Artificial Mixture Methods for Correlated Nominal Responses and Discrete Failure Time

by

Shufang Wang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2012

Doctoral Committee:

Professor Alexander Tsodikov, Chair
Professor Hal Morgenstern
Professor Bin Nan
Associate Professor Timothy D. Johnson

To Orson

ACKNOWLEDGEMENTS

I would never have been able to finish my dissertation without the guidance of my advisor, suggestions from committee members, help from friends and support from my family.

I would like to express my deepest gratitude to my academic advisor Dr. Alexander Tsodikov for his excellent guidance, endless help, patience, caring, encouragement and providing me with an excellent atmosphere for doing research. Without him, I would never have accomplished this work.

Sincere appreciation also goes to my committee memebers, Dr. Timothy Johnson, Dr. Hal Morgenstern and Dr. Bin Nan, for being on my committee and providing valuable comments and suggestions.

I would like to extend my gratitude to Dr. John Kalbfleisch, Dr. Thomas Braun , Dr. Peter Song and all other faculty members in the Department of Biostatistics at the University of Michigan for their help and support throughout my study at the University of Michigan.

I would also like to thank my parents, two elder sisters and many friends. They were always supporting and encouraging me with their best wishes.

Finally, I want to thank my son, Orson, for always being there cheering me up and my husband, Zhao, for being supportive all the time.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

ABSTRACT

Artificial Mixture Methods for Correlated Nominal Responses and Discrete Failure Time

by

Shufang Wang

Chair: Alexander Tsodikov

Multinomial logit model with random effects is a common choice for modeling correlated nominal responses. But due to the presence of random effects and the complex form of the multinomial probabilities, the computation is often costly. We generalize the artificial mixture method for independent nominal response (*Tsodikov and Chefo* (2008)) to correlated nominal responses. Our method transforms the complex multinomial likelihood to Poisson-type likelihoods and hence allows for the estimates to be obtained iteratively solving a set of independent low-dimensional problems. The methodology is applied to real data and studied by simulations.

For discrete failure time data in large data sets, there are often many ties and a large number of distinct event time points. This poses a challenge of a high-dimensional model. We explore two ideas with the discrete proportional odds(PO) model due to its methodological and computational convenience. The log-likelihood function of discrete PO model is the difference of two convex functions, hence difference convex algorithm(DCA) carries over and brings computational efficiency. An alternative method proposed is a recursive procedure. As a result of simulation studies, these two methods work better than Quasi-Newton method in terms of both accuracy and computational time .

The results from the research on the discrete PO model motivate us to develop artificial mixture methods to discrete failure time data. We consider a general discrete transformation model and mediate the high-dimensional optimization problem by changing the model form at the “complete-data”level (conditional on artificial variables). Two complete data representations are studied: proportional hazards(PH) and PO mixture frameworks. In the PH mixture framework, we reduce the high-dimensional optimization problem to many one-dimensional problems. In the PO mixture framework, both recursive solution and DCA can be synthesized into the Expectation-Maximization(EM)-type algorithm leading to simplification in the optimization. PO mixture method is superior to the PH mixture method as a result of this study. It is applied to real data sets to fit a discrete PH and PH-PH models. Simulation study fitting discrete PH model shows that the advocated PO mixture method outperforms Quasi-Newton method in terms of both accuracy and speed.

CHAPTER I

Introduction

Nominal or polytomous response data and discrete failure time data are very common in many fields of research. In this research, we develop a few artificial mixture methods to model such response data.

1.1 Artificial Mixture Method

Artificial mixture methods are modeling techniques that simplify the likelihood function and hence the computation by introducing artificial random variables to the original model. It was first developed in *Tsodikov* (2003a), where the author introduced frailty term in the form of PH frailty model under survival data setting. It is natural to use EM-type algorithms after the introduction of artificial random variables, treating them as missing data or part of missing data. However, the “complete-data ”level likelihood sometimes does not correspond to a legitimate probabilistic model. *Tsodikov* (2003a) proposed a well-formularized and theoretically vigorous Quasi-Expectation-Maximization(QEM) algorithm which inherits all the benefits of Expectation-Maximization(EM) algorithm. The QEM algorithm justifies the artificial mixture method even when the “complete-data ”level likelihood is probabilistically illegitimate.

It was named “Fake mixture method ”in *Tsodikov and Chefo* (2008) as an approach to simplify the computation in the standard multinomial logit model. It transforms the multinomial likelihood to Poisson one without introducing nuisance parameters, which makes the method superior to many other approaches that achieved likelihood simplification at the cost of model augmentation leaving the model with a lot of extra parameters to cope with. Besides, other Poisson-based approaches only work with discrete covariates.

1.2 Correlated Nominal Responses

As an example of nominal response, types of health services utilization may include in-patient, out-patient and day clinic, as in *Kuss and McLerran* (2007) and *Wang and Tsodikov* (2010). Sometimes the multinomial outcome of interest may be constructed from two or more categorical variables as in the real data application in *Tsodikov and Chefo* (2008), where the four-category response was constructed from stage and histologic grade of the tumor of prostate cancer. Correlated observations may be present due to measures within a cluster or taken repeatedly for the same subject. Assuming independence of observations and using the standard multinomial logit model will lead to bias unless correlation between observations is modeled correctly. One popular method to model the correlation is to use multinomial logit model with random effects. However, the presence of random effects and the complex form of the multinomial likelihood often make the computation costly.

A number of approaches have been proposed in the literature to tackle the computational complexity in both the standard multinomial logit model and mixed effects multinomial logit model. They can be grouped into the following three types:

- Augmentation methods, where a large number of nuisance parameters were added.

  For example, *Baker* (1994)'s Multinomial-Poisson (MP) transformation is most important and essential in this type of approaches. Methods resembling MP transformation includes the Bayesian version of MP transformation in *Gosh et al.* (2006), a Poisson log-linear or nonlinear model as in *Chen and Kuo* (2001). *Lang* (1996) showed that the idea can be interpreted as using Lagrange multipliers to normalize multinomial probabilities. Because normalization needs to be enforced for each distinct covariate pattern, this type of transformation is restricted to discrete covariates. If the methods are used for continuous covariates, a separate parameter corresponding to the distinct covariate value is needed. Then the dimension of the parameter space is comparable to the number of observations, the maximum likelihood estimates become biased or can be inconsistent. Other augmentation methods applicable to data with continuous covariates are also available such as *Scott* (2011). However, they introduce high-dimensional augmentation adding many nuisance parameters.

- Approximate methods

  In order to estimate the parameters in a multinomial logit model with random effects, we need to integrate the “complete-data ”likelihood function with respect to the random effects. Therefore methods dealing with integration naturally apply here, such as (1) using approximate Taylor-series expansion to linearize the integrand resulting from the random effects such as *Breslow and Clayton* (1993) and (2) using quadrature methods to approximately evaluate the integrals, such as *Rabe-Hesketh et al.* (2002), *Hedeker* (2003) and *Clarkson and Zhan* (2002) etc. The third type of approximate approaches is the Monte Carlo EM (MCEM) algorithm developed recently, where random effects were treated as missing data and EM-type algorithm is used with Monte Carlo numerical methods in the E step for the evaluation of integrals(*McCulloch* (1997), *Booth and Hobert* (1999) and *Chen et al.* (2002) etc.).

- Artificial mixture method for standard multinomial logit model

  Different from the aforementioned approaches, *Tsodikov and Chefo* (2008) developed an artificial mixture method to simplify the computation in standard multinomial logit model by transforming the multinomial likelihood to a set of Poisson likelihoods through the introduction of artificial random variables.

  This approach does not add any new parameters to the model as normalizarion restrictions are enforced by averaging over artificial variables rather than by Lagrange multipliers.

  In this thesis we generalize this approach to correlated observations.

Among methods available for nominal responses with correlation in multinomial logit model with random effects, quadrature methods are most commonly used. However, as pointed out in *Hartzel et al.* (2001), quadrature methods for integration are feasible only for integrals with dimension up to about 5 or 6. MCEM algorithm may be used to handle such situations, which synthesizes Monte Carlo approximation for integrals and EM algorithm to obtain consistent parameter estimates.

We target at situations where quadrature methods are not feasible. We generalize the artificial mixture method for independent nominal response to correlated nominal responses. Our method is comparable to MCEM in terms of convergence rate, but faster and easier to implement due to the simplificity of the likelihood function and the dimension reduction of the parameter space in the M-step in the optimization procedure. Our method transforms the complex multinomial likelihood to Poisson-type likelihoods and hence allows for the estimates to be obtained iteratively solving a set of independent low-dimensional problems. The methodology is applied to real data and studied by simulations. Research results on this topic was published in *Wang and Tsodikov* (2010).

1.3 Discrete Failure Time Data

For large-scale studies, failure time data are often collected in a discrete time scale, implying a discrete failure time model or a model for grouped failure time data (*Pipper and Ritz* (2006)). Such studies are characterized by a large number of ties and distinct event time points. if the failure is truly discrete or grouped on purpose to a larger time unit, treating such times as discrete is more appropriate. As a consequence, discrete failure time models have gained more attention lately. For large data sets, this implies that the parameter space is high-dimensional when we adopt a nonparametric form of the baseline hazards, since the number of baseline hazard parameters equals the number of distinct event time points. Absent partial likelihood machinery that is specific to the continuous form of the likelihood, optimization of the likelihood in such high-dimensional parameter spaces becomes a challenge.

Researchers explored methods to deal with many ties and/or treat failure time as discrete, such as *Prentice and Gloeckler* (1978),*Stewart and Pierce* (1982), *Johnson and Christensen* (1986), *Sinha et al.* (1994), *Yu et al.* (2004), *Pipper and Ritz* (2006), *Zhao and Zhou* (2008), *Li et al.* (2008), *Yu et al.* (2009) and etc. Most of existing studies focused on a specific model for discrete failure time data such as Cox-type models. Even with the Cox model in discrete form problems with method stability were reported in high-dimentional cases *Prentice and Gloeckler* (1978).

We propose a general approach for a class of transformation models that shows stable behavior in high-dimensional optimization problems.

We first study the use of discrete proportional odds(PO) model in this situation. We propose a Minorization-maximization(MM)-type algorithm and a recursive procedure for the high-dimensional optimization problem and compare these two methods with a traditional full likelihood maximization method - Quasi-Newton method.

We then apply artificial mixture technique to extend the base discrete PO model and proportional hazards(PH) model to a class of discrete transformation models.

Our proposed methods are superior in terms of accuracy and speed for parameter estimation in such a high-dimensional cases.

1.3.1 Discrete proportional odds model

The Cox PH model (*Cox* (1972)) has been the most common choice for failure time data. The PO model, which has been a popular tool for ordinal data, has become widely accepted since its first application to failure time data by *Bennett* (1983a). Although there are situations where the Cox PH model and PO model differ, in most studies with limited follow up, the survivor function is close to one resulting in little difference in these two models. We study the PO model for discrete failure time data due to its methodological and computational convenience.

As pointed out earlier, models for discrete failure time data in large data sets often contain a large number of parameters, including regression parameters and the baseline survivor functions. Joint estimation of regression parameters and baseline survivor functions has been subject to the curse of dimensionality. Researchers has explored methods to simplify the computation, such as MM algorithm proposed in *Hunter and Lange* (2002).

In our research, we further developed two methods with the discrete PO model: Difference Convex Algorithm(DCA) and a recursive procedure. We compare these two methods with Quasi-Newton method. Both DCA and recursive procedure work better than Quasi-Newton method in terms of accuracy and speed. The results of this research provide basic support for the PO mixture method proposed in Chapter IV.

1.3.2 Artificial Mixture Methods for Discrete Failure Time Data

The presence of many ties and a large number of distinct event time points in discrete failure time data poses a challenge of a high-dimensional model without the simplicity of the continuous model. Conventional methods to treat ties or jointly fit the full model are computationally prohibitive in large samples. We consider a general discrete transformation model and mediate the problem by changing the model form at the “complete-data”level (conditional on artificial variables). Two complete data representations of a given discrete transformation model are studied: PH and PO mixture frameworks. In the PH mixture framework, we reduce the high-dimensional optimization problem to many one-dimensional problems. In the PO mixture framework, an MM-type algorithm can be applied to simplify the optimization. Meanwhile, a recursive procedure can also be utilized. We advocate the PO mixture method as a result of this study. We apply our advocated PO mixture method to real data sets and conduct simulation studies fitting discrete PH model using PO mixture method. Simulation studies support our findings.

CHAPTER II

A Self-consistency Approach to Multinomial Logit Model with Random Effects

Key Words: QEM algorithm, Multinomial logit model with random effects

2.1 Introduction

Multinomial logit model with random effects is a common choice in the analysis of correlated nominal data in biomedical science. Such correlation could come from repeated measures or clustered observations. The complex form of the likelihood function and the presence of random effects make the computation costly. The presence of random effects implies computationally expensive multi-dimensional integrals.

A number of approaches have been proposed in the literature to overcome computational difficulties both in standard multinomial logit model and multinomial logit mixed effects model. *Breslow and Clayton* (1993) advocated penalized quasi-likelihood estimation approach to avoid the complex form of multinomial likelihood. *Chen and Kuo* (2001) used the fact that the multinomial distribution can be derived from a set of Poisson random variables conditionally on their total being fixed (*Mccullagh and Nelder* (1989)) and suggested transforming the multinomial problem to Poisson log-linear or non-linear model. *Lang* (1996) showed that the idea can be interpreted as using Lagrange multipliers to normalize multinomial probabilities. Because normalization needs to be enforced for each distinct covariate pattern, the Poisson log-linear transformation is restricted to discrete covariates. This introduces high-dimensional augmentation adding nuisance parameters.

It is convenient that this method can be implemented with standard software such as SAS NLMIXED. *Chen and Kuo* (2001) and *Kuss and McLerran* (2007) also considered a general non-linear modeling approach in conjunction with SAS NLMIXED. A related series of methods used multinomial likelihood directly applying numerical approximation to multidimensional integrals, such as Gaussian quadrature (*Rabe-Hesketh et al.* (2002), *Hedeker* (2003) etc.), spherical-radial quadrature (*Clarkson and Zhan* (2002)), first-order Taylor series expansion of the integrand (*Breslow and Clayton* (1993)), and Bayesian methods (*Daniels and Gatsonis* (1997)). *Tsodikov and Chefo* (2008) introduced artificial mixing variables to transform the multinomial likelihood to Poisson-type likelihood and reduce the complexity of the likelihood function. They apply EM-type algorithm that enjoys factorization of the model dimension at the M step that represents Poisson regressions. In this chapter, we extend the method to the multinomial logit model with random effects. Treating random effects and the artificial variables as missing data, we apply the generalized self-consistency approach described in *Tsodikov* (2003b) (Quasi-EM algorithm) to parameter estimation. The key benefit of this approach is that the M-step reduces to a set of low-dimensional problems as described in Section 2.3.

EM-type methods are advocated because of their stability with complex models. As pointed out in *Hartzel et al.* (2001), quadrature method for integration is feasible only for integrals with dimensions up to about 5 or 6. There may be a slight increase in the computation capacity these days, while it is still true that multivariate quadrature method is not computationally feasible for evaluating high-dimensional integrals. A number of Monte Carlo(MC) EM procedures were recently proposed in the literature under a more general framework - generalized linear mixed models, such as *McCulloch* (1997), *Booth and Hobert* (1999) and *Chen et al.* (2002), where Monte Carlo approximation for integration and EM algorithm are synthesized to obtain consistent parameter estimates. For ease of evaluating Monte Carlo error and automatically adjusting Monte Carlo sample size, M, *Booth and Hobert* (1999) proposed to use the MC approximations to high-dimensional integrals with independent random samples. As a consequence, M can be automatically increased until algorithm converges. *Brian S. Caffo* (2005) developed a more general method to increase the MC sample size which requires less computation, while it roughly preserves the ascent property of the observed likelihood function.

Unlike MCEM in *McCulloch* (1997), automated MCEM in *Booth and Hobert* (1999) and ascent-based MCEM in *Brian S. Caffo* (2005), where the missing data contain only the random effects, in this chapter, we introduce artificial random variables and treat them, together with the random effects, as missing data. We then use MC approximation to high-dimensional integrals with independent random samples in E-step. Our approach is different in the treatment of the M-step - the introduction of artificial variables simplifies and factorizes the complex likelihood function in the M-step which contributes to greater computational efficiency in a mixed multinomial subclass of generalized linear mixed models. To accelerate the MCEM algorithm, we propose a simple method to increase M, which does not require us to evaluate the MC error. At the same time, it roughly preserves the ascent property of observed likelihood function, since it is very similar to a special case of *Brian S. Caffo* (2005), as shown later in the paper.

In Section 2, we specify the multinomial logit model with mixed effects. In Section 3, we specify estimation steps. Variance estimation is described in Section 4. We summarize the detailed estimation procedure in Section 5, followed by simulation studies in Section 6. The method is applied to a real dataset in Section 7.

