,127.00	11,242.89
POI	16	−5,553.20	11,140.41	11,249.49
GEO	16	−5,864.68	11,763.37	11,872.44
GEO-LOG	17	−5,876.26	11,786.52	11,902.41
LOG-BELL	17	−5,967.94	11,969.88	12,085.77
LOG	16	−5,978.38	11,990.77	12,099.84
BER	16	−6,160.02	12,354.05	12,463.12

The Table 2.2 compares the fitting of the proposed model to the dataset into the one that deal with only one concurrent cause into the PS modelling (the particular case), proposed by Gallardo et al. (2017), was obtained the two penalized likelihood criteria Akaike information criterion (AIC), see Akaike, (1973) and Bayesian information criterion (BIC), see Schwarz (1978). Although the proposed model has one more parameter than

in the particular case, it is possible to observe that for the Melanoma dataset, four of the proposed mixture models presented better performance than the fitting using the particular case (Bell), they are Bernoulli-Poisson, Geometric-Bell, Poisson-Geometric and Bernoulli-Bell, showing the flexibility of the proposed model on choosing the best combination of distributions. The mixture considering the case where the competitive causes come from the Bernoulli and Geometric distribution was chosen as the best fit. The Weibull distribution was used to model the time to event of interest.

Table 2.3: ML estimates and SE obtained by fitting the four best combinations of PSMCC model and for BELL model to melanoma dataset.

Parameter	BER-GEO		GEO-BELL		POI-GEO		BER-BELL		BELL	
	ML	SE	ML	SE	ML	SE	ML	SE	ML	SE
β_0:Intercept	1.038	0.118	1.054	0.107	1.041	0.110	1.231	0.144	1.787	0.170
β_1:Surgery	0.921	0.046	0.950	0.041	0.928	0.042	1.064	0.065	1.007	0.070
β_2:Element. School	0.430	0.068	0.434	0.060	0.435	0.062	0.464	0.083	0.439	0.097
β_2:High School	0.566	0.081	0.567	0.073	0.570	0.075	0.527	0.098	0.571	0.117
β_2:Col. graduate	0.896	0.092	0.904	0.083	0.905	0.086	0.921	0.107	0.897	0.132
β_2:Not Informed	0.548	0.070	0.554	0.062	0.554	0.065	0.551	0.085	0.578	0.101
β_3:Stage II	-1.333	0.050	-1.319	0.047	-1.327	0.048	-1.355	0.049	-1.283	0.069
β_3:Stage III	-2.323	0.050	-2.291	0.046	-2.310	0.047	-2.382	0.052	-2.197	0.069
β_3:Stage IV	-3.970	0.056	-3.989	0.051	-3.969	0.052	-6.606	0.083	-4.216	0.082
β_4:Age	-0.011	0.001	-0.011	0.001	-0.011	0.001	-0.019	0.001	-0.014	0.002
β_5:Female	0.466	0.033	0.482	0.030	0.471	0.031	0.622	0.039	0.557	0.048
β_6:Service 9	-0.254	0.034	-0.260	0.031	-0.256	0.032	-0.270	0.039	-0.292	0.049
β_7:Radiotherapy	-0.420	0.050	-0.472	0.044	-0.434	0.046	-1.051	0.071	-0.748	0.074
β_8:Chemotherapy	-0.175	0.042	-0.222	0.037	-0.195	0.039	-0.736	0.056	-0.511	0.061
α	-2.670	0.030	-2.650	0.030	-2.659	0.030	-2.557	0.030	-2.226	0.031
ν	1.433	0.019	1.422	0.019	1.429	0.019	1.260	0.019	1.236	0.019
γ	0.036	0.021	0.857	0.028	0.057	0.019	0.228	0.117	-	-

All coefficient were significant at 5% of significance level

Through the Table 2.3, the estimates and SE for the parameters of the five best models chosen by the AIC and BIC selection criteria considered in Table 2.2 can be better appreciated. As the AIC and BIC criterion pointed the BER-GEO model as the best fit, let us focus on the interpretation of the results for the parameters estimated in this case. All the covariates considered in the modeling have significant coefficients of level 5%, which means that these variables contribute positively or negatively on lifetime of patients with carcinogenic cutaneous melanoma. By the estimates obtained, it can be inferred that patients who had surgery have 0.9212 times more chances to survive than patients who did not. With each passing year, patients who are in stage IV of the disease are 3.9696 times more likely to die compared with the ones on the initial stage. The same can be observed with patients on the Stage III of disease, where the chances of been cured 2.3233 times less. The higher the stage of the disease, the greater the

negative impact on the lifetime of patients. The educational level of people considered in modeling is a factor that contributes positively on the time of life. Those who have college graduate have 0.8958 more lifetime years than people without a degree. The study also reveals that the chances of women's survival are 0.4663 times higher than men's. Another interesting fact is with patients who did radiotherapy and chemotherapy, looking for the estimates of these parameters can be seen that lifetime is less in the group who did only radiotherapy, than those who had also chemotherapy, but the impact on time of life is not too significant.

For instance, for the BER-GEO model, looking for the estimate for γ, 3.6% of the population of the number of carcinogenic cells follows a BER distribution, whereas 96.4% of the population follows a GEO model. It can be observed that for all combinations and for the particular case among the best models studied, all presented significant coefficients at 5% of significance.

Table 2.4: Hypothesis test for $\gamma = 0$ and $\gamma = 1$ in different combinations for PSMCC model in melanoma data set.

Mixture model	$H_0 : \gamma = 0$ Model in H_0	LR	p-value	$H_0 : \gamma = 1$ Model in H_0	LR	p-value
BER-GEO	BER	1356.52	$<$0.001	GEO	765.84	$<$0.001
GEO-BELL	GEO	765.56	$<$0.001	BELL	57.76	$<$0.001
POI-GEO	POI	140.56	$<$0.001	GEO	763.52	$<$0.001
BER-BELL	BER	1312.10	$<$0.001	BELL	13.64	$<$0.001

We also considered to test $H_0 : \gamma = 0$ $(\gamma = 1)$ versus $H_1 : \gamma > 0$ $(\gamma < 1)$ in the four mixture models selected in Table 2.3 in order to select any of the particular models. Note that the null hypothesis is at the boundary of the parameter space. Therefore, the usual likelihood ratio test (LR) statistic, $\mathrm{LR} = 2(\ell_1^{ML} - \ell_0^{ML})$, where ℓ_i^{ML} denotes the log-likelihood function evaluated at the ML estimator under the hypothesis H_i, $i = 0, 1$, does not have the usual chi-squared distribution with 1 degrees of freedom ($\chi^2_{(1)}$). In this case, the asymptotic distribution is $(1/2) + (1/2)\chi^2_{(1)}$; see, for instance Stram and Lee (1994). Table 2.4 shows the p-value for such cases. We highlight that in all the cases the particular models are rejected in favour of the general model with any traditional significance level.

Figure 2.3 shows through the randomized quantile residuals (Dunn and Smyth,

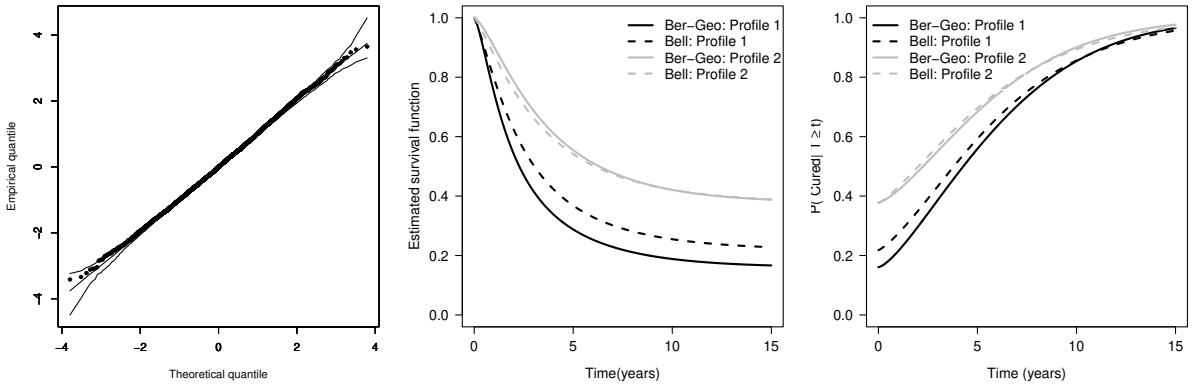

Figure 2.3: Randomized quantile residuals (left) for mixture of concurrent causes for Bernoulli and Geometric model. Estimated SF (center) and $P(\text{Cured}|T \geq t)$ (right) for BER-GEO and BELL models to patients with (Profile 1) and without (Profile 2) chemotherapy, and both profiles considering that patients have made surgery, radiotherapy, full primary school, were women and have Stage 1 of disease.

1996) that its behaviour is like a random sample from the standard normal distribution for the model for mixture of BER-GEO of concurrent causes model, which seems to yield an adequate fit to this data set. The estimated survival function for BER-GEO and Bell models to patients with (Profile 1) and without (Profile 2) chemotherapy, and both profiles considering that patients have made surgery, radiotherapy, completed primary school, were women, receive the category 9 of service and was ate Stage 1 of disease suggests a long-term survival model. In addition, there is evidence that patients who have undergone chemotherapy are more likely to survive the first 5 years. Furthermore, there are indications that the probability of individuals being cured is higher among patients who underwent chemotherapy in the first years compared to those who did not undergo chemotherapy throughout the study.

2.5 Concluding remarks

In this work, we proposed a new model for survival data assuming competing

causes of the event of interest and following a mixture of two PS distributions. The proposed methodology assumes that the distribution of the number of competing is different from individual to individual. The proposed model includes several well-known models as special cases and defines at least ten new special cases. We consider covariates on the cure rate of M, so allowing a direct modeling for the cure rate through covariates, facilitating the comparison with other models. We also supposed that the time to the event of interest follow Weibull distribution. Special cases are studied in some detail. Maximum likelihood inference is implemented straightforwardly for estimating the model parameters. In particular, we proposed a simplified estimation procedure based on the EM algorithm. A Monte Carlo simulation study has shown that the estimates based on the EM method of the model parameters tend to their true values, whereas the distributions of these estimators converge to normality, when the sample size increases, as expected. Finally, an empirical illustration for the cutaneous melanoma diagnosed in the state of São Paulo, Brazil, was analyzed by considering the models introduced here. The sub-model chosen as the best fit through the information criterion suggest that number of carcinogenic cells in cutaneous melanoma comes from a mixture of Bernoulli and Geometric distributions, where it could be inferred that most cells come from Geometric population, totalling 96.4%. The other 3.6% follow a Bernoulli distribution. An interesting fact suggests that in cure fraction, patients that has made surgery has more lifetime than those who has not. Albeit chemotherapy treatment reduces the time of life of most patients, it invariably presents a higher probability of cure in comparison with those patients who do not attend chemotherapy. The regression structure on the cure fraction revealed relevant issues about the socioeconomic aspects of individuals with cutaneous melanoma cancer. Women survive longer than men and the higher the educational level, the longer the survival time. In addition, the higher the stage of the disease, the greater the negative impact on patient's survival time. The empirical results showed the potentiality of this methodology and, particularly, the use of the diagnostic tools derived in the work. In fact, the proposed model fits the data set well. Future work should explore other estimation methods for the proposed cure rate model, for instance, the Bayesian approach similarly as developed by Chen et al. (1999), and also consider other parametric forms for promotion times, such as beta prime and gamma.

Appendix A: Proofs

Proof of Proposition 3.1

Following Yakovlev and Tsodikov (1996), we have that

$$\begin{aligned} f(t_i,\delta_i,m_{1i},m_{2i},y_i,p_i,\boldsymbol{\eta},\gamma) &= f(t_i,\delta_i,y_i|M_{1i}=m_{1i},M_{1i}=m_{2i},\boldsymbol{\eta},\gamma)P(M_{1i}=m_{1i};\theta_{1i}) \\ &\quad \times P(M_{2i}=m_{2i};\theta_{2i})P(Y_i=y_i;\gamma) \\ &= \left\{S(t_i;\boldsymbol{\eta})^{m_{1i}-\delta_i}[m_{1i}f(t_i;\boldsymbol{\eta})]^{\delta_i}\right\}^{y_i} \\ &\quad \times \left\{S(t_i;\boldsymbol{\eta})^{m_{2i}-\delta_i}[m_{2i}f(t_i;\boldsymbol{\eta})]^{\delta_i}\right\}^{1-y_i} \\ &\quad \times \frac{a_1(m_{1i})\theta_{1i}^{m_{1i}}}{A_1(\theta_{1i})}\frac{a_2(m_{2i})\theta_{2i}^{m_{2i}}}{A_2(\theta_{2i})}\gamma^{y_i}(1-\gamma)^{1-yi}, \end{aligned} \tag{2.14}$$

where $m_{1i},m_{2i}=\delta_i,\delta_i{+}1,\ldots$. Define $\mathbb{E}_{\theta_{1i}S(t_i;\boldsymbol{\eta})}\left[M_{1i}^d\right]=\sum_{m_{ji}=0}^{\infty}m_{1i}^d\frac{a_1(m_{1i})(\theta_{1i}S(t_i;\boldsymbol{\eta}))^{m_{1i}}}{A_1(\theta_{1i}S(t_i;\boldsymbol{\eta}))}$ and $\mathbb{E}_{\theta_{2i}S(t_i;\boldsymbol{\eta})}\left[M_{2i}^d\right]=\sum_{m_{2i}=0}^{\infty}m_{2i}^d\frac{a_2(m_{2i})(\theta_{2i}S(t_i;\boldsymbol{\eta}))^{m_{2i}}}{A_2(\theta_{ji}S(t_i;\boldsymbol{\eta}))}$. By using (2.14), we have that

$$\begin{aligned} f(t_i,\delta_i,y_i,p_i,\boldsymbol{\eta},\gamma) &= \sum_{m_{1i}=0}^{\infty}\sum_{m_{2i}=0}^{\infty}\left\{S(t_i;\boldsymbol{\eta})^{m_{1i}-\delta_i}[m_{1i}f(t_i;\boldsymbol{\eta})]^{\delta_i}\gamma\right\}^{y_i}\left\{S(t_i;\boldsymbol{\eta})^{m_{2i}-\delta_i}[m_{2i}f(t_i;\boldsymbol{\eta})]^{\delta_i}\right. \\ &\quad \left.\times(1-\gamma)\right\}^{1-y_i}\frac{a_1(m_{1i})\theta_{1i}^{m_{1i}}}{A_1(\theta_{1i})}\frac{a_2(m_{2i})\theta_{2i}^{m_{2i}}}{A_2(\theta_{2i})} \\ &= \left[\left(\frac{f(t_i;\boldsymbol{\eta})}{S(t_i;\boldsymbol{\eta})}\right)^{\delta_i}\frac{A_1(\theta_{1i}S[t_i;\boldsymbol{\eta}])}{A_1(\theta_{1i})}\sum_{m_{1i}=0}^{\infty}m_{1i}^{\delta_i}\frac{a_1(m_{1i})(\theta_{1i}S[t_i;\boldsymbol{\eta}])^{m_{1i}}}{A_1(\theta_{1i}S[t_i;\boldsymbol{\eta}])}\gamma\right]^{y_i} \\ &\quad \times\left[\left(\frac{f(t_i;\boldsymbol{\eta})}{S(t_i;\boldsymbol{\eta})}\right)^{\delta_i}\frac{A_2(\theta_{2i}S[t_i;\boldsymbol{\eta}])}{A_2(\theta_{2i})}\sum_{m_{2i}=0}^{\infty}m_{2i}^{\delta_i}\frac{a_2(m_{1i})(\theta_{2i}S[t_i;\boldsymbol{\eta}])^{m_{2i}}}{A_2(\theta_{2i}S[t_i;\boldsymbol{\eta}])}(1-\gamma)\right]^{1-y_i} \\ &= \omega_i^{y_i}\left(\frac{f(t_i;\boldsymbol{\eta})\mathbb{E}_{\theta_{2i}S(t_i;\boldsymbol{\eta})}\left[M_{2i}\right]}{S[t_i;\boldsymbol{\eta}]}\right)^{\delta_i}\frac{A_2(\theta_{2i}S[t_i;\boldsymbol{\eta}])}{A_2(\theta_{2i})}(1-\gamma). \end{aligned} \tag{2.15}$$

