The use of submodels as a basis for efficient estimation of complex models

by

Abdollah Safari

M.Sc., University of Tehran, Iran, 2012
B.Sc., University of Tehran, Iran, 2010

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Spring 2018

Approval

Name: **Abdollah Safari**

Degree: **Doctor of Philosophy (Statistics)**

Title: **The use of submodels as a basis for efficient estimation of complex models**

Examining Committee: **Chair:** Jinko Graham
Professor

Rachel MacKay Altman
Senior Supervisor
Associate Professor

Thomas M. Loughin
Supervisor
Professor

Dave Campbell
Internal Examiner
Associate Professor
Department of Statistics and Actuarial Science

John Neuhaus
External Examiner
Professor
School of Medicine
University of California, San Francisco

Date Defended: **November 8, 2017**

Abstract

In this thesis, we consider problems where the true underlying models are complex and obtaining the maximum likelihood estimator (MLE) of the true model is challenging or time-consuming.

In our first paper, we investigate a general class of parameter-driven models for time series of counts. Depending on the distribution of the latent variables, these models can be highly complex. We consider a set of simple models within this class as a basis for estimating the regression coefficients in the more complex models. We also derive standard errors (SEs) for these new estimators. We conduct a comprehensive simulation study to evaluate the accuracy and efficiency of our estimators and their SEs. Our results show that, except in extreme cases, the maximizer of the Poisson generalized linear model (the simplest estimator in our context) is an efficient, consistent, and robust estimator with a well-behaved standard error.

In our second paper, we work in the context of display advertising, where the goal is to estimate the probability of conversion (a pre-defined action such as making a purchase) after a user clicks on an ad. In addition to accuracy, in this context, the speed with which the estimate can be computed is critical. Again, computing the MLEs of the true model for the observed conversion statuses (which depends on the distribution of the delays in observing conversions) is challenging, in this case because of the huge size of the data set. We use a logistic regression model as a basis for estimation, and then adjust this estimate for its bias. We show that our estimation algorithm leads to accurate estimators and requires far less computation time than does the MLE of the true model.

Our third paper also concerns the conversion probability estimation problem in display advertising. We consider a more complicated setting where users may visit an ad multiple times prior to taking the desired action (e.g., making a purchase). We extend the estimator that we developed in our second paper to incorporate information from such visits. We show that this new estimator, the DV-estimator (which accounts for the distributions of both the conversion delay times and the inter-visit times) is more accurate and leads to better confidence intervals than the estimator that accounts only for delay times (the D-estimator). In addition, the time required to compute the DV-estimate for a given data set is only moderately greater than that required to compute the D-estimate – and is substantially less than that required to compute the MLE.

In summary, in a variety of settings, we show that estimators based on simple, misspecified models can lead us to accurate, precise, and computationally efficient estimates of both the key model parameters and their standard deviations.

Keywords: Kullback-Leibler information; Bias-adjustment; Parameter-driven models; Time series; Count data; Generalized linear model; Generalized linear mixed model; Hidden Markov model; Display advertising; Conversion probability; Survival times; Censoring.

To my mother and my father!

“How strange and foolish is man. He loses his health in gaining wealth. Then, to regain his health he wastes his wealth. He ruins his present while worrying about his future, but weeps in the future by recalling his past. He lives as though death shall never come to him, but dies in a way as if he were never born!”

– Ali ibn Abi Talib

Acknowledgements

My deepest gratitude goes to Dr. Rachel Altman. Rachel pushed and inspired me whenever I needed motivation yet managed to be exceptionally patient and understanding. Always kind, always supportive, and always approachable, she was more of a friend. I have learned so much not only from her expertise in her field, but also from her character as an incredible human being. I could not have asked for, or imagined, a better supervisor, and I still do not think I deserved it. I am and forever will be grateful.

I would also like to thank Dr. Thomas Loughin. His advice and guidance proved to be essential to my work. I am thankful for that, and for all his good stories that brought us joy and laughter.

I am grateful to my master's supervisor, Dr. Hamid Pezeshk, whose encouragements played in key role in my decision to pursue a PhD. I would also like thank Dr. Ahmad Parsian for inspiring me to study statistics in the first place. I still vividly remember my very first course in statistics with him in my first year at university. Had it not been for his energetic and motivating class, I would not have pursued a career in statistics.

And of course, not of this would have been possible without my family and my friends who were always there for me. I am lucky to have you.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

Estimating the parameters of a complex model can be a challenging problem. The challenges can have a variety of root causes, including intractability of the likelihood or the size of the data set. One way of simplifying estimation in such cases is to maximize the likelihood of a simpler model chosen to approximate the complex model, especially if we are interested in estimating of a subset of the complex model's parameters.

In this thesis, we introduce three settings where the true models are complex and obtaining the maximum likelihood estimator (MLE) can be difficult. We propose some alternative simple models that have likelihoods that can be maximized efficiently. Since these simple models are misspecified, their maximizers are sometimes biased for the parameters of the true models. In these cases, we adjust the estimates for their bias to obtain accurate estimates of the parameters of the true models. We use the Kullback-Leibler information criterion (KLIC) approach to determine the adjustment. In addition, we derive standard errors (SEs) of the adjusted estimators. Our estimators are, in general, accurate and efficient, and require a relatively short computation time.

1.2 Previous work

Since the general idea of this thesis is approximating complex models using simple misspecified models, we list some previous work in this area. This section is intended to supplement the literature review in each paper, which is, in general, specific to the settings we consider there.

The results in our papers rely on two key works by Halbert White. White [1982] shows that when the observations are independent, under mild conditions, the MLE of the parameters based on a misspecified model is consistent for the minimizer of the KLIC (Theorem 2.2 of White 1982). In a second paper, White [1984] develops asymptotic covariance matrices for estimators based on general dynamic models that are accurate even when the models are misspecified.

Much additional research has been conducted on the impact of model misspecification on parameter estimation or prediction in different contexts. Neuhaus et al. [1992] consider a misspecified mixing distribution in logistic-normal models. They conclude that, although the MLE of the regression coefficients can be biased, the magnitude of this bias is typically small. Gustafson [2001] considers the estimation of the scale parameter in a general class of models, under model misspecification, using Kullback-Leibler divergence. Chow [1984] builds on the work of White [1984], deriving the asymptotic covariance matrix of the ML estimator of a misspecified model in an economics context. Heagerty and Kurland [2001] investigate the impact of model violations on the MLE of regression coefficients in a generalized linear mixed model. Bates and White [1985] present a general theory of consistent estimation for possibly misspecified dynamic models. ten Have et al. [1999] show that fitting a misspecified model to binary data with multiple levels of clustering can lead to bias in the regression coefficient estimators. Tan et al. [1999] show that, in clustered binary data, marginal quantities (such as sensitivity and specificity) are unbiased under misspecification of the distribution of the random effects. Many authors point out that the regression coefficient estimates may not be sensitive to the violation of the random effect assumptions in generalized linear mixed models (see, e.g., Huber 1967 and Heagerty and Zeger 2000). Rizopoulos and Verbeke [2008] suggest an alternative parameterization for random effects and parameters, and investigate the effect of misspecifying the random effects distribution on the MLEs in a longitudinal response process. All of the mentioned work, among others (see, e.g., Takeuchi 1976, Zeger and Liang 1986 and Liang and Zeger 1986), are either not directly relevant to our problems, especially for the correlated observations probelm in Chapter 2 (see, e.g., Berk 1966, Berk 1970 and Huber 1967), or they lead us to the same results as the ones proposed by White.

1.3 Motivating references for this dissertation

We have two main motivating references for this dissertation. The first is Davis et al. [2000], who propose the maximizer of the simple Poisson generalized linear model as an estimator of the regression coefficients in a complex model for time series of counts. The second is Chapelle [2014], who considers the context of display advertising, and studies the MLE of the probability that a click on an online ad converts (i.e., results in a pre-defined action by the user, such as a purchase). The computation of the MLE in the first setting can be challenging due to the complex structure of the likelihood, and in the second setting due to the large size of the data set.

1.4 Display advertising overview

The main motivating application in chapters 3 and 4 is display advertising. Since this topic is fairly new, we briefly introduce some key terms and definitions in this section.

A newly developed method to advertise products (for selling or buying) is display ads on the Internet (see Muthukrishnan 2009). A diverse set of problems have arisen within this context (see, e.g., McAfee et al. 2010, Tietenberg 2010, and Varian 2007).

The two most common forms of online advertising are **paid search advertising** and **display advertising**. The former is the the most common advertising form where advertisers pay to have their ads displayed to users as they type queries into search engines. The latter is a newly developed, popular form of online advertising where ads can be displayed as banners in various sizes and positions in different webpages. These banners can have different forms, including text, images, and video. In this thesis, although we focus on display advertising, our findings can be applied to paid search advertising with some small modifications[1]. Display advertising has four main components: advertisers, publishers, users, and connectors (or ad networks). Advertisers want to promote their products. Publishers own webpages with some available spots for ads. Users consume services provided by publishers on different webpages. Finally, the main component, connectors (e.g., Google) look for the "best" match of the other three components, in a way to maximize their own profit. In this thesis, we consider an advertiser point of view, and track the display advertising procedure only after a click event occurs. After such an event, the advertiser can track the conversion status of the click (converted or not), and the conversion delay (the time until conversion, should it occur). Figure 1.1 shows the observed conversion statuses and conversion delays of users after they click an ad, over time.

Some key terms in display advertising are defined as follows[2]:

- **Reach:** Number of people who can potentially view the online ad.
- **Click-through Rate (CTR):** The chance that a user clicks on an online ad.
- **Conversion Rate (CVR):** Percentage of users who clicked on an ad who eventually convert.
- **Campaign:** Process of planning, creating, buying, and tracking an advertising procedure.
- **Impressions:** Number of times a banner or text ad was requested and presumably seen by a user.

[1] Source: Google Ads Display Network, available at http://www.google.com/ads/displaynetwork

[2] Source: Google Analytics Partners, available at http://www.kasatria.com

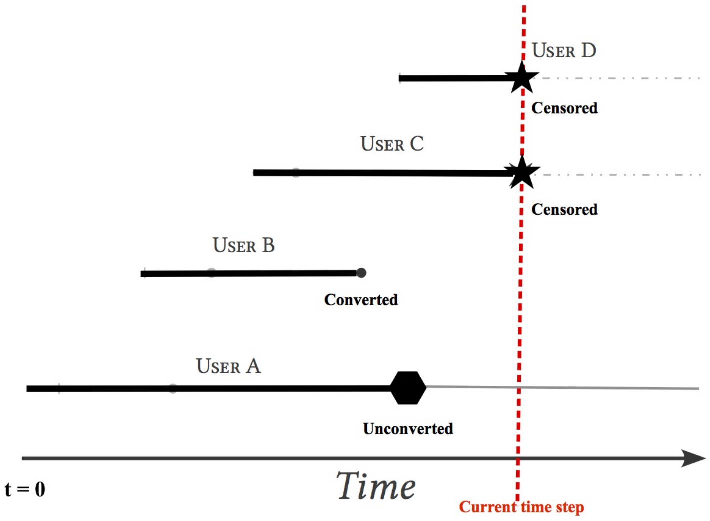

Figure 1.1: Illustration of observed conversion statuses and conversion delays in display advertising

- **Traffic:** Number of visitors and visits that a website receives.

One attractive feature of online display advertising is that advertisers can control the advertising method and the associated costs according to their budget and goals. The three main payment approaches for online ad placement are[3]:

- Cost per thousand impressions (CPM): Advertisers pay for every 1000 impressions of their ads. This approach is appropriate if the number of people who see their banner needs to be guaranteed.

- Cost per click (CPC): Advertiser pays only for the traffic that goes to their website (i.e., via a click on their ad). The CPC (which is the most popular payment way) is a preferred method of payment for advertisers who need to guarantee that they pay only for those users who click on the advertising link or banner and then go to their site.

- Cost per action (CPA): Advertisers pay for every converted click. The CPA is a preferred method of payment for advertisers who want to guarantee the number of conversions generated as a result of an advertisement.

We focus on only the CPA approach in this thesis.

[3] Source: Ontario advertising company, available at https://www.ontario.ca

1.5 Thesis structure

This thesis is structured as a series of five chapters, each with a unique focus. The current chapter and Chapter 5 provide introductory information and a summary of our conclusions, respectively. The other three chapters are written as stand-alone papers – though they have related themes, as described above. The three papers have been submitted in different top statistical journals. I am the primary author of all the papers, but others have contributed in each, as described in the following sections. My supervisor, Dr. Rachel MacKay Altman, offered comments on the entire thesis.

1.5.1 Overview of Chapter 2

Chapter 2 is entitled "Parameter-Driven Models for Time Series of Count Data", by Safari A., Altman R. M., Leroux B., and has been submitted to Biostatistics.

The focus of the paper is the estimation of the regression coefficients of time series of count data. We consider a general class of parameter-driven models for such data. While this class is highly flexible (and includes some common models as special cases), maximum likelihood (ML) estimation of the regression coefficients of models in this class can be challenging. We study the behaviour of three simple estimators of the regression coefficients. We show that the estimator based on the Poisson generalized linear model performs remarkably well in terms of bias and efficiency, even if the data are overdispersed or autocorrelated. We also derive a standard error (SE) for our estimators that is simpler and more accurate than those suggested in the literature. At the end, we show how our methods and results can be applied in practice, and include a detailed analysis of polio and epileptic seizure data sets.

1.5.2 Overview of Chapter 3

Chapter 3 is entitled "Display advertising: Estimating conversion probability efficiently", by Safari A., Altman R. M., Loughin T. M., and has been submitted to the Annals of Applied Statistics.

The context of this paper is display advertising. Our focus is the development of an accurate and computationally efficient estimator of the probability that a click on an online ad will convert. Computational efficiency is critical in this Big Data setting where publishers need to re-estimate conversion probability rapidly and frequently as time progresses and more data accrue. At any time point, conversion probability can be estimated based on a model for data collected on clicks observed prior to this time, namely the conversion statuses and the delays in observing conversions. The maximum likelihood estimate (MLE) of the parameters of this model is computationally expensive. Instead, we consider a different estimator based on a simple logistic regression model (which we call the naive model) that treats the current conversion statuses of the clicks as their eventual ones. The naive model ignores the delays in conversion and the maximizer of the likelihood of this model leads to an under-estimate of conversion probability. However, we can adjust this estimator for its bias by incorporating information from the delay times. We also propose an algorithm to compute the standard error (SE) of our bias-adjusted estimator efficiently. We show how our methods and results can be applied to a real data set. We also conduct a simulation study where we show that our adjusted estimator (which we call the D-adjusted estimator) has relatively low bias and much lower computational time than the MLE, and has an accurate SE.

1.5.3 Overview of Chapter 4

Chapter 4 is entitled "Conversion probability estimation in display advertising: Incorporating information from multiple visits to the same ad", by Safari A., Altman R. M., and has been submitted to the Journal of Computational and Graphical Statistics.

This paper is an extension of the paper described in Chapter 3. In particular, we consider the case where users may visit an ad multiple times after their click prior to making a purchase or taking some other pre-defined action. Our goal in this paper is to incorporate both conversion delay time and users' previous visits to estimate the conversion probability more accurately. As in Chapter 3, we build our estimator based on a naive logistic regression model that treats the current conversion statuses of the clicks as their eventual ones, and then adjust its bias. In this paper, however, we not only adjust the estimator based on the delay time distribution, but also based on the distribution of the number of visits. Similar to Chapter 3, our new bias-adjusted estimator (DV-adjusted) has much lower computational time relative to the MLE of the true model. With a simulation study, we show that although the DV-adjusted estimator is slightly more computationally expensive than the D-adjusted estimator, it's more accurate with almost zero bias at the end of the observation period in our simulation study, whereas the D-adjusted estimator remains biased for longer.

Chapter 2

Parameter-Driven Models for Time Series of Count Data

The chapter derives, with few modifications, from:
Safari, A., Altman, R. M., Leroux, B., 2018. *Parameter-Driven Models for Time Series of Count Data*, In preparation.

2.1 Abstract

This paper considers a general class of parameter-driven models for time series of counts. The maximum likelihood estimator (MLE) of such models can be challenging to obtain. Therefore, we instead propose the maximizers of two simple intermediate models: the 2-state Poisson mixture model and the 2-state Poisson hidden Markov model (HMM). We evaluate the accuracy and efficiency of these estimators relative to the maximizer of the Poisson generalized linear model (GLM) considered in the literature. Moreover, we derive standard errors (SEs) for all three estimators and propose a simple algorithm to compute the SEs efficiently. Our results, based on a comprehensive simulation study, show that except in extreme cases, the MLE of the GLM is an efficient, consistent, and robust estimator with a well-behaved estimated standard error. The MLE of the HMM is appropriate only when the true model is extreme relative to the GLM. Our results are applied to problems concerning polio incidence and daily numbers of epileptic seizures.

2.2 Introduction

Cox [1981] defines two classes of models for time series data: parameter-driven models (PDMs) and observation-driven models (ODMs). In an ODM, autocorrelation is introduced through the dependence of the conditional expectation of the current observation on the past observations (e.g. an AR(1)), whereas a PDM uses a set of latent variables to explain the autocorrelation. A key feature of PDMs is that the observed data are assumed to be independent given the values of the latent variables. The class of PDMs includes moving average models, hidden Markov models (HMMs), generalized linear mixed models (GLMMs), and hierarchical generalized linear models (see Lee and Nelder 1996). ODMs and PDMs have different properties and usages (see Davis and Yam 2004 for more details). For example, in the case of ODMs, evaluating the likelihood and estimating the model parameters are typically straightforward, and, for this reason, ODMs are often used for forecasting. On the other hand, the interpretation of PDMs is usually simpler than that of ODMs. Specifically, on the marginal level, properties of PDMs are often easier to establish than those of ODMs, especially when the data are non-normally distributed. PDMs thus provide a convenient way of modeling overdispersion and autocorrelation. However, PDMs tend to have more complicated likelihoods, and hence estimation of these models is challenging.

In this paper, we focus solely on the study of PDMs for time series of count data. In particular, we use the setting of Zeger [1988] where, conditional on latent variables, the observations are Poisson distributed with log mean specified by a linear function of predictors and latent variables. Zeger [1988] introduces a consistent quasi-maximum likelihood estimator (QMLE) of the regression coefficients in such models. Others have used this approach in different areas of application (e.g., Campbell 1994, Brannas and Johansson 1994, Albert et al. 1994, and McShane et al. 1997). Although computationally more feasible than maximum likelihood estimation, the quasi-maximum likelihood estimation is still a complex approach, and is not available in standard statistical software packages. Davis, Dunsmuir and Wang 2000 (henceforth called DDW) suggest estimating the regression coefficients using the maximizer of the likelihood of a Poisson generalized linear model (GLM). They prove under general conditions that, although this estimator is based on a misspecified model, it is consistent for the true regression coefficients. They also derive its asymptotic variance-covariance matrix. The GLM-based estimator has advantages like simplicity and robustness. However, to the best of our knowledge, its efficiency has not been studied. Similarly, the consistency of estimators of other PDMs is typically the only property studied (e.g., Davis and Wu 2009 and Neuhaus et al. 2013).

Our goal in the present paper is to investigate the efficiency of the GLM estimator relative to that of two other simple estimators. In particular, we introduce two approximating models that can be considered intermediate to the simple GLM and the complex true PDM: Poisson finite mixture models (FMMs) and Poisson HMMs. We then evaluate the bias and standard errors (SEs) of our three estimators of the regression coefficients. These estimators are all within the class of PDMs that we consider, but unlike the maximum likelihood estimators (MLEs) of some models, are easy to compute. In addition, we derive an estimator of the asymptotic variance-covariance matrices for these estimators and propose a simple algorithm to compute these matrices efficiently.

The remainder of the paper is organized as follows. In §2.3, we specify the Poisson PDM of Zeger [1988] and present some special cases of his model. We emphasize the Poisson FMM and HMM, the two new intermediate models that we consider as a basis for estimating the regression parameters of this model. In §2.4, we derive estimators of the asymptotic SEs of the GLM, FMM, and HMM estimators. In §2.5, we illustrate the performance of the estimators and their standard errors for certain choices of the true model. In §2.6, we present results on the relative performances of the estimators for a broad class of true models and provide some general rules for efficient estimation of regression coefficients in the context of time series of counts. In §2.7, we present applications of our results to problems concerning polio incidence (in Zeger 1988 and DDW) and daily numbers of epileptic seizures. We conclude with a discussion in §2.8.

2.3 Model specification and estimation

Let $Y = (Y_1, \ldots, Y_n)$ be the observed counts, and let $\alpha = (\alpha_1, \ldots, \alpha_n)$ be the latent variables (discrete or continuous). We assume, as for all PDMs, that $\{Y_t\}$ are independent given $\{\alpha_t\}$. Following Zeger [1988], we model $Y_t \mid \alpha_t$ as having a Poisson distribution with

$$\log\left(E\left[Y_t \mid \alpha_t\right]\right) = X_t'\beta + \alpha_t \tag{2.1}$$

where X_t is a d-dimensional vector of covariates at time t and $\{\alpha_t\}$ is a general stationary process with $E\left[\exp\left(\alpha_t\right)\right] = 1$, $Var\left(\exp(\alpha_t)\right) = \sigma_\alpha^2$, $Cov\left(\exp(\alpha_t), \exp(\alpha_s)\right) = \gamma_{t-s}$, and $Corr(\exp(\alpha_t), \exp(\alpha_s)) = \rho_{t-s}$, where $t \geq s$. We assume that the first entry of X_t is a 1, i.e., that the first entry of β is the intercept. Then the likelihood of the true model can be written as

$$\mathcal{L} = \int_\alpha \prod_{t=1}^{n} P(y_t \mid \alpha_t) dG(\alpha_1, \ldots, \alpha_n) \tag{2.2}$$

where P is the Poisson probability mass function and G is the joint cumulative distribution function of the latent variables. Depending the distribution of the latent variables (G), (2.2) can be a complex function with no closed form, and obtaining its maximizer can be challenging.