2.2 Model specification

Suppose the categorical response, Y_{ij}^{*}, has R categories, indexed as $r(r = 1,...,R)$. Clusters of correlated data are indexed by i (i = 1,..., I). Repeated measures within a cluster are indexed by j (j =1,...,T_i). Let Y_{ijr}=$I(Y_{ij}^{*} = r)$. Hence, response category probability $p_{ijr} = \mathrm{E}[Y_{ijr}] = P(Y_{ij}^{*} = r)$. Let $\boldsymbol{X}_{ij}$ denote the column vector of exploratory variables for the j-th observation in the i-th cluster, in which the first element is one, corresponding to the intercept in model (II.1). And let $\boldsymbol{Z}_{ij}$ denote the column vector of random effects for the j-th observation in the i-th cluster.

Assume category R is the reference category. The logits compare any category $r = 1, ..., R-1$ with the reference category. The model is called reference-cell logit random effects model as described in *Agresti* (2002).

$$log(\frac{p_{ijr}}{p_{ijR}}) = \boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta_r} + \boldsymbol{Z}_{ij}^{\mathrm{T}}\boldsymbol{b}_{ir}, r = 1, ..., R-1, \tag{II.1}$$

where $\boldsymbol{\beta}_r$ is the fixed effects coefficient vector of length $p+1$, corresponding to an intercept and p covariates. $\boldsymbol{\beta}_R$ is set to be zero. $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{R-1}$ are to be estimated. $\boldsymbol{b}_{ir}$ is the random effects coefficient vector. The assumption on the distribution of $\{\boldsymbol{b}_{ir}\}_{r=1}^{R-1}$ is arbitrary and our choice will be specified in Section 3.

Define Y_{ij} as a column vector with r^{th} element $(Y_{ij})_r$ being Y_{ijr} . Under a GLM setting, p_{ijr} is linked with the linear predictor, $\eta_{ijr} = \boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta_r} + \boldsymbol{Z}_{ij}^{\mathrm{T}}\boldsymbol{b}_{ir}$, through the reference-cell logit function $g(.) = \log(\frac{p_{ijr}}{p_{ijR}})$.

$$p_{ijr} = \frac{\theta_{ijr}}{1 + \sum_{l=1}^{R-1} \theta_{ijl}}, r = 1, ..., R-1 \tag{II.2}$$

In the rest of the paper, we use the following notations: θ_{ijr}=exp$\{\eta_{ijr}\}$. $\boldsymbol{b}_i$ is a vector consisting of all elements in $\{\boldsymbol{b}_{ir}\}_{r=1}^{R-1}$. $\boldsymbol{Y}_i = \{\boldsymbol{Y}_{ij}\}_{j=1}^{T_i}$. E_X indicates statistical expectation taken with respect to random variable X.

2.3 Estimation

2.3.1 Artificial Mixture

Tsodikov and Chefo (2008) introduced artificial variables $\{U_{ij}\} \overset{iid}{\sim} \mathrm{Exp}(1)$ to standard multinomial logit model, to transform the multinomial likelihood to Poisson-type likelihood. This approach also works when repeated measures are present. For standard exponentially distributed variable, U, the Laplace transform has the form

$$\mathcal{L}(s) = \mathrm{E}_U \left\{ e^{-Us} \right\} = \frac{1}{1+s}. \tag{II.3}$$

Observing the similarity between (II.3) and the multinomial probabilities (II.2), we can write p_{ijr} in the artificial mixture form

$$p_{ijr} = \theta_{ijr} \mathrm{E}_{U_{ij}} \left\{ \exp \left[-U_{ij} \sum_{l=1}^{R-1} \theta_{ijl} \right] \right\}.$$

If we pretend that $U_{ij}(\forall i, j)$ are observed, the expectation sign, E, can be removed in the above equation. Denote the complete-data expression by $\tilde{p}_{ijr}(\cdot|\cdot, U_{ij})$.

$$\tilde{p}_{ijr} = \theta_{ijr} \exp \left\{ -U_{ij} \sum_{l=1}^{R-1} \theta_{ijl} \right\}. \tag{II.4}$$

It is clear that $\tilde{p}_{ijr}(\cdot|\cdot, U_{ij})$ is no longer a legitimate probability. This indicates that the likelihood function constructed from $\tilde{p}_{ijr}(\cdot|\cdot, U_{ij})$ is not a legitimate probabilistic model. The self-consistency theory developed in *Tsodikov* (2003b) justifies the approach. Denote $\boldsymbol{U}_i$ as a column vector, $\boldsymbol{U}_i = (U_{i1}, ..., U_{iT_i})^T$.

2.3.2 Complete-data Likelihood

For model (II.4) with the artificial variables $\{U_{ij}\}$, the complete data include $\{\boldsymbol{X}_{ij}, \boldsymbol{Y}_{ij}, \boldsymbol{b}_{ir}, U_{ij}, \forall i, j, r\}$. The complete data likelihood function is

$$L_{\text{CD}}(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \prod_i f(\boldsymbol{y}_i, \boldsymbol{b}_i, \boldsymbol{U}_i; \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \prod_i f(\boldsymbol{y}_i \,|\, \boldsymbol{b}_i, \boldsymbol{U}_i; \boldsymbol{\beta}) \times f(\boldsymbol{b}_i) \times f(\boldsymbol{U}_i),$$

where $f(\boldsymbol{y}_i \,|\, \boldsymbol{b}_i, \boldsymbol{U}_i; \boldsymbol{\beta}) = \prod_{j,r} \{\tilde{p}_{ijr}\}^{y_{ijr}}$ is the conditional multinomial density for the i-th cluster with illegitimate probabilities $\{\tilde{p}_{ijr}\}$. It can be derived that

$$\log \{f(\boldsymbol{y}_i \,|\, \boldsymbol{b}_i, \boldsymbol{U}_i; \boldsymbol{\beta})\} = \sum_{j,r} \{y_{ijr} \log(\theta_{ijr}) - U_{ij}\theta_{ijr}\}$$

Under the following assumptions: (1)$\{U_{ij}\} \overset{iid}{\sim}$ Exp(1); (2)U_{ij} and $\boldsymbol{b}_{ir}$ are independent for any j, r; and (3)$\boldsymbol{b}_i$ follows multivariate normal distribution, $\boldsymbol{b}_i \sim MVN(\boldsymbol{0}, \boldsymbol{\Sigma})$, the complete data likelihood can be written as

$$L_{\text{CD}}(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \prod_i \{f(\boldsymbol{y}_i \,|\, \boldsymbol{b}_i, \boldsymbol{U}_i; \boldsymbol{\beta}) \times f(\boldsymbol{b}_i; \boldsymbol{\Sigma}) \times f(\boldsymbol{U}_i)\}. \tag{II.5}$$

Denote the corresponding log-likelihood function as ℓ_{CD} and the contribution of the i-th cluster to the log likelihood function $\ell_{\text{CD},i}$.

In the likelihood function (II.5), $\{\boldsymbol{b}_i\}$ and artificial variables $\{U_{ij}\}$ can be treated as missing data, hence EM-type algorithm can be used for parameter estimation.

2.3.3 Estimation steps

Let $\boldsymbol{\phi}$=$(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ be the combined parameter vector. The EM algorithm is an iterative process. Given parameter estimates from the k-th iteration, the $(k+1)$-th iteration can be formulated as $\text{Argmax}_\phi \left\{Q(\boldsymbol{\phi} \,|\, \boldsymbol{\phi}^{(k)})\right\}$, where $Q(\boldsymbol{\phi} \,|\, \boldsymbol{\phi}^{(k)})$ forms the E-step.

$$Q(\boldsymbol{\phi} \,|\, \boldsymbol{\phi}^{(k)}) = \text{E}\left\{\log(L_{\text{CD}}(\phi)) \,|\, \boldsymbol{y}; \boldsymbol{\phi}^{(k)}\right\}. \tag{II.6}$$

The M-step finds $\boldsymbol{\phi}^{(k+1)}$ such that

$$Q(\boldsymbol{\phi}^{(k+1)} \,|\, \boldsymbol{\phi}^{(k)}) \geq Q(\boldsymbol{\phi} \,|\, \boldsymbol{\phi}^{(k)}), \tag{II.7}$$

for any $\boldsymbol{\phi}$ in the parameter space.

To evaluate the expectation in E-step, we need to evaluate the multi-dimensional integrals with respect to $\boldsymbol{b}_i$ and $\boldsymbol{U}_i$.

$$Q(\boldsymbol{\phi} \,|\, \boldsymbol{\phi}^{(k)}) = \mathrm{E}\left\{\ell_{\mathrm{CD}} \,|\, y; \boldsymbol{\phi}^{(k)}\right\} = \sum_i \left\{\int \ell_{\mathrm{CD},i} \times f(\boldsymbol{b}_i, \boldsymbol{U}_i \,|\, \boldsymbol{y}_i, \phi^{(k)}) d\boldsymbol{b}_i d\boldsymbol{U}_i\right\},$$

where $f(\boldsymbol{b}_i, \boldsymbol{U}_i \,|\, \boldsymbol{y}_i, \phi^{(k)}) \propto f(\boldsymbol{y}_i \,|\, \boldsymbol{b}_i, \boldsymbol{U}_i; \boldsymbol{\beta}^{(k)}) \times f(\boldsymbol{b}_i; \boldsymbol{\Sigma}^{(k)}) \times f(\boldsymbol{U}_i)$. The normalizing constant is $\int f(\boldsymbol{y_i} \,|\, \boldsymbol{b_i}, \boldsymbol{U_i}; \boldsymbol{\beta}^{(k)}) \times f(\boldsymbol{b}_i; \boldsymbol{\Sigma}^{(k)}) \times f(\boldsymbol{U}_i) d\boldsymbol{b}_i d\boldsymbol{U}_i$, denoted as $a_i^{(k)}$.

Note that the dimension of the integrals easily exceeds the limit that the computer can handle using quadrature methods when either R or the number of random effects is large. In this case, Monte Carlo (MC) approximations to the intractable integrals are often suggested. *Booth and Hobert* (1999) and *Hartzel et al.* (2001) proposed to use rejection sampling or importance sampling based on $f(\boldsymbol{b}_i, \boldsymbol{U}_i \,|\, \boldsymbol{y}_i, \phi^{(k)})$ as it's not easy to directly sample from f. By doing so, one does not need to evaluate the normalizing constants $a_i^{(k)}$. In this research, we randomly select M independent samples from multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}^{(k)}$ for $\boldsymbol{b}_i$ and from standard exponential distribution for U_{ij}. Denote these samples as $\boldsymbol{b}_{i,m}, U_{ij,m}$ for $m = 1, ..., M$. Hence, $Q(\boldsymbol{\phi} \,|\, \boldsymbol{\phi}^{(k)})$ can be approximated by $\tilde{Q}(\boldsymbol{\phi} \,|\, \boldsymbol{\phi}^{(k)})$, and $a_i^{(k)}$ by $\tilde{a}_i^{(k)}$.

$$\tilde{Q}(\boldsymbol{\phi} \,|\, \boldsymbol{\phi}^{(k)}) = \frac{1}{M} \sum_i \sum_{m=1}^{M} \left\{\frac{f(\boldsymbol{y_i} \,|\, \boldsymbol{b}_{i,m}, \boldsymbol{U}_{i,m}; \boldsymbol{\beta}^{(k)})}{\tilde{a}_i^{(k)}} \log f(\boldsymbol{y}_i, \boldsymbol{b}_{i,m}, \boldsymbol{U}_{i,m}; \boldsymbol{\phi})\right\}, \tag{II.8}$$

where $\tilde{a}_i^{(k)} = \frac{1}{M}\sum_{m=1}^{M} f(\boldsymbol{y_i} \,|\, \boldsymbol{b}_{i,m}, \boldsymbol{U}_{i,m}; \boldsymbol{\beta}^{(k)})$. Compared to rejection sampling and/or importance sampling methods widely used in the literature on generalized mixed effects models such as *Booth and Hobert* (1999) and *Hartzel et al.* (2001), the main advantage of our sampling scheme is that one does not need to look for a trial distribution as we sample from the exact distributions, $MVN(\mathbf{0}, \mathbf{\Sigma}^{(k)})$ and Exp(1). The method synthesizes MC approximation and QEM algorithm developed in *Tsodikov* (2003b), hence we will call it MCQEM algorithm to differentiate it from a regular M-step procedure.

Denote $w_{i,m}^{(k)} = \frac{f(\boldsymbol{y_i}\,|\,\boldsymbol{b}_{i,m}, \boldsymbol{U}_{i,m}; \boldsymbol{\beta}^{(k)})}{\tilde{a}_i^{(k)}}$. The M-step becomes

$$\underset{\boldsymbol{\beta},\, \mathbf{\Sigma}}{\text{Argmax}} \left\{ \sum_{i,m} w_{i,m}^{(k)} \left\{ \log f(\boldsymbol{y}_i \,|\, \boldsymbol{b}_{i,m}, \boldsymbol{U}_{i,m}; \boldsymbol{\beta}) + \log f(\boldsymbol{b}_i; \mathbf{\Sigma}) + \log f(\boldsymbol{U}_i) \right\} \right\}$$

We obtain $\boldsymbol{\beta}^{(k+1)}$ and $\mathbf{\Sigma}^{(k+1)}$ with a two-stage procedure. First, we estimate $\boldsymbol{\beta}^{\boldsymbol{(k+1)}}$ with $\mathbf{\Sigma}$ fixed at $\mathbf{\Sigma}^{(k)}$, achieved by maximizing the first term in $\tilde{Q}(\boldsymbol{\phi} \,|\, \boldsymbol{\phi}^{(k)})$, which is the log-likelihood function as if $\boldsymbol{b}_{i,m}, \boldsymbol{U}_{i,m}$ are observed. Plugging in $\log f(\boldsymbol{y}_i \,|\, \boldsymbol{b}_i, \boldsymbol{U}_i; \boldsymbol{\beta})$ specified early, the objective function in this step can be written as

$$\begin{aligned} &\sum_{i,m} w_{i,m}^{(k)} \left\{ \log f(\boldsymbol{y}_i \,|\, \boldsymbol{b}_{i,m}, \boldsymbol{U}_{i,m}; \boldsymbol{\beta}) \right\} \\ &= \sum_r \sum_{i,m} w_{i,m}^{(k)} \sum_j \left\{ y_{ijr}(\boldsymbol{X}_{ij}^T \boldsymbol{\beta_r} + \boldsymbol{Z}_{ij}^T \boldsymbol{b_{ir}}) - U_{ij} \exp(\boldsymbol{Z}_{ij}^T \boldsymbol{b_{ir}}) \exp(\boldsymbol{X}_{ij}^T \boldsymbol{\beta_r}) \right\}. \end{aligned}$$

It's obvious that we can view the objective function as a sum of $R-1$ components, each corresponding to the contribution of a Poisson likelihood with $I\left\{y_{ij} = r\right\}$ as the response and $U_{ij,m} \exp(\boldsymbol{Z}_{ij}^T \boldsymbol{b}_{ir,m})$ as the offset term. As a consequence, we can obtain $\boldsymbol{\beta}_r^{(k+1)} (\forall r)$ from a weighted Poisson regression with weight $w_{i,m}^{(k)}$. For each Poisson regression, the original data are replicated M times.

Second, we estimate $\mathbf{\Sigma}^{(k+1)}$ with $\boldsymbol{\beta}$ fixed at $\boldsymbol{\beta}^{(k+1)}$. This step only relates to the second term in $\tilde{Q}(\boldsymbol{\phi} \,|\, \boldsymbol{\phi}^{(k)})$, which is simply the likelihood contribution of a linear combination of multivariate normal samples with weight $w_{i,m}^{(k)}$.

Suppose we have p exploratory variables. Then, we have $(p+1)\times(R-1)$ coefficient parameters to estimate in total. The M-step of our method estimates $p+1$ parameters at a time, and repeats the process $R-1$ times. In contrast, without introducing the artificial variables, one has to estimate all $(p+1)\times(R-1)$ coefficient parameters in one step. In this sense, our method results in greater simplicity in the M-step. This advantage is demonstrated by a simulation study in Section 2.6.

The process iterates between E-step and M-step until a convergence criterion is met. In this research, the process will stop when the relative change of parameter estimates does not exceed a given tolerance, for instance, 1e-8.