Therefore,

$$Y_i \mid t_i,\delta_i \sim \text{BER}\left(\frac{\omega_i}{1+\omega_i}\right). \tag{2.16}$$

On the other hand, by definition, we have that

$$P(M_{1i}=m_{1i},M_{2i}=m_{2i}\mid t_i,\delta_i,y_i,\boldsymbol{\eta},\gamma)=\frac{f(t_i,\delta_i,m_{1i},m_{2i},y_i,\boldsymbol{\eta},\gamma)}{f(t_i,\delta_i,y_i,\boldsymbol{\eta},\gamma)}.$$

Plugging the expressions (2.14) and (2.15) in the latter definition, we obtain

$$P(M_{1i}=m_{1i},M_{2i}=m_{2i}\mid y_i,t_i,\delta_i,\boldsymbol{\eta})=\frac{a_1(m_{1i})\left(\theta_{1i}\left[S(t_i;\boldsymbol{\eta})\right]^{y_i}\right)^{m_{1i}}}{A_1\left(\theta_{1i}\left[S(t_i;\boldsymbol{\eta})\right]^{y_i}\right)}\left(\frac{m_{1i}}{\mathbb{E}_{\theta_{1i}S(t_i;\boldsymbol{\eta})}[M_{1i}]}\right)^{y_i\delta_i}$$
$$\times\frac{a_2(m_{2i})\left(\theta_{2i}\left[S(t_i;\boldsymbol{\eta})\right]^{1-y_i}\right)^{m_{2i}}}{A_2\left(\theta_{2i}\left[S(t_i;\boldsymbol{\eta})\right]^{1-y_i}\right)}\left(\frac{m_{2i}}{\mathbb{E}_{\theta_{2i}S(t_i;\boldsymbol{\eta})}[M_{2i}]}\right)^{(1-y_i)\delta_i}.$$

Therefore, M_{1i} and M_{2i} are conditionally independent given (Y_i,t_i,δ_i) such that

$$P(M_{1i}=m_{1i}\mid y_i,t_i,\delta_i)=\frac{a_1(m_{1i})\left(\theta_{1i}\left[S(t_i;\boldsymbol{\eta})\right]^{y_i}\right)^{m_{1i}}}{A_1\left(\theta_{1i}\left[S(t_i;\boldsymbol{\eta})\right]^{y_i}\right)}\left(\frac{m_{1i}}{\mathbb{E}_{\theta_{1i}S(t_i;\boldsymbol{\eta})}[M_{1i}]}\right)^{y_i\delta_i}\quad\text{and}\quad(2.17)$$

$$P(M_{2i}=m_{2i}\mid y_i,t_i,\delta_i)=\frac{a_2(m_{2i})\left(\theta_{2i}\left[S(t_i;\boldsymbol{\eta})\right]^{1-y_i}\right)^{m_{2i}}}{A_2\left(\theta_{2i}\left[S(t_i;\boldsymbol{\eta})\right]^{1-y_i}\right)}\left(\frac{m_{2i}}{\mathbb{E}_{\theta_{2i}S(t_i;\boldsymbol{\eta})}[M_{2i}]}\right)^{(1-y_i)\delta_i}.\quad(2.18)$$

Details for Corollary 3.1

In this section, some details are presented for E-step of the EM algorithm. Using the conditional distribution found in the previous section for M_{ij} given D_{obs}, for $j=1,2$, on Equations (2.17) and (2.18) and the conditional distribution for Y_i given t_i,δ_i deduced on Equation (2.16), the steps that are needed for E-step can be obtained in close form. By definition, the expected value for the number of initial causes M_{ji} given (y_i,t_i,δ_i), for $j=1,2$, can be written in close form as

$$\mathbb{E}[M_{1i}\mid y_i,t_i,\delta_i]=\left(\frac{\mathbb{E}_{\theta_{1i}S(t_i;\boldsymbol{\eta})}[M_{1i}^{1+\delta_i}]}{\mathbb{E}_{\theta_{1i}S(t_i;\boldsymbol{\eta})}[M_{1i}]^{\delta_i}}\right)^{y_i}\left(\mathbb{E}_{\theta_{1i}}[M_{1i}]\right)^{1-y_i},\quad\text{for } y_i,\delta_i=0,1,$$

and

$$\mathbb{E}[M_{2i}\mid y_i,t_i,\delta_i]=\left(\mathbb{E}_{\theta_{2i}}[M_{2i}]\right)^{y_i}\left[\frac{\mathbb{E}_{\theta_{2i}S(t_i;\boldsymbol{\eta})}[M_{2i}^{1+\delta_i}]}{\mathbb{E}_{\theta_{2i}S(t_i;\boldsymbol{\eta})}[M_{2i}]^{\delta_i}}\right]^{1-y_i},\quad\text{for } y_i,\delta_i=0,1.$$

Note that the expressions above for $\mathbb{E}[M_{1i}\mid y_i,t_i\,\delta_i]$ and $\mathbb{E}[M_{2i}\mid y_i,t_i\,\delta_i]$ are in function of the random variable Y_i, here and after lets call them $g_1(Y_i)$ and $g_2(Y_i)$, respectively .

Additional details for E-step

Since $Y_i \sim$ BER $\left(\frac{\omega_i}{1+\omega_i}\right)$, where ω_i is in the Equation (2.8), the only values that Y_i takes are 0 and 1. By using the definition of expected value of function of random variables, it comes

$$\begin{aligned}
\mathbb{E}\left[M_{1i} \mid \mathsf{D}_{obs}\right] &= \mathbb{E}\left[g_1(Y_i) \mid \mathsf{D}_{obs}\right] = g_1(0)P(Y_i = 0 \mid \mathsf{D}_{obs}) + g_1(1)P(Y_i = 1 \mid \mathsf{D}_{obs}) \\
&= \left(\frac{1}{1+\omega_i}\right)\mathbb{E}_{\theta_{1i}}[M_{1i}] + \left(\frac{\omega_i}{1+\omega_i}\right)\left[\delta_i \frac{\mathbb{E}_{\theta_{1i}S(t_i;\boldsymbol{\eta})}[M_{1i}^2]}{\mathbb{E}_{\theta_{1i}S(t_i;\boldsymbol{\eta})}[M_{1i}]} + (1-\delta_i)\mathbb{E}_{\theta_{1i}S(t_i;\boldsymbol{\eta})}[M_{1i}]\right].
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left[M_{2i} \mid \mathsf{D}_{obs}\right] &= \mathbb{E}\left[g_2(Y_i) \mid \mathsf{D}_{obs}\right] = g_2(0)P(Y_i = 0 \mid \mathsf{D}_{obs}) + g_2(1)P(Y_i = 1 \mid \mathsf{D}_{obs}) \\
&= \left(\frac{1}{1+\omega_i}\right)\left[\delta_i \frac{\mathbb{E}_{\theta_{2i}S(t_i;\boldsymbol{\eta})}[M_{2i}^2]}{\mathbb{E}_{\theta_{2i}S(t_i;\boldsymbol{\eta})}[M_{2i}]} + (1-\delta_i)\mathbb{E}_{\theta_{2i}S(t_i;\boldsymbol{\eta})}[M_{2i}]\right] + \left(\frac{\omega_i}{1+\omega_i}\right)\mathbb{E}_{\theta_{2i}}[M_{2i}].
\end{aligned}$$

Separating the cases $\delta_i = 0$ and $\delta_i = 1$ and considering the convenient distribution for each component of the mixture M_{1i} and M_{2i}, the result is obtained for $\widetilde{M_{1i}}^{(k)} = \mathbb{E}\left[M_{1i} \mid \mathsf{D}_{obs}, \boldsymbol{\psi}^{(k)}\right]$ and $\widetilde{M_{2i}}^{(k)} = \mathbb{E}\left[M_{2i} \mid \mathsf{D}_{obs}, \boldsymbol{\psi}^{(k)}\right]$. Similarly, by using the same properties for $\widetilde{Y_i M_{ji}}^{(k)} = \mathbb{E}\left[Y_i M_{ji} \mid \mathsf{D}_{obs}; \boldsymbol{\psi}^{(k)}\right]$, we have that

$$\begin{aligned}
\mathbb{E}\left[Y_i M_{1i} \mid \mathsf{D}_{obs}\right] &= \mathbb{E}\left[\mathbb{E}\left(Y_i M_{1i} \mid \mathsf{D}_{obs}\right) \mid \mathsf{D}_{obs}\right] \\
&= \mathbb{E}\left[Y_i \mathbb{E}\left(M_{1i} \mid \mathsf{D}_{obs}\right) \mid \mathsf{D}_{obs}\right] = \mathbb{E}\left[Y_i \mathbb{E}[g_1(Y_i)] \mid \mathsf{D}_{obs}\right] \\
&= \left(\frac{\omega_i}{1+\omega_i}\right)\left[\delta_i \frac{\mathbb{E}_{\theta_{1i}S(t_i;\boldsymbol{\eta})}[M_{1i}^2]}{\mathbb{E}_{\theta_{1i}S(t_i;\boldsymbol{\eta})}[M_{1i}]} + (1-\delta_i)\mathbb{E}_{\theta_{1i}S(t_i;\boldsymbol{\eta})}[M_{1i}]\right] \qquad (2.19)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left[Y_i M_{2i} \mid \mathsf{D}_{obs}\right] &= \mathbb{E}\left[\mathbb{E}\left(Y_i M_{2i} \mid \mathsf{D}_{obs}\right) \mid \mathsf{D}_{obs}\right] \\
&= \mathbb{E}\left[Y_i \mathbb{E}\left(M_{2i} \mid \mathsf{D}_{obs}\right) \mid \mathsf{D}_{obs}\right] = \mathbb{E}\left[Y_i \mathbb{E}[g_2(Y_i)] \mid \mathsf{D}_{obs}\right] \\
&= \left(\frac{\omega_i}{1+\omega_i}\right)\mathbb{E}_{\theta_{2i}}[M_{2i}]. \qquad (2.20)
\end{aligned}$$

Similarly, by separating the cases $\delta_i = 0$ and $\delta_i = 1$ and considering the

convenient distribution for each component of the mixture M_1 and M_2, the result is obtained for the conditional expected values $\widetilde{Y_i M_{1i}}^{(k)} = \mathbb{E}\left[Y_i M_{1i} \mid \mathsf{D}_{obs}; \boldsymbol{\psi}^{(k)}\right]$ and $\widetilde{Y_i M_{2i}}^{(k)} = \mathbb{E}\left[Y_i M_{2i} \mid \mathsf{D}_{obs}; \boldsymbol{\psi}^{(k)}\right]$. For $\widetilde{Y}_i^{(k)} = \mathbb{E}\left[Y_i \mid \mathsf{D}_{obs}; \boldsymbol{\psi}^{(k)}\right]$ the result is straightforward of the expectation from the distribution of $Y_i \mid t_i, \delta_i$ deduced on Equation (2.16). Since it is Bernoulli $\left(\frac{\omega_i}{1+\omega_i}\right)$, it follows

$$\mathbb{E}\left[Y_i \mid \mathsf{D}_{obs}\right] = \left(\frac{\omega_i}{1+\omega_i}\right).$$

Appendix B: Additional results for simulation study

Table 2.5: Empirical bias, SE, RMSE and CP of the ML estimators for the Weibull distribution to time-to-event in the concurrent causes regression.

Distribution for M_1 and M_2	$\boldsymbol{\psi}$	Bias	SE	RMSE	CP	Bias	SE	RMSE	CP
		$n = 200$				$n = 400$			
BER and GEO	β_0	0.020	0.665	0.637	0.955	0.017	0.422	0.406	0.963
	β_1	−0.106	0.497	0.511	0.943	−0.044	0.323	0.319	0.962
	β_2	−0.015	0.084	0.084	0.953	−0.005	0.055	0.055	0.949
	α	−0.139	0.763	0.685	0.963	−0.078	0.499	0.472	0.948
	ν	0.026	0.258	0.230	0.963	0.019	0.174	0.166	0.961
	γ	−0.007	0.503	0.311	0.956	−0.024	0.369	0.277	0.950
		$n = 600$				$n = 1000$			
	β_0	0.020	0.337	0.320	0.962	0.015	0.255	0.245	0.962
	β_1	−0.022	0.262	0.257	0.952	−0.015	0.200	0.198	0.949
	β_2	−0.005	0.045	0.043	0.958	−0.004	0.034	0.033	0.952
	α	−0.022	0.407	0.373	0.952	−0.013	0.312	0.284	0.956
	ν	−0.002	0.142	0.131	0.955	−0.001	0.106	0.100	0.961
	γ	−0.001	0.302	0.244	0.955	0.001	0.232	0.195	0.962
		$n = 200$				$n = 400$			
GEO and LOG	β_0	−0.006	0.510	0.515	0.959	0.033	0.351	0.336	0.957
	β_1	−0.030	0.350	0.341	0.951	−0.028	0.253	0.239	0.955
	β_2	−0.002	0.056	0.053	0.948	−0.005	0.041	0.039	0.955
	α	−0.242	0.665	0.788	0.968	−0.097	0.441	0.393	0.964
	ν	0.086	0.249	0.262	0.948	0.037	0.178	0.166	0.964
	γ	−0.083	0.627	0.323	0.965	−0.043	0.452	0.306	0.930
		$n = 600$				$n = 1000$			
	β_0	−0.004	0.284	0.266	0.970	0.015	0.219	0.204	0.958
	β_1	0.001	0.205	0.192	0.965	−0.011	0.160	0.149	0.964
	β_2	−0.002	0.033	0.032	0.951	−0.003	0.026	0.024	0.959
	α	−0.082	0.356	0.331	0.961	−0.036	0.275	0.254	0.960
	ν	0.023	0.141	0.136	0.951	0.013	0.112	0.104	0.957
	γ	−0.044	0.330	0.268	0.949	−0.018	0.294	0.223	0.951
		$n = 200$				$n = 400$			
POI and LOG	β_0	−0.010	0.524	0.531	0.967	0.009	0.356	0.352	0.947
	β_1	−0.024	0.368	0.358	0.951	−0.013	0.258	0.253	0.958
	β_2	−0.004	0.058	0.058	0.941	−0.003	0.041	0.041	0.949
	α	−0.259	0.692	0.865	0.943	−0.110	0.454	0.446	0.942
	ν	0.087	0.248	0.273	0.949	0.034	0.167	0.171	0.943
	γ	−0.050	0.365	0.274	0.937	−0.030	0.264	0.221	0.932
		$n = 600$				$n = 1000$			
	β_0	0.008	0.285	0.268	0.961	0.012	0.218	0.215	0.950
	β_1	−0.008	0.209	0.203	0.953	−0.014	0.161	0.158	0.956
	β_2	−0.002	0.033	0.033	0.946	−0.001	0.026	0.026	0.955
	α	−0.063	0.364	0.350	0.956	−0.031	0.279	0.267	0.949
	ν	0.020	0.135	0.135	0.957	0.009	0.103	0.101	0.950
	γ	−0.018	0.213	0.194	0.938	−0.010	0.165	0.154	0.959

Table 2.6: Empirical bias, SE, RMSE and CP of the ML estimators for the Weibull distribution to time-to-event in the concurrent causes regression.

