OPTIMIZATION AND STATISTICAL ESTIMATION FOR THE POST RANDOMIZATION METHOD

A Dissertation in
Statistics
by
Yong Ming Jeffrey Woo

© 2013 Yong Ming Jeffrey Woo

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2013
The dissertation of Yong Ming Jeffrey Woo was reviewed and approved* by the following:

Aleksandra B. Slavković  
Associate Professor of Statistics and Public Health Sciences  
Dissertation Co-Advisor, Co-Chair of Committee

Donald P. Richards  
Professor of Statistics  
Dissertation Co-Advisor, Co-Chair of Committee

Naomi S. Altman  
Professor of Statistics

Stephanie T. Lanza  
Research Associate Professor of Health and Human Development

David Hunter  
Professor of Statistics  
Department Head

*Signatures are on file in the Graduate School.
Abstract

The field of Statistical Disclosure Control (SDC) aims at developing methodology that balances the objectives of providing data for valid statistical inference and safeguarding confidential information. One of the SDC methods for categorical variables is the Post Randomization Method (PRAM). The basic idea underlying PRAM is to misclassify values of the categorical variables, via a known probability mechanism captured by a PRAM matrix. This thesis focuses on three primary methodological developments that enable PRAM to become a more theoretically and practically viable SDC method.

First, we focus on the issue of obtaining valid statistical analysis with data subject to PRAM. The application of PRAM is known to produce biased parameter estimates in generalized linear models (GLMs). We develop and implement EM-type algorithms that take into account the effect of PRAM and obtain asymptotically unbiased estimators of parameters in GLMs, when both covariates and response variables are subject to PRAM. The basic ideas are based on the “EM by method of weights” in the missing data literature. Second, we extend the proposed methodology in order to deal with dependent covariates when estimating parameters in GLMs by relaxing the assumption of independence of covariates. This is done by modeling the distribution of the covariates subject to PRAM as a product of univariate conditional distributions. This approach advances the PRAM methodology by making it more applicable in practice and results in more accurate estimators of the regression parameters. Results from simulation studies and application to the 1993 Current Population Survey are presented. Lastly, we address the issue of obtaining optimal PRAM matrices which produce safe files and maximize data utility with respect to a widely-used utility measure for PRAM: entropy-based information loss, \( EBIL \), a variant of Shannon’s entropy. We show that for a certain class of PRAM matrices, \( EBIL \) displays monotonic properties, which implies the minimum of \( EBIL \) occurs at an extreme point of the convex region that satisfies a pre-determined rule for safe files. Using these
properties, we present an algorithm that obtains PRAM matrices which produce safe files with higher data utility when compared to PRAM matrices obtained using built-in numerical methods and routines.
# Table of Contents

List of Figures viii

List of Tables ix

Acknowledgments xi

Chapter 1
   Introduction to Statistical Disclosure Control 1
      1.1 Issues in SDC 3
      1.2 Contributions of Dissertation 5

Chapter 2
   Overview of Statistical Disclosure Control (SDC) 7
      2.1 SDC Terminology 7
      2.2 SDC Methods for Categorical Data 8
      2.3 Disclosure Risk and Data Utility 13
      2.4 SDC Software 17

Chapter 3
   Overview of PRAM 19
      3.1 Randomized Response 20
      3.2 The Post Randomization Method (PRAM) 22
         3.2.1 The Method-of-Moments Estimator for Population Proportions 23
         3.2.2 The Maximum Likelihood Estimator for Population Proportions 24
         3.2.3 Invariant PRAM 28
      3.3 Methodological Challenges with PRAM 29
         3.3.1 Issues with Estimation 29
         3.3.2 Issues with Dependencies between Variables 32
         3.3.3 Issues with Obtaining Optimal PRAM Matrices 33
      3.4 Proposed Work & Contributions 38
### Chapter 4
**Generalized Linear Model Estimation with PRAM, with Independent Covariates**

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>40</td>
</tr>
<tr>
<td>4.2</td>
<td>Generalized Linear Models with Variables Subject to PRAM</td>
<td>43</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Categorical Covariate Subject to PRAM</td>
<td>44</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Response Variable Subject to PRAM</td>
<td>47</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Covariate and Response Subject to PRAM</td>
<td>48</td>
</tr>
<tr>
<td>4.3</td>
<td>Generalized Linear Model Estimation with Variables Subject to PRAM</td>
<td>49</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Example I: Logistic Regression</td>
<td>50</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Example II: Poisson Regression</td>
<td>56</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Varying the Probability of Success &amp; Distribution of Covariates</td>
<td>60</td>
</tr>
<tr>
<td>4.3.4</td>
<td>Applying PRAM to Two Independent Covariates</td>
<td>61</td>
</tr>
<tr>
<td>4.4</td>
<td>Application to 1993 CPS Dataset</td>
<td>63</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Disclosure Risk Assessment</td>
<td>66</td>
</tr>
<tr>
<td>4.5</td>
<td>Discussion</td>
<td>69</td>
</tr>
</tbody>
</table>

---

### Chapter 5
**Generalized Linear Model Estimation with PRAM, with Dependent Covariates**

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>73</td>
</tr>
<tr>
<td>5.2</td>
<td>Generalized Linear Models with Dependent Categorical Covariates Subject to PRAM</td>
<td>75</td>
</tr>
<tr>
<td>5.3</td>
<td>Comparison of EM Algorithms, With and Without Assumption of Independence</td>
<td>79</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Algorithm IV and Algorithm I With Dependent Data</td>
<td>80</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Algorithm IV and Algorithm I With Independent Data</td>
<td>84</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Covariates as Series of Conditional Distributions and as Joint Distribution</td>
<td>86</td>
</tr>
<tr>
<td>5.4</td>
<td>Application to CPS Data</td>
<td>88</td>
</tr>
<tr>
<td>5.5</td>
<td>Discussion</td>
<td>91</td>
</tr>
</tbody>
</table>

---

### Chapter 6
**Optimal PRAM with Respect to Disclosure Control**

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>95</td>
</tr>
<tr>
<td>6.2</td>
<td>Motivation</td>
<td>96</td>
</tr>
<tr>
<td>6.3</td>
<td>Notation and Definitions</td>
<td>98</td>
</tr>
<tr>
<td>6.3.1</td>
<td>A Measure of Disclosure Risk</td>
<td>98</td>
</tr>
<tr>
<td>6.3.2</td>
<td>A Measure of Data Utility</td>
<td>99</td>
</tr>
<tr>
<td>6.4</td>
<td>Obtaining Optimal PRAM Matrix for a Binary Variable</td>
<td>99</td>
</tr>
</tbody>
</table>
List of Figures

2.1 Risk-Utility map. ......................................................... 13

4.1 Plot of coefficient estimates for logistic regression with covariate subject to PRAM, when $n = 10000$, $p = 0.90$. .......................... 55
4.2 Plot of coefficient estimates for logistic regression with covariate subject to PRAM, when $n = 10000$, $p = 0.80$. .......................... 56
4.3 Plot of coefficient estimates for Poisson regression with covariate subject to PRAM, when $n = 10000$, $p = 0.90$. .......................... 59
4.4 Plot of coefficient estimates for Poisson regression with covariate subject to PRAM, when $n = 10000$, $p = 0.80$. .......................... 59
4.5 Plot of estimates for coefficient of marital status, Case 1: marital subject to PRAM. .................................................. 66
4.6 Plot of estimates for coefficient of marital status, Case 2: salary subject to PRAM. .................................................. 66
4.7 Plot of estimates for coefficient of marital status, Case 3: marital and salary subject to PRAM. .................................................. 67

5.1 Plot of estimates for $\beta_1$, using Algorithm I and Algorithm IV. ...... 84
5.2 Plot of estimates for $\beta_3$, when Marital Status subject to PRAM. Estimates from Algorithm I (left) and Algorithm IV(right). ........... 91
5.3 Plot of estimates for $\beta_0$, when Marital Status subject to PRAM. ...... 91
5.4 Plot of estimates for $\beta_1$, when Marital Status subject to PRAM. ...... 92
5.5 Plot of estimates for $\beta_2$, when Marital Status subject to PRAM. ...... 92

6.1 Risk-Utility map. ......................................................... 97
6.2 Plot of Entropy-Based Information Loss versus diagonal entries in PRAM matrix. .................................................. 104
List of Tables

4.1 Average bias, standard error, mean squared error, & closeness proportion of MLEs for logistic regression with variables subject to PRAM. ......................................................... 52
4.2 Average bias, standard error, mean squared error, & closeness proportion of MLEs for Poisson regression with covariate subject to PRAM. ......................................................... 58
4.3 Average bias of MLEs with closeness proportion in parentheses, when accounting for PRAM applied to a single categorical covariate, with distribution of covariate and probability of success varied. .......................... 60
4.4 Average bias, with closeness proportion in parentheses, of MLEs for logistic regression with two independent covariates. ................................................................. 62
4.5 Parameter Estimates from Original Data (O.D.), data subject to PRAM without EM algorithms ($\hat{\beta}_{k, \text{noadjust}}$), and data subject to PRAM with EM Algorithms ($\hat{\beta}_{k, \text{adjust}}$). Average ML estimates with standard errors in parentheses for O.D.. Average ML estimates with closeness proportion in parentheses for $\hat{\beta}_{k, \text{noadjust}}$ and $\hat{\beta}_{k, \text{noadjust}}$. Case 1: marital subject to PRAM; Case 2: salary subject to PRAM; Case 3: both marital and salary subject to PRAM. ........................................ 64
4.6 Three-way Table for the CPS 1993 data by Marital, Sex and Race. For $d = 800$, values in parenthesis are $\hat{R}_{PRAM}(k)$ and $T_{k}^{(d)}$, respectively. ........................................... 69

5.1 Average bias, standard error, mean squared error, & closeness proportion of MLEs for logistic regression when dependent covariates subject to PRAM. Comparison between Algorithm I, which assumed independence, and Algorithm IV, which did not assume independence. 82
5.2 Average bias, standard error, mean squared error, & closeness proportion of MLEs for logistic regression when independent covariates subject to PRAM. Comparison between Algorithm I, which assumed independence, and Algorithm IV, which did not assume independence. ........................................ 85
5.3 Average bias, standard error, mean squared error, & closeness proportion of MLEs for logistic regression when independent covariates subject to PRAM. Comparison between specifying distribution of covariates subject to PRAM, as series of conditional distributions, and as joint distribution. ................................. 87

5.4 Average bias, standard error, mean squared error, & closeness proportion of parameter estimates from data subject to PRAM without using either EM algorithm (\(\hat{\beta}_{k,\text{noadjust}}\)), data subject to PRAM with EM Algorithm I assuming independence of covariates (\(\hat{\beta}_{k,\text{ind}}\)), and data subject to PRAM with EM Algorithm IV without assumption of independence of covariates (\(\hat{\beta}_{k,\text{dep}}\)). ............................. 89

6.1 Results comparing our methodology with built in constrOptim routine in R. ................................................................. 122
Acknowledgments

First and foremost, I can never thank my advisors Dr. Sesa Slavković and Dr. Don Richards enough, for working closely with me and for their advice, encouragement, and help in getting my dissertation done. They have gone beyond the call of duty and my success is possible because of them.

I also want to thank my committee members, Dr. Naomi Altman and Dr. Stephanie Lanza, for reading my dissertation and for their valuable comments and suggestion for future work.

I would also like to express my thanks to Bill Winkler, Bimal Sinha, and Tommy Wright at the U.S. Census Bureau for their suggestions and comments on my research.

The Department of Statistics at Penn State is great because of the people that make up the department. I would like to extend my gratitude to: Dr. Jim Rosenberger and Byran Smucker, for their help when I was applying for the U.S. Census Bureau Fellowship; Dr. Sonja Petrovic, Dr. Murali Haran, and Vishesh Karwa, for their constructive comments during my practice presentation; and Lucia Tabacu for subbing in for me as an instructor while I was away presenting at conferences. I also want to thank my friends at the department and at Penn State for making my time at PSU enjoyable.

Thanks also to my friends at Stepping Stones Community Church, for their prayers, support, and encouragement.

And of course I want to thank the people that have always been there for me: my parents, Bih Li and Choi Yin, and my sister, Sharon. It is always nice having a reliable source of food from back home mailed to Happy Valley.

This work was supported in part by: United States Census Bureau Dissertation Fellowship YA1323-12-SE-0303 and NSF grant BCS-0941553.
Chapter 1

Introduction to Statistical Disclosure Control

Statistical agencies face the challenge of providing high-quality data products to aid statistical research, while also guarding against the risk of disclosing confidential information on individual respondents. The field of Statistical Disclosure Control (SDC) has been developed to balance the objectives of providing data for valid inference and safeguarding confidential information.

The development of SDC methodology and its practice are important for many reasons. First, statisticians have a moral and professional obligation to safeguard the release of data that will allow a subject’s confidential information to be disclosed, e.g., International Statistical Institute’s Declaration on Professional Ethics (2010).

Second, if subjects have confidence that their confidential information are adequately protected, they will be more inclined to participate in surveys. Subject confidence in data confidentiality is a legitimate issue for statistical agencies, as concerns about threats to data privacy have led to the postponement of the census in the former West Germany, and even the abolition of the census in the
Netherlands in 1971. The conventional census in the Netherlands has since been replaced by a virtual census in 2001 and 2011, which makes use of administrative registers and household sample surveys; see van der Laan (2000) and Nordholt (2005).

Third, there is an increase in demand for more detailed information, which makes it easier for identities or sensitive attributes to be disclosed. Statistical agencies can ensure confidentiality by not releasing data for public use, but such a practice will significantly limit researchers from accessing the massive amounts of data and running statistical analyses that can benefit society.

Fourth, with the improvement in computer technology, researchers are able to perform analyses on large microdata files on their own, without needing a statistical agency to summarize the data for them. However, with the higher detail in datasets and improvement in computer technology, there is greater risk of disclosure; see Willenborg and de Waal (1996), Reiter (2004), Fienberg and Jin (2007), Museux et al. (2008), Abowd et al. (2009), Slavković (2009), McCaa et al. (2010) and Ramanayake and Zayatz (2010).

There are also legal issues regarding SDC. For example, Title 13 of the United States Code, the law under which the United States Census Bureau operates, guarantees the confidentiality of census information and establishes penalties for disclosing such information (Zayatz, 2002). In the United Kingdom, the Census Act of 1991 governs the U.K. Census Offices, and states the disclosure of any personal census information without proper authorization is an offense (Longhurst et al., 2007). In the Netherlands, the Registration Chamber enforces the Personal Data Protection act, which regulates the maintenance and use of registration of personal data (van der Laan, 2000). Several legal acts have been passed in the European Union (E.U.) to protect the right to privacy in personal data processing (Museux
et al., 2008).

Thus, there are professional, practical, and legal reasons, for statistical agencies to safeguard the confidentiality of information. Many methods have been developed to reduce the risk of disclosing confidential information, and these methods result in publishing redacted data in place of the original data. While the redacted data reduces the risk of disclosing confidential information, challenges regarding obtaining valid statistical inference arise.

1.1 Issues in SDC

Consider the following example. Suppose that a microdata file is released for public use. In this file, there is a respondent with the following values: \textit{Occupation} = \textit{Mayor}, \textit{Residence} = \textit{Boalsburg}, and \textit{CriminalRecord} = yes. Since such a combination is rare (and indeed unique, assuming there is only one mayor of Boalsburg), the public is able to learn something new about this particular individual even if their name is not published in the file, and their privacy is at risk as a result. SDC methods have been developed to prevent disclosure of such individual information. Perhaps instead of publishing this individual’s occupation as ‘mayor’, it could be published instead as ‘governmental service’. As a result, all government employees will be labeled accordingly under \textit{Occupation}.

When SDC methods are applied to a dataset, typically only the perturbed or masked dataset will be published; the original dataset will not be released for use by others. An obvious drawback is that the utility of the perturbed data has possibly been decreased. In other words, will statistical analyses on the perturbed dataset provide results similar to analyses on the original dataset, and if the results are not similar, what kind of adjustments should be made so that similar results can be
achieved?

One of the perturbation-based SDC methods is the Post Randomization Method (PRAM), which is originally proposed by Gouweleeuw et al. (1998). The main idea behind PRAM is to publish redacted data after the values of categorical variables in the original dataset have been misclassified by a known probability mechanism. This probability mechanism is described by a transition matrix, i.e., PRAM matrix.

PRAM provides advantages compared to other SDC methodologies: unlike non-perturbative methods which lead to loss of detail, PRAM maintains detail in the variables; and unlike most perturbative methods, the application of PRAM is probabilistic in nature and hence intruders cannot determine which records have been perturbed. While PRAM provides certain advantages compared to other SDC methodologies, it has seen limited use in practice due to a number of unresolved issues related to both disclosure risk and data utility. The three most commonly discussed issues are:

(1) Adjustments to estimates of frequency counts and models needed when performing statistical analyses (e.g., de Wolf et al. (1998), Gouweleeuw et al. (1998), van den Hout and van der Heijden (2002), van den Hout and Elamir (2006), van den Hout and Kooiman (2006));

(2) Dealing with dependent variables and inconsistencies in the perturbed dataset after the application of PRAM (e.g., de Wolf et al. (1998), Gouweleeuw et al. (1998), Shlomo and de Waal (2008)); and

(3) Finding optimal PRAM matrices that result in a file that safeguards against disclosure of confidential information and maximizes data utility (e.g., Cator et al. (2005), de Wolf et al. (1998), de Wolf and van Gelder (2004), Mares and Torra (2010), Shlomo and Skinner (2010)).
1.2 Contributions of Dissertation

The rest of this dissertation is organized as follows. Chapter 2 introduces the key
terms, methods and available software in statistical disclosure control; Chapter 3
introduces the Post Randomization Method (PRAM) and reviews the seminal
paper by Gouweleeuw et al. (1998), as well as an overview on PRAM research since
its introduction in Gouweleeuw et al. (1998). Chapters 4, 5, and 6 focus on three
primary methodological developments in this thesis.

In Chapter 4, we focus on the issue of obtaining valid statistical analysis with
data subject to PRAM. The application of PRAM is known to produced biased
parameter estimates in generalized linear models (GLMs); therefore the effect of
PRAM has to be taken into account. We develop and implement EM-type
algorithms to obtain asymptotically unbiased estimators, that is the maximum
likelihood estimators of parameters in GLMs, when chosen variables are subject
to PRAM. The basic ideas are based on the “EM by method of weights” developed by
Ibrahim (1990) for GLMs with covariates missing at random, and on the approach
proposed by van den Hout and Kooiman (2006) for linear regression with covariates
subject to randomized response. There is an extensive literature for missing data
with either missing covariate or missing response variable. We extend these ideas by
developing an EM-type algorithm that obtains unbiased estimates of GLMs when
both covariate and response variables are subject to PRAM. This is a more difficult
problem than either case when covariates or response variables are subject to
PRAM, and has received little attention in statistical literature.

In Chapter 5, we address the issue of dealing with dependent variables. Building
on the results from Chapter 4, we relax the condition of independence of covariates.
A more complex EM algorithm to estimate GLMs when chosen variables are subject
to PRAM is needed to take into account the dependence of covariates. This is done by modeling the distribution of the covariates subject to PRAM as a product of univariate conditional distributions. This approach advances the methodology by making it more applicable in practice by allowing dependence between the covariates, and results in more accurate estimators of the regression parameters.

In Chapter 6, we address the issue of obtaining optimal PRAM matrices, which produce safe files and maximize data utility upon their application to the data. One widely-used measure of data utility for PRAM is the entropy-based information loss (EBIL) (Domingo-Ferrer and Torra, 2001a), which is a variant of Shannon’s entropy. The mathematical properties of EBIL have received little attention in the literature, possibly because it consists of many variables which make mathematical derivations complicated. Thus far, most evaluations of the effect of PRAM on EBIL have been done computationally; see Domingo-Ferrer and Torra (2001a), de Wolf and van Gelder (2004), and also Mares and Torra (2010), who noted that these evaluations were computationally expensive. We establish monotonic properties of EBIL for certain classes of PRAM matrices. As such, an optimal PRAM matrix, one which produces a safe file and maximizes data utility, can be derived explicitly and results in safe files with higher data utility than using built-in numerical methods and routines.
Chapter 2

Overview of Statistical Disclosure Control (SDC)

2.1 SDC Terminology

In this section we review important definitions from the field of Statistical Disclosure Control (SDC). Many of these definitions can be found in Hundepool et al. (2010), Willenborg and de Waal (1996) and Fienberg and Jin (2007).

*Confidentiality* is the promise not to share data in an identifiable form.

*Disclosure* happens when something new can be learned about a subject through released data. *Identity disclosure* happens when the identity of a subject can be linked with data containing confidential information. *Attribute disclosure* happens with an attribute value in the data can be linked with the subject. *Inferential disclosure* happens when information can be inferred with a high degree of confidence from the released data.

*Microdata* is the file that contains the data at the individual level. Typically, each row corresponds to each individual, while each column corresponds to a variable.
Aggregated data refer to summary statistics that result from performing statistical analyses on microdata.

Keys are defined by the cross-classification of categorical variables. Disclosure risk measurements are usually a function of keys.

Data utility refers to the ability to get similar results when doing statistical analyses on the released data and on the original data. In some of the literature, information loss is a term that is used instead.

Direct identifying variables are variables that can lead immediately to a certain reidentification of a unit. As such, direct identifiers should not be published with the released data. Some examples are name and address.

Indirect identifying variables are variables that, based on a single variable, may not lead to reidentification of a unit. However, several indirect identifiers, when used together (cross-classified), may lead to reidentification. This usually happens when a certain cross-classification is rare. For the sake of brevity, since direct identifying variables are to be dropped from the released data, we will refer to indirect identifying variables as identifying variables in the rest of this thesis.

We will next give a brief overview of SDC methods for categorical data.

2.2 SDC Methods for Categorical Data

There exists an extensive list of SDC methodology, some of which apply only to categorical variables. According to Slavković (2009), traditional methods for categorical data include perturbation, swapping, sampling, recoding and suppression. Modern methods include simulating synthetic data (Rubin, 1993), using remote access servers (Keller-McNulty and Unger, 1998), partial information release (Fienberg and Slavković, 2004), and using tools from secure computation
(Slavković et al., 2007) to share statistical products.

Most SDC methods can be expressed in the form of $V' = AVB + C$, where $V$ is the original file, $V'$ is the altered file, $A$ is the record-transforming matrix, $B$ the variable-transforming matrix, and $C$ is noise (Duncan and Pearson, 1991). These methods are called matrix masking methods. Some common matrix methods for categorical data include the Post Randomization Method (PRAM) (Gouweleeuw et al., 1998), data swapping (Dalenius and Reiss, 1982), sampling, global recoding, top and bottom coding, and local suppression. PRAM will be discussed in detail in Chapter 3.

The idea behind data swapping is to exchange values of certain variables between records in a way to preserve lower-order marginals. Fienberg and McIntyre (2004) mention that while the original procedure proposed by Dalenius and Reiss (1982) has hardly been used in practice, later variations have been developed and used by agencies like the U.S. Census Bureau. In the approach used by the Census Bureau, a small proportion of records are swapped with other records from a different geographic region in the file on a set of predetermined variables. This approach guarantees that marginals involving the matching variables remain the same (Ramanayake and Zayatz, 2010). Researchers at the National Institute of Statistical Sciences (NISS) have developed NISS WebSwap (Sanil et al., 2002), a computer program which swaps attributes between records in a file provided by a user. For the swapping algorithm in NISS WebSwap, pairs of records are randomly selected, and their attributes are swapped if swapping constraints are satisfied. The algorithm continues until a predetermined number of swaps have been made. Fienberg and McIntyre (2004) noted that the Census Bureau focuses on preserving summary statistics in released margins, while the NISS approach focuses on the risk utility trade off without focusing on preserving any statistic. Several drawbacks exist in data swapping and their variants: improbable combinations may be created
relationships between variables can be distorted (Moore (1996), and Fienberg and McIntyre (2004)); and the effect of swapping on inference like regression is ambiguous (Fienberg and McIntyre, 2004).

Non-perturbative masking methods for categorical data include sampling, global recoding, top and bottom coding, and local suppression. In sampling, a subsample of the original file is released instead of releasing the entire original file. Shlomo and Skinner (2012) suggested that sampling should be used in conjunction with another SDC method like PRAM, since using both methods generally leads to greater protection than using either method on its own. In global recoding, levels of a categorical variable are combined to form less specific levels. For example, levels of an age category '21-30' and '31-40' can be combined to form a less specific level '21-40'. Top and bottom coding are variants of global recoding, and are used primarily for ordinal variables. The idea is the the top (or bottom) levels of the ordinal variable will be combined to form a new level. For example, the two largest levels of an annual income variable could be 'more than $2 million' and 'between $1 and $2 million' could be combined to form a new level 'more than $1 million'. In local suppression, certain values of variables are removed, to increase the number of records that will have the same set of scores. A drawback in non-perturbative masking methods is that they lead to a loss of detail in the variables.

Releasing synthetic data with characteristics similar to the collected microdata was originally proposed in Rubin (1993). Producing synthetic data might appear to be a promising method, since records in the microdata are manufactured and do not come directly from the original data, which solves the problem of reidentification. However, a published synthetic record could potentially take on the same values for some variables as a real record, which may lead to subjects fearing that privacy may be compromised. It is also difficult to maintain a high level of data utility, since
the synthetic data is generated to preserve only certain properties depending on the model used (Reiter, 2002).

Remote access servers are computers that hold the collected microdata files. Users submit requests for statistical analyses, and when a request is considered safe, the server will provide the results of the statistical analyses, such as estimated model parameters and standard errors. The user will not see the actual or even an altered version of the microdata, thus safeguarding the confidentiality of the file more effectively than releasing altered data. Remote access servers, however, still do not guarantee total confidentiality. For example, residuals are often used to check for model diagnostics, and providing residuals can lead to disclosure of the values of the response variable. Reiter (2004) provides a less technical discussion of remote access servers; for a more technical discussion, see Gomatam et al. (2005). In some cases, users are provided with secure remote access by agencies. Users will obtain an encryption device and sign legal documents promising to keep data confidential. As noted in Abowd and Vilhuber (2011), remote access servers are widely used in European agencies. In some cases, remote access servers also use synthetic data, for example, the Cornell Virtual Research Data Center (RDC) (Abowd and Vilhuber, 2011). Access is given to any user who wants to use the synthetic data in the Synthetic Data Server at the RDC.