2.3.4 Escalation of the MC sample size

Convergence may not be obtained when using a constant MC sample size, M, due to a persistent MC error (*Brian S. Caffo* (2005)). On the other hand, it's not efficient to use large M at an early stage of the EM iterations. Sometimes, it's critical to increase M automatically over iterations. *Booth and Hobert* (1999) proposed to use normal approximation to MC error and construct a $100(1-\alpha)\%$ confidence ellipsoid for $\boldsymbol{\phi}^{(k+1)}$ and hence determine to increase M or not, by comparing whether $\boldsymbol{\phi}^{(k)}$ lies in the confidence ellipsoid. *Brian S. Caffo* (2005) developed a data-driven strategy for increasing M which preserves the ascent property of the likelihood function over iterations with large probabilities. They constructed an approximate confidence interval for the change of $Q(\boldsymbol{\phi}^{(k+1)}\,|\,\boldsymbol{\phi}^{(k)})$ over two consecutive iterations and hence determine whether to increase M.

Methods proposed in the aforementioned articles all require a large amount of computation such as approximating MC error, second-order derivatives of $Q(\boldsymbol{\phi}\,|\,\boldsymbol{\phi}^{(k)})$ and so on. There does not exist standard software to implement these methods. It's obviously tedious to derive these formulas for those who want only to apply these methods to real problems. In this research, we suggest a simpler strategy for the escalation of M, which is similar to a special case as in *Brian S. Caffo* (2005) with α=0.5. We increase M by a given factor whenever observed likelihood function at $\boldsymbol{\phi^{(k+1)}}$ is less than that evaluated at $\boldsymbol{\phi^{(k)}}$. The observed likelihood function can be easily approximated using MC method.

2.4 Variance Estimation

Louis (1982) described a variance estimation method based on the observed information matrix for EM algorithm. Let S be the complete data score vector. $S = \frac{\partial \log L_{\text{CD}}}{\partial \phi}$, where L_{CD} is given by (II.5). The observed information matrix, I_{obs},can be estimated through the formula: $I_{\text{obs}} = \text{E}_{\boldsymbol{U},\boldsymbol{b}} \left\{ I_{\text{CD}}(\boldsymbol{U}, \boldsymbol{b}) - SS^T(U,W) | L_{\text{CD}}(\boldsymbol{U}, \boldsymbol{b}) \right\}$, where $I_{\text{CD}} = -\frac{\partial^2 \log L_{\text{CD}}}{\partial \phi \partial \phi^{\text{T}}}$ is the complete data information matrix. I_{CD} overestimates the sample information as it assumes that missing data are known. The expression for I_{obs} represents the so-called missing information principle that observed information is complete information minus the missing information. Inverting the observed information matrix, we get the estimated covariance matrix. The procedure is tedious as it requires the calculation of the first and second derivatives. Hence, we seek for an alternative method to estimate the covariance matrix.

We estimate the variance components based on an approximation to the likelihood curvature by a quadratic form fit by linear regression based on the sampled points on likelihood surface around MLE. Related idea was used in *Neilsen et al.* (1992).

Applying the Taylor expansion to the log-likelihood function $\ell(\boldsymbol{\phi})$ at $\boldsymbol{\phi}_{\text{MLE}}$ up to the second order term, we get

$$\ell(\boldsymbol{\phi}) \approx \ell(\boldsymbol{\phi}_{\text{MLE}}) + (\boldsymbol{\phi} - \boldsymbol{\phi}_{\text{MLE}})^T G(\boldsymbol{\phi}_{\text{MLE}}) + (\boldsymbol{\phi} - \boldsymbol{\phi}_{\text{MLE}})^T D(\boldsymbol{\phi}_{\text{MLE}})(\boldsymbol{\phi} - \boldsymbol{\phi}_{\text{MLE}})/2.$$

where $G(\cdot)$ and $D(\cdot)$ are the gradient vector and the Hessian matrix, respectively.

Since $G(\boldsymbol{\phi}_{\text{MLE}}) \approx 0$,

$$\ell(\boldsymbol{\phi}) - \ell(\boldsymbol{\phi}_{\text{MLE}}) \approx (\boldsymbol{\phi} - \boldsymbol{\phi}_{\text{MLE}})^T D(\boldsymbol{\phi}_{\text{MLE}})(\boldsymbol{\phi} - \boldsymbol{\phi}_{\text{MLE}})/2.$$

Denote $\ell(\boldsymbol{\phi}) - \ell(\boldsymbol{\phi}_{\text{MLE}})$ as $\Delta\ell$ and $\boldsymbol{\phi} - \boldsymbol{\phi}_{\text{MLE}}$ as $\Delta\boldsymbol{\phi}$. Let $\{D(\cdot)\}_{ij} = d_{ij}$. If we randomly sample K points for $\boldsymbol{\phi}$ within the close neighborhood of $\boldsymbol{\phi}_{\text{MLE}}$, we then have

$$\Delta\ell_k = \frac{1}{2}\sum_{i,j} d_{ij}\Delta\phi_{ki}\Delta\phi_{kj} + \epsilon_k, \forall k \in 1, ..., K, \tag{II.9}$$

where $\Delta\phi_{ki}$ is the i^{th} element of $\Delta\boldsymbol{\phi}$ of the k^{th} sampled point on the log-likelihood surface. The distribution of ϵ_k is Gaussian with mean zero.

We can estimate d_{ij} in the above equation by minimizing the following sum of squares:

$$\begin{aligned} SS &= \sum_k \left\{\Delta\ell_k - \frac{1}{2}\sum_{i,j} d_{ij}\Delta\boldsymbol{\phi}_{ki}\Delta\boldsymbol{\phi}_{kj}\right\}^2 \\ &= \sum_k \left\{\Delta\ell_k - \frac{1}{2}\sum_i d_{ii}\Delta\boldsymbol{\phi}_{ki}^2 - \sum_{i<j} d_{ij}\Delta\boldsymbol{\phi}_{ki}\Delta\boldsymbol{\phi}_{kj}\right\}^2. \end{aligned}$$

Observe that the above sum of squares has the form of objective function in linear regression with response being $\Delta\ell_k$, independent variables $\frac{1}{2}\Delta\boldsymbol{\phi}_i^2 (i = 1, ..., m)$ and $\Delta\boldsymbol{\phi}_i\Delta\boldsymbol{\phi}_j$ for any $i < j$, and the $m(m+1)/2$ unknown elements in matrix D being the corresponding regression coefficients. Or we can directly recognize (II.9) as a linear regression model without intercept term. Denote the unknown elements in the matrix D as a vector, $\boldsymbol{\gamma}$. Therefore, $\boldsymbol{\gamma}$ can be estimated, when K is large enough such that the design matrix of the linear regression model is non-singular. In other words, the Hessian matrix can be estimated through a multiple linear regression through the origin. Hence, approximations of the standard errors of the parameter estimates can be obtained easily.

Elements of the K $\Delta\phi$ points are sampled independently from a symmetric distribution with mean 0 on the log-likelihood surface. They need to be close enough to $\boldsymbol{\phi}_{\text{MLE}}$ but not too close causing numerical concerns. They are chosen to be in the $\frac{1}{\sqrt{n}}$-range around $\boldsymbol{\phi}_{\text{MLE}}$.

The remaining part of the method is to solve (II.9). Simulation study in Section 2.6 suggests that the variance approximation method works well.

2.5 Algorithm

In this section, we outline the MCQEM algorithm in details.

1. Set initial values for $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$ and MC sample size M. We may set $\boldsymbol{\beta}^{(0)}$ to be the coefficient estimates by fitting a standard multinomial logit model, ignoring the within-cluster correlation. It is a good set of starting values especially when the within-cluster correlation is small. $\boldsymbol{\Sigma}^{(0)}$ can be identity matrix. $M^{(0)}$=200. Let k=0.

2. Calculate the observed log-likelihood function at $\boldsymbol{\phi}^{(k)}$, denoted as $\ell_{obs}(\boldsymbol{\phi}^{(k)})$.

3. Randomly draw $M^{(k)}$ independent samples from MVN$(\mathbf{0}, \boldsymbol{\Sigma}^{(k)})$ for $\boldsymbol{b}_{i,m}$ and from Exp(1) for $U_{ij,m}$.

4. Calculate the weight $w_{i,m}^{(k)}$ and offset $U_{ij,m}\exp(\boldsymbol{Z}_{ij,m}^T\boldsymbol{b}_{ir})$. Estimate $\boldsymbol{\beta}$ (denoted as $\boldsymbol{\beta}^{(new)}$ through $R-1$ Poisson regressions with the above weight and offset and $\boldsymbol{\Sigma}^{(new)}$ from a set of multivariate random samples with the weight.

5. Evaluate the observed log-likelihood function at $\boldsymbol{\phi}^{(new)}$, denoted as $\ell_{obs}(\boldsymbol{\phi}^{(new)})$ and compare it with $\ell_{obs}(\boldsymbol{\phi}^{(k)})$. If $\ell_{obs}(\boldsymbol{\phi}^{(k)}) < \ell_{obs}(\boldsymbol{\phi}^{(new)})$, accept $\boldsymbol{\phi}^{(new)}$ as $\boldsymbol{\phi}^{(k+1)}$, update k, $\ell_{obs}(\boldsymbol{\phi}^{(k)})$ and $M^{(k+1)} = M^{(k)}$. Go back to step 3. Otherwise, increase $M^{(k)}$ by a give factor, for instance, 5/4, without updating other quantities and go back to Step 3.

6. Iterate over these steps until convergence.

7. Estimate the Hessian matrix as described in Section 2.4 and obtain the standard errors.

2.6 Simulation studies

In this section, two simulation studies are performed. First, a simulation study is conducted to illustrate the benefit of introducing artificial variables to simplify the M-step. Second, we do a simulation study to illustrate the MCQEM algorithm using a random intercept multinomial logit model.

2.6.1 Simulation study 1

The introduction of artificial variables makes MCQEM different from MCEM in estimating the parameters at the M-step. In the usual MCEM algorithm, a standard multinomial logit model is fitted for the M-fold duplicated dataset, treating simulated $\boldsymbol{b}_{i,m}$ as observed. All $(p+1) \times (R-1)$ coefficient parameters are estimated jointly. In contrast, in MCQEM algorithm, we fit $(R-1)$ independent Poisson regressions, each having $(p+1)$ parameters as if $\boldsymbol{b}_{i,m}$ and $\boldsymbol{U}_{i,m}$ were observed. This simulation study is designed to show the advantage of doing so.

We simulate three independent random variables of size 1000 from standard normal distribution, and dichotomize the third variable using zero as a cutpoint. We vary the number of categories in the $R = 3, ..., 15$ range to highlight the computational benefit achieved at the M-step. The set of true parameter values for $R = 15$ is a 4 by 14 matrix. For $R < 15$ we use the corresponding $R-1$ column submatrix.

For a given R, we apply two methods to twenty simulated datasets. The first method is to estimate all the parameters jointly in a multinomial logit model. The second is to fit $R-1$ Poisson regressions with prespecified artificial variables, which

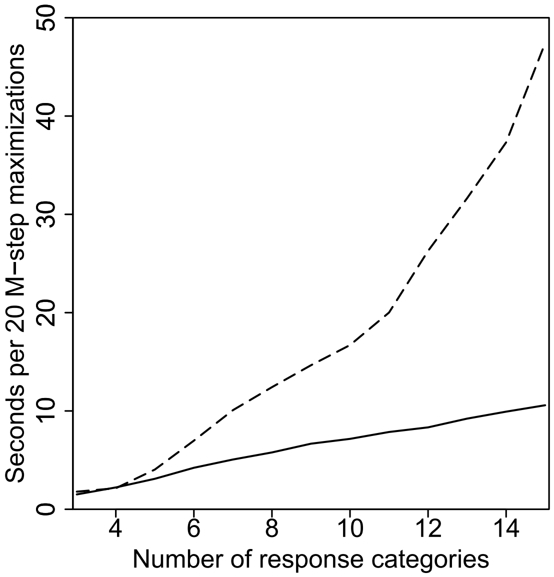

Figure 2.1: Number of response categories vs. time used per twenty estimation procedures (dashed line: multinomial logit model; solid line: $R-1$ Poisson models with prespecified artificial variables)

are drawn from standard exponential distribution. The convergence criteria for both methods are set to be the same. The setting mimics what we have at the M-steps of MCEM and MCQEM. Hence, it can be used to compare the M-steps of these two methods.

Figure 1 shows the relationship between the number of response categories, R, and total time used in parameter estimation for all twenty simulated datasets.

As shown in the figure 2.1, fitting $R-1$ Poisson regressions always take less time than estimating all parameters jointly in the multinomial logit model. The advantage of reduced dimension increases rapidly with the number of categories R.

2.6.2 Simulation study 2

We explore the performance of the MCQEM algorithm and the variance approximation method for correlated multinomial data using a random intercept model. Four exploratory variables are included along with random intercepts with the following parameterization.

$$\log(\frac{p_{ijr}}{p_{ijR}}) = \boldsymbol{X_{ij}}^{\mathrm{T}}\boldsymbol{\beta_r} + b_i\alpha_r, r = 1, ..., R-1, \tag{II.10}$$

where the random intercept b_{ir} is re-parameterized as $b_i\alpha_r$. This is a special case of the general model (II.1) with random intercepts only and full dependency between $b_{i1}, ..., b_{i(R-1)}$.

We consider a categorical response with three categories, regressed on the covariates of the real data described in Section 2.7. In this data set, there are 36 clusters of size seven to ten. The true parameters used to simulate the responses are taken to be the parameter estimates of the real dataset obtained through the proposed MCQEM algorithm. It is compared with adaptive Gauss-Hermite quadrature methods using NLMIXED in SAS.

Table 1 summarizes results of 500 simulations. The second column shows the true parameter values used to simulate data. The third column lists the empirical mean of parameter estimates using MCQEM algorithm proposed in this chapter. The first number in the parenthesis is the standard error of the parameter estimate using variance approximation method described in Section 2.4. The second number in the parenthesis indicates the percentage of the estimated 95% confidence intervals covering the true parameter value. The fourth column contains those from NLMIXED in SAS, using adaptive Gauss-Hermite quadrature method to approximation the intractable integrals.

The empirical mean of parameter estimates of 500 simulations are very close to the true parameter values, taking the small sample size (36 clusters in total) into consideration. This supports the proposed methodology for multinomial logit model with random effects.

As seen in Table 1, among all the 95% confidence intervals constructed based on our standard error estimates, about 95% contain the true parameter values. The

Parameters	True parameter values	QEM(se, coverage rate*100)	NLMIXED(se, coverage rate*100)
α_1	1.652	1.609 (0.387, 95.2)	1.566 (0.365,90.8)
α_2	1.293	1.253 (0.326, 94.8)	1.224 (0.307,92.6)
β_{10}	-2.954	-3.049 (0.848, 93.8)	-3.057 (0.835,96.4)
β_{11}	1.333	1.368 (1.007, 92.4)	1.370 (0.917,92.4)
β_{12}	0.316	0.323 (0.903, 91.8)	0.320 (0.852,92.0)
β_{13}	4.073	4.219 (0.611, 94.0)	4.220 (0.607,94.8)
β_{14}	-2.567	-2.660 (0.556, 95.2)	-2.660 (0.543,95.2)
β_{20}	-1.097	-1.133 (0.634, 94.4)	-1.138 (0.611,95.0)
β_{21}	0.283	0.293 (0.812, 92.0)	0.293 (0.723,90.6)
β_{22}	-0.428	-0.469 (0.712, 97.0)	-0.471 (0.706,94.8)
β_{23}	2.390	2.452 (0.429, 95.0)	2.450 (0.424,95.8)
β_{24}	-1.612	-1.642 (0.496, 94.8)	-1.642 (0.491,95.2)

Table 2.1: Summary statistics of parameter estimates obtained by QEM algorithm and standard errors by the variance approximation method described in Section 2.4, compared with adaptive Gauss-Hermite quadrature method through NLMIXED in SAS. Results are based on 500 simulations.

empirical means of the standard errors are also very close to the empirical standard errors of the parameter estimates. These evidences suggest that the variance approximation method works well and is stable. Numerical errors with evaluation of standard errors lead to problems less than 0.6% of all simulations.

Comparing the third and fourth column in Table 1, our method is competitive to adaptive Gauss-Hermite quadrature method for low-dimensional problems. It is worth pointing out again that quadrature methods are no longer feasible for high-dimensional integrals due to computation cost.

2.7 Application to a real data set

Kuss and McLerran (2007) analyzed a data set on physician's recommendations and preferences in traumatic brain injury (TBI) rehabilitation. They kindly made their data available for our study. For each of multiple TBI disease histories, 36 physicians were asked to choose an optimal rehabilitation setting from the following: in-patient, day-clinic and out-patient. Four binary covariates were measured as well. The four covariates are answers to the following 4 questions: (1) Is the physician a neurologist? (2) Is the physician a specialist? (3) Is time since last event longer than three months? (4) Is the patient severely handicapped after TBI? The four covariates are named neuro, special, time and severity respectively. Recommendations within the same physician are expected to be correlated. Treating the rehabilitation setting as nominal, this forms a typical multinomial problem with repeated measures.