One way to understand the implications of the assumed (conditional) model on the resulting distribution for the observed data is to examine the marginal moments. In Zeger's model, the effect of the covariates on the marginal moments and the mean-variance and mean-covariance relationships are easy to determine. In particular, the first and second marginal moments are:

$$\begin{aligned}
\mu_t \equiv E\left(Y_t\right) &= \exp\left(X_t'\beta\right) && (2.3)\\
Var\left(Y_t\right) &= \mu_t + \mu_t^2\sigma_\alpha^2 && (2.4)\\
Cov(Y_s, Y_t) &= \mu_s\mu_t\gamma_{t-s}, \quad t \geq s && (2.5)\\
Corr(Y_s, Y_t) &= \frac{\mu_s\mu_t\gamma_{t-s}}{\sqrt{\left(\mu_s + \mu_s^2\sigma_\alpha^2\right)\left(\mu_t + \mu_t^2\sigma_\alpha^2\right)}}, \quad t \geq s && (2.6)
\end{aligned}$$

Equation (2.4) shows that $\mu_t^2 \sigma_\alpha^2$ is the extra-Poisson variation and is a function of σ_α^2. In addition, from (2.5), Davis et al. [2000] show that $\mid Corr(Y_s, Y_t) \mid \leq \mid \rho_{t-s} \mid$, regardless of the distribution of the latent process. Consequently, detecting and analyzing the latent process based on the observed data can be challenging.

Zeger's model is quite general, and, for this reason, includes many common models for time series of counts. In §2.3.1, we discuss some estimators of the parameters of these models that have been studied in the literature. In §2.3.2 and §2.3.3, we present two new estimators of the regression parameters in (2.1), the FMM and HMM estimators, using the idea of intermediate models. Like the GLM estimator, these new estimators are based on (usually) misspecified but simple models.

2.3.1 Existing estimators

In this section, we describe two estimators of the regression parameters in (2.1) proposed in the literature: DDW's estimator (which we call the GLM estimator) and the MLE.

We first describe the GLM estimator in our notation. If we set $\alpha_t \equiv 0$ in (2.1), then $\{Y_t\}$ are treated as independent and as following a Poisson GLM. The Poisson GLM correctly specifies the marginal mean of (2.1), but misspecifies its conditional mean. The GLM estimator is the maximizer of the associated likelihood function. DDW suggest using this (d-dimensional) estimator to estimate the regression coefficients in Zeger's model and show that, for any stationary non-negative or mixing process α_t, the estimator is consistent for the regression parameters. However, the efficiency of this estimator is questionable since it may not use all of the information in the autocorrelation and overdispersion in the observations.

Another existing estimator of the parameters in (2.1) is the MLE based on the true model (at least when the true model is a GLM, FMM, a HMM with few hidden states, or a GLMM). Depending on the complexity of the true model, obtaining the MLE can be challenging. For instance, Nelson and Leroux [2008] study the MLE when the true model is a Poisson GLMM. To describe this estimator in our notation, let $\{\alpha_t\}$ be an unobserved $AR(1)$ process and set $\alpha_t = c + \phi\alpha_{t-1} + \delta_t, t = 1, \ldots, n$, where $\delta \sim N(0, \sigma^2)$. To satisfy the constraint $E(\exp(\alpha_t)) = 1$, set $c = -\frac{\sigma^2}{2(1+\phi)}$. Consequently, $\alpha_t \sim N(-\frac{\sigma^2}{2(1-\phi^2)}, \frac{\sigma^2}{1-\phi^2})$ (DDW). Then, the process $\{Y_t\}$ is treated as following a Poisson GLMM. We call the maximizer of the associated likelihood function the GLMM estimator. The GLMM estimator can incorporate information in the overdispersion and autocorrelation. However, the associated likelihood function does not have a simple algebraic form, and numerical methods typically do not perform well for such high dimensional integrals. Nelson and Leroux [2008] use Markov chain Monte Carlo methods to approximate the likelihood and obtain the maximizer. Their simulation studies suggest that, when the true model is a Poisson GLMM, the GLMM estimator (i.e., the MLE) is consistent for the regression parameters. However, computations are expensive, particularly for large sample sizes. Thus, we do not consider their context further in this paper.

In practice, the true underlying model is almost always unknown. Even when the true model is specified, obtaining the MLE is usually difficult. Thus, in the following two subsections, we propose two new simple estimators based on models that are intermediate to the GLM and true model, the FMM and HMM. The FMM and HMM capture overdispersion and/or autocorrelation that the GLM does not; our goal in this paper is to explore the properties of the maximizers of the likelihoods associated with these models relative to those of the GLM estimator.

2.3.2 Poisson FMM estimator

Since Poisson GLMs can't capture overdispersion in the observations, we suggest basing estimation on the Poisson FMM, a model that is within the class (2.1), but more flexible than the GLM. In particular, we consider the model where the latent variables $\{\alpha_t\}$ are independent and multinomial distributed with k possible outcomes ($\alpha_t \in \{S_1, \ldots, S_k\}$), and corresponding probabilities $p_i = P(\alpha_t = S_i)$, $i = 1, \ldots, k$. The process $\{Y_t\}$ is thus treated as following a Poisson mixture model. The likelihood associated with this model can be expressed as

$$\mathcal{L} = \prod_{t=1}^{n} \sum_{i=1}^{k} P(Y_t = y_t \mid \alpha_t = S_i) p_i, \tag{2.7}$$

where $P(Y_t = y_t \mid \alpha_t = S_i)$ is the Poisson probability mass function with corresponding mean parameter $\mu_t \exp(S_i)$. We define the FMM estimator as the maximizer of (2.7). For the purpose of simplicity, we restrict our study to the case where $k = 2$ (henceforth called FMM2). Optimization of the likelihood function is easy in this case and the estimator is of dimension just $d + 2$ for $k = 2$.

Like the GLM estimator, the FMM2 estimator may not use all of the information in the autocorrelation in the observations. However, it can capture the information in the overdispersion. Therefore, we expect it to be more efficient than the GLM estimator when overdispersion is present.

2.3.3 Poisson HMM estimator

The second model we suggest for estimation purposes is the Poisson HMM, which allows for both autocorrelation and overdispersion. A stationary Poisson HMM can be defined by allowing α_t in (2.1) to follow a Markov chain (MC) with k hidden states ($\alpha_t \in \{S_1, \ldots, S_k\}$), transition probabilities P_{ij}, $i, j = 1, \ldots, k$, and limiting probabilities π_{S_j}, $j = 1, \ldots, k$. As an aside, the Poisson FMM is a special case of the Poisson HMM where the rows of the transition probability matrix, $\{P_{i,j}\}$, are identical.

The likelihood function of the Poisson HMM can be expressed as

$$\begin{aligned}\mathcal{L} &= \sum_{i_1=1}^{k} \pi_{S_{i_1}} P(Y_1 = y_1 \mid \alpha_1 = S_{i_1}) \sum_{i_2=1}^{k} P_{i_1,i_2} P(Y_2 = y_2 \mid \alpha_2 = S_{i_2}) \dots \\ & \sum_{i_n=1}^{k} P_{i_{n-1},i_n} P(Y_n = y_n \mid \alpha_n = S_{i_n})\end{aligned} \tag{2.8}$$

This expression is equivalent to a product of matrices. We define the HMM estimator as the maximizer of (2.8). For a small number of hidden states, k, the HMM estimator is easy to compute using numerical methods. As in the case of the FMM estimator, we thus restrict attention to the case where $k = 2$ (henceforth called HMM2). We expect the HMM2 estimator, which can use the information in the overdispersion and autocorrelation in the data but has dimension of just $d + 3$, to be more efficient than the GLM and FMM2 estimators when overdispersion and autocorrelation are present.

2.4 SE of the estimators

In this section, we develop approximate SEs for our two new estimators (the HMM2 and FMM2 estimators) and the previously developed GLM estimator (DDW). In addition, we propose a simple and computationally efficient algorithm to compute these SEs. White [1984] develops asymptotic covariance matrices for estimators based on dynamic models (which include Poisson HMMs, FMMs and GLMs) even when the models are misspecified. Let M be the (possibly misspecified) model that we fit to the data, θ be the vector of parameters in this model, $\widehat{\theta_n}$ be the MLE of θ based on the assumed model M, and g_t be the conditional distribution of Y_t given $Y_1, Y_2, \dots, Y_{t-1}$ under the model M. White [1984] shows that asymptotically $\sqrt{n} I_n^{*-1/2} H_n^* (\widehat{\theta_n} - \theta^*) \sim N(0, 1)$, where

$$\begin{aligned}H_n^* &= E\left[n^{-1} \sum_{t=1}^{n} \frac{\partial^2 \log g_t\left(Y_t | Y_1, \dots, Y_{t-1}, X, \theta_n^*\right)}{\partial {\theta_n^*}^2}\right] \\ \text{and} & \\ I_n^* &= Var\left[n^{-1/2} \sum_{t=1}^{n} \frac{\partial \log g_t\left(Y_t | Y_1, \dots, Y_{t-1}, X, \theta_n^*\right)}{\partial \theta_n^*}\right].\end{aligned} \tag{2.9}$$

Under mild conditions, White [1984] shows that $\widehat{H_n}$ and $\widehat{I_n}$ are consistent estimators of H_n and I_n, where

$$\widehat{H_n} = n^{-1} \sum_{t=1}^{n} \left. \frac{\partial^2 \log g_t \left(Y_t | Y_1, \ldots, Y_{t-1}, X, \theta_n\right)}{\partial \theta_n} \right|_{\theta_n = \widehat{\theta_n}}$$

and

$$\begin{aligned} \widehat{I_n} &= n^{-1} \sum_{t=1}^{n} \frac{\partial \log g_t \left(Y_t | Y_1, \ldots, Y_{t-1}, X, \theta_n\right)}{\partial \theta_n} \frac{\partial \log g_t \left(Y_t | Y_1, \ldots, Y_{t-1}, X, \theta_n\right)^t}{\partial \theta_n} \\ &+ n^{-1} \sum_{\tau=1}^{\ell} \sum_{t=\tau+1}^{n} \left\{ \frac{\partial \log g_t \left(Y_t | Y_1, \ldots, Y_{t-1}, X, \theta_n\right)}{\partial \theta_n} \frac{\partial \log g_t \left(Y_{t-\tau} | Y_1, \ldots, Y_{t-\tau-1}, X, \theta_n\right)^t}{\partial \theta_n} \right. \\ &+ \left. \left. \frac{\partial \log g_t \left(Y_t | Y_1, \ldots, Y_{t-1}, X, \theta_n\right)^t}{\partial \theta_n} \frac{\partial \log g_t \left(Y_{t-\tau} | Y_1, \ldots, Y_{t-\tau-1}, X, \theta_n\right)}{\partial \theta_n} \right\} \right|_{\theta_n = \widehat{\theta_n}} . \quad (2.10) \end{aligned}$$

White [1984] suggests using $\ell < n^{1/3}$.

To approximate the SE of the HMM2 estimator, we use the algorithm of Lystig and Hughes [2002] to compute partial derivatives of the conditional pmf, g_t, in $\widehat{H_n}$ and $\widehat{I_n}$ efficiently (g_t is equivalent to $\Lambda_t = P(Y_t | Y_1, \ldots, Y_{t-1})$ in the notation of Lystig and Hughes [2002]).

For the GLM estimator, DDW propose a SE that depends on well-behaved estimates of the latent process' moments, which are not easy to obtain. As an alternative, since the GLM is a specific case of the HMM, we can use (2.10) to obtain a SE of the GLM estimator that is independent of the latent process' parameters. Evaluating (2.10) in the GLM case is straightforward since $\widehat{H_n}$ is the usual estimated covariance matrix of the GLM estimator when the GLM is the true model, and the elements of $\widehat{I_n}$ are products of the first derivatives of the Poisson GLM log-likelihood function.

Equation 2.10 can in fact be used to obtain SEs for all of our estimators (and for the FMM and HMM estimators based on any number of hidden states), since they are all special cases of the HMM.

2.5 Performance of the estimators and their standard errors for special cases of the true model

This section details a simulation study of the three estimators described in §2.3: the GLM, FMM2, and HMM2 estimators. We consider certain true models in the class (2.1) to illustrate our key points, and report our general results concerning models in this class in §2.6.

The true models considered in this secion include GLMMs and HMMs, with parameters chosen to achieve different degrees of variation and autocorrelation in the observations (see (2.4) and (2.6)). Specifically, we consider values of σ_α^2 that lead to a variety of overdispersion

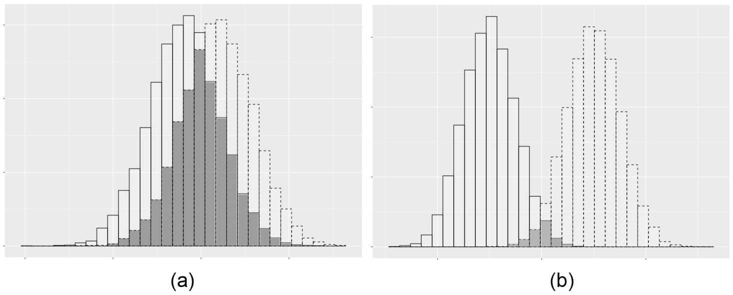

Figure 2.1: Two distributions with low (a) and high (b) separation probability (shaded area).

(OD) factors, defined as

$$\begin{aligned} OD_t &\equiv \frac{Var(Y_t)}{E(Y_t)} \\ &= 1 + \sigma_\alpha^2 \mu_t \end{aligned} \tag{2.11}$$

Likewise, we choose a variety of values for $Corr(Y_t, Y_s)$ (emphasizing in particular $Corr(Y_t, Y_{t-1})$, which we call AC1) in order to investigate the impact of the autocorrelation on the performance of our estimators.

In addition, we consider a third property of the data, which we call “separation probability” (SP). Specifically, in the case where the latent variable takes on k different values, we have k different conditional distributions for the observations. For each t, we define $SP_{S_j,S_{j+1}}$ as one minus the overlap probability of each “adjacent” pair of conditional distributions, i.e., if $S_1 < \cdots < S_k$, we have the following $k-1$ values of SP:

$$SP_{S_j,S_{j+1}} = 1 - \sum_{i=0}^{\infty} \min\{P\left(Y_t = i \mid \alpha_t = S_j\right), P\left(Y_t = i \mid \alpha_t = S_{j+1}\right)\} \tag{2.12}$$

$j = 1, \ldots, k-1$. SP represents the closeness of the conditional Poisson distributions. Figure 2.1 shows two examples with low (a) and high (b) SP. In the case where the latent variable is continuous (e.g., the Poisson GLMM described in §2.3), we define $SP = 0$.

In summary, we consider three data properties (OD, AC1 and SP) as factors that may affect the efficiency of our estimators. In our simulation studies, we choose the parameters of the true models to achieve different levels of these factors, as described in table 2.1. AC1 and SP lie between 0 and 1 (we consider only positive values for AC1 since we are interested only in the effect of its magnitude). OD, on the other hand, can be any value greater than one. We choose the levels of OD and AC1 based on real data sets (the ones used in this paper and those provided by Weib 2009, Zhu 2012, and Liboschik et al. 2017). We define low, medium, and high levels of SP relative to the boundaries of the [0, 1] interval. Note that OD, AC1, and SP cannot always be manipulated separately. For instance, in the Poisson

HMM with $k = 2$ and a fixed transition probability matrix, all the three data properties are increasing functions of the first hidden state, S_1.

We focus on two main cases. In study 1, the true models are chosen such that the MLEs are easy to compute. We then compare the sample bias (i.e., the average value of $\beta_i - \widehat{\beta}_i$ across replicates for each i) and sample variances (SVs) of our estimates to those of the MLEs of the true models. In this way, we get a sense of how much information we lose by fitting a misspecified model (under the assumption that the MLE is the most efficient estimator). In study 2, the MLE of the true models are too expensive to compute. Therefore, we simply compare the bias and SVs of our estimators (which are all based on misspecified models).

In our studies, we consider covariates of different forms, including binary, normally distributed, and a trend. We investigate two sample sizes, $n = 100$ and $n = 1000$. We fix $\beta_0 = 2$ and $\beta_1 = 0.5$. We tried other values for the regression coefficients, but they seemed to have limited effect on the bias and efficiency results, as long as the levels of the three factors remain approximately constant. Therefore, for simplicity, we fix these values and change only the values of the other parameters to achieve different levels of the three factors of interest. We generate 4000 replicates for each run.

2.5.1 Study 1

We first consider the case where the true model is a stationary Poisson HMM with $K = 2$ hidden states. The 2-state HMM normally has 3 free parameters (say S_1, p_{11}, and p_{22}). However in this study, for simplicity, we fix the elements of the transition probability matrix at $p_{11} = p_{22} = 0.9$ and vary only the value of S_1. We use a binary covariate.

The simulation results show that all the estimators are approximately unbiased even for $n = 100$; the magnitude of the estimated bias was always less than or equal to 0.005, i.e., very small relative to the true values of β_0 and β_1 (see online material, table SM 2). Figure 2.2 shows the ratios of the SVs of the HMM2 estimator (the MLE of the true model, in this case) to those of the GLM and FMM2 estimators for different sample sizes and different levels of the factors. The GLM and FMM2 estimators perform well at the low level of the factors where the true model is close to a GLM. Otherwise, the HMM2 estimator outperforms the GLM and FMM2 estimators. The efficiency of the GLM and FMM2 estimators relative to the HMM2 estimator is lower at the high level of the factors, where relative efficiency can be as low as 55%. However, based on the real data sets we have examined, we expect the levels of the factors in practice to be less than our high level. The trends in SV ratios for different sample sizes are similar.

2.5.2 Study 2

In this study, we investigate the performance of the GLM, FMM2, and HMM2 estimators when the true model is complex. In particular, we consider the case where the data gen-

Table 2.1: Factor levels in simulation study

	Low	Medium	High
OD*	1.5	3	5
AC1*	0.15	0.25	0.5
SP*+	0.25	0.45	0.7

* Averaged over the covariate values
+ This factor is 0 for GLMMs

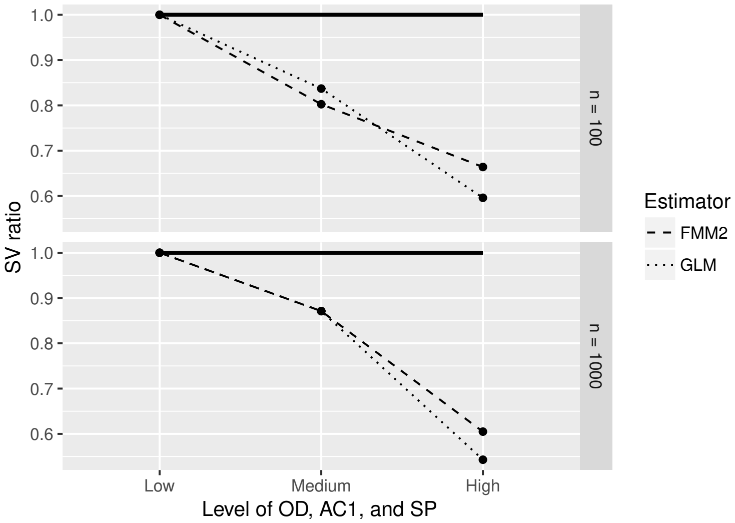

Figure 2.2: Ratio of the SVs of the HMM estimator (the MLE) of the slope to that of the GLM and FMM estimators when a 2-state Poisson HMM is the true model. The solid black line indicates where the efficiency of the estimators would be equal to that of the MLE of the true model. All three factors are simultaneously set to the levels indicated on the x-axis.

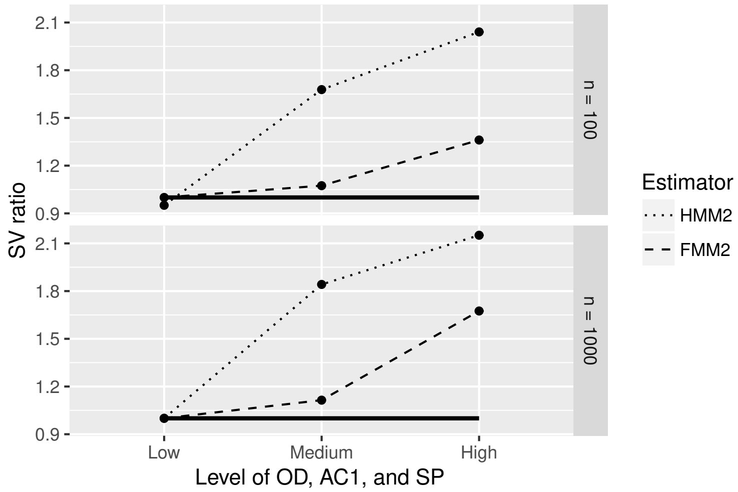

Figure 2.3: Ratio of the SVs of the GLM estimator to those of the HMM and FMM estimators when the 4-state Poisson HMM is the true model. The solid black line indicates where the efficiency of the estimators would be equal to that of the GLM. All three factors are simultaneously set to the levels indicated on the x-axis.

erating mechanism is a Poisson HMM with $K = 3$ or $K = 4$, or a Poisson GLMM. The 3-state and 4-state HMMs have $d + 8$ and $d + 15$ free parameters, respectively, and the likelihood associated with the GLMM consists of an n-dimensional integral. For these reasons, computing the MLEs of the parameters of these models and their SEs is computationally challenging.