Secure computation with distributed databases typically involves a database that is partitioned among several agencies. This strategy will allow the agencies to perform statistical analyses on the database while keeping their own part of the database private. There are two main kinds of partitioning: horizontal partitioning, where different agencies have data on the same attributes on different subjects; and vertical partitioning, where different agencies have data on different attributes on the same subjects. A simple example of a secure computation will be to calculate \( t = \sum_i t_i \),
where $t_i$ is a value belonging to agency $i$. Agency 1 will generate a number $R$ randomly, uniformly from $[0, m)$ where $m$ is a large number, and adds $R$ to $t_1$ to obtain $s_1 = R + t_1$. The result, $s_1$, is then sent to agency 2, which then adds its value $t_2$ to $s_1$ to obtain $s_2$, and $s_2$ is then sent to agency 3 and so on. The last agency then computes $s_k$ and sends it back to agency 1. Agency 1 subtracts $R$ from $s_k$ to obtain $t$, and shares $t$ with all the agencies. Some drawbacks with secure computation is that it requires the agencies to be honest, and methods to assess disclosure risk need to be developed. For more examples and details, see Karr et al. (2005) and Slavković et al. (2007).

For the partial information release strategy, agencies release marginal and conditional tables instead of full contingency tables. This will allow log-linear models to be fitted and association measures to be preserved. Issues regarding disclosure risk and data utility with partial information release are still ongoing. See Slavković (2009) for an overview and Slavković and Lee (2010) for application to two-way tables and logistic regression.

Newer trends include the concept of differential privacy, which was originally proposed in the computer science literature (Dwork, 2006). As defined in Dwork (2006), a randomized function $A$ gives $\epsilon$-differential privacy if for all data sets $D_1$ and $D_2$ differing in at most one element, and all $S \subseteq \text{Range}(A)$,

$$P[A(D_1) \in S] \leq \exp(\epsilon) \times P[A(D_2) \in S].$$

Conceptually, what differential privacy means for an individual in a database is that the output based on a randomized function will not differ significantly if the value of the individual changes.

Next, we will introduce the measures of disclosure risk and data utility, two
important measures to consider when applying SDC methods.

### 2.3 Disclosure Risk and Data Utility

With increased computing power and built-in statistical packages, many researchers nowadays prefer to work with microdata instead of aggregated data. However, releasing microdata instead of aggregated data increases the risk of disclosure. The goal of SDC for microdata is that given an original microdata set $V$, a protected microdata set $V'$ is released in its place so that disclosure risk is low and data utility is high. We typically want data utility to be as high as possible subject to constraints on disclosure risk. However, a method that results in a file that has less risk of disclosure typically leads to the file having less utility. The idea of risk-utility trade-off is usually represented with a risk-utility (RU) map as in Figure 2.1 (Duncan et al., 2001).

**Figure 2.1.** Risk-Utility map.

Disclosure risk measures are usually a function of keys, defined earlier as the
cross-classification of categorical variables. According to Hundepool et al. (2010), disclosure risk measures can be broadly defined into three types.

The first type of disclosure risk measures is based on keys in the sample, where a subject is at risk if its combination of scores on identifying variables in the sample is below a certain threshold. An example of such a measure was proposed by de Wolf and van Gelder (2004). They proposed calculating $P(\xi = k | X = k)$, the conditional probability that, given a score $k$ in the released file, the original score is $k$ as well. In the context of PRAM, this conditional probability can be estimated by 

$$
\frac{\sum_l p_{lk} T_\xi(l)}{\sum_l p_{lk}},
$$

where $p_{lk}$ are known transition probabilities for all $l$, and $T_\xi(l)$ are the frequency counts in the sample for each level of the variable.

The second type of disclosure risk measures is based on keys in the population, where a subject is at risk if its combination of scores on identifying variables in the population is below a certain threshold. An example arises from estimating the conditional probability that a unique match is a correct match (Skinner and Elliot, 2002). Let $X$ be the cross-classification of the variables of interest, with levels $j = 1, ..., J$. Let $f_j$ and $F_j$ denote the frequency counts of level $j$ in the sample and population respectively, and $I$ is the indicator function. Skinner and Elliot (2002)’s measure was

$$
\theta = P(\text{correct match} | \text{unique match}) = \frac{\sum_{j=1}^J I(f_j = 1)}{\sum_{j=1}^J F_j I(f_j = 1)}.
$$

Suppose a subject, called the chosen unit, is chosen randomly from the population, and the value of $X$ for the chosen unit equals the value of $X$ for some subjects in the sample. This becomes a unique match if there is only one such subject in the sample, the matching unit, with the same value of $X$. This becomes a correct match if the chosen unit and matching unit are the same. The number of possible chosen subjects
for which a unique match will exist is \( \sum_j F_j I(f_j = 1) \), and the number of subjects for which the match is correct is \( \sum_j I(f_j = 1) \).

These first two types of measures deal with rare combinations of categorical variables, and are usually expressed as probabilities of reidentification.

The third type of disclosure risk is based on record linkage, which is usually used for continuous variables. One such example is distance-based record linkage, where each record in \( V' \) is linked to the nearest record in \( V \), based on a defined distance function (Hundepool et al., 2010).

These measures usually assume that there is no access to an external data file with which the released data file can be linked.

In general, the greater the difference between results of statistical analyses from the original data \( V \) and the protected data \( V' \), the greater the loss of information. Evaluation of information loss should be based on the desired statistical uses of the protected data. In Hundepool et al. (2010), it is mentioned that it is difficult to perform data protection for specific uses of the data because it is difficult to identify all potential uses of the data, and even if all the uses are identified, releasing several versions of \( V' \) to minimize information loss for each use may result in unexpected disclosure.

According to Domingo-Ferrer and Torra (2001a), measures of information loss for categorical data typically fall into three types: direct comparison of categorical values, comparison of contingency tables, and entropy-based measures. The following examples come from Domingo-Ferrer and Torra (2001a):

1. For direct comparison of categorical values, the equality between pairs of categories in \( V \) and \( V' \) can be computed in the following definition. For a
categorical variable $\xi$, define

$$d_\xi(c, c') = \begin{cases} 
0 & \text{if } c = c' \\
1 & \text{if } c \neq c'
\end{cases}$$

where $c$ is the level in $V$ and $c'$ is the level corresponding to $c$ in $V'$.

2. For comparison of contingency tables, given files $V$ and $V'$ and their corresponding $t$-dimensional contingency tables for $t \leq K$, we define, for a subset $W$ of variables

$$CTBIL(V, V'; W, K) = \sum_{\{\xi_{j_1} \ldots \xi_{j_t}\} \subseteq W; |\{\xi_{j_1} \ldots \xi_{j_t}\}| \leq K} \sum_{i_1, \ldots, i_t} |x^V_{i_1, \ldots, i_t} - x^{V'}_{i_1, \ldots, i_t}|$$

where $x^V_{i_1, \ldots, i_t}$ is the entry of the contingency table of $V$ at position given by $i_1, \ldots, i_t$.

3. An entropy-based measure can be defined as

$$EBIL = -\sum_{r \in V'} \sum_{i=1}^{n} P(\xi = i | X = j) \log P(\xi = i | X = j)$$

for all individuals $r$ in $V'$, where $X$ is the categorical variable in $V'$ that corresponds to $\xi$ in $V$.

Different measures of information loss can lead to different interpretations. For example, compare the original and altered files below.
In this example, $d_{\xi}(c, c') = 2$ and $CTBIL(V, V') = 0$. Only the first measure will indicate that there is a difference in the two files, while the second measure will indicate the corresponding $2 \times 2$ tables are the same, but not indicate the fact that values in the file have changed for some subjects. If the goal is to preserve marginal counts then $CTBIL$ may be a preferred measure.

In many papers in the literature, measures of data utility mainly quantify the difference between the original and the perturbed file. However, statistical inference can be brought to bear in measures of data utility (Rubin 1993, Reiter 2004, Gomatam et al. 2005, Slavković and Lee 2010), for example, how close are the estimates from a linear regression model, when using the original data compared with using the altered data.

Next, we introduce software that have been developed for SDC.

### 2.4 SDC Software

There are two general software packages that implement PRAM and other SDC methods relevant for microdata. Statistics Netherlands, with the European Statistical System net project, has developed a software program, $\mu$-Argus, version 4.2 (Hundepool et al., 2010). The program allows users to produce an altered file after applying various SDC methods such as global recoding and PRAM to
categorical variables; micro aggregation, rank swapping, top and bottom coding, rounding and noise addition can be applied to quantitative variables, as well as local suppression for both categorical and quantitative variables. The specification of PRAM is limited only to band matrices and fully filled matrices with equal off-diagonal entries. The software can be downloaded from

http://neon.vb.cbs.nl/casc/mu.htm

An R package, sdcMicro, has been developed by Templ (2008). SDC methods that can be applied to categorical data are PRAM, local suppression and global recoding; microaggregation, top and bottom coding, rank swapping and noise addition can be applied to quantitative variables. The specification of PRAM is limited only to the invariant PRAM method, which we define in Chapter 3.2.3.

The next chapter gives a detailed introduction to PRAM, with the literature review and highlights current methodological challenges.
Chapter 3

Overview of PRAM

One of the SDC methods for categorical variables is the Post Randomization Method (PRAM), which was first introduced by Gouweleeuw et al. (1998). The basic idea underlying PRAM is to misclassify the values of categorical variables via a known probability mechanism, after the data have been collected. This mechanism is expressed in the form of a transition matrix, and in the context of PRAM, it is called a PRAM matrix.

PRAM is a perturbative masking method. Since the PRAM matrix is known, characteristics of the original file can potentially be estimated from the perturbed version which is the version that is released for use by others.

PRAM has been used in conjunction with other SDC methods in the 2001 UK census (Longhurst et al., 2007). PRAM has also been implemented in SDC software such as $\mu$-Argus (Hundepool et al., 2010) and in the R package sdcMicro (Templ, 2008). However, in both $\mu$-Argus and sdcMicro, the manner in which PRAM can be applied is limited. PRAM has yet to be applied extensively primarily because finding a way to apply PRAM optimally is still not known (Mares and Torra, 2010).

PRAM is related to the randomized response (RR) technique developed and used
in survey sampling. The next subsection gives a brief introduction to RR.

3.1 Randomized Response

Randomized Response (RR) is an interview technique introduced by Warner (1965), which can be used when sensitive questions have to be asked. For more details, also see Chaudhuri and Mukerjee (1988).

The following example demonstrates the application of RR to a survey question. Let a sensitive question be

\[ Q = \text{“Have you ever cheated on an exam?”} \]

The interviewer asks the respondent to roll a die and keep the outcome hidden from the interviewer. If the outcome is 1, 2, 3 or 4, the respondent answers question \( Q \). Otherwise, he is asked to answer \( Q^c \), where

\[ Q^c = \text{“Have you never cheated on an exam?”} \]

In this example, the respondent answers \( Q \) with probability \( p = \frac{2}{3} \) and \( Q^c \) with probability \( \frac{1}{3} \). The goal of RR is that the respondent feels comfortable giving a truthful answer to a sensitive question, since the interviewer will not know what question the respondent is answering. However, if the respondent does not feel that he is sufficiently protected, he may still be reluctant to give an honest answer.

Let \( \pi \) be the unknown probability of observing a “yes” to \( Q \), and \( \lambda \) is the population proportion of “yes” responses to both \( Q \) and \( Q^c \). So by the Law of Total Probability

\[ \lambda = p\pi + (1 - p)(1 - \pi). \]
Solving for \( \pi \), we have

\[
\pi = \frac{\lambda - (1 - p)}{2p - 1}.
\]

Note that \( p \neq \frac{1}{2} \). Using \( \hat{\lambda} \), the observed proportion of “yes” in the sample to estimate \( \lambda \), the method-of-moments estimator of \( \pi \) is

\[
\hat{\pi}_{\text{MOM}} = \frac{\hat{\lambda} - (1 - p)}{2p - 1}.
\]  

(3.1.1)

In this example, let \( \lambda = (\lambda, 1 - \lambda)' \), \( \pi = (\pi, 1 - \pi)' \), and

\[
P = \begin{pmatrix}
p & 1 - p \\
1 - p & p
\end{pmatrix}.
\]

Hence RR designs can be expressed in matrix form

\[
\lambda = P'\pi
\]  

(3.1.2)

and the method-of-moments estimator can be expressed in matrix form

\[
\hat{\pi}_{\text{MOM}} = (P^{-1})'\hat{\lambda}.
\]  

(3.1.3)

From (3.1.1), it can be shown that under certain conditions, \( \hat{\pi} \in [0, 1] \) if and only if \( 1 - p < \hat{\lambda} < p \) and \( p > \frac{1}{2} \), or \( p < \hat{\lambda} < 1 - p \) and \( p < \frac{1}{2} \).

Chaudhuri and Mukerjee (1988) showed that the MLE, \( \hat{\pi}_{\text{MLE}} \), of \( \pi \) is the same as the method-of-moments estimator of \( \pi \) under the conditions specified above. In other
circumstances,

\[ \hat{\pi}_{MLE} = \begin{cases} 
1, & \text{if } \hat{\lambda} \geq p > \frac{1}{2} \text{ or } \hat{\lambda} \leq p < \frac{1}{2} \\
0, & \text{if } \hat{\lambda} \leq (1 - p) < \frac{1}{2} \text{ or } \hat{\lambda} \geq (1 - p) > \frac{1}{2} 
\end{cases} \]

Next, we introduce PRAM and discuss its relation to RR.

### 3.2 The Post Randomization Method (PRAM)

Let \( \xi \) denote a categorical variable in the original file to which PRAM will be applied, and let \( X \) denote the same categorical variable in the perturbed file. The probability that an original score \( \xi = k \) is changed into the score \( X = l \) is denoted by the transition probability \( p_{kl} = P(X = l \mid \xi = k) \) for all \( k, l = 1, \ldots, K \). So PRAM, like RR, is described by a \( K \times K \) transition matrix with entries \( p_{kl} \). Let \( T_\xi \) be the \( K \times 1 \) vector of frequencies of the \( K \) categories of \( \xi \) observed in the original file, and \( T_X \) be the \( K \times 1 \) vector of frequencies in the perturbed file. Let \( P \) be a \( K \times K \) transition matrix of conditional misclassification probabilities \( p_{ij} \), with \( \sum_{j=1}^{K} p_{ij} = 1 \). Based on the Law of Total Probability, frequency estimation with PRAM is often presented in the form

\[ E(T_X \mid \xi) = P^T T_\xi. \]  

From (3.1.2) and (3.2.1), it can be seen that RR and PRAM are based on similar mathematical formulas.

Let \( \xi(r) \) and \( X(r) \) denote the value of the \( r \)-th record in the microdata file. Given \( \xi(r) = k \), the score on \( X(r) \) is drawn from the probability distribution \( p_{k1}, p_{k2}, \ldots, p_{kK} \).

Suppose we want to apply PRAM to the categorical variables \( \xi_1 \) and \( \xi_2 \), with \( K_1 \) and \( K_2 \) levels respectively. Let \( X_s \) denote the value of \( \xi_s, s = 1, 2 \), in the perturbed file.
Applying PRAM means that given $\xi_1^{(r)} = k_1$ and $\xi_2^{(r)} = k_2$, the scores on $X_1^{(r)}$ and $X_2^{(r)}$ are simultaneously drawn from the probability distribution $P(X_1 = l_1, X_2 = l_2 \mid \xi_1 = k_1, \xi_2 = k_2)$. Alternatively, $\xi_1$ and $\xi_2$ can be viewed as a single compounded variable $\xi$ with $K_1 K_2$ levels. If PRAM is applied sequentially the transition probability matrix can be written as $P = P_2 \otimes P_1$ where $\otimes$ denotes the Kronecker product.

While PRAM and RR have similar mathematical formulas, they have different purpose in practice (van den Hout and van der Heijden, 2002): (1) PRAM is applied to identifiers that are categorical whereas RR is applied to response variables in surveys; (2) for RR to be implemented successfully, the respondent needs to understand the protection RR provides in order to give a truthful answer for subsequent estimations of proportions not to be biased while frequency estimates after PRAM is applied are unbiased; (3) in RR, the transition matrix has to be determined before the data are collected whereas in PRAM the transition matrix is determined after the data have been collected.

### 3.2.1 The Method-of-Moments Estimator for Population Proportions

Let $\pi = (\pi_1, \pi_2, ..., \pi_K)'$ be the $K \times 1$ vector of population proportions of the $K$ categories of $\xi$ in the original file, and $\lambda = (\lambda_1, \lambda_2, ..., \lambda_K)'$ be the $K \times 1$ vector of proportions in the perturbed file. In Gouweleeuw et al. (1998), based on the Law of Total Probability, it is shown that

$$E(\hat{\lambda} \mid \xi) = P' \hat{\pi}.$$
Assuming \( \mathbf{P} \) is non-singular, the method-of-moments estimator of \( \mathbf{\pi} \) is given in (3.1.3)

\[
\hat{\mathbf{\pi}}_{\text{MOM}} = (\mathbf{P}^{-1})' \hat{\mathbf{\lambda}}.
\]

Since \( \mathbf{P} \) is non-singular, the rows in \( \mathbf{P} \) must be linearly independent.

### 3.2.2 The Maximum Likelihood Estimator for Population Proportions

Since the rows of \( \mathbf{P}^{-1} \) may not sum to one, the estimated proportions of \( \mathbf{\xi} \), \( \hat{\mathbf{\pi}}_{\text{MOM}} \), may not always belong to the parameter space \([0, 1]\). As defined earlier, let \( \mathbf{\pi} \) be the proportions of \( \mathbf{\xi} \), let \( \mathbf{\lambda} \) be the proportions of \( \mathbf{X} \), and let \( \hat{\mathbf{\lambda}} \) be the proportions of \( \mathbf{X} \) in the sample. Also, let \( \mathbf{T}_\mathbf{\xi} \) and \( \mathbf{T}_\mathbf{X} \) denote the vectors \((\mathbf{T}_\mathbf{\xi}(1), ..., \mathbf{T}_\mathbf{\xi}(\mathbf{K}))\) and \((\mathbf{T}_\mathbf{X}(1), ..., \mathbf{T}_\mathbf{X}(\mathbf{K}))\), the vectors of frequency counts in the original file and perturbed file respectively. Thus, \( \mathbf{T}_\mathbf{\xi} \) has a multinomial distribution with parameters \( n = \sum_{i=1}^{\mathbf{K}} \mathbf{T}_\mathbf{\xi}(i) = \sum_{i=1}^{\mathbf{K}} \mathbf{T}_\mathbf{X}(i) \) and \( \mathbf{\pi} \), and \( \mathbf{T}_\mathbf{X} \) is multinomial with parameters \( n \) and \( \mathbf{\lambda} \). Based on Law of Total Probability, \( \lambda_i = \sum_{k=1}^{\mathbf{K}} p_{ki} \pi_k \).

We proceed to outline the proof by van den Hout and van der Heijden (2002) that when the MLE is in the interior of the parameter space then the MLE is identical with the method-of-moments estimator.

Since \( \mathbf{\pi} \) is a parameter from the multinomial distribution, the likelihood is given by

\[
l^*(\mathbf{\pi}) = \frac{n!}{\mathbf{T}_\mathbf{X}(1)! \cdots \mathbf{T}_\mathbf{X}(\mathbf{K})!} \lambda_1^{\mathbf{T}_\mathbf{\xi}(1)} \cdots \lambda_{\mathbf{K}}^{\mathbf{T}_\mathbf{\xi}(\mathbf{K})} \tag{3.2.2}
\]

and the log-likelihood is given by

\[
\log l^*(\mathbf{\pi}) = \sum_{i=1}^{\mathbf{K}} n_i^* \log \lambda_i + C,
\]
where $C$ is a constant. Likelihoods from exponential families take the form

$$l^*(\pi) = a(\pi)b(T_X) \exp\{\theta'(\pi)t(T_X)\}. \quad (3.2.3)$$

Let $a(\pi) = 1$, $b(T_X) = n! / T_X(1)!...T_X(K)!$, $t(T_X) = (T_X(1), ..., T_X(K))'$ and $\theta'(\pi) = (\log \lambda_1, ..., \log \lambda_K)$. Equation (3.2.3) becomes

$$l^*(\pi) = \frac{n!}{T_X(1)!...T_X(K)!} \exp\{\theta_1 T_X(1) + \cdots + \theta_K T_X(K)\}$$

$$= \frac{n!}{T_X(1)!...T_X(K)!} \lambda_1^{T_X(1)} \cdots \lambda_K^{T_X(K)}$$

which is equivalent to (3.2.2). Hence, (3.2.2) belongs to the exponential family. Note that $t(T_X)$ is a sufficient statistic and $\theta'(\pi)$ is the canonical parameter.

Since $n = \sum_{i=1}^K T_X(i)$ and $\sum_{i=1}^K \lambda_i = 1$, an alternative definition of (3.2.3) for (3.2.2) exists. Rewriting (3.2.2), we obtain

$$l^*(\pi) = \frac{n!}{T_X(1)!...T_X(K)!} \lambda_1^{T_X(1)} \cdots \lambda_K^{T_X(K)}$$

$$= \frac{n!}{T_X(1)!...T_X(K)!} \left( \sum_{i=1}^K \lambda_i \right)^n \binom{T_X(1) \cdots T_X(K)}{T_X(K-1)} \left( \frac{\lambda_1}{\lambda_K} \right)^{T_X(1)} \cdots \left( \frac{\lambda_{K-1}}{\lambda_K} \right)^{T_X(K-1)}$$

$$= (1 + \frac{\lambda_1}{\lambda_K} + \cdots + \frac{\lambda_{K-1}}{\lambda_K})^{-n} \frac{n!}{T_X(1)!...T_X(K)!} \exp \left[ \sum_{i=1}^{K-1} T_X(i) \log \frac{\lambda_i}{\lambda_K} \right]$$

$$\times \exp \left[ \theta_1 T_X(1) + \cdots + \theta_{K-1} T_X(K-1) \right] \quad (3.2.4)$$

Comparing (3.2.4) with (3.2.3), with

$$t(T_X) = (T_X(1), ..., T_X(K-1))'$$
\[ \theta'(\pi) = (\log \frac{\lambda_1}{\lambda_K}, \ldots, \log \frac{\lambda_{K-1}}{\lambda_K}), \]
\[ a(\pi) = (1 + \exp \theta_1 + \cdots + \exp \theta_{K-1})^{-n}, \text{ and} \]
\[ b(T_X) = \frac{n!}{T_X(1)! \cdots T_X(K)!}, \]

we have an exponential representation of (3.2.2) in \( K - 1 \) dimensions.

Let \( \Omega \) be the domain of variation for \( \pi \) and \( \Theta = \theta(\Omega) \) the canonical parameter domain. To show that (3.2.4) is from a regular exponential family, we need to show the following (Barndorff-Nielsen, 1986, p. 116):

i) \( \Theta \) is an open subset of \( \mathbb{R}^{K-1} \), and

ii) \( \Theta = \{ \theta : \int_{T_X} \frac{n!}{T_X(1)! \cdots T_X(K)!} \exp [\theta' t(T_X)] dT_X < \infty \} \).

Since \( \Theta = \mathbb{R}^{K-1} \), (i) is satisfied. For (ii), since \( 0 < \lambda_k < 1 \) for \( k = 1, \cdots, K \) and \( n = \sum_{i=1}^{K-1} T_X(i) \),

\[
\int_{T_X} \frac{n!}{T_X(1)! \cdots T_X(K)!} \exp [\theta' t(T_X)] dT_X
\]
\[
= n! \int_{T_X} \frac{1}{T_X(1)! \cdots T_X(K)!} \exp \left[ T_X(1) \log \frac{\lambda_1}{\lambda_K} + \cdots + T_X(K-1) \log \frac{\lambda_{K-1}}{\lambda_K} \right] dT_X
\]
\[
\leq n! \int_{T_X} \left( \frac{\lambda_1}{\lambda_K} \right)^{T_X(1)} \cdots \left( \frac{\lambda_{K-1}}{\lambda_K} \right)^{T_X(K-1)} dT_X
\]
\[
= n! \int_{T_X} \lambda_1^{T_X(1)} \cdots \lambda_K^{T_X(K)} \frac{1}{\lambda_K^n} dT_X
\]
\[
\leq n! \int_{T_X} \left( \frac{1}{\lambda_K} \right)^n dT_X
\]
\[
< \infty.
\]

A property of the regular exponential family states that the MLE and the method-of-moments estimator are equal when both are in the interior of the parameter space (Barndorff-Nielsen, 1986). To find the MLE of (3.2.4), we minimize the log-likelihood
log \( l^n(\pi) \). Using the representation in (3.2.3), we want to solve

\[
\frac{\partial}{\partial \theta} \log [a(\pi)b(T_X) \exp\{\theta'(\pi)t(T_X)\}] = 0,
\]

which upon simplification gives

\[
\frac{\partial}{\partial \theta} \theta'(\pi)t(T_X) = \frac{\partial}{\partial \theta} [-\log a(\pi)].
\]

We have

\[
\frac{\partial}{\partial \theta} \theta'(\pi)t(T_X) = (T_X(1), \ldots, T_X(K - 1))'
\]

and since \( \lambda_i = \sum_{k=1}^K p_{ki} \pi_k \),

\[
\frac{\partial}{\partial \theta} [-\log a(\pi)] = n \begin{pmatrix}
\frac{\exp \theta_1}{1+\exp \theta_1+\cdots+\exp \theta_{K-1}} \\
\vdots \\
\frac{\exp \theta_{K-1}}{1+\exp \theta_1+\cdots+\exp \theta_{K-1}}
\end{pmatrix} = n \begin{pmatrix}
\lambda_1 \\
\vdots \\
\lambda_{K-1}
\end{pmatrix}
\]

\[
= n \begin{pmatrix}
p_{1,1} & \cdots & p_{1,K} \\
\vdots & \ddots & \vdots \\
p_{K-1,1} & \cdots & p_{K-1,K}
\end{pmatrix}' \begin{pmatrix}
\pi_1 \\
\vdots \\
\pi_K
\end{pmatrix} \quad (3.2.5)
\]

Comparing with (3.2.1), (3.2.5) is equivalent to (3.2.1) on which the method-of-moments estimator is based. Therefore, the MLE and method-of-moments estimator for \( \pi \) are equal in the parameter space.
3.2.3 Invariant PRAM

Thus far, the only restriction imposed on the matrix $P$ is non-singularity. An additional condition could be placed on $P$ such that the distribution of $\xi$ over different categories is invariant with respect to $P'$, i.e.,

$$P'T_\xi = T_\xi.$$

A non-trivial solution, as suggested in Gouweleeuw et al. (1998) is as follows. Assume $T_\xi(k)$ is smallest of $T_\xi(i)$, without loss of generality, hence $T_\xi(k) \geq T_\xi(K) > 0$ for $k = 1, ..., K$, and consider $0 < \theta < 1$, such that

$$p_{kl} = \begin{cases} 1 - (\theta T_\xi(K)/T_\xi(k)) & \text{if } l = k, \\ \theta T_\xi(K)/(K - 1) & \text{if } l \neq k. \end{cases}$$

Thus $E(T_X | \xi) = P'T_\xi = T_\xi$, and $T_\xi$ can be estimated by $\hat{T}_\xi = T_X$.