To identify factors that influence setting recommendations, we apply our method to this data set. Random intercepts only model with unstructured covariance matrix and a model in the form of (II.10) are compared by likelihood ratio test with a p-value 0.32, which shows no evidence against model (II.10) for this real dataset. Hence, we fit model (II.10) for the real dataset.

Results from our method are compared with (1) independence model, namely, standard multinomial logit model, ignoring the correlation within cluster(shown in Column 2 in Table 2); (2) model (II.10), implemented in SAS using PROC NLMIXED with Gauss-Hermite quadrature method for integral approximation(shown in Column 3 in Table 2). Estimates of the independence model are obtained through PROC LOGISTIC in SAS and results from PROC NLMIXED are based on SAS codes modified from those provided in *Kuss and McLerran* (2007). Results of our method are shown in MCQEM column in Table 2. In order to be consistent with *Kuss and McLerran* (2007), we choose in-patient as the reference category.

From Table 2, comparing the likelihood values, we see that the independence model does not provide as good a fit as the correlated model. For the MCQEM algorithm and PROC NLMIXED with Gaussian quadrature method for multi-dimensional integration, the parameter estimates are all very close. These two methods also provide almost the same maximum log-likelihood. The estimated standard errors from

Parameters	IM	NLMIXED	MCQEM
α_{OP}	-	1.611(0.370)	1.652(0.442)
α_{DC}	-	1.246(0.311)	1.293(0.337)
(Out-patient)			
Intercept	-2.429(0.566)	-2.948(0.810)	-2.954(0.756)
Neuro	1.073(0.481)	1.319(0.910)	1.333(0.898)
Special	0.296(0.426)	0.280(0.851)	0.316(0.855)
Time	3.150(0.456)	4.088(0.573)	4.073(0.551)
Severity	-2.022(0.441)	-2.571(0.526)	-2.567(0.522)
(Day-clinic)			
Intercept	-0.879(0.430)	-1.073(0.595)	-1.097(0.580)
Neuro	0.088(0.395)	0.269(0.713)	0.283(0.753)
Special	-0.446(0.403)	-0.453(0.697)	-0.428(0.744)
Time	1.723(0.324)	2.383(0.403)	2.390(0.401)
Severity	-1.204(0.414)	-1.609(0.471)	-1.612(0.461)
-2ℓ	492.9	462.2	462.2
No. of parameters	10	12	12

Table 2.2: Comparison of Parameter Estimates: Multinomial logit model (with or without random effects) fit to physician's recommendations and preferences data in traumatic brain injury (TBI) rehabilitation (IM: Independence Model; NLMIXED: Proc NLMIXED in SAS)

PROC NLMIXED and the quadratic variance estimation method proposed in this chapter are also very close. The benefit of our method is that it simplifies the M steps in estimating the coefficient parameters and avoids derivatives for variance estimation.

2.8 Discussion

The MCQEM algorithm proposed in this research is a MCEM-type algorithm with artificial variables introduced to break the dimension of the M-step into a series of Poisson regression sub-problems. Artificial variables slightly increase the dimension of the E-step requiring larger MC sample size to achieve the same accuracy. Therefore, the MCQEM algorithm is best suited for problems with smaller cluster size.

When the number of categories becomes larger, quadrature methods become prohibitively slow. While our method as well as MCEM would require a larger number of MC sample size M in this case, it still remains feasible with larger number of categories. The MCQEM algorithm converges even when the model is non-identifiable due to empty/sparse categories.

Similar to EM algorithm, the MCQEM algorithm does not provide variance estimates automatically. We propose the variance approximation method based on the idea of *Neilsen et al.* (1992), which only requires knowledge of the log-likelihood function and MLE. This allows us to get variance estimates avoiding taking derivatives. It works well when the log-likelihood function is approximately quadratic around the MLE. When the condition is not met, the variance estimates are not reliable. The traditional variance estimates are no good in this case either. We have examined a few extreme cases, where the parameters are near the boundary of the parameter space, and the log-likelihood function is far from quadratic. Neither our method nor numerical differentiation (SAS) is reliable in those extreme cases.

When missing observations are present, if the mechanism is missing at random (MAR) or missing completely at random (MCAR), the method of this study could be extended to deal with missing observations by incorporating them into the EM framework. If missing is not at random, delicate care may be needed. Methods proposed for this purpose in the literature such as sensitivity analysis and multiple imputation (*Fitzmaurice et al.* (2008) and *Roderick J.A. Little* (2002)) should apply to our setting.

Acknowledgments

The authors thank Oliver Kuss to allow us to use their dataset analyzed in *Kuss and McLerran* (2007). The authors are also grateful to the reviewers for important and helpful comments and suggestions.

CHAPTER III

On the Estimation of Proportional Odds Model for Discrete Failure Time

3.1 Introduction

The proportional odds (PO) model has long been a popular tool of ordinal categorical data analysis (*Agresti* (2007)). With the recognition of its utility in survival analysis targeting continuous data (*Bennett* (1983a)) the model has gained more widespread use as an alternative to the Cox model (*Cox* (1972)).

Let $G(t|z)$ be a survivor function, given covariates z (assumed here to be time-independent). The PO model is built from the assumption that cumulative odds of survival $G/(1-G)$ are proportional to a baseline survival function G_0 that is assumed to be arbitrary,

$$\frac{G(t|z)}{1-G(t|z)} = \frac{G_0(t)}{1-G_0(t)}\,\theta(z^{\mathrm{T}}\beta), \tag{III.1}$$

uniformly with respect to t, where usually $G_0(t) = G(t|0)$ or $\theta(0) = 1$. This implies that the odds ratio $\theta(z^{\mathrm{T}}\beta) = \exp(z^{\mathrm{T}}\beta)$ does not depend on time, the PO assumption. The Cox model is defined based on a similar concept with a different measure of relative risk, the hazard ratio $\theta(z^{\mathrm{T}}\beta) = \exp(z^{\mathrm{T}}\beta)$,

$$\frac{\log G(t|z)}{\log G_0(t)} = \theta(z^{\mathrm{T}}\beta), \tag{III.2}$$

uniformly with respect to t, where $\Lambda(t|z) = -\log G(t|z)$ is a cumulative hazard. The hazard ratio (HR) is usually defined as an instantaneous characteristic

$$\frac{\lambda(t|z)}{\lambda_0(t)} = \theta(z^{\mathrm{T}}\beta), \tag{III.3}$$

where $\lambda(t|z) = \frac{d\Lambda(t|z)}{dt}$ is the hazard function. The cumulative and instantaneous HR effect measures (III.2) and (III.3) are the same as long as the true model is Cox.

While most of the time the family of responses reproduced by the two models are very similar, there are situations when they do differ. Let $\tilde{\theta}(t,z)$ be the time-dependent HR when the true model is PO (can be considered an HR in a misspecified Cox model),

$$\tilde{\theta}(t,z) = 1 - \frac{1-\theta(z)}{\frac{\theta(z)}{G(t|z)} + 1 - \theta(z)}, \tag{III.4}$$

where $\theta(z)$ is the odds ratio. Follows from (III.4) is a well known fact (*Kirmani and Gupta* (2001)) that the hazard ratio in the PO model attenuates to 1 in follow-up time as $G(t|z)$ decreases in t. With a proper PO model ($G(\infty|z) = 0$), the hazard ratio has a limit of 1, while with improper model (cure models, $G(\infty|z) > 0$), the limit stops short of 1, but the attenuation of HR is still happening. So the PO model is potentially useful in situations where covariates become less relevant with time as compared to the Cox model. One such example is shown in Figure 3.1. However, with limited follow-up in most studies G is close to 1 resulting in little difference between the two models, $\tilde{\theta} = \theta$ when $G = 1$ in (III.4). That makes the choice between the two a matter of methodological and computational convenience.

With continuous data, the Cox model has been by far the most convenient choice due to the advent of the partial likelihood, the martingale machinery (*Andersen et al.* (1993); *Fleming and Harrington* (2005)), and the fact that a finite dimensional maximization is used to get the maximum likelihood estimates (MLE).

Asymptotic theory for the continuous PO model has been a test case of empirical processes (*Murphy et al.* (1997); *Murphy* (2000)), estimating equations (*Cheng et al.* (1995)), rank-based transformation model methods (*Cuzick* (1988)), use of marginal likelihood (*Pettitt* (1984)), and sieve maximum likelihood (*Shen* (1998); *Huang and Rossini* (1996)).

Computationally, joint estimation of β and the baseline survivor function via semi-parametric maximum likelihood (*Bennett* (1983b)) has also been subject to the curse of dimensionality that made researchers look for alternative approaches. *Lange et al.* (2000); *Hunter and Lange* (2002) proposed to use MM algorithm for joint estimation of β and G_0 in PO model. They reparameterized the model to ensure existence of a simpler surrogate objective function minorising the likelihood and touching the likelihood surface at the current iteration point. EM algorithm (*McLachlan and Krishnan* (1997)) is a particular case of the MM where the construction of the surrogate objective function corresponds to the E-step. Maximization of the surrogate objective function corresponds to the M-step. *Tsodikov* (2003a) used an artificial mixture formulation for the PO model as

$$G(t \mid \boldsymbol{z}) = \mathrm{E}\left\{ F(t)^{U(\boldsymbol{\beta},\boldsymbol{z})} \;\middle|\; \boldsymbol{z} \right\} = \frac{\theta(z^{\mathrm{T}}\beta)}{\theta(z^{\mathrm{T}}\beta) + H(t)}, \tag{III.5}$$

where $F = \exp(-H)$ is a (transformed but still arbitrary) baseline survival function ($G_0 = (1 + H)^{-1}$), H is an arbitrary cumulative hazard, and $U = U(\boldsymbol{\beta}, \boldsymbol{z})$ is an exponential random variable with the rate $\theta(z^{\mathrm{T}}\beta)$. This approach represents the PO model as an average over artificially mixed Cox models that leads to an EM algorithm with the M-step being a computationally efficient continuous Cox model solution.

The discrete Cox model or a model for grouped data has proven to be a challenge due to the fact that the likelihood contribution of exact observations (failures) is no longer log linear in the baseline hazard function that erases the partial likelihood advantage. *Prentice and Gloeckler* (1978) proposed to maximize the full likelihood. However problems were reported when the dimension was high. *Prentice and Kalbfleisch* (2003) explored a different discrete model enforcing the log-linear likelihood structure and preserving the multiplicative form of the model in terms of the cumulative hazard. This convenience is reached at the cost of having to observe a restriction on the cumulative hazard to keep the model probabilistically consistent. Also, their model is not a Cox model for grouped data.

Observing the difficulties with the Cox model in the discrete (grouped data) case, we explore the use of the PO model in this situation. In doing so we are targeting situations when the dimension of the model is high so that the traditional full likelihood approach to fit the discrete PO model may be problematic. In this chapter we further explore two ideas with the discrete PO model. First, we observe that the grouped data likelihood for the discrete PO model retains a log-linear form in the differential of the baseline cumulative hazard resembling the likelihood of a continuous transformation model in survival analysis. This allows us to develop a Difference Convex Algorithm (DCA) (*de Leeuw* (1994); *An and Pham* (1997)) motivated by the artificial mixture method for continuous transformation models (*Tsodikov* (2003a)). The second is the idea of a recursive procedure applied to an artificially unrestricted model (relieved of the normalization restriction on probabilities) and subsequent enforcement of the restriction by a Lagrange multipliers method, initially explored in (*Tsodikov et al.* (1998)) in a two-sample test context. Both procedures will be compared to the traditional full likelihood maximization by a generic (conjugate gradients) method.

3.2 Likelihood

Let i index discrete time points, t_i, $i = 1, \ldots, K$, that define the support of a discrete survival distribution or a set of grouping intervals. Let C_i be a set of subjects who are censored at t_i, and D_i be the set of subjects who fail at t_i; j will index subjects in these sets. So, each subject is indexed by a pair (i, j) and the set that j belongs to.

The event indicator is denoted by c_{ij}, $c_{ij} = 1$, if (i, j)-th subject fails at t_i $(j \in D_i)$ and 0 otherwise.

Note that a discrete baseline cumulative hazards function $H_i = \sum_{k \leq i} \Delta H_k$ is a sum of its jumps ΔH_k at times t_k.

The log-likelihood function of a discrete (grouped) model can be written as

$$\ell = \sum_i \sum_{j \in C_i} \log[G_{ij}(H_i)] + \sum_i \sum_{j \in D_i} \log[G_{ij}(H_{i-1}) - G_{ij}(H_i)] \qquad \text{(III.6)}$$

where $G_{ij}(H_{i-1}) = G(t_{i-1}|z_{ij})$. Here it is understood that $G(t|\cdot)$ depends on t only through a baseline cumulative hazard function $H(t)$, a typical assumption in nonlinear transformation models. Also z_{ij} means a covariate vector for the jth subject whose event (censoring or failure) is associated with t_i, where j is in C_i if $c_{ij} = 0$ or in D_i if $c_{ij} = 1$.

The key distinction from a continuous likelihood is that here the contribution of failures $G(H_{i-1}) - G(H_i)$ cannot be approximated by the first term of Taylor series, $G(H_i)\lambda(H_i)\Delta H_i$, since the residual term $o(\Delta H_i)$ of the series does not become small asymptotically because ΔH_i is fixed. Note that availability of Nelson-Aalen-Breslow type estimators and associated computationally efficient processing of high-dimensional nuisance function H with continuous models is contingent upon the "linearity" of the failure contribution in ΔH (*Tsodikov* (2003a)). We note that this "linearity" is preserved by the PO model. In other words, the form of the failure contribution with PO model is invariant to whether the model is discrete or continuous so that in both cases we have the contribution as $\varphi(H)\Delta H$ for different but similar smooth functions φ of H. Indeed, from (III.5), with the survivor function

$$G_{ij} = \frac{\theta_{ij}}{\theta_{ij} + H_i}$$

we have

$$G_{ij}(H_{i-1}) - G_{ij}(H_i) = \frac{\theta_{ij}}{(\theta_{ij} + H_{i-1})(\theta_{ij} + H_i)} \Delta H_i. \tag{III.7}$$

Because of this property there is little difference in estimation methods between discrete and continuous versions of the PO model.

Finally, the log-likelihood function of discrete PO model takes the form

$$\ell(\boldsymbol{\beta}) = \sum_i \left\{ \sum_{j \in C_i \cup D_i} \log \frac{\theta_{ij}}{\theta_{ij} + H_i} + \sum_{j \in D_i} \log \frac{1}{\theta_{ij} + H_{i-1}} + \sum_{j \in D_i} \log \Delta H_i \right\} \tag{III.8}$$

3.3 Methods

3.3.1 Difference Convex Algorithm (DCA)

Note that the first two terms of (III.8) are convex functions (denote by $B(h)$) of the vector

$$h = (\Delta H_1, \dots, \Delta H_K)^{\mathrm{T}}$$

while the last term is a concave function (denote by $A(h)$). The log-likelihood function (III.8) is therefore the difference of two concave functions, $\ell(\boldsymbol{x}) = B(\boldsymbol{x}) - A(\boldsymbol{x})$. The iterative maximization procedure,

$$\nabla B \left(\boldsymbol{h}^{(m+1)} \right) = \nabla A \left(\boldsymbol{h}^{(m)} \right), \tag{III.9}$$

where m counts iterations and $\nabla A(\boldsymbol{h}) = \partial A / \partial \boldsymbol{h}$ is the gradient of A, represents an MM algorithm, as follows from convexity arguments (*de Leeuw* (1994); *An and Pham* (1997); *Tsodikov* (2003a)). The surrogate objective function for the above construction has the form

$$\mathcal{Q} \left(\boldsymbol{h} \,|\, \boldsymbol{h}^{(m)} \right) = B \left(\boldsymbol{h}^{(m)} \right) - A(\boldsymbol{h}) + \nabla^{\mathrm{T}} A \left(\boldsymbol{h}^{(m)} \right) \left(\boldsymbol{h} - \boldsymbol{h}^{(m)} \right). \tag{III.10}$$

Specifically, we have the following iterations

$$\Delta H_k^{(m+1)} = \frac{d_k}{\sum\limits_{i \geq k} \sum\limits_{j \in C_i \cup D_i} \frac{1}{\theta_{ij} + H_i^{(m)}} + \sum\limits_{i > k} \sum\limits_{j \in D_i} \frac{1}{\theta_{ij} + H_{i-1}^{(m)}}} \qquad \text{(III.11)}$$

for $k = 1, \ldots, K$, where d_k is the number of failures associated with t_k. Note that with H_{i-1} substituted by H_i we would have (III.11) become the algorithm proposed by *Tsodikov* (2003a) for the continuous PO model. As an MM algorithm each iteration of (III.11) will improve the likelihood.

Estimation of regression coefficients β jointly with H can be accomplished in a variety of ways. Here we use a stable if not the fastest Gauss-Seidel type two-stage procedure.