Distribution for M_1 and M_2	$\boldsymbol{\psi}$	Bias	SE	RMSE	CP	Bias	SE	RMSE	CP
		$n = 200$				$n = 400$			
BER and POI	β_0	0.104	0.715	0.786	0.973	0.044	0.449	0.400	0.973
	β_1	−0.287	0.722	1.407	0.970	−0.084	0.382	0.611	0.970
	β_2	−0.039	0.119	0.264	0.972	−0.009	0.063	0.067	0.965
	α	−0.119	0.826	0.537	0.987	−0.033	0.573	0.375	0.991
	ν	0.028	0.268	0.233	0.960	0.007	0.187	0.161	0.967
	γ	0.004	1.224	0.362	0.996	0.013	0.874	0.333	0.999
		$n = 600$				$n = 1000$			
	β_0	0.028	0.350	0.312	0.969	0.006	0.267	0.236	0.967
	β_1	−0.037	0.283	0.279	0.961	−0.007	0.212	0.199	0.963
	β_2	−0.005	0.049	0.047	0.959	−0.003	0.037	0.035	0.958
	α	−0.031	0.446	0.310	0.987	−0.021	0.341	0.244	0.984
	ν	0.015	0.148	0.136	0.952	0.005	0.111	0.103	0.966
	γ	0.005	0.690	0.305	0.999	−0.011	0.538	0.265	0.999
		$n = 200$				$n = 400$			
POI and GEO	β_0	0.043	0.589	0.545	0.975	0.015	0.403	0.357	0.969
	β_1	−0.068	0.446	0.391	0.974	−0.030	0.309	0.286	0.965
	β_2	−0.012	0.076	0.072	0.966	−0.005	0.052	0.048	0.965
	α	−0.149	0.773	0.575	0.980	−0.054	0.549	0.368	0.983
	ν	0.042	0.272	0.230	0.958	0.013	0.191	0.157	0.971
	γ	−0.012	1.007	0.352	0.995	0.007	0.747	0.324	0.997
		$n = 600$				$n = 1000$			
	β_0	0.011	0.324	0.289	0.969	0.005	0.245	0.217	0.975
	β_1	−0.024	0.247	0.225	0.961	−0.015	0.188	0.167	0.976
	β_2	−0.002	0.042	0.038	0.963	−0.002	0.032	0.030	0.959
	α	−0.029	0.433	0.300	0.987	−0.016	0.330	0.247	0.990
	ν	0.001	0.150	0.127	0.973	0.002	0.114	0.099	0.979
	γ	−0.009	0.596	0.297	0.987	0.001	0.455	0.099	0.998
		$n = 200$				$n = 400$			
GEO and BELL	β_0	0.020	0.575	0.490	0.976	0.010	0.4003	0.316	0.982
	β_1	−0.069	0.422	0.382	0.970	−0.030	0.300	0.250	0.975
	β_2	−0.006	0.072	0.065	0.968	−0.003	0.052	0.045	0.964
	α	−0.147	0.706	0.489	0.987	−0.058	0.504	0.312	0.988
	ν	0.048	0.276	0.224	0.973	0.019	0.201	0.149	0.978
	γ	0.054	2.084	0.350	0.999	0.012	1.560	0.321	0.998
		$n = 600$				$n = 1000$			
	β_0	0.007	0.311	0.270	0.967	0.010	0.233	0.210	0.965
	β_1	−0.008	0.235	0.210	0.955	−0.002	0.177	0.162	0.969
	β_2	−0.002	0.041	0.036	0.961	−0.002	0.031	0.028	0.960
	α	−0.053	0.383	0.254	0.994	−0.023	0.284	0.193	0.993
	ν	0.017	0.155	0.124	0.982	0.009	0.116	0.093	0.982
	γ	0.013	1.171	0.302	0.985	−0.003	0.865	0.256	0.975

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Petrov B.N., Csáki F. (Eds.), Proceedings of the second international symposium on information theory*, pages 267–281.

Berkson, J. and Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47(259):501–515.

Borchers, H. W. (2022). *pracma: Practical Numerical Math Functions* R package version 2.3.8.

Cancho, V. G., Louzada, F. and Ortega, E. M. (2013). The Power Series Cure Rate Model: An Application to a Cutaneous Melanoma Data. *Communications in Statistics-Simulation and Computation*, 42(3):586–602

Chen, M.-H., Ibrahim, J., and Sinha, D. (1999). A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, 94(447):909–919.

Dempster, A. P., Laird. N. M, and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38

Dunn, P. K., & Smyth, G. K. (1996). Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, 5(3), 236–244.

Feller, W. (2008) An introduction to probability theory and its applications. *John Wiley & Sons.*

Gallardo, D. I., de Castro, M. and Gómez, H. W. (2021) An alternative promotion time cure model with overdispersed number of competing causes: An application to melanoma data. *Mathematics*, 9(15):1815,

Gallardo, D. I., Gómez, H. W., and Bolfarine, H. (2017a). A new cure rate model based on the Yule–Simon distribution with application to a melanoma data set. *Journal of Applied Statistics*, 44(7):1153–1164.

Gallardo, D. I., Gómez, Y. M., and de Castro, M. (2018). A flexible cure rate model based on the polylogarithm distribution. *Journal of Statistical Computation and Simulation*, 88(11):2137–2149.

Gallardo, D. I., Romeo, J. S., and Meyer, R. (2017b). A simplified estimation procedure based on the em algorithm for the power series cure rate model. *Communications in Statistics-Simulation and Computation*, 46(8):6342–6359.

Gómez, Y. M., Gallardo, D. I.,Leão, J. and Calsavara, V. F. (2021). On a new piecewise regression model with cure rate: Diagnostics and application to medical data. *Statistics in Medicine*, 40(29):6723–6742.

Ibrahim, J. G., Chen, M. H. and Sinha, D. (2001). *Bayesian survival analysis*, volume 2. Springer.

Kalbfleisch, J. D and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley.

Leão, J., Bourguignon, M., Gallardo, D. I., Rocha, R., and Tomazella, V. (2020) A new cure rate model with flexible competing causes with applications to melanoma and transplantation data. *Statistics in Medicine*, 39(24):3272–3284

Noack, A. (1950). A class of random variables with discrete distributions. *Annals of Mathematical Statistics*, 21:127–132.

Ortega, E. M., Cordeiro, G. M., Campelo, A. K., Kattan, M. W., Cancho, V. G. (2015). A power series beta weibull regression model for predicting breast carcinoma. *Statistics in medicine*, 34(8):1366–1388.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rodrigues, J. Cancho, V. G., de Castro, M., and Louzada-Neto, F. (2009). On the unification of long-term survival models. *Statistics & Probability Letters*, 79(6):753–759.

Rodrigues, J., de Castro, M., Cancho. V. G. and Balakrishnan, N. (2009). Com–poisson cure rate survival models and an application to a cutaneous melanoma data. *Journal of Statistical Planning and Inference*, 139(10):3605–3611.

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

Stram, D. and Lee, J. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50(4):1171–1177.

Yakovlev, A. Y. and Tsodikov, A. (1996). *Stochastic Models of Tumor Latency and their Biostatistical Applications*. World Scientific, New Jersey,US.

CHAPTER 3

Poisson-Birnbaum-Saunders mixture cure rate model

Resumo

Introduzimos uma nova modelagem para modelos de sobrevivência de longa duração, assumindo que o número de causas competitivas segue uma mistura da distribuição Poisson e Birnbaum-Saunders (BS) (Gonçalves et al., 2022). Neste contexto, apresentamos algumas propriedades estatísticas e demonstramos que o modelo do tempo de promoção surge como um caso limite. Aprofundamo-nos em discussões detalhadas de modelos específicos nesta classe. Particularmente, examinamos o número esperado de causas concorrentes, o qual depende de covariáveis. Permitindo uma modelagem direta da fração de cura como função destas covariáveis. Apresentamos um algoritmo do tipo EM, com a finalidade de discutir estimação por máxima verossimilhança e fornecer compreensão acerca da inferência dos parâmetros para este modelo. Além disso, delineamos condições suficientes para garantir a consistência e a normalidade assintótica dos estimadores de máxima verossimilhança. Para avaliar o desempenho do nosso método de estimação, conduzimos uma simulação de MC para estudar as propriedades assintóticas e examinamos o poder do teste da razão

de verossimilhanças, contrastando nossa metodologia com o modelo por tempo de promoção. Para demonstrar a relevância prática do nosso modelo, aplicamo-lo a um conjunto de dados médicos reais de um estudo populacional sobre a incidência de câncer de mama no estado de São Paulo, Brasil. Nossos resultados puderam ilustrar que o modelo proposto pode superar as abordagens tradicionais em termos de ajuste, destacando sua utilidade potencial em cenários do mundo real.

palavras-chave: Birnbaum-Saunders; Dados de câncer de mama; Causas concorrentes; Modelo de fração de cura; Algoritmo EM.

Abstract

We introduce a new modelling for long-term survival models, assuming that the number of competing causes follows a mixture of Poisson and the BS distribution (Gonçalves et al., 2022). In this context, we present some statistical properties of our model and demonstrate that the promotion time model emerges as a limiting case. We delve into detailed discussions of specific models within this class. Notably, we examine the expected number of competing causes, which depends on covariates. This allows for direct modeling of the cure rate as a function of covariates. We present an EM algorithm for parameter estimation, to discuss the estimation via maximum likelihood (ML) and provide insights into parameter inference for this model. Additionally, we outline sufficient conditions for ensuring the consistency and asymptotic normal distribution of ML estimators. To evaluate the performance of our estimation method, we conduct a MC simulation to provide asymptotic properties and a power study of LR test by contrasting our methodology against the promotion time model. To demonstrate the practical applicability of our model, we apply it to a real medical dataset from a population-based study of incidence of breast cancer in São Paulo, Brazil. Our results illustrate that the proposed model can outperform traditional approaches in terms of model fitting, highlighting its potential utility in real-world scenarios

keywords: Birnbaum-Saunders; Breast cancer data; Competing causes; Cure rate

model; EM algorithm.

3.1 Introduction

Cancer represents a significant global public health issue, as it is a leading cause of death and poses a major obstacle to the increase in life expectancy. In most countries, it ranks as the first or second leading cause of premature death before the age of 70. Both the incidence and mortality rates are rapidly increasing worldwide, driven by demographic and epidemiological transitions (Sung et al., 2021). The significant increase in disease rates directly reflects the lifestyle choices that most families have been adopting over time. The adoption of certain behavioral and environmental changes, such as dietary habits and exposure to environmental pollutants, contributes to the rise in cancer incidence and mortality. These factors also impact mobility, recreation, and overall structural conditions that influence health and quality of life (Wild et al., 2020).

Effective interventions have been implemented for the prevention, early detection, and treatment of the disease in countries with high human development indices. These efforts have had a substantial impact on reducing the incidence and mortality rates associated with cancer (Sung et al., 2021). According to the Global Cancer Estimates Observatory (GLOBOCAN), a web-based platform presenting global cancer statistics prepared by the International Agency for Research on Cancer (IARC), the impact of cancer on the world in 2020 was significant. There were 19.3 million new cancer cases worldwide (18.1 million if cases of non-melanoma skin cancer are excluded). This means that one in five individuals receive a cancer diagnosis during their lifetime (Ferlay et al. (2013); Sung et al.(2021)).

The long-term survivors of breast cancer patient have significantly improved over the past 50 years, due to advancements in the field of medical science and the introduction of new treatment approaches. As a result, an increasing number of patients are now considered “cured” or “immune” to the event of interest. It is anticipated that a certain percentage of patients will respond positively to treatment, leading to an improvement in overall survival. The long-term survival or cure rate models are specifically designed to account for this characteristic. It is essential to understand that

these models do not apply to overall survival since even if a patient is cured of a specific disease, they remain vulnerable to other diseases, making it impossible to achieve a complete cure for all illnesses.

Cure rate models are appropriate when there are individuals in the population who will never experience the event of interest. The pioneering model in this context was proposed by Berkson and Gage (1952), where it is assumed that there are two distinct groups: those who are immune and those who are susceptible to the event of interest. Chen et al. (1999) discussed an alternative model with a biological interpretation. In this model, the authors assumed the existence of some carcinogenic cells (latent variable), denoted as M for each individual. The classification of subjects into cured and susceptible categories is determined by $M = 0$ and $M \geq 1$, respectively. In their initial proposal, the authors considered M to follow a Poisson distribution with mean θ [$M \sim$ POI(θ)].

In the literature, various alternative models to M have been proposed. Notable examples include the negative binomial (NB) as particular cases and some well-known distributions such as the Bernoulli (BER), Binomial (Bin), Poisson (POI) and Geometric (GEO) (Rodrigues et al., 2009); COM-Poisson (Rodrigues et al., 2009); Power Series (PS) (Cancho et al., 2013); Yule-Simon (Gallardo et al., 2017); Polylogarithm (Gallardo et al., 2018); Zero-modified Geometric (ZMG) (Leão et al., 2020), compound Poisson (Gómez et al., 2023), mixture of power series (Brandão et al., 2023), among others.

An interesting class of models was discussed in Barreto-Souza (2015), where it is assumed that, conditional on a latent variable Z, $M \mid Z \sim$ POI(θZ). The author considered the Exponential Family (EF) of distributions for Z with mean 1, which encompasses a wide class of models, including the gamma, inverse gaussian, and generalized hyperbolic secant, among others. In this chapter, we explore a similar concept, but we consider the BS distribution for Z. The BS model does not belong to the EF, but possesses many interesting properties: it can be directly parameterized in terms of the mean, it has a moment-generating function in a closed and simple form, and it can be expressed as a mixture of distributions, among other characteristics. The proposed model presents itself as a compelling alternative to the widely acknowledged NB model. Both models exhibit the common feature of overdispersion of simultaneous

causes in relation to the mean. This incorporation not only addresses overdispersion but also introduces versatility to the array of modeling options at our disposal.

The chapter is organized as follows. In Section 3.2, we introduce the Poisson-Birnbaum-Saunders (PBS) mixture cure rate model . Section 3.3 provides a comprehensive review of the maximization of the log-likelihood function for this model, and we propose an estimation procedure based on the EM algorithm. The performance of our proposed model is thoroughly examined in Section 3.4 through two simulation studies. To illustrate the practical application of the methodology, in Section 3.5, we analyze a dataset comprising survival times of patients with breast cancer in the state of São Paulo, Brazil. Finally, in Section 3.6, we present a detailed discussion of the main findings and implications of this study.

3.2 The proposed model

In this Section, we provide an overview of the BS distribution and introduce our proposed modeling approach.

3.2.1 Birnbaum-Saunders model (BS)

The BS distribution has been widely considered in the literature due to its physical arguments, favourable statistical properties, and its connection with the normal distribution. The BS model was proposed by Birnbaum and Saunders (1969) and has been extensively applied for modelling failure times in engineering. However, novel applications have emerged in biological, environmental and financial sciences as well; for instance, Desmond (1985), Kotz et al. (2010), Saulo et al. (2013) and Leiva et al. (2014a),Leiva et al. (2014b),Leiva et al. (2015a),Leiva et al. (2015b),Leiva et al. (2017).

In the context of the BS distribution, Santos-Neto et al. (2012) introduced various parameterizations. One such parameterizations is defined by the parameters $\mu = \beta(1 + \alpha^2/2)$ and $\phi = 2/\alpha^2$, where $\alpha > 0$ and $\beta > 0$ are the original BS parameters (Birnbaum and Saunders, 1969), $\mu > 0$ is a scale parameter and represents the mean of the distribution, while $\phi > 0$ acts as a shape and precision parameter. We use the

notation $Z \sim \mathrm{BS}(\mu, \phi)$ to denote a random variable (RV) Z following this distribution. If $Z \sim \mathrm{BS}(\mu, \phi)$, its PDF is as follows

$$f_Z(z;\mu,\phi) = \frac{\exp(\phi/2)\sqrt{\phi+1}}{4\,z^{\frac{3}{2}}\,\sqrt{\pi\mu}}\left(z+\frac{\phi\mu}{\phi+1}\right) \times \exp\left(-\frac{\phi}{4}\left(\frac{z(\phi+1)}{\phi\mu}+\frac{\phi\mu}{z(\phi+1)}\right)\right), \quad z>0. \tag{3.1}$$

For the particular case $\mu = 1$ (which will be of our interest) and defining $a = (\phi+1)/2$ and $b = \frac{1}{2}\phi^2/(\phi+1)$, we have that

$$\begin{aligned} f_Z(z;1,\phi) &= \frac{1}{2}\frac{\exp(\phi/2)(\phi+1)^{1/2}}{2\sqrt{\pi}} z^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(za+b/z\right)\right) \\ &\quad + \frac{1}{2}\frac{\exp(\phi/2)(\phi+1)^{1/2}}{2\sqrt{\pi}} z^{-\frac{3}{2}} \left(\frac{\phi}{\phi+1}\right) \exp\left(-\frac{1}{2}\left(za+b/z\right)\right) \\ &= \frac{1}{2}\left[\text{GIG}\left(a,b,\frac{1}{2}\right)+\text{GIG}\left(a,b,-\frac{1}{2}\right)\right], \end{aligned} \tag{3.2}$$

where GIG(a,b,p) denotes the generalized inverse gamma distribution with PDF given by

$$f(z;a,b,p) = \frac{(a/b)^{p/2}}{2\mathcal{K}_p(\sqrt{ab})} z^{p-1} \exp\left(-\frac{1}{2}\left(az+b/z\right)\right).$$

The result in (3.2) indicates that the BS model is a mixture of two GIG distributions. The supposition on the distribution BS with $\mu = 1$ and $\phi > 0$ is done to ensure identifiability to the model, which will be very useful in the estimation process that will be developed and does not reduce the applicability of the model in practical cases. This results was also presented in Equation (22) of Balakrishnan and Kundu (2019).