The advantage of invariant PRAM is that we can obtain frequency counts in the original file from the perturbed file without needing the PRAM matrix, and there will be no negative frequencies. However, there are some drawbacks, such as a limited choice of the PRAM matrix. If we want to preserve the distribution of all possible cross-tabulations and consider all variables $\xi_1, ..., \xi_m$ as a single compounded variable $\xi$ with $K = K_1 \times \cdots \times K_m$ categories, then $K$ may be very large and sparse since many categories may contain no observations and a non-trivial choice of $P$ may be difficult.
3.3 Methodological Challenges with PRAM

There are a number of issues which arise in applying PRAM. In this section, we review the current PRAM literature and focus on three commonly discussed issues.

1. Adjustments to estimates of frequency counts, and models necessary for statistical analyses (e.g., de Wolf et al. (1998), Gouweleeuw et al. (1998), van den Hout and van der Heijden (2002), van den Hout and Elamir (2006), van den Hout and Kooiman (2006));

2. Dealing with dependent variables and inconsistencies in the perturbed dataset after the application of PRAM (e.g., de Wolf et al. (1998), Gouweleeuw et al. (1998), Shlomo and de Waal (2008)); and

3. Finding an optimal PRAM matrix that produces a safe file and maximizes data utility (e.g., Cator et al. (2005), de Wolf et al. (1998), de Wolf and van Gelder (2004), Mares and Torra (2010), Shlomo and Skinner (2010)).

3.3.1 Issues with Estimation

In this section, we discuss some issues with estimation of frequency counts and models and dependencies between variables in PRAM. From a computational standpoint, we may only want scores on each record to be able to change into a few predetermined number of scores. Gouweleeuw et al. (1998) suggest partitioning a variable $\xi$ with $K$ levels into groups $C_1, ..., C_G$ so that each level can only be replaced by a level within the same group. This could be useful especially when the groups can be constructed in a way that the levels within each group are similar. The same strategy was also proposed in Shlomo and de Waal (2008), although their goal was to ensure consistency in the perturbed file.
Invariant PRAM was developed to avoid the issue of getting negative estimated frequency counts. However, an issue regarding the use of invariant PRAM is that there appears to be less freedom of choice for the PRAM matrix (Gouweleeuw et al., 1998). A “two-stage” PRAM to obtain an invariant PRAM matrix was first proposed in de Wolf et al. (1998) to allow for more freedom.

The following algorithm to derive such a matrix is described in Willenborg and de Waal (2001) and Shlomo and Skinner (2010). The PRAM matrix $R = PQ$ is applied, where $P$ is the usual PRAM matrix, and $Q$ is the reclassification matrix, with entries $q_{ij} = P(\xi = j|X = i)$. The matrix $Q$ can be estimated by

$$
\hat{Q} = \frac{p_{ji}P(X = j)}{\sum_l p_{li}P(X = l)}.
$$

and $\hat{Q}$ can be obtained by transposing $P$, multiplying each column $j$ by $P(X = j)$ and then normalizing each row so that each row sums to one. Let $R^* = \alpha R + (1 - \alpha)I$, where $I$ is the identity matrix. $R^*$ is also invariant, and the level perturbation can be controlled by $\alpha$. This algorithm, for example, is used in the PRAM function in the sdcMicro package for R. The main purpose of using invariant is preserving marginals.

Another issue with PRAM is obtaining unbiased estimators of $\pi$, the proportions for $T_\xi$. It is possible that the method-of-moment estimator, $\hat{\pi}_{MOM}$ (3.1.3), of the proportions is not a maximum likelihood estimator when it falls outside the parameter space $[0, 1]$. van den Hout and van der Heijden (2002) discuss adjustments that need to be made for certain statistical analyses in order to obtain MLEs. For example, they proposed an EM algorithm to find the MLE of the true contingency table after PRAM has been applied to the file. The EM algorithm can be used as an iterative technique to find MLEs when the data are missing or incomplete. We can consider finding MLEs after PRAM has been applied to the data as an incomplete data problem.
Since the perturbed file is released and the original file is not, the observed values in the released file can be linked with their true latent values. The EM algorithm works well for data subject to PRAM since there is a unique maximum in the interior of the parameter space. The incomplete-data likelihood is also from a regular exponential family and is therefore strictly concave, so finding the maximum should not be an issue when the starting point is chosen in the interior of the parameter space.

Another method to avoid getting negative estimated frequency counts was proposed by van den Hout and Elamir (2006), who proposed using calibration probabilities (or reclassification probabilities in other literature) $P(\xi|X)$ instead of the misclassification probabilities $P(X|\xi)$ to avoid negative estimates of frequency counts. Using calibration probabilities will yield the MLE, and will lead to estimates that have smaller variance than estimates using misclassification probabilities. However, a disadvantage is that it only works well for univariate analysis, whereas in practice, multivariate analysis is far more important. van den Hout and Elamir (2006) also suggest using misclassification proportions instead of misclassification probabilities. Misclassification proportions give information about the actual changes due to PRAM, whereas misclassification probabilities are about the expected changes due to PRAM. Their experiments show that calibration probabilities do not work as well as misclassification probabilities when it comes to bivariate frequency estimation, and working with misclassification proportions, compared to misclassification probabilities, tended to work better in bivariate frequency estimations in terms of root mean square error. Working with misclassification proportions, instead of misclassification probabilities, appear to be quite promising.

van den Hout and Kooiman (2006) propose an EM algorithm to obtain the MLE of normal linear regression models with categorical covariates subject to randomized
response. This idea can be extended to linear regression where categorical covariates are subject to PRAM. van den Hout and Kooiman (2006) show in their experiments that MLEs are biased when the misclassification is not taken into account, and their proposed EM algorithm is able to obtain unbiased estimates when taking the misclassification into account. Their EM algorithm is similar to the ideas developed in Ibrahim (1990), who developed EM algorithms to obtain MLEs of generalized linear models (GLMs) with missing covariates. The proposed EM algorithm by van den Hout and Kooiman (2006) actually converges quickly since the maximization in the M-step can be expressed analytically. This EM algorithm is not sensitive with respect to starting values, as long as the starting values lie in the parameter space.

3.3.2 Issues with Dependencies between Variables

Gouweleeuw et al. (1998) also brought up issues with dependencies between variables in the original file. The variables in the original file cannot be perturbed in a way that results in inconsistencies between variables in the perturbed file. For example, a record could have the values

\[
\text{age} = 26-35 \text{ and marital status} = \text{Married},
\]

and with the application of PRAM, the record becomes

\[
\text{age} = \text{under 12} \text{ and marital status} = \text{Divorced}.
\]

Gross et al. (2004) proposed a stratification method as a way to preserve dependencies between multivariate distributions. To preserve the relationship between two variables, PRAM is applied to the first variable within each stratum
defined by values of the second variable. PRAM is then applied to the second variable within each stratum of the first variable. Gross et al. (2004) showed that their stratification method worked well in preserving multivariate distributions on a sample of anonymized records from the 2001 UK census.

Shlomo and de Waal (2008) proposed dividing the variable which will be subject to PRAM into subgroups $C_1, ..., C_G$ and developing invariant PRAM matrices to each subgroup. The PRAM matrices for each subgroup should be placed on the main diagonals of the final matrix, where the off-diagonal entries will be 0; this ensures that the variable is perturbed only within that subgroup. Shlomo and de Waal (2008) also suggest that variables that are highly correlated with the perturbed variable should be compounded into a single variable, and PRAM should be carried out on the compound variable. These steps should reduce some inconsistencies but not all. Remaining inconsistencies should be imputed via a hot-deck imputation, which is implemented by using a neighboring donor matching on the control variables. Shlomo and de Waal (2008) applied the proposed strategy to the 1995 Israeli census, and showed that the number of inconsistencies were largely reduced.

### 3.3.3 Issues with Obtaining Optimal PRAM Matrices

A general SDC paradigm suggests that a technique should be applied such that the file becomes safe while minimizing information loss. Gouweleeuw et al. (1998) suggest that PRAM should be applied such that when a rare combination occurs in the perturbed file, then it has a low probability of being true. One way of measuring this is by means $ER(k)$, of the expected ratio of the score $k$, defined as

$$ER(k) = \frac{p_{kk}T_\xi(k)}{\sum_{l \neq k} p_{lk}T_\xi(l)},$$
where the numerator represents the expected number of records that remain at $k$, and the denominator represents the expected “inflow” of records that were not originally $k$ and now take the value $k$. If this ratio is below a chosen threshold then the perturbed file is considered safe.

Empirical evaluations of disclosure risk and information loss for various SDC methods were carried out in Domingo-Ferrer and Torra (2001b). PRAM was compared to top coding, bottom coding, and global recoding for categorical variables. They found negative correlation between disclosure risk and information loss, and PRAM appeared to perform more poorly than other SDC methods. However, this could be due to the fact that non-optimal PRAM matrices were chosen, and only invariant PRAM matrices of the form specified in Gouweleeuw et al. (1998) were evaluated.

The effects of different PRAM matrices on disclosure control as well as information loss were investigated in de Wolf and van Gelder (2004). Another measure of disclosure risk was used, which is the conditional probability the original score was $k$ given a score $k$ in the released file, $P(\xi = k \mid X = k)$. This measure can be estimated by

$$\hat{R}_{PRAM}(k) = \frac{p_{kk} T_{k}(k)}{\sum_{l} p_{lk} T_{k}(l)}. \quad (3.3.1)$$

Using traditional threshold rules, a record is considered safe whenever a certain combination of scores on identifying variables occur at least $d$ times. A safe record can be linked with at least $d$ records in the population. If done randomly, the probability that the record is linked correctly in the population is less than or equal to $d^{-1}$, or the risk of disclosure is at most $d^{-1}$. de Wolf and van Gelder (2004) also
suggested considering a record safe whenever

\[
\hat{R}_{PRAM}(k) \leq \frac{T_\zeta(k)}{d}.
\]  

(3.3.2)

In the empirical evaluations performed by de Wolf and van Gelder (2004), information loss measures proposed by Domingo-Ferrer and Torra (2001a) were used. The PRAM matrices investigated were of three types:

- Band matrices \(nB(p;b)\). Here \(p\) denotes the value of the diagonal entries, \(b\) the bandwidth and \(n\) the size of the matrix. Off-diagonal elements in the band have equal probability mass. For example, \(4B(0.5;2)\) is

\[
\begin{pmatrix}
0.5 & 0.5 & 0 & 0 \\
0.25 & 0.5 & 0.25 & 0 \\
0 & 0.25 & 0.5 & 0.25 \\
0 & 0 & 0.5 & 0.5 \\
\end{pmatrix}
\]

- Fully filled matrices with equal off diagonal entries, \(nE(p)\). Here \(n\) denotes the size of the matrix and \(p\) the value of the diagonal entries. For example, \(3E(0.6)\) is

\[
\begin{pmatrix}
0.6 & 0.2 & 0.2 \\
0.2 & 0.6 & 0.2 \\
0.2 & 0.2 & 0.6 \\
\end{pmatrix}
\]

- Fully filled matrices with off-diagonal entries depending on corresponding frequencies in the original file \(nF(p)\). Here \(n\) denotes the size of the matrix and \(p\) the value of the diagonal entries. The off-diagonal entries are computed
by
\[ p_{kl} = \frac{(1 - p_{kk})(\sum_{i=1}^{K} T_\xi(i) - T_\xi(k) - T_\xi(l))}{(n - 2)(\sum_{i=1}^{K} T_\xi(i) - T_\xi(k))}. \]

For example, with \( T_\xi = (50, 80, 120)' \), \( 3F(0.5) \) is
\[
\begin{pmatrix}
0.5 & 0.3 & 0.2 \\
0.35 & 0.5 & 0.15 \\
0.31 & 0.19 & 0.5
\end{pmatrix}.
\]

The first two kinds of matrices are readily available in \( \mu \)-Argus (Hundepool et al., 2010). It appeared that in most cases, increasing the number of non-zero entries resulted in a decrease of unsafe combinations.

Cator et al. (2005) formalize a number of conditions that enables a file to be protected against disclosure. Firstly, there should be no rare combinations in the original file. Secondly, provided the size of the file is large enough, there always exist PRAM matrices that will result in safe data. Finally, if certain combinations appear frequently enough, these records will always be safe, regardless of choice of PRAM matrices. Cator et al. (2005) suggest using a quadratic loss function when evaluating information loss since it leads to an easier optimization problem for finding the optimal PRAM matrix that minimizes information loss and produces a safe file. In their experiments, Cator et al. (2005) found that for large files, the optimal PRAM matrix will be close to the identity matrix. A strategy they suggested is to give combinations that appear frequently small probabilities to be changed into combinations that appear rarely, and leave other combinations unchanged. Then the occurrence of rare combinations will be due to PRAM and will not lead to disclosure of an individual.

Skinner and Elliot (2002) proposed another measure of disclosure risk, \( \theta \), which
estimates the proportion of correct matches among population units which match a sample unique. van den Hout and Elamir (2006) compared $\theta$ with $\hat{R}_{PRAM}(k)$, (3.3.1), in a simulation experiment. It turned out that as sample size increases, $\theta$ increases whereas (3.3.1) decreases, seemingly conflicting results at first glance. For $\theta$, with a larger sample size, a sample unique is more likely to also be a population unique which increases the probability a sample unique is correctly matched to the unit in the population. For (3.3.1), a larger sample size means that there are more records that can be perturbed to take on the value of a rare combination. This shows how different concepts of disclosure can result in different methods of disclosure control. Thus, when using sampling with PRAM, there is a dilemma of choosing a sample size for the released data.

Mares and Torra (2010) suggest using evolutionary algorithms to find an optimal PRAM matrix, one that when applied to variables maximizes data utility and minimizes disclosure risk. Such algorithms are stochastic optimization and search methods that mimic the metaphor of natural biological evolution. Two basic operators were used: crossover and mutation. Crossover is performed by swapping two ranges of values, and mutation replaces a simple value by its negation. In this experiment, both crossover and mutation have rates of 0.5. They carried out experiments on the two kinds of PRAM matrices, fully-filled matrices with equal off-diagonal entries, and fully-filled matrices, with off-diagonal elements depending on corresponding frequencies in the original file, as defined above in de Wolf and van Gelder (2004). It is interesting to point out that the final PRAM matrices do not have dominant diagonal entries i.e., do not have diagonal entries close to 1. This appears surprising because a common paradigm is to have PRAM matrices that are diagonal dominant. A drawback in the proposed algorithm is that it is computationally expensive.
Shlomo and Skinner (2012) linked PRAM with a concept of differential privacy, which was originally proposed in the computer science literature (Dwork, 2006). As defined in Dwork (2006), a randomized function $A$ gives $\epsilon$-differential privacy if for all data sets $D_1$ and $D_2$ differing in at most one element, and all $S \subseteq \text{Range}(A)$,

$$P[A(D_1) \in S] \leq \exp(\epsilon) \times P[A(D_2) \in S].$$

Shlomo and Skinner (2012) prove that PRAM preserves $\epsilon$-differential privacy provided there are no zero-elements in the perturbation matrix, and using both sampling and perturbation generally leads to greater protection than using either method singly. Shlomo and Skinner (2012) argue the practicality of using $\epsilon$-differential privacy for statistical agencies, since breaching $\epsilon$-differential privacy will require the intruder having knowledge of counts in a particular cell for the whole population excluding the target individual, which is unrealistic. They also propose the use of a less strict condition, $(\epsilon, \delta)$-probabilistic differential privacy (Chaudhuri and Mishra, 2006), which allows $\epsilon$-differential privacy to fail with a small probability of no more than $\delta$.

### 3.4 Proposed Work & Contributions

In the next three chapters, we propose solutions to three methodological issues with PRAM outlined in the previous section.

In Chapter 4, we focus on the first issue of obtaining valid statistical analysis with data subject to PRAM. The application of PRAM is known to produce biased estimators of parameters in generalized linear models (GLMs); therefore the effect of PRAM has to be taken into account in order to maintain data utility. We develop
expectation-minimization (EM) algorithms to obtain unbiased maximum likelihood estimators (MLEs) of GLMs when PRAM has been applied to: 1) covariates, 2) response variable, and 3) both covariates and response variables, under the assumption of independent covariates. In particular, Case 3 is a more difficult problem and has received little attention in the literature.

In Chapter 5, we address the second raised issue of dealing with dependent variables. Building on results from Chapter 4, we relax the condition of independent covariates. This is done by modeling the distribution of the covariates subject to PRAM as a product of univariate conditional distributions. This approach advances the methodology by making it more applicable in practice by allowing dependence between the covariates, and results in more accurate estimators of the regression parameters.

In Chapter 6, we address the third issue of obtaining optimal PRAM matrices. We establish monotonicity properties of the entropy-based information loss measure for data utility and for certain classes of PRAM matrices. In this context, an optimal PRAM matrix, one which produces a file with an acceptable level of disclosure risk and maximizes data utility, can be derived explicitly. We derive an algorithm that obtains a PRAM matrix that is superior to existing numerical methods and routines, since safe files with higher data utility will be produced.

These three chapters are the basis for three papers to be submitted, and hence some repetitions will be present.
Chapter 4

Generalized Linear Model

Estimation with PRAM, with Independent Covariates

4.1 Introduction

Statistical agencies face the challenge of providing high-quality data products to aid valid statistical research and policy-making, while also guarding against the risk of disclosing confidential information on individual respondents. Statistical Disclosure Control (SDC) methods aim at finding the best compromise between data utility and disclosure risk. For more details on SDC methodology and its importance to official statistics, see Willenborg and de Waal (1996); Fienberg and Slavković (2010); Ramanayake and Zayatz (2010).

Microdata are sets of records containing detailed information on individual respondents and there is a substantial demand for such high quality data products nowadays. Many SDC methods have been developed for microdata, and when
applied they typically lead to publishing of one or more altered datasets by introducing bias and variance to the original data. The Post Randomization Method (PRAM) is one such method originally proposed by Gouweleeuw et al. (1998). The main idea behind PRAM is to publish redacted data after the values of categorical variables in the original dataset have been misclassified by a known probability mechanism. This probability mechanism is described by a transition matrix, i.e., PRAM matrix.

While PRAM provides certain advantages compared to other SDC methodologies, it has seen a limited use in practice due to a number of unresolved issues related to data utility. More specifically, parameter estimates and summary statistics based on PRAMed data are typically biased and need to be adjusted to take the effects of PRAM into account in order to obtain valid inference. To address this issue, for example, Gouweleeuw et al. (1998) proposed an unbiased moment estimator for frequency counts, van den Hout and van der Heijden (2002) proposed estimates of odds ratios for data subject to PRAM, and most recently Woo and Slavković (2012) proposed an EM algorithm to obtain unbiased estimates of regression coefficients in the logistic regression model. In the related literature on randomized response, the issue of estimation was addressed by van den Hout and Kooiman (2006) who proposed an EM algorithm to estimate the linear regression model with covariates subject to randomized response, and van den Hout et al. (2007) who proposed using a Newton-Raphson method to estimate the logistic regression model with response variable subject to randomized response.

In this chapter, we develop and implement EM-type algorithms to obtain asymptotically unbiased estimators, that is the maximum likelihood estimators of parameters in generalized linear models GLMs, when chosen variables are subject to

---

1This chapter is a basis for two papers: Woo and Slavković (2012, 2013).
PRAM. This work extends the methodology and results presented in Woo and Slavković (2012) on logistic regression, and adds new simulation results. The basic ideas are based on the “EM by method of weights” developed by Ibrahim (1990) for GLMs with covariates missing at random, and on the approach proposed by van den Hout and Kooiman (2006) for linear regression with covariates subject to randomized response. There is an extensive literature for missing data with either missing covariate or missing response variable. We build on these ideas by developing an EM-type algorithm that obtains unbiased estimates of GLMs when both covariate and response variables are subject to PRAM. This is a more difficult problem than either case when covariates or response variables are subject to PRAM, and has received little attention in statistical literature.

The rest of this chapter is organized as follows. Section 4.2 presents the EM-type methodology to obtain asymptotically unbiased estimates of GLMs when variables are subject to PRAM. Three cases are considered: (1) categorical covariates subject to PRAM, (2) response variable subject to PRAM, and (3) both the covariates and response variable subject to PRAM. Section 4.3 reports the results of simulation studies intended to evaluate the performance of the proposed algorithms using examples from logistic and Poisson regression, and to evaluate the effects of varying the parameters of the logistic regression model on the proposed methodology. Section 4.4 applies the proposed methodology to data from the 1993 Current Population Survey (CPS) (data from The National Bureau of Economic Research), and Section 4.5 contains further discussion of the results and open problems.
4.2 Generalized Linear Models with Variables Subject to PRAM

Let $Y$ denote the response variable that comes from an exponential family distribution,

$$f(y; \theta, \psi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\psi)} + c(y, \psi) \right\}$$

for some functions $a(.), b(.), c(.)$, where $\theta$ is the canonical parameter, and $\psi$ is the dispersion parameter.

Let $X$ denote a design matrix with $p$ covariates, and $\beta$ denote a $(p + 1) \times 1$ vector of the regression parameters. For generalized linear models (GLMs), the mean, $\mu_i$, of $y_i$ for observation $i$ depends on the linear predictor $\eta_i = x_i\beta$, i.e., $g(\mu_i) = \eta_i = x_i\beta$, where $g(.)$ is the link function. Thus GLMs take the following form

$$E(y_i|x_i) = \mu_i = g^{-1}(x_i\beta). \tag{4.2.1}$$

Since under the application of PRAM only the perturbed data are published, estimating GLMs with variables subject to PRAM can be viewed as an incomplete data problem, with the perturbed, released data associated with the true, unreleased data. An attractive method to estimate GLMs with variables subject to PRAM would be the EM algorithm, which is a standard iterative procedure for computing maximum likelihood estimates with incomplete data (Dempster et al., 1977; Wu, 1983). In each iteration of the algorithm, the E step derives the conditional expectation of the complete data likelihood given the observed data and the current estimates of the parameters, and the M step follows by maximizing the conditional expectation of the likelihood with respect to the parameters. Building on the ideas from Ibrahim (1990)
and van den Hout and Kooiman (2006) for covariates missing at random and for covariates subject to randomized response, respectively, we develop and implement an EM-type algorithm to obtain unbiased estimates of $\beta$ in GLMs for three cases: (1) covariates subject to PRAM; (2) response variables subject to PRAM; and (3) both covariates and response variables subject to PRAM. This work extends the work of Woo and Slavković (2012) which developed similar EM algorithms but only for logistic regression models. In the next three subsections, we present the proposed algorithm as three different algorithms to highlight the differences across the three aforementioned cases.

### 4.2.1 Categorical Covariate Subject to PRAM

We present an EM algorithm to obtain unbiased estimates of regression coefficients in case GLMs when categorical covariates are subject to PRAM. This method is similar to the “EM by method of weights” proposed by Ibrahim (1990), which is used to estimate parameters in GLMs with missing covariates; a related algorithm for the linear regression model with covariates subject to randomized response was proposed by van den Hout and Kooiman (2006). For ease of exposition, we present the case where PRAM is applied to one categorical covariate, but the methodology generalizes to applying PRAM to multiple categorical covariates.

Let $\mathbf{X} = (W, Z)$, where $\mathbf{X}$ denotes all the covariates, $W$ denotes the categorical covariate to which PRAM is applied, with $W^*$ denoting the observed and released version of private $W$, and $Z$ denoting the covariates which are not subject to PRAM, and can be both categorical and continuous. The levels of $W$ and $W^*$ are $\{w_1, ..., w_J\}$. Let $\mathbf{P}_W$ be the $J \times J$ PRAM matrix that contains the probabilities $p_{W_{jk}} = P(W^* = w_k | W = w_j)$, and $\pi^*_j = P(W^* = w_j)$ and $\pi_j = P(W = w_j)$ be the
marginal probabilities.

The joint distribution of \((x_i, y_i)\) is specified via the conditional distribution of \(y_i\) given \(x_i\) and the distribution of \(x_i\) which is a joint distribution of \((w_i, z_i)\), for observations \(i = 1, ..., n\). The complete data log-likelihood can be expressed as

\[
\ell(\phi; W, Z, y) = \sum_{i=1}^{n} \ell(\phi; x_i, y_i) = \sum_{i=1}^{n} \left\{ \ell_{y_i|x_i}(\beta) + \ell_{w_i|z_i}(\pi) + \ell_{z_i}(\gamma) \right\}, \tag{4.2.2}
\]

where \(\phi = (\beta, \gamma, \pi)\), and the distribution of \(W\) is multinomial with parameter \(\pi\). At iteration \(\nu\) of the EM algorithm, the E-step can be written as

\[
Q(\phi|\phi^{(\nu)}) = \sum_{i=1}^{n} E\left( \ell(\phi; x_i, y_i) \mid \text{data}, \phi^{(\nu)} \right) = \sum_{i=1}^{n} \sum_{j=1}^{J} P(W(i) = w_j \mid w^*(i), z(i), y(i), \phi^{(\nu)}) \ell(\phi; x_i, y_i) \tag{4.2.3}
\]

where \(\phi^{(\nu)}\) is the value of parameter \(\phi\) at iteration \(\nu\), for observation \(i\). The first part of (4.2.2) is the log-likelihood of the GLM, and for example, when we have a single covariate the last two parts reduce to the log-likelihood of a multinomial distribution with parameter \(\pi\).

Then, the M-step maximizes (4.2.3). This can be done via a weighted regression, by creating a “new” dataset, with each observation \(i\) taking on all possible values of \(W\), i.e. \((W(i) = w_1), (W(i) = w_2), ..., (W(i) = w_J)\), with weights

\[
q_j(i) = P(W(i) = w_j \mid w^*(i), z(i), \phi^{(\nu)}),
\]

where \(w^*\) is the observed value of the variable subject to PRAM. Using Bayes’ rule,
the conditional distribution of the weights $q_j$ is

$$P \left( W = w_j | W^* = w_k, Y, Z, \phi^{(\nu)} \right) = \frac{P \left( Y| w_j, Z, \phi^{(\nu)} \right) p_{Wjk} \pi_j^{(\nu)}}{\sum_{l=1}^J P \left( Y| w_l, Z, \phi^{(\nu)} \right) p_{Wlk} \pi_l^{(\nu)}}. \tag{4.2.4}$$

Note that $P(Y|.)$ in (4.2.4) is the probability density for the GLM. In the algorithm by Ibrahim (1990) for GLMs with missing covariates, the inner sum in (4.2.3) is taken over all levels of $W$, but only for observations with missing covariates. In our setting, since PRAM is applied to all observations, the inner sum is taken for all observations.