DCA for PO model

0: Initialize the baseline $\boldsymbol{h}$ vector and β.

1: Exercise (III.11) until convergence with the fixed β.

2: Maximize the likelihood with respect to β with fixed h as found at the previous step by a general numerical maximization algorithm (Conjugate Gradients).

3: Check convergence. If not satisfied return to Step 1 with the β as found at Step 2.

3.3.2 Recursive procedure

The algorithm of this section is based on the idea of relieving the model of normalization restriction and subsequently enforcing it through the method of Lagrange multipliers. We take the normalization restriction on the baseline hazard function in the form of $H(0) = 0$.

In order to keep the likelihood maximization problem from being ill-defined, another restriction must be placed on the model. Let us pretend that the cumulative hazard function is known $H_K = x$.

Consider the score equation

$$\frac{d_k}{\Delta H_k} - \sum_{i \geq k} \left\{ \sum_{j \in C_i \cup D_i} \frac{1}{\theta_{ij} + H_i} + \sum_{j \in D_i} \frac{1}{\theta_{ij} + H_{i-1}} \right\} = 0, \qquad \text{(III.12)}$$

for $k = 1, \ldots, K$. Note that the left part of (III.12) is a function of

$$H_k, H_{k+1}, \ldots, H_K = x.$$

Starting with $k = K - 1$ we can solve the equation for H_{K-1} obtaining it as a function of $H_K = x$, say $\varphi_{K-1}(x)$. Next, we can take the equation at $k = K - 2$ and solve it for H_{K-2} obtaining it as a function of $H_{K-1} = \varphi_{K-1}(x)$ and $H_K = x$, and in the end also as a function of x. Repeating the process until $k = 0$ we get the recursively obtained equation

$$H_0 = \varphi_0(x) = 0. \qquad \text{(III.13)}$$

Solving this algebraic equation with respect to x gives the solution x^*.

Having obtained x^* we run the recursion again with $x = x^*$ for $k = K, K-1, \ldots, 1$ to get the full vector h.

This recursive procedure replaces (III.11), and the rest of the algorithm is similar to DCA.

It can be shown that the above algorithm is implementing the method of Lagrange multipliers for maximization of the likelihood under the restriction $\Delta H_0 = 0$.

Note that neither of the proposed two procedures involve high-dimensional generic maximization with respect to h, that ensures their computational efficiency.

3.3.3 Conjugate Gradient method for PO model

In order to compare the proposed two algorithms with the traditional approach, a Quasi-Newton method is used to estimate the baseline hazards in discrete PO model.

To make this procedure comparable with the other two methods examined in this study, we use two-stage maximization procedure and iterate between β and $\mathbf{\Delta H}$. We first estimate $\mathbf{\Delta H}$, given $\beta = \beta$, using a Quasi-Newton method with box constraints. Then, we estimate β, given h found at the previous step, and repeat the cycle until convergence is satisfied.

3.4 Real Data Example

As an example, we apply these three methods to fit a discrete PO model to prostate cancer data from the National Cancer Institute's Surveillance epidemiology and end results (SEER) programme. The dataset is similar to the one used in *Tsodikov* (2003a) except that we used full data with 3 stages in the present paper.

In this data set, 11621 cases of primary prostate cancer diagnosed in the state of Utah between 1988 and 1999 were identified. The following selection criteria were applied to the original 19819 Utah cases registered in the database: valid positive survival time, valid stage of the disease and age 18 years or more. Prostate cancer specific survival was analyzed by the stage of the disease (localized, regional and distant).

Survival time is measured in months resulting in all observations grouped into 143 time intervals with ties reaching a few hundred observations for some intervals.

All methods agree up to 3rd digit in the point estimates given in the following table.

We note a better fit of the PO model to this dataset that shows survival in a distant stage slightly attenuated with time. We will use this model fit to furnish a

Table 3.1: Point estimates and standard errors (in brackets) resulting from fitting discrete PO model to Surveillance, Epidemiology and End Results (SEER) prostate cancer survival data

Point Estimates	DCA	Recursive	Quasi-Newton
Regional	-0.5817(0.0925)	-0.5817(0.0925)	-0.5814(0.0925)
Distant	-3.4328(0.0895)	-3.4328(0.0895)	-3.4327(0.0895)

realistic simulation example.

3.5 Simulation studies

3.5.1 Simulation setting

The purpose of the simulation studies is to examine the performance of three methods in fitting a discrete PO model: DCA, recursive solution and Quasi-Newton method. For Quasi-Newton method, we use R function *optim()* with finite-difference approximation to the gradient vector, which makes it comparable to the other two methods since neither of them requires the specification of gradient vector in the optimization procedure.

We generate data sets of size 11621 and use the parameter estimates from the prostate cancer data as the true parameters. Average censoring proportions of about 50% and 80% are examined. The results are based on 500 simulation replicates. The tolerance is set to be 1e-4.

3.5.2 Simulation results

Simulation results are presented in the following tables. Parameter estimates are shown in Tables 3.2 and 3.3. Tables 3.5 and 3.4 summarizes the time used by recursive method and Quasi-Newton method relative to that used by DCA. In the simualtion study for time, we examined three different accuracy levels as distances of parameter estimates from the true parameters(estimated from DCA with tolerance 1e-10).

Table 3.2: Simulation results with average censoring proportion around 80%: discrete PO model fit using three methods: Recursive procedure, DCA algorithm and Quasi-Newton method

Stage	True parameters	DCA	Recursive	Quasi-Newton
β_1	-0.5817	-0.5814(0.0694)	-0.5814(0.0694)	-0.5791(0.0695)
β_2	-3.4329	-3.4366(0.0719)	-3.4366(0.0719)	-3.4331(0.0724)

Table 3.3: Simulation results with average censoring proportion around 50%: discrete PO model fit using three methods: Recursive procedure, DCA algorithm and Quasi-Newton method

Stage	True parameters	DCA	Recursive	Quasi-Newton
β_1	-0.5817	-0.5798(0.0671)	-0.5797(0.0671)	-0.5667(0.0668)
β_1	-3.4329	-3.4330(0.0616)	-3.4330(0.0616)	-3.4204(0.0611)

Table 3.4: Time used by recursive and Quasi-Newton methods relative to that used by DCA - discrete PO model, censoring proportion = 80%

Distance from true parameters	Recursive	Quasi-Newton
0.1	2.9	7
0.01	2.9	9.2
0.001	3.9	12

Table 3.5: Time used by recursive and Quasi-Newton methods relative to that used by DCA - discrete PO model, censoring proportion = 50%

Distance from true parameters	Recursive	Quasi-Newton
0.1	3	6.3
0.01	3.4	5.8
0.001	2.8	6

We observed that both DCA and the recursive procedure outperformed the traditional Quasi-Newton approach in terms of accuracy and speed. The Quasi-Newton methods was an order of magnitude slower as measured by the relative computation time in Tables 3.4 and 3.5 and less accurate at that as given by the comparison of the estimates in Tables 3.2 and 3.3. All methods showed decent stability to censoring perhaps a result of ample sample size available.

3.6 Discussion and Conclusion

With this study we have revisited the point estimation algorithms for the discrete proportional odds model. We found the PO model naturally suited for the discrete or grouped data setting because the form of its likelihood function with right censored data is invariant with respect to the type of data (continuous or grouped) in the sense that the contribution of failures is still proportional to the jump of the baseline cumulative hazard functions. This property ensured that the artificial mixture and MM approaches carry over with slight modification from the continuous setting and bring their computational efficiency with them. This stands in stark contrast with the Cox model that has seen a number of challenges with discrete data.

A novel recursive approach has been developed as an alternative method. Although a little slower than the MM-type algorithms it has potentially superior accuracy and stability as it is based on recursive solution to algebraic equations that can be accomplished by stable bi-section algorithms if needed.

The research targeted the most difficult niche of data applications when the dimensionality was high despite the model still being discrete. Cancer registry such as SEER gives such an example. The fact that traditional maximization methods fail to exploit the specific likelihood structure and are subject to the curse of dimensionality makes them slow. In particular, Newton type methods have $O(n^3)$ complexity due to a direct or indirect (iterative) information matrix inverse. At the same time MM-type methods show approximately linear complexity curve with increase in the dimension of h.

There is probably little difference between these methods if the dimensionality of the problem is low.

The results of this study hold much promise for the development of efficient computational methods for the broad class of transformation models. The artificial mixture approach (*Tsodikov* (2003a)) allowed to extend the numerical efficiency of the Cox model estimates to transformation models. The Cox model has shown itself poorly in the discrete setting. The PO model, on the contrary, takes to the discrete setting naturally because of its ordinal heritage. It appears therefore that we might be able to develop computationally efficient procedures for a broad class of models that can be represented as an (artificial) mixture of PO models just like continuous transformation models are usually mixtures of PH models (univariate frailty models). With this study we have shown that efficient methods exist for the PO model and it is therefore a good candidate as the base model spawning a mixture family for the discrete and grouped data setting.

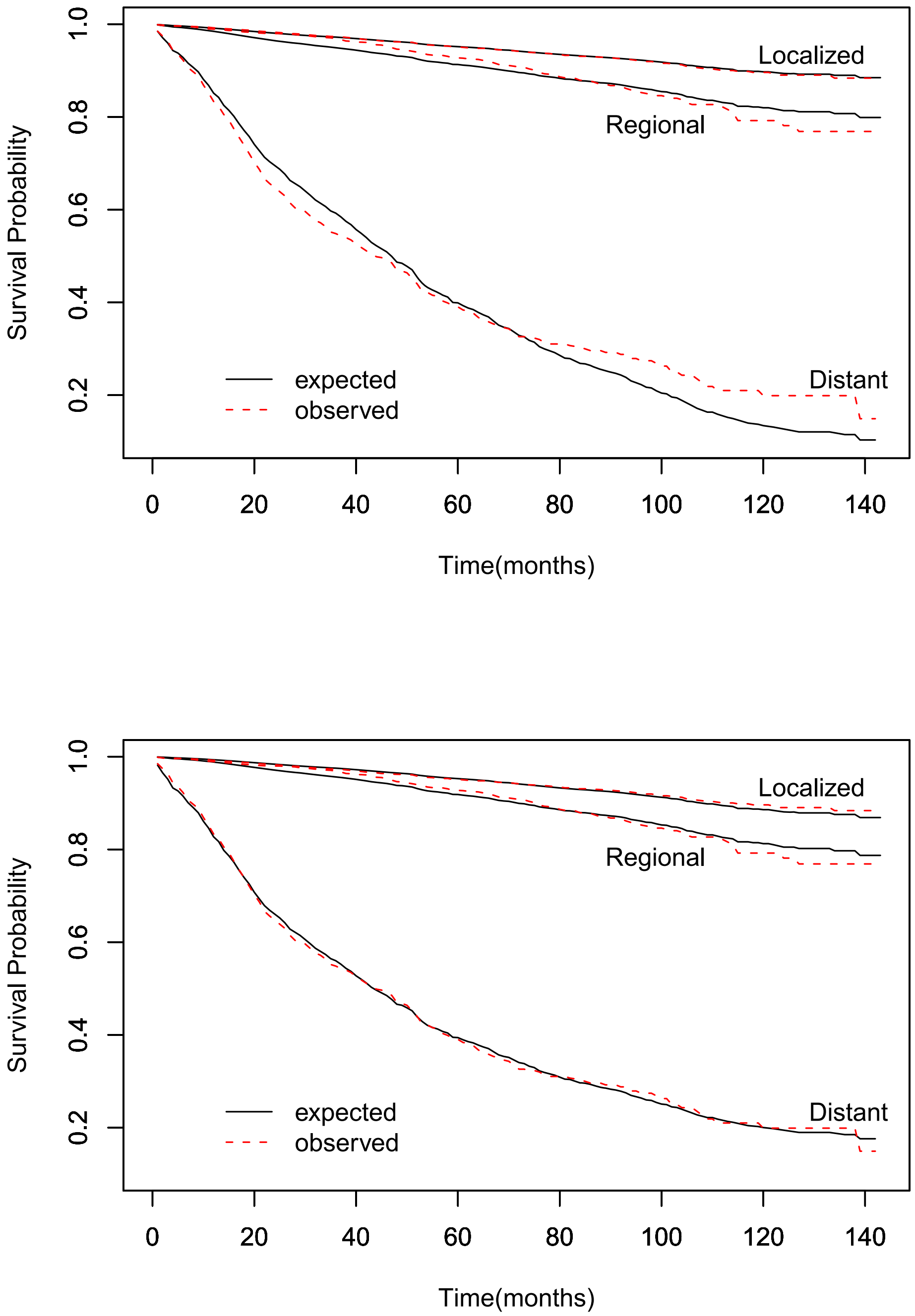

Figure 3.1: Discrete Cox model (top) and PO model (bottom) fit and observed Kaplan-Meier curves; Surveillance, Epidemiology and End Results (SEER) prostate cancer survival data by stage of the disease at diagnosis.

CHAPTER IV

Aritificial Mixture Methods for Discrete Failure Time

Key Words: QEM algorithm, discrete failure time data, artificial mixture methods, PH mixture method, PO mixture method

4.1 Introduction

Most research on analyzing failure time data considers time as a continuous measure, a basic underlying assumption of which is that failure times are untied. As a result, the exact method of treating ties has to consider all possible orderings of tied events (See *Kalbfleisch and Prentice* (1973) and *Kalbfleisch and Prentice* (2002)). In practice, failure time is always measured in a certain time unit, such as month or year, so that ties can occur. A moderate number of ties can be handled appropriately by slight modifications to the exact likelihood such as *Breslow* (1974) and *Efron* (1977). If time is truly discrete, or if there exist many ties, then treating such failure time as discrete is more appropriate. Analyzing such event times as if they were exact may introduce bias and hence leads to incorrect inferences. Sometimes the original failure time data are grouped to a larger time unit for various purposes such as simplification of computation in interim analysis. Under such circumstance, treating failure time as discrete is also more consistent with the data. As a consequence, discrete failure time models have gained more attention lately.

Researchers explored methods to deal with many ties and/or treat failure time as discrete, such as *Prentice and Gloeckler* (1978), *Stewart and Pierce* (1982), *Johnson and Christensen* (1986), *Sinha et al.* (1994), *Yu et al.* (2004), *Pipper and Ritz* (2006), *Zhao and Zhou* (2008), *Li et al.* (2008), *Yu et al.* (2009), etc. For example, *Prentice and Gloeckler* (1978) proposed a method to fit a discrete PH model by a Newton-Raphson algorithm applied to maximize the full model likelihood. They reported instability and difficulties in situations when the problem was high-dimensional. *Yu et al.* (2004) studied mixture cure models for grouped failure time data and focused on the estimation of cure fraction. *Prentice and Kalbfleisch* (2003) proposed a mixed discrete and continuous Cox model that enjoys computational efficiency but does not apply to the grouped data likelihood. Besides it requires a restriction on the cumulative hazard that has to be observed in order to make the model probabilistically consistent. *Zhao and Zhou* (2008) studied a discrete PH cure model as an extension of the latter framework.

Most of the methods mentioned above are case by case studies focused on a specific model. With this study we propose a new approach that serves a general class of discrete transformation models.

Depending on the assumptions made for baseline hazards, they can either be treated as nuisance parameters and hence the regression parameters can be estimated semi-parametrically, or have to be estimated jointly with regression parameters. Since there does not exist a criterion to select the baseline hazard functions, we apply a more general and less restrictive representation of the baseline hazards and assume them to be piecewise constant. This inevitably gives rise to a high-dimensional optimization problem.

We propose two artificial mixture methods for a general discrete transformation model based on the idea of artificial random variables and the QEM algorithm (*Tsodikov* (2003a)): PH mixture method and PO mixture method. Applications of both methods are not restricted to Cox PH model or Cox-type models. Most importantly, both methods simplify the optimization procedure. Due to its simplicity, we advocate PO mixture method as a result of this study. We incorporate the PO mixture framework with the recursive solution and DCA for discrete PO model, proposed in Chapter III. Those two procedures are applied after missing data are imputed. PO mixture method with either recursive solution or DCA is superior to Quasi-Newton method, in terms of accuracy and speed for parameter estimation in such a high-dimensional parameter space.

We introduce the idea of artificial mixtures and the survival model in 4.2 and 4.3. We then present PH and PO artificial mixture methods in Section 4.4 and 4.5 respectively. In Section 4.6, we compare our advocated PO mixture method with Quasi-Newton method in two simulation studies, followed by discussion in Section 4.7.