From (3.1) we can obtain the survival (SF) and hazard rate (HR) functions of Z which are, respectively,

$$\begin{aligned} S_Z(z;\mu,\phi) &= \frac{1}{2}\Phi\left((z+\phi(z-\mu))/(2\sqrt{z(1+\phi)\mu})\right), \quad z>0, \\ h_Z(z;\mu,\phi) &= \frac{\exp\left(-(-\phi\mu+\phi z+z)^2/(4(\phi+1)\mu z)\right)(\phi\mu+\phi z+z)}{(\pi\mu(\phi+1))^{\frac{1}{2}}2\mu^{\frac{1}{2}}\,z^{\frac{3}{2}}\Phi\left((z+\phi(z-\mu))/(2\sqrt{z(1+\phi)\mu})\right)}, \quad z>0, \end{aligned}$$

where Φ is the N(0, 1) cumulative distribution function (CDF). Finally, the moment generating function (MGF) for the BS distribution can be expressed as

$$M_Z(t) = \frac{1}{2}\left(1 + \sqrt{\frac{\phi+1}{\phi+1-4t\mu}}\right)\exp\left(\frac{\phi\left(\sqrt{\phi+1}-\sqrt{1+\phi-4t\mu}\right)}{2\sqrt{1+\phi}}\right).$$

3.2.2 Poisson-Birnbaum-Saunders mixture model

In this subsection, we introduce a novel cure rate model based on a mixture of the Poisson and BS distributions. The PBS mixture model was proposed by Gonçalves et al. (2022). In addition to exploring its properties, we also discuss a method for generating values from this model.

Let M be an unobserved variable denoting the initial number of competing causes related to the occurrence of an event of interest. In a medical context, such as with cancer patients, M represents the number of carcinogenic cells in patients undergoing cancer treatment. We assume that, conditional on $Z = z$, $M \mid Z = z \sim$ POI(θz). We further assume that $Z \sim$ BS$(1, \phi)$, i.e., $\mathbb{E}(Z) = 1$ and Var$(Z) = (2\phi + 5)/(\phi + 1)^2$. It is straightforward to see that $\lim_{\phi\to\infty}$ Var$(Z) = 0$ and then, Z is degenerated at 1 when $\phi \to \infty$.

Under this scheme, $\mathbb{E}(M) = \mathbb{E}(\mathbb{E}(M|Z)) = \theta$ and Var$(M) = \mathbb{E}($Var$(M|Z)) +$ Var$(\mathbb{E}(M|Z)) = \mathbb{E}(\theta Z) +$ Var$(\theta Z) = \theta + \theta^2(2\phi + 5)/(\phi + 1)^2 > \theta$, i.e., the distribution of M is over-dispersed. Furthermore, we can readily compute the PGF of M as follows

$$\begin{aligned}
G_M(s) &= \mathbb{E}(s^M) = \mathbb{E}\left(\mathrm{e}^{M\log(s)}\right) = M_M(\log s) = \mathbb{E}(M_M(\log(s))|Z) \\
&= \mathbb{E}\left(\mathrm{e}^{\theta Z(s-1)}\right) = M_Z(\theta(s-1)) \\
&= \frac{1}{2}\left(1 + \sqrt{\frac{\phi+1}{\phi+1+4\theta(1-s)}}\right)\exp\left(\frac{\phi\left(\sqrt{\phi+1}-\sqrt{1+\phi+4\theta(1-s)}\right)}{2\sqrt{\phi+1}}\right),
\end{aligned}$$

where $M_M(\cdot)$ represents the moment generating function of the distribution for the variable M for competing causes.

The usual scheme here is the assumption that $V_1, \ldots, V_m$, the time to produce a detectable cancer for each of the m carcinogenic cells, are conditionally independent

given $M = m$ with common SF $S(\cdot;\boldsymbol{\eta})$. In addition, if $M = 0$ the individual is considered as cured an then it is defined $P(V_0 = \infty) = 1$. With those notations, the time-to-event for the individual can be represented as $T = \min(V_0, V_1, \ldots, V_M)$. Under this usual competing risks framework, and per Theorem 2 in Rodrigues et al. (2009), we have that the (improper) population SF and PDF of the PBS mixture cure rate model are given by

$$\begin{aligned} S_{pop}(t;\theta,\phi,\boldsymbol{\eta}) &= \Pr(T > t;\theta,\phi,\boldsymbol{\eta}) = G_M\big(S(t;\boldsymbol{\eta});\theta,\phi\big) \\ &= \frac{1}{2}\left(1 + \sqrt{1 - \frac{4\theta(1 - S(t;\boldsymbol{\eta}))}{\phi + 1 + 4\theta(1 - S(t;\boldsymbol{\eta}))}}\right) \\ &\quad \times \exp\left\{\frac{\phi}{2}\left(1 - \frac{\sqrt{\phi + 1 + 4\theta(1 - S(t;\boldsymbol{\eta}))}}{\sqrt{\phi + 1}}\right)\right\} \end{aligned} \tag{3.3}$$

and

$$f_{pop}(t;\theta,\phi,\boldsymbol{\eta}) = \frac{f(t;\boldsymbol{\eta})\theta u^{-3/2}\exp\left\{\frac{\phi}{2}\left[1 - \sqrt{u}\right]\right\}}{\phi + 1}\left[1 + \frac{\phi}{2}u(1 + u^{-1/2})\right], \tag{3.4}$$

where $u = 4\theta(1 - S(t;\boldsymbol{\eta}))/(\phi + 1)$ and $S(t,\boldsymbol{\eta})$ and $f(t,\boldsymbol{\eta})$ are the SF and PDF of time-to-event and $\boldsymbol{\eta}$ denotes a vector of unknown parameters. We assume that the time to the event of interest follows a Weibull distribution with $\boldsymbol{\eta} = (\alpha, \nu)$ unknown parameter vector. An important detail concerning the relationship between the Poisson-Birnbaum-Saunders mixture model and the Poisson model with mean θ is that when the ϕ parameter from the BS distribution in our proposal tends to infinity, the population SF in Equation (3.3), denoted as $S_{pop}(t;\theta,\phi,\boldsymbol{\eta})$, converges in the limit to the population SF $S_{pop}(t;\theta,\boldsymbol{\eta})$ of the Cure Rate Poisson model (also knows in the literature as the promotion time cure rate model). In other words, $\lim_{\phi\to\infty} S_{pop}(t;\theta,\phi,\boldsymbol{\eta}) = \mathrm{e}^{-\theta(1-S(t;\boldsymbol{\eta}))}$, as described in Rodrigues et al. (2009). It is immediate that the cure rate of the model is given by

$$\begin{aligned} p &= \lim_{t\to\infty} S_{pop}(t;\theta,\phi,\boldsymbol{\eta}) \\ &= \frac{1}{2}\left(1 + \sqrt{\frac{\phi + 1}{\phi + 1 + 4\theta}}\right)\exp\left(\frac{\phi\left(\sqrt{\phi + 1} - \sqrt{1 + \phi + 4\theta}\right)}{2\sqrt{1 + \phi}}\right). \end{aligned} \tag{3.5}$$

For heterogeneous populations with varying characteristics, we can introduce explanatory variables into the cure rate using the cure parameter θ from Poisson distribution in our mixing approach for this model. When these factors are integrated, a distinct cure rate parameter is assigned to each patient or subject, represented as p_i, where i ranges from 1 to n, being n the number of individuals or subjects in the study. To capture the influence of these explanatory factors on the cure rate, different link functions can be employed.

Remark 3.2.1. *Hashimoto et al. (2014) present a model named Poisson Birnbaum-Saunders. Such a model corresponds to considering $M \sim$ POI(θ) and $S(t;\boldsymbol{\eta})$ as the SF for the BS distribution. Despite the similarity in name, for our approach PBS mixture model considers a very different assumption than in the aforementioned work, namely $M \mid Z = z \sim$ POI$(\theta z), Z \sim$ BS$(1, \phi)$ and, up to this moment, any particular choice for $S(t;\boldsymbol{\eta})$.*

3.3 Estimation

In this Section, we focus on estimating the model parameters. Let us consider the situation when the time to an event is not completely observed and is subject to right censoring. Let c_i be the censoring time for the ith individual and t_i^* being the failure time. We observe $t_i = \min(t_i^*, c_i)$ and $\delta_i = \mathbb{I}(t_i \leq c_i)$, where $\delta_i = 1$ if t_i is a time-to-event and $\delta_i = 0$ if t_i is right-censored, for $i = 1, \ldots, n$. Based on the observed vectors $\mathsf{D}_{obs} = ((t_1, \delta_1, \mathbf{x}_1^\top)^\top, \ldots, (t_n, \delta_n, \mathbf{x}_n^\top)^\top)$, where $\mathbf{x}_i^\top$ is the covariate vector of dimension $(q+1) \times 1$ related to the cure of the ith individual. These covariates on the cure fraction p_i in (3.5) can be modeled via a link function $g(\cdot)$ in θ_i. In order to deal with the effect of the explanatory variables on the cure, let $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_q)^\top$ be the vector of regression coefficients to be estimated. Note that $\boldsymbol{\beta}$ is related to explanatory variables with observed values for the patient i denoted by $\mathbf{x}_i = (1, x_{1i}, \ldots, x_{qi})^\top$, which are associated with the cured fraction. Observe that different kinds of link functions can be considered, so that, the choice of the link function depends on the parameter space. In this particular case, the transformation $\theta_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$ have been used to that end. The

variables $\mathbf{M} = (M_1, \ldots, M_n)^\top$, and $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$ are non-observable and thus the complete data are denoted through vector $\mathsf{D}_{comp} = (\mathbf{t}, \boldsymbol{\delta}, \mathbf{x}, \mathbf{M}, \mathbf{Z})$.

To obtain the estimates for $\boldsymbol{\psi} = (\boldsymbol{\eta}^\top, \boldsymbol{\beta}, \phi)$, where $\boldsymbol{\eta} = (\alpha, \nu)^\top$, we can use the corresponding log-likelihood function method under uninformative censoring that is expressed as

$$\ell(\boldsymbol{\psi}|\mathsf{D}_{obs}) = \sum_{i=1}^{n} \left[\delta_i \log f_{pop}(t_i, \boldsymbol{\psi}) + (1 - \delta_i) \log S_{pop}(t_i, \boldsymbol{\psi})\right]. \tag{3.6}$$

To obtain the ML estimators, it is necessary to maximize (3.6) in relation to $\boldsymbol{\psi}$, i.e., a maximization of dimension $q + 3$. In the following subsection, we discuss an EM-type algorithm in order to provide a more attractive and robust estimation procedure.

3.3.1 EM algorithm

Let us focus on estimating the model parameters when it involves incomplete data, latent variables or missing data by using the ML method proposed by Dempster et al. (1977). The EM algorithm is commonly employed to handle ML estimates of the parameters of interest. It uses incomplete data to deal with the estimation process. This algorithm iteratively intend of the conditional distribution of the latent variables given the observed data and actual parameter estimates in the E-step to obtain ML estimates of the parameters. Thereafter, in the M-step, this conditional expectation is maximized to obtain ML estimates of the parameters studied.

To derive the formula for the E-step, the following Proposition and Corollary can be employed. The proofs of the latter can be found in Appendix A.

Proposition 3.3.1. *For the PBS model, the conditional distribution of i)* $M_i \mid z_i, t_i, \delta_i$*; ii)* $Z_i \mid y_i, t_i, \delta_i$ *and; iii)* $Y_i \mid t_i, \delta_i$ *are, respectively, given by*

$$\begin{aligned} M_i - \delta_i \mid z_i, t_i, \delta_i &\sim \textit{POI}\left(\theta_i z_i S(t_i; \boldsymbol{\eta})\right), \\ Z_i \mid y_i, t_i, \delta_i &\sim \textit{GIG}\left(p_i(y_i), a_i, b_i\right), \ \textit{and} \\ Y_i \mid t_i, \delta_i &\sim \textit{BER}\left(\omega_i\right), \end{aligned}$$

with $p_i(y_i) = \delta_i - y_i + 1/2$, $a_i = a_i(\phi) = 2\theta_i + (\phi+1)/2$, $b_i = b_i(\phi) = \phi^2/[2(\phi+1)]$, *and*

$$\omega_i = \frac{\mathcal{K}_{p_i(1)}(\sqrt{a_i b_i})\left(\frac{\phi}{\phi+1}\right)^{1+p_i(1)}}{\mathcal{K}_{p_i(0)}(\sqrt{a_i b_i})\left(\frac{\phi}{\phi+1}\right)^{p_i(0)} + \mathcal{K}_{p_i(1)}(\sqrt{a_i b_i})\left(\frac{\phi}{\phi+1}\right)^{1+p_i(1)}}, \tag{3.7}$$

where $\mathcal{K}_{p_i(\cdot)}$ *is a modified Bessel function of the second kind Abramowitz and Stegun (1972).*

Corollary 3.3.1. *The expected value for* Y_i, Z_i, Z_i^{-1} *and* M_i *given* (t_i, δ_i) *are respectively*

$$\begin{aligned}
\mathbb{E}(Y_i \mid t_i, \delta_i) &= \omega_i, \\
\mathbb{E}(Z_i \mid t_i, \delta_i) &= \frac{\sqrt{b_i}\mathcal{K}_{p_i(0)+1}(\sqrt{a_i b_i})}{\sqrt{a_i}\mathcal{K}_{p_i(0)}(\sqrt{a_i b_i})}(1-\omega_i) + \frac{\sqrt{b_i}\mathcal{K}_{p_i(1)+1}(\sqrt{a_i b_i})}{\sqrt{a_i}\mathcal{K}_{p_i(1)}(\sqrt{a_i b_i})}\omega_i, \\
\mathbb{E}(Z_i^{-1} \mid t_i, \delta_i) &= \left[\frac{\sqrt{a_i}\mathcal{K}_{p_i(0)+1}(\sqrt{a_i b_i})}{\sqrt{b_i}\mathcal{K}_{p_i(0)}(\sqrt{a_i b_i})} - \frac{2p_i(0)}{b_i}\right](1-\omega_i) + \left[\frac{\sqrt{a_i}\mathcal{K}_{p_i(1)+1}(\sqrt{a_i b_i})}{\sqrt{b_i}\mathcal{K}_{p_i(1)}(\sqrt{a_i b_i})} - \frac{2p_i(1)}{b_i}\right]\omega_i, \\
\mathbb{E}(M_i \mid t_i, \delta_i) &= \delta_i + \theta_i S(t_i; \boldsymbol{\eta})\mathbb{E}[Z_i \mid t_i, \delta_i].
\end{aligned}$$

More details of the results presented above are provided in Appendix section.