In the algorithm by van den Hout and Kooiman (2006), the weights in (4.2.3) are multiplied by the number of levels $J$; however, no justification was provided for this.

EM Algorithm I runs as follows:

**EM Algorithm I**: Initial values can be the estimates of $\beta$ from the regression of $Y \sim X^*$, where $X^* = (W^*, Z)$. $\pi^*$ can be used as the initial estimate of $\pi$.

**E-step**: Compute $q_j^{(\nu)}(i)$ using (4.2.4) for $i = 1, \ldots, n$ and $j = 1, \ldots, J$.

**M-step**: Carry out weighted regression with weights $q_j^{(\nu)}(i)$, using standard software.

**Update $\phi^{(\nu)}$:**

- $\beta^{(\nu+1)} = \hat{\beta}$ from weighted regression.
- $\pi_j^{(\nu+1)} = \frac{\sum_{i=1}^n q_j^{(\nu)}(i)}{n}$ for $j = 1, \ldots, J$.

With the updated $\phi^{(\nu+1)}$, new weights can be computed in the E-step, and the algorithm continues until convergence.

Note that for the initial estimate of $\pi$, one can also use $\hat{\pi} = (P_W^{-1})^T \pi^*$ which eliminates an update step of $\pi$; see EM Algorithm III. Furthermore, the proposed weights assume that the covariates are independent; see Appendix A. The initial estimates of $\gamma$ can come from the marginal distribution of $Z$. 


4.2.2 Response Variable Subject to PRAM

Here we present the second case when the response variable is subject to PRAM. Let $Y$ denote the binary response variable to which PRAM is applied, with $Y^*$ denoting the observed and released version of private $Y$. Let $P_Y$ be the PRAM matrix that contains the probabilities, with $p_{Yjk} = P(Y^* = k|Y = j)$, $k, j \in \{0, 1\}$. Following the method proposed in Section 4.2.1, the parameter $\pi$ in the complete data log-likelihood (4.2.2) can be estimated directly since $X$ is not subject to PRAM. Thus, at iteration $\nu$ of the algorithm, the E-step simplifies to

$$Q(\phi|\phi^{(\nu)}) = \sum_{i=1}^{n} \sum_{j=0}^{1} P(Y(i) = j|y^*(i), x(i), \beta^{(\nu)}) \{ \ell_y(x_i, \beta) \}, \quad (4.2.5)$$

Then, the M-step maximizes (4.2.5). This can be done via a weighted regression, by creating a “new” dataset, with each observation $i$ taking on $(Y = 0)$, $(Y = 1)$ with weights $r_{j}(i) = P(Y(i) = j|y^*(i), x(i), \beta^{(\nu)})$. Using Bayes’ rule, the weights $r_{j}(i)$ can be computed as

$$P(Y = j|Y^* = k, X, \beta^{(\nu)}) = \frac{P(Y = j|X, \beta^{(\nu)}) p_{Yjk}}{\sum_{l=0}^{1} p_{Ylk} P(Y = l|X, \beta^{(\nu)})}. \quad (4.2.6)$$

EM Algorithm II runs as follows:
EM Algorithm II: Initial values can be the estimates of \( \beta \) from the regression of \( Y^* \sim X \), where \( Y^* \) is the response variable subject to PRAM.

E-step:
Compute \( r_j^{(\nu)}(i) \) using (4.2.6) for \( i = 1, \ldots, n \) and \( j = 0, 1 \).

M-step:
Carry out weighted regression with weights \( r_j^{(\nu)}(i) \), using standard software. Update \( \beta^{(\nu)} \): \( \beta^{(\nu+1)} = \hat{\beta} \) from weighted regression.
With the updated \( \beta^{(\nu+1)} \), new weights can be computed in the E-step, and the algorithm continues until convergence.

4.2.3 Covariate and Response Subject to PRAM

Next, we seek to obtain unbiased parameter estimates of model (4.2.1) when both covariates and response variables are subject to PRAM. While this setting is important in practice, it is more complex than the other two and has received little attention in both PRAM and missing data literatures.

The weighted regression is done by creating a “new” dataset with each observation \( i \) taking on all possible values of \( Y \) and \( W \), i.e. \( (Y = 0), (Y = 1) \) and \( (W = w_1), (W = w_2), \ldots, (W = w_J) \) with weights \( s_{ml}(i) = P(Y(i) = m, W(i) = w_l|Y^*(i) = k, W^*(i) = w_j, Z, \phi^{(\nu)}) \). The weights can be computed as

\[
s_{ml} = \frac{p_{Y_{mk}} P(Y = m|W = w_l, Z, \phi^{(\nu)})}{\sum_{a=0}^{1} \sum_{b=0}^{1} \sum_{l=1}^{J} \sum_{c=1}^{J} \sum_{d=0}^{1} p_{W_{lj}} \pi(l) p_{Y_{ak}} P(Y = a|W = w_l, Z, \phi^{(\nu)})}
\]

\[
\times \frac{p_{W_{lj}} \pi(l) \sum_{b=0}^{1} \sum_{d=0}^{1} p_{Y_{bk}} P(Y = b|W = w_l, Z, \phi^{(\nu)})}{\sum_{c=1}^{J} \sum_{d=0}^{1} p_{W_{cj}} \pi(c) p_{Y_{dk}} P(Y = d|W = w_c, Z, \phi^{(\nu)})}.
\]

(4.2.7)
The weights in (4.2.7) are more difficult to derive mathematically than the weights presented in (4.2.4) and (4.2.6); the technical details are given in Appendix C. Again, for ease of exposition, independence of covariates is assumed. EM Algorithm III runs as follows:

**EM Algorithm III:** Initial values can be the estimates of $\beta$ from the regression of $Y^* \sim X^*$, where $X^* = (W^*, Z^*)$. $\pi$ can be estimated by $\hat{\pi} = (P_{W}^{-1})^T \pi^*$. 

**E-step:**
Compute $s_{ml}^{(\nu)}(i)$ using (4.2.7) for $i = 1, \ldots, n$, $m = 0, 1$ and $l = 1, \cdots, J$.

**M-step:**
Carry out weighted regression with weights $s_{ml}^{(\nu)}(i)$, using standard software.

*Update $\beta^{(\nu)}$:*

$\beta^{(\nu+1)} = \hat{\beta}$ from weighted regression

With the updated $\beta^{(\nu+1)}$, a new dataset with new weights can be computed in the E-step, and the algorithm continues until convergence.

### 4.3 Generalized Linear Model Estimation with Variables Subject to PRAM

We carried out a number of simulations to demonstrate the effect of PRAM and the proposed algorithm on maximum likelihood estimates of regression coefficients in generalized linear models (GLMs) for the above three cases. The main goal of the algorithm is to enable the users to obtain the same inference with PRAMed data as they would have if they had access to the original confidential data, and more specifically to reduce the bias that arises due to application of PRAM. Thus we focus our evaluation on reporting the mean bias of the estimates. We also report the standard errors and mean squared errors, along with how well the new confidence
intervals capture the estimated values obtained by performing regression on the original data. In general, our algorithms perform extremely well, in terms of low mean bias and coverage, that is it offers solid guarantees that the users would reach the same inference with PRAMed data subject to our algorithms as they would have if they had access to the original data. We present two examples: logistic regression and Poisson regression.

4.3.1 Example I: Logistic Regression

In the first example, experiments were carried out to evaluate the proposed EM algorithms in estimating the logistic regression model when variables are subject to PRAM. Let \( X \) denote a design matrix with \( p \) covariates and \( n \) observations, and let \( Y \) be the binary response variable. Let \( x_i = (x_{0,i}, \ldots, x_{p+1,i}) \) denote a vector of the covariates for observation \( i = 1, \ldots, n \). The logistic regression model can be written as

\[
E(y_i|x_i) = \frac{\exp(x_i\beta)}{1 + \exp(x_i\beta)}
\]  

(4.3.1)

where \( \beta = (\beta_0, \beta_1, \ldots, \beta_p)' \) are the regression coefficients and parameters of interest.

We fix \( x_{0,i} = 1 \) for \( i = 1, \ldots, n \) so \( \beta_0 \) is the intercept.

We ran 500 simulations, with varying sample size \( n = 100, 1000, 10000 \). For each simulation, we generate a random sample from a population of interest, with \( \beta = (0.5, 0.5)' \), a binary covariate \( X \) with \( x_{1,i} \) sampled from Bernoulli(0.4), and \( y_i \) sampled from Bernoulli\( (\frac{\exp(x_i\beta)}{1 + \exp(x_i\beta)}) \). We treat this as the confidential original data, and fit a logistic regression as if we had access to them to obtain the estimate of the slope, \( \hat{\beta}_1 \). PRAM was then applied to each of the following three cases:

1. PRAM applied to \( x_{1,i} \) to obtain \( x_{1,i}^* \).
2. PRAM applied to \( y_i \) to obtain \( y_i^* \).

3. PRAM applied to both \( x_{1,i} \) and \( y_i \) to obtain \( x_{1,i}^* \) and \( y_i^* \), respectively.

In all three cases, only the variables that were not subject to PRAM, and the variables that had been applied with PRAM are released; the original variables before the application of PRAM are considered private and unobserved. We used PRAM matrices of the following form

\[
P = \begin{pmatrix}
p & 1 - p \\
1 - p & p
\end{pmatrix},
\]

where the level of perturbation was varied by varying the value of \( p = 0.8, 0.9 \).

After the application of PRAM to the data, we fit a logistic regression to the PRAMed data, and compute the estimate of the slope, \( \hat{\beta}_{1,\text{noadjust}} \). We compare the estimates \( \hat{\beta}_{1,\text{noadjust}} \) to \( \hat{\beta}_1 \), in terms of bias, i.e., \( \hat{\beta}_{1,\text{noadjust}} - \hat{\beta}_1 \). The average bias, standard error, and the mean squared error of the estimates of \( \hat{\beta}_{1,\text{noadjust}} \) over 500 simulations were computed and are reported in Table 4.1 in the row labeled “\( \hat{\beta}_{1,\text{noadjust}} \)”. We also report a measure called “closeness proportion”, which gives an idea of how often \( \hat{\beta}_{1,\text{noadjust}} \) is close to \( \hat{\beta}_1 \). We evaluated the closeness proportion in the following manner: for each simulation, we constructed a confidence interval by adding and subtracting two standard errors from the estimate of \( \hat{\beta}_{1,\text{noadjust}} \). The proportion of intervals that contained the estimate \( \hat{\beta}_1 \) from the regression with the original data is reported under the column “closeness” in Table 4.1.

Next, we fit logistic regression to the PRAMed data but using the appropriate proposed EM algorithms to obtain the estimate of the slope, i.e., \( \hat{\beta}_{1,\text{adjust}} \). We computed the average bias, standard error, mean squared error, and the closeness proportion of \( \hat{\beta}_{1,\text{adjust}} \) with respect to \( \hat{\beta}_1 \) over 500 simulations. We report the results
Table 4.1. Average bias, standard error, mean squared error, & closeness proportion of MLEs for logistic regression with variables subject to PRAM.

<table>
<thead>
<tr>
<th>Case</th>
<th>$n$</th>
<th>$\hat{\beta}_1$, noadjust</th>
<th>$\hat{\beta}_1$, adjust</th>
<th>$p = 0.9$</th>
<th>$p = 0.8$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Bias</td>
<td>SE</td>
<td>MSE</td>
<td>closeness</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>-0.1145</td>
<td>0.4600</td>
<td>0.2247</td>
<td>0.938</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0163</td>
<td>0.5359</td>
<td>0.2874</td>
<td>0.998</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>-0.1079</td>
<td>0.1351</td>
<td>0.0299</td>
<td>0.902</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0068</td>
<td>0.1595</td>
<td>0.0255</td>
<td>0.994</td>
</tr>
<tr>
<td></td>
<td>10000</td>
<td>-0.1087</td>
<td>0.0428</td>
<td>0.0137</td>
<td>0.302</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>-0.1360</td>
<td>0.4401</td>
<td>0.2122</td>
<td>0.946</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0227</td>
<td>0.6344</td>
<td>0.4030</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>-0.1213</td>
<td>0.1319</td>
<td>0.0321</td>
<td>0.864</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0012</td>
<td>0.1831</td>
<td>0.0335</td>
<td>0.998</td>
</tr>
<tr>
<td></td>
<td>10000</td>
<td>-0.1232</td>
<td>0.0447</td>
<td>0.0172</td>
<td>0.210</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>-0.2039</td>
<td>0.4413</td>
<td>0.2363</td>
<td>0.954</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0974</td>
<td>0.9031</td>
<td>0.8250</td>
<td>0.982</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>-0.2021</td>
<td>0.1306</td>
<td>0.0579</td>
<td>0.660</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0011</td>
<td>0.2250</td>
<td>0.0506</td>
<td>0.988</td>
</tr>
<tr>
<td></td>
<td>10000</td>
<td>-0.2003</td>
<td>0.0423</td>
<td>0.0419</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0018</td>
<td>0.0726</td>
<td>0.0053</td>
<td>0.984</td>
</tr>
</tbody>
</table>
in Table 4.1 in the row labeled “$\hat{\beta}_{1,\text{adjust}}$”. The following stopping criterion
$|\beta_{1,\text{adjust}}^{(p+1)} - \beta_{1,\text{adjust}}^{(p)}| < 10^{-4}$ was used to ensure convergence of the EM algorithm.

To estimate the covariance matrix, we used the inverse of the observed information matrix. For Algorithm I, following the method described in Ibrahim et al. (2005), the estimated observed information matrix is given by

$$I(\hat{\phi}) = - \bar{Q}(\hat{\phi}|\hat{\phi}^\nu) - \sum_{i=1}^{n} \sum_{j=1}^{J} q_{j}(i) S_{i}(\hat{\phi}|x_{i}, y_{i}) S_{i}(\hat{\phi}|x_{i}, y_{i})'$$

$$+ \sum_{i=1}^{n} \hat{Q}_{i}(\hat{\phi}|\hat{\phi}^\nu) \hat{Q}_{i}(\hat{\phi}|\hat{\phi}^\nu)'$$

(4.3.2)

where $\bar{Q}(\phi|\phi^\nu) = \sum_{i=1}^{n} \sum_{j=1}^{J} q_{j}(i) \frac{\partial^{2} \ell(\phi|x_{i}, y_{i})}{\partial \phi \partial \phi^\nu}$, $\hat{Q}_{i}(\phi|\phi^\nu) = \sum_{j=1}^{J} q_{j}(i) \frac{\partial \ell(\phi|x_{i}, y_{i})}{\partial \phi}$, and

$S_{i}(\phi|x_{i}, y_{i}) = \frac{\partial \ell(\phi|x_{i}, y_{i})}{\partial \phi}$. To estimate the covariance matrices when running EM algorithms II and III, (4.3.2) was adjusted, with weights $q_{j}(i)$ replaced by weights from (4.2.6) and (4.2.7), for cases (2) and (3), respectively.

From the summary of the results presented in Table 4.1, we observe across all three cases that the estimates $\hat{\beta}_{1,\text{noadjust}}$ are more biased than $\hat{\beta}_{1,\text{adjust}}$, but also that within each case the mean bias appears to be unaffected by sample size. The mean bias increases in magnitude with a higher level of perturbation as expected, e.g., for case 1 and $n = 100$, the mean bias is -0.1145 and -0.2058 when $p = 0.9$ and $p = 0.8$, respectively. Furthermore, case 3, where we apply PRAM to both the response variable and the covariates, produces estimates which are more biased. The estimates using the proposed EM algorithms, $\hat{\beta}_{1,\text{adjust}}$, perform well in reducing the bias and thus clearly outperform $\hat{\beta}_{1,\text{noadjust}}$. For example, for case 1, with $n = 10000$ and $p = 0.9$, the mean bias of $\hat{\beta}_{1,\text{adjust}}$ is 0.0015, as compared to the mean bias of $\hat{\beta}_{1,\text{noadjust}}$, which is -0.1087. Furthermore, as the sample size increases, the reduction in bias due to use of the algorithms is even more pronounced, e.g., in case 1 when
$p = 0.9$, the mean absolute bias of $\hat{\beta}_{1, \text{adjust}}$ is 0.0163, 0.0068, and 0.0015 compared to mean absolute bias of $\hat{\beta}_{1, \text{noadjust}}$ 0.1145, 0.1079, and 0.1087 for $n = 100$, $n = 1000$, and $n = 10000$, respectively. The algorithms also perform slightly better when $p = 0.9$ compared to when $p = 0.8$, in terms of smaller mean bias. For example, in case 1 when $n = 10000$ and $p = 0.9$, the mean bias is 0.0015, and when $p = 0.8$, the mean bias is -0.0037.

When we first considered the issue of GLM parameter estimation with variables subject to PRAM, our main goal in maintaining data utility was to obtain unbiased estimates of the regression coefficients. However, depending on the goals of the data analyst, other measures such as the mean-squared error (MSE) that capture both bias and variance may be more applicable in judging the quality of the estimation. While our algorithms produce estimators with small mean biases, these estimators may have slightly higher standard errors and MSEs than in the cases when no adjustments are made to account for the affect of PRAM. For example, for $n = 100$, while the estimates $\hat{\beta}_{1, \text{adjust}}$ have smaller mean biases, their MSEs are higher in all three cases. As sample size increases, $\hat{\beta}_{1, \text{adjust}}$, however, have smaller MSEs than $\hat{\beta}_{1, \text{noadjust}}$. This is due to the reduction in bias as sample size increases. A more thoughtful consideration may need to be given if we are willing to accept the smaller bias at the expense of variance. Another consideration may be to find a PRAM matrix that minimizes variance subject to a pre-specified level of bias that can be tolerated.

We introduced the closeness proportion measure as a way to gauge how often $\hat{\beta}_{1, \text{noadjust}}$ and $\hat{\beta}_{1, \text{adjust}}$ were close to $\hat{\beta}_1$, the estimates using the original simulated data. Our simulations consistently show that $\hat{\beta}_{1, \text{adjust}}$ have a higher closeness proportion than $\hat{\beta}_{1, \text{noadjust}}$, and in some instances significantly higher. For example, for case 2, with $n = 10000$, when no adjustments are made, only about 21% of 95% confidence intervals out of the 500 runs would cover the regression estimate that a user would get.
from fitting a regression to the original data, thus leading to false inference, whereas
this number goes to 99% using the proposed algorithms thus leading to valid inference.
While there is no discernible effect of sample size on closeness proportion for $\hat{\beta}_{1,\text{adjust}}$, for $\hat{\beta}_{1,\text{noadjust}}$ the closeness proportion decreases as sample size increases, due to the
mean bias staying roughly the same and standard error decreasing as sample size
increases.

Figure 4.1 displays the plots of $\hat{\beta}_1$ using the original data (in black), along with
the estimates of $\beta_1$ with data subject to PRAM (in red; $\hat{\beta}_{1,\text{noadjust}}$ in left column,
$\hat{\beta}_{1,\text{adjust}}$ in right column), and the intervals for $\hat{\beta}_{1,\text{noadjust}}$ and $\hat{\beta}_{1,\text{adjust}}$, for $n = 10000$
and $p = 0.90$, when a single covariate is subject to PRAM. We only show plots for
the first 50 simulations. The intervals for $\hat{\beta}_{1,\text{adjust}}$ are much more likely to contain
$\hat{\beta}_1$ than the intervals for $\hat{\beta}_{1,\text{noadjust}}$, further supporting the results captured by the
closeness proportion measure. Figure 4.2 displays similar plots, but with $p = 0.80$.
When comparing Figure 4.1 and Figure 4.2, we can see that the bias for $\hat{\beta}_{1,\text{noadjust}}$
is greater with a higher level of perturbation, and the algorithm works well for both
$p = 0.90$ and $p = 0.80$.

![Figure 4.1. Plot of coefficient estimates for logistic regression with covariate subject to PRAM, when $n = 10000$, $p = 0.90$.](image-url)
Figure 4.2. Plot of coefficient estimates for logistic regression with covariate subject to PRAM, when $n = 10000$, $p = 0.80$.

4.3.2 Example II: Poisson Regression

In the next example, we carried out experiments to evaluate the proposed EM algorithms in estimating the Poisson regression model when variables are subject to PRAM. The Poisson regression model can be written as

$$E(y_i|x_i) = \exp x_i \beta$$

where $\beta = (\beta_0, \beta_1, ..., \beta_p)'$ are the regression coefficients and parameters of interest. We fix $x_{0,i} = 1$ so $\beta_0$ is the intercept.

Similarly to the logistic regression experiments, we ran 500 simulations with varying sample sizes $n = 100, 1000, 10000$, and with the PRAM matrices of the same form and the perturbation parameters, $p = 0.8, 0.9$. For each simulation, we generate a random sample from a population of interest with $\beta = (0.2, 0.6)'$, a binary covariate $X$ with $x_{1,i}$ sampled from Bernoulli(0.5), and $y_i$ sampled from Poisson($x_i \beta$). We treated this as the confidential original data, and fit Poisson regression as if we had access to them to obtain the estimate of $\hat{\beta}_1$. PRAM was then
applied to the covariate \( x_{1,i} \) to obtain \( x_{1,i}^* \). Like in the logistic regression example, we evaluate the effect of PRAM on Poisson regression by computing and reporting the average bias, standard error, MSE, and closeness proportion measure of \( \hat{\beta}_{1,\text{noadjust}} \) with respect to \( \hat{\beta}_1 \) over 500 simulations, and of \( \hat{\beta}_{1,\text{adjust}} \) after applying the proposed EM algorithm to adjust for the affect of PRAM. The results are reported in Table 4.2.

The results show similar trends to what was discussed in the case of logistic regression, for both the values of \( \hat{\beta}_{1,\text{noadjust}} \) and \( \hat{\beta}_{1,\text{adjust}} \) with respect to \( \hat{\beta}_1 \). The estimates from the EM algorithm, \( \hat{\beta}_{1,\text{adjust}} \), significantly outperform \( \hat{\beta}_{1,\text{noadjust}} \), in terms of smaller mean bias and higher closeness proportion measure across all sample sizes, and smaller MSE as the sample size increase. For example, for \( p = 0.8 \), the mean absolute bias of \( \hat{\beta}_{1,\text{adjust}} \) is 0.0220, 0.0030, and 0.0006 in comparison to the mean absolute bias of \( \hat{\beta}_{1,\text{noadjust}} \) with values of 0.2460, 0.2468 and 0.2478 for \( n = 100 \), \( n = 1000 \), and \( n = 10000 \), respectively. As expected, the mean bias increases in magnitude with increasing level or perturbation, e.g., when \( n = 100 \), the mean bias of \( \hat{\beta}_{1,\text{noadjust}} \) is -0.1196 and -0.2460 for \( p = 0.9 \) and \( p = 0.8 \) respectively, and the proposed algorithms do better in adjusting the bias for lower level of perturbation with smaller sample size, but as \( n \) increases, there is no discernible difference in the performance of the algorithm for different levels of perturbation in adjusting the bias and lowering the MSE. The values of the closeness of proportion measure indicate that the adjustments that algorithm provides are needed in many cases in order to obtain valid inference, e.g., closeness value of 0.642 versus 0.982 when no adjustment and adjustment is made for case \( n = 100, p = 0.8 \).

Figures 4.3 and 4.4 support the above observations. Figures 4.3 displays the plots of \( \hat{\beta}_1 \) using the original data (in black), along with the estimates of \( \beta_1 \) with data subject to PRAM (in red; \( \hat{\beta}_{1,\text{noadjust}} \) in left column, \( \hat{\beta}_{1,\text{adjust}} \) in right column), and the
Table 4.2. Average bias, standard error, mean squared error, & closeness proportion of MLEs for Poisson regression with covariate subject to PRAM.

<table>
<thead>
<tr>
<th>n</th>
<th>Estimate</th>
<th>Bias</th>
<th>SE</th>
<th>MSE</th>
<th>Closeness</th>
<th>Bias</th>
<th>SE</th>
<th>MSE</th>
<th>Closeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>n = 100</td>
<td>$\hat{\beta}_{1,\text{noadjust}}$</td>
<td>-0.1196</td>
<td>0.1664</td>
<td>0.0420</td>
<td>0.864</td>
<td>-0.2460</td>
<td>0.1642</td>
<td>0.0868</td>
<td>0.642</td>
</tr>
<tr>
<td></td>
<td>$\hat{\beta}_{1,\text{adjust}}$</td>
<td>-0.0010</td>
<td>0.2079</td>
<td>0.0432</td>
<td>0.990</td>
<td>-0.0220</td>
<td>0.2565</td>
<td>0.0663</td>
<td>0.982</td>
</tr>
<tr>
<td>n = 1000</td>
<td>$\hat{\beta}_{1,\text{noadjust}}$</td>
<td>-0.1227</td>
<td>0.0507</td>
<td>0.0176</td>
<td>0.316</td>
<td>-0.2468</td>
<td>0.0504</td>
<td>0.0635</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>$\hat{\beta}_{1,\text{adjust}}$</td>
<td>0.0008</td>
<td>0.0616</td>
<td>0.0038</td>
<td>1</td>
<td>-0.0030</td>
<td>0.0747</td>
<td>0.0056</td>
<td>0.998</td>
</tr>
<tr>
<td>n = 10000</td>
<td>$\hat{\beta}_{1,\text{noadjust}}$</td>
<td>-0.1251</td>
<td>0.0165</td>
<td>0.0159</td>
<td>0</td>
<td>-0.2478</td>
<td>0.0157</td>
<td>0.0616</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>$\hat{\beta}_{1,\text{adjust}}$</td>
<td>0.0005</td>
<td>0.0206</td>
<td>0.0004</td>
<td>1</td>
<td>-0.0006</td>
<td>0.0239</td>
<td>0.0006</td>
<td>0.992</td>
</tr>
</tbody>
</table>
intervals for $\hat{\beta}_{1,\text{noadjust}}$ and $\hat{\beta}_{1,\text{adjust}}$, for $n = 10000$ and $p = 0.90$. We only show plots for the first 50 simulations. Figure 4.4 displays similar plots of the estimates of $\hat{\beta}_1$, but with $p = 0.80$. When comparing Figure 4.3 and Figure 4.4, we can see that the bias for $\hat{\beta}_{1,\text{noadjust}}$ is greater with a higher level of perturbation, and the algorithm works well for both $p = 0.90$ and $p = 0.80$.

**Figure 4.3.** Plot of coefficient estimates for Poisson regression with covariate subject to PRAM, when $n = 10000$, $p = 0.90$.

**Figure 4.4.** Plot of coefficient estimates for Poisson regression with covariate subject to PRAM, when $n = 10000$, $p = 0.80$. 