4.2 Artificial Mixtures

Let $p(x\,|\,z)$ be a family of probability distributions describing a model for the random response X regressed on covariates z. The idea of an artificial mixture is to represent $p(x\,|\,z)$ as a marginal probability (a mixture model)

$$p(x\,|\,z) = \mathrm{E}\left\{p_0(x\,|\,z, U)|z\right\}, \tag{IV.1}$$

where U is a mixing variable, possibly a vector, representing artificial missing data, and $p_0(\cdot|\cdot, U)$ are some probabilities conditional on U that define the latent model. The expectation is taken conditional on z implying that U is generally itself a regression on z. In other words, an artificial mixture model is considered such that one gets the original target model when missing data U are integrated out. Representation (IV.1) can be considered as a form of an integral transform of r.v. U. The potential utility of (IV.1) exploited in this study lies in the simplicity of the latent model p_0 contrasted with the complexity of the original marginal model p. This paradigm invites an application of an EM algorithm to fit the marginal model. Imputation of U and/or some functions of U, given observed data, the E-Step, and fitting the latent model by maximum likelihood, the M-Step, are the two steps of the iterative procedure.

A key to simplifying the E-Step comes through recognition that imputed U is some kind of a conditional moment. Moments of random variables can be found by differentiating a transform (ex. a Laplace transform, L). The form of the transform is readily available from (IV.1), and this leads to M-step expressed through L and its derivatives. The M-Step is simplified by the design of the latent model. Partial identifiability of univariate frailty models makes this design flexible. In the semi-parametric framework infinitely many latent models (p_0, U) satisfying (IV.1) exist for any fixed marginal model p, and the problem is to find one associated with some computational advantage. Keeping the development general in terms of L makes the method applicable to a wide variety of transformation models. All of them will enjoy the same computational advantage once a suitable base latent model p_0 is found, specific to the general form of the likelihood under study. Note that the form of the distribution of U (inverse of the transform L) need not be known explicitly, as the algorithm is specified in terms L and its derivatives, not the inverse.

4.3 Marginal survival model and likelihood

Define the survival model through the survival function

$$G(t|z) = L(H(t)|\omega, z), \tag{IV.2}$$

where L is a parametrically specified survival function (with parameters ω), and time argument t has been transformed by an arbitrary nondecreasing discrete cumulative hazard function H.

Define the notation: t_i, $i = 1, 2, \ldots, K$, are distinct time points specifying the support of the discrete survival distribution to be estimated; C_i is the set of subjects who are censored at t_i; D_i is the set of subjects who fail at t_i; R_i is the set of subjects at risk at t_i; j indexes subjects in a given set; $H_i = H(t_i) = \sum_{k \leq i} \Delta H_k$; ΔH_k is the jump of the baseline cumulative hazard at time t_k that induces the discrete mass at t_i. For any function $A(H|z)$ of arbitrary baseline cumulative hazard H and covariates z define $A_{ij}(H) = A(H|z_{ij})$, where z_{ij} is the covariate vector for the jth subject who failed at t_i if $j \in D_i$ or who is censored at t_i if $j \in C_i$. The log-likelihood function of a discrete (grouped) model can be written as

$$\ell = \sum_i \sum_{j \in C_i} \log \left[L_{ij} \left(\sum_{k \leq i} \Delta H_k \right) \right] + \sum_i \sum_{j \in D_i} \log \left[L_{ij} \left(\sum_{k < i} \Delta H_k \right) - L_{ij} \left(\sum_{k \leq i} \Delta H_k \right) \right].$$

Hence, the score equation for H is

$$\frac{\partial \ell}{\partial \Delta H_k} = \sum_{i \geq k} \left\{ \sum_{j \in C_i} \left[\frac{L'_{ij}(H_i)}{L_{ij}(H_i)} \right] + \sum_{j \in D_i} \left[\frac{L'_{ij}(H_{i-1,j}) I(i > k) - L'_{ij}(H_i)}{L_{ij}(H_{i-1}) - L_{ij}(H_i)} \right] \right\} = 0,$$

where $i, k = 1, \ldots, K$, and L' is the derivative of L with respect to its H argument. Solving the above equation system presents a dimensionality challenge for large K.

4.4 PH artificial mixture method

Consider the PH mixture framework defined by the PH model chosen as a latent base model. This corresponds to the the univariate PH frailty form of the (marginal) survivor function with L being the Laplace transform of U. We have (IV.1) turn into

$$G(t|z) = \mathrm{E}\left\{ e^{-UH(t)} \middle| z \right\} = L(H(t)|z), \tag{IV.3}$$

where p_0 is based on the latent survival function $\exp\{-UH(t)\}$, and the artificial random variable U is regressed on covariates z. The complete-data (latent model) log-likelihood function, conditional on subject-specific U_{ij} and z_{ij}, can be written as

$$\ell_0 = \sum_i \sum_{j \in C_i} \log[e^{-U_{ij}H_i}] + \sum_i \sum_{j \in D_i} \log[e^{-U_{ij}H_{i-1}} - e^{-U_{ij}H_i}] + C, \tag{IV.4}$$

where C does not depend on H.

This artificial mixture representation yields computationally efficient algorithms in the continuous case where the contribution of failures is replaced by the first term of the Taylor expansion (*Tsodikov* (2003a))

$$e^{-U_{ij}H_{i-1}} - e^{-U_{ij}H_i} = e^{-U_{ij}H_i} U_{ij} \Delta H_i + o(\Delta H_i), \tag{IV.5}$$

and is therefore worth studying as a potential solution for the discrete one.

4.4.1 Parameter estimation in the PH mixture method

To estimate the hazard H and regression coefficients (hidden in the distribution of U), we may apply a Gauss-Seidel type two-stage procedure.

Step 1: Given regression coefficients β, estimate $\{H_i\}_{i=1}^K$ by an EM algorithm;

Step 2: Estimate β given the hazard obtained from Step 1.

Iterate between these two steps until convergence.

Step 2 is handled by the Newton-Raphson method. Alternatively, Step 1 may be considered as nested in Step 2, the latter defined as maximization of the profile likelihood $\ell(\beta, H(\beta))$ with $H(\beta)$ defined as a solution to Step 1. In Step 1, EM-type algorithm is invoked, treating U_{ij} as missing data. Taking derivative of (IV.4) with respect to ΔH_k, we obtain the k-th score equation

$$\frac{\partial \ell_0}{\partial \Delta H_k} = \sum_{j \in R_i} (-U_{ij}) + \sum_{j \in D_k} \frac{U_{kj}}{1 - e^{-U_{kj}\Delta H_k}} = 0. \tag{IV.6}$$

Once functions of U_{ij} entering the above score equations are imputed (the E-step), they are solved for ΔH_k at the M-step. EM iterations continue until H used in the imputation and H as a solution of (IV.6) become sufficiently close (self-consistency) at which point it is reported as the outcome of Step 1. As a result the high-dimensional problem is reduced to many one-dimensional ones.

Imputation in the score equation (IV.6). The bottleneck here is the E-step that involves the imputation of functional forms of U_{ij}, given the observed data for the subject (i, j): (1) E$[U_{ij}|$censored$]$; (2)E$[U_{ij}|$failure$]$; and (3) E $\left[U_{ij}/\{1 - e^{-U_{ij}\Delta H_i}\}\middle|\text{failure}\right]$. The part “censored”or “failure”in the conditional expectation represents the observed data on the (i, j)th subject. The first two conditional expectations are obtained in terms of derivatives of L similar to *Tsodikov* (2003a):

$$\mathrm{E}[U_{ij}|\text{censored}] = \frac{\mathrm{E}\{U_{ij}e^{-U_{ij}H_i}\}}{\mathrm{E}\{e^{-U_{ij}H_i}\}} = \frac{-L'_{ij}(H_i)}{L_{ij}(H_i)}, \tag{IV.7}$$

$$\mathrm{E}[U_{ij}|\text{failure}] = \frac{\mathrm{E}\{U_{ij}[e^{-U_{ij}H_{i-1}} - e^{-U_{ij}H_i}]\}}{\mathrm{E}\{e^{-U_{ij}H_{i-1}} - e^{-U_{ij}H_i}\}} = \frac{-L'_{ij}(H_{i-1}) + L'_{ij}(H_i)}{L_{ij}(H_{i-1}) - L_{ij}(H_i)}. \tag{IV.8}$$

However, E $\left[\frac{U_{ij}}{1-e^{-U_{ij}\Delta H_i}}\middle|\text{failure}\right]$ does not have a closed-form expression. Expanding the fraction into a power series and imputing each term gives the closed form

$$\frac{1}{1 - e^{-U_{kj}\Delta H_k}} = \sum_{s=0}^{\infty} e^{-sU_{kj}\Delta H_k}, \text{ and,} \tag{IV.9}$$

$$\mathrm{E}\left[U_{kj}\frac{1}{1 - e^{-U_{kj}\Delta H_k}}\middle|\text{failure}\right] = \sum_{s=0}^{\infty} \frac{-L'_{kj}\left(s\Delta H_k + H_{k-1}\right) + L'_{kj}\left(s\Delta H_k + H_k\right)}{L_{kj}\left(H_{k-1}\right) - L_{kj}\left(H_k\right)}.$$

However, a singularity at $\Delta H \to 0$ (small ΔH are expected in high-dimensional discrete problems) makes the convergence of the series slow. Applying numerical integration in (IV.7), (IV.8) is equally attractive slowing down the convergence of the algorithm. To remedy the situation we isolate the singularity rewriting the likelihood as

$$\ell_0 = \sum_i \left\{ \sum_{j \in C_i \cup D_i} (-U_{ij} H_i) + \sum_{j \in D_i} \log \Delta H_i - \sum_{j \in D_i} \log \frac{U_{ij} \Delta H_i}{e^{U_{ij} \Delta H_i} - 1} \right\}, \qquad \text{(IV.10)}$$

and dropping terms independent of H. Note that the term $\log \frac{U_{ij}\Delta H_i}{e^{U_{ij}\Delta H_i} - 1}$ does not have the singularity any more and is zero at $\Delta H_k = 0$. Its expansion around $\Delta H_i = 0$ promises reasonable convergence rates with small ΔH_k. The Power series expansion has coefficients γ_k that are determined explicitly and recurrently based on so-called Bernoulli numbers (*Gradshteyn et al.* (2007)) (see Appendix IV.9)

$$\log \frac{t}{e^t - 1} = \sum_{k=1}^{\infty} \gamma_k t^k, \quad t = U_{ij} \Delta H_i.$$

This method reduces the imputation of $\log \frac{U_{ij}\Delta H_i}{e^{U_{ij}\Delta H_i} - 1}$ to the imputation of powers of U_{ij} by

$$\mathrm{E}[U_{ij}^s | \text{failure}] = (-1)^s \frac{L_{ij}^{(s)}(H_{i-1}) - L_{ij}^{(s)}(H_i)}{L_{ij}(H_{i-1}) - L_{ij}(H_i)},$$

where $L_{ij}^{(s)}(x) = \frac{d^s}{dx^s} L_{ij}(x)$.

4.5 PO mixture method

Now consider the Proportional Odds (PO) model as the basis for the latent model p_0 in the general artificial mixture formulation (IV.1) resulting in the PO mixture framework. The choice here is motivated by the widespread use of the PO model with categorical data. We find that in this case the imputation has closed form and does not require costly series approximations. Also, the procedure can be specified in terms of the Laplace transform L, not requiring any derivatives. Last but not least, the M-step enjoys a recurrent structure of the score equation for H that ensures its computational efficiency. All these facts make the PO mixture method the preferred one (see below).

4.5.1 Discrete PO model

In this section we consider an algorithm to fit the base PO model as it is used at the M-step. The survival function of the PO model takes the form

$$G = L(H(t)|z) = \frac{\theta(z)}{\theta(z) + H(t)}, \tag{IV.11}$$

where θ is the odds ratio of survival relative to the baseline survival function characterized by $\theta = 1$. Here L is a Laplace transform of an exponential distribution with rate θ. The likelihood for a discrete PO model is

$$\ell(\boldsymbol{\beta}) = \sum_i \sum_{j \in C_i} \log \frac{\theta_{ij}}{\theta_{ij} + H_i} + \sum_i \sum_{j \in D_i} \log \left[\frac{\theta_{ij}}{\theta_{ij} + H_{i-1}} - \frac{\theta_{ij}}{\theta_{ij} + H_i} \right]. \tag{IV.12}$$

It can be rewritten as follows:

$$\ell(\beta) = \sum_i \sum_{j \in D_i} \log \Delta H_i + \sum_i \sum_{j \in C_i \cup D_i} \log \frac{\theta_{ij}}{\theta_{ij} + H_i} + \sum_i \sum_{j \in D_i} \log \frac{1}{\theta_{ij} + H_{i-1}} \tag{IV.13}$$

4.5.1.1 Parameter estimation

Two alternative procedures to fit the PO model are used following *Tsodikov and Wang* (2011). The first procedure is a recurrent solution to the score equation derived through Lagrange multipliers. The score equations with respect to ΔH_k can be written as

$$\sum_{i\geq k}\sum_{j\in C_i\cup D_i}\frac{-1}{\theta_{ij}+H_i}+\frac{d_k}{H_i-H_{i-1}}+\sum_{i>k}\sum_{j\in D_i}\frac{-1}{\theta_{ij}+H_{i-1}}=0,\quad k=1,\ldots,K, \quad \text{(IV.14)}$$

where d_k is the multiplicity of failures at t_k. This system of equations has recurrent structure in that H_k can be found when H_i, $i=k+1,\ldots,K$ are known. We can initially consider H_0 as unrestricted and derive it as a function of H_K where $H_1,\ldots,H_{K-1}$ are derived by solving $(IV.14)$ for $k=K-1,K-2,\ldots,1$. With $k=K$, (IV.14) involves H_K and H_{K-1} and is solved for H_{K-1}, given H_K, resulting in H_{K-1} being a function of H_K. Then H_{K-2} is obtained as a function of H_K from (IV.14) at $k=K-2$, and the previous solution. This continues until $k=1$ when we solve for H_0 as a function of H_K. The solution to this system of equations emerges when the equation $H_0(H_K)=0$ is solved enforcing the restriction. The procedure can be interpreted as a method of Lagrange multipliers for maximizing the likelihood over H under the restriction $H_0=0$.

The second procedure is an MM algorithm *Lange et al.* (2000). Consider a Nelson-Aalen-type estimator

$$\Delta H_k=\frac{d_k}{\sum\limits_{i\geq k}\sum\limits_{j\in C_i\cup D_i}\frac{1}{\theta_{ij}+H_i}+\sum\limits_{i>k}\sum\limits_{j\in D_i}\frac{1}{\theta_{ij}+H_{i-1}}} \quad \text{(IV.15)}$$

that is a consequence of (IV.14). The MM algorithm treats the right part of (IV.15) as based on the previous iteration copy of H and updates it getting the next iteration copy in the left part of (IV.15). Iterations proceed until the left and the right part of (IV.15) are self-consistent in the sense that they are based on the same H^*, the fixed-point of (IV.15). By the MM theory and the property that the likelihood (III.8) can be represented as a difference between two concave functions, each such iteration will improve the likelihood.

4.5.2 PO mixture model setting

The PO artificial mixture family emerges as we expand the PO model (IV.11) by randomizing its predictor θ similar to the PH mixture idea (IV.1), (IV.3). Substitute θ in (IV.11) by an artificial random variable $U \sim \Pr(u|z)$, regressed on covariates z, such that we can view the new model survivor function as

$$G(t|z) = L(H(t)|z) = \mathrm{E}\left\{\frac{U}{U + H(t)}\right\}, \tag{IV.16}$$

where H is the arbitrary baseline cumulative hazards (cumulative odds, to be precise) function. The function L is still a Laplace transform of random variable $V = W/U$, where W has a unit exponential distribution given U. Indeed,

$$\mathrm{E}\left\{e^{-\frac{W}{U}H}\right\} = \mathrm{E}\left\{\mathrm{E}\left[e^{-\frac{W}{U}H}\middle| U\right]\right\} = \mathrm{E}\left[\frac{U}{U + H}\right], \tag{IV.17}$$

same as (IV.16). This means that (IV.16) could be written in the form (IV.3) with the frailty variable V instead of U, and the PH mixture method of Section 4.4 is still applicable. However, here we base our algorithm on the artificial mixture form (IV.16) implying the PO base model at the M-Step.