As in study 1, we vary the S_j's and fix the elements of the transition probability matrices at $p_{ii} = 0.9$ and $p_{ij} = 0.1/(k-1)$, where $i \neq j$. We use different forms of covariates, including binary, seasonal, and trend, in this study. We choose the values of the parameters in the true models to achieve different levels of the factors described in Table 2.1 (except that SP is fixed at 0 when the Poisson GLMM is the true model).

Our results show that all the estimators are approximately unbiased; the magnitude of the estimated bias was always less than or equal to 0.005 (see online materials, tables SM 3-6 and 8-12). When the Poisson HMM (with 3 or more hidden states) is the true model but the number of hidden states is assumed to be 2, again the GLM and FMM2 estimators perform well at the low level of the factors. The HMM2 estimator outperforms all the other estimators (see fig. 2.3) and the FMM2 estimator has the second lowest SV at higher levels of the factors. The results are similar when the true model is a 3-state Poisson HMM (see online supplementary material, table SM 3).

In contrast, when the true model is a Poisson GLMM (especially with a trend covariate), the efficiency of the GLM estimator is at least as good as those of the HMM2 and FMM2 estimators (fig. 2.4).

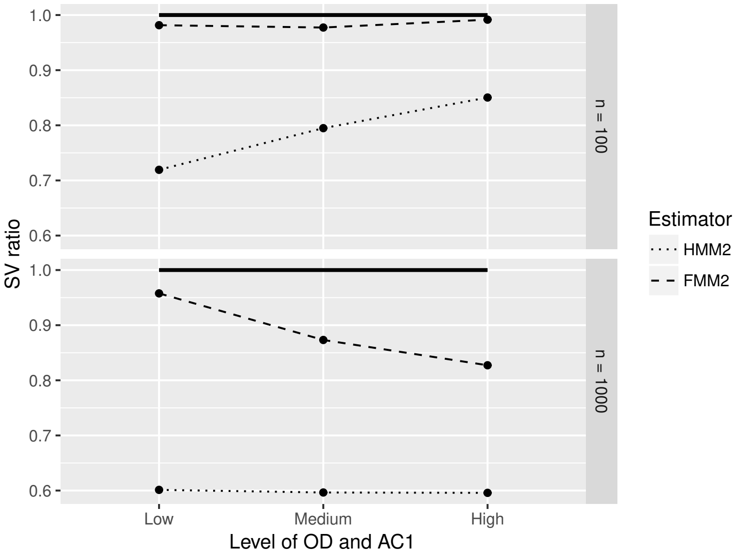

Figure 2.4: Ratio of the SVs of the GLM estimator of the slope to those of the HMM and FMM estimators when the Poisson GLMM is the true model. The solid black line indicates where the efficiency of the estimators would be equal to that of the GLM. OD and AC1 are simultaneously set to the levels indicated on the x-axis. SP is treated as 0.

We also looked at the performance of the 3-state Poisson HMM and FMM estimators in our simulation study. However, because these estimators were poorly behaved (even for $n = 1000$), we have omitted the details of these findings.

2.5.3 Sample SD and SE of the estimators

In this section, we use a simulation study to evaluate the performance of the SEs that we developed in §2.4. We use the same true models considered in studies 1 and 2. We assess the SEs of only the GLM and HMM2 estimators, i.e., the most efficient estimators identified in this context. We treat the sample SDs (SSDs) of the estimates (across replicates) as the true SDs for the purpose of this section, and then compare the SSDs to the average SEs of the estimators (computed as proposed in §2.4). We choose $\ell = 1$ in (2.10) after having experimented with different values.

Table 2.2 shows the SSD and the average SE of the GLM and HMM2 estimators of β_1 for the high level of the factors, $n = 1000$, and one covariate (binary or trend). The results are similar for other more moderate levels of the factors and for $n = 100$ (see online material, tables SM 21-22). In addition, we compare the average values of the DDW and White [1984] SEs of the GLM estimator. Both SEs are quite accurate when we have one binary covariate in true model, regardless of the latent process distribution. When the true model has a trend, both underestimate the true SE, but our SE based on the method of White [1984] always performs better than that of DDW, sometimes dramatically. (See online material for other cases, tables SM 21-22, 25-26, and 28.)

Table 2.2: Performance of the SEs of the estimators based on different (misspecified) models

		HMM Estimator		GLM Estimator		
True model	Covariate	Sample SD	Asy SE (White)	Sample SD	Asy SE (DDW)	Asy SE (White)
Poisson 2-state HMM	Binary	0.012	0.012	0.025	0.025	0.026
Poisson 2-state HMM	Trend	0.012	0.012	0.068	0.034	0.039
Poisson GLMM	Binary	0.030	0.030	0.025	0.024	0.025
Poisson GLMM	Trend	0.054	0.046	0.040	0.019	0.033

2.6 General results concerning the estimators and their standard errors

Our simulation studies in the previous section illustrate the effect of three factors (OD, AC1 and SP) on the performance of the HMM2, FMM2 and GLM estimators and their SEs for certain true models that have a single covariate. In this section, we report on the performance these estimators more generally, i.e., for any choice of true model in class (2.1). We study the effect of including multiple covariates as well as the effect of the level of OD, AC1, and SP. With respect to the latter, of the 27 possible combinations of levels of these factors, we study 22. (See online material for details of each setting, table SM 16.) Due to constraints inherent to models in class (2.1), we are not able to simulate data for five combinations, e.g. low OD and high AC1 and SP. We call each combination of factor levels a run.

Table 2.3 shows a summary of the most efficient estimator (indicated by $*$) for each run where the model had a single covariate. All the estimators are approximately unbiased. In this table, we include only the most extreme runs where the GLM estimator is the most efficient estimator (since models associated with the less extreme runs are even closer to the GLM), and the least extreme runs where the HMM2 estimator is the most efficient estimator (since models associated with more extreme runs are even closer to the HMM). See online material (table SM 16) for results from other runs. In general, we recommend the GLM estimator as a consistent, efficient and robust estimator. The HMM2 estimator performs better only if the SP factor is at its high level, or the SP and AC1 factors are both at at least their medium level. In other words, when the data arise from model (2.1) with a latent variable that is either continuous or takes on closely spaced values, we recommend the GLM estimator. The HMM2 estimator is appropriate only when the latent variable takes on highly separated values. We never recommend the FMM2 estimator because, even when the FMM is the true model, the FMM2 estimator is not substantially more efficient than the GLM estimator in the cases we considered (see online materials, table SM 1). Interestingly, the OD factor doesn't have a big impact on the efficiency of the estimators.

Table 2.3: Most efficient estimators for different levels of the factors when the model has one covariate

	Factor			Estimator	
Run	OD	AC1	SP	GLM	HMM2
1	High	Medium	Low	*	
2	Medium	High	Low	*	
3	High	Low	Medium	*	
4	Low	Medium	Medium		*
5	Medium	High	Medium		*
6	Medium	Low	High		*

We also conducted studies where the true model had multiple covariates of different forms. Our results show that the number and form of covariates affect the efficiency of the estimators (see online materials, tables SM 17-28). For instance, estimating a trend effect is challenging and the chance of underestimating its SE is high when using either the HMM2 or GLM estimator, especially when other covariates are present in the model. In general, the SE of the GLM estimator is better-behaved than that of the HMM2 estimator when we have many covariates in the model. In particular, when the HMM2 estimator is more efficient (e.g. runs 4-6 in table 2.3), the SE of the HMM2 estimator can be severely negatively biased (as low as 50% of the SSD), whereas the SE of the GLM estimator is approximately unbiased (except the estimator of the coefficient of the trend, where both SEs are negatively biased). Therefore, we recommend the GLM estimator, regardless of the levels of the factors in table 2.1.

In practice, when the response is a function of one covariate, the factors in table 2.1 can be estimated. Thus, the problem can be classified according to the runs in table 2.3 and the best estimator identified. However, when we have multiple covariates in the model, estimating factors described in table 2.1 could be challenging. As an alternative, we can fit a GLM to the observed counts, compute the standardized residuals, and then estimate the factors for these residuals. In this way, we can classify models with any number of covariates according to table2.3.

To summarize, we present some guidelines for choosing the "best" estimator (i.e., the estimator with high relative efficiency and well-behaved SE) among the three estimators in practice (where the true model is unknown). In particular, we recommend the HMM2 estimator only when the model has fewer than 4 covariates and SP is at its high level (or SP and AC1 are both at their medium level). Otherwise, in light of its consistency, efficiency, robustness, and well-behaved SE estimate, we recommend the GLM estimator.

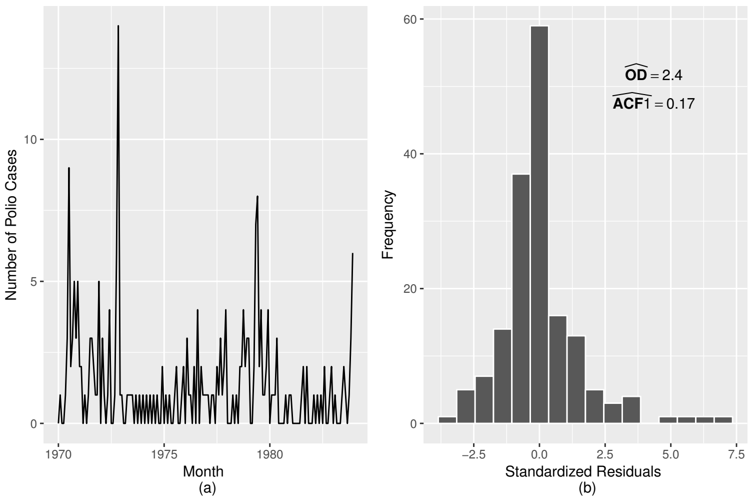

Figure 2.5: (a) Time series plot of polio data and (b) histogram of standardized residuals obtained after fitting the Poisson GLM

2.7 Application

We now apply our findings from the previous sections to real applications. We first re-analyze the polio data considered by Zeger [1988] and DDW. We then present a second application concerning the daily numbers of epileptic seizures.

2.7.1 Polio Incidence Data

Zeger [1988] analyzed the monthly number of cases of polio reported by the U.S. Centers for Disease Control from 1970 - 1983 (168 months) to assess whether the data provided evidence of a long-term decrease in the rate of infection. His analysis was based on a QMLE of the time trend. The data are displayed in fig. 2.5 (a). The polio counts display overdispersion relative to the Poisson distribution ($\widehat{OD} = 2.40$). To compare our approach with that of DDW, we consider the same set of covariates.

No separation among latent distribution is visible (see the Poisson GLM residuals in fig. 2.5 (b)), estimated AC1 and OD of the Poisson GLM standardized residuals are low (0.17) and medium (2.40), respectively, suggesting that the true data generating mechanism is less extreme than that in run 1 in table 2.3. In addition, we have multiple covariates, including a trend. Therefore, the GLM estimator is the best choice in terms of efficiency and accuracy of its SE.

Table 2.4 shows the GLM estimates of the regression parameters along with their SEs based on our approach and on those of DDW. The SEs of all the covariate coefficient estimates (especially the trend) vary considerably across methods. To give more context for these differences, DDW reported the SEs of the GLM estimates (Table 1 of DDW and third column in table 2.4) based on some additional model assumptions (including the assumption

Table 2.4: GLM estimates and SEs (polio dataset). The values in the third column are DDW's SEs, computed after making additional assumptions about the model for the latent process. The values in the fourth column are also DDW's SEs, but computed without these assumptions. The values in the fifth column are our SEs, based on the approach of White [1984].

Covariate	Est	Asy SE (DDW - extra assumptions)	Asy SE (DDW)	Asy SE (White)
Intercept	0.207	0.205	0.040	0.112
Trend $\times 10^{-3}$	-4.799	4.115	1.788	2.548
$cos(2\pi t/12)$	-0.149	0.157	0.009	0.136
$sin(2\pi t/12)$	-0.532	0.168	0.082	0.191
$cos(4\pi t/12)$	0.169	0.122	0.035	0.149
$sin(4\pi t/12)$	-0.432	0.125	0.046	0.149

that the latent process follows an AR(1) model; see §4 of Zeger 1988). They didn't offer a way to check these assumptions and we can't verify them easily. We therefore focus on SEs computed without making further assumptions. Specifically, we compute DDW's SEs of the GLM estimates using method of moment estimates of the latent process covariances (see DDW and Zeger 1988). These values (fourth column of table 2.4) differ substantially from the model-based values in the third column. We recommend using our SE based on the approach of White [1984] (fifth column of table 2.4) because it performs better in general (according to our simulation study). Importantly, the trend based on our estimated SE is not significant (at $\alpha = 0.05$), whereas it is significant based on that of DDW. This difference in conclusions is notable; given the poor performance of DDW's SE, particularly in the case of a trend, presumably the conclusion based on our SE (non-significance) is the more reliable of the two.

2.7.2 Daily number of epileptic seizures

The second application that we consider is a series of counts of myoclonic seizures suffered by one patient on 204 consecutive days. In the neurology literature, Poisson HMMs appear to be common for the analysis of seizure counts (see e.g. Hopkins et al. 1985 or Franke and Seligmann 1993). In particular, Albert [1991] and Le et al. [1992] fit a Poisson 2-state HMM to these counts. Since other predictor variables for this patient are not available, we consider only a trend effect in the model.

The data are illustrated in fig. 2.6 (a). We can see medium estimated AC1 and high estimated OD in the GLM residuals. In addition, the histogram of the Poisson GLM standardized residuals suggests at least two components (i.e., the true model might be a Poisson HMM), but with low SP (fig. 2.6 (b)). Since the properties of the residuals are similar to those of run 1 in table 2.3, we recommend the GLM estimator. Table 2.5 shows the esti-

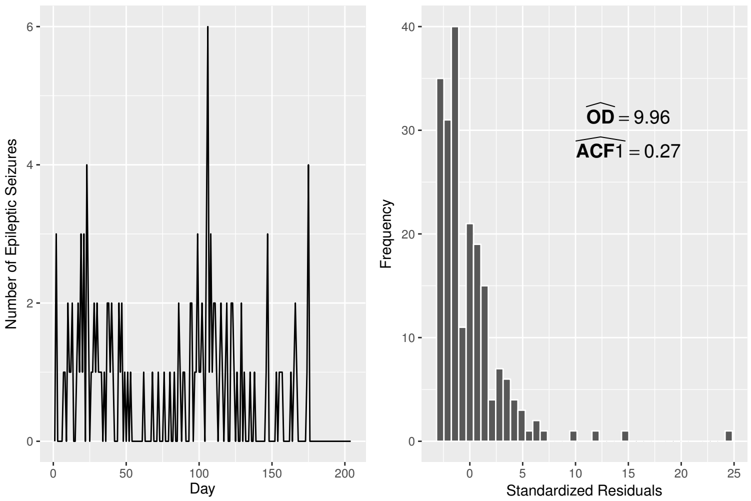

Figure 2.6: (a) Time series plot of epileptic seizure counts and (b) histogram of standardized residuals obtained after fitting the Poisson GLM

Table 2.5: GLM and HMM estimates and SEs (seizure dataset).

	GLM Estimate			HMM Estimate	
Covariate	Est	Asy SE (DDW)	Asy SE (White)	Est	Asy SE (White)
Intercept	0.179	0.481	0.42	-0.170	0.283
Day	-0.006	0.363	0.172	-0.933	0.820

mated coefficients and their SEs. As expected, the GLM estimator does have smaller SEs than does the HMM2 estimator.

These results are consistent with our findings that the HMM2 estimator is more efficient only when the true model is an “extreme” HMM.

2.8 Discussion

In this paper, we considered a general class of models for time series of counts (2.1). We conducted a comprehensive study of the accuracy and efficiency of three estimators, the GLM, FMM2, and HMM2 estimators, and the accuracy of their SEs.

Our results showed that except in extreme cases, the GLM estimator is the most efficient. In addition, the GLM estimator is consistent and robust, has a well-behaved estimated SE, and is easy to compute using standard software (although computing its SE requires specialized code). The GLM estimator has particular advantages over the MLE in the usual case where the distribution of the latent variable is unknown (and may be complex) since specification of this distribution is unnecessary for the purposes of computing both the estimate and its SE.

We considered other estimators including the 3-state FMM and HMM estimators, negative binomial (Davis and Wu 2009), and Zeger's QMLE, as well, but they were less efficient than the HMM2 estimator, or poorly behaved, or their SEs were poorly behaved (see §2.5).

Initially, we thought of the HMM with x number of hidden states (HMMx) as an approximation to the GLMM. Thus, we expected that the HMMx estimator (for $x > 2$) would perform better than the GLM or HMM2 estimator when the true model was a GLMM, HMM3, HMM4, etc. But that was not the case, at least for the sample sizes we considered. Interesting future work could include the exploration of complex true models where the HMMx estimator does perform well.

We proposed three main factors (OD, AC1 and SP) associated with the true model that could affect the performance of the estimators. We then considered a broad set of simulation runs (including extreme cases) by changing the levels of our factors. Surprisingly, OD, which is an obvious violation of the assumption of a Poisson GLM, had limited effect on the relative efficiency of the estimators – even the Poisson GLM estimator. In other words, the extra variability impacts the variance of all estimators, but seems to inflate all variances similarly.

In addition, we developed SEs for our estimators. To the best of our knowledge, our SEs for the Poisson HMM and FMM estimators are unique. Our SE for the Poisson GLM estimator is easier to compute and more accurate than the existing SEs. Moreover, we showed in §2.7 that we might obtain different (possibly wrong) conclusions by using less accurate SEs.

We focused on the efficiency of the covariate coefficient estimators in this paper. However, in our simulation studies, we also considered the intercept estimators. Our results show that all our intercept estimators are approximately unbiased. Both DDW's and our SEs underestimate the true SD of both the GLM and HMM estimators. But, in general, our SE is relatively closer to the SSD of the GLM intercept estimates. Developing a SE that is approximately unbiased for all the coefficient estimators (including the intercept), is a topic for future work.

Chapter 3

Display advertising: Estimating conversion probability efficiently

The chapter derives, with few modifications, from:
Safari, A., Altman, R. M., Loughin, T. M., 2018. *Display advertising: Estimating conversion probability efficiently*, In preparation.

3.1 Abstract

The goal of online display advertising is to entice users to "convert" (i.e., take a pre-defined action such as making a purchase) after clicking on the ad. An important measure of the value of an ad is the probability of conversion. The focus of this paper is the development of a computationally efficient, accurate, and precise estimator of conversion probability. The challenges associated with this estimation problem are the delays in observing conversions and the size of the data set (both number of observations and number of predictors). Two models have previously been considered as a basis for estimation: A logistic regression model and a joint model for observed conversion statuses and delay times. Fitting the former is simple, but ignoring the delays in conversion leads to an under-estimate of conversion probability. On the other hand, the latter is less biased but computationally expensive to fit. Our proposed estimator is a compromise between these two estimators. We apply our results to a data set from Criteo, a commerce marketing company that personalizes online display advertisements for users.

3.2 Introduction

Display advertising is a relatively new type of online advertisement where advertisers pay publishers to present their ads (also known as impressions) on different webpages. Depending on the purpose of the advertisement, different payment options can be used. These options include cost per impression, where the advertisers pay the publishers to display their ads (whether the user clicks the ad or not), cost per click, where the advertisers pays for an impression only if a user clicks on it, and cost per action (CPA), where advertisers pay only if the user takes a predefined action (conversion) after clicking the ad, such as purchasing a product or service Muthukrishnan [2009], Chapelle [2014].

For profitability, the CPA option requires that publishers make a “good” match between advertisers and customers. In particular, they should display ads with high expected earnings per impression, i.e., ads where the customer’s probability of clicking and the subsequent probability of the click’s converting are high. McAfee [2011]. The entire process of ad selection needs to be completed in the time between when a user opens a page and when the page is fully rendered. Thus, the publisher has a very short time in which to choose which ad(s) to display to the user. Great progress has been made predicting whether a user will click on an impression in the context of search advertising (see for example Hillard et al. [2010], or McMahan et al. [2013]) and display advertising (see for example Chapelle et al. [2014], or Agarwal et al. [2010]). However, little is known about estimating the probability of conversion. For instance, Rosales et al. [2012] perform an experimental analysis (on a private Yahoo data set) to show the advantage of conversion probability over the click probability as a measure of profitability in display advertising, and point out the lack of inference about this new measurement in the literature.

The main issue in conversion probability estimation is the delay between the click and the eventual conversion status of the click (called the conversion delay), which can vary from a few milliseconds to months. In other words, eventual conversion status (converted or unconverted) is unknown for clicks where the conversion delay is censored. Chapelle [2014] proposed using the maximum likelihood estimator (MLE) of the conversion probability based on a delay feedback model (DFM), a mixture model for observed conversion status that depends on the delay distribution. Although his estimator is accurate when the model is correctly specified, his approach is not computationally efficient. Efficiency is critical in this Big Data setting where publishers need to re-estimate conversion probability rapidly and frequently as time progresses and more data accrue (i.e., real-time updating). In addition, the performance of his estimator is unknown when the delay distribution is not exponential.

Our goal in this paper is to develop a method for estimating probability of conversion with high accuracy and in a computationally efficient manner. In particular, we introduce a new estimator based on the logistic regression model that (wrongly) treats all conversion statuses as known, and then reduce the bias of this estimator through a novel application of the Kullback-Leibler distance. We evaluate the accuracy and computational efficiency of this new estimator compared to those of Chapelle’s estimator. In addition, we study the performance of these estimators when the delay distribution is misspecified.

In §3.3 we define some notation and present the DFM of Chapelle [2014]. In §3.4, we introduce our estimator along with an algorithm to evaluate it efficiently for a given data set. Section §3.5 presents an application of our results to a data set released by Criteo Chapelle

[2014], and §3.6 describes a simulation study that illustrates the accuracy, precision, and computational efficiency of the estimators.