### 4.3.3 Varying the Probability of Success & Distribution of Covariates

We present additional results on performance of the proposed algorithms subject to varying the probabilities of success for a binary response variable and varying the distribution of a binary covariate. We carried out a simulation study, with one binary covariate $X$ with $x_{1,i}$ sampled from $\text{Bernoulli}(\pi)$ and $y_i$ sampled from $\text{Bernoulli}(\frac{\exp(x_i\beta)}{1+\exp(x_i\beta)})$, where $\beta = (1, \beta_1)'$, with the sample size $n = 1000$. The following values of $\beta_1 = \{-2, -0.5, 0.5, 2\}$ and $\pi = \{0.1, 0.2, 0.3, 0.4, 0.5\}$ were used to assess their effect on the performance of the EM Algorithm I. PRAM is applied to $x_{1,i}$ to obtain $x_{1,i}^*$ using the same form of the PRAM matrix as in the previous sections and varying the perturbation proportions as $p = 0.9$ and $p = 0.8$. We computed the estimates of the regression coefficients $\hat{\beta}_{1,\text{adjust}}$ over 500 simulations of model (4.3.1), and report the mean bias and closeness proportion measure of $\hat{\beta}_{1,\text{adjust}}$ in Table 4.3.

#### Table 4.3. Average bias of MLEs with closeness proportion in parentheses, when accounting for PRAM applied to a single categorical covariate, with distribution of covariate and probability of success varied.

<table>
<thead>
<tr>
<th>$\pi$</th>
<th>$p$</th>
<th>$\beta_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>-2</td>
</tr>
<tr>
<td>0.1</td>
<td>0.9</td>
<td>-0.0385 (0.970)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.1467 (0.952)</td>
</tr>
<tr>
<td>0.2</td>
<td>0.9</td>
<td>-0.0318 (0.978)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>-0.0046 (0.976)</td>
</tr>
<tr>
<td>0.3</td>
<td>0.9</td>
<td>-0.0204 (0.988)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>-0.0039 (0.982)</td>
</tr>
<tr>
<td>0.4</td>
<td>0.9</td>
<td>0.0002 (0.998)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>-0.0162 (0.984)</td>
</tr>
<tr>
<td>0.5</td>
<td>0.9</td>
<td>-0.0151 (0.994)</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.0029 (0.980)</td>
</tr>
</tbody>
</table>

The results suggest that the proposed EM algorithm works better when the
distribution of the response variable and covariate are less skewed. In terms of lower mean bias and higher closeness proportion, the algorithm gives better estimates when the value of \( \pi \) approaches 0.5, when \( \beta_1 = \{-0.5, 0.5\} \), as well as when \( p = 0.9 \). The estimates are more biased when we consider more extreme values such as \( \pi = 0.1 \) or \( \beta_1 = 2 \) which correspond to highly skewed distributions for \( X \) and \( Y \). For example, for \( p = 0.8 \), when \( \pi = 0.5 \) and \( \beta_1 = -0.5 \), the mean bias and closeness proportion is -0.0042 and 0.982, respectively; when \( \pi = 0.1 \) and \( \beta_1 = 2 \), the mean bias and closeness proportion is -0.6188 and 0.904, respectively.

### 4.3.4 Applying PRAM to Two Independent Covariates

In this section, we consider the effectiveness of the proposed algorithm when PRAM is applied to more than one covariate. We carried out an experiment with two binary covariates \( X_1 \) and \( X_2 \), with \( x_{1,i}, x_{2,i} \) sampled independently from \( \text{Bernoulli}(0.45) \) and \( \text{Bernoulli}(0.55) \), and \( y_i \) sampled from \( \text{Bernoulli}(\frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}) \). We set \( \beta = (-0.6, 0.8, -0.3)' \) and ran 500 simulations for various sample sizes \( n = 100, 1000, 10000 \), and perturbation levels of \( p = 0.8, 0.9 \) as before. Here PRAM was applied independently to the two covariates and logistic regression was performed with released data with and without using EM Algorithm I. The results are displayed in Table 4.4.

The estimates of both covariates obtained using Algorithm I, i.e., \( \hat{\beta}_{k,\text{adjust}} \) for \( k = 1, 2 \), as with the case of one covariate, are consistently less biased and have better closeness proportion measures than the estimates, \( \hat{\beta}_{k,\text{noadjust}} \), when no adjustment is applied. For example, when \( n = 1000 \) and \( p = 0.8 \), the mean bias and closeness proportion for \( \hat{\beta}_{1,\text{adjust}} \) are 0.0159 and 0.982 while for \( \hat{\beta}_{1,\text{noadjust}} \) they are -0.3154 and 0.290, and for \( \hat{\beta}_{2,\text{adjust}} \) are 0.0014 and 0.980 while for \( \hat{\beta}_{2,\text{noadjust}} \) they are...
Table 4.4. Average bias, with closeness proportion in parentheses, of MLEs for logistic regression with two independent covariates.

<table>
<thead>
<tr>
<th></th>
<th>$p = 0.9$</th>
<th>$p = 0.8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 100$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{\beta}_1$, noadjust</td>
<td>-0.1992 (0.974)</td>
<td>-0.3239 (0.886)</td>
</tr>
<tr>
<td>$\hat{\beta}_1$, adjust</td>
<td>-0.0176 (0.998)</td>
<td>0.0944 (0.976)</td>
</tr>
<tr>
<td>$\hat{\beta}_2$, noadjust</td>
<td>0.0583 (0.996)</td>
<td>0.1373 (0.946)</td>
</tr>
<tr>
<td>$\hat{\beta}_2$, adjust</td>
<td>-0.0177 (0.998)</td>
<td>-0.0093 (0.968)</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{\beta}_1$, noadjust</td>
<td>-0.1660 (0.852)</td>
<td>-0.3154 (0.290)</td>
</tr>
<tr>
<td>$\hat{\beta}_1$, adjust</td>
<td>-0.0024 (0.998)</td>
<td>0.0159 (0.982)</td>
</tr>
<tr>
<td>$\hat{\beta}_2$, noadjust</td>
<td>0.0599 (0.986)</td>
<td>0.1314 (0.826)</td>
</tr>
<tr>
<td>$\hat{\beta}_2$, adjust</td>
<td>-0.0040 (0.996)</td>
<td>0.0014 (0.980)</td>
</tr>
<tr>
<td>$n = 10000$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{\beta}_1$, noadjust</td>
<td>-0.1670 (0)</td>
<td>-0.3269 (0)</td>
</tr>
<tr>
<td>$\hat{\beta}_1$, adjust</td>
<td>-0.0019 (0.998)</td>
<td>0.0011 (0.988)</td>
</tr>
<tr>
<td>$\hat{\beta}_2$, noadjust</td>
<td>0.0645 (0.740)</td>
<td>0.1255 (0.114)</td>
</tr>
<tr>
<td>$\hat{\beta}_2$, adjust</td>
<td>0.0004 (1)</td>
<td>0.0000 (0.984)</td>
</tr>
</tbody>
</table>

0.1314 and 0.826. As before, the effect of the adjustment is more pronounced as the sample size increases, e.g. smaller bias with increasing sample size, while the closeness proportion measure is very stable although a bit higher with smaller perturbation, e.g., for for $n = 100$, the closeness proportion for $\hat{\beta}_2$, adjust is 0.998 for $p = 0.9$, and the closeness proportion is 0.968 when $p = 0.8$. $p = 0.9$. When no adjustment is made, the closeness proportion measure decreases, further supporting the case that an adjustment is needed if a user is to make a valid inference with PRAMed data. Finally, the estimates of $\hat{\beta}_2$, noadjust appear to be less affected by PRAM, when compared to the estimates of $\hat{\beta}_1$, noadjust, in terms of smaller mean bias and higher closeness proportion. For example, when $n = 10000$ and $p = 0.9$, the mean bias and closeness proportion of $\hat{\beta}_1$, noadjust are -0.1670 and 0, respectively, while the mean bias and closeness proportion of $\hat{\beta}_2$, noadjust are 0.0645 and 0.740. This observation is likely due to attenuation of the regression slope, where estimates of slopes tend towards zero when measurement error is present in the covariates. In other words, the association between the response and covariate is weakened.
Applying PRAM to covariates is similar to introducing measurement error in the covariates, so it should not be surprising to observe attenuation of the slope. In this example, we set $\beta_1 = 0.8$ and $\beta_2 = -0.3$, thus we have a more significant gradient for $\beta_1$, and hence $\hat{\beta}_{2,\text{noadjust}}$ is less affected by PRAM than $\hat{\beta}_{1,\text{noadjust}}$.

### 4.4 Application to 1993 CPS Dataset

We implemented the methodology described in Section 4.2 on data from the 1993 Current Population Survey (CPS). The data are from The National Bureau of Economic Research. The dataset contains 48,842 records on 8 categorical variables. We performed logistic regression for salary (0 = < $50,000 or 1 = > $50,000) on the covariates Sex (0 = Female or 1 = Male), Race (0 = Non White or 1 = White), and Marital Status (0 = Married or 1 = Unmarried). The parameter estimates from fitting the logistic regression with the original data are displayed in the first line of Table 4.5, labeled as “O.D.”.

We considered the following three cases: 1) marital status subject to PRAM; 2) salary subject to PRAM; and 3) both marital status and salary subject to PRAM. In each case, the following PRAM matrix was applied to the variables that were subject to PRAM

$$ P = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}. $$

We ran 500 simulations for each case with fitting the standard logistic regression (without adjustment) on the data subject to PRAM. The mean estimates and closeness proportions of $\hat{\beta}_{k,\text{noadjust}}$ for $k = 0, \cdots, 3$ are computed and are reported in Table 4.5 in the rows labeled “$\hat{\beta}_{k,\text{noadjust}}$”. Then using the EM Algorithms I, II and III, we obtain the MLEs of model (4.3.1). The algorithms typically converged at
Table 4.5. Parameter Estimates from Original Data (O.D.), data subject to PRAM without EM algorithms ($\hat{\beta}_{k,\text{noadjust}}$), and data subject to PRAM with EM Algorithms ($\hat{\beta}_{k,\text{adjust}}$). Average ML estimates with standard errors in parentheses for O.D.. Average ML estimates with closeness proportion in parentheses for $\hat{\beta}_{k,\text{noadjust}}$ and $\hat{\beta}_{k,\text{adjust}}$. Case 1: marital subject to PRAM; Case 2: salary subject to PRAM; Case 3: both marital and salary subject to PRAM.

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\beta}_0$ (O.D.)</th>
<th>$\hat{\beta}_1$ (gender)</th>
<th>$\hat{\beta}_2$ (race)</th>
<th>$\hat{\beta}_3$ (marital status)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{\beta}_{k,\text{noadjust}}$</td>
<td>-0.5475 (0)</td>
<td>0.1550 (0)</td>
<td>0.2323 (0)</td>
<td>-1.4539 (0)</td>
</tr>
<tr>
<td>$\hat{\beta}_{k,\text{adjust}}$</td>
<td>-0.7785 (0.508)</td>
<td>0.2138 (0.350)</td>
<td>0.3745 (0.928)</td>
<td>-2.3282 (0.944)</td>
</tr>
<tr>
<td>Case 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{\beta}_{k,\text{noadjust}}$</td>
<td>-1.2807 (0)</td>
<td>0.7283 (0)</td>
<td>0.2721 (0)</td>
<td>-1.0581 (0)</td>
</tr>
<tr>
<td>$\hat{\beta}_{k,\text{adjust}}$</td>
<td>-1.2469 (0.128)</td>
<td>0.4372 (0.098)</td>
<td>0.444 (0.468)</td>
<td>-2.1863 (0.262)</td>
</tr>
<tr>
<td>Case 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{\beta}_{k,\text{noadjust}}$</td>
<td>-1.4458 (0)</td>
<td>0.7398 (0)</td>
<td>0.4400 (0)</td>
<td>-1.6009 (0)</td>
</tr>
<tr>
<td>$\hat{\beta}_{k,\text{adjust}}$</td>
<td>-1.0477 (0.290)</td>
<td>0.3851 (0.124)</td>
<td>0.4249 (0.734)</td>
<td>-2.2562 (0.510)</td>
</tr>
</tbody>
</table>
around 15 steps, based on the criterion $|\beta_k^{(\nu+1)} - \beta_k^{(\nu)}| < 10^{-4}$, although we stopped it after 20 steps. The mean estimates and closeness proportions of $\hat{\beta}_{k,\text{adjust}}$ are also reported in Table 4.5 in the rows labeled “$\hat{\beta}_{k,\text{adjust}}$”.

In terms of mean bias and closeness proportion measure, EM Algorithm II appears to work best, followed by the algorithms I and III. EM Algorithm III being least effective is no surprise, since PRAM is applied to both the response variable and one of the covariates. For example, the mean biases of $\hat{\beta}_{3,\text{adjust}}$ across the 500 runs for algorithm I, II and III are 0.0604, −0.0116, 0.1303 and the closeness proportions are 0.510, 0.944, and 0.262, respectively. A plausible explanation for these results is that algorithms I and III assume independence of covariates, and in this dataset, the distribution of the covariates may not be independent. When we carried out the chi-squared test of association for all two-way tables for the covariates Sex, Race, and Marital Status, we obtained p-values that were all close to 0. Algorithm II does not require the distribution of the covariate to be specified. This issue regarding dependence of the covariates may also explain why the estimates from the algorithms do not appear to perform as well as in the simulation studies in Section 4.3. Furthermore, there is more sparseness in these data, for example, 33% of the sample were female, and 14% of the sample were non-white. Based on the simulation results from Section 3.3., we can expect that the algorithm’s effectiveness will drop although it still leads to much better results than when no adjustment is considered.

Figures 4.5, 4.6 and 4.7 display the plots of the estimate of $\beta_3$ based on the original data (in black), along with the estimates of $\beta_3$ with data subject to PRAM (in red; $\hat{\beta}_{3,\text{noadjust}}$ in left column, $\hat{\beta}_{3,\text{adjust}}$ in right column), and the intervals for $\hat{\beta}_{3,\text{noadjust}}$ and $\hat{\beta}_{3,\text{adjust}}$, for case 1, case 2 and case 3 respectively. When not adjusting for PRAM, the estimates $\beta_3$ from the original logistic regression fall outside the intervals. Indeed, for
case 3, the estimates for $\beta_3$ fall outside the range of the plot. When the adjustment is made using the EM algorithm, the value of $\hat{\beta}_3$ from the original logistic regression is more likely to fall within the intervals.

**Figure 4.5.** Plot of estimates for coefficient of marital status, Case 1: *marital* subject to PRAM.

**Figure 4.6.** Plot of estimates for coefficient of marital status, Case 2: *salary* subject to PRAM.

### 4.4.1 Disclosure Risk Assessment

When applying any SDC methodology, both disclosure risk and data utility should be evaluated. This chapter focuses on preserving data utility of microdata subject to
PRAM when fitting a GLM on the data subject to PRAM. We now briefly discuss disclosure risk assessment and how a PRAM matrix may be chosen. Ideally, the chosen PRAM matrix is one that maximizes data utility under some predetermined levels of disclosure risk set by the statistical agency.

Since the estimates from the proposed EM algorithm converge to the maximum likelihood estimates, data utility is preserved for all choices of PRAM matrices in this case. Thus, the next step is to find a PRAM matrix that satisfies some level of disclosure control or risk set by the statistical agency. A traditional measure of disclosure risk was proposed by de Wolf and van Gelder (2004), which involves calculating the conditional probability that given a record with level \( k \) in the perturbed file, the original level was \( k \) as well, \( P(X = k|X^* = k) \).

In the context of PRAM, this conditional probability can be estimated by

\[
\hat{R}_{\text{PRAM}}(k) = P(X = k|X^* = k) = \frac{p_{kk}T_\xi(k)}{\sum_l p_{lk}T_\xi(l)},
\]

where \( T_\xi(l) \) are the frequency counts in the sample for level \( l \). The numerator of (4.4.1) estimates the number of records with level \( k \) in the original file that will remain as \( k \)
in the perturbed file, and the denominator estimates the number of records with level $k$ in the original file that remain as $k$ in the perturbed file plus the number of records that were not $k$ in the original file that take on the level $k$ in the perturbed file. Using traditional threshold rules, a record is considered safe whenever a certain combination of scores on identifying variables occur at least $d$ times. A safe record can be linked with at least $d$ records in the population. If done randomly, the probability that the record is linked correctly in the population is less than or equal to $d^{-1}$; or the risk of disclosure is at most $d^{-1}$. de Wolf and van Gelder (2004) also suggested considering a record safe whenever

$$\hat{R}_{PRAM}(k) \leq \frac{T_\xi(k)}{d}. \quad (4.4.2)$$

We use Case 1, marital subject to PRAM as an example. The three-way table of counts and $\hat{R}_{PRAM}(k)$ for marital, sex, and race are displayed in Table 4.6. $\hat{R}_{PRAM}(k)$ for married females (both non-white and white) is much lower than unmarried females (non-white and white). This is expected since the number of unmarried females is greater than the number of married females in this example. Thus, even though both unmarried and married females had a 10% probability of being misclassified as married and unmarried females, respectively, the actual number of unmarried females misclassified as married females is much higher than married females being misclassified. This leads to an observation that a higher proportion of married females in the perturbed file were originally unmarried females in the original file. Depending on the disclosure rules set by the agency, such values of $\hat{R}_{PRAM}(k)$ may provide sufficient disclosure control. For example, if the agency decides that the threshold is $d = 800$, the $\hat{R}_{PRAM}(k)$ is less than $\frac{T_\xi(k)}{d}$ for all $k$, satisfying (4.4.2).
Table 4.6. Three-way Table for the CPS 1993 data by Marital, Sex and Race. For $d = 800$, values in parenthesis are $\hat{R}_{PRAM}(k)$ and $\frac{T_k}{d}$, respectively.

<table>
<thead>
<tr>
<th></th>
<th>Married</th>
<th></th>
<th>Unmarried</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-White</td>
<td>White</td>
<td>Non-White</td>
<td>White</td>
</tr>
<tr>
<td>Female</td>
<td>521 (0.6394, 0.65)</td>
<td>2288 (0.6572, 2.86)</td>
<td>2644 (0.9786, 3.31)</td>
<td>10739 (0.9769, 13.42)</td>
</tr>
<tr>
<td>Male</td>
<td>1990 (0.9029, 2.49)</td>
<td>18245 (0.9400, 22.81)</td>
<td>1925 (0.8970, 2.41)</td>
<td>10490 (0.8380, 13.11)</td>
</tr>
</tbody>
</table>

4.5 Discussion

In this chapter, we propose using the quality of inference as the yardstick by comparing the estimates of regression coefficients in GLMs computed on the data subject to PRAM, with and without using the proposed EM algorithms that aim at adjusting the bias introduced by application of PRAM. PRAM is applied to the original confidential data to reduce the risk of disclosure of sensitive attributes and make the microdata more readily available to the users outside the official statistical agencies. As expected, from our examples with logistic and Poisson regressions, the estimates are biased when fitting GLMs on data that have been subject to PRAM, thus affecting data utility, and in some instances may lead to wrong inference. In response to this issue, by using the proposed EM-type methodology, we can obtain estimates that are much closer to their true values and asymptotically unbiased, thus leading to valid statistical inference.

In general, based on the simulation results, the proposed algorithms produce rather accurate regression coefficient estimates especially with larger sample sizes (when $n = 1000$ or $n = 10000$) given that the underlying distribution is not severely skewed. This is in part due to the fact that the weights computed in the E-step are more accurate estimators with larger sample sizes. This suggests that the
algorithms will work well for data at the larger level like at national or state levels, but may not perform as well for samples from very small populations. The algorithms also more efficiently account for the bias with less skewed or sparse data. As expected, when comparing the estimates using the proposed EM algorithms with different levels of perturbation, our results suggest that the algorithms work better with a lower level of perturbation. Since we used the same stopping criterion, this should not be surprising. Perhaps using a stricter stopping criterion, when we have a higher level of perturbation, may lead to more significant bias adjustments and better closeness proportion measures.

It should be noted the algorithms did not perform as well on the 1993 Current Population Survey data. One obvious reason is range of underlying cell proportions, where \( \pi \) varies from 0.01 to 0.37. Another plausible explanation for these results is the algorithms assume independence of covariates, which may not be a reasonable assumption to make. An algorithm that allows for dependency between the covariates may be more applicable. This issue is discussed in detail in Chapter 5.

There seems to be a trend with the biasness of the slope in the regression models fitted to the data subject to PRAM. In all cases, the estimates for the slope had negative bias when the slope was positive, and positive bias when the slope was negative. This is a similar trend to what is observed with regression models with measurement error on covariates, where it has been established that the estimated slope goes towards 0 when there is measurement error in the covariate; for more on attenuation issues, see Frost and Thompson (2000). In our analysis, however, such a phenomenon was also present when the response variable was subject to PRAM. A more careful analysis needs to be carried out on the direction of bias, and how perturbation of response variable and/or covariate affects the direction of bias, and to consider if such findings can improve the proposed algorithms.
The main objective of the proposed methodology is to reduce the bias in GLM estimation due to application of PRAM to data in order to obtain more valid results. However, our results also show that with the application of PRAM, the estimates from our algorithms have higher standard errors when compared to the estimates from the GLM fitted to the PRAMed data without making any adjustments. More careful analyses of the variance trends need to be done; e.g., How much of the variance is due to PRAM and how much due to the algorithm? Would a stricter stopping criterion have any significant effect on reducing the additional variance? Furthermore, due to this additional variance, users may want to consider using MSE or some other measures to evaluate data utility and quality of statistical inference when dealing with PRAMed data and decide if our proposed algorithms are needed in their particular context.

Thus far, we have worked with complete data with no missing observations. A next step would be to expand our methodology to also include missing data, and evaluate the impact of missingness in data on data utility and disclosure risk in this setting.

Recent work by Shlomo and Skinner (2010) claims that combining sampling with a perturbation method like PRAM offers greater protection than using either method on its own. The same authors have also pointed out that PRAM itself may guarantee $\epsilon$-differential privacy (Dwork, 2006), as long as the PRAM matrix does not contain zero elements. An interesting study would be to evaluate performance of PRAM only, PRAM with sampling and our EM methodology with respect to data utility. Depending on the results of the evaluations, we may need to consider adjusting the proposed EM algorithms to account for sampling. Similarly, the adjustments could be considered to ensure that the methodology satisfies the definition of $\epsilon$-differential privacy, or possibly the more relaxed version of $(\epsilon, \delta)$-differential privacy.
There are other SDC methodologies that can be applied to categorical variables in microdata. Data swapping is used by agencies like the U.S. Census Bureau, and their approach guarantees that marginals involving the matching variables remain the same. However, the effect on regression analysis is ambiguous (Fienberg and McIntyre, 2004). Synthetic data methods are also becoming increasingly popular. Reiter (2005) carried out an empirical study using fully synthetic data with the 2000 Current Population Study, and found that the coverage probabilities for the logistic regression are extremely low. An interesting next step would be to compare the performance of synthetic data methodology to our proposed EM algorithms for PRAM.
Chapter 5

Generalized Linear Model

Estimation with PRAM, with Dependent Covariates

5.1 Introduction

In Chapter 4, we showed that parameter estimates of generalized linear models (GLMs) are biased when fitting GLMs on data that have been subject to PRAM, thus affecting data utility. To maintain data utility when data are subject to PRAM, we proposed using a methodology based on an Expectation-Maximization (EM) algorithm to obtain asymptotically unbiased parameter estimates for GLMs when variables had been subject to PRAM. This methodology is based on the “EM by method of weights” proposed by Ibrahim (1990), which was proposed for GLMs with missing covariates. A similar approach was developed in van den Hout and Kooiman (2006) for the linear regression model with covariates subject to randomized response. Our approach in Chapter 4 assumed independence of
covariates. In practice, such an assumption may not be feasible. Dealing with dependent variables that are subject to PRAM is a commonly discussed issue in the literature. In this chapter, we build on the ideas from Chapter 4 to handle the case when the covariates are dependent.

The main methodological advance in this chapter is in the specification for the distribution of the covariates subject to PRAM. In Chapter 4, since independence of covariates was assumed, this distribution was modeled as a product of univariate marginal distributions. In this chapter, the distribution of the covariates is modeled as a product of univariate conditional distributions, which allow for dependence between the covariates, thus making it a more practical specification. This specification also leads to more accurate estimators of the regression parameters. In van den Hout and Kooiman (2006), the distribution of the covariates is modeled as a joint multinomial distribution. Under most circumstances, our specification as a product of univariate conditional distributions reduces the number of parameters to be estimated for the distribution of the covariates subject to PRAM, and also results in more precise estimators for regression parameters for GLMs.

The rest of this chapter is organized as follows. Section 5.2 presents the EM-type methodology to obtain estimates of GLMs when dependent covariates are subject to PRAM, Section 5.3 presents the results of simulation studies to compare the EM-type methodology with and without the assumption of independence, Section 5.4 applies the methodology to the 1993 Current Population Survey (data from The National Bureau of Economic Research), and Section 5.5 contains a brief discussion.
5.2 Generalized Linear Models with Dependent Categorical Covariates Subject to PRAM

First, we will re-introduce some of the notation that will be used for the rest of this chapter. Many of these were defined earlier in Chapter 4. Let $Y$ denote the response variable that comes from an exponential family distribution,

$$f(y; \theta, \psi) = \exp \left\{ \frac{y \theta - b(\theta)}{a(\psi)} + c(y, \psi) \right\}$$

for some functions $a(.), b(.), c(.)$. $\theta$ is the canonical parameter, and $\psi$ is the dispersion parameter.

Let $X$ denote a design matrix with $p$ covariates, and $\beta$ denote a $(p+1) \times 1$ vector of the regression parameters. For GLMs, the mean, $\mu_i$, of $y_i$ for observation $i$ depends on the linear predictor $\eta_i = x_i \beta$, i.e., $g(\mu_i) = \eta_i = x_i \beta$, where $g(.)$ is the link function.

Let $X = (W, Z)$, where $W = (W_1, \cdots, W_h)$ denotes the categorical covariates to which PRAM is applied, with $W^*$ denoting the observed and released version of private $W$ and $h \leq p$, and $Z$ denotes the covariates which are not subject to PRAM, and can be both categorical and continuous. Let the number of levels for $W_c$ be $J_c$, for $c = 1, \cdots, h$. Let $P_{W_c}$ be the $J_c \times J_c$ PRAM transition matrix that contains the transition probabilities $p_{W_c(j,k)} = P(W^*_c = w_{ck}|W_c = w_{cj})$.