4.5.3 Parameter estimation in the PO mixture method

4.5.3.1 Estimating hazards

With the latent survivor function at t_i for the subject (i, j) $L_{ij}(H_i|U_{ij}) = \frac{U_{ij}}{U_{ij}+H_i}$, the complete data log-likelihood function (omitting terms that do not depend on H) takes the form (IV.13) with U_{ij} instead of θ_{ij}. The score equations with respect to ΔH_k, for $k = 1, ..., K$, become

$$\frac{\partial \ell}{\partial \Delta H_k} = \frac{d_k}{\Delta H_k} + \sum_{i \geq k} \sum_{j \in C_i \cup D_i} \frac{-1}{U_{ij} + H_i} + \sum_{i > k} \sum_{j \in D_i} \frac{-1}{U_{ij} + H_{i-1}} = 0 \tag{IV.18}$$

The E-step involves the imputation of the following functional forms of U_{ij}. 1) $\mathrm{E}\left\{\frac{1}{U_{ij}+H_i}|\text{censored}\right\}$; 2) $\mathrm{E}\left\{\frac{1}{U_{ij}+H_i}|\text{failure}\right\}$; and 3) $\mathrm{E}\left\{\frac{1}{U_{ij}+H_{i-1}}|\text{failure}\right\}$. "censored"or "failure"in the conditional expectation represents the observed data on the (i,j)th subject. We have an imputation for censored observation as

$$\mathrm{E}\left\{\frac{1}{U_{ij}+H_i}\middle|\,\text{censored}\right\} = \frac{\mathrm{E}\left\{\frac{1}{U_{ij}+H_i}\frac{1}{U_{ij}+\tilde{H}_i}\right\}}{\mathrm{E}\left\{\frac{1}{U_{ij}+\tilde{H}_i}\right\}} = \frac{L(H_i|z_{ij}) - L(\tilde{H}_i|z_{ij})}{(\tilde{H}_i - H_i)L(\tilde{H}_i|z_{ij})}, \qquad \text{(IV.19)}$$

where tilde is used to mark a copy of H used in the missing data distribution for the imputation.

Similarly for the failure we obtain

$$\mathrm{E}\left\{\frac{1}{U_{ij}+H_m}\middle|\,\text{failure}\right\} = \frac{\mathrm{E}\left\{\frac{1}{U_{ij}+H_m}\left[\frac{U_{ij}}{U_{ij}+\tilde{H}_{i-1}} - \frac{U_{ij}}{U_{ij}+\tilde{H}_i}\right\}\right]}{\mathrm{E}\left\{\frac{U_{ij}}{U_{ij}+\tilde{H}_{i-1}} - \frac{U_{ij}}{U_{ij}+\tilde{H}_i}\right\}} =$$

$$\frac{\frac{L(H_m|z_{ij})-L(\tilde{H}_{i-1}|z_{ij})}{(\tilde{H}_{i-1}-H_m)L(\tilde{H}_{i-1}|z_{ij})} - \frac{L(H_m|z_{ij})-L(\tilde{H}_i|z_{ij})}{(\tilde{H}_i-H_m)L(\tilde{H}_i|z_{ij})}}{L(\tilde{H}_{i-1}|z_{ij}) - L(\tilde{H}_i|z_{ij})}, \qquad \text{(IV.20)}$$

$m = i-1$ or $m = i$, dependent on whether second or third term of (IV.18) is being imputed. Note that all the imputations in the PO mixture method have closed-form expressions. Now, the imputed form of the score equation (IV.18) becomes

$$\frac{d_k}{\Delta H_k} - \sum_{i\geq k}\sum_{j\in C_i}\frac{L(H_i|z_{ij}) - L(\tilde{H}_i|z_{ij})}{(\tilde{H}_i - H_i)L(\tilde{H}_i|z_{ij})} - \sum_{j\in D_i}\frac{\frac{L(H_i|z_{ij})-L(\tilde{H}_{i-1}|z_{ij})}{(\tilde{H}_{i-1}-H_i)L(\tilde{H}_{i-1}|z_{ij})} - \frac{L(H_i|z_{ij})-L(\tilde{H}_i|z_{ij})}{(\tilde{H}_i-H_i)L(\tilde{H}_i|z_{ij})}}{L(\tilde{H}_{i-1}|z_{ij}) - L(\tilde{H}_i|z_{ij})} -$$

$$\sum_{i>k}\sum_{j\in D_i}\frac{\frac{L(H_{i-1}|z_{ij})-L(\tilde{H}_{i-1}|z_{ij})}{(\tilde{H}_{i-1}-H_{i-1})L(\tilde{H}_{i-1}|z_{ij})} - \frac{L(H_{i-1}|z_{ij})-L(\tilde{H}_i|z_{ij})}{(\tilde{H}_i-H_{i-1})L(\tilde{H}_i|z_{ij})}}{L(\tilde{H}_{i-1}|z_{ij}) - L(\tilde{H}_i|z_{ij})} = 0 \qquad \text{(IV.21)}$$

Note that the imputed form of the score equation is not the same as in the PO model due to the non-linear (in U) form of the terms that were imputed. Nevertheless, solution to this equation with respect to H given β and $\tilde{H}$ is similar to that of the PO model (Section 4.5.1).

First, in terms of H, the imputed score equations (IV.21) have the form

$$\varphi_i(H_i, H_{i-1}) = 0, i = 1, \ldots, K, \tag{IV.22}$$

and the recursive procedure of Section 4.5.1.1 will work: Set H_K aside, solve (IV.22) sequentially for H_{K-1} with $i = K-1$, then for H_{K-2} with $i = K-2$ using $H_{K-1}(H_K)$ from the previous solution, etc. until with $i = 1$ the equation $H_0(H_K) = 0$ is obtained. Solving it for H_K and reconstructing all $H_i(H_K)$ gives the final solution.

Alternatively, the MM iterative procedure may be employed. As discussed in Section 4.5.1, the score equation for the PO model (similar to (IV.18) prior to imputation) has a representation as a difference between derivatives of two convex functions $A'(H) - B'(H) = 0$, where H is understood as a vector of ΔH_i, $i = 1, \ldots, K$. Difference Convex Algorithm (DCA) (a version of the MM algorithm) finds the next iteration H^{m+1} as the solution to the equation $A'(H^{m+1}) = B'(H^m)$, where m counts iterations (this is what (IV.15)-based algorithm is). We note that the imputation operator, a conditional expectation is a linear one and does not alter convexity properties. Hence, iterations based on (IV.21) written in the (IV.15) form

$$\Delta H_k^{m+1} = \frac{d_k}{B'(H^m)} \tag{IV.23}$$

will also constitute an MM algorithm monotonically improving the interim M-Step likelihood and converging to the fixed point of the imputed score equation.

Once the EM algorithm does its job of finding H given β with either recursive or DCA implementation of the M-Step, regression coefficients need to be estimated. This can be done in two ways: either using Gauss-Seidel iterations alternating between maximization over H given β and over β given H, or using the profile likelihood approach of obtaining the $H(\beta)$ solution by the EM algorithm and plugging it into the likelihood, then maximizing the profile likelihood over β.

Newton-Raphson method or conjugate gradient search is used to maximize over β in any context.

4.5.4 Applications

We provide two applications as an example, both derived from the Utah Cancer Registry (UCR) survival data, which is part of the SEER database (http://www.seer.cancer.gov/). One is a prostate cancer dataset with stage (1=localized, 2=regional and 3=distant), and another a breast cancer dataset of patients in localized stage with age group as a covariate. The data are described and analyzed using continuous models in *Tsodikov* (2003a) and *Tsodikov* (2002), respectively. In this study we recognize that the data are grouped (coarsened) because survival time is measured in months and apply discrete models to study the problem.

4.5.4.1 Application 1: fit a PH model using the PO mixture method

Various challenges were reported fitting a discrete proportional hazards (PH) model. Here we use the stable algorithms of this chapter to fit the PH model by representing it as a mixed PO model. In the sense of (IV.17) this artificial mixture emerges when U is an exponential distribution coupled with unit-exponential W so that the ratio $W/U = \theta(z)$ is non-random with $G(t|z) = L(H(t)|z) = e^{-\theta(z)H(t)}$. Two-stage iterative process is used to estimate the baseline hazards and the regression coefficients. Given estimates of regression coefficients at a previous iteration, we apply the EM algorithm, with closed form expressions for the imputation in E-step.

Results from fitting discrete PH model using our PO mixture method are listed in the second column in Table 4.1. Numerical derivatives are used to obtain the standard error estimates. Results from fitting PH model using the method described in *Tsodikov* (2003a) and treating the failure time data as continuous are given in the third column for reference.

Table 4.1: PH model fit to UCR prostate cancer data. Standard error are shown in parentheses.

Estimates (se)	discrete model using PO mixture method	continuous model
stage 2	1.23(0.04)	1.22(0.04)
stage 3	2.67(0.06)	2.66(0.06)

We are not surprised to see that the results from treating failure time as discrete and continuous do not differ much, since the jumps of the cumulative hazard are small. In this situation, continuous model works well as an approximation to discrete model. However, this situation presents a challenge for discrete models because of high dimensionality, and it makes sense to use as an example. Besides, a continuous model is still only an approximation in this case providing a minor error in the second digit.

4.5.4.2 Application 2: PH-PH model fit using the PO mixture method

In PH-PH model *Tsodikov* (2002), the survivor function has the form:

$$G(\cdot|z) = \exp\{-\theta(z)[1 - F^{\eta(z)}]\},$$

where F is an arbitrary survival function. This cure model is often used to reproduce dissimilar long-term and short-term effects on survival, where $\theta(z)$ models the long-term effect, while $\eta(z)$ models the short-term effect. The PH-PH name comes from the fact that PH model is used twice as a composition, once in a cure form $G = \exp\{-\theta(z)[1 - F]\}$, and then in a non-cure form $F^{\eta(z)}$ to model departures from the proportional hazards assumption.

The same two-stage estimation procedure is used to estimate the baseline hazards and the regression coefficients. Given regression coefficients, we estimate the baseline hazards using the EM algorithm based on artificially random variables U. In this example $L(H|z) = \exp\{-\theta(z)[1-\exp(-\eta(z)H)]\}$. A cure model implies the restriction

$H_K = \infty$ to avoid unidentifiability issue. Therefore the recursive procedure starts at $k = K - 1$ instead of K. Also, no imputations are done at or after K. We use the UCR breast cancer data to illustrate our method. We restrict our analysis to patients with localized disease and include age as a covariate. Long-term effect are removed since *Tsodikov* (2002) showed that they are not significant.

Results using our PO mixture method are listed in the second column in Table 4.2, together with results from the continuous model using the method described in *Tsodikov* (2003a) in the third column for reference. Numerical derivatives of the likelihood are used for variance estimation.

Table 4.2: PH-PH model fit to UCR breast cancer data with patients at localized stage only and age as the only covariate. Standard errors are shown in parentheses.

Estimates (se)	PO mixture method	continuous model
age 46-55(ST)	-0.22(0.15)	-0.37(0.14)
age 56-65(ST)	-0.49(0.14)	-0.65(0.13)
age $\geq$ 66(ST)	-0.54(0.13)	-0.68(0.12)
long-term cure rate	-1.05(0.06)	-1.10(0.05)

Survival curves for all four age groups are shown in Figure 1. They provide a good fit to the observed survival curves (Kaplan-Meier).

4.6 Simulation studies

As discussed earlier, the PO mixture method has computation advantages over the PH mixture method. As a result, we recommend the PO mixture method. In this section, we examine our advocated PO mixture method in two simulation studies. We consider two scenarios where the censoring proportion is about 80% and 50% respectively.

The simulation studies are based on the SEER prostate cancer data. We first fit discrete PH model to the data set and obtained the estimates of the regression

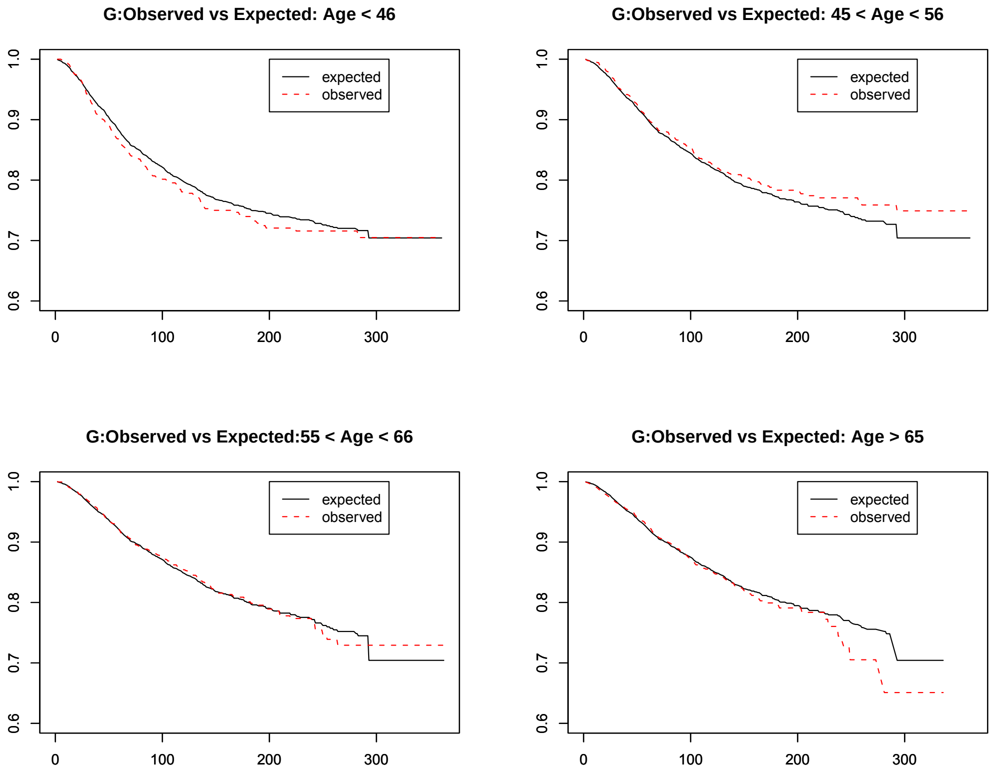

Figure 4.1: Survival curves for UCR breast cancer patients at localized stage in four age groups

parameters and baseline hazards. We then use those estimates as the true values of parameters in the simulation studies. For the simulation study where the average censoring proportion is about 80%, we use Stage in the real data set as the covariate. For the second scenario, we simulate the covariate such that the average censoring proportion is about 50%. In both cases, the censoring is assumed to be exponentially distributed.

For each scenario, we simulate 500 data sets, with each of sample size 11621. We then fit a discrete PH model to those simulated data sets using PO mixture method with recursive solution, PO mixture method with DCA and Quasi-Newton method. The results are shown in the following two tables.

Based on the results from 500 simulations, we can see the PO mixture method

Table 4.3: Simulation studies to compare PO mixture method with Quasi-Newton method - fit PH model with average censoring proportion around 80%

	True parameter	PO Mixture with Recursive Solution	PO Mixture with DCA	Quasi-Newton Method
Regional Vs. localized	0.606	0.607(0.067)	0.607(0.066)	0.601(0.066)
Distant Vs. localized	2.919	2.918(0.049)	2.919(0.049)	2.911(0.049)

Table 4.4: Simulation studies to compare PO mixture method with Quasi-Newton method - fit PH model with average censoring proportion around 50%

	True parameter	PO Mixture with Recursive Solution	PO Mixture with DCA	Quasi-Newton Method
Regional Vs. localized	0.606	0.611(0.066)	0.608(0.066)	0.593 (0.065)
Distant Vs. localized	2.919	2.927(0.056)	2.924(0.056)	2.909 (0.055)

with both recursive solution and DCA is more accurate than Quasi-newton method.

Tables 4.5 and 4.6 list the simulation results on the time used by PO mixture method with recursive solution and Quasi-Newton method relative to that used by PO mixture method with DCA. We examined three different accuracy levels as distances of parameter estimates from the true parameters (estimated from PO mixture method with DCA with tolerance being 1e-10). Among these methods, PO mixture method with DCA in the M-step works the fastest, while the Quasi-Newton method takes the longest time.

Table 4.5: Time used by PO mixture with recursive method and Quasi-Newton method relative to that used by PO mixture with DCA - discrete PH model, censoring proportion = 50%

Distance from true parameters	PO mixture with recursive method	Quasi-Newton method
0.1	1.6	7.2
0.01	1.7	6.1
0.001	1.7	6.2

Table 4.6: Time used by PO mixture with recursive method and Quasi-Newton method relative to that used by PO mixture with DCA - discrete PH model, censoring proportion = 80%

Distance from true parameters	PO mixture with recursive method	Quasi-Newton method
0.1	1.9	8
0.01	2	10
0.001	2.8	14.2

4.7 Discussion

We proposed two methods for modeling discrete failure time data by introducing artificial random variables to the model and treating them as missing data. EM algorithm based on imputation of the artificial missing data is used for the parameter estimation. Two procedures were proposed for solving the M-Step, the recursive one and a version of the MM algorithm, both nested within the M-Step of the original EM. We targeted a family of discrete transformation models and kept the development general with respect to model specification that was done using a Laplace transform $L-$function. The approach allowed considerable flexibility as to the choice of the basis model. This flexibility was exploited in search for computational efficiency. We have explored using the PH and the PO latent basis models, each giving rise to a different family of estimation procedures. The PH mixture method reduces the dimension of the optimization procedure. However, it requires numerical approximation in the imputation step. For large data sets, this adds considerable computational burden to the problem.