The complete log-likelihood for $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\eta}, \phi)$, with $\boldsymbol{\eta} = (\alpha, \nu)^\top$ and and thus the complete data are denoted by $\mathsf{D}_{comp} = (\mathbf{t}, \boldsymbol{\delta}, \mathbf{x}, \mathbf{M}, \mathbf{M}_2, \mathbf{Y})$ is given by

$$\ell(\boldsymbol{\psi}; \mathsf{D}_{comp}) = \ell_1(\boldsymbol{\eta}; \mathsf{D}_{comp}) + \ell_2(\boldsymbol{\beta}; \mathsf{D}_{comp}) + \ell_3(\phi; \mathsf{D}_{comp}), \tag{3.8}$$

where

$$\begin{aligned}
\ell_1(\boldsymbol{\eta}; \mathsf{D}_{comp}) &= \sum_{i=1}^{n}\left[(M_i - \delta_i)\log S(t_i; \boldsymbol{\eta}) + \delta_i \log f(t_i; \boldsymbol{\eta})\right], \\
\ell_2(\boldsymbol{\beta}; \mathsf{D}_{comp}) &= \sum_{i=1}^{n}\left[M_i \log\theta_i - Z_i\theta_i\right], \quad \text{and} \\
\ell_3(\phi; \mathsf{D}_{comp}) &= \frac{n}{2}\left[\phi + \log(1+\phi)\right] + \sum_{i=1}^{n} Y_i\left[\log\phi - \log(1+\phi)\right] \\
&\quad - \frac{1}{4}\sum_{i=1}^{n}\left[(1+\phi)Z_i + \frac{\phi^2}{Z_i(1+\phi)}\right].
\end{aligned}$$

Let $\boldsymbol{\psi}^{(k)}$ be the estimate of $\boldsymbol{\psi}$ at the kth iteration and denote $Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}^{(k)})$ as the conditional expectation of $\ell(\boldsymbol{\psi}; \mathsf{D}_{comp})$ in (3.8) given the observed data and $\boldsymbol{\psi}^{(k)}$. Then this conditional expectation can be decomposed as

$$Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}^{(k)}) = Q_1(\boldsymbol{\eta} \mid \boldsymbol{\psi}^{(k)}) + Q_2(\boldsymbol{\beta} \mid \boldsymbol{\psi}^{(k)}) + Q_3(\phi \mid \boldsymbol{\psi}^{(k)}),$$

with

$$Q_1(\boldsymbol{\eta} \mid \boldsymbol{\psi}^{(k)}) = \sum_{i=1}^{n} \left[\left(\widetilde{M}_i^{(k)} - \delta_i \right) \log S(t_i; \boldsymbol{\eta}) + \delta_i \log f(t_i; \boldsymbol{\eta}) \right], \tag{3.9}$$

$$Q_2(\boldsymbol{\beta} \mid \boldsymbol{\psi}^{(k)}) = \sum_{i=1}^{n} \left[\widetilde{M}_i^{(k)} \log \theta_i - \widetilde{Z}_i^{(k)} \theta_i \right], \tag{3.10}$$

$$\begin{aligned} Q_3(\phi \mid \boldsymbol{\psi}^{(k)}) = {} & \frac{n}{2} \left[\phi + \log(1+\phi) \right] + \sum_{i=1}^{n} \widetilde{Y}_i^{(k)} \left[\log(\phi) - \log(1+\phi) \right] \\ & - \frac{1}{4} \sum_{i=1}^{n} \left[(1+\phi) \widetilde{Z}_i^{(k)} + \frac{\widetilde{\kappa}_i^{(k)} \phi^2}{(1+\phi)} \right], \end{aligned} \tag{3.11}$$

where $\widetilde{M}_i^{(k)} = \mathbb{E}(M_i \mid t_i, \delta_i, \boldsymbol{\psi} = \widehat{\boldsymbol{\psi}}^{(k)})$, $\widetilde{Z}_i^{(k)} = \mathbb{E}(Z_i \mid t_i, \delta_i, \boldsymbol{\psi} = \widehat{\boldsymbol{\psi}}^{(k)})$, $\widetilde{\kappa}_i^{(k)} = \mathbb{E}(Z_i^{-1} \mid t_i, \delta_i, \boldsymbol{\psi} = \widehat{\boldsymbol{\psi}}^{(k)})$ and $\widetilde{Y}_i^{(k)} = \mathbb{E}(Y_i \mid t_i, \delta_i, \boldsymbol{\psi} = \widehat{\boldsymbol{\psi}}^{(k)})$. Note that all the expected values required can be computed using corollary 1 and the functions Q_1, Q_2 and Q_3 depend only on $\boldsymbol{\eta}$, $\boldsymbol{\beta}$ and ϕ, respectively.

In short, the kth iteration of the EM algorithm is given by

- E-step: Following the Corollary 1, for $i = 1, \ldots, n$, update the values of the following latent variables: $\widetilde{M}_i^{(k)}$, $\widetilde{Z}_i^{(k)}$, $\widetilde{\kappa}_i^{(k)}$ and $\widetilde{Y}_i^{(k)}$.
- M-step: Given the actual values of $\widetilde{M}_i^{(k)}$, $\widetilde{Z}_i^{(k)}$, $\widetilde{\kappa}_i^{(k)}$ and $\widetilde{Y}_i^{(k)}$, find the values of $\boldsymbol{\eta}^{(k)}$, $\boldsymbol{\beta}^{(k)}$ and $\phi^{(k)}$ that maximizes (3.9), (3.10) and (3.11), in relation to $\boldsymbol{\eta}$, $\boldsymbol{\beta}$ and ϕ, respectively.

The E-step and M-step are performed iteratively until a predefined convergence criterion is met, specifically when the difference between consecutive estimates reaches a predetermined tolerance level. Conversely, the standard errors for the estimator $\widehat{\boldsymbol{\psi}}$ can be derived from the Hessian matrix of the observed log-likelihood function in (3.6),

which is given by

$$\boldsymbol{\Sigma}(\widehat{\boldsymbol{\psi}}) = -\frac{\partial^2 \ell(\boldsymbol{\psi}|\mathbf{D}_{\text{obs}})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^\top}\bigg|_{\boldsymbol{\psi}=\widehat{\boldsymbol{\psi}}}.$$

This matrix can be computed using the `hessian` function included in the `pracma` (Borchers, 2022) package of the R Core Team (2024) software. Under appropriated regularity conditions, it was shown by Kalbfleisch and Prentice (2022) that the asymptotic distribution of the estimator $\widehat{\boldsymbol{\psi}}$ follows:

$$\sqrt{n}\left[\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\psi})\right]^{-1/2}\left(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right) \xrightarrow{\mathcal{D}} N_{q+3}(\mathbf{0}_{q+3}, \mathbf{I}_{q+3}), \qquad \text{as } n \to \infty, \tag{3.12}$$

where $\mathbf{0}_q$ represents a vector of zeros with a dimension of q, and $\mathbf{I}_q$ denotes the identity matrix of order q. In addition, if $\widehat{\sigma}^2_\phi$ denotes the estimated variance of $\widehat{\phi}$, then by the delta method (Hajek et al., 1999), for $\theta = g(\phi) = (2\phi + 5)/(\phi + 1)^2$, it is obtained

$$\frac{\sqrt{n}\left(\widehat{\phi} - \phi\right)\left(\widehat{\phi} + 1\right)^2}{\sqrt{2}\widehat{\sigma}_\phi\sqrt{(\widehat{\phi} + 1)^2 - (2\widehat{\phi} + 5)(\widehat{\phi} + 1)}} \xrightarrow{\mathcal{D}} N\left(0, 1\right). \tag{3.13}$$

Results in Eq. (3.12) and (3.13) allows to build confidence intervals for each parameter and/or θ.

3.4 Monte Carlo simulation studies

In this Section, we present the results of two simulation studies. The first is related to assessing the performance of the ML estimates for the PBS model through the EM algorithm. The second study is devoted to evaluating the performance of the likelihood ratio (LR) test to decide between the PBS model and the traditional promotion time cure rate model.

3.4.1 Asymptotic properties

In this study, we assess the effectiveness of parameter estimation using the EM algorithm by recovering parameter values for simulated datasets. To facilitate the study's conduction, the following data structure was created. The time-to-event values were drawn from the Weibull distribution with fixed parameters $\alpha = -3.52$ and $\nu = 1.4$. As our investigation involves studying covariates within the cure fraction, we generated a sample for the number of competing causes using the PBS mixture model, with fixed regression coefficients $\beta_0 = -1.67$, $\beta_1 = 1.21$, $\beta_2 = 2.53$, and $\phi = 1.63$, which provides Var$(Z) = 1.38$. To evaluate different sample sizes (n), set at 200, 400, 600, 800, 1000, 1200, 1400, 1600 and 5000, we conducted a MC study comprising 1,000 replications for each size.

For each individual, a categorical covariate with three levels was considered. This variable was denoted as x_{11i}, x_{12i}, and x_{13i} for $i = 1, \ldots, n$. The values of these covariates were sampled from a Multinomial distribution with probabilities 0.15, 0.26, and 0.59, respectively. The censoring times for all sample sizes were drawn from a Uniform distribution between 0 and 20, resulting in an average censoring percentage of approximately 20%. The selected parameter values were approximated from the estimates obtained for the application of our proposed model in the next section.

For each parameter value and sample size, we presented the empirical estimates for the standard deviation of $\boldsymbol{\psi} = (\beta_0, \beta_1, \beta_2, \alpha, \nu, \phi)^\top$, as well as the estimated bias and root mean squared error (RMSE) of the ML estimators and and the CP of the asymptotic 95% confidence intervals, all based on the asymptotic distribution given by Equation (3.12). The results are shown in Table 3.1. The standard error used to compute the RMSE was obtained using the Hessian matrix computed using the `hessian` function included in the `pracma` package (Borchers ,2022) of R Core Team (2024), considering the asymptotic distribution of the ML estimators based on EM estimates.

The performance evaluation of the proposed model is based on results obtained in a MC study. Table 3.1 summarizes the simulation study of model parameter estimates from 1,000 replicates of experiments. Evaluating the estimates as the sample size increases, the biases and the RMSEs decrease for most cases. This shows us the efficiency of the ML estimates of the proposed model. The estimate of variance for the

Table 3.1: Empirical standard deviation (SD), Bias, Root of MSE and CP of the ML estimators for PBS mixture model using the Weibull distribution to time-to-event in the concurrent causes regression.

Sample size	Parameter	β_0	β_1	β_2	α	ν	ϕ	$Var(Z)$
	Real Value	-1.7	1.21	2.53	-3.52	1.40	1.36	1.386
200	Bias	-0.157	0.300	0.446	-0.331	0.088	8.626	0.333
	RMSE	1.440	1.520	1.585	0.807	0.237	32.488	1.167
	EM SD	1.432	1.490	1.521	0.736	0.220	31.322	1.118
	CP	0.994	0.971	0.970	0.940	0.961	0.809	0.749
400	Bias	-0.009	0.072	0.141	-0.155	0.042	5.194	0.141
	RMSE	0.448	0.492	0.592	0.577	0.167	20.010	0.947
	EM SD	0.448	0.487	0.575	0.556	0.162	19.324	0.936
	CP	0.983	0.971	0.956	0.920	0.945	0.855	0.729
600	Bias	-0.023	0.026	0.067	-0.064	0.021	3.726	0.019
	RMSE	0.361	0.390	0.473	0.457	0.133	13.470	0.810
	EM SD	0.360	0.389	0.469	0.452	0.131	12.944	0.810
	CP	0.974	0.961	0.952	0.927	0.948	0.892	0.724
800	Bias	-0.010	0.023	0.041	-0.040	0.015	2.426	0.024
	RMSE	0.316	0.347	0.412	0.392	0.111	11.095	0.714
	EM SD	0.316	0.346	0.410	0.390	0.110	10.826	0.714
	CP	0.962	0.950	0.954	0.941	0.955	0.898	0.749
1000	Bias	-0.024	0.037	0.041	-0.032	0.012	1.796	0.001
	RMSE	0.274	0.299	0.360	0.349	0.096	7.792	0.656
	EM SD	0.273	0.297	0.357	0.348	0.095	7.583	0.656
	CP	0.968	0.961	0.959	0.947	0.949	0.920	0.757
1200	Bias	-0.011	0.023	0.020	-0.017	0.007	1.290	-0.018
	RMSE	0.254	0.272	0.325	0.317	0.093	6.452	0.602
	EM SD	0.254	0.271	0.324	0.317	0.092	6.321	0.602
	CP	0.960	0.955	0.958	0.949	0.940	0.925	0.759
1400	Bias	-0.014	0.016	0.015	-0.012	0.005	0.932	-0.035
	RMSE	0.228	0.252	0.319	0.297	0.083	3.955	0.570
	EM SD	0.228	0.252	0.319	0.297	0.083	3.843	0.569
	CP	0.969	0.954	0.944	0.960	0.952	0.933	0.767
1600	Bias	-0.008	-0.001	-0.004	0.003	0.001	0.774	-0.038
	RMSE	0.209	0.225	0.274	0.276	0.077	3.621	0.535
	EM SD	0.208	0.225	0.273	0.276	0.077	3.537	0.534
	CP	0.960	0.959	0.956	0.957	0.951	0.944	0.771
5000	Bias	0.002	-0.006	-0.007	-0.002	0.002	0.112	-0.016
	RMSE	0.117	0.131	0.159	0.153	0.043	0.501	0.300
	EM SD	0.117	0.131	0.159	0.153	0.043	0.488	0.300
	CP	0.958	0.956	0.960	0.963	0.962	0.962	0.829

random variable Z, which is a function only of the ϕ parameter, is more efficient for estimating the dispersion of data generated from the model, yielding favorable results across all studied sample sizes. When analyzing all the regression coefficient estimates, the average bias approaches zero as the sample size increases.

The standard deviation (SD) and the root mean squared error (RMSE) are closer to each other, it suggests that the standard errors of parameters are well estimated. The biases of the time-to-event distribution parameters α and ν become smaller as the sample size n increases. Furthermore, it can be noted that these estimated values are greater compared to the ones considered in the regression structure. This happened

because the values chosen for the simulation study were larger, in absolute terms, than the ones for the $\boldsymbol{\beta}$ vector of regression parameters. Last but not least, Despite the estimated values for $\mathrm{Var}(Z)$, which are greatly influenced by the estimated values for ϕ, the CP's for all scenarios studied were close to the nominal value (95%). Additional MC simulation results for different set of parameters can be found in Tables 3.8 and 3.7 of Appendix B.

In addition to the simulation study presented in Table 3.1 above and considering the relationship between the PBS mixture model and the promotion time model through the parameter values ϕ, a more challenging scenario for parameter estimation can be observed in Table 3.2. In this study, 1000 MC replicates were generated for fixed sample sizes $n = 5000$, varying the true values of ϕ in $\{0.5, 1, 5, 10\}$ while fixing the other model parameters at $\beta_0 = -1.67$, $\beta_1 = 1.21$, $\beta_2 = 2.53$, $\alpha = -3.52$, and $\nu = 1.40$. With the obtained EM estimated values, it was observed that, as expected, the estimation of the precision parameter ϕ becomes more biased as the values of ϕ increase.

Additionally, the standard deviation (SD) and the root mean squared error (RMSE) are closer to each other, suggesting that the standard errors of parameters are well estimated. Each simulated scenario for the values of ϕ considered in the study had CP's close to the nominal level (95%). The study suggests that the estimates of the other parameters were not relatively affected as the parameter ϕ increases.

3.4.2 Hypothesis testing

In this subsection, a MC study was conducted to evaluate the performance of the LR test in comparing the proposed PBS mixture model with the Poisson model when the parameter ϕ increases. The main purpose of this simulation study is to assess the performance of our proposed model for different values for ϕ within the parameter space. Drawing inspiration from the findings of Barreto-Souza (2015), where the promotion model is presented as a limiting case, the focus here is on investigating the model's efficacy through the likelihood ratio (LR) test. Our exploration centers on testing the

Table 3.2: Empirical Bias, Root of MSE, standard deviation (SD) and CP of the ML estimators for PBS mixture model using the Weibull distribution to time-to-event in the concurrent causes regression for variations of ϕ parameter.