Following the methodology proposed in Chapter 4, we consider the joint distribution of $(x_i, y_i)$, for observations $i = 1, \cdots, n$. This joint distribution can be specified by the conditional distribution for $y_i$ given $x_i$ and the joint distribution of $x_i$. In a GLM setting, the main interest is in estimating the parameter vector $\beta$ of $y_i$ given $x_i$. $x_i$ itself can be written as the conditional distribution of $w_i|z_i$ and the joint distribution of $z_i$ with parameters $\pi$ and $\gamma$ respectively. The complete data
log-likelihood can be expressed as
\[
\ell (\phi; W, Z, y) = \sum_{i=1}^{n} \ell (\phi; x_i, y_i)
\]
\[
= \sum_{i=1}^{n} \left\{ \ell_{y_i|x_i}(\beta) + \ell_{w_i|z_i}(\pi) + \ell_{z_i}(\gamma) \right\}, \quad (5.2.1)
\]
where \(\phi = (\beta, \pi, \gamma)\), and the distribution of \(W|Z\) is multinomial with parameter \(\pi\).

Since \(W|Z\) can be written as a series of conditional distributions \((W_h|W_1, \cdots, W_{h-1}, Z), (W_{h-1}|W_1, \cdots, W_{h-2}, Z), \cdots, (W_1|Z)\), \(\ell_{W|Z}(\pi)\) can be decomposed as
\[
\ell_{W|Z}(\pi) = \ell_{W_h|W_1, \cdots, W_{h-1}, Z}(\pi_h)
\]
\[
+ \ell_{W_{h-1}|W_1, \cdots, W_{h-2}, Z}(\pi_{h-1})
\]
\[
+ \cdots
\]
\[
+ \ell_{W_1|Z}(\pi_1), \quad (5.2.2)
\]
where \(\pi = (\pi_1, \cdots, \pi_h)\) with \(\pi_k\) being a vector of parameters for the \(k\)th conditional distribution of \(W\). Since \(W_c\) is categorical for all \(c\), each conditional distribution of \((W_c|W_1, \cdots, W_{c-1})\) for all \(c\) can be modeled via a multinomial logistic regression. This specification of the distribution of the covariates subject to PRAM, \(W|Z\), as a series of conditional distributions in (5.2.2) differs from the specification in (4.2.2) from Chapter 4. In Chapter 4, independence of the covariates \(W\) was assumed, hence the distribution of \(W|Z\) was expressed as a product of the distributions \((W_1|Z), (W_2|Z), \cdots, (W_h|Z)\). In the specification in (5.2.2), independence is not assumed, thus making it a more practical specification.

It should also be noted that in Ibrahim (1990) and van den Hout and Kooiman
(2006), the distribution of the missing covariates and covariates subject to randomized response was expressed as a joint multinomial distribution. By expressing the distribution of the covariates subject to PRAM as a series of conditional distributions as in (5.2.2), the number of parameters to be estimated for the distribution of the covariates subject to PRAM can be reduced, especially when the number of levels of the covariates get large. For example, in the conditional specification in (5.2.2), if the best fit for the conditional distributions $(W_{c|W_1, \cdots, W_{c-1}})$ for all $c$ are multinomial logistic regression with main effects, the number of parameters is reduced to 
\[
\sum_{i=1}^{h} (J_i - 1) \times \left[ 1 + \sum_{j=1}^{i-1} (J_j - 1) \right],
\]
whereas a joint multinomial distribution will require $(\prod_{i=1}^{h} J_i) - 1$ parameters. The proof of the reduction in number of parameters is in Appendix C.

We denote the conditional probabilities of
\[
P(W_{c|W_1} = w_{c,1}, \cdots, W_{c|W_{c-1}} = w_{c-1,j_{c-1}}, Z)\]
and
\[
P(W_{c} = w_{c,j_{c}}|W_1 = w_{1,j_1}, \cdots, W_{c-1} = w_{c-1,j_{c-1}}, Z)\]
by $\pi^*_c(j_1, \cdots, j_{c-1})$ and $\pi_{c,j_{c}}(j_1, \cdots, j_{c-1})$ respectively, for $c = 1, \cdots, h$.

At iteration $\nu$ of the EM algorithm, the E step derives the conditional expectation of the complete data log-likelihood given the observed data and current estimates of the parameters. Thus the E step can be written as
\[
Q(\phi|\phi^{(\nu)}) = \sum_{i=1}^{n} E\left( \ell(\phi; x_i, y_i) | \text{data}, \phi^{(\nu)} \right)
= \sum_{i=1}^{n} \sum_{c_1=1}^{J_1} \cdots \sum_{c_h=1}^{J_h} P(W_1(i) = w_{1,c_1}, \cdots, W_h(i) = w_{h,c_h} | \text{data}, \phi^{(\nu)}) \ell(\phi; x_i, y_i),
\]
where $\text{data} = (w^*(i), z(i), y(i))$ and $\phi^{(\nu)}$ is the estimate of parameter $\phi$ at iteration $\nu$. The first part of the complete data log-likelihood (5.2.1) is the log-likelihood of the GLM, the second part becomes a series of log-likelihood of $h$ multinomial distributions (since $W_1, \cdots, W_h$ are all categorical), and the last part is the log-likelihood of the
covariates that were not subject to PRAM.

The M-step maximizes (5.2.3), which can be obtained via a weighted regression, by creating a “new” dataset, with each observation \( i \) taking on all possible values of \( W_1, \ldots, W_h \), with weights

\[
q_{j_1, \ldots, j_h}(i) = P(W(i) = (w_{1j_1}, \ldots, w_{hj_h})|y(i), w^*(i), z(i), \phi^{(\nu)}),
\]

where \( w^* \) is the observed value of the covariates subject to PRAM. Using Bayes’ rule, the conditional distribution of the weights \( q_{j_1, \ldots, j_h} \) is

\[
P\left( W = (w_{1j_1}, \ldots, w_{hj_h})|W^* = (w_{1k_1}, \ldots, w_{kh_h}), Y, Z, \phi^{(\nu)} \right) = \\
P\left( Y|w_{1j_1}, \ldots, w_{hj_h}, Z, \phi^{(\nu)} \right) \prod_{m=1}^{h} p_{wm}(j_m, k_m) \prod_{m=1}^{h} \pi_{m,jm}(j_1, \ldots, j_{m-1})
\]

\[
\sum_{c_1=1}^{J_1} \cdots \sum_{c_h=1}^{J_h} \left[ P\left( Y|w_{1c_1}, \ldots, w_{hc_h}, Z, \phi^{(\nu)} \right) \prod_{l=1}^{h} p_{w_l}(c_l, k_l) \prod_{l=1}^{h} \pi_{l,cl}(c_1, \ldots, c_{l-1}) \right].
\]

(5.2.4)

EM Algorithm IV runs as follows:
**EM Algorithm IV:** Initial values can be the estimates of \( \beta \) from the regression of \( Y \sim X^* \), where \( X^* = (W^*, Z) \). \( \pi^* \) can be used as the initial estimate of \( \pi \).

**E-step:**
Compute \( q_{j_1,\ldots,j_h}^{(\nu)}(i) \) using (5.2.4) for \( i = 1, \ldots, n \) and \( j_c = 1, \ldots, J_c, c = 1, \ldots, h \).

**M-step:**
Carry out weighted regression with weights \( q_{j_1,\ldots,j_h}^{(\nu)}(i) \), using standard software.

Update \( \pi^{(\nu)}, \beta^{(\nu)} \):
- \( \beta^{(\nu+1)} = \hat{\beta} \) from weighted regression of \( Y \) on \( (W, Z, \beta^{(\nu)}) \).
- \( \pi^{(\nu+1)} \) from the weighted multinomial logistic regressions of \( W_1 \) on \( (Z, \pi_1^{(\nu)}) \), \( W_2 \) on \( (W_1, Z, \pi_2^{(\nu)}) \), \ldots, \( W_h \) on \( (W_1, \ldots, W_{h-1}, Z, \pi_h^{(\nu)}) \).

With the updated \( \beta^{(\nu+1)} \) and \( \pi^{(\nu+1)} \), a new dataset with new weights can be computed in the E-step, and the algorithm continues until convergence.

### 5.3 Comparison of EM Algorithms, With and Without Assumption of Independence

We carried out experiments to evaluate Algorithm IV from Chapter 5.2 and Algorithm I from Chapter 4.2.1. The difference in the algorithms is that Algorithm IV does not assume independence of the covariates that were subject to PRAM. We also evaluate an algorithm which models the covariates subject to PRAM as a joint multinomial distribution. The main goal of the algorithm is to enable the users to obtain the same inference with PRAMed data as they would have if they had access to the original confidential data, and more specifically to reduce the bias that arises
due to application of PRAM. Thus we focus our evaluation on reporting the mean bias of the estimates. We also report the standard errors and mean squared errors, along with how well the new confidence intervals capture the estimated values obtained by performing regression on the original data. In general, Algorithm IV performs extremely well, in terms of low mean bias and coverage, that is it offers solid guarantees that the users would reach the same inference with PRAMed data subject to our algorithms as they would have if they had access to the original data. Algorithm IV also improves on Algorithm I when we have dependent covariates. We will compare three different scenarios:

1. Compare Algorithm IV with Algorithm I on dependent covariates.

2. Compare Algorithm IV with Algorithm I on independent covariates.

3. Compare Algorithm IV with the approach in van den Hout and Kooiman (2006), where covariates subject to PRAM were modeled as a joint multinomial distribution.

5.3.1 Algorithm IV and Algorithm I With Dependent Data

Let $Y$ denote the binary response variable, and let $X$ denote a design matrix with $p$ covariates. Let $x_i = (x_{0,i}, x_{1,i}, \ldots, x_{p+1,i})$ denote a vector of the covariates for observation $i = 1, \ldots, n$. The logistic regression model can be written as

$$E(y_i|x_i) = \frac{\exp(x_i\beta)}{1 + \exp(x_i\beta)} \quad (5.3.1)$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ are the regression coefficients and parameters of interest. We fix $x_{0,i} = 1$ so $\beta_0$ is the intercept.
We ran 500 simulations, with varying sample size \( n = 100, 1000, 10000 \). For each simulation, we generate a random sample from a population of interest, with \( \beta = (-0.6, 0.8, 0.3)' \), two binary covariates \( X_1 \) and \( X_2 \) with \( x_{1,i} \) sampled from \( \text{Bernoulli}(0.5) \), \( x_{2,i} | x_{1,i} \) sampled from \( \text{Bernoulli}(0.3 + 0.4X_1) \), and \( y_i \) sampled from \( \text{Bernoulli}(\frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}) \). We treat this as the confidential original data, and fit a logistic regression as if we had access to them to obtain the estimates of the slopes, \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \).

PRAM was then applied independently to \( x_{1,i} \) and \( x_{2,i} \) to obtain \( x_{1,i}^* \) and \( x_{2,i}^* \) respectively, for \( i = 1, \cdots, n \). Only the variables that were not subject to PRAM, and the variables that had been applied with PRAM are released; the original variables before the application of PRAM are considered private and unobserved. We used PRAM matrices of the following form

\[
P = \begin{pmatrix}
0.9 & 0.1 \\
0.1 & 0.9 \\
\end{pmatrix}.
\]

After the application of PRAM to the data, we fit a logistic regression to the PRAMed data, and compute the estimates of the slopes, \( \hat{\beta}_{1,\text{noadjust}} \) and \( \hat{\beta}_{2,\text{noadjust}} \). We compare the estimates \( \hat{\beta}_{1,\text{noadjust}} \) and \( \hat{\beta}_{2,\text{noadjust}} \) to \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \) respectively, in terms of bias, i.e., \( \hat{\beta}_{k,\text{noadjust}} - \hat{\beta}_k \), for \( k = 1, 2 \). The average bias, standard error, and mean squared error of \( \hat{\beta}_{k,\text{noadjust}} \) over 500 simulations were computed and are reported in Table 5.1 in the rows labeled \( \text{“} \hat{\beta}_{1,\text{noadjust}} \text{“} \) and \( \text{“} \hat{\beta}_{2,\text{noadjust}} \text{“} \). We also report a measure called \( \text{“} \text{closeness proportion} \text{“} \), which gives an idea of how often \( \hat{\beta}_{k,\text{noadjust}} \) is close to \( \hat{\beta}_k \). We evaluated the closeness proportion in the following manner: for each simulation, we constructed a confidence interval by adding and subtracting two standard errors from the estimate of \( \hat{\beta}_{k,\text{noadjust}} \). The proportion of intervals that contained the estimate
\( \hat{\beta}_k \) from the regression with the original data is reported under the column “closeness” in Table 5.1.

Table 5.1. Average bias, standard error, mean squared error, & closeness proportion of MLEs for logistic regression when dependent covariates subject to PRAM. Comparison between Algorithm I, which assumed independence, and Algorithm IV, which did not assume independence.

<table>
<thead>
<tr>
<th></th>
<th>Bias</th>
<th>SE</th>
<th>MSE</th>
<th>closeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n = 100 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \hat{\beta}_{1,\text{noadjust}} )</td>
<td>-0.2785</td>
<td>0.4235</td>
<td>0.2369</td>
<td>0.934</td>
</tr>
<tr>
<td>( \hat{\beta}_{1,\text{ind}} )</td>
<td>-0.1237</td>
<td>0.5536</td>
<td>0.3218</td>
<td>0.976</td>
</tr>
<tr>
<td>( \hat{\beta}_{1,\text{dep}} )</td>
<td>0.0281</td>
<td>0.8897</td>
<td>0.7923</td>
<td>0.986</td>
</tr>
<tr>
<td>( \hat{\beta}_{2,\text{noadjust}} )</td>
<td>0.2012</td>
<td>0.4518</td>
<td>0.2447</td>
<td>0.966</td>
</tr>
<tr>
<td>( \hat{\beta}_{2,\text{ind}} )</td>
<td>0.1657</td>
<td>0.5897</td>
<td>0.3752</td>
<td>0.976</td>
</tr>
<tr>
<td>( \hat{\beta}_{2,\text{dep}} )</td>
<td>-0.0457</td>
<td>0.9212</td>
<td>0.8507</td>
<td>0.988</td>
</tr>
<tr>
<td>( n = 1000 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \hat{\beta}_{1,\text{noadjust}} )</td>
<td>-0.2273</td>
<td>0.1340</td>
<td>0.0713</td>
<td>0.660</td>
</tr>
<tr>
<td>( \hat{\beta}_{1,\text{ind}} )</td>
<td>-0.0783</td>
<td>0.1779</td>
<td>0.0378</td>
<td>0.992</td>
</tr>
<tr>
<td>( \hat{\beta}_{1,\text{dep}} )</td>
<td>0.0078</td>
<td>0.2085</td>
<td>0.0435</td>
<td>0.994</td>
</tr>
<tr>
<td>( \hat{\beta}_{2,\text{noadjust}} )</td>
<td>0.1678</td>
<td>0.1394</td>
<td>0.0476</td>
<td>0.830</td>
</tr>
<tr>
<td>( \hat{\beta}_{2,\text{ind}} )</td>
<td>0.1316</td>
<td>0.1771</td>
<td>0.0487</td>
<td>0.970</td>
</tr>
<tr>
<td>( \hat{\beta}_{2,\text{dep}} )</td>
<td>-0.0121</td>
<td>0.2088</td>
<td>0.0437</td>
<td>0.990</td>
</tr>
<tr>
<td>( n = 10000 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \hat{\beta}_{1,\text{noadjust}} )</td>
<td>-0.2268</td>
<td>0.0411</td>
<td>0.0531</td>
<td>0</td>
</tr>
<tr>
<td>( \hat{\beta}_{1,\text{ind}} )</td>
<td>-0.0802</td>
<td>0.0521</td>
<td>0.0091</td>
<td>0.720</td>
</tr>
<tr>
<td>( \hat{\beta}_{1,\text{dep}} )</td>
<td>0.0007</td>
<td>0.0611</td>
<td>0.0037</td>
<td>1</td>
</tr>
<tr>
<td>( \hat{\beta}_{2,\text{noadjust}} )</td>
<td>0.1743</td>
<td>0.0429</td>
<td>0.0322</td>
<td>0.004</td>
</tr>
<tr>
<td>( \hat{\beta}_{2,\text{ind}} )</td>
<td>0.1414</td>
<td>0.0543</td>
<td>0.0229</td>
<td>0.188</td>
</tr>
<tr>
<td>( \hat{\beta}_{2,\text{dep}} )</td>
<td>0.0022</td>
<td>0.0632</td>
<td>0.0040</td>
<td>0.996</td>
</tr>
</tbody>
</table>

Next, we fit logistic regression to the PRAMed data but using Algorithm I and Algorithm IV to obtain the estimate of the slope, i.e., \( \hat{\beta}_{k,\text{ind}} \) and \( \hat{\beta}_{k,\text{dep}} \). We computed the average bias, standard error, mean squared error, and the closeness proportion of \( \hat{\beta}_{k,\text{ind}} \) and \( \hat{\beta}_{k,\text{dep}} \) with respect to \( \hat{\beta}_k \) over 500 simulations. We report the results in Table 5.1. The rows labeled “\( \hat{\beta}_{k,\text{ind}} \)” display the results from Algorithm I which assumed independence of covariates, and the rows labeled “\( \hat{\beta}_{k,\text{dep}} \)” display the results from Algorithm IV which does not assume independence of covariates.

From the summary of results presented in Table 5.1, we observe for all sample
sizes that the estimates $\hat{\beta}_{k,\text{dep}}$ are less biased than $\hat{\beta}_{k,\text{ind}}$. This suggests $\hat{\beta}_{k,\text{dep}}$ from Algorithm IV reduces bias and outperform $\hat{\beta}_{k,\text{ind}}$ from Algorithm I. For example, when $n = 10000$, the mean bias of $\hat{\beta}_{1,\text{dep}}$ is 0.0007, as compared to the mean bias of $\hat{\beta}_{1,\text{ind}}$, which is -0.0802. As the sample size increases, the reduction in bias due to Algorithm IV is more pronounced. For example, the mean absolute bias of $\hat{\beta}_{1,\text{dep}}$ is 0.0281, 0.0078, and 0.0007 compared to the mean absolute bias of $\hat{\beta}_{1,\text{ind}}$ 0.1237, 0.0783, and 0.0802.

When we first considered the issue of GLM parameter estimation with variables subject to PRAM, our main goal in maintaining data utility was to obtain unbiased estimates of the regression coefficients. However, depending on the goals of the data analyst, other measures such as the mean-squared error (MSE) that capture both bias and variance may be more applicable in judging the quality of the estimation. While Algorithm IV produce estimators with small mean biases, these estimators may have slightly higher standard errors and MSEs than in the estimators from Algorithm I. As sample size increases, $\hat{\beta}_{k,\text{dep}}$ have smaller MSEs than $\hat{\beta}_{k,\text{ind}}$, for example, when $n = 10000$. This is due to the reduction in bias as sample size increases.

We introduced the closeness proportion as a way to gauge how often $\hat{\beta}_{k,\text{noadjust}}$, $\hat{\beta}_{k,\text{ind}}$, and $\hat{\beta}_{k,\text{dep}}$ were close to $\hat{\beta}_k$, the estimates using the original simulated data. Our results show that $\hat{\beta}_{k,\text{dep}}$ have higher closeness proportions than $\hat{\beta}_{k,\text{ind}}$, and the proportions are significantly higher when $n = 10000$. For example, the closeness proportions of $\hat{\beta}_{k,\text{dep}}$ are 1 and 0.996, whereas the closeness proportion of $\hat{\beta}_{k,\text{ind}}$ are 0.720 and 0.188. The results show that estimates from Algorithm IV lead to valid inference, while the estimates from Algorithm I may lead to false inference, especially with large sample sizes.

Figure 5.1 displays the plots of the estimates of $\hat{\beta}_1$ using the original data (in black), along with the estimates of $\beta_1$ with data subject to PRAM using algorithm
I and IV (in red; $\hat{\beta}_{1,\text{ind}}$ in left column, $\hat{\beta}_{1,\text{dep}}$ in right column), and the intervals for $\hat{\beta}_{1,\text{ind}}$ and $\hat{\beta}_{1,\text{dep}}$, for $n = 10000$. We only show plots for the first 50 simulations. The intervals for $\hat{\beta}_{k,\text{dep}}$ are much more likely to contain $\hat{\beta}_k$, which support the results from the closeness proportion measure.

![Figure 5.1. Plot of estimates for $\beta_1$, using Algorithm I and Algorithm IV.](image)

### 5.3.2 Algorithm IV and Algorithm I With Independent Data

Next, we compare Algorithm I and Algorithm IV, when we have independent covariates. Similarly to the previous experiment with dependent covariates, we ran 500 simulations with varying sample sizes $n = 100, 1000, 10000$, and with the PRAM matrices of the same form. For each simulation, we generate a random sample from a population of interest with $\beta = (-0.6, 0.8, -0.3)'$, two covariates, $X_1$ and $X_2$, with $x_{1,i}$ and $x_{2,i}$ sampled independently from Bernoulli(0.5), and $y_i$ sampled from Bernoulli($\frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}$). We treated this as the confidential original data, and fit logistic regression as if we had access to them to obtain the estimate of $\hat{\beta}_1$ and $\hat{\beta}_2$. PRAM was then applied independently to $x_{1,i}$ and $x_{2,i}$ to obtain $x^*_1,i$ and $x^*_2,i$. 


respectively, for \( i = 1, \ldots, n \). Like in the previous example, we evaluate Algorithm I and Algorithm IV with independent covariates by computing and reporting the average bias, standard error, MSE, and closeness proportion measure of the slopes, \( \hat{\beta}_{k,\text{ind}} \) and \( \hat{\beta}_{k,\text{dep}} \), over 500 simulations. The rows labeled \( \hat{\beta}_{k,\text{ind}} \) display the results from Algorithm I which assumed independence of covariates, and the rows labeled \( \hat{\beta}_{k,\text{dep}} \) display the results from Algorithm IV which does not assume independence of covariates.

Table 5.2. Average bias, standard error, mean squared error, & closeness proportion of MLEs for logistic regression when independent covariates subject to PRAM. Comparison between Algorithm I, which assumed independence, and Algorithm IV, which did not assume independence.

| \( n \) | \( \hat{\beta}_{1,\text{ind}} \) | Bias | SE | MSE | closeness | \( \hat{\beta}_{1,\text{dep}} \) | Bias | SE | MSE | closeness | \( \hat{\beta}_{2,\text{ind}} \) | Bias | SE | MSE | closeness | \( \hat{\beta}_{2,\text{dep}} \) | Bias | SE | MSE | closeness |
| 100 | 0.0175 | 0.5569 | 0.3104 | 0.994 | 0.0419 | 0.5686 | 0.3251 | 0.990 | -0.0108 | 0.5559 | 0.3102 | 0.996 | -0.0339 | 0.6091 | 0.3721 | 0.992 |
| 1000 | 0.0021 | 0.1635 | 0.0267 | 1 | 0.0021 | 0.1713 | 0.0293 | 0.996 | 0.0006 | 0.1646 | 0.0271 | 1 | -0.0014 | 0.1673 | 0.0280 | 1 |
| 10000 | -0.0008 | 0.0511 | 0.0026 | 1 | 0.0011 | 0.0531 | 0.0028 | 0.998 | -0.0004 | 0.0532 | 0.0028 | 0.998 | 0.0002 | 0.0533 | 0.0028 | 0.996 |

In this experiment, we observe different results to what was discussed in the previous subsection with dependent covariates. While both algorithms reduce the bias in the slopes significantly, the bias of \( \hat{\beta}_{k,\text{ind}} \) is significantly smaller than the bias of \( \hat{\beta}_{k,\text{dep}} \) for small sample size, \( n = 100 \). For example, the mean bias of \( \hat{\beta}_{1,\text{ind}} \) is 0.0175 compared to the mean bias of \( \hat{\beta}_{1,\text{dep}} \) 0.0419. Similar to the previous experiment, the standard errors of \( \hat{\beta}_{k,\text{ind}} \) are smaller than the standard errors of \( \hat{\beta}_{k,\text{dep}} \), for all sample sizes, although the difference is more significant for \( n = 100 \). The MSEs
of $\hat{\beta}_{k,\text{ind}}$ are also smaller than the MSEs of $\hat{\beta}_{k,\text{dep}}$. These results suggest that with independent data, Algorithm I performs better than Algorithm IV, in terms of smaller bias, standard errors and MSE.

### 5.3.3 Covariates as Series of Conditional Distributions and as Joint Distribution

Next, we carried out an experiment to evaluate Algorithm IV, which specifies the covariates subject to PRAM as a series of conditional distributions, with the approach in van den Hout and Kooiman (2006), which specified the distribution of the covariates subject to PRAM as a joint multinomial distribution. In this experiment, we ran 500 simulations with varying sample sizes $n = 100, 1000, 10000$, and with the PRAM matrices of the same form. For each simulation, we generate a random sample from a population of interest with $\beta = (-0.6, 0.8, -0.3, 0.2)'$, three binary covariates $X_1, X_2, X_3$ with $x_{1,i}, x_{2,i}, x_{3,i}$ sampled from Bernoulli(0.5), and $y_i$ sampled from Bernoulli($\frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}$). We treated this as the confidential original data, and fit logistic regression as if we had access to them to obtain the estimates of $\hat{\beta}_k$, for $k = 1, 2, 3$. PRAM was then applied independently to $x_{1,i}$, $x_{2,i}$, and $x_{3,i}$ to obtain $x^*_{1,i}$, $x^*_{2,i}$ and $x^*_{3,i}$ respectively, for $i = 1, \cdots, n$.

Next, we ran Algorithm IV and an algorithm that specifies the distribution of the covariates subject to PRAM as a joint multinomial distribution to obtain the MLEs of the logistic regression model. The average bias, standard error, mean squared error, and the closeness proportion of the estimates of $\beta_1$, $\beta_2$, and $\beta_3$ over 500 simulations are computed and reported in Table 5.3. The rows labeled “$\hat{\beta}_{k,\text{cond}}$” display the results from Algorithm IV, and the rows labeled “$\hat{\beta}_{k,\text{joint}}$” display the results from the algorithm which modeled the covariates subject to PRAM as a joint distribution.
Table 5.3. Average bias, standard error, mean squared error, & closeness proportion of MLEs for logistic regression when independent covariates subject to PRAM. Comparison between specifying distribution of covariates subject to PRAM, as series of conditional distributions, and as joint distribution.