In the PO mixture method, on the contrary, all imputations have closed-form expressions making the algorithm precise, simple and stable. Therefore, the PO mixture method is recommended for use with discrete data, while the PH mixture method is better suited for continuous survival models.

For the specification of the algorithm for a particular transformation model the knowledge of the specific distribution of the artificial random variables, U, is not required, and in fact we have not defined them in the applications presented.

In fact the number of algorithms that can be built using the proposed approach is without limits. While we believe the two classes spawned by the latent PH and PO models are perhaps most interesting and serve the variety of discrete and continuous models well, the optimal choice of the algorithm is an open and challenging question.

The applicability of the PO mixture framework goes beyond discrete models. The fact that the form of the PO model likelihood is virtually the same for discrete or continuous situation (unlike the PH model) makes it particularly useful for data coming from a mixed discrete-continuous distribution.

CHAPTER V

Conclusion and Discussion

The MCQEM algorithm proposed in this research is a MCEM-type algorithm with artificial variables introduced to break the dimension of the M-step into a series of Poisson regression sub-problems. Artificial variables slightly increase the dimension of the E-step requiring larger MC sample size to achieve the same accuracy. Therefore, the MCQEM algorithm is best suited for problems with smaller cluster size.

When the number of categories becomes larger, quadrature methods become prohibitively slow. While our method as well as MCEM would require a larger number of MC sample size M in this case, it still remains feasible with larger number of categories. The MCQEM algorithm converges even when the model is non-identifiable due to empty/sparse categories.

Similar to EM algorithm, the MCQEM algorithm does not provide variance estimates automatically. We propose the variance approximation method based on the idea of *Neilsen et al.* (1992), which only requires knowledge of the log-likelihood function and MLE. This allows us to get variance estimates avoiding taking derivatives. It works well when the log-likelihood function is approximately quadratic around the MLE. When the condition is not met, the variance estimates are not reliable. The traditional variance estimates are no good in this case either. We have examined a few extreme cases, where the parameters are near the boundary of the parameter space, and the log-likelihood function is far from quadratic. Neither our method nor numerical differentiation (SAS) is reliable in those extreme cases.

When missing observations are present, if the mechanism is missing at random (MAR) or missing completely at random (MCAR), the method could be extended to deal with missing observations by incorporating them into the EM framework. If missing is not at random, delicate care may be needed. Methods proposed for this purpose in the literature such as sensitivity analysis and multiple imputation (*Fitzmaurice et al.* (2008) and *Roderick J.A. Little* (2002)) should apply to our setting.

We have revisited the point estimation algorithms for the discrete proportional odds model. We found the PO model naturally suited for the discrete or grouped data setting because the form of its likelihood function with right censored data is invariant with respect to the type of data (continuous or grouped) in the sense that the contribution of failures is still proportional to the jump of the baseline cumulative hazard function. This property ensured that the artificial mixture and MM approaches carry over with slight modification from the continuous setting and bring their computational efficiency with them. This stands in stark contrast with the Cox model that has seen a number of challenges with discrete data.

A novel recursive approach has been developed as an alternative method. Although a little slower than the MM-type algorithms it has potentially superior accuracy and stability as it is based on recursive solution to algebraic equations that can be accomplished by stable bi-section algorithms if needed.

Our approach targeted the most difficult niche of data applications when the dimensionality was high despite the model still being discrete. Cancer registry such as SEER gives such an example. The fact that traditional maximization methods fail to exploit the specific likelihood structure and are subject to the curse of dimensionality makes them slow. In particular, Newton type methods have $O(n^3)$ complexity due to a direct or indirect (iterative) information matrix inverse. At the same time MM-type methods show approximately linear complexity curve with increase in the dimension of h.

There is probably little difference between these methods if the dimensionality of the problem is low.

The results hold much promise for the development of efficient computational methods for the broad class of transformation models. The artificial mixture approach (*Tsodikov* (2003a)) allowed to extend the numerical efficiency of the Cox model estimates to transformation models. The Cox model has shown itself poorly in the discrete setting. The PO model, on the contrary, takes to the discrete setting naturally because of its ordinal heritage. We used this observation to develop computationally efficient procedures for a broad class of models that can be represented as an (artificial) mixture of PO models just like continuous transformation models are usually mixtures of PH models (univariate frailty models). We have shown that efficient methods exist for the PO model and it is therefore a good candidate as the base model spawning a mixture family for the discrete and grouped data setting.

We proposed two methods for modeling discrete failure time data by introducing artificial random variables to the model and treating them as missing data. EM algorithm based on imputation of the artificial missing data is used for the parameter estimation. Two procedures were proposed for solving the M-Step, the recursive one and a version of the MM algorithm, both nested within the M-Step of the original EM.

We targeted a family of discrete transformation models and kept the development general with respect to model specification that was done using a Laplace transform $L-$function. The approach allowed considerable flexibility as to the choice of the basis model. This flexibility was exploited in search for computational efficiency. We have explored using the PH and the PO latent basis models, each giving rise to a different family of estimation procedures.

The PH mixture method reduces the dimension of the optimization procedure. However, it requires numerical approximation in the imputation step. For large data sets, this adds considerable computational burden to the problem.

In the PO mixture method, on the contrary, all imputations have closed-form expressions making the algorithm precise, simple and stable. Therefore, the PO mixture method is recommended for use with discrete data, while the PH mixture method is better suited for continuous survival models.

For the specification of the algorithm for a particular transformation model the knowledge of the specific distribution of the artificial random variables, U, is not required, and in fact we have not defined them in the applications presented.

In fact the number of algorithms that can be built using the proposed approach is without limits. While we believe the two classes spawned by the latent PH and PO models are perhaps most interesting and serve the variety of discrete and continuous models well, the optimal choice of the algorithm is an open and challenging question.

The applicability of the PO mixture framework goes beyond discrete models. The fact that the form of the PO model likelihood is virtually the same for discrete or continuous situation (unlike the PH model) makes it particularly useful for data coming from a mixed discrete-continuous distribution.

APPENDICES

APPENDIX A

Derivation of power series method in artificial mixture method in IV

The expansion of $\frac{t}{e^t-1}$ involves so-called Bernoulli numbers B_n and has the form (*Gradshteyn et al.* (2007))

$$\frac{t}{e^t-1} = \sum\nolimits_{n=0}^{\infty} B_n \frac{t^n}{n!} = 1 + \sum\nolimits_{n=1}^{\infty} B_n \frac{t^n}{n!}, \tag{A.1}$$

where B_n have the recursive relationship $B_n = \sum_{k=0}^{n} \binom{n}{k} B_k$, $B_0 = 1$. With $t = U_{ij}\Delta H_i$ and $a = \sum_{n=1}^{\infty} B_n \frac{t^n}{n!}$ we obtain the expansion

$$\log(1+a) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{a^n}{n}, \tag{A.2}$$

for $-1 < a < 1$. Combining the above expressions and utilizing the following formula for powers of power series

$$\begin{aligned} &\left(\sum\nolimits_{k=0}^{\infty} a_k x^k\right)^n = \sum\nolimits_{k=0}^{\infty} c_k x^k, \text{where} \\ &c_0 = a_0^n, c_m = \tfrac{1}{ma_0} \sum\nolimits_{k=1}^{m} (kn - m + k) a_k c_{m-k}, \end{aligned} \tag{A.3}$$

for $m \geq 1$, see *Gradshteyn et al.* (2007), we can finally express $\log \frac{t}{e^t-1}$ as a power series of t. Denote its coefficients by γ_k, so that $\log \frac{t}{e^t-1} = \sum_{k=1}^{\infty} \gamma^k t^k$.

BIBLIOGRAPHY

BIBLIOGRAPHY

Agresti, A. (2002), *Categorical Data Analysis*, John Wiley and Sons, New Jersey.

Agresti, A. (2007), *An Introduction to Categorical Data Analysis*, John Wiley & Sons.

An, L. T. H., and D. T. Pham (1997), Solving a class of linearly constrained indefinite quadratic problems by d.c. algorithms, *Journal of Global Optimization*, *11*(3), 253–285.

Andersen, P., O. Borgan, R. Gill, and N. Keiding (1993), *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.

Baker, S. (1994), The Multinomial-Poisson transformation, *The Statistician*, *43*, 495–504.

Bennett, S. (1983a), Analysis of survival data by the proportional odds model, *Statistics in medicine*, *2*(2), 273–277.

Bennett, S. (1983b), Log-logistic regression models for survival data, *Applied Statistics*, *32*(2), 165–171.

Booth, J., and J. Hobert (1999), Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm, *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *61*, 265–285.

Breslow, N. E. (1974), Covariance analysis of censored survival data, *Biometrics*, *30*, 89–99.

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

Brian S. Caffo, G. L. J., Wolfgang Jank (2005), Ascent-based mcem, *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *67*, 235–251.

Chen, J., D. Zhang, and M. Davidian (2002), A monte carlo em algorithm for generalized linear mixed models with flexible random effects distribution, *Biostatistics*, *3*(3), 347–360.

Chen, Z., and L. Kuo (2001), A note on the estimation of multinomial logit models with random effects, *The American Statistician*, *55*, 89–95.

Cheng, S. C., L. J. Wei, and Z. Ying (1995), Analysis of transformation models with censored data, *Biometrika*, *82*(4), 835–845.

Clarkson, D. B., and Y. Zhan (2002), Using spherical-radial quadrature to fit generalized linear mixed effects models, *Journal of Computational and Graphical Statistics*, *11*, 639–659.

Cox, D. R. (1972), Regression models and life-tables, *Journal of the Royal Statistical Society. Series B*, *34*, 187–220.

Cuzick, J. (1988), Rank regression, *Ann.Statist.*, *16*, 1369–1389.

Daniels, M., and C. Gatsonis (1997), Hierarchical polytomous regression models with applications to health services research, *Statistics in Medicine*, *16*, 2311–2325.

de Leeuw, J. (1994), *Information Systems and Data Analysis (ed. H. H. Bock, W. Lenski, and M. M. Richter)*, chap. Block relaxation algorithms in statistics, pp. 308–325, Berlin: Springer-Verlag.

Efron, B. (1977), The efficiency of coxs likelihood function for censored data, *Journal of the American Statistical Association*, *72*, 557–565.

Fitzmaurice, G., M. Davidian, G. Verbeke, and G. Molenberghs (2008), *Longitudinal Data Analysis*, Chapman and Hall/CRC, Boca Raton,FL.

Fleming, T., and D. Harrington (2005), *Counting process and survival analysis*, Wiley Series in Probability and Statistics.

Gosh, M., L. Zhang, and B. Mukherjee (2006), Equivalence of posteriors in the bayesian analysis of the multinomial-poisson transformation, *Metron-International Journal of Statistics LXIV*, *1*, 19–28.

Gradshteyn, I., I. Ryzhik, A. Jeffrey, and D. Zwillinger (2007), *Table of Integrals, Series, and Products*, Academic Press.

Hartzel, J., A. Agresti, and B. Caffo (2001), Multinomial logit random effects models, *Statistical Modelling*, *1*, 81–102.

Hedeker, D. (2003), A mixed-effects multinomial logistic regression model, *Statistics in Medicine*, *22*, 1433–1446.

Huang, J., and A. J. Rossini (1996), Sieve estimation for the proportional odds failure-time model with interval censoring, *Technical Report 250*, *250*.

Hunter, D. R., and K. Lange (2002), Computing estimates in the proportional odds model, *Annals of the Institute of Statistical Mathematics*, *54*(1), 155–168.

Johnson, W., and R. Christensen (1986), Bayesian nonparametric survival analysis for grouped data, *The Canadian Journal of Statistics*, *14*, 307–314.

Kalbfleisch, J. D., and R. L. Prentice (1973), Marginal likelihoods based on cox's regression and life model, *Biometrika*, *60*, 267–278.

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

Kirmani, S., and R. Gupta (2001), The proportional odds model in survival analysis, *Annals of the Institute of Statistical Mathematics*, *53*, 203–216.

Kuss, O., and D. McLerran (2007), A note on the estimation of the multinomial logistic model with correlated responses in sas, *Computer Methods and Programs in Biomedicine*, *87*, 262–269.

Lang, J. (1996), On the comparison of multinomial and Poisson log-linear models, *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *58*, 253–266.

Lange, K., D. Hunter, and I. Yang (2000), Optimization transfer using surrogate objective functions (with discussion), *Journal of Computational and Graphical Statistics*, pp. 1–59.

Li, Z., P. Gilbert, and B. Nan (2008), Weighted likelihood method for grouped survival data in case-cohort studies with application to hivv accine trials, *Biometrics*, *64*, 1247C1255.

Louis, T. A. (1982), Finding the observed information matrix when using the EM algorithm, *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *4*(2), 226–233.

Mccullagh, P., and J. Nelder (1989), *Generalized linear models*, Chapman and Hall, London.

McCulloch, C. E. (1997), Maximum likelihood algorithms for generalized linear mixed models, *Journal of the American Statistical Association*, *92*, 162–170.

McLachlan, G., and T. Krishnan (1997), *The EM Algorithm and Extensions*, John Wiley and Sons, New York.

Murphy, S. (2000), On profile likelihood, *Journal of the American Statistical Association*, *95*, 449–485.

Murphy, S. A., A. J. Rossini, and A. W. Van der Vaart (1997), Mle in the proportional odds model, *Journal of the American Statistical Association*, *92*, 968–976.

Neilsen, G., R. Gill, P. Andersen, and S. TI (1992), A counting process approach to maximum likelihood estimation in frailty models, *Scand J Statist*, *19*, 25–43.

Pettitt, A. N. (1984), Proportional odds model for survival data and estimates using ranks, *Applied Statistics*, *33*, 169–175.

Pipper, C. B., and C. Ritz (2006), Checking the grouped data version of the cox model for interval-grouped survival data, *Scandinavian Journal of Statistics*, *34*, 405–418.

Prentice, R. L., and L. A. Gloeckler (1978), Regression analysis of grouped survival data with application to breast cancer data, *Biometrics*, *34*, 57–68.

Prentice, R. L., and J. D. Kalbfleisch (2003), Mixed discrete and continuous cox regression model, *Lifetime Data Analysis*, *9*, 195–210.

Rabe-Hesketh, S., A. Skrondal, and A. Pickles (2002), Reliable estimation of generalized linear mixed models using adaptive quadrature, *The Stata Journal*, *2*, 1C2.

Roderick J.A. Little, D. B. R. (2002), *statistical analysis with missing data*, Wiley-interscience, Hoboken, NJ.

Scott, S. L. (2011), Data augmentation, frequentist estimation, and the bayesian analysis of multinomial logit models, *Statistical Papers*, *52*, 87–109.

Shen, X. (1998), Proportional odds regression and sieve maximum likelihood estimation, *Biometrika*, *85*(1), 165–177.

Sinha, D., M. A. Tanner, and W. J. Hall (1994), Maximization of the marginal likelihood of grouped survival data, *Biometrika*, *81*, 53–60.

Stewart, W. H., and D. A. Pierce (1982), Efficiency of cox's model in estimating regression parameters with grouped survival data, *Biometrika*, *69*, 539–545.

Tsodikov, A. (2002), Semi-parametric models of long- and short-term survival: an application to the analysis of breast cancer survival in utah by age and stage, *Statistics in Medicine*, *21*, 895–920.

Tsodikov, A. (2003a), Semiparametric models: A generalized self-consistency approach, *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *65*(3), 759–774.

Tsodikov, A. (2003b), Semiparametric models: A generalized self-consistency approach, *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *65*(3), 759–774.

Tsodikov, A., and S. Chefo (2008), Generalized self-consistency:multinomial logit model and poisson likelihood, *Journal of Statistical Planning and Inference*, *138*, 2380–2397.

Tsodikov, A., and S. Wang (2011), On the estimation of po model for discrete failure time data, *submitted*.

Tsodikov, A., D. Hasenclever, and M. Loeffler (1998), Regression with bounded outcome score: Evaluation of power by bootstrap and simulation in a chronic myelogenous leukaemia clinical trial, *Statistics in Medicine*, *17*, 1909–1922.

Wang, S., and A. Tsodikov (2010), A self-consistency approach to multinomial logit model with random effects, *Journal ofStatisticalPlanningandInference*, *140*, 1939–1947.

Yu, B., R. Tiwari, K. Cronin, and E. Feuer (2004), Cure fraction estimation from the mixture cure models for grouped survival data, *Statistics in Medicine*, *23*, 1733–1747.

Yu, B., L. Huang, R. C. Tiwari, E. J. Feuer, and K. A. Johnson (2009), Modelling population-based cancer survival trends by using join point models for grouped survival data, *Journal of the Royal Statistical Society: Series A*, *172*, 405–425.

Zhao, X., and X. Zhou (2008), Discrete-time survival models with long-term survivors, *Statistics in Medicine*, *27*, 1261–1281.