Real value	**Estimates**	ϕ	β_0	β_1	β_2	α	ν
$\phi = 0.5$	Bias	0.022	-0.009	0.004	-0.001	0.003	0.000
	RMSE	0.125	0.136	0.154	0.160	0.131	0.041
	EM SD	0.123	0.136	0.154	0.160	0.131	0.041
	CP	0.985	0.957	0.948	0.973	0.978	0.975
$\phi = 1$	Bias	0.072	-0.010	0.005	0.004	0.006	-0.001
	RMSE	0.336	0.122	0.134	0.156	0.152	0.042
	EM SD	0.329	0.121	0.134	0.156	0.152	0.042
	CP	0.972	0.957	0.956	0.968	0.972	0.969
$\phi = 5$	Bias	1.480	-0.005	0.009	0.017	-0.010	0.003
	RMSE	5.946	0.111	0.117	0.149	0.144	0.039
	EM SD	5.759	0.111	0.117	0.148	0.144	0.039
	CP	0.908	0.948	0.948	0.953	0.951	0.945
$\phi = 10$	Bias	3.864	0.002	0.011	0.021	-0.024	0.006
	RMSE	14.368	0.100	0.109	0.133	0.124	0.033
	EM SD	13.839	0.100	0.109	0.131	0.121	0.032
	CP	0.859	0.970	0.960	0.960	0.947	0.969

hypothesis:

$$\begin{cases} H_0: & \text{the POI model is the true ,} \\ H_1: & \text{the PBS mixture model is the true .} \end{cases}$$

The LR test allows to discern substantial deviations from the null hypothesis, positing the PBS mixture model as the true model. Through this study, the aim is to gauge the power of the LR test in identifying significant deviations from the null hypothesis. Notably, the null hypothesis above lies at the boundary of the parameter space for ϕ, making the usual LR test statistic, denoted as $\mathrm{LR} = 2(\ell_1^{ML} - \ell_0^{ML})$, where ℓ_i^{ML} represents the log-likelihood function evaluated at the ML estimator under hypothesis H_i, $i = 0, 1$, unable to follow the standard chi-squared distribution with 1 degree of freedom ($\chi^2_{(1)}$). Instead, in this scenario, the asymptotic distribution is given by $(1/2) + (1/2)\chi^2_{(1)}$, as demonstrated by Stram and Lee (1994).

The simulation scenario used consists of the following configuration: In the data generation we consider the case when the alternative hypothesis is true, that is, the

competitive causes comes from the PBS mixture model. The time-to-event values were sampled from the Weibull distribution with fixed parameters $\alpha = -4$ and $\nu = 1.8$. We generated a sample for the number of competing causes using the PBS mixture model, with fixed regression coefficients in $\beta_0 = 1.9$, $\beta_1 = -1.5$, $\beta_2 = -0.2$, and to analyse those significant deviation of the null hypothesis mentioned later, varying the ϕ parameter at $\phi \in \{0.6,\ 1,\ 10,\ 25,\ 100\}$. The sample sizes were defined in $n \in \{200;\ 400;\ 600;\ 1,000\}$. The significance levels were set at $\xi \in \{0.01,\ 0.05,\ 0.1\}$. The percentage of rejection of the null hypothesis for the fixed cases are displayed in Table 3.3.

Table 3.3: Power (%) of LR Test for different values of ϕ and sample sizes.

ϕ	Significance level (%) and Sample size (n)											
	1%				5%				10%			
	200	400	600	1000	200	400	600	1000	200	400	600	1000
0.6	0.253	0.438	0.641	0.857	0.508	0.706	0.838	0.963	0.631	0.826	0.898	0.986
1	0.209	0.364	0.495	0.763	0.437	0.631	0.778	0.915	0.616	0.741	0.872	0.960
10	0.040	0.053	0.084	0.129	0.177	0.190	0.225	0.297	0.285	0.271	0.345	0.464
25	0.026	0.028	0.033	0.048	0.117	0.104	0.119	0.127	0.198	0.178	0.208	0.254
100	0.020	0.017	0.011	0.012	0.078	0.059	0.061	0.066	0.127	0.124	0.133	0.139

From Table 3.3, we can deduce that the power of the LR test rises proportionally with larger sample sizes, as anticipated. For small values of ϕ, the study shows that the test is more powerful in identifying the PBS mixture model as the best fit, indicating its effectiveness in detecting significant deviations from the null hypothesis. It is also possible to observe that as the value of ϕ increases, the percentage of rejection of the LR test decreases, indicating that for large values of ϕ, the Poisson model is the most adequate to adjust data with this feature, as expected.

All the scenarios chosen for this simulations study have a considered computational effort to compute these rejection percentages. The higher the values of ϕ parameters and the sample sizes considered in this study, the greater the computational effort employed in obtaining the calculated percentages.

3.5 Application with breast cancer data

In this section, we present a real data problem related to melanoma cancer in the state of São Paulo, Brazil. Female breast cancer is the most incidence in the world, with 2.3 million new cases (11.7%), followed by lung cancer with 2.2 million cases (11.4%).

Additionally, colon and rectum cancer accounted for 1.9 million cases (10.0%), prostate cancer for 1.4 million cases (7.3%), and non-melanoma skin cancer for 1.2 million cases (6.2%) (INCA, 2022). Like in the rest of the world, cancer plays a significant role in public health in Brazil, contributing to a substantial number of deaths and placing pressure on the public healthcare system's costs. Among the various prevalent types of cancer in the country, breast cancer stands out as a prominent contributor to this increasing mortality trend. Globally, it is estimated that 70% of breast cancer deaths occur in women from low- and middle-income countries (Goss et al., 2013). According to data from the National Cancer Institute, in the State of São Paulo, Brazil, the estimated crude and adjusted incidence rates per 100,000 inhabitants, as well as the number of new cancer cases for the year 2023, were 97.72 and 58.90 cases, respectively.

Swaminathan et al. (2023) conducted a comprehensive assessment of treatments aimed at improving survival rates for breast cancer. Therapeutic decisions are typically based on the recognition of the unique characteristics of tumours. In cases of non-metastatic breast cancer, the standard approach involves surgical excision and the removal of axillary lymph nodes, followed by postoperative radiotherapy as a local therapy. In essence, the primary procedures for treating non-metastatic breast cancer include removing the tumour and regional lymph nodes from the breast and preventing metastatic recurrence. Surgery and radiotherapy are commonly employed as regional approaches for early-stage tumours. Chemotherapy is considered a gold-standard treatment strategy, utilizing combinations of cytotoxic drugs to either destroy or reduce the growth of breast cancer cells. The selection or combination of these medical approaches depends on the overall condition of the patients.

As cited in the Introduction section, in this chapter, we utilize a dataset from the Oncology Foundation of São Paulo (FOSP), São Paulo, Brazil. The dataset comprises observations from a retrospective survey involving patients diagnosed with breast cancer in the State of São Paulo, Brazil, during the years 2009 to 2016, with follow-up conducted until 2021. The event of interest was defined as death due to breast cancer, and the time-to-event was calculated as the period between the date of diagnosis and the date of death attributed to cancer. Patients who did not experience cancer-related mortality during the follow-up period were considered as right-censored observations. The

dataset contains a total of $59,300$ patients and the explanatory variables considered in our analysis are as follows: X_1: Clinical cancer stage (Stage I: $n = 14,988\ (25.3\%)$; Stage II: $n = 21,987\ (37.1\%)$; Stage III: $n = 16,094\ (27.1\%)$; and Stage IV, $n = 6,231\ (10.5\%)$), patient's treatment, being X_2: Surgery (Yes: $n = 45,719\ (77.1\%)$), X_3: Radiotherapy (Yes: $n = 26,479\ (44.7\%)$), X_4: Chemotherapy (Yes: $n = 39,747\ (67\%)$), and X_4: Age at diagnosis in years (mean$\pm$ standard deviation, 56.3$\pm$13.62). The maximum observed follow-up time was 13.85 years. The median and mean follow-up times were approximately 5.27 and 5.35 years, respectively. The percentage of censored observations was 77.67%.

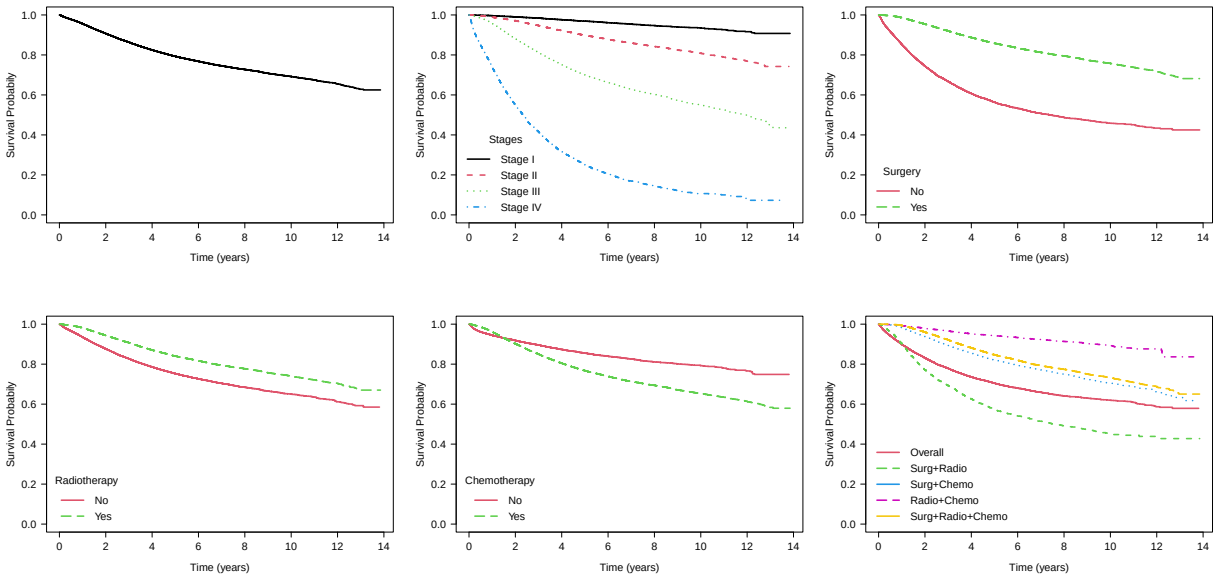

Figure 3.1: Estimated SF obtained from the Kaplan-Meier (KM) estimator for overall patients diagnosed with breast cancer, by clinical stage, surgery, radiotherapy, chemotherapy and combinations of treatments.

Figure 3.1 shows the estimated KM curves associated with the clinical stage and type of treatment. Higher survival rates were observed in early clinical stages (I and II), while poorer prognoses were noted in advanced clinical Stage IV. Notably, higher survival rates were observed in patients who underwent surgery, received radiotherapy, and did not receive chemotherapy. For patients undergoing chemotherapy, better survival rates were observed within the first 2 years, with improved long-term survival for those who did not receive chemotherapy. This result is expected, as a significant percentage of

patients at clinical stages III and IV received chemotherapy (80.8%), which typically presents a more challenging prognosis.

A hypothesis test proposed by Maller and Zhou (1992) available in the package `npcure` López et al. (2022) of R Core Team (2024) was carried out with the aim of verifying whether there are presence of "immune" individuals in the study, this estimated value is used to test whether the study has enough follow-up time. Based on the results for this application, the test has provided evidences through p-value < 0.0001, that there is presence of immune individuals and that the follow-up time is sufficient at 5% significance level. The same characteristic for long-term survivals in the KM-curves in data for patients with breast cancer were discussed in Rodrigues et al. (2016), Makdissi et al. (2019) and S. Pal (2021).

In this section, we fitted the proposed model as well as various cure rate models from the existing literature to the real breast cancer data presented in the previous section. The event of interest was defined as death due to breast cancer. Our objective was to assess the effect of variables such as age at diagnosis, clinical stage, surgery, radiotherapy, and chemotherapy on survival rates.

We obtained ML estimates by employing the EM algorithm, as detailed in Section 3.3.1. The EM algorithm has been implemented in the `R` language (R Core Team, 2024) and is available to the community upon request. Furthermore, we computed the ML estimates for the parameters of the compared models, including NB, POI and BER (standard mixture) using the `EM.PScr` function included in the `PScr` (Gallardo and Azimi, 2023) package of `R` software.

It is important to emphasize that, in our proposal, the time-to-event distribution chosen for individuals at risk is the Weibull distribution, following parameterizations as detailed in Table 2 of Gallardo et al. (2017) for both $S(t, \boldsymbol{\eta})$ and $f(t, \boldsymbol{\eta})$, survival and density functions, which are defined by $\boldsymbol{\eta} = (\alpha, \nu)$ parameters. These values differ from the typical Weibull distribution included in the `PScr` Package used to compute EM estimates for the models under comparison, which are parameterized by $\boldsymbol{\eta} = (\nu_{(PScr)}, \sigma_{(PScr)})$, where $\nu_{(PScr)}$ and $\sigma_{(PScr)}$ are the conventional shape and scale parameters, respectively.

We conducted a comparative analysis of the proposed PBS mixture model against the NB, POI and BER (standard mixture) models to evaluate their fitting performances

Table 3.4: AIC, BIC and BF values obtained by fitting the PBS mixture, NB, POI and BER (standard mixture) models to the breast cancer dataset.

Models	AIC	BIC	Estimated Log-likelihood	BF
PBS mixture	92,666.60	92,765.53	−46,322.32	−
NB	92,705.90	92,804.77	−46,341.94	39.24
POI	93,144.70	93,234.56	−46,562.33	469.02
BER (Std. Mixture)	95,413.70	95,503.65	−47,696.87	2738.1

on the dataset, considering that the time-to-event comes from the Weibull distribution.

In Table 3.4, values for Akaike information criterion (AIC), as introduced by Akaike (1973), Bayesian information criterion (BIC), proposed by Schwarz (1978), and Bayes factor (BF) are provided. We use the BF to evaluate the magnitude of the difference between two BIC values; see Kass and Raftery (1995). We compute the AIC and BIC in all models but the BF is obtained for the comparison between the PBS versus NB, PBS versus POI and PBS versus BER. Decision about the best fit is made according to the interpretation of the BF presented in Table 6 of Leiva et al. (2015b). Table 3.4 indicates that the PBS mixture model provides the best overall fit in terms of AIC, BIC and BF.

The ML estimates of the model parameters, accompanied by their corresponding standard errors and p-values for each model, can be found in Table 3.5. Given that both AIC and BIC criteria have indicated our proposal as the most suitable among the four fitted models, the interpretation of the results will be based on the estimated parameters of this specific model.

Based on the results provided in Table 3.5, all covariates included in the analysis are statistically significant associated (at the 5% significance level) with the time-to-event, except for the chemotherapy, which can be considered significant at 8%. Positive estimated regression coefficients were obtained for the clinical stage and age at diagnosis, indicating that higher clinical stages and older age at diagnosis are associated with worse survival rates. Conversely, negative estimated values were obtained for the surgery and radiotherapy indicating that patients who undergo surgery and radiotherapy have better survival rates compared to those who did not receive surgery and radiotherapy. Examining the fitted values for chemotherapy treatment, notable variations in the estimated coefficient values emerge, particularly in the model fitted through the Bernoulli

Table 3.5: ML estimates, standard error (SE) and respective p-value obtained by fitting of cure rate models for PBS mixture, NB, POI and BER (standard mixture) applied to breast cancer.

Parameter	PBS		NB		POI		BER	
	ML	SE	ML	SE	ML	SE	ML	SE
$\beta_{0:\text{Intercept}}$	-1.673	0.092	-1.477	0.154	-1.905	0.081	-2.649	0.092
p-value	< 0.001		< 0.001		< 0.001		< 0.001	
$\beta_{1:\text{Stage II}}$	1.219	0.050	1.194	0.048	1.119	0.046	1.169	0.050
p-value	< 0.001		< 0.001		< 0.001		< 0.001	
$\beta_{2:\text{Stage III}}$	2.533	0.055	2.4216	0.051	2.219	0.045	2.562	0.051
p-value	< 0.001		< 0.001		< 0.001		< 0.001	
$\beta_{3:\text{Stage IV}}$	4.086	0.075	3.986	0.064	3.361	0.046	7.953	0.830
p-value	< 0.001		< 0.001		< 0.001		< 0.001	
$\beta_{4:\text{ Surgery}}$	-0.755	0.030	-0.748	0.028	-0.561	0.020	-0.732	0.038
p-value	< 0.001		< 0.001		< 0.001		< 0.001	
$\beta_{5:\text{ Radiotherapy}}$	-0.327	0.024	-0.326	0.023	-0.252	0.018	-0.319	0.030
p-value	< 0.001		< 0.001		< 0.001		< 0.001	
$\beta_{6:\text{Chemotherapy}}$	0.052	0.030	0.026	0.030	0.091	0.023	0.357	0.039
p-value	0.080		0.459		< 0.001		< 0.001	
$\beta_{7:\text{ Age}}$	0.008	0.001	0.008	0.001	0.006	0.001	0.009	0.001
p-value	< 0.001		< 0.001		< 0.001		< 0.001	
α	-3.526	0.070	–	–	–	–	–	–
ν and $\nu_{(PScr)}$	1.396	0.019	1.331	0.016	1.182	0.012	1.167	0.009
$\sigma_{(PScr)}$	–	–	17.233	1.497	12.499	0.795	5.006	0.057
ϕ	1.358	0.141	–	–	–	–	–	–
q	–	–	1.120	0.080	–	–	–	–

(standard mixture). Despite its relevance to the estimated model, this discrepancy may arise from the observations associated with this variable. Furthermore, it is possible to note that the adjustment for the Bernoulli model presents a very unsatisfactory result for the estimated value of the log-likelihood presenting in Table 3.4, which is much lower than the other models compared. This fact may be impacting the estimated values for the parameters, making them to differ from the other settings.