<table>
<thead>
<tr>
<th>n</th>
<th>$\hat{\beta}_{1,\text{cond}}$</th>
<th>$\hat{\beta}_{1,\text{joint}}$</th>
<th>$\hat{\beta}_{2,\text{cond}}$</th>
<th>$\hat{\beta}_{2,\text{joint}}$</th>
<th>$\hat{\beta}_{3,\text{cond}}$</th>
<th>$\hat{\beta}_{3,\text{joint}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.0655</td>
<td>0.1150</td>
<td>0.0349</td>
<td>-0.1345</td>
<td>0.1090</td>
<td>0.2253</td>
</tr>
<tr>
<td></td>
<td>0.8325</td>
<td>0.8833</td>
<td>0.8223</td>
<td>0.9955</td>
<td>0.6794</td>
<td>0.7645</td>
</tr>
<tr>
<td></td>
<td>0.6973</td>
<td>0.7934</td>
<td>0.6774</td>
<td>1.009</td>
<td>0.4735</td>
<td>0.6351</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>-0.0453</td>
<td>0.0130</td>
<td>0.0126</td>
<td>-0.0156</td>
<td>0.0802</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>0.1934</td>
<td>0.2023</td>
<td>0.1867</td>
<td>0.2094</td>
<td>0.1765</td>
<td>0.1945</td>
</tr>
<tr>
<td></td>
<td>0.0395</td>
<td>0.0411</td>
<td>0.0350</td>
<td>0.0441</td>
<td>0.0376</td>
<td>0.0378</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10000</td>
<td>-0.0059</td>
<td>0.0000</td>
<td>0.0130</td>
<td>0.0015</td>
<td>0.0083</td>
<td>-0.0004</td>
</tr>
<tr>
<td></td>
<td>0.0601</td>
<td>0.0637</td>
<td>0.0584</td>
<td>0.0622</td>
<td>0.0538</td>
<td>0.0544</td>
</tr>
<tr>
<td></td>
<td>0.0036</td>
<td>0.0041</td>
<td>0.0036</td>
<td>0.0039</td>
<td>0.0030</td>
<td>0.0030</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Our results show some interesting observations. The estimates $\hat{\beta}_{k,\text{cond}}$ and $\hat{\beta}_{k,\text{joint}}$ reduce the bias significantly. For small sample size, $n = 100$, the biases of $\hat{\beta}_{k,\text{cond}}$ are smaller than the biases of $\hat{\beta}_{k,\text{joint}}$. For example, the mean absolute biases of $\hat{\beta}_{k,\text{cond}}$ are 0.0655, 0.0349, and 0.1090, compared to the absolute mean biases of $\hat{\beta}_{k,\text{joint}}$ 0.1150, 0.1345, and 0.2253. However, for larger sample sizes, $n = 1000, 10000$, the biases of $\hat{\beta}_{k,\text{cond}}$ are larger than the biases of $\hat{\beta}_{k,\text{joint}}$. For example, when $n = 10000$, the mean absolute biases of $\hat{\beta}_{k,\text{cond}}$ are 0.0053, 0.0130, and 0.0083, compared to the mean absolute biases of $\hat{\beta}_{k,\text{joint}}$ 0, 0.0015, and 0.0004. In terms of standard errors, $\hat{\beta}_{k,\text{cond}}$ have smaller standard errors compared to $\hat{\beta}_{k,\text{joint}}$ for all sample sizes. Due to this reduction in standard errors, MSEs of $\hat{\beta}_{k,\text{cond}}$ are smaller. It appears that Algorithm
IV performs better with smaller sample sizes, where the distribution of the covariates subject to PRAM are specified by a product of univariate conditional distributions, than the algorithm that specifies this distribution as a joint distribution. This is due to small cell counts for each level of the joint distribution.

5.4 Application to CPS Data

We implemented the methodology using Algorithm IV as described in Section 5.2 on data from the 1993 Current Population Survey (CPS). The data are from The National Bureau of Economic Research. We seek to compare with the results from the methodology described in Section 4.4, which used Algorithm I where independence of covariates was assumed. The dataset contains 48842 records. We perform logistic regression for salary \( (0 = <$50,000 or 1 = >$50,000) \) on the covariates Sex \((0 = \text{Female} or 1 = \text{Male})\), Race \((0 = \text{Non White} or 1 = \text{White})\), and Marital Status \((0 = \text{Married} or 1 = \text{Unmarried})\). The model we were fitting was

\[
\log \frac{P(\text{Salary} = 1|X)}{P(\text{Salary} = 0|X)} = \beta_0 + \beta_1 \text{Sex} + \beta_2 \text{Race} + \beta_3 \text{Marital Status} . \tag{5.4.1}
\]

The parameter estimates from fitting the logistic regression with the original data, before the application of PRAM, are denoted by \(\hat{\beta}_k\) for \(k = 0, 1, 2, 3\).

The following PRAM matrix was applied to the variable Marital Status:

\[
P = \begin{pmatrix}
0.9 & 0.1 \\
0.1 & 0.9
\end{pmatrix} .
\]

We carried out 500 simulations. In each simulation, standard logistic regression is performed on the data subject to PRAM. We also used Algorithm I from Chapter 4.4
Table 5.4. Average bias, standard error, mean squared error, & closeness proportion of parameter estimates from data subject to PRAM without using either EM algorithm ($\hat{\beta}_{k,\text{noadjust}}$), data subject to PRAM with EM Algorithm I assuming independence of covariates ($\hat{\beta}_{k,\text{ind}}$), and data subject to PRAM with EM Algorithm IV without assumption of independence of covariates ($\hat{\beta}_{k,\text{dep}}$).

<table>
<thead>
<tr>
<th>$\hat{\beta}_0$</th>
<th>Bias</th>
<th>SE</th>
<th>MSE</th>
<th>CP</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.5873</td>
<td>0.0125</td>
<td>0.3451</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>$\hat{\beta}_1$</td>
<td>0.4543</td>
<td>0.0079</td>
<td>0.2064</td>
<td>0</td>
</tr>
<tr>
<td>$\hat{\beta}_2$</td>
<td>0.0475</td>
<td>0.0070</td>
<td>0.0023</td>
<td>0</td>
</tr>
<tr>
<td>$\hat{\beta}_3$</td>
<td>0.7158</td>
<td>0.0189</td>
<td>0.5127</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\hat{\beta}_{k,\text{noadjust}}$</th>
<th>Bias</th>
<th>SE</th>
<th>MSE</th>
<th>CP</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.1892</td>
<td>0.0131</td>
<td>0.0360</td>
<td>0.290</td>
<td></td>
</tr>
<tr>
<td>0.0996</td>
<td>0.0081</td>
<td>0.0100</td>
<td>0.124</td>
<td></td>
</tr>
<tr>
<td>0.0324</td>
<td>0.0072</td>
<td>0.0011</td>
<td>0.734</td>
<td></td>
</tr>
<tr>
<td>0.0604</td>
<td>0.0289</td>
<td>0.0045</td>
<td>0.510</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\hat{\beta}_{k,\text{ind}}$</th>
<th>Bias</th>
<th>SE</th>
<th>MSE</th>
<th>CP</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.0077</td>
<td>0.0206</td>
<td>0.0005</td>
<td>0.988</td>
<td></td>
</tr>
<tr>
<td>-0.0086</td>
<td>0.0124</td>
<td>0.0002</td>
<td>0.938</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\hat{\beta}_{k,\text{dep}}$</th>
<th>Bias</th>
<th>SE</th>
<th>MSE</th>
<th>CP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0164</td>
<td>0.0249</td>
<td>0.0009</td>
<td>0.938</td>
<td></td>
</tr>
<tr>
<td>-0.0077</td>
<td>0.0206</td>
<td>0.0005</td>
<td>0.988</td>
<td></td>
</tr>
<tr>
<td>-0.0086</td>
<td>0.0124</td>
<td>0.0002</td>
<td>0.938</td>
<td></td>
</tr>
<tr>
<td>-0.0385</td>
<td>0.0350</td>
<td>0.0027</td>
<td>0.908</td>
<td></td>
</tr>
</tbody>
</table>
and Algorithm IV on the data subject to PRAM. The mean bias, standard error, mean
squared error, and closeness proportion with respect to $\hat{\beta}_k$ of $\hat{\beta}_{k,\text{noadjust}}$, $\hat{\beta}_{k,\text{ind}}$, and $\hat{\beta}_{k,\text{dep}}$ are computed and are reported in Table 5.4 in the columns labeled “$\hat{\beta}_{k,\text{noadjust}}$”, “$\hat{\beta}_{k,\text{ind}}$”, and “$\hat{\beta}_{k,\text{dep}}$” respectively.

The results show similar trends to what was discussed in Section 5.3.1, for large
sample sizes. Algorithm IV is a better algorithm than Algorithm I. While both $\hat{\beta}_{k,\text{ind}}$ and $\hat{\beta}_{k,\text{dep}}$ significantly outperform $\hat{\beta}_{k,\text{noadjust}}$ for $k = 0, 1, 2, 3$ in terms of smaller mean
bias, the mean biases of $\hat{\beta}_{k,\text{dep}}$ are also smaller than the mean biases of $\hat{\beta}_{k,\text{ind}}$. The
mean absolute biases of $\hat{\beta}_{k,\text{dep}}$ are 0.0164, 0.0077, 0.0086, and 0.0385, compared to $\hat{\beta}_{k,\text{dep}}$ 0.1892, 0.0996, 0.0324, and 0.0604. As expected, the standard errors of $\hat{\beta}_{k,\text{dep}}$
are larger than the standard errors of $\hat{\beta}_{k,\text{ind}}$. Due to the reduction in bias, the MSEs of $\hat{\beta}_{k,\text{dep}}$ are smaller than the MSEs of $\hat{\beta}_{k,\text{ind}}$. $\hat{\beta}_{k,\text{dep}}$ also has high closeness proportions.

This indicates that $\hat{\beta}_{k,\text{dep}}$ will lead to valid inference most of the times. The closeness
proportions of $\hat{\beta}_{k,\text{dep}}$ are 0.938, 0.988, 0.938, and 0.908, compared to $\hat{\beta}_{k,\text{ind}}$, 0.290, 0.124, 0.734, and 0.510.

Figure 5.2 compares the plots of the estimate of $\beta_3$ from Algorithm I (left) and
Algorithm IV (right). Estimates from the original data (in black), along with the
estimates of $\beta_3$ using EM algorithm (in red), and the confidence intervals for the
estimates via EM algorithm (in green). The estimates from Algorithm IV are more
accurate, on average. The intervals for $\hat{\beta}_{3,\text{dep}}$ contain $\hat{\beta}_3$ most of the time, supporting
the closeness proportion measure reported earlier.

Figures 5.3, 5.4, and 5.5 display the plots of the estimate of $\beta_0$, $\beta_1$, and $\beta_2$ on
the original data (in black), along with the estimates of $\beta_i$ using EM Algorithm IV
(in red), and the confidence intervals for the estimates via EM Algorithm IV (in
green). The intervals from Algorithm IV capture $\hat{\beta}_k$ most of the time, leading to
valid inference most of the time.
5.5 Discussion

In this chapter, we propose an EM-type methodology to estimate regression parameters of GLMs when variables are subject to PRAM. The methodology builds on the ideas from Chapter 4, Ibrahim (1990), and van den Hout and Kooiman (2006). The main difference in the methodologies lies in the specification for the
Figure 5.4. Plot of estimates for $\beta_1$, when Marital Status subject to PRAM.

Figure 5.5. Plot of estimates for $\beta_2$, when Marital Status subject to PRAM.

covariates subject to PRAM. In this chapter, we specified the distribution of the covariates subject to PRAM as a product of univariate conditional distributions, whereas in Chapter 4, independence between the covariates was assumed, and thus the distribution was specified as a product of univariate marginal distributions. Thus, the specification in this chapter is more practical, and also leads to more
accurate estimators. In the specification in van den Hout and Kooiman (2006), this distribution was specified as a joint multinomial distribution, which in effect, was compounding the covariates. Our specification as a product of multinomial conditional distributions reduces the number of parameters to be estimated, and also results in more precise estimators for regression parameters.

In general, based on our results, Algorithm IV outperforms Algorithm I when we have dependent covariates. Algorithm IV produce more accurate regression coefficient estimates than Algorithm I. Another consideration that needs to be accounted for is computing time. In our experiment in Section 5.3.1, when $n = 10000$, Algorithm IV took half an hour longer to run than Algorithm I. This is due to more parameters that need to be estimated and updated in the algorithm. We expect the difference in computing time to be much more pronounced when we have more covariates with higher number of levels.

When we have independent covariates, Algorithm I slightly performs Algorithm IV, in terms of slightly higher accuracy and slightly higher precision, especially for small sample size, $n = 100$. This suggests that if we have large sample sizes, Algorithm IV can still be used.

We also compared the estimates of the regression coefficients when using Algorithm IV and when using an algorithm which modeled the distribution of the covariates subject to PRAM as a joint multinomial distribution (which follows the approach in van den Hout and Kooiman (2006)). We have shown that by specifying the distribution of the covariates subject to PRAM as a series of conditional distributions instead of a multinomial distribution, the number of parameters for the distribution is reduced. Algorithm IV also produces to more precise estimates of the regression parameters, especially in smaller sample sizes. One explanation for this is that by specifying the distribution of the covariates subject to PRAM as a
joint multinomial distribution, there are too many parameters to estimated with small numbers in the cross-classification of the covariates, thus affecting the estimates of the proportions of the cross-classifications.

Since the distribution for the covariates subject to PRAM is specified as a series of univariate conditional distributions (or series of multinomial logistic regressions), the question of finding the best fit for each of the multinomial logistic regression arises: how are variable selection techniques affected with variables subject to PRAM? Also, does the order of the conditional distributions matter?

Thus far, in Chapters 4 and 5, we have focused on using the quality of inference as a measure of data utility. Our algorithms result in unbiased estimators of regression parameters for large sample sizes, thus maintaining data utility. Disclosure risk also needs to be considered when applying any SDC methodology. In Chapter 6, we focus on obtaining an optimal PRAM matrix, one which maximizes data utility under an acceptable level of disclosure risk.
Chapter 6

Optimal PRAM with Respect to Disclosure Control

6.1 Introduction

A commonly-discussed issue in applying PRAM which has contributed to its limited use in official statistics is the problem of obtaining an optimal PRAM matrix that produces a safe file and at the same time maximizes data utility, see Cator et al. (2005), and Mares and Torra (2010). In the previous chapters with fitting GLMs to PRAMed data, since the estimates from the proposed EM algorithms converge to the maximum likelihood estimates for GLMs, data utility is preserved for all choices of PRAM matrices. However, an agency or a user may want to consider alternative analysis and measures of both the risk and utility. This chapter will focus on methodological developments to address the issue of obtaining an optimal PRAM matrix that will lead to maximizing the pre-specified utility measure and minimizing the pre-specified disclosure risk measure, that could be applied in more general setting that goes beyond the fitting of GLMs.
The chapter is organized as follows. Section 6.2 provides more background on the risk-utility consideration in SDC, Section 6.3 provides the terminology, definitions, and notation used in this chapter, Section 6.4 derives an optimal PRAM matrix for a binary variable. In Section 6.5 we extend this result to cases when PRAM is applying to variables with more than two levels, given we restrict our attention to a particular class of PRAM matrices, and in Section 6.6, this restriction is relaxed. Section 6.7 contains a discussion of results and future work.

6.2 Motivation

In the context of PRAM, an optimal PRAM matrix is a matrix that when applied to data, produces a safe file (or analogously a file within an acceptable level of disclosure risk) and maximizes data utility at the same time.

One widely-used measure of data utility for PRAM is the entropy-based information loss (EBIL) (Domingo-Ferrer and Torra, 2001a), which is a variant of Shannon’s entropy. The mathematical properties of EBIL have received little attention in the literature, possibly because it consists of many variables which make mathematical derivations complicated. Thus far, most evaluations of the effect of PRAM on EBIL have been done computationally; see Domingo-Ferrer and Torra (2001a), de Wolf and van Gelder (2004), and also Mares and Torra (2010), who noted that these evaluations are computationally expensive. We shall show that under certain conditions, the optimal PRAM matrix can be found algebraically. Using these results, we present an algorithm that obtains PRAM matrices which produce safe files with higher data utility when compared to PRAM matrices obtained using built-in numerical methods and routines.

With SDC methodology, a trade off between disclosure risk and data utility is
usually expressed in an R-U framework; see Duncan and Pearson (1991). This concept can be illustrated via a risk-utility (R-U) map (Figure 6.1). In general, applying a higher level of perturbation will result in a safer file and a file that has less data utility. The original unperturbed file will always have the highest data utility, but may contain records that are not safe, according to a pre-specified rule to determine the safety of the records. The other extreme is to consider the situation that will be safest: not releasing a file at all. This minimizes disclosure risk, but provides no use to anyone. A maximum disclosure risk constraint should be specified, and any released perturbed file must satisfy this constraint. The statistical agency holding the original file will want to perturb the file in a way that ensures the safety of all the records, and at the same time maximize the utility of the file.

\begin{figure}
\centering
\includegraphics[width=0.5\textwidth]{risk_utility_map}
\caption{Risk-Utility map.}
\end{figure}

Next, the notation for the rest of this chapter is defined, along with definitions of disclosure risk and information loss.
6.3 Notation and Definitions

As defined earlier in Chapter 3, let $\xi$ denote a categorical variable in the *original* file, and let $X$ denote the same categorical variable in the *perturbed* file. Assume $\xi$ and $X$ have the same levels: $1, \ldots, k+1$. The transition probability that an original score $\xi = k$ becomes $X = l$ in the perturbed file is denoted by $p_{kl} = P(X = l | \xi = k)$, and let $P$ be the $(k + 1) \times (k + 1)$ PRAM-matrix, with entries $p_{kl}$ as defined above. Note that the rows must sum to 1, i.e. $\sum_{l=1}^{k+1} p_{kl} = 1$. Let $T_\xi = (T_\xi(1), \ldots, T_\xi(k+1))'$ and $T_X = (T_X(1), \ldots, T_X(k+1))'$ denote the vectors of frequency counts from the original and perturbed files respectively.

6.3.1 A Measure of Disclosure Risk

de Wolf and van Gelder (2004) carried out some empirical experiments to examine the effect the entries in PRAM matrices have on disclosure risk and data utility. To evaluate disclosure risk, they estimated the conditional probability the original score was $m$ given a score $m$ in the released file, $P(\xi = m \mid X = m)$. This can be estimated by

$$\hat{R}_{PRAM}(m) = \frac{p_{mm}T_\xi(m)}{\sum_{l=1}^{k+1} p_{lm}T_\xi(l)}.$$  \hspace{1cm} (6.3.1)

Using traditional threshold rules, a record is considered safe whenever a certain combination of scores on identifying variables occur at least $d$ times. A safe record can be linked with at least $d$ records in the population. If done randomly, the probability that the record is linked correctly in the population is less than or equal to $d^{-1}$; or the risk of disclosure is at most $d^{-1}$. Using the rule defined in de Wolf and van Gelder (2004), a record is safe whenever

$$\hat{R}_{PRAM}(m) \leq \frac{T_\xi(m)}{d}.$$  \hspace{1cm} (6.3.2)
6.3.2 A Measure of Data Utility

In most of the SDC literature, data utility is measured by information loss. One common measure of information loss used for categorical data, is the Entropy-Based Information Loss (EBIL), as defined in Domingo-Ferrer and Torra (2001a),

$$EBIL = -\sum_{l=1}^{k+1} \sum_{m=1}^{k+1} T_X(l) P(\xi = m \mid X = l) \log P(\xi = m \mid X = l). \quad (6.3.3)$$

Further, the term $P(\xi = m \mid X = l)$ can be estimated by

$$\frac{p_{ml} T_\xi(m)}{\sum_{j=1}^{k+1} p_{jl} T_\xi(j)}. \quad (6.3.4)$$

6.4 Obtaining Optimal PRAM Matrix for a Binary Variable

When applying PRAM to variables in a file, the goal is to obtain a PRAM-matrix that minimizes (6.3.3), subject to the constraints (6.3.2) being satisfied. Suppose PRAM is to be applied to a binary variable, with $T_\xi = (T_\xi(1), T_\xi(2))'$. The PRAM matrix, $P$, can be written as

$$P = \begin{pmatrix} p_{11} & 1 - p_{11} \\ 1 - p_{22} & p_{22} \end{pmatrix}.$$

Using (6.3.4), the expression for $EBIL$, (6.3.3), becomes

$$EBIL(P) = -T_X(1) \frac{p_{11} T_\xi(1)}{p_{11} T_\xi(1) + (1 - p_{22}) T_\xi(2)} \log \left[ \frac{p_{11} T_\xi(1)}{p_{11} T_\xi(1) + (1 - p_{22}) T_\xi(2)} \right]$$

$$-T_X(1) \frac{(1 - p_{22}) T_\xi(2)}{p_{11} T_\xi(1) + (1 - p_{22}) T_\xi(2)} \log \left[ \frac{(1 - p_{22}) T_\xi(2)}{p_{11} T_\xi(1) + (1 - p_{22}) T_\xi(2)} \right]$$
\begin{align*}
-T_X(2) & \frac{(1 - p_{11})T_\xi(1)}{(1 - p_{11})T_\xi(1) + p_{22}T_\xi(2)} \log \left[ \frac{(1 - p_{11})T_\xi(1)}{(1 - p_{11})T_\xi(1) + p_{22}T_\xi(2)} \right] \\
- T_X(2) & \frac{p_{22}T_\xi(2)}{(1 - p_{11})T_\xi(1) + p_{22}T_\xi(2)} \log \left[ \frac{p_{22}T_\xi(2)}{(1 - p_{11})T_\xi(1) + p_{22}T_\xi(2)} \right].
\end{align*}

(6.4.1)

We want to obtain the values of $p_{11}$ and $p_{22}$ that minimize $EBIL(P)$, (6.4.1), subject to constraints in (3.3.2). We consider the following two cases:

1. Both $T_\xi(1), T_\xi(2)$ are less than the threshold $d$.

2. Only one of $T_\xi(1), T_\xi(2)$ is less than the threshold $d$.

### 6.4.1 Both Components Less Than Threshold

In this case, we consider both $T_\xi(1), T_\xi(2)$ are less than the threshold $d$. We state the following proposition, which obtains the optimal PRAM matrix when applied to a binary variable:

**Proposition 6.4.1.** If $T_\xi(1) < d$, $T_\xi(2) < d$, and the diagonal elements in the PRAM matrix are diagonally dominant, setting $p_{11} = T_\xi(2)/d$ and $p_{22} = T_\xi(1)/d$ will minimize $EBIL$, (6.3.3), and result in a safe perturbed file according to the constraints as specified in (6.3.2).

We proceed with the proof of Proposition 6.4.1.

*Proof.* The constraints (6.3.2) can be written as

\begin{align*}
\frac{p_{11}T_\xi(1)}{p_{11}T_\xi(1) + (1 - p_{22})T_\xi(2)} & \leq \frac{T_\xi(1)}{d} \quad (6.4.2) \\
\frac{p_{22}T_\xi(2)}{(1 - p_{11})T_\xi(1) + p_{22}T_\xi(2)} & \leq \frac{T_\xi(2)}{d}. \quad (6.4.3)
\end{align*}
From Gouweleeuw et al. (1998), $E(T_x \mid \xi) = P'T_\xi$. Also, let $a = p_{11}T_\xi(1)$ and $b = p_{22}T_\xi(2)$, so $EBIL(P)$, (6.4.1), becomes

$$
EBIL(a, b) = -a \log \left[ \frac{a}{a - b + T_\xi(2)} \right] - (T_\xi(2) - b) \log \left[ \frac{T_\xi(2) - b}{a - b + T_\xi(2)} \right] - (T_\xi(1) - a) \log \left[ \frac{T_\xi(1) - a}{T_\xi(1) - a + b} \right] - b \log \left[ \frac{b}{T_\xi(1) - a + b} \right],
$$

(6.4.4)

and the constraints (6.4.2) and (6.4.3) become

$$
b \leq a(1 - \frac{d}{T_\xi(1)}) + T_\xi(2)
$$

(6.4.5)

and

$$
a \leq b(1 - \frac{d}{T_\xi(2)}) + T_\xi(1).
$$

(6.4.6)

The goal is to obtain $a$ and $b$ that minimize $EBIL(a, b)$, (6.4.4), subject to constraints (6.4.5) and (6.4.6). In the application of PRAM, it is common to consider PRAM matrices that are diagonally dominant, i.e. $p_{11} + p_{22} \geq 1$ in this instance. Consequently, we obtain the following results. Since we have $p_{11} + p_{22} \geq 1$, $a = p_{11}T_\xi(1)$, and $b = p_{22}T_\xi(2)$, it follows that

$$
bT_\xi(1) + aT_\xi(2) \geq T_\xi(1)T_\xi(2).
$$

Adding $aT_\xi(1) - a^2 + ab$ and subtracting $aT_\xi(2)$ and $bT_\xi(1)$ from both sides of the
inequality, we have

\[ aT_\xi(1) - a^2 + ab \geq aT_\xi(1) - bT_\xi(1) + T_\xi(1)T_\xi(2) - a^2 + ab - aT_\xi(2). \]

Dividing throughout by \( aT_\xi(1) - bT_\xi(1) + T_\xi(1)T_\xi(2) - a^2 + ab - aT_\xi(2) \) and simplifying, we obtain

\[ \frac{a}{a - b + T_\xi(2)} \frac{T_\xi(1) - a + b}{T_\xi(1) - a} \geq 1. \]

Taking logarithms,

\[ \log \left[ \frac{a}{a - b + T_\xi(2)} \frac{T_\xi(1) - a + b}{T_\xi(1) - a} \right] \geq 0, \]

and then expanding the logarithm,

\[ \log \left[ \frac{a}{a - b + T_\xi(2)} \right] + \frac{T_\xi(2) - b}{a - b + T_\xi(2)} + \frac{b - T_\xi(2)}{a - b + T_\xi(2)} - \log \left[ \frac{T_\xi(1) - a}{T_\xi(1) - a + b} \right] \frac{b}{T_\xi(1) - a + b} \frac{b}{T_\xi(1) - a + b} \geq 0. \tag{6.4.7} \]

The expression on the LHS of (6.4.7) is the derivative of the negative of \( EBIL(a, b) \) (6.4.4). Similarly, since we have \( p_{11} + p_{22} \geq 1, a = p_{11}T_\xi(1), \) and \( b = p_{22}T_\xi(2), \) it follows that

\[ bT_\xi(1) + aT_\xi(2) \geq T_\xi(1)T_\xi(2). \]

Adding \( bT_\xi(2) - b^2 + ab \) and subtracting \( aT_\xi(2) \) and \( bT_\xi(1) \) from both sides of the inequality, we have

\[ bT_\xi(2) - b^2 + ab \geq bT_\xi(2) - aT_\xi(2) + T_\xi(1)T_\xi(2) - b^2 + ab - bT_\xi(1). \]

Dividing throughout by \( bT_\xi(2) - aT_\xi(2) + T_\xi(1)T_\xi(2) - b^2 + ab - bT_\xi(1) \) and simplifying,
we obtain
\[
\frac{b}{b - a + T_\xi(1)} \frac{T_\xi(2) - b + a}{T_\xi(2) - b} \geq 1.
\]

Taking logarithms,
\[
\log \left[ \frac{b}{b - a + T_\xi(1)} \frac{T_\xi(2) - b + a}{T_\xi(2) - b} \right] \geq 0,
\]
and then expanding the logarithm,
\[
\log \left[ \frac{b}{b - a + T_\xi(1)} \right] + \frac{T_\xi(1) - a}{b - a + T_\xi(1)} + \frac{a - T_\xi(1)}{b - a + T_\xi(1)} - \log \left[ \frac{T_\xi(2) - b}{T_\xi(2) - b + a} \right] - \frac{a}{T_\xi(2) - b + a} + \frac{a}{T_\xi(a) - b + a} \geq 0. \tag{6.4.8}
\]

The expression on the LHS of (6.4.8) is the derivative of the negative of $EBIL(a, b)$ (6.4.4). From (6.4.7) and (6.4.8), it follows that $EBIL(a, b)$, (6.4.4), is monotonic decreasing in $a$ and $b$, i.e. $\nabla EBIL(a, b) \leq 0$, and hence $EBIL(P)$, (6.4.1), is also monotonic decreasing in $p_{11}$ and $p_{22}$. A plot of $EBIL(P)$, (6.4.1), versus the diagonal entries of the PRAM matrix, $p_{11}$ and $p_{22}$, is shown in Figure 6.2.