3.3 Model specification

In this section, we describe the DFM developed by Chapelle [2014]. The assumptions of the DFM (and of our methods that follow) are: **i.** the true conversion probability is fixed over time, **ii.** the predictors don't depend on time, **iii.** a converted click can never become unconverted, **iv.** an unconverted click with delay time less than a fixed time period (the conversion window) can convert, **v.** an unconverted click with delay time greater than the conversion window cannot convert (in other words, an unconverted click can convert any time within the conversion window, W), and **vi.** W is long enough that only a negligible proportion of conversions occur outside this window.

Let the data collection start at time 0. Label clicks sequentially in time as $1, 2, 3, \ldots$. Let $t_{i,0}$ be the time of click i – treated as non-random for the purposes of this paper. Throughout, we use bold letters to denote vectors. For instance, $\boldsymbol{x_i} \equiv (x_{i,1}, \ldots, x_{i,k})$ is a $1 \times k$ vector of covariates associated with click i, e.g., attributes of the user and/or origin website. We define $x_{i,1} = 1 \; \forall i$ so as to include an intercept. Define C_i to be the eventual conversion status indicator for click i, i.e. $C_i = 1$ if click i ever converts and $C_i = 0$ otherwise. Let T_i^c be the time at conversion if $C_i = 1$; if $C_i = 0$, then fix $T_i^c = t_{i,0} + W$. Then the delay time D_i is defined as $D_i = T_i^c - t_{i,0}$ (so that $D_i = W$ if $C_i = 0$). Given $\boldsymbol{x_i}$ and $C_i = 1$, let $h_i(d) = h(d \mid \boldsymbol{x_i}, C_i = 1)$ and $H_i(d) = H(d \mid \boldsymbol{x_i}, C_i = 1)$ be the conditional pdf and cdf, respectively, of D_i.

Now suppose that at a given moment $t > 0$ we wish to estimate the conversion probability of a click with covariates $\boldsymbol{x_i}$. Define $a_i(t) = \min\{t - t_{i,0}, W\}$ to be the age of click i. Since we treat $t_{i,0}$ as non-random, $a_i(t)$ is non-random as well.

For subsequent derivations, we consider a given fixed time t and suppress t in our notation for convenience. At this time, say n clicks have accumulated. Let Y_i be the current conversion status indicator of click $i = 1, \ldots, n$, i.e., $Y_i = 1$ if click i converted prior to time t and $Y_i = 0$ otherwise. Note that D_i is observed prior to t if $Y_i = 1$, and is greater than or equal to a_i (right censored) if $Y_i = 0$.

To the best of our knowledge, the DFM is the only model for conversion probability in the literature that incorporates conversion delays, i.e., that is based on the bivariate response for each click, (Y_i, D_i). In this model, C_i is assumed to follow a logistic regression model with $p_i = P(C_i = 1|\boldsymbol{x_i}) = \frac{exp(\boldsymbol{\beta_c'}\boldsymbol{x_i})}{1+exp(\boldsymbol{\beta_c'}\boldsymbol{x_i})}$. Given $C_i = 1$ and $\boldsymbol{x_i}$, delay times are assumed to follow an exponential distribution with rate $\lambda_i = \exp(\boldsymbol{\beta_d'}\boldsymbol{x_i})$. The log-likelihood function of the DFM is then

$$\begin{aligned}\ell\left(\boldsymbol{\beta}_c, \boldsymbol{\beta}_d | \boldsymbol{y}, \boldsymbol{d}\right) = \; & - \sum_{i:y_i=1} \left\{\log\left[P(C_i = 1|\boldsymbol{x_i})\right] + \log(\lambda_i) - \lambda_i d_i\right\} \\ & - \sum_{i:y_i=0} \log\left[1 - P(C_i = 1|\boldsymbol{x_i}) + P(C_i = 1|\boldsymbol{x_i})\exp(-\lambda_i a_i)\right]\end{aligned}$$

For later derivations in this paper, we will require a different form for ℓ. Specifically, we define Z_i as

$$Z_i(a_i) \equiv Z_i := \min(D_i, a_i), \tag{3.1}$$

so that $0 \leq Z_i \leq a_i$. Note that Z_i is a function of a single random variable, D_i. We can define an equivalence relationship between Y_i and Z_i as

$$\begin{aligned} Z_i < a_i &\iff Y_i = 1 \\ Z_i = a_i &\iff Y_i = 0 \end{aligned} \tag{3.2}$$

Then, the likelihood function of the DFM can be rewritten in terms of the z_i's (realizations of the Z_i's) as

$$L_g\left(\boldsymbol{\beta}_c | \boldsymbol{z}\right) = \prod_i \left(p_i h(z_i)\right)^{I(z_i < a_i)} \left(1 - p_i H(z_i)\right)^{I(z_i \geq a_i)}. \tag{3.3}$$

(See Appendix 3.8.1 for the proof.)

3.4 Estimation

As discussed by Chapelle [2014], the likelihood function in (3.3) is non-convex with no closed form for the MLE. Therefore, its optimization is very slow. For this reason, we consider alternative estimators in this section.

3.4.1 Naive estimator

A simple (but misspecified model) for observed conversion status is the logistic regression model where the current conversion statuses of the clicks are treated as their eventual conversion statuses. In other words, conversion delay time (and the possibility that unconverted clicks with age less than W could convert) are ignored. Chapelle [2014] calls this model the "naive model". The likelihood function of this model is

$$\begin{aligned} L_f(\boldsymbol{\alpha}|\boldsymbol{z}) &= f(\boldsymbol{z}|\boldsymbol{\alpha}) \\ &= \prod_{i=1}^{n} f(z_i|\boldsymbol{\alpha}) \\ &= \prod_{i=1}^{n} \left[\theta_i^{I(z_i<a_i)}(1-\theta_i)^{I(z_i\geq a_i)}\right], \end{aligned} \tag{3.4}$$

where $\theta_i = \frac{\exp(\boldsymbol{\alpha}'\boldsymbol{x_i})}{1+\exp(\boldsymbol{\alpha}'\boldsymbol{x_i})}$ is the conversion probability of the i^{th} click and $\boldsymbol{\alpha}$ is a vector of regression coefficients. The likelihood function of the naive model is convex and computationally efficient to optimize. However, the MLE of θ_i is biased low for the true probability of conversion, since some unconverted clicks could convert later.

3.4.2 Bias - adjusted estimator

In this section, we introduce a new estimator to adjust for the bias in the naive estimator. We use the Kullback-Leibler information criterion (KLIC) approach White [1982]. Suppose that $g(\boldsymbol{z}|\boldsymbol{\beta}_c)$ is the true data-generating distribution, but that $f(\boldsymbol{z}|\boldsymbol{\alpha})$ is the assumed model. Then the KLIC can be computed as follows:

$$\begin{aligned} KLIC\,(g|f\ ;\ \boldsymbol{\alpha},\boldsymbol{\beta}_c) &= E_g\left(\ln\left[\frac{g(\boldsymbol{z}|\boldsymbol{\beta}_c)}{f(\boldsymbol{z}|\boldsymbol{\alpha})}\right]\right) \\ &= E_g\left(\ln\left[g(\boldsymbol{z}|\boldsymbol{\beta}_c)\right]\right) - E_g\left(\ln\left[f(\boldsymbol{z}|\boldsymbol{\alpha})\right]\right) \\ &= E_g\left(\ln\left[g(\boldsymbol{z}|\boldsymbol{\beta}_c)\right]\right) - E_g\left(\ln\left[\prod_i f(z_i|\boldsymbol{\alpha})\right]\right) \\ &= E_g\left(\ln\left[g(\boldsymbol{z}|\boldsymbol{\beta}_c)\right]\right) - \sum_i E_g\left(\ln\left[f(z_i|\boldsymbol{\alpha})\right]\right) \\ &= E_g\left(\ln\left[g(\boldsymbol{z}|\boldsymbol{\beta}_c)\right]\right) \\ &- \sum_i \left[p_i \ln(\theta_i) H_i(a_i) + \ln(1-\theta_i)(1-H_i(a_i)\, p_i)\right] \end{aligned}$$

White [1982] shows that the MLE of the parameters in the misspecified model is consistent for the minimizer of the KLIC. We use his results to adjust the naive estimator and remove its asymptotic bias relative to the true model. In other words, we assume that the true model is (3.3) and treat the parameters of the true model, $\boldsymbol{\beta}_c$, as known. Then we minimize the KLIC with respect to the parameters of the misspecified model, $\boldsymbol{\alpha}$, resulting in estimating equations that depend on both $\boldsymbol{\beta}_c$ and the unknown KLIC minimizer, $\widetilde{\boldsymbol{\alpha}}$. We then solve for $\boldsymbol{\beta}_c$.

The details are as follows. First, we have

$$\left.\frac{\partial\ KLIC\,(g|f\ ;\ \boldsymbol{\alpha},\boldsymbol{\beta}_c)}{\partial\alpha_j}\right|_{\widetilde{\boldsymbol{\alpha}}}=0 \tag{3.5}$$

$$\Rightarrow\quad \sum_i p_i H_i(a_i)x_{i,j}=\sum_i x_{i,j}\widetilde{\theta}_i\ ,\quad j=1,\dots,k \tag{3.6}$$

where $\widetilde{\theta}_i=\theta_i|_{\widetilde{\boldsymbol{\alpha}}}$, and k is the number of regression coefficients. Treat H_i and $\widetilde{\boldsymbol{\alpha}}$ as known for the moment. Note that the equations in (3.6) are algebraically equivalent to the weighted quasi-score equations associated with a logistic regression model (with $\widetilde{\theta}_i$ taking the place of the usual response variable). Thus, they can be solved efficiently for $\boldsymbol{\beta}_c$.

In the usual case where H_i and $\widetilde{\boldsymbol{\alpha}}$ are unknown, we plug in consistent estimates. In particular, we compute $\hat{\theta}_i$, the MLE of θ_i from (3.4), which is consistent for $\widetilde{\theta}_i$ (by White's theorem).

To estimate H_i, given the family of distributions of the delay (e.g., exponential), we can find the MLE of the delay distribution parameters. However, since the censoring rate could be very high, especially when t is small, this MLE can be quite biased (see, e.g., Shen and Yang [2015], Wan et al. [2015b], and Hirose [1999]). As a remedy, we can adjust for the delay rate estimator bias as well. Firth [1993] proposes a general approach to bias reduction using on a penalized score function. Pettitt [1998] apply Firth's approach to obtain the penalized likelihood when the responses are exponentially distributed and possibly censored at a fixed censoring time for all the observations. We extend bias adjustment approach of Pettitt [1998] by relaxing fixed censoring time for all the observation (i.e., each click has its own censoring time) to adjust the delay rate estimator for its bias. In other words, using our notation, the penalized likelihood can be written as

$$L^*(\boldsymbol{\lambda}|\boldsymbol{z})=\prod_{i\in S^*}(\lambda_i)^{-2}h_i(z_i)H_i(a_i), \tag{3.7}$$

where $\lambda_i=\exp(\boldsymbol{\beta}_d'\boldsymbol{x_i})$ as before,

$$h_i(z_i)\quad=\quad\begin{cases}\lambda_i\exp(-\lambda_i z_i), & z_i<a_i\\ \exp(-\lambda_i z_i), & z_i=a_i\end{cases}\ , \tag{3.8}$$

$H_i(a_i)=1-\exp(-a_i\lambda_i)$, and $S^*=\{i:\ C_i=1\}$. Note that since in the application, we don't know the eventual conversion status of clicks (especially for recent clicks), we approximate S^* by $\hat{S^*}=\{i:\ y_i=1\}\cup\{i:\ y_i=0,a_i<W\}$, which the approximation improves as time goes on. In other words, we exclude only unconverted clicks with a_i longer than W in (3.7) since we assume they never convert, and thus don't contribute information about the delay distribution.

When the delays follow a Weibull distribution, we cannot obtain a closed form for the Firth [1993] penalized likelihood function. However, if we make the usual assumption that only the scale parameter of the Weibull distribution depends on the covariates, we can obtain the Weibull penalized likelihood function for a fixed shape parameter as

$$L_w^*(\boldsymbol{\gamma}, \nu|\boldsymbol{z}) = \prod_{i \in S^*} \left(\frac{\nu}{\gamma_i}\right)^2 h_i^w(z_i) H_i^w(z_i), \tag{3.9}$$

where h_i^w and H_i^w are the pdf and cdf, respectively, of the Weibull distribution with scale parameter $\gamma_i = \exp(\boldsymbol{\beta}_d^t \boldsymbol{x_i})$ and shape parameter ν. We suggest first estimating the shape parameter, ν, by its MLE, and then treating it as a known parameter in (3.9).

We call the convergence probability estimator based on exponential and Weibull distributions for the delays the E-bias-adjusted and W-bias-adjusted estimators, respectively.

To summarize, we obtain our bias-adjusted estimate of $\boldsymbol{\beta_c}$ as follows:

1. Compute the MLE of $\boldsymbol{\alpha}$ based on the naive model (3.4).

2. Compute the maximum penalized likelihood estimates of the delay distribution parameters using Firth's approach (i.e. (3.7) if the delay distribution is exponential or (3.9) if the delay distribution is Weibull).

3. Compute the bias-adjusted estimate $\hat{\boldsymbol{\beta}}_c$ by solving the equations in (3.6), substituting the estimates of $\boldsymbol{\alpha}$ and the delay distribution for their true values.

Standard GLM software can be used to compute the estimates in Steps 1 and 3, while packages such as brglm in R can be used to compute the estimates in Step 2. Thus, an advantage of the bias-adjusted estimator is that it can be computed efficiently and easily.

The similarity between (3.6) and the weighted quasi-score equations associated with a logistic regression model suggests that a SE for $\hat{\boldsymbol{\beta}}_c$ (or $\hat{p}_i$, the estimator of p_i) could be efficiently computed as a function of the derivative of the left side of (3.5). We explore the validity of this SE in §3.6.

3.5 Application

In this section, we apply our results from the previous sections to a publicly available data set released by Criteo, a commerce marketing company that connects publishers and advertisers[1]. The data concern a collection of clicks that accrued over a period of two months, with $W = 30$ days Chapelle [2014]. The eventual conversion statuses of the clicks are also included in the data set.

[1] The data set is available at http://research.criteo.com/outreach/

In this data set, each row corresponds to a display ad chosen by Criteo and subsequently clicked by the user. The first two columns are click time and conversion time, where the latter is blank for unconverted clicks. The data set has 17 covariates (8 integer-valued and 9 categorical variables). Except for campaign ID (one of the categorical variables), the definitions of the covariates are undisclosed (due to confidentiality issues).

To evaluate the performance of our bias-adjusted estimators (i.e., E-bias-adjusted and W-bias-adjusted estimators) in this section, we investigate their bias, SE, and computation time relative to three other estimators: the **naive** estimator, the **oracle** estimator (the MLE of the logistic regression model based on the eventual conversion statuses of the clicks), and the maximizer of the DFM when the distribution of the delays is treated as exponential (Chapelle's estimator). Note that the oracle estimate is not obtainable in practice, where at any time t, the delay distribution parameters will be unknown. However, we include this estimator as a "gold standard" to which we compare the other estimators.

Following Chapelle [2014], we use log-loss to measure the bias of each estimator. Log-loss is a measure of the distance between parameter estimates and the true quantity of interest. In our case, log-loss is algebraically equivalent to the negative log-likelihood (NLL) of the logistic regression model (treating eventual conversion statuses as the true quantities of interest).

Estimating the parameters in the DFM can be very slow, depending on the number of covariates in the model. For instance, say we choose a subset of the covariates in the full data set such that we have 300 covariate coefficients (corresponding to the continuous covariates and the dummy variables that represent the categorical covariates) in the model. Obtaining the MLE of the DFM is approximately 500 times slower than computing the bias-adjusted estimator. To keep the parameter estimation time feasible in our data analysis, we use only the first 100 covariates from the data set in our analyses in this section. Note that, in this paper, we are interested only in the relative performances of the estimators given a set of covariates; variable selection in this context is an important topic for future research.

Since the data set is huge, we use data splitting and use only a random sample (approximately 10%) of the data set as our training set. We then obtain our estimates on the training set and compute NLL on the rest of the data set (our test set). We repeat this procedure 40 times and report the average of the NLLs.

Figure 3.1 shows the average (over the 40 random splits of the data) NLL of the estimators at different time steps. The DFM estimator has convergence problems, especially when the number of known conversions is not large relative to the number of parameters (i.e., over the first two weeks of the observation period). After excluding the problematic estimates, the DFM estimator still behaves poorly (top plot of figure 3.1). To illustrate the differences among the other estimators better, we omit the DFM estimator from the plot (bottom plot of figure 3.1). The E-bias-adjusted estimator appears to outperform the other estimators. Specifically, the E-bias-adjusted estimator appears to outperform the W-bias-

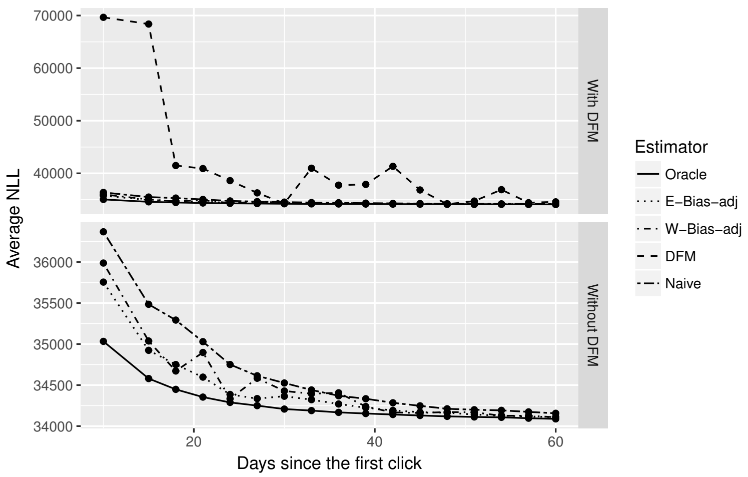

Figure 3.1: Two views of the average NLL of the estimators at different time steps (averaged over 40 random splits of the data): with the DFM estimator (top) and without the DFM estimator (bottom)

Table 3.1: Average computation time (in seconds) of each estimate over repeated data splitting and different time steps. Approximately 100 covariates are included in the model.

Estimator	Run time
Naive	4.67
E-Bias-adjusted	120.67
W-Bias-adjusted	158.30
DFM	2545.82

adjusted in the first month, and they perform similarly in the second month. In addition, as we obtain more new clicks and more information about the old clicks, the NLL of the estimators appears to decrease and get closer to that of the oracle estimator.

Table 3.1 shows the average computation time (in seconds) of the estimates based on repeated data splitting when we have approximately 100 covariates in the model. The computation time of the DFM estimator is about 21 times longer than that of the E-bias-adjusted estimate, and the computation time of W-bias-adjusted estimator is about 30% longer than that of the E-bias-adjusted estimator.

As a final note, the distribution of the observed delays of converted clicks looks closer to Weibull than exponential (see online material and also, e.g., Ji et al. [2016]) and thus the assumption of exponential delay times in the DFM is unreasonable. However, the MLE (i.e., of the maximizer of the model based on a Weibull distribution for the delays) has serious convergence issues and very long computation time. Thus we did not study the performance of this estimator in detail.

3.6 Accuracy, precision and computational efficiency of the bias-adjusted estimators

In this section, we use a simulation study to evaluate the performance of our estimators. We investigate their bias, SE, and computation time. Besides the estimators mentioned in §3.5, since we know the delay distribution in our simulation study, we consider the **true-bias-adjusted** estimator (the bias-adjusted estimator computed using the true cdf of the delay distribution, so that the weights in (3.6) are known). The true-bias-adjusted estimator helps us to gauge how much we lose by estimating the delay distribution parameters with (3.7) (or (3.9)).

We suggest using bias of the estimated probabilities as a measure of error in the simulation study. Average bias at time t is defined as $\frac{1}{n}\sum_i (p_i - \hat{p}_i)$ for an estimator of p_i, $\hat{p}_i$. Recall that the p_i's vary according to covariates; average bias can be interpreted as an estimate of the *marginal bias* of the estimator of probability of conversion (in contrast with $\mathrm{E}[p_i - \hat{p}_i]$, which represents the bias of $\hat{p}_i$ *conditional* on $\boldsymbol{x}_i$).

3.6.1 Simulation study design

Since our focus in this paper is display advertising, we use a real data set (the Criteo data described in Section 3.5) to inform the design of our simulation study. Specifically, we pick approximately 8500 clicks ($n \approx 8500$) from a campaign with a large number of clicks. For this campaign, the average conversion probability was moderate ($\sim 30\%$). We use the covariates values given in the Criteo data set by Chapelle [2014] and keep these values the same across runs. Since for the selected clicks some covariates have only one value (or have only a few values that differ from the mode), we use only three of the categorical variables (resulting in 16 dummy variables) and four of the integer-valued covariates in the original data set.

We conduct two simulation studies. In the first, we generate exponential-distributed conversion delays. In the second, we generate Weibull-distributed conversion delays. The parameters of these distributions are set to their estimated values based on the observed delays in the chosen campaign (using only converted clicks). In other words, the estimated parameters based on the Criteo data become the true parameter values in the simulation study. Similarly, we estimate the regression coefficients of the conversion probability model by fitting a logistic regression to the final conversion status of the clicks in the data set. We then use these estimated coefficients as the true coefficients in the simulation studies (see online material for the covariates and coefficients we use).