All of the findings in this study are consistent with observations made in routine clinical practice. Clinical stage and age at diagnosis have previously been reported as prognostic factors, indicating that younger patients in early clinical stages who undergo surgery and radiotherapy tend to have a better prognosis (Makdissi et al., 2019).

The estimated long-term survivors for Equation (3.5) considering patients with fixed ages at diagnosis of 20, 56 (the average age of patients), and 70 years, undergoing

Table 3.6: ML estimates of cure rate and 95% Confidence Interval (IC) obtained by Delta Method for PBS mixture cure rate model applied to breast cancer dataset through Stage of disease and treatments.

Age: 20 years old

Stages of disease		Stage I	Stage II	Stage III	Stage IV
No treatment	Estimate	0.826	0.593	0.278	0.046
	IC 95%	(0.802; 0.849)	(0.555; 0.631)	(0.243; 0.314)	(0.036; 0.055)
Surgery	Estimate	0.908	0.751	0.457	0.129
	IC 95%	(0.896; 0.921)	(0.724; 0.778)	(0.418; 0.496)	(0.107; 0.152)
Radiotherapy	Estimate	0.867	0.666	0.353	0.075
	IC 95%	(0.848; 0.886)	(0.632 0.701)	(0.314; 0.392)	(0.061; 0.089)
Chemotherapy	Estimate	0.818	0.581	0.267	0.042
	IC 95%	(0.795; 0.842)	(0.546; 0.615)	(0.236; 0.298)	(0.034; 0.050)
Surgery and radiotherapy	Estimate	0.932	0.806	0.537	0.183
	IC 95%	(0.922; 0.941)	(0.784; 0.828)	(0.499; 0.575)	(0.155; 0.212)
Surgery and chemotherapy	Estimate	0.904	0.742	0.444	0.122
	IC 95%	(0.892; 0.917)	(0.717; 0.766)	(0.410; 0.479)	(0.103; 0.141)
Radiotherapy and chemotherapy	Estimate	0.861	0.655	0.341	0.070
	IC 95%	(0.842; 0.880)	(0.623; 0.687)	(0.306; 0.375)	(0.057; 0.082)
Surgery, radiotherapy and chemotherapy	Estimate	0.929	0.798	0.525	0.174
	IC 95%	(0.919; 0.938)	(0.778; 0.817)	(0.492; 0.558)	(0.149; 0.199)

Age: 56 years old

Stages of disease		Stage I	Stage II	Stage III	Stage IV
No treatment	Estimate	0.784	0.528	0.221	0.028
	IC 95%	(0.759; 0.810)	(0.487; 0.568)	(0.192; 0.250)	(0.023; 0.034)
Surgery	Estimate	0.884	0.699	0.390	0.093
	IC 95%	(0.870; 0.898)	(0.672; 0.726)	(0.355; 0.425)	(0.077; 0.109)
Radiotherapy	Estimate	0.833	0.605	0.290	0.050
	IC 95%	(0.813; 0.854)	(0.572; 0.639)	(0.257; 0.323)	(0.040; 0.059)
Chemotherapy	Estimate	0.776	0.515	0.211	0.026
	IC 95%	(0.749; 0.802)	(0.481; 0.549)	(0.185; 0.237)	(0.021; 0.031)
Surgery and radiotherapy	Estimate	0.913	0.761	0.471	0.138
	IC 95%	(0.902; 0.924	(0.738; 0.783)	(0.436; 0.505)	(0.116; 0.159)
Surgery and chemotherapy	Estimate	0.878	0.688	0.378	0.087
	IC 95%	(0.864 0.893	(0.663; 0.713)	(0.346 0.409)	(0.073; 0.101)
Radiotherapy and chemotherapy	Estimate	0.826	0.593	0.279	0.046
	IC 95%	(0.805; 0.847	(0.561; 0.625)	(0.249; 0.309)	(0.037; 0.054)
Surgery, radiotherapy and chemotherapy	Estimate	0.909	0.752	0.458	0.130
	IC 95%	(0.898; 0.920)	(0.730; 0.773)	(0.426; 0.489)	(0.111; 0.149)

Age: 70 years old

Stages of disease		Stage I	Stage II	Stage III	Stage IV
No treatment	Estimate	0.766	0.502	0.201	0.023
	IC 95%	(0.739; 0.794)	(0.459; 0.545)	(0.174; 0.228)	(0.018; 0.028)
Surgery	Estimate	0.873	0.677	0.365	0.080
	IC 95%	(0.857; 0.888)	(0.649; 0.705)	(0.330; 0.399)	(0.066; 0.095)
Radiotherapy	Estimate	0.818	0.581	0.267	0.042
	IC 95%	(0.796; 0.841)	(0.546; 0.615)	(0.235; 0.299)	(0.034; 0.050)
Chemotherapy	Estimate	0.757	0.489	0.191	0.021
	IC 95%	(0.729; 0.786)	(0.454; 0.524)	(0.166; 0.216)	(0.017; 0.025)
Surgery and Radiotherapy	Estimate	0.904	0.742	0.444	0.122
	IC 95%	(0.892; 0.916)	(0.718; 0.766)	(0.409; 0.479)	(0.102; 0.142)
Surgery and Chemotherapy	Estimate	0.867	0.666	0.352	0.075
	IC 95%	(0.851; 0.883)	(0.639; 0.693)	(0.321; 0.384)	(0.062; 0.087)
Radiotherapy and Chemotherapy	Estimate	0.811	0.568	0.256	0.038
	IC 95%	(0.787; 0.834)	(0.535; 0.602)	(0.227; 0.285)	(0.031; 0.046)
Surgery, radiotherapy and chemotherapy	Estimate	0.900	0.732	0.432	0.115
	IC 95%	(0.887; 0.912)	(0.709; 0.755)	(0.399; 0.464)	(0.097; 0.132)

various types of treatments across the four clinical stages are shown in Table 3.6. The study revels that estimated long-term survivors decrease as age increases, indicating that younger patients have better survival rates when diagnosed early. As expected, patients in clinical stage IV exhibited a poorer prognosis, regardless of their age at diagnosis and the type of treatment received. In some cases, physicians have opted for submitting the patients to more than one treatment, providing in some cases higher probability of cure than a specific treatment isolated.

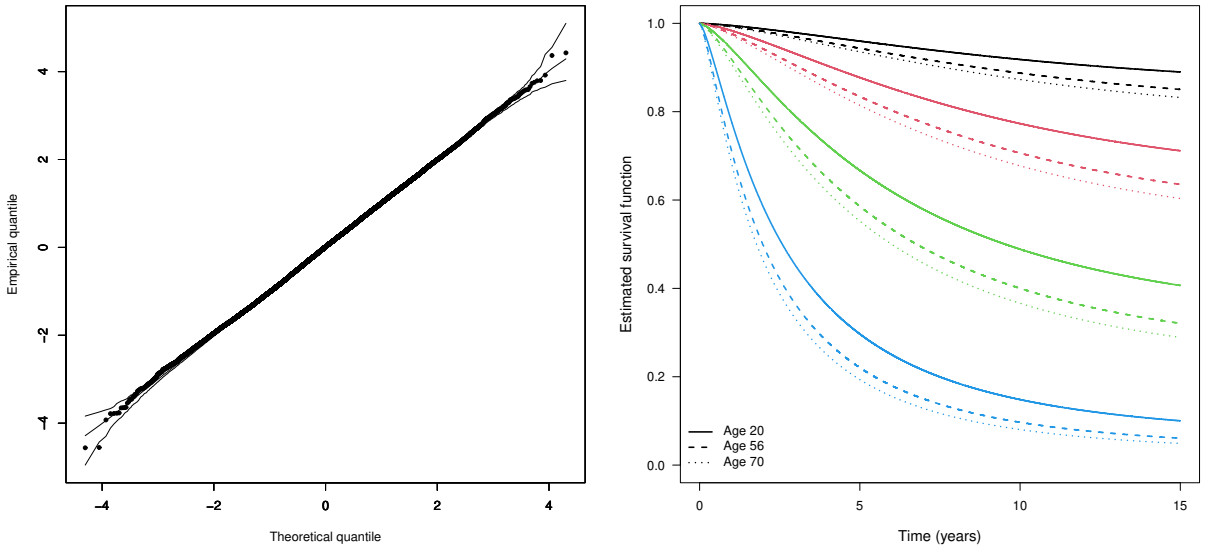

Figure 3.2: Normalized randomized quantile residuals for PBS mixture applied to breast cancer dataset. Estimated SF for PBS Mixture, for patients with 20, 56 and 70 years old, who underwent radiotherapy and chemotherapy through stages of disease: Stage I (black), Stage II (red), Stage III (green) and Stage IV (blue).

Figure 3.2 shows the quantile versus quantile plot of the normalized randomized quantile residuals (Dunn and Smyth, 1996) for the PBS mixture model, which suggests that the proposed model shows a good agreement with the expected standard normal distribution. Figure 3.2 also illustrated the estimated survival curve associated to patients who underwent radiotherapy and chemotherapy for three different ages through the stage of disease. These estimates also indicate that younger patients in all clinical stages who undergo radiotherapy and chemotherapy tend to have a better prognosis. Furthermore, the estimated survival decays for all ages studied as the clinical stage

grows up. in all scenarios, older patients had worse cure rates.

3.6 Concluding Remarks

In this chapter, we propose a new model for long-term survival data, assuming that the number of concurrent causes for event of interest is a mixture of the Poisson and BS distributions. This approach represents a significant innovation, as the BS model does not belong to the exponential family, presenting several interesting properties and applications in medical and biological research. A distinguishing feature of the proposed model is the existence of closed-form equations for all conditional expectations, allowing an efficient estimation via ML. In addition, the developed estimation algorithm is remarkably simple to implement, as all the steps are completely defined. This model emerges as a competitive alternative to the NB model, which is widely recognized in the literature. Both models share the characteristic of overdispersion of concurrent causes relative to the mean. However, our model utilizes the BS as an innovative and popular alternative in recent literature, adding versatility to the available modeling options.

The simulation study suggests that the ML estimators have good performance in terms of bias, RMSE, and CP, despite the heightened complexities inherent in modeling with small samples, particularly concerning the estimation of the ϕ parameter and in terms of overall fitting. Through the power study for the likelihood test, the authors have estimated the statistical test through samples generated from the model under investigation contrasting these results with the Poisson model. The study has shown great percentages of rejection of the Poisson model for minor values of the dispersion parameter ϕ, in addition to a decrease in this percentage according to the higher values set for this parameter. This result is expected as the authors have argued in Section 3.2 for the development of the proposed model.

The proposed methodology has fitted well the dataset provided by FOSP with a retrospective survey with 59,300 patients diagnosed with breast cancer in the State of São Paulo, Brazil. Criterion's, AIC, BIC and BF have shown that the PBS mixture cure rate model had a better fit as compared to the POI, NB and BER models. Furthermore, had well-fitting through the normalized randomized quantiles residuals.

In short, our methodology was able to yield more precise inferences regarding the impact of disease stages, different types of applied treatments, and patient ages than the commonly used promotion time model in survival data analysis with cure fraction, in addition to the NB and the standard mixture models. Furthermore, This study underscores the significance of early disease detection in achieving treatment success, emphasizing the importance of both breast self-examination and regular screening examinations in enhancing treatment efficacy and attaining higher rates of recovery through therapeutic interventions.

Appendix A

In this Section, we provide details for Proposition 3.1 and Corollary 3.1.

A.1. Proof of Proposition 3.3.1

Using the results in Gallardo et al. (2017), we obtain For $m_i = 0, 1, \ldots, z_i > 0, y_i = 0, 1$.

$$
\begin{aligned}
f(t_i, \delta_i, m_i, z_i, y_i) = S(t_i; \boldsymbol{\eta})^{m_i - \delta_i} [m_i f(t_i; \boldsymbol{\eta})]^{\delta_i} \frac{(\theta_i z_i)^{m_i} \mathrm{e}^{-\theta_i z_i}}{m_i!} \\
\times \frac{\exp(\phi/2)\sqrt{\phi+1}}{4\sqrt{\pi}} z_i^{-y_i + \frac{1}{2} - 1} \left(\frac{\phi}{\phi+1}\right)^{y_i} \\
\times \exp\left\{-\frac{1}{2}\left[\frac{(\phi+1)z_i}{2} + \frac{\phi^2}{2z_i(\phi+1)}\right]\right\}.
\end{aligned}
$$

$$
\begin{aligned}
f(m_i, z_i, y_i \mid t_i, \delta_i) &\propto \frac{(\theta_i z_i S(t_i; \boldsymbol{\eta}))^{m_i - \delta_i})}{(m_i - \delta_i)!} z_i^{p_i(y_i) - 1} \exp\left\{-\frac{1}{2}\left[a_i z_i + \frac{b_i}{z_i}\right]\right\} \left(\frac{\phi}{\phi+1}\right)^{y_i}, \\
&\propto \underbrace{\frac{(\theta_i z_i S(t_i; \boldsymbol{\eta}))^{m_i - \delta_i} \mathrm{e}^{-\theta_i z_i S(t_i; \boldsymbol{\eta})}}{(m_i - \delta_i)!}}_{f(m_i - \delta_i | z_i, t_i, \delta_i)} \\
&\quad \times \underbrace{\frac{(a/b)^{p_i(y_i)/2}}{2\mathcal{K}_{p_i(y_i)}(\sqrt{a_i b_i})} z_i^{p_i(y_i) - 1} \exp\left\{-\frac{1}{2}\left[a_i z_i + \frac{b_i}{z_i}\right]\right\}}_{f(z_i | y_i, t_i, \delta_i)} \\
&\quad \times \underbrace{\mathcal{K}_{p_i(y_i)}(\sqrt{a_i b_i}) \left(\frac{\phi}{\phi+1}\right)^{y_i} \times \left(\frac{b_i}{a_i}\right)^{p(y_i)/2}}_{f(y_i | t_i, \delta_i)},
\end{aligned}
$$

with $m_i = \delta_i, \delta_i + 1, \ldots, z_i > 0, y_i = 0, 1,$, furthermore, $p_i(y_i) = \delta_i - y_i + 1/2$, $a_i = a_i(\phi) = 2\theta_i F(t_i; \eta) + (\phi+1)/2$, $b_i = b_i(\phi) = \phi^2/[2(\phi+1)]$. The result is obtained recognizing the distributions in each case.