We consider two factors affecting the performance of the conversion probability estimators: average conversion probability and average delay time, where average means across all clicks of the campaign. We choose the levels of the factors based on the range of the conversion probabilities and delays in the real data set; see table 3.2 for details. To keep

Table 3.2: Levels of the factors in the simulation studies

Factor	Low	Medium	High
Conversion probability[†]	0.1	0.3	0.6
Delay mean[*]	2	4	7

[†] averaged across all clicks
[*] in days

the simulation study feasible and the number of parameters in the model manageable, we assume no interaction among the covariates – in particular no interaction between campaign and the other covariates. Under this assumption, we can vary the factors of interest (average conversion probability and average delay time) simply by varying the values of the intercepts and of the campaign effects in both the delay and conversion models while keeping the other covariate coefficients (and the shape parameter, in the Weibull case) fixed.

To create a realistic scenario in our simulation studies, we track the clicks since start of the data collection at $t = 0$, and evaluate the estimators at 17 different time steps over a two month period (with time steps spaced far enough apart such that approximately equal numbers of clicks occur in each interval). At each time step t, we consider only clicks that occurred by t. Similarly, we treat a click as converted only if we observe its conversion by t and its age is less than $W = 30$ days. Otherwise, we treat it as unconverted.

3.6.2 Study 1

We first consider the case where the conversion delays follow an exponential distribution. In other words, we generate data from Chapelle's DFM. Thus, the MLE of the DFM and the E-bias-adjusted estimator are based on the correct model.

Figure 3.2 shows the average bias of the estimators over time when both factors (average conversion probability and average delay) are at their medium level. As expected, since the DFM is the true model in this study, its maximizer (the MLE) outperforms all other estimators (except the oracle estimator). In particular, it appears to be less biased than the E-bias-adjusted estimator (especially over the first month). That said, the bias of both estimators seems quite small in the second month (less than 0.007 on average). The true-E-bias-adjusted estimator appears to perform slightly better than the E-bias-adjusted estimator (especially over the first month), and the naive estimator appears to remain biased even after two months by approximately 0.05. The overall trend in bias is similar when we use other levels of the factors given in table 3.2. As expected, when the average delay is at its low level, the accuracy of the naive estimator appears to be almost as high as the other estimators. Moreover, the MLE of the DFM behaves poorly when the average delay is high and average conversion probability is low (see online materials).

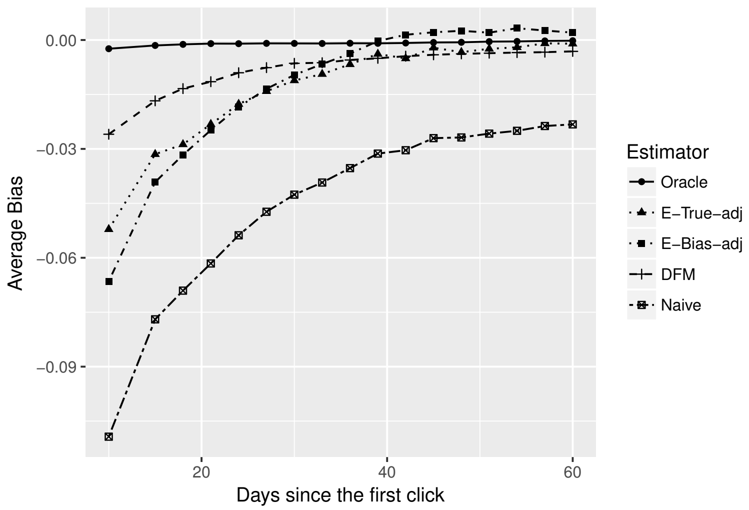

Figure 3.2: Average bias of the estimators over time for the medium level of the factors when the delays follow an exponential distribution

3.6.3 Study 2

In this study, we consider a Weibull distribution for the delays.

We compute all the estimators (including the W-bias-adjusted estimator) over time as in study 1. Note that in this case, the maximizer of the DFM and the E-bias-adjusted estimator are both based on the (misspecified) exponential distribution for the delay times. Thus, the former is no longer the MLE and we call it DFM estimator in this study. We do not study the MLE (i.e., the maximizer of the DFM modified to allow a Weibull distribution for the delays) due to convergence issues and very long computation times. In addition, Since the true-E-bias-adjusted estimator hasn't been derived for this study, we don't consider the estimator here.

Figure 3.3 shows the average bias of the E-bias-adjusted and W-bias-adjusted estimators, along with that of the oracle and naive estimators over time when both factors (average conversion probability and average delay) are at their medium level. Over the first three weeks, the E-bias-adjusted estimator appears to slightly outperform the W-bias-adjusted estimator. However, both estimators perform similarly after the third week. In addition, the computation time of the W-bias-adjusted is approximately 30% more than that of the E-bias-adjusted estimator. Therefore, we consider only the E-bias-adjusted estimator for the remainder of this paper.

Figure 3.4 shows the bias of the estimators over time when both factors, average conversion probability and average delay, are at their medium level. In contrast with study 1, the E-bias-adjusted estimator appears to outperform the DFM estimator. In particular, as time goes on, the bias of the E-bias-adjusted estimator nearly disappears, whereas the bias of the DFM estimator does not. In addition, the bias of the E-bias-adjusted estimator shows that

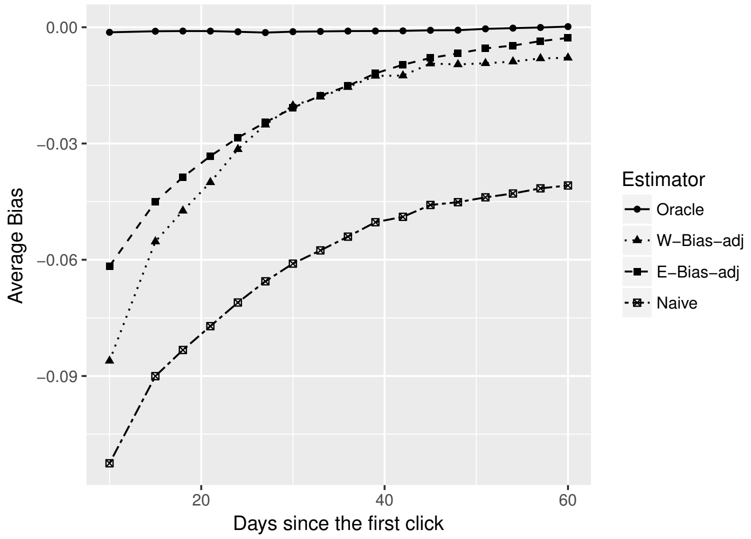

Figure 3.3: Average bias of the bias-adjusted estimators over time for the medium level of the factors when the delays follow a Weibull distribution

the maximum penalized likelihood estimator of the parameters in the delay time model (see (3.7)) performs well even when the delay distribution is misspecified, especially for $t \geq 30$. The trend in bias is similar for other levels of the factors given in table 3.2. Again, when the delay mean is in its low level, the accuracy of the naive estimator is almost as high as the other estimators. Similar to study 1, the DFM estimator behaves poorly when the average delay is high and average conversion probability is low (see online materials).

3.6.4 Coverage probability of the bias-adjusted estimator

As mentioned in §3.4.2, we can efficiently compute a SE for $\hat{p}_i$ as a function of the derivative of the left side of (3.5) by using the delta method (built-in *predict.glm* function in R). In this section, we study the validity of our SE.

Figure 3.5 shows the average coverage probability (CP) associated with 95% confidence intervals for conversion probability based on the E-bias-adjusted estimator over time when the delays follow exponential or Weibull distributions. In the first month, the average CP is below the nominal level (approximately 88%). However, in the second month, the average CP is more than 92%. To show the closeness of the average CP to the nominal value of 0.95 at each time point more carefully, we add the non-rejection region for the score test of whether CP differs from 0.95. This region is defined as $(0.95 - 2\sqrt{0.95(1-0.95)/R}, 0.95 + 2\sqrt{0.95(1-0.95)/R}) \approx (0.94, 0.96)$, where $R = 2000$ is the number of replicates. The CP when the delays are exponential-distributed (so that the E-bias-adjusted estimator is based on the correct model) is not significantly different the nominal coverage level at the last 4 time steps. In contrast, when the delays are Weibull-distributed, the CP differs significantly from 0.95 except at the last time step. In other words, CP is lower when the E-bias-adjusted

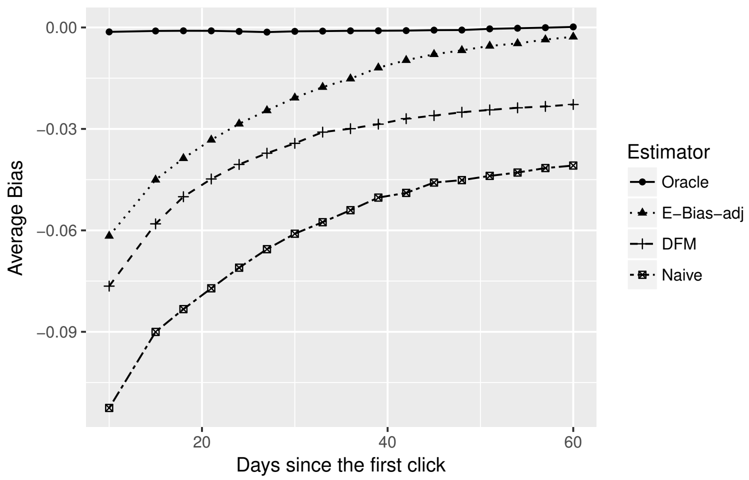

Figure 3.4: Average bias of the estimators over time for the medium level of the factors when the delays follow a Weibull distribution

Table 3.3: Average computation times (in seconds) of each estimator over different time steps along with its SSD for the medium level of the factors (Study 2)

Estimator	Run time average (SSD)
Naive	0.09 (0.04)
Bias-adjusted	5.16 (2.38)
DFM	24.54 (4.02)

estimator is based on a misspecified model, but our results suggests that it converges to 0.95.

To compute the SE (and CP) associated with the DFM estimator, we could compute the Hessian matrix of the estimates for each replicate. However, this matrix was non-positive-definite for most replicates. For this reason, we omit results concerning the DFM estimator. The results for the other runs are similar (see online material).

3.6.5 Computation time

Given the very short time available for choosing an ad and publishing it on the host website – and the huge number of ad requests and new campaigns at any time – publishers need to refit the model and obtain the conversion probabilities frequently. Therefore, computation time is a critical issue in display advertising. Table 3.3 shows the average computation times of the estimates along with their sample standard deviation (SSD) when the true delay distribution is Weibull (Study 2) and the factors are at their medium level. In particular, the computation time of the DFM estimator is more than 5 times that of the E-bias-adjusted estimate. For the levels of the factors that we considered, this ratio can be between 4 and 8. The computation time of the estimates is similar for Study 1.

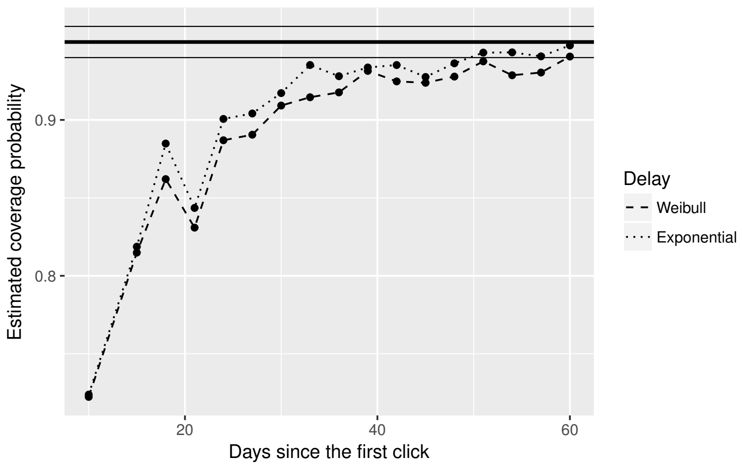

Figure 3.5: Average coverage probability of the 95% CI for conversion probability based on the E-bias-adjusted estimator over time for the medium level of the factors (studies 1 and 2)

3.7 Discussion

In this paper, we developed a method for estimating probability of conversion efficiently and with high accuracy. In particular, we introduced a bias-adjusted estimator based on a simple (misspecified) logistic model, and evaluated its accuracy and computational efficiency.

As an alternative, we could obtain the MLE and bias-adjusted estimators by assuming a Weibull distribution for the delays, which would allow greater flexibility in the model and would, in particular, provide a better description of the the delays in the Criteo data (see online material). However, the MLE of this model suffers from both convergence issues and lengthy computation times. Moreover, the W-bias-adjusted estimator is not as consistent and efficient as the E-bias-adjusted estimator. Therefore, we recommend the E-bias-adjusted estimator even when the delays follow a Weibull distribution.

Since clicks have different associated true probability of conversion, the estimators of these probabilities (and their bias) have different variances. When computing the average of bias, one may account for these differences by weighting each bias by its true SD, especially when the range of the true probabilities is large. In our case, there was no difference between behaviour of the estimators in bias and weighted bias.

In some cases, the conversion probability can be close to zero, i.e., the boundary of the parameter space. To check the behaviour of the estimators in such cases, we considered not only the simulation runs at the low level of average conversion probability in our study (i.e., 0.1), but also an additional run with the average conversion probability equal to 0.05. Lower true conversion probability seems to lead to a lower absolute average bias for all the estimators (as expected). However, the trends in average bias across estimators are similar to those when average conversion probability is higher.

To reduce overall computation time in the example, we used data splitting to obtain the estimates in our application. Comparing the performance of the estimators over the entire data set could be another interesting problem.

Our estimation method incorporates data only from users' final click on an ad. In other words, we ignore users' previous (unconverted) clicks on the same ad. Interesting future work could be a model that can capture the information in the historical unconverted clicks of the users.

3.8 Appendix

3.8.1 Likelihood of Z

To prove (3.3), we first derive the cdf of Z_i as

$$\begin{aligned} G_{Z_i}(z_i) &= P\left(Z_i \leq z_i\right) \\ &= P\left(Z_i \leq z_i | C_i = 1\right) P(C_i = 1) \\ &+ P\left(Z_i \leq z_i | C_i = 0\right) P(C_i = 0) \\ &= \begin{cases} 0 & \text{if } z_i \leq 0 \\ H_i(z_i)p_i & \text{if } 0 < z_i < a_i \\ 1 & \text{if } z_i \geq a_i \end{cases}, \end{aligned} \tag{3.10}$$

where $p_i = \frac{\exp(\boldsymbol{\beta}_c'\boldsymbol{x}_i)}{1+\exp(\boldsymbol{\beta}_c'\boldsymbol{x}_i)}$. Therefore, the likelihood function is

$$\begin{aligned} L_g\left(\boldsymbol{\beta}_c|\boldsymbol{z}\right) &= g\left(\boldsymbol{z}\right) \\ &= \prod_i g\left(z_i|a_i\right) \\ &= \prod_i \left(p_i h(z_i)\right)^{I(z_i < a_i)} \left(1 - p_i H(z_i)\right)^{I(z_i \geq a_i)}. \end{aligned} \tag{3.11}$$

Chapter 4

Conversion probability estimation in display advertising: Incorporating information from multiple visits to the same ad

The chapter derives, with few modifications, from:
Safari, A., Altman, R. M., 2018. *Conversion probability estimation in display advertising: Incorporating information from multiple visits to the same ad*, In preparation.

4.1 Abstract

One important problem in display advertising is estimating the probability of conversion (i.e., the probability that a user takes a pre-defined action such as making a purchase) after the user clicks on an ad. The main challenges involved in this estimation are the delays in observing conversions and the size of the data set. Earlier estimators of conversion probability include information from the distribution of the delays in observing conversions. However, these estimators are based on the assumption that a user can click only once on a given ad prior to taking the desired action. We propose a new estimator that uses information not only in the distribution of the delays, but also in the distribution of inter-visit times. Using a simulation study, we show that our estimator is computationally efficient and more accurate than estimators that ignore the information in the inter-visit times. In addition, the coverage probabilities of confidence intervals based on this estimator are close to their nominal values.

4.2 Introduction

Given the important role of online advertising, research on display advertising (displaying ads on different webpages) is expanding quickly. Depending on the purpose of the advertising, different payment methods can be used. We focus on the cost per action (CPA) method in this paper, where the publisher is paid only if the user takes a pre-defined action on the advertiser website after clicking on the ad (conversion). See, e.g., Brea [2014] for other payment methods.

In the CPA setting, publishers need to choose ads with high conversion probability for a given user. Therefore, estimating the conversion probability for a given ad and user is an important question – one that, due to the delays between clicks and conversions, can be challenging to address. Safari et al. and Chapelle [2014] incorporate these conversion delay times when estimating conversion probability. They use a "last-touch attribution" (LTA) method, which assumes that only the last publisher influences the probability of a click's transforming into a conversion. LTA ignores users' previous viewings of the same ad (possibly displayed by different publishers on different webpages) and assumes that conversion is a result of the attribution of the last visit before the conversion. However, many authors have shown the benefits of using another online conversion attribution method, "multi-touch attribution" (MTA) Atlas [2008], over LTA. That literature emphasizes the information we can gain by considering users' previous visits. For instance, Li and Kannan [2014] recommend that advertisers consider users' previous visits to gain a better understanding of the contribution of different publishers and ads to conversion probability. In a data-driven analysis, Berman [2013] concludes that estimating the attribution of each visit (and of its associated publisher) using the MTA method, provides publishers with greater accuracy in choosing the "best" ad for a given user and publisher. Diemert et al. [2017] (using unpublished data from Criteo) argue that using the MTA method improves the accuracy of choosing the "best" match between ad and user. Jordan et al. [2011] (one of our main motivating papers) suggest based on their (unpublished) data that conversion is a result not only of the actions of the *last* publisher, but rather of a combination of the actions of *all* publishers. Finally, Ji et al. [2016] propose a probabilistic model to predict the eventual conversion status via MTA. Specifically, they extend the setting of Chapelle's model to a context where visits can occur. However, their final model is very complex. Even though they make some further assumptions to simplify this model, they end up with a two stage estimator where the estimating equation in the second stage is essentially the same as that of Chapelle [2014], and similarly computationally inefficient. Nevertheless, they advocate the importance of the information in users' previous visits in the analysis of conversion probability. Although all the above papers discuss the importance of the potential information in users' visit history, they do not consider conversion probability estimation via MTA.

Our goal in this paper is to use both the conversion delay times and users' previous visits to estimate conversion probability accurately and in a computationally efficient way. Similar to Safari et al., we build our estimator based on a naive logistic regression model that treats the current conversion statuses of the clicks as their eventual ones, and then use the Kullback-Leibler information criterion (KLIC) to adjust the bias in the estimator. In this paper, however, we adjust the estimator based on not only the delay time distribution, but also the distribution of the number of visits. In addition, we show how much accuracy we lose by ignoring the extra information from the users' previous visits.

The remainder of this paper is organized as follows. In §4.3, we define our notation and explain how previous models for conversion statuses and delay times are misspecified in the case when visits can occur. In §4.4, we introduce a new approach for estimating conversion probability, along with an algorithm to compute these estimates efficiently for a given data set. Section 4.5 describes a simulation study that illustrates the accuracy, precision, and computational efficiency of the estimators. We conclude with a discussion in §4.6.

4.3 Model specification

In this section, we define our notation and present a model for observed conversion statuses, visit times, and delay times.

Label clicks sequentially in time as $1, 2, 3, \ldots$, and let $\boldsymbol{x}_i$ be a $1 \times k$ vector of covariates associated with click i, e.g., attributes of the user and/or origin website. (We use bold letters, e.g., $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,k})$, for vectors throughout this paper.) We define $x_{i,1} = 1 \ \forall i$ so as to include an intercept. Define C_i to be the eventual conversion status indicator of click i, i.e., $C_i = 1$ if click i ever converts and $C_i = 0$ otherwise. Thus, we are interested in estimating $P(C_i = 1)$. We define V_i to be the eventual number of visits associated with click i. Note that $V_i = 0$ if the user clicked and never visited. Let $T_{i,j}$ be the age of click i at its j^{th} visit, $j = 1, \ldots, V_i$. Let $t_{i,0}$ be the observed time of the i^{th} click (treated as non-random for the purposes of this paper). In addition, for clicks that eventually convert, we define T_i^c as the age of click i at conversion, and $S_i = T_i^c - T_{i,V_i}$ as the time between the last visit and conversion. Under this definition, conversion, if it occurs, is assumed to occur *after* the last visit, i.e., conversion does *not* coincide with an additional visit.

Let W be the chosen "conversion window", i.e., the maximum length of the delay permitted between click and conversion. (We assume that W is selected to be long enough that only a negligible proportion of conversions occur outside this window.) The delay time for click i, D_i, is defined as

$$D_i = \begin{cases} T_i^c, & C_i = 1 \\ W, & C_i = 0 \end{cases} .$$

We then define $T^*_{i,j} = T_{i,j} - T_{i,j-1}$ as the time between the $(j-1)^{th}$ and j^{th} visits associated with the i^{th} click.

We make the following assumptions (using the framework of Chapelle [2014] and Safari et al.): **i.** the true conversion probability is fixed over time, **ii.** the predictors don't depend on time, **iii.** a converted click can never become unconverted, **iv.** an unconverted click with delay time less than W could convert.

An additional assumption – which differentiates the present work from that of Safari et al. and Chapelle [2014] – is that a user can visit an ad multiple times within the conversion window and that inter-visit times are independent. We refer to the user's first click as simply "click" and all subsequent clicks as "visits". Thus, a user can "click" only once but can "visit" repeatedly. As with delays, we assume that only a negligible proportion of visits occur outside of the conversion window.

Now suppose that, at a given moment $t > 0$, we wish to estimate the conversion probability p_i of click i (with covariates $\boldsymbol{x}_i$) based on the conversion statuses of the clicks observed prior to time t. Due to the conversion delay issue, we may not know the eventual conversion statuses of these clicks at time t. Therefore, we define additional variables to represent the observed characteristics of these clicks at time t. Specifically, at this time, we say that $N(t)$ clicks have accumulated. Let $Y_i(t)$ be the current (observed) conversion status indicator of click $i = 1, \ldots, N(t)$, where $Y_i(t) = 1$ if click i converted prior to time t and $Y_i(t) = 0$ otherwise. Define $a_i(t) = t - t_{i,0}$ to be the age of click i, i.e., the time since the user clicked. Since $t_{i,0}$ is treated as non-random, $a_i(t)$ is treated as non-random as well. Note that D_i is observed prior to t if and only if $Y_i(t) = 1$ or $a_i(t) \geq W$. Similarly, D_i is right censored (i.e., greater than or equal to $a_i(t)$) if and only if $Y_i(t) = 0$ and $a_i(t) < W$. Finally, let $O_i(t)$ be the observed number of visits associated with click i by time t, $\boldsymbol{T}_i^{O_i}(t) = (T_{i,1}, \ldots, T_{i,O_i(t)})$, and let

$$R_i(t) = \begin{cases} S_i & \text{if } Y_i(t) = 1 \\ a_i(t) - T_{i,O_i(t)} & \text{if } Y_i(t) = 0 \end{cases} \tag{4.1}$$

be the "remainder" (the time elapsed since the last observed visit, truncated at conversion, if it occurs). We can define an equivalence relationship between Y_i and (R_i, T_{i,O_i}) as

$$\begin{aligned} R_i(t) < a_i(t) - T_{i,O_i}(t) &\iff Y_i(t) = 1 \\ R_i(t) = a_i(t) - T_{i,O_i}(t) &\iff Y_i(t) = 0 \end{aligned} \tag{4.2}$$

We treat $(O_i(t), \boldsymbol{T}_i^{O_i}(t), R_i(t))$ as the observed data for the i^{th} click at time t. For subsequent derivations, we consider a fixed time t and suppress t in our notation for convenience.

Let $(\boldsymbol{\beta}_c, \boldsymbol{\beta}_v, \boldsymbol{\beta}_t)$ be the covariate coefficients in the models for the distribution of conversion status, number of visits, and inter-visit times and remainders, respectively. For simplicity, we assume that the $T^*_{i,j}$'s, S_i's, and V_i's are independent, and that for each i, the $T^*_{i,j}$'s and S_i (if $C_i = 1$) are exponential distributed with rate parameter $\lambda_i = \exp(\boldsymbol{\beta}_t' \boldsymbol{x}_i)$. We assume a logistic regression model with covariate coefficients $\boldsymbol{\beta}_c$ for conversion status. (The exponential and logistic models are consistent with the approaches of Safari et al. and Chapelle [2014].) Finally, we assume that the number of visits, V_i, follows a Poisson distribution with mean $\gamma_i = \exp(\boldsymbol{\beta}_v' \boldsymbol{x}_i)$.

Let f_X and F_X be generic representations of the pmf (or pdf) and cdf, respectively, of a random variable X. We can then write the contribution of the i^{th} click to the likelihood as

$$\begin{aligned} & \mathcal{L}_i\left(\boldsymbol{\beta}_c, \boldsymbol{\beta}_v, \boldsymbol{\beta}_t | r_i, o_i, \boldsymbol{t}_i^{o_i}\right) \\ = & \begin{cases} p_i f_{V_i}(o_i) f_{\boldsymbol{T}_i^{o_i}}(\boldsymbol{t}_i^{o_i}) f_{S_i}(r_i), & t_{i,o_i} + r_i < a_i \\ f_{\boldsymbol{T}_i^{o_i}}(\boldsymbol{t}_i^{o_i}) \left\{p_i[1 - F_{S_i}(r_i)] f_{V_i}(o_i) + \right. & \\ \left. (1 - p_i) f_{V_i}(o_i) + [1 - F_{V_i}(o_i)][1 - F_{S_i}(r_i)]\right\}, & t_{i,o_i} + r_i = a_i \end{cases} \end{aligned} \tag{4.3}$$

See appendix 4.7.1 for the derivation.

Let $\boldsymbol{T}^O$ be the concatenated vector $(\boldsymbol{T}_1, \ldots, \boldsymbol{T}_N)$. We define the full likelihood as

$$\mathcal{L}\left(\boldsymbol{\beta}_c, \boldsymbol{\beta}_v, \boldsymbol{\beta}_t | \boldsymbol{r}, \boldsymbol{o}, \boldsymbol{t}^{\boldsymbol{o}}\right) = \prod_{i=1}^{N} \mathcal{L}_i\left(\boldsymbol{\beta}_c, \boldsymbol{\beta}_v, \boldsymbol{\beta}_t | r_i, o_i, \boldsymbol{t}_i^{o_i}\right) \tag{4.4}$$

Maximizing the full likelihood (4.4) is computationally burdensome. Therefore, in § 4.4, we present some alternatives to the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}_c$ (the parameter of primary interest). We describe three estimators from the literature and a fourth novel estimator obtained by approximating the full likelihood with tractable objective functions.

4.4 Estimation of $\boldsymbol{\beta}_c$

In order to compute the MLE of $\boldsymbol{\beta}_c$, we would need to maximize (4.4) with respect to all the parameters simultaneously. Since maximization of (4.4) can be slow and inefficient (cf. Safari et al.), in this paper, we consider alternative estimators of $\boldsymbol{\beta}_c$. In § 4.4.1, we discuss some existing estimators of $\boldsymbol{\beta}_c$. In § 4.4.2, we develop a new estimator that incorporates information in visits as well as delay times.

4.4.1 Existing estimators of $\boldsymbol{\beta}_c$

In this section, we present three estimators of $\boldsymbol{\beta}_c$ proposed in the literature: the "naive estimator", the maximizer of the likelihood associated with the Chapelle's model, and the bias-adjusted estimator developed by Safari et al..

Naive estimator

The naive estimator of $\boldsymbol{\beta}_c$ is based on the simple idea of treating current conversion statuses of the clicks as their eventual conversion statuses (an assumption that may or may not be correct for unconverted clicks), and ignoring possible conversion delays. In other words, C_i is treated as equal to $Y_i(t)$ for each i and t. The likelihood function associated with this model is then

$$L_f\left(\boldsymbol{\alpha}|\boldsymbol{r},\boldsymbol{o},\boldsymbol{t^o}\right) \quad = \quad f_{\boldsymbol{O},\boldsymbol{T^o}}(\boldsymbol{o},\boldsymbol{t^o})\prod_{i=1}^{N}\left[\theta_i^{I(t_{i,o_i}+r_i<a_i)}(1-\theta_i)^{I(t_{i,o_i}+r_i=a_i)}\right], \tag{4.5}$$

where $\theta_i = \frac{\exp(\boldsymbol{\alpha}'\boldsymbol{x}_i)}{1+\exp(\boldsymbol{\alpha}'\boldsymbol{x}_i)}$ is the conversion probability of the i^{th} click, and $\boldsymbol{\alpha}$ is a vector of regression coefficients. Note that the product in (4.5) is equivalent to the likelihood associated with a logistic regression model. The naive estimator of $\boldsymbol{\beta}_c$ is defined as the maximizer of this part of the likelihood.

The function (4.5) is convex and computationally efficient to optimize. However, the estimator of conversion probability based on this model is biased low, since some of the unconverted clicks could convert later.

Delay feedback model estimator

The delay feedback model (DFM) developed by Chapelle [2014] and studied by Safari et al. forms the basis of a second estimator of $\boldsymbol{\beta}_c$. The DFM describes the distribution of conversion statuses at time t under the assumption that no visits can occur. Under this assumption, $\boldsymbol{r}$ represents the entirety of the observed data. Safari et al. show that the likelihood function associated with the DFM is

$$L_d\left(\boldsymbol{\beta}_c|\boldsymbol{r}\ \right) = \prod_{i=1}^{N}\left[p_i f_{S_i}(r_i)\right]^{I(r_i<a_i)}\left[1-p_i F_{S_i}(r_i)\right]^{I(r_i=a_i)}. \tag{4.6}$$

Chapelle [2014] applies the DFM in a context where visits can occur but are not recorded. If a visit occurs, he assumes that it is the last (i.e., that $O_i = V_i$). He then treats the time of this visit as click time – and measures age, delay time, and conversion window from this newly defined click time. However, when visits can occur, delays are no longer exponential random variables, but rather a random sum of exponential distributed random variables. Thus, the likelihood (4.6) is misspecified in the case when visits can occur. Given the possible impact of this misspecification on the MLE of $\boldsymbol{\beta}_c$ based on (4.6), and given that maximizing (4.6) can be very time-consuming (Safari et al.), we do not consider this estimator further in this paper.

Delay-adjusted estimator

A third estimator of $\boldsymbol{\beta}_c$ in the literature is that proposed by Safari et al.. This estimator is based on the naive estimator, but with a bias adjustment computed using the KLIC approach. Specifically, Safari et al. assume that the true model is the DFM and treat its parameters as known. Then they minimize the KLIC with respect to the parameters of the naive model. Therefore, they obtain parameters of the DFM in terms of the naive model's parameters, which can be estimated by maximizing the likelihood function associated with the naive model. Since this adjustment is based only on the delay distribution, we call this estimator the delay-adjusted (D-adjusted) estimator (equivalent to the E-bias-adjusted estimator in the terminology of Safari et al.).

If a user can visit only once (so that delay time is exponential), this bias adjustment is asymptotically exact. However, if a user can visit multiple times (so that delay time is a random sum of exponential random variables), this adjustment is inaccurate. In the other hand, as time goes on, the bias correction disappears (i.e., the D-adjusted estimator approaches to the naive estimator). Since the naive estimator is a consistent estimator for the true conversion probability after substantial time has passed (no adjustments needed), the D-adjusted estimator is asymptotically consistent for the true conversion probability as well.

4.4.2 Delay-visit-adjusted estimator of $\boldsymbol{\beta}_c$

In this section, we develop a new estimator based on the naive estimator adjusted for bias using both the delay and inter-visit time distributions. Rather than working with the complicated likelihood (4.4), our strategy is to specify three simpler objective functions that we maximize to estimate $\boldsymbol{\beta}_t$, $\boldsymbol{\beta}_v$, and $\boldsymbol{\beta}_c$ sequentially. In particular, we first consider an approximation of (4.4) that allows $\hat{\boldsymbol{\beta}}_t$ to be estimated separately from $\boldsymbol{\beta}_c$ and $\boldsymbol{\beta}_o$. We then maximize an approximation of $L\left(\boldsymbol{\beta}_c, \boldsymbol{\beta}_v, \hat{\boldsymbol{\beta}}_t | \boldsymbol{r}, \boldsymbol{o}, \boldsymbol{t}^{\boldsymbol{o}}\right)$ that allows $\hat{\boldsymbol{\beta}}_v$ to be estimated separately from $\boldsymbol{\beta}_c$. Finally, we estimate $\boldsymbol{\beta}_c$ using the naive approach described in Section 4.4.1, with a bias correction determined using the KLIC approach of Safari et al. applied to the distribution of $\boldsymbol{R}$ conditional on $\boldsymbol{O}$ and $\boldsymbol{T}^{\boldsymbol{O}}$, evaluated at $\hat{\boldsymbol{\beta}}_v$ and $\hat{\boldsymbol{\beta}}_t$. We call $\hat{\boldsymbol{\beta}}_c$ the delay-visit-adjusted (DV-adjusted) estimator.

We justify our approach as follows. First, we note that for every click i for which $a_i \geq W$, the random variables $O_i = V_i$, $\boldsymbol{T}_i^{O_i} = \boldsymbol{T}_i^{V_i}$, and R_i (which equals S_i if the i^{th} click converts and $a_i - T_{i,V_i}$ otherwise) are independent. As time goes on, the proportion of such clicks approaches 100%, and both the likelihood (4.4) and the product of our chosen objective functions approach the product of the marginal distributions of O_i, R_i, and $\boldsymbol{T}_i^{V_i}$. Therefore, our estimate of $\boldsymbol{\beta}_c$ is asymptotically equivalent to the MLE (and to the naive estimator). However, we expect that after only a relatively short time has passed (our scenario of interest), our estimator will outperform the MLE.

We now describe our approach in detail. We select an objective function for estimating $\boldsymbol{\beta}_t$ in two steps. First, we assume that $p_i \equiv 1$ in (4.4). With this approximation, $\boldsymbol{\beta}_t$ can be estimated separately from $\boldsymbol{\beta}_c$ and $\boldsymbol{\beta}_v$. In particular, taking $p_i \equiv 1$, the portion of (4.4) that depends on $\boldsymbol{\beta}_t$ is

$$\begin{aligned}
& L_t(\boldsymbol{\beta}_t|\boldsymbol{o}, \boldsymbol{t^*}, \boldsymbol{r}) \\
= & \prod_{i=1}^{N} \left\{ \prod_{j=1}^{o_i} f_{T^*_{i,j}}(t^*_{i,j}) \right\} [f_{S_i}(r_i)]^{I(t_{i,o_i}+r_i<a_i)} [1 - F_{S_i}(r_i)]^{I(t_{i,o_i}+r_i=a_i)} \\
= & \prod_{i=1}^{N} \prod_{j=1}^{o_i} \lambda_i \exp(-\lambda_i t^*_{i,j}) \cdot \prod_{i=1}^{N} [\lambda_i \exp(-\lambda_i r_i)]^{I(t_{i,o_i}+r_i<a_i)} [\exp(-\lambda_i r_i)]^{I(t_{i,o_i}+r_i=a_i)} . \quad (4.7)
\end{aligned}$$

Note that, as a consequence of our approximation, the last factor in (4.7) (wrongly) includes contributions from clicks that may not convert or have another associated visit. However, as time passes, the proportion of these clicks goes to 0 and the approximation improves.

The second product in (4.7) – the objective function associated with the remainders – is equivalent to the usual likelihood associated with right-censored exponential observations. In our context, when t is small, the censoring rate can be very high. In the presence of substantial censoring, the estimator of $\boldsymbol{\beta}_t$ can be quite biased (see, e.g., Wan et al. [2015a], Shen and Yang [2014]). Thus, our second step in choosing an objective function for estimating $\boldsymbol{\beta}_t$ is to adjust this portion of (4.7) according to the penalized score method of Firth [1993]. Firth's method is a general approach to bias reduction. Pettitt [1998] use this idea to obtain the penalized likelihood when the responses are exponentially distributed and possibly censored at a fixed censoring time for all observations. We extend their approach by relaxing the assumption of a fixed censoring time for all observations, and allowing the remainder associated with each click to have its own censoring time. The adjusted form of (4.7) is then

$$\begin{aligned}
& L^*_t(\boldsymbol{\beta}_t|\boldsymbol{o}, \boldsymbol{t^*}, \boldsymbol{r}) \\
= & \prod_{i=1}^{N} \prod_{j=1}^{o_i} \lambda_i \exp(-\lambda_i t^*_{i,j}) \cdot \\
& \prod_{i=1}^{N} [\lambda_i \exp(-\lambda_i r_i)]^{I(t_{i,o_i}+r_i<a_i)} [\exp(-\lambda_i r_i)]^{I(t_{i,o_i}+r_i=a_i)} (\lambda_i)^{-2} [1 - \exp(-\lambda_i r_i)] . \quad (4.8)
\end{aligned}$$

We maximize the objective function L^*_t to estimate $\boldsymbol{\beta}_t$.

Given $\hat{\boldsymbol{\beta}}_t$, we now develop an objective function for estimating $\boldsymbol{\beta}_v$. Specifically, in (4.4), we assume $p_i \equiv 0$. With this approximation, $\boldsymbol{\beta}_v$ can be estimated separately from $\boldsymbol{\beta}_c$. The portion of (4.4) that depends on $\boldsymbol{\beta}_v$ is then

$$\begin{aligned}
L_v(\boldsymbol{\beta}_v|\boldsymbol{o}, \boldsymbol{t^o}, \boldsymbol{r} \ ; \ \hat{\boldsymbol{\beta}}_t) \quad = & \prod_{i \in \{t_{i,o_i}+r_i<a_i\}} f_{V_i}(o_i) \cdot \\
& \prod_{i \in \{t_{i,o_i}+r_i=a_i\}} \{ f_{V_i}(o_i) + [1 - F_{S_i}(r_i)][1 - F_{V_i}(o_i)] \} , \quad (4.9)
\end{aligned}$$

The second product in (4.9) (wrongly) assumes that unconverted clicks never convert. However, again, as time passes, the proportion of these clicks goes to 0 and the approximation improves.

Finally, we estimate $\hat{\boldsymbol{\beta}}_c$ given $\hat{\boldsymbol{\beta}}_t$ and $\hat{\boldsymbol{\beta}}_v$. One option is to maximize the portion of (4.4) that depends on $\boldsymbol{\beta}_c$, i.e.,

$$\begin{aligned} & L_c\left(\boldsymbol{\beta}_c|\boldsymbol{r},\boldsymbol{o},\boldsymbol{t}_i^{\boldsymbol{o}}\ ;\ \hat{\boldsymbol{\beta}}_v,\hat{\boldsymbol{\beta}}_t\right) \\ = & \prod_{i=1}^{N}\left\{p_i\hat{f}_{V_i}(o_i)\hat{f}_{S_i}(r_i)\right\}^{I(t_{i,o_i}+r_i<a_i)} \cdot \\ & \left\{p_i[1-\hat{F}_{S_i}(r_i)]\hat{f}_{V_i}(o_i)+(1-p_i)\hat{f}_{V_i}(o_i)+[1-\hat{F}_{S_i}(r_i)][1-\hat{F}_{V_i}(o_i)]\right\}^{I(t_{i,o_i}+r_i=a_i)} \end{aligned} \tag{4.10}$$

However, the process of maximizing (4.10) is relatively slow. Instead, we use a version of the KLIC approach of Safari et al.. Specifically, we derive a bias-adjustment for $\hat{\boldsymbol{\alpha}}$ (the estimator of the naive model). To this end, we estimate the expected value (under the true model) of the log ratio of the distributions of $(\boldsymbol{R},\boldsymbol{O},\boldsymbol{T})$ under the true and naive models. We find the minimizer, $\widetilde{\boldsymbol{\alpha}}$, of this function (in terms of $\boldsymbol{\beta}_c$, $\boldsymbol{\beta}_o$, and $\boldsymbol{\beta}_t$). Using this relationship and our estimates $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}_t$ and $\hat{\boldsymbol{\beta}}_v$, we obtain an estimate for $\boldsymbol{\beta}_c$.

The details are as follows. Noting that the distribution of $(\boldsymbol{O},\boldsymbol{T})$ is the same under both the true and naive models, the ratio of interest is equivalent to the ratio of the distributions of $\boldsymbol{R}$ conditional on $(\boldsymbol{O},\boldsymbol{T})$ under the two models. The KLIC (where the expectation is with respect to the distribution in (4.4)) is then

$$\begin{aligned} & KLIC\left(\boldsymbol{\alpha};\boldsymbol{\beta}_c,\boldsymbol{\beta}_o,\boldsymbol{\beta}_t\right) \\ = & E\left\{\ln\left[\frac{\mathcal{L}\left(\boldsymbol{\beta}_c,\boldsymbol{\beta}_v,\boldsymbol{\beta}_t|\boldsymbol{R},\boldsymbol{O},\boldsymbol{T}^{\boldsymbol{O}}\right)}{L_f\left(\boldsymbol{\alpha}|\boldsymbol{R},\boldsymbol{O},\boldsymbol{T}^{\boldsymbol{O}}\right)}\right]\right\} \\ = & E\left\{\ln\left[\frac{\mathcal{L}\left(\boldsymbol{\beta}_c,\boldsymbol{\beta}_v,\boldsymbol{\beta}_t|\boldsymbol{T},\boldsymbol{O},\boldsymbol{T}^{\boldsymbol{O}}\right)/f_{\boldsymbol{O},\boldsymbol{T}}(\boldsymbol{O},\boldsymbol{T})}{L_f\left(\boldsymbol{\alpha}|\boldsymbol{R},\boldsymbol{O},\boldsymbol{T}^{\boldsymbol{O}}\right)/f_{\boldsymbol{O},\boldsymbol{T}}(\boldsymbol{O},\boldsymbol{T})}\right]\right\} \\ = & \text{constant} - E\left\{\prod_{i=1}^{N}\left[\theta_i^{I(T_{i,O_i}+R_i<a_i)}(1-\theta_i)^{I(T_{i,O_i}+R_i=a_i)}\right]\right\}, \end{aligned} \tag{4.11}$$

where “constant” means independent of $\boldsymbol{\alpha}$.

To evaluate (4.11), we separate the joint distribution of $(\boldsymbol{R},\boldsymbol{O},\boldsymbol{T})$ into the product of the conditional distribution of $(\boldsymbol{R}\mid\boldsymbol{O},\boldsymbol{T})$ and the joint distribution of $(\boldsymbol{O},\boldsymbol{T})$. We define the expectation with respect to the former as the “conditional KLIC” (CKLIC).

The conditional distribution of $\boldsymbol{R}$ under the true model is

$$f_{\boldsymbol{R}|\boldsymbol{O},\boldsymbol{T}}\left(\boldsymbol{r}\mid\boldsymbol{o},\boldsymbol{t}\right) = \prod_{i=1}^{N}\frac{f_{R_i,O_i,\boldsymbol{T}_i^{O_i}}(r_i,o_i,\boldsymbol{t}_i^{o_i})}{f_{\boldsymbol{T}_i^{o_i}}(\boldsymbol{t}_i^{o_i})\left\{f_{V_i}(o_i)+[1-F_{V_i}(o_i)][1-F_{S_i}(a_i-t_{i,o_i})]\right\}}. \tag{4.12}$$

See appendix 4.7.2 for the derivation.