A.2. Proof of Proposition 3.3.1

As described later, if the conditional distribution of $Z_i \mid y_i, t_i, \delta_i \sim$ GIG$(a_i, b_i, p_i(y_i))$, we have the following results

$$\mathbb{E}[Z_i \mid y_i, t_i, \delta_i] = \frac{\sqrt{b_i}\mathcal{K}_{p_i(y_i)+1}(\sqrt{a_i b_i})}{\sqrt{a_i}\mathcal{K}_{p_i(y_i)}(\sqrt{a_i b_i})} = g_1(y_i),$$

$$\mathbb{E}[Z_i^{-1} \mid y_i, t_i, \delta_i] = \frac{\sqrt{a_i}\mathcal{K}_{p_i(y_i)+1}(\sqrt{a_i b_i})}{\sqrt{b_i}\mathcal{K}_{p_i(y_i)}(\sqrt{a_i b_i})} - \frac{2p_i(y_i)}{b_i} = g_3(y_i).$$

Using the properties of conditional expectation, it is easy to see that

$$\mathbb{E}[Z_i \mid t_i, \delta_i] = \mathbb{E}[\mathbb{E}(Z_i \mid y_i, t_i, \delta_i) \mid t_i, \delta_i].$$

Since the distribution of $Y_i \mid t_i, \delta_i \sim$ BER(ω_i) with ω_i defined in Equation (3.7) and using the expressions calculated later it is straightforward that

$$\begin{aligned}\mathbb{E}[Z_i \mid t_i, \delta_i] &= \mathbb{E}[g_1(Y_i)|t_i, \delta_i] = g_1(0)\mathbb{P}[Y_i = 0|t_i, \delta_i] + g_1(1)\mathbb{P}[Y_i = 1|t_i, \delta_i] \\ &= \frac{\sqrt{b_i}\mathcal{K}_{p_i(0)+1}(\sqrt{a_i b_i})}{\sqrt{a_i}\mathcal{K}_{p_i(0)}(\sqrt{a_i b_i})}(1-\omega_i) + \frac{\sqrt{b_i}\mathcal{K}_{p_i(1)+1}(\sqrt{a_i b_i})}{\sqrt{a_i}\mathcal{K}_{p_i(1)}(\sqrt{a_i b_i})}\omega_i;\end{aligned}$$

$$\begin{aligned}\mathbb{E}[Z_i^{-1} \mid t_i, \delta_i] &= \mathbb{E}[g_3(Y_i)|t_i, \delta_i] = g_3(0)\mathbb{P}[Y_i = 0|t_i, \delta_i] + g_3(1)\mathbb{P}[Y_i = 1|t_i, \delta_i] \\ &= \left[\frac{\sqrt{a_i}\mathcal{K}_{p_i(0)+1}(\sqrt{a_i b_i})}{\sqrt{b_i}\mathcal{K}_{p_i(0)}(\sqrt{a_i b_i})} - \frac{2p_i(0)}{b_i}\right](1-\omega_i)\left[\frac{\sqrt{a_i}\mathcal{K}_{p_i(1)+1}(\sqrt{a_i b_i})}{\sqrt{b_i}\mathcal{K}_{p_i(1)}(\sqrt{a_i b_i})} - \frac{2p_i(1)}{b_i}\right]\omega_i.\end{aligned}$$

From the Proposition 3.3.1 conditional distribution for $M_i - \delta_i \mid z_i, t_i, \delta_i \sim$ POI$(\theta_i z_i S(t_i; \boldsymbol{\eta}))$ so that its expected value is $\mathbb{E}[M_i - \delta_i \mid z_i, t_i, \delta_i] = \theta_i z_i S(t_i; \boldsymbol{\eta})$. Similary, using the properties of conditional expectation, we have that

$$\begin{aligned}\mathbb{E}[M_i \mid z_i, t_i] &= \mathbb{E}[\mathbb{E}(M_i \mid z_i, t_i, \delta_i) \mid z_i, t_i] = \mathbb{E}[\delta_i + \theta_i Z_i S(t_i; \boldsymbol{\eta}) \mid z_i, t_i] \\ &= \delta_i + \theta_i S(t_i; \boldsymbol{\eta})\mathbb{E}[Z_i \mid t_i, \delta_i].\end{aligned}$$

Appendix B: Simulation Study

Table 3.7: Empirical, Bias, Root of MSE, standard error (SE) and CP of the ML estimators for PBS mixture model using the Weibull distribution to time-to-event in the concurrent causes regression.

Sample size	**Parameter**	β_0	β_1	β_2	α	ν	ϕ	Var(Z)
	Real Value	-0.81	1.26	2.64	-3.93	1.43	0.94	1,8
400	Bias	-0.036	0.024	0.005	0.004	0.004	2.153	-0.120
	RMSE	0.399	0.422	0.485	0.512	0.145	12.717	0.737
	SE	0.462	0.442	0.579	0.685	0.174	6.036	0.985
	CP	0.968	0.960	0.958	0.949	0.955	0.980	0.788
600	Bias	-0.017	-0.001	-0.015	0.012	0.001	0.662	-0.091
	RMSE	0.316	0.332	0.379	0.389	0.116	4.046	0.580
	SE	0.368	0.357	0.471	0.557	0.143	3.029	0.761
	CP	0.971	0.961	0.962	0.963	0.968	0.977	0.836
800	Bias	-0.019	-0.001	-0.020	0.021	-0.004	0.366	-0.090
	RMSE	0.283	0.293	0.341	0.358	0.110	2.147	0.536
	SE	0.318	0.307	0.406	0.479	0.123	1.879	0.695
	CP	0.971	0.962	0.961	0.964	0.955	0.979	0.841
1000	Bias	-0.035	0.010	-0.011	0.035	-0.006	0.376	-0.084
	RMSE	0.253	0.263	0.299	0.328	0.095	3.278	0.492
	SE	0.282	0.276	0.364	0.426	0.110	1.038	0.636
	CP	0.968	0.958	0.969	0.958	0.950	0.982	0.858
1200	Bias	-0.002	-0.008	-0.013	0.009	0.000	0.180	-0.048
	RMSE	0.237	0.243	0.280	0.299	0.090	0.858	0.448
	SE	0.259	0.251	0.333	0.392	0.101	0.639	0.600
	CP	0.967	0.949	0.962	0.964	0.957	0.972	0.885
1400	Bias	-0.012	0.000	-0.006	0.007	0.002	0.121	-0.047
	RMSE	0.217	0.226	0.252	0.251	0.077	0.495	0.381
	SE	0.238	0.233	0.308	0.361	0.093	0.494	0.504
	CP	0.965	0.955	0.979	0.969	0.971	0.987	0.893
1600	Bias	-0.013	0.001	-0.012	0.017	-0.004	0.101	-0.046
	RMSE	0.190	0.203	0.235	0.228	0.069	0.421	0.345
	SE	0.223	0.218	0.288	0.338	0.087	0.446	0.454
	CP	0.977	0.963	0.972	0.979	0.975	0.991	0.899
5000	Bias	0.0003	-0.006	-0.008	0.006	-0.002	0.024	-0.015
	RMSE	0.101	0.115	0.123	0.109	0.038	0.160	0.168
	SE	0.125	0.122	0.163	0.190	0.049	0.216	0.224
	CP	0.982	0.958	0.977	0.980	0.980	0.997	0.950

Table 3.8: Empirical, Bias, Root of MSE, standard error (SE) and CP of the ML estimators for PBS mixture model using the Weibull distribution to time-to-event in the concurrent causes regression.

Sample size (n)	**Parameter**	β_0	β_1	β_2	α	ν	ϕ	Var(Z)
	Real Value	1.9	-1.5	-0.2	-4	1.8	0.6	2.42
400	Bias	0.017	-0.015	-0.002	-0.043	0.019	1.554	-0.105
	RMSE	0.672	0.346	0.052	0.535	0.181	10.714	0.971
	SE	0.760	0.366	0.055	0.602	0.202	5.854	2.244
	CP	0.942	0.933	0.950	0.947	0.952	0.896	0.843
600	Bias	0.022	-0.010	-0.002	-0.032	0.016	0.462	-0.064
	RMSE	0.539	0.280	0.042	0.439	0.152	3.385	0.785
	SE	0.631	0.302	0.045	0.497	0.166	1.723	1.801
	CP	0.957	0.945	0.948	0.954	0.953	0.930	0.885
800	Bias	0.013	-0.011	-0.002	-0.022	0.009	0.263	-0.064
	RMSE	0.451	0.231	0.036	0.360	0.123	1.699	0.665
	SE	0.551	0.262	0.039	0.432	0.145	0.952	1.494
	CP	0.964	0.962	0.967	0.965	0.966	0.937	0.915
1000	Bias	0.023	-0.018	-0.001	-0.024	0.008	0.143	-0.031
	RMSE	0.444	0.221	0.034	0.348	0.118	0.685	0.642
	SE	0.493	0.235	0.035	0.386	0.129	0.522	1.475
	CP	0.953	0.951	0.952	0.957	0.957	0.919	0.924
1200	Bias	0.019	-0.013	-0.002	-0.019	0.008	0.118	-0.030
	RMSE	0.382	0.193	0.030	0.294	0.102	0.798	0.548
	SE	0.449	0.213	0.032	0.352	0.118	0.473	1.243
	CP	0.964	0.967	0.953	0.969	0.965	0.947	0.941

References

Abramowitz, M. and Stegun, I. (1972). Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, *Dover Publications, NY.*

Akaike, H. (1973). nformation theory and an extension of the maximum likelihood principle. *In: Petrov, B.N. and Csaki, F., Eds., International Symposium on Information Theory*, pages 267–281.

Balakrishnan, N. and Kundu, D. (2019). Birnbaum-saunders distribution: A review of models, analysis, and applications. *Applied Stochastic Models in Business and Industry*, 35(1):4–49.

Barreto-Souza, W. (2015). Long-term survival models with overdispersed number of competing causes. *Computational Statistics and Data Analysis*, 91(1):51–63.

Berkson, J. and Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47(259):501–515.

Birnbaum, Z. W. and Saunders, S. C. (1969). A new family of life distributions. *Journal of Applied Probability*, 6:319–327.

Borchers, H. W. (2022). *pracma: Practical Numerical Math Functions*. R package version 2.3.8.

Brandão, M., Leão, J., Gallardo, D., and Bourguignon, M. (2023). Cure rate models for heterogeneous competing causes. *Statistical Methods in Medical Research*.

Cancho, V. G., Louzada, F., and Ortega, E. M. (2013). The power series cure rate model: an application to a cutaneous melanoma data. *Communications in Statistics-Simulation and Computation*, 42(3):586–602.

Chen, M.-H., Ibrahim, J., and Sinha, D. (1999). A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, 94(447):909–919.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Desmond, A. (1985). Stochastic models of failure in random environments. *Canadian Journal of Statistics*, 13(13):171–183.

Dunn, P. K., & Smyth, G. K. (1996). Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, 5(3), 236–244.

Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., and Bray, F. (2021). Cancer statistics for the year 2020: An overview. *International Journal of Cancer*, 149(4):778–789.

Gallardo, D. I. and Azimi, R. (2023). *PScr:Estimation for the Power Series Cure Rate Model*. R package version 1.1.

Gallardo, D. I., Gómez, H. W., and Bolfarine, H. (2017a). A new cure rate model based on the Yule–Simon distribution with application to a melanoma data set. *Journal of Applied Statistics*, 44(7):1153–1164.

Gallardo, D. I., Gómez, Y. M., and de Castro, M. (2018). A flexible cure rate model based on the polylogarithm distribution. *Journal of Statistical Computation and Simulation*, 88(11):2137–2149.

Gallardo, D. I., Romeo, J. S., and Meyer, R. (2017b). A simplified estimation procedure based on the em algorithm for the power series cure rate model. *Communications in Statistics-Simulation and Computation*, 46(8):6342–6359.

Gómez, Y., Gallardo, D., Bourguignon, M., Bertolli, E., and Calsavara, V. (2023). A general class of promotion time cure rate models with a new biological interpretation. *Lifetime Data Analysis*, 29:66–86.

Gonçalves, J., Barreto-Souza, W., and Ombao, H. (2022). Poisson-Birnbaum-Saunders regression model for clustered count data. *https://doi.org/10.48550/arXiv.2202.10162*.

Goss, P. E., Lee, B. L., and Badovinac-crnjevic, T. (2013). Planning cancer control in latin america and the caribbean. *The Lancet Oncology*, 14(5):391–436.

Hajek, J., Sidak, Z., and Sen, P. K. (1999). *Theory of Rank Tests*. Academic Press, San Diego, London.

Hashimoto, E., Ortega, E., Cordeiro, G., and Cancho, V. (2014). The poisson birnbaum-saunders model with long-term survivors. *Statistics: A Journal of Theoretical and Applied Statistics*, pages 1394–1413.

INCA. (2022). *Estimativa 2023: Incidência do Câncer no Brasil*. INCA-Instituto Nacional do Cancer, Rio de Janeiro.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Taylor & Francis*, 90: 773–795.

Kotz, S., Leiva, V., and Sanhueza, A. (2010). Two new mixture models related to the inverse gaussian distribution. *Methodology and Computing in Applied Probability*, 12:199–212.

Leão, J., Bourguignon, M., Gallardo, D. I., Rocha, R., and Tomazella, V. (2020). A new cure rate model with flexible competing causes with applications to melanoma and transplantation data. *Statistics in Medicine*, 39(24):3272–3284.

Leiva, V., Marchant, C., Ruggeri, F., and Saulo, H. (2015a). A criterion for environmental assessment using Birnbaum-Saunders attribute control charts. *Environmetrics*, 26:463–476.

Leiva, V. and Tejo, M. and Guiraud, P. and Schmachtenberg, O. and Orio, P. and Marmolejo-Ramos, F. (2015b). Modeling neural activity with cumulative damage distributions. *Biological Cybernetics*, 109:421–433.

Leiva, V., Ruggeri, F., Saulo, H., and Vivanco, J. F. (2017). A methodology based on the Birnbaum-Saunders distribution for reliability analysis applied to nano-materials. *Reliability Engineering and System Safety*, 157:192–201.

Leiva, V., Santos-Neto, M., Cysneiros, F. J. A., , and Barros, M. (2014a). Birnbaum-Saunders statistical modelling: A new approach. *Statistical Modelling*, 14:21–48.

Leiva, V., Saulo, H., Leão, J., and Marchant, C. (2014b). A family of autoregressive conditional duration models applied to financial data. *Computational Statistics and Data Analysis*, 79:175–191.

Leiva, V., Tejo, M., Guiraud, P., Schmachtenberg, O., Orio, P., and Marmolejo, F. (2015b). Modeling neural activity with cumulative damage distributions. *Biological Cybernetics*, 109:421–433.

López-de-Ullibarri, I., López, C., Jácome, M. A., (2022). *npcure: Non Parametric Estimation in Mixture Cure Models*. R package version 0.1-5.

Maller, R. A., & Zhou, S. (1992). Estimating the Proportion of Immunes in a Censored Sample. *Biometrika*, 79(4): 731–739.

Makdissi, F. B., Leite, F. P. M., Peres, S. V., Mendonça, D. R., de Oliveira, M. M., Lopez, R. V. M., Sanches, S. M., Gondim, G. R. M., Iyeyasu, H., Calsavara, V. F., et al. (2019). Breast cancer survival in a brazilian cancer center: a cohort study of 5,095 patients. *Mastology*, 29(1):37–46.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rodrigues, J., de Castro, M., Cancho, V. G., and Balakrishnan, N. (2009b). Com–poisson cure rate survival models and an application to a cutaneous melanoma data. *Journal of Statistical Planning and Inference*, 139(10):3605–3611.

Rodrigues, R., Cancho, V., De Castro, M., and Louzada-Neto, F. (2009c). On the unification of long-term survival models. *Statistics and Probability Letters*, 79(6):753–759.

Rodrigues, J., Cordeiro, Gauss. M., Cancho, V. G. and Balakrishnan, N. (2016). Relaxed Poisson cure rate models *Biometrical Journal (2016)* 58:397–4157.

S. Pal (2021). A simplified stochastic EM algorithm for cure rate model with negative binomial competing risks: An application to breast cancer data, *Statistics in Medicine (2021)*, 28: 0277-6715.

Santos-Neto, M., Cysneiros, F. J. A., Leiva, V., and Ahmed, S. (2012). On new parameterizations of the Birnbaum-Saunders distribution. *Pakistan Journal of Statistics*, 28:1–26.

Saulo, H., Leiva, V., Ziegelmann, F. A., and Marchant, C. (2013). A nonparametric method for estimating asymmetric densities based on skewed birnbaum-saunders distributions applied to environmental data. *Stochastic Environmental Research and Risk Assessment*, 27:1479–1491.

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

Stram, D. and Lee, J. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50:1171–1177.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249.

Swaminathan, H., Saravanamurali, K., and Yadav, S. A. (2023). Extensive review on breast cancer its etiology, progression, prognostic markers, and treatment. *Medical Oncology*, 40.

Wild, C. P., Weiderpass, E., and Stewart, B. (2020). *World cancer report: cancer research for cancer prevention*. International Agency for Research on Cancer, Lyon, France.