Therefore, the CKLIC is

$$\begin{aligned} & CKLIC\left(\boldsymbol{\alpha}; \boldsymbol{\beta}_c, \boldsymbol{\beta}_o, \boldsymbol{\beta}_t\right) \\ = \; & \text{constant} - \sum_{i=1}^{N} \left\{ \ln(\theta_i) \frac{p_i f_{V_i}(O_i) F_{S_i}(a_i - T_{i,O_i})}{f_{V_i}(O_i) + [1 - F_{V_i}(O_i)][1 - F_{S_i}(a_i - T_{i,O_i})]} \right. \\ & \left. + \ln(1-\theta_i) \left[1 - \frac{p_i f_{V_i}(O_i) F_{S_i}(a_i - T_{i,O_i})}{f_{V_i}(O_i) + [1 - F_{V_i}(O_i)][1 - F_{S_i}(a_i - T_{i,O_i})]} \right] \right\} \end{aligned} \tag{4.13}$$

See appendix 4.7.3 for the derivation.

Evaluating the expectation of the CKLIC with respect to the joint distribution of $(\boldsymbol{O}, \boldsymbol{T})$ to obtain the KLIC is challenging. We thus instead compute the empirical estimate of the KLIC by evaluating the CKLIC at the values of $\boldsymbol{O}$ and $\boldsymbol{T}$ observed in our sample:

$$\begin{aligned} & \widehat{KLIC}\left(\boldsymbol{\alpha}; \boldsymbol{\beta}_c, \boldsymbol{\beta}_o, \boldsymbol{\beta}_t\right) \\ = \; & \text{constant} - \sum_{i=1}^{N} \left\{ \ln(\theta_i) \frac{p_i f_{V_i}(o_i) F_{S_i}(a_i - t_{i,o_i})}{f_{V_i}(o_i) + [1 - F_{V_i}(o_i)][1 - F_{S_i}(a_i - t_{i,o_i})]} \right. \\ & \left. + \ln(1-\theta_i) \left[1 - \frac{p_i f_{V_i}(o_i) F_{S_i}(a_i - t_{i,O_i})}{f_{V_i}(o_i) + [1 - F_{V_i}(o_i)][1 - F_{S_i}(a_i - t_{i,o_i})]} \right] \right\} \end{aligned} \tag{4.14}$$

White [1982] shows that the MLE of a misspecified model is consistent for the minimizer of the KLIC. The minimizer of (4.14), $\widetilde{\boldsymbol{\alpha}}$, can be estimated using the equations

$$\begin{aligned} & \left. \frac{\partial \; \widehat{KLIC}\left(\boldsymbol{\alpha}; \boldsymbol{\beta}_c, \boldsymbol{\beta}_o, \boldsymbol{\beta}_t\right)}{\partial \alpha_j} \right|_{\widetilde{\boldsymbol{\alpha}}} = 0 \qquad (4.15) \\ \Rightarrow \; & \sum_i \frac{p_i f_{V_i}(o_i) F_{S_i}(a_i - t_{i,o_i})}{f_{V_i}(o_i) + [1 - F_{V_i}(o_i)][1 - F_{S_i}(a_i - t_{i,o_i})]} x_{i,j} = \sum_i x_{i,j} \widetilde{\theta}_i \;, \quad j = 1, \ldots, k. \end{aligned}$$

Here $\widetilde{\theta}_i = \theta_i|_{\widetilde{\boldsymbol{\alpha}}}$, and k is the number of regression coefficients. If we solve the equations in (4.15) for $\boldsymbol{\beta}_c$, we will obtain a formula to estimate the parameters of the true model, $\boldsymbol{\beta}_c$. In other words, using the estimates $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}_t$ and $\hat{\boldsymbol{\beta}}_v$ and estimating $\widetilde{\boldsymbol{\alpha}}$ by $\hat{\boldsymbol{\alpha}}$, we can compute an estimate of $\boldsymbol{\beta}_c$. Note that the equations in (4.15) are equivalent to the weighted quasi-score estimates and thus can be solved efficiently – much more quickly than the score equations based on (4.10).

To summarize, we follow the steps below to obtain our bias-adjusted estimate of $\boldsymbol{\beta}_c$:

1. Compute the estimate of $\boldsymbol{\beta}_t$ from (4.8).

2. Compute the estimate of $\boldsymbol{\beta}_v$ from (4.9).

3. Compute the MLE of $\boldsymbol{\alpha}$ based on the naive model (4.5).

4. Compute the bias-adjusted estimate $\hat{\boldsymbol{\beta}}_c$ by solving the equations in (4.15), substituting in the estimates of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_v$, and $\boldsymbol{\beta}_t$ obtained in steps 1–3.

4.5 Simulation study

In this section, we use a simulation study to evaluate the performance of our **DV-adjusted** estimator. We investigate its bias, standard error (SE), and computation time relative to four other estimators: the **naive** estimator, the **oracle** estimator (the MLE of the logistic regression model based on the eventual conversion statuses of the clicks), the **true-DV-adjusted** estimator (the delay-visit adjusted estimator computed using the true distribution of the inter-visit and delay times), and the delay adjusted **D-adjusted** estimator due to Safari et al..

The oracle and true-DV-adjusted estimates are not obtainable in practice, where at any time t, the delay, inter-visit, and remainder distribution parameters – as well as some C_i's – will be unknown. The true-DV-adjusted estimator helps us to gauge how much we lose by estimating the parameters involved in (4.8) and (4.9), while the oracle estimator is the most accurate estimator that we can apply to a given data set.

To evaluate the D-adjusted estimator in our setting (i.e., where visits can occur), we assume that $V_i = O_i$, and treat time of the last visit as the click time. This assumption, also made by Chapelle [2014], is incorrect for some clicks. However, as time goes on, the proportion of clicks for which $O_i < V_i$ goes to 0. Therefore, the bias correction is asymptotically accurate, even in our setting. By comparing the performance of the D- and DV-adjusted estimators, we can observe the increase in accuracy of the estimate of $\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}$ resulting from using the information in the observed number of visits.

We use bias of the estimated probabilities of conversion as a measure of error in the simulation study. Bias at time t is defined as $\frac{1}{N}\sum_i (p_i - \hat{p}_i)$ for an estimator of p_i, $\hat{p}_i$. Recall that the p_i's vary according to covariates; average bias can be interpreted as an estimate of the *marginal bias* of the estimator of probability of conversion (in contrast with $\mathrm{E}[p_i - \hat{p}_i]$, which represents the bias of $\hat{p}_i$ *conditional* on $\boldsymbol{x}_i$).

We were unable to procure a data set that contains users' visit history of an ad. Instead, to inform the design of our simulation study, we pick approximately 8500 clicks ($N \approx 8500$) from a campaign with a large number of clicks and moderate average conversion probability ($\sim 30\%$) from the publicly available data[1] originally studied by Chapelle [2014] (sub-sampled as described in Safari et al.). The data include four continuous and three categorical covariates that are measured on each click.

We consider four factors affecting the performance of the conversion probability estimators: average conversion probability, average delay time, average number of visits (where

[1] The data set is available at http:/research.criteo.com/outreach

Table 4.1: Levels of the factors in the simulation study – averaged across all clicks

Factor	Low	Medium	High
Conversion probability	0.1	0.3	0.6
Delay mean*	7	4	2
Number of visits	3	7	10

* in days

average means across all clicks in the campaign), and delay distribution (exponential or Weibull). Table 4.1 shows the levels of the factors we chose based on the range of the conversion probabilities and delays in the real data set (we use some online resources to choose values for the average number of visits – including online statistical reports [2], realistic simulation algorithms [3] and display marketing reports Brea [2014]). To keep the simulation study feasible with a manageable number of unknown parameters in the model, we assume no interaction between the covariates. Consequently, the covariate coefficients remain fixed across different campaigns. I.e, we generate data for each simulation run by varying only the intercepts in the delay, visit, and conversion models, while keeping all other coefficients fixed.

We use these data to choose the true parameter values in our simulation study. Specifically, we take $\boldsymbol{\beta}_t$ as the MLE of the parameters of the delay distribution (assumed to be either exponential or Weibull) based on the observed delays in the chosen campaign (using only converted clicks). Similarly, we choose $\boldsymbol{\beta}_c$ as the MLE of the parameters of the logistic regression model fit to the final conversion status of the clicks in the data set. The data set does not provide information about the distribution of number of visits. So, for simplicity, we choose $\boldsymbol{\beta}_{vj} = \boldsymbol{\beta}_{cj}$ for all $j > 1$. See online material for details regarding the covariates and coefficients we used.

We use the $\boldsymbol{T}_0$'s and the $\boldsymbol{x}$'s from the available data set. For each click i, we generate V_i, then independent inter-visit times $\boldsymbol{T}_i^{V_i}$, and then C_i. For any click i with $C_i = 1$, we also generate S_i. Following Chapelle [2014], we assume $W = 30$ is large enough that all visits occur within the conversion window. In other words, we assume that a negligible proportion of clicks convert outside the conversion window. To create a realistic scenario in our simulation studies, we track the clicks and the subsequent visits since click time of the first click (which we call $t = 0$), and evaluate the estimators at 10 different, equally spaced time steps over a two month period. At each time step t, we consider only clicks that occurred by t. Similarly, we treat a click as converted only if we observe its conversion

[2] E.g. Econsultancy, available at https://econsultancy.com

[3] E.g. MacPaw Inc., available at http://analyzecore.com

by t (and that its age is less than W). Otherwise, we treat it as unconverted. In addition, for each click, we consider only visits that occurred by time t.

4.5.1 Study 1

We first consider the case where the S_i's and inter-visit times follow an exponential distribution. Thus, the DV-adjusted and true-DV-adjusted estimators are based on the correct model.

Figure 4.1 shows the average bias of the estimators over time when all three factors, average conversion probability, average delay, and average number of visits, are at their medium level. Except for the oracle estimator, the DV-adjusted estimator seems to outperform all the other estimators until three weeks after the first click, after which the true-DV-adjusted estimator appears to perform better. To show the closeness of the average bias to the nominal value of 0 at each time point more carefully, we add the non-rejection region (NRR) for the score test of whether average bias differs from 0. The NRR is defined as $(2\sqrt{0.95(1-0.95)/R}, 2\sqrt{0.95(1-0.95)/R}) \approx (0.013, 0.013)$, where $R = 1000$ is the number of replicates. The average biases of the true-DV-adjusted and DV-adjusted estimators are not significantly different the nominal bias level after one month and six weeks, respectively. The D-adjusted estimator appears to be less biased than the naive estimator, but neither converges to the bias NRR after two months. The trends in average bias seem to be similar when we use other levels of the factors given in table 4.1. As expected, when the average number of visits is low, the accuracy of the DV-adjusted and D-adjusted estimators appear to be similar, and when both average delay time and number of visits are at their low level, the accuracy of all the estimators are similar (see online materials).

4.5.2 Study 2

In this study, we consider a Weibull distribution for the S_i's and inter-visit times. Consequently, all of our estimators, even the true-DV-adjusted estimator, are based on a misspecified model. Since the true-DV-adjusted estimator hasn't been derived for this study, we don't consider this estimator here.

We compute the five estimates over time as in study 1. Note that in this case, the D-adjusted and DV-adjusted estimators are based on the (misspecified) exponential distribution. Figure 4.2 shows the average bias of the estimators over time when all the factors (average conversion probability, average delay, and average number of visits) are at their medium level with the bias NRR (shaded area). Again, the DV-adjusted estimator appears to outperform the other estimators (except the oracle estimator). In particular, the average bias of the DV-adjusted estimator essentially disappears within the first two months, whereas the average biases of the D-adjusted and naive estimators do not. The trends in the bias of the estimators appear to be similar when we use other levels of the factors given in table 4.1. As in study 1, when average number of visits is low, the accuracies of the

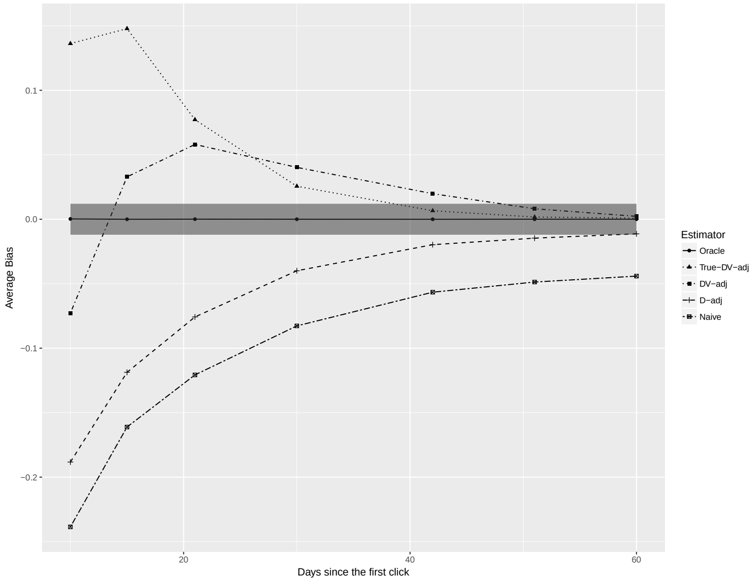

Figure 4.1: Average bias of the estimators along with the bias NRR over time for the medium level of the factors when the S_i's and inter-visit times follow an exponential distribution

DV-adjusted and D-adjusted estimators seem similar, and when both average delay time and number of visits are at their low level, the accuracies of all the estimators seem similar (see online materials).

The results of study 2 are important in that they show the robustness of the DV-adjusted estimator to misspecification of the delay distribution. In particular, even if the delay distribution is not exponential (as appears to be the case in the Criteo data set), the DV-adjusted estimator has minimal bias after two months.

4.5.3 Coverage probability of the bias-adjusted estimators

The estimating equations (4.15), which are equivalent to weighted quasi-score functions, suggest an efficient method for obtaining a standard error (SE) for the DV-adjusted and D-adjusted (see Safari et al.) estimators. In this section, we study the validity of our SE (via the coverage probability of confidence intervals for the probability of conversion).

Figure 4.3 shows the estimated coverage probability (CP) associated with (nominal) 95% CIs for conversion probability (for each estimator at each time point) when the S_i's and inter-visit times follow a Weibull distribution (study 2). In the first month, as expected, the estimated CPs of the estimators are below the nominal level (approximately 75%). However, in the second month, especially after 40 days, the estimated CPs of the estimators are more than 92%. As expected, the estimated CPs based on the D-adjusted estimator appear to be always smaller than those of the DV-adjusted estimator. To show the closeness of the average CP to the nominal value of 0.95 at each time point more carefully, we add the NRR for the score test of whether CP differs from 0.95, which is approximately $(0.94, 0.96)$

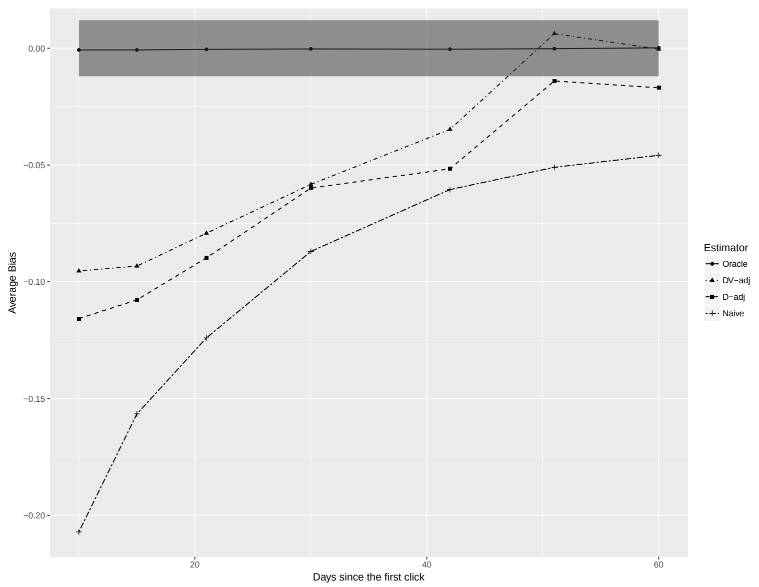

Figure 4.2: Average bias of the estimators with the bias NRR over time for the medium level of the factors when the S_i's and inter-visit times follow a Weibull distribution

for $R = 2000$. The estimated CP of the DV-adjusted estimator is not significantly different from 0.95 at the last 2 time steps. In contrast, the estimated CP of the D-adjusted estimator differs significantly from 0.95 even after two months. The results for other runs are similar (see online material).

4.5.4 Computation time

Since publishers need to re-compute the estimates of the regression coefficients frequently (using the new clicks that arrive), and since the number of clicks associated with a given campaign can be huge, computational efficiency of the conversion probability estimate is critical. Table 4.2 shows the average and sample standard deviation (SSD) of computation times (in seconds) for the estimators when the true distribution of the S_i's and inter-visit times is Weibull (study 2) and the factors are at their medium level. As expected, since the DV-adjusted estimator has an extra step (the estimation of $\boldsymbol{\beta}_v$ in the visits model) relative to the D-adjusted estimator, its computation time is almost 30% greater than that of the D-adjusted estimator. However, given the high accuracy of the DV-adjusted estimator relative to the other estimators presented in §4.5.1 and §4.5.2, and also a fairly small computation time difference relative to other estimators, we recommend the DV-adjusted estimator in practice. The computation time of the estimates is similar for study 1.

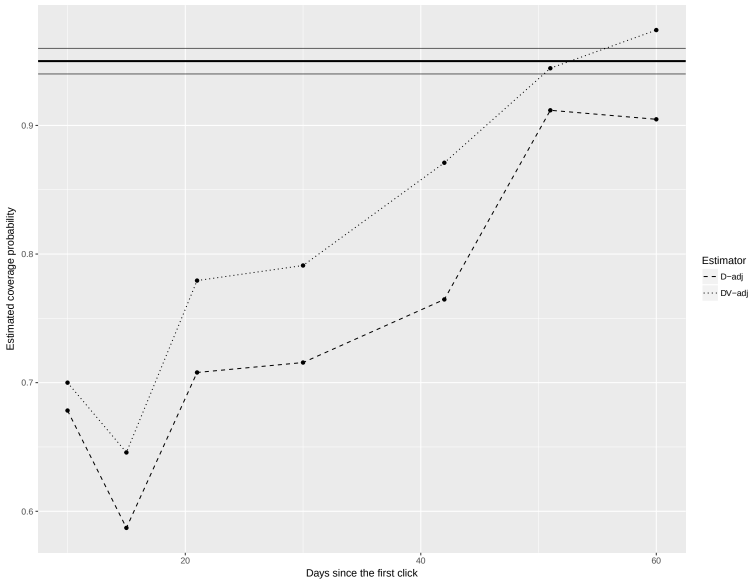

Figure 4.3: Estimated coverage probability of the 95% CI based on the DV-adjusted and D-adjusted estimators along with the NRR over time for the medium level of the factors (study 2)

Table 4.2: Average (SSD) computation time (in seconds) for each estimator over different time steps for the medium level of the factors (study 2).

Estimator	Average (SSD) Run time
Naive	0.10 (0.03)
D-adjusted	6.61 (3.22)
DV-adjusted	8.98 (3.58)

4.6 Discussion

In this paper, we developed a method for estimating the probability of conversion efficiently and with high accuracy. In particular, we introduced a bias-adjusted estimator based on a simple (misspecified) logistic model that can incorporate information from the conversion delay distribution and also from users' visits. Our estimator requires far less time to compute than does the MLE, and is more accurate than the bias-adjusted estimator that uses the information in the delay distribution alone.

To simplify calculations, we assumed that the $T^*_{i,j}$'s and S_i are identically distributed, and that the V_i's are Poisson distributed. Extending our methods to allow for alternative distributional assumptions is a promising avenue for future research.

Since clicks have different associated true probabilities of conversion, the estimators of these probabilities (and their biases) have different variances. When computing the average bias, accounting for these differences by weighting each bias by its true SD could be desirable, especially when the range of the true probabilities is large. In our case, however, the difference in behaviour of the bias and weighted bias was minimal.

Another challenging question in display advertising is to assign the contribution of each ad displaying source (e.g., search, video, mobile app) when a conversion occurs. Currently, most algorithms give full credit to only the source associated with the last visit prior to conversion. Interesting future work would be to build on the model described in this paper and obtain a "fair" contribution algorithm.

4.7 Appendix

4.7.1 True likelihood

We find the contribution to the likelihood of the i^{th} click in two different cases.

- Case 1: $t_{i,o_i} + r_i < a_i$

$$\mathcal{L}_i\left(\boldsymbol{\beta}_c, \boldsymbol{\beta}_v, \boldsymbol{\beta}_t | r_i, o_i, \boldsymbol{t}_i^{o_i}\right) \quad = \quad p_i f_{V_i}(o_i) f_{\boldsymbol{T}_i^{O_i}}(\boldsymbol{t}_i^{o_i}) f_{S_i}(r_i) \tag{4.16}$$

- Case 2: $t_{i,o_i} + r_i = a_i$

$$\begin{aligned}
& \mathcal{L}_i\left(\boldsymbol{\beta}_c, \boldsymbol{\beta}_v, \boldsymbol{\beta}_t | r_i, o_i, \boldsymbol{t}_i^{o_i}\right) \\
= \quad & p_i f_{O_i, \boldsymbol{T}_i^{O_i}, R_i}(o_i, \boldsymbol{t}_i^{o_i}, r_i \mid C_i = 1) \\
+ \quad & (1-p_i) f_{O_i, \boldsymbol{T}_i^{O_i}, R_i}(o_i, \boldsymbol{t}_i^{o_i}, r_i \mid C_i = 0) \\
= \quad & p_i \sum_v f_{O_i, \boldsymbol{T}_i^{O_i}, R_i}(o_i, \boldsymbol{t}_i^{o_i}, r_i \mid C_i = 1, V_i = v) f_{V_i}(v) \\
+ \quad & (1-p_i) \sum_v f_{O_i, \boldsymbol{T}_i^{O_i}, R_i}(o_i, \boldsymbol{t}_i^{o_i}, r_i \mid C_i = 0, V_i = v) f_{V_i}(v) \\
= \quad & p_i f_{\boldsymbol{T}_i^{O_i}}(\boldsymbol{t}_i^{o_i})[1 - F_{S_i}(r_i)] \sum_v f_{V_i}(v) 1_{(o_i <= v)} \\
+ \quad & (1-p_i) f_{\boldsymbol{T}_i^{O_i}}(\boldsymbol{t}_i^{o_i}) \sum_v f_{V_i}(v) \left\{ 1_{(o_i = v)} + 1_{(o_i < v)} [1 - F_{T^*_{i,o_i}}(r_i)] \right\} \\
= \quad & p_i f_{\boldsymbol{T}_i^{O_i}}(\boldsymbol{t}_i^{o_i})[1 - F_{S_i}(r_i)] P(V_i \geq o_i) \\
+ \quad & (1-p_i) f_{\boldsymbol{T}_i^{O_i}}(\boldsymbol{t}_i^{o_i}) \left\{ f_{V_i}(o_i) + [1 - F_{V_i}(o_i)][1 - F_{T^*_{i,o_i}}(r_i)] \right\} \\
\overset{(*)}{=} \quad & f_{\boldsymbol{T}_i^{O_i}}(\boldsymbol{t}_i^{o_i}) p_i \left\{ [1 - F_{S_i}(r_i)] P(V_i \geq o_i) - f_{V_i}(o_i) - [1 - F_{V_i}(o_i)][1 - F_{S_i}(r_i)] \right\} \\
+ \quad & f_{\boldsymbol{T}_i^{O_i}}(\boldsymbol{t}_i^{o_i}) \left\{ f_{V_i}(o_i) + [1 - F_{V_i}(o_i)][1 - F_{S_i}(r_i)] \right\} \\
= \quad & f_{\boldsymbol{T}_i^{O_i}}(\boldsymbol{t}_i^{o_i}) p_i \left\{ [1 - F_{S_i}(r_i)] f_{V_i}(o_i) - f_{V_i}(o_i) \right\} \\
+ \quad & f_{\boldsymbol{T}_i^{O_i}}(\boldsymbol{t}_i^{o_i}) \left\{ f_{V_i}(o_i) + [1 - F_{V_i}(o_i)][1 - F_{S_i}(r_i)] \right\} \\
= \quad & f_{\boldsymbol{T}_i^{O_i}}(\boldsymbol{t}_i^{o_i}) \left\{ p_i [1 - F_{S_i}(r_i)] f_{V_i}(o_i) \right. \\
+ \quad & \left. (1-p_i) f_{V_i}(o_i) + [1 - F_{V_i}(o_i)][1 - F_{S_i}(r_i)] \right\}
\end{aligned}$$

(*): Note that the factor $1 - F_{S_i}(r_i)$ at the end of the line relies on the assumption that, for each i, the $T^*_{i,j}$'s and S_i are iid.

4.7.2 Conditional pdf of R_i

The conditional distribution of R_i under the true model is

$$f_{\boldsymbol{R}|\boldsymbol{O},\boldsymbol{T}}(\boldsymbol{r} \mid \boldsymbol{o}, \boldsymbol{t}) = \frac{f_{R_i, O_i, \boldsymbol{T}_i^{O_i}}(r_i, o_i, \boldsymbol{t}_i^{o_i})}{f_{O_i, \boldsymbol{T}_i^{O_i}}(o_i, \boldsymbol{t}_i^{o_i})}, \tag{4.17}$$

where the numerator is the true joint distribution given in (4.3). The denominator can be found by integrating the joint distribution over r_i:

$$\begin{aligned}
f_{O_i,\boldsymbol{T}_i^{O_i}}(o_i,\boldsymbol{t}_i^{o_i}) &= \int_0^{a_i-t_{i,o_i}} f_{R_i,O_i,\boldsymbol{T}_i^{O_i}}(r_i,o_i,\boldsymbol{t}_i^{o_i})dr_i \\
&= p_i f_{V_i}(o_i) f_{\boldsymbol{T}_i^{o_i}}(\boldsymbol{t}_i^{o_i}) F_{S_i}(a_i - t_{i,o_i}) \\
&+ f_{\boldsymbol{T}_i^{o_i}}(\boldsymbol{t}_i^{o_i}) \left\{ p_i[1 - F_{S_i}(a_i - t_{i,o_i})] f_{V_i}(o_i) \right. \\
&+ \left. (1-p_i) f_{V_i}(o_i) + [1 - F_{V_i}(o_i)][1 - F_{S_i}(a_i - t_{i,o_i})] \right\} \\
&= f_{\boldsymbol{T}_i^{o_i}}(\boldsymbol{t}_i^{o_i}) \left\{ f_{V_i}(o_i) + [1 - F_{V_i}(o_i)][1 - F_{S_i}(a_i - t_{i,o_i})] \right\} \qquad (4.18)
\end{aligned}$$

4.7.3 CKLIC computation

$$\begin{aligned}
& CKLIC\left(\boldsymbol{\alpha};\boldsymbol{\beta}_c,\boldsymbol{\beta}_o,\boldsymbol{\beta}_t\right) \\
= \; & E\left\{\ln\left[\frac{\mathcal{L}\left(\boldsymbol{\beta}_c,\boldsymbol{\beta}_v,\boldsymbol{\beta}_t|\boldsymbol{R},\boldsymbol{O},\boldsymbol{T}^{\boldsymbol{O}}\right)}{L_f\left(\boldsymbol{\alpha}|\boldsymbol{R},\boldsymbol{O},\boldsymbol{T}^{\boldsymbol{O}}\right)}\right]\right\} \\
= \; & E\left\{\ln\left[\frac{\mathcal{L}\left(\boldsymbol{\beta}_c,\boldsymbol{\beta}_v,\boldsymbol{\beta}_t|\boldsymbol{T},\boldsymbol{O},\boldsymbol{T}^{\boldsymbol{O}}\right)/f_{\boldsymbol{O},\boldsymbol{T}}(\boldsymbol{O},\boldsymbol{T})}{L_f\left(\boldsymbol{\alpha}|\boldsymbol{R},\boldsymbol{O},\boldsymbol{T}^{\boldsymbol{O}}\right)/f_{\boldsymbol{O},\boldsymbol{T}}(\boldsymbol{O},\boldsymbol{T})}\right]\right\} \\
= \; & E\left\{\ln\left[\frac{\mathcal{L}\left(\boldsymbol{\beta}_c,\boldsymbol{\beta}_v,\boldsymbol{\beta}_t|\boldsymbol{R}\;;\;\boldsymbol{O},\boldsymbol{T}^{\boldsymbol{O}}\right)}{L_f\left(\boldsymbol{\alpha}|\boldsymbol{R}\;;\;\boldsymbol{O},\boldsymbol{T}^{\boldsymbol{O}}\right)}\right]\right\} \\
= \; & E\left\{\ln\left[\mathcal{L}\left(\boldsymbol{\beta}_c,\boldsymbol{\beta}_v,\boldsymbol{\beta}_t|\boldsymbol{R}\;;\;\boldsymbol{O},\boldsymbol{T}^{\boldsymbol{O}}\right)\right]\right\}-\sum_{i=1}^{N}E\left\{\ln\left[L_f\left(\boldsymbol{\alpha}|R_i\;;\;O_i,\boldsymbol{T}_i^{O_i}\right)\right]\right\} \\
= \; & \text{constant} \\
- \; & \sum_{i=1}^{N}\left\{\int_0^{a_i-t_{i,o_i}}\ln[f(r_i\mid\boldsymbol{o}_i,\boldsymbol{t}_i^{\boldsymbol{o}_i}\;;\;\boldsymbol{\alpha})]g(r_i\mid o_i,\boldsymbol{t}_i^{o_i}\;;\;\boldsymbol{\beta}_c)dr_i\right\} \\
= \; & \text{constant} \\
- \; & \sum_{i=1}^{N}\left\{\frac{\ln(\theta_i)}{f_{O_i,\boldsymbol{T}_i^{O_i}}(o_i,\boldsymbol{t}_i^{o_i})}\int_0^{a_i-t_{i,o_i}}f_{R_i,O_i,\boldsymbol{T}_i^{O_i}}(r_i,o_i,\boldsymbol{t}_i^{o_i})dr_i\right. \\
+ \; & \left.\frac{\ln(1-\theta_i)}{f_{O_i,\boldsymbol{T}_i^{O_i}}(o_i,\boldsymbol{t}_i^{o_i})}\left[1-\int_0^{a_i-t_{i,o_i}}f_{R_i,O_i,\boldsymbol{T}_i^{O_i}}(r_i,o_i,\boldsymbol{t}_i^{o_i})dr_i\right]\right\} \\
= \; & \text{constant} \\
- \; & \sum_{i=1}^{N}\left\{\ln(\theta_i)\frac{p_i f_{V_i}(o_i)f_{\boldsymbol{T}_i^{o_i}}(\boldsymbol{t}_i^{o_i})F_{S_i}(a_i-t_{i,o_i})}{f_{\boldsymbol{T}_i^{o_i}}(\boldsymbol{t}_i^{o_i})\left\{f_{V_i}(o_i)+[1-F_{V_i}(o_i)][1-F_{S_i}(a_i-t_{i,o_i})]\right\}}\right. \\
+ \; & \left.\ln(1-\theta_i)\left[1-\frac{p_i f_{V_i}(o_i)f_{\boldsymbol{T}_i^{o_i}}(\boldsymbol{t}_i^{o_i})F_{S_i}(a_i-t_{i,o_i})}{f_{\boldsymbol{T}_i^{o_i}}(\boldsymbol{t}_i^{o_i})\left\{f_{V_i}(o_i)+[1-F_{V_i}(o_i)][1-F_{S_i}(a_i-t_{i,o_i})]\right\}}\right]\right\} \\
= \; & \text{constant} \\
- \; & \sum_{i=1}^{N}\left\{\ln(\theta_i)\frac{p_i f_{V_i}(o_i)F_{S_i}(a_i-t_{i,o_i})}{f_{V_i}(o_i)+[1-F_{V_i}(o_i)][1-F_{S_i}(a_i-t_{i,o_i})]}\right. \\
+ \; & \left.\ln(1-\theta_i)\left[1-\frac{p_i f_{V_i}(o_i)F_{S_i}(a_i-t_{i,o_i})}{f_{V_i}(o_i)+[1-F_{V_i}(o_i)][1-F_{S_i}(a_i-t_{i,o_i})]}\right]\right\} \qquad (4.19)
\end{aligned}$$

Chapter 5

Conclusion

In this thesis, we consider model parameter estimation problems where the true underlying models are complex. We build estimators based on a simpler, misspecified models, and adjust for bias in the estimators, if necessary.

The focus of Chapter 2 is the estimation of the regression coefficients in models for time series of count data. We consider a general class of parameter-driven models for such data. While this class is highly flexible (and includes some common models as special cases), computation of the MLE of the regression coefficients of models in this class can be challenging. We study the behaviour of three simple estimators of the regression coefficients. We show that the estimator based on the Poisson generalized linear model performs remarkably well in terms of bias and efficiency, even if the data are overdispersed or autocorrelated. We also derive a SE for our estimators that is simpler and more accurate than those suggested in the literature. We show how our methods and results can be applied in practice, and include a detailed analysis of polio and epileptic seizure data sets.

The context of Chapter 3 is display advertising. Our focus is the development of a computationally efficient and accurate estimator of the probability that a click on an online ad will convert. We show how to obtain a consistent estimate of this probability based on a simple logistic regression model (which ignores the conversion delay times) and a bias adjustment determined using the KLIC approach (which uses the information in the conversion delay distribution). With a simulation study, we show that our adjusted estimator (which we call the D-adjusted estimator) has relatively low bias and computational time, even when the bias adjustment is based on incorrect assumptions. We apply these findings to the problem of estimating the probability of conversion in a real data set.

In Chapter 4, we extend our estimator in Chapter 3 to allow for visits to the ad after the click of a user. We propose a method for estimating the probability of conversion with high accuracy. In particular, we use the KLIC approach to obtain a bias-adjusted estimator based on a simple (misspecified) logistic model. In this way, we incorporate information from both the conversion delay and inter-visit time distributions. This new estimator has a reasonable computational time (relative to that of the MLE of the true model), and has a smaller bias than that of either the maximizer of the logistic regression model or the D-adjusted estimator.

In conclusion, we show that, for some problems, the maximizers of the likelihoods of simple, misspecified models are accurate, efficient estimators parameters of the true models. In other cases, where these estimators are biased for the parameters of interest, we can examine the true underlying model, and employ bias correction techniques (e.g., the KLIC approach and the penalized likelihood approach of Firth [1993]). By adjusting the estimators and their SEs in this way, we can obtain approximately unbiased estimators with accurate SEs.

Bibliography

D. Agarwal, R. Agrawal, and R. Khanna. Estimating rates of rare events with multiple hierarchies through scalable log-linear models. *The 16th ACM SIGKDD international conference on Knowledge discovery and data mining.*, pages 213–222, 2010.

P. S. Albert. A two-state markov mixture model for a time series of epileptic seizure counts. *Biometrics*, 47:1371–1381, 1991.

a Microsoft Subsidiary Atlas. Engagement mapping: A new measurement standard is emerging for advertisers. *Microsoft Atlas Institute.*, 2, 2008.

C. Bates and H. White. A unified theory of consistent estimation for parametric models. *Econometric Theory*, 1(2):151–178, 1985.

Robert H. Berk. Limiting behavior of posterior distributions when the model is incorrect. *Ann. Math. Statist.*, 37(1):51–58, 1966.

Robert H. Berk. Consistency a posteriori. *Ann. Math. Statist.*, 41(3):894–906, 1970.

Ron Berman. Beyond the last touch: attribution in online advertising. *Social Media Marketing Industry Report*, 2013.

Cesar A. Brea. Marketing and sales analytics: proven techniques and powerful applications. *Pearson Education LTD*, 2014.

O. Chapelle. Modeling delayed feedback in display advertising. *20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2014)*, 1:1097–1105, 2014.

O. Chapelle, E. Manavoglu, and R. Rosales. Simple and scalable response prediction for display advertising. *ACM Transactions on Intelligent Systems and Technology.*, 5, 2014.

C. Gregory Chow. Maximum-likelihood estimation of misspecified models. *Economic Modelling*, 1(2):134–138, 1984.

D. R. Cox. Statistical analysis of time series: some recent developments. *Scandinavian Journal of Statistics*, 8:93–115, 1981.

R. A. Davis and R. Wu. A negative binomial model for time series of counts. *Biometrika*, 96:735–749, 2009.

R. A. Davis and G. R. Yam. Parameter- and observation-driven state space models. *Lecture available at http://www.stat.columbia.edu/~rdavis/lectures/Cyprus1_04.pdf*, 2004.

R. A. Davis, W. T. Dunsmuir, and Y. Wang. On autocorrelation in a poisson regression model. *Biometrika*, 87:491–505, 2000.

Eustache Diemert, Julien Meynet, Pierre Galland, and Damien Lefortier. Attribution modeling increases efficiency of bidding in display advertising. *The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2017, New York, NY, USA., ACM.*, 2017.

D. Firth. Bias reduction of maximum likelihood estimated. *Biometrika*, 80:27–38, 1993.

Paul Gustafson. On measuring sensitivity to parametric model misspecification. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(1):81–94, 2001.

J. Patrick Heagerty and F. Brenda Kurland. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, 88(4):973–985, 2001.

P. J. Heagerty and S. L. Zeger. Marginalized multilevel model and likelihood inference. *Statist. Sci.*, 15:1–26, 2000.

D. Hillard, S. Schroedl, and E. Manavoglu. Improving ad relevance in sponsored search. *The 3rd ACM international conference on Web search and data mining.*, pages 361–370, 2010.

H. Hirose. Bias correction for the maximum likelihood estimates in the two-parameter weibull distribution. *IEEE Transactions on Dielectrics and Electrical Insulation.*, 6:66–69, 1999.

P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. *the 5th Berkeley Symposium on Mathematical Statistics and Probability.*, 1967.

Wendi Ji, Xiaoling Wang, and Dell Zhang. A probabilistic multi-touch attribution model for online advertising. *The 25th ACM International on Conference on Information and Knowledge Management.*, pages 1373–1382, 2016.

P. Jordan, M. Mahdian, and S. Vassilvitskii. The multiple attribution problem in pay-per-conversion advertising. *Algorithmic Game Theory (SAGT 2011)*, 2011.

N. D. Le, B. G. Leroux, and M. L. Puterman. Reader reaction: Exact likelihood evaluation in a markov mixture model for time series of seizure counts. *Biometrics*, 48:317–323, 1992.

Y. Lee and J. A. Nelder. Hierarchical generalized linear models. *Journal of the Royal Statistical Society*, 58:619–678, 1996.

Hongshuang Li and PK Kannan. Attributing conversions in a multichannel online marketing environment: An empirical model and a field experiment. *Journal of Marketing Research.*, 51(1):40–56, 2014.

K. Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986.

T. Liboschik, K. Fokianos, and R. Fried. tscount: An r package for analysis of count time series following generalized linear models. *Mathematical and Computer Modelling*, 82(2): 1–51, 2017.

C. Lystig and J. P. Hughes. Exact computation of the observed information matrix for hidden markov models. *Journal of Computational and Graphical Statistics*, 11(3):678–689, 2002.

P. McAfee, J. McMillan, and S. Wilkie. The greatest auction in history. *Cambridge, MA: Harvard University Press.*, pages 168–184, 2010.

R. P. McAfee. The design of advertising exchanges. *Review of Industrial Organization*, 39 (3):169–185, 2011.

H. B. McMahan, G. Holt, and D. Sculley. Ad click prediction: a view from the trenches. *19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1222–1230, 2013.

S. Muthukrishnan. Ad exchange: Research issues. *5th International Workshop on Internet and Network Economics*, 1:1097–1105, 2009.

K. P. Nelson and B. G. Leroux. Properties and comparison of estimation methods in a log-linear generalized linear mixed model. *Journal of Statistical Computation and Simulation*, 78(3):367–384, 2008.

J. M. Neuhaus, W. W. Hauk, and J. D. Kalbfleisch. The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika*, 79(4):755–762, 1992.

J. M. Neuhaus, C. E. McCulloch, and R. Boylan. Estimation of covariate effects in generalized linear mixed models with a misspecified distribution of random intercepts and slopes. *Statistics in Medicine*, 32:2419–2429, 2013.

A. N. Pettitt. Bias correction for censored data with exponential lifetimes. *Statistica Sinica*, 8:941–963, 1998.

Dimitris Rizopoulos and Geert Verbeke. Shared parameter models under random effects misspecification. *Biometrika*, 95(1):63–74, 2008.

R. Rosales, H. Cheng, and E. Manavoglu. Post-click conversion modeling and analysis for non-guaranteed delivery display advertising. *fifth ACM international conference on Web search and data mining*, pages 293–302, 2012.

A. Safari, R. M. Altman, and T. M. Loughin. Efficient conversion probability estimator for display advertising. *In preparation.*

Y. Shen and Z. L. Yang. Bias-correction for weibull common shape estimation. *Journal of Statistical Computation and simulation*, pages 3017–3046, 2014.

Y. Shen and ZL Yang. Bias-correction for weibull common shape estimation. *Journal of Statistical Computation and simulation.*, 85(15):3017–3046, 2015.

K. Takeuchi. Distribution of information statistics and criteria for adequacy of models. *Math. Sci.*, 153:12–18, 1976.

M. Tan, Y. Qu, and J. S. Rao. Robustness of the latent variable model for correlated binary data. *Biometrics*, 55:258–263, 1999.

T. R. ten Have, A. R. KUNSELMAN, and L. TRAN. A comparison of mixed effects logistic regression models for binary response data with two nested levels of clustering. *Statist. Med.*, 18:947–960, 1999.

T. Tietenberg. The evolution of emissions trading. *Cambridge, MA: Harvard University Press.*, pages 42–58, 2010.

H. Varian. Position auctions. *International Journal of Industrial Organization.*, 25:1163–1178, 2007.

Xiaomin Wan, Liubao Peng, and Yuanjian Li. A review and comparison of methods for recreating individual patient data from published kaplan-meier survival curves for economic evaluations: A simulation study. *PLoS One*, 2015a.

Xiaomin Wan, Liubao Peng, and Yuanjian Li. A review and comparison of methods for recreating individual patient data from published kaplan-meier survival curves for economic evaluations: A simulation study. *Hills RK, ed. PLoS ONE.*, 10(3), 2015b.

Christian H. Weib. Modelling time series of counts with overdispersion. *Statistics Methods Application*, 40:507–519, 2009.

H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.

H. White. *Maximum Likelihood Estimation of Misspecified Dynamic Models*. Springer, 1984.

S. L. Zeger. A regression model for time series of counts. *Biometrika*, 75:621–629, 1988.

S. L. Zeger and K. Y. Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42:121–130, 1986.

Fukang Zhu. Modeling time series of counts with com-poisson ingarch models. *Mathematical and Computer Modelling*, 56(9):191–203, 2012.

Appendix A

Supplementary materials

Supplementary materials for all the three papers are available at
http://researchdata.sfu.ca/pydio$_public/0734e$7