As a result, finding the values of $a$ and $b$ that minimize $EBIL(a, b)$, (6.4.4), amounts to finding the boundary solution to the constraints (6.4.5) and (6.4.6), which results in solving the following linear equations
\[
b = a(1 - \frac{d}{T_\xi(1)}) + T_\xi(2) \tag{6.4.9}
\]
and
\[
a = b(1 - \frac{d}{T_\xi(2)}) + T_\xi(1). \tag{6.4.10}
\]
Solving the linear equations (6.4.9) and (6.4.10) result in $p_{11} = T_\xi(2)/d$ and $p_{22} = T_\xi(1)/d$.

6.4.2 Only One Component Less Than Threshold

In this case, we consider only one of $T_\xi(1)$, $T_\xi(2)$ is less than the threshold $d$. Without loss of generality, assume $T_\xi(1) \geq d$ and $T_\xi(2) < d$. For this case, we state the following proposition on how to obtain the optimal PRAM matrix.

**Proposition 6.4.2.** Without loss of generality, if $T_\xi(1) \geq d$ and $T_\xi(2) < d$, and the diagonal elements in the PRAM matrix are diagonally dominant, setting $p_{11} = 1 + (T_\xi(2) - d)T_\xi(1)^{-1}$ and $p_{22} = 1$ will minimize $EBIL$, (6.3.3), and result in a safe perturbed file according to the constraints as specified in (6.3.3).
Proof. The constraints (6.3.2) can be written as

\[ p_{22} \leq 1 \quad (6.4.11) \]

and

\[ \frac{p_{22}T_\xi(2)}{(1 - p_{11})T_\xi(1) + p_{22}T_\xi(2)} \leq \frac{T_\xi(2)}{d}. \quad (6.4.12) \]

The goal is to minimize \( EBIL(P), \) (6.4.1), subject to the constraints (6.4.11) and (6.4.12). As shown in Section 6.4.1, as long as the PRAM matrix is diagonal dominant, \( EBIL(P), \) (6.4.1), is monotonic decreasing in \( p_{11} \) and \( p_{22}. \) As a result, we only need to find the boundary solutions to the constraints (6.4.11) and (6.4.12), which results in \( p_{11} = 1 + (T_\xi(2) - d)T_\xi(1)^{-1} \) and \( p_{22} = 1. \]

Next, we consider the case for variables with more than two levels.

### 6.5 Obtaining Optimal PRAM Matrix for Variables with More Than Two Levels

For this subsection 6.5, we consider a class of \((k + 1) \times (k + 1)\) PRAM matrices, \( P = (p_{ij}), \) where

\[ p_{ij} = \begin{cases} p_i, & i = j \\ k^{-1}(1 - p_i), & i \neq j \end{cases} \quad (6.5.1) \]
for $i, j = 1, ..., k + 1$. Hence, $\mathbf{P}$ takes the form

\[
\mathbf{P} = \begin{pmatrix}
p_1 & k^{-1}(1-p_1) & k^{-1}(1-p_1) & \cdots & k^{-1}(1-p_1) & k^{-1}(1-p_1) \\
k^{-1}(1-p_2) & p_2 & k^{-1}(1-p_2) & \cdots & k^{-1}(1-p_2) & k^{-1}(1-p_2) \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
k^{-1}(1-p_k) & k^{-1}(1-p_k) & k^{-1}(1-p_k) & \cdots & p_k & k^{-1}(1-p_k) \\
k^{-1}(1-p_{k+1}) & k^{-1}(1-p_{k+1}) & k^{-1}(1-p_{k+1}) & \cdots & k^{-1}(1-p_{k+1}) & p_{k+1}
\end{pmatrix}.
\]

It needs to be noted that in the application of PRAM, it is typical to consider PRAM matrices that are diagonally dominant, i.e. $\frac{1}{2} \leq p_i \leq 1$, $i = 1, 2, ..., k + 1$. The practical implication of considering PRAM matrices that are diagonally dominant is to restrict the misclassification probability to be less than half and, in so doing, to reduce the loss of information.

### 6.5.1 Conditions for the Existence of Safe PRAM Matrices

We define a record to be safe whenever (6.3.2) is satisfied. For brevity, we now denote $T_{\xi}(i)$ by $T_i$. Also, assume that all entries of the vector of frequency counts, $\mathbf{T}$, is less than a pre-determined threshold $d$, as defined in (6.3.2), i.e. $T_i < d$, $i = 1, ..., k + 1$. We state the following proposition that provides a condition for an optimal PRAM matrix to exist.

**Proposition 6.5.1.** A PRAM matrix that results in safe records exists if

\[
d < (1 - k^{-1})T_{(1)} + (1 + k^{-1})\bar{T},
\]

where $T_{(1)} = \min\{T_i\}$ and $\bar{T} = (k + 1)^{-1}\sum_{i=1}^{k+1} T_i$. 

Proof. Using (3.3.1), it follows from (6.3.2) that a PRAM matrix is safe whenever

\[
\frac{p_i}{p_i T_i + k^{-1} \sum_{l=1,l \neq i}^{k+1} (1 - p_l) T_l} \leq \frac{1}{d},
\]

(6.5.2)

for \(i = 1, \ldots, k + 1\). Simplifying expression (6.5.2), we obtain

\[
(d - T_i) p_i + k^{-1} \sum_{l=1,l \neq i}^{k+1} p_l T_l \leq k^{-1} \sum_{l=1,l \neq i}^{k+1} T_l,
\]

(6.5.3)

for \(i = 1, \ldots, k + 1\). Let \(r_i \equiv 2(p_i - \frac{1}{2})\), so \(p_i \equiv \frac{1}{2}(r_i + 1)\). Then, \(\frac{1}{2} \leq p_i \leq 1, 0 \leq r_i \leq 1\) and (6.5.3) becomes

\[
(d - T_i) r_i + k^{-1} \sum_{l=1,l \neq i}^{k+1} r_l T_l \leq \left( T_i + k^{-1} \sum_{l=1,l \neq i}^{k+1} T_l \right) - d,
\]

(6.5.4)

for \(i = 1, \ldots, k + 1\). For solutions to (6.5.4) to exist, the right-hand side of expression (6.5.4) must be positive, i.e.

\[
d < T_i + k^{-1} \sum_{l=1,l \neq i}^{k+1} T_l
\]

\[
= T_i + k^{-1} \left( \sum_{l=1}^{k+1} T_l - T_i \right)
\]

\[
= (1 - k^{-1}) T_i + (1 + k^{-1}) T,
\]

for \(i = 1, \ldots, k + 1\). Because \(T_{(1)} = \min\{T_i\}\), then \(T_{(1)} \leq T_i, i = 1, \ldots, k + 1\). Hence,

\[
(1 - k^{-1}) T_{(1)} + (1 + k^{-1}) T \leq (1 - k^{-1}) T_i + (1 + k^{-1}) T, i = 1, \ldots, k + 1.
\]
As a result, there exists a PRAM matrix resulting in safe records if

\[ d < (1 - k^{-1})T(1) + (1 + k^{-1})\bar{T}. \]

6.5.2 Obtaining Optimal PRAM Matrix

The problem we seek to solve is: Minimize EBIL (6.3.3) subject to the constraints \( \hat{R}_{PRAM}(m) \leq \frac{T_m}{d} \) (6.3.2) being satisfied, for \( m = 1, \ldots, k + 1 \).

We shall prove that the minimum value of the EBIL function is attained at an extreme point of the set defined by the constraints \( \hat{R}_{PRAM}(m) \leq \frac{T_m}{d} \), for \( m = 1, \ldots, k + 1 \). To prove this result, we shall show that:

1. The EBIL function (6.3.3) is monotonic decreasing.
2. Monotonic functions are quasi-convex and quasi-concave.
3. The set defined by the constraints in (6.3.2) is compact and convex.
4. The EBIL function is continuous and hence attains a minimum in the set defined by (6.3.2).

Hence, by applying a theorem of Kelner and Nikolova (2007), the minimum of EBIL is attained at an extreme point of the set defined by (6.3.2).

6.5.3 Monotonicity of Entropy-Based Information Loss Function

In this subsection, we state and prove that EBIL is monotonic decreasing in each \( p_i, i = 1, \ldots, k + 1 \), for the class of PRAM matrices (6.5.1).
Proposition 6.5.2. The measure $EBIL$ in (6.3.3) is monotonic decreasing in each $p_i$, $i = 1, ..., k + 1$, for the class of PRAM matrices (6.5.1).

Proof. Consider $NEBIL = -EBIL$, i.e.

$$NEBIL(P) = \sum_{l=1}^{k+1} \sum_{m=1}^{k+1} p_{ml} T_m \log \left( \frac{p_{ml} T_m}{\sum_{j=1}^{k+1} p_{jl} T_j} \right). \quad (6.5.5)$$

Using the expression for $p_{ij}$ as defined in (6.5.1), $NEBIL$ (6.5.5) becomes

$$\begin{align*}
\sum_{i=1}^{k+1} p_i T_i \log \left( \frac{p_i T_i}{p_i T_i + k^{-1} \sum_{j=1}^{k+1} (1 - p_j) T_j} \right) \\
+ k^{-1} \sum_{i=1}^{k+1} \sum_{h=1, h \neq i}^{k+1} (1 - p_h) T_h \log \left( \frac{k^{-1}(1 - p_h) T_h}{p_i T_i + k^{-1} \sum_{j=1, j \neq i}^{k+1} (1 - p_j) T_j} \right). \quad (6.5.6)
\end{align*}$$

Therefore, the derivative of NEBIL with respect to $p_i$ is

$$\frac{\partial NEBIL}{\partial p_i} = T_i \log \left( \frac{kp_i T_i}{kp_i T_i + \sum_{j=1}^{k+1} (1 - p_j) T_j} \right)$$

$$- k^{-1} T_i \sum_{h=1, h \neq i}^{k+1} \log \left( \frac{(1 - p_i) T_i}{(1 - p_i) T_i + kp_h T_h + \sum_{m=1, m \neq h, m \neq i}^{k+1} (1 - p_m) T_m} \right). \quad (6.5.7)$$
For more technical details on derivation in (6.5.7), see Appendix C.1. Let \( \rho_{i,1} \equiv p_i \) and \( \rho_{i,2} \equiv k^{-1}(1 - p_i) \). Hence, expression (6.5.7) becomes

\[
T_i \log \left( \frac{k \rho_{i,1} T_i}{k \rho_{i,1} T_i + \sum_{j=1; j \neq i}^{k+1} k \rho_{j,2} T_j} \right) \\
- k^{-1} T_i \sum_{h=1; h \neq i}^{k+1} \log \left( \frac{k \rho_{i,2} T_i}{k \rho_{i,2} T_i + k \rho_{h,1} T_h + \sum_{m=1; m \neq h, m \neq i}^{k+1} k \rho_{m,2} T_m} \right) .
\]

(6.5.8)

Combining the log terms in expression (6.5.8), we obtain

\[
T_i \log \left\{ \frac{\rho_{i,1} T_i}{\rho_{i,1} T_i + \sum_{j=1; j \neq i}^{k+1} \rho_{j,2} T_j} \left[ \prod_{h=1; h \neq i}^{k+1} \left( \rho_{i,2} T_i + \rho_{h,1} T_h + \sum_{m=1; m \neq h, m \neq i}^{k+1} \rho_{m,2} T_m \right) \right]^{1/k} \right\}
\]

(6.5.9)

If expression (6.5.8) is greater than 0, then NEBIL is strictly increasing. This is equivalent to having the terms inside the log in expression (6.5.9) to being than 1 which, in turn, is equivalent to

\[
(\rho_{i,1} T_i)^{k+1} \prod_{h=1; h \neq i}^{k+1} \left( \rho_{i,2} T_i + \rho_{h,1} T_h + \sum_{m=1; m \neq h, m \neq i}^{k+1} \rho_{m,2} T_m \right) > (\rho_{i,2} T_i)^{k+1} \prod_{j=1; j \neq i}^{k+1} \rho_{j,2} T_j .
\]

(6.5.10)
for $i = 1, \ldots, k + 1$. We proceed to show that (6.5.10) is true. Let

$$A_h \equiv \rho_{h,1} T_h + \sum_{m=1, m \neq h, m \neq i}^{k+1} \rho_{m,2} T_m, \forall h \neq i,$$

and

$$B_i \equiv \sum_{j=1, j \neq i}^{k+1} \rho_{j,2} T_j.$$

Since $\rho_{h,1} > \rho_{h,2}$, it follows that, for all $h \neq i$,

$$A_h = \rho_{h,1} T_h + \sum_{j=1, j \neq i}^{k+1} \rho_{j,2} T_j > \rho_{h,2} T_h + \sum_{j=1, j \neq i}^{k+1} \rho_{j,2} T_j = B_i$$

and

$$(\rho_{i,1} T_i)^k \prod_{h=1, h \neq i}^{k+1} (\rho_{i,2} T_i + A_h) > (\rho_{i,1} T_i)^k (\rho_{i,2} T_i + B_i)^k. \quad (6.5.11)$$

Note that the left-hand sides of inequalities (6.5.10) and (6.5.11) are identical.

To complete the proof, we next show that the right-hand side of (6.5.11) is greater than the right-hand side of (6.5.10). Since $\rho_{i,1} > \rho_{i,2}$, it follows that

$$\rho_{i,1} T_i B_i + \rho_{i,1} \rho_{i,2} T_i^2 > \rho_{i,2} T_i B_i + \rho_{i,1} \rho_{i,2} T_i^2$$

and after simplifying and raising both sides to the power $k$,

$$(\rho_{i,1} T_i)^k (\rho_{i,2} T_i + B_i)^k > (\rho_{i,2} T_i)^k (\rho_{i,1} T_i + B_i)^k.$$
Hence, we have shown that (6.5.10) holds, i.e.,

\[(\rho_i, T_i)^k \prod_{h=1, h\neq i}^{k+1} (\rho_i, 2T_i + A_h) > (\rho_i, T_i)^k (\rho_i, 2T_i + B_i)^k > (\rho_i, 2T_i)^k (\rho_i, 1T_i + B_i)^k\]

This implies that \(\frac{\partial NEBIL}{\partial p_i}\) is positive, and that EBIL is monotonic decreasing in each \(p_i\), for \(i = 1, ..., k + 1\). \(\square\)

### 6.5.4 Quasi-concavity of Monotonic Functions

In this subsection, we show that monotonic functions are *quasi-convex* and *quasi-concave*. We proceed with some definitions.

Let \(S\) be a convex subset of \(\mathbb{R}^n\). A function \(f : S \rightarrow \mathbb{R}\) is *quasi-convex* on \(S\) if for all \(x, y \in S\) and all \(\lambda \in [0, 1]\),

\[f(\lambda x + (1 - \lambda)y) \leq \max(f(x), f(y)).\]

A function \(f : S \rightarrow \mathbb{R}\) is *quasi-concave* on \(S\) if for all \(x, y \in S\) and all \(\lambda \in [0, 1]\),

\[f(\lambda x + (1 - \lambda)y) \geq \min(f(x), f(y)).\]

**Proposition 6.5.3.** A monotonic function is both quasi-convex and quasi-concave.

**Proof.** Without loss of generality, let \(f\) be a monotonic increasing function and let \(x \leq y\), where \(x = (x_1, \cdots, x_n)'\), \(y = (y_1, \cdots, y_n)'\) and \(x_i \leq y_i\) for \(i = 1, \cdots, n\). Then,

\[f(\lambda x + (1 - \lambda)y) \leq f(\lambda y + (1 - \lambda)y) = f(y) = \max(f(x), f(y))\]
and
\[
f(\lambda x + (1 - \lambda)y) \geq f(\lambda x + (1 - \lambda)x) = f(x) = \min(f(x), f(y)).
\]

Therefore, any monotonic function is both quasi-convex and quasi-concave.

Proposition 6.5.3 implies that since $EBIL$ (6.3.3) is monotonic decreasing in each $p_i$ for $i = 1, \cdots, k + 1$, $EBIL$ is both quasi-convex and quasi-concave.

6.5.5 Compactness and Convexity of the Region for Safe Records

We state and prove the following proposition, that the region defined by the inequalities (6.3.2) is a compact convex set.

**Proposition 6.5.4.** The region $C$, defined by the inequalities given in (6.3.2), is a compact convex set.

**Proof.** The region, $C$, for safe records is defined by the inequalities given in (6.3.2). Using (3.3.1), it follows from (6.3.2) that $C$ is given by

\[
\frac{p_{mm}}{\sum_{l=1}^{k+1} p_{lm} T_l} \leq \frac{1}{d'}
\]

for $m = 1, \ldots, k + 1$, which after cross-multiplying, becomes

\[
\sum_{l=1}^{k+1} p_{lm} T_l - d p_{mm} \geq 0,
\]

(6.5.12)
for $m = 1, \ldots, k + 1$. The inequalities from (6.5.12) take the form

$$\alpha' p_m + \beta_m \geq 0,$$

for $m = 1, \ldots, k + 1$, where $\alpha = (T_1, \ldots, T_{k+1})'$, $p_m = (p_{1m}, \ldots, p_{(k+1)m})'$ and $\beta_m = -dp_{mm}$. With this notation, the region for safe records is

$$C = \left\{ P : \alpha' p_m + \beta_m \geq 0, m = 1, \ldots, k + 1 \right\}.$$

(6.5.13)

Since $0 \leq p_{ij} \leq 1$ for $i, j = 1, \cdots, k + 1$, the set $C$ is bounded. The set $C$ is a closed set, by definition. Hence, by the Heine-Borel theorem, the set $C$ is compact.

To show that $C$ is convex, we need to show that if $u, v \in C$ then $\lambda u + (1 - \lambda) v \in C, \lambda \in [0, 1]$. Suppose $u, v \in C$, then we have

$$\alpha' u + \beta_m \geq 0$$

and

$$\alpha' v + \beta_m \geq 0.$$

Then,

$$\alpha' [\lambda u + (1 - \lambda) v] + \beta_m = \alpha' [\lambda u + (1 - \lambda) v] + \lambda \beta_m + (1 - \lambda) \beta_m$$

$$= \lambda [\alpha' u + \beta_m] + (1 - \lambda) [\alpha' v + \beta_m] \quad (6.5.14)$$

Since $u, v \in C$, then expression (6.5.14) is nonnegative; hence $C$ is a convex set. Thus, $C$ is a compact convex set.
6.5.6 Continuity of Entropy-Based Information Loss Function

The $EBIL$ function is continuous in $p_i$, for $i = 1, \cdots, k + 1$. From (6.5.6), $EBIL$ can be written as

$$- \sum_{i=1}^{k+1} p_i T_i \log \left( \frac{p_i T_i}{p_i T_i + k^{-1} \sum_{j=1}^{k+1} (1 - p_j) T_j} \right)$$

Recall that $\frac{1}{2} \leq p_i \leq 1$ for all $i = 1, \cdots, k + 1$. It is trivial to show that $EBIL$ is continuous in $\frac{1}{2} \leq p_i < 1$, for all $i = 1, \cdots, k + 1$. When $p_i = 1$ for any $i = 1, \cdots, k + 1$, some of the terms in $EBIL$ take the form $0 \log 0$. Note that $\lim_{p \to 0^+} p \log p = 0$, so $EBIL$ is continuous in $p_i$ for all $i = 1, \cdots, k + 1$ on the interval $[\frac{1}{2}, 1]$.

Since $EBIL$ is continuous, it attains a maximum and a minimum in the compact region $C$ (6.5.13).

6.5.7 Minimization of Monotonic Functions Over a Compact Convex Set

Recall we are minimizing $EBIL(P)$ (6.3.3), with entries of $P$ given by (6.5.1), over a set $C$ given by (6.5.13). We state the following corollary.

**Corollary 6.5.5.** The minimum of $EBIL(P)$ occurs at an extreme point of $C$. 
Before we prove Corollary 6.5.5, we define the following: hyperplane, closed halfspace, extreme point, relative interior, separation, and proper separation. These definitions can be found in Rockafellar (1972) and Bertsekas et al. (2003).

A hyperplane $H$ in $\mathbb{R}^n$ is a set that takes the form $H = \{ x : a^t x = b \}$, where $a$ is a nonzero vector in $\mathbb{R}^n$ and $b$ is a scalar.

The sets $\{ x : a^t x \geq b \}$ and $\{ x : a^t x \leq b \}$ are the closed halfspaces associated with the hyperplane $H$.

A point $x \in C$ is an extreme point of $C$ if and only if there is no way to express $x$ as a convex combination $\lambda y + (1 - \lambda)z$ such that $y, z \in C$ and $\lambda \in (0, 1)$, except by taking $y = z = x$.

A point $x$ is in the relative interior of $C$ if $x \in C$ and there exists an open sphere $S$ centered at $x$ such that $S \subset C$.

A hyperplane $H$ separates $C_1 \subset \mathbb{R}^n$ and $C_2 \in \mathbb{R}^n$ if each lies in a different closed halfspace associated with $H$.

A hyperplane $H$ properly separates $C_1 \subset \mathbb{R}^n$ and $C_2 \in \mathbb{R}^n$ if it separates $C_1$ and $C_2$ and $H$ does not fully contain both $C_1$ and $C_2$.

To prove Corollary 6.5.5, we use the following theorems and lemmas. The following theorem is from Rockafellar (1972, p. 97).

**Theorem 6.5.6.** (Rockafellar, 1972) Let $C_1$ and $C_2$ be non-empty convex sets in $\mathbb{R}^n$. In order that there exist a hyperplane separating $C_1$ and $C_2$ properly, it is necessary and sufficient that the relative interiors of $C_1$ and $C_2$ are disjoint.

The following lemma and its proof are adapted from Bertsekas et al. (2003).

**Lemma 6.5.7.** (Bertsekas et al., 2003) Let $C$ be a nonempty convex subset of $\mathbb{R}^n$. If a hyperplane $H$ contains $C$ in one of its closed halfspaces, then every extreme point of $C \cap H$ is also an extreme point of $C$. 
Proof. (Bertsekas et al., 2003) Let $x$ be an extreme point of $C \cap H$, but not an extreme point of $C$. Then $x = \lambda y + (1 - \lambda)z$ for some $\lambda \in (0, 1)$ and $y, z \in C$, $y \neq x$, $z \neq x$.

Since $x \in H$, the closed halfspace containing $C$ is of the form $\{w : a'w \geq a'x\}$ where $a$ is a nonzero vector. The hyperplane $H$ takes the form $\{w : a'w = a'x\}$.

Thus $a'y \geq a'x$ and $a'z \geq a'x$. Since $x = \lambda y + (1 - \lambda)z$, we have $a'y = a'x$ and $a'z = a'x$. This implies $y = x$ and $z = x$, which is a contradiction. Hence, $x$ is also an extreme point of $C$.

The following lemma and its proof are adapted from Nikolova et al. (2006).

**Lemma 6.5.8.** (Nikolova et al., 2006) Let $C$ be a nonempty convex subset of $\mathbb{R}^n$. A quasi-concave function $f : C \to \mathbb{R}$ that attains a minimum in the relative interior of $C$, attains the minimum at an extreme point of $C$.

**Proof.** (Nikolova et al., 2006) Suppose $f$ attains a minimum at $x^*$ in the relative interior of $C$. Since $C$ is compact, by Caratheodory’s theorem, $x^*$ is a convex combination of finitely many extreme points $z_i$ of $C$. If $f(z_i) = f(x^*)$ for at least one $z_i$, then $f$ attains a minimum at the extreme point $z_i$.

Otherwise, $f(z_i) > f(x^*)$ for all $i$ so $\min_i f(z_i) > f(x^*)$. Consider the set $U = \{x : f(x) \geq \min_i f(z_i)\}$. The set $U$ contains $z_i$ and is convex, therefore it contains the convex hull of $z_i$. Since $x^*$ is a convex combination of finitely many $z_i$’s, $x^*$ is contained in the convex hull of $z_i$, which is a subset of $U$. Hence, $f(x^*) \geq \min_i f(z_i)$ which is a contradiction. Therefore, $f(z_i) = f(x^*)$ for at least one extreme point.

The following theorem is due to Kelner and Nikolova (2007). The proof is adapted from Nikolova et al. (2006) and Bertsekas et al. (2003), and uses theorem 6.5.6, lemma 6.5.7 and lemma 6.5.8.
Theorem 6.5.9. (Kelner and Nikolova, 2007) Let \( C \subset \mathbb{R}^n \) be a compact convex set. A quasi-concave function \( f : C \to \mathbb{R} \) that attains a minimum over \( C \), attains the minimum at some extreme point of \( C \).

Proof. (Nikolova et al. (2006) and Bertsekas et al. (2003)) Let \( x^* \) denote the minimizer of \( f \) over \( C \). If \( x^* \) belongs to the relative interior of \( C \), \( x^* \) occurs at an extreme point of \( C \), by lemma 6.5.8.

Otherwise, \( x^* \) does not belong to the relative interior of \( C \). By Theorem 6.5.6, there exists a hyperplane \( H_1 \) which properly separates \( x^* \) and \( C \). Since \( x^* \in C \), \( x^* \in H_1 \). Since \( H_1 \) properly separates \( x^* \) and \( C \), \( C \not
\subset H_1 \). So, \( \text{dim}(C \cap H_1) \) is less than \( \text{dim}(C) \).

If \( x^* \) belongs in the relative interior of \( C \cap H_1 \), \( x^* \) occurs at an extreme point of \( C \cap H_1 \), by lemma 6.5.8. Using lemma 6.5.7, since \( C \) is contained in one of the closed halfspaces of \( H_1 \), this extreme point of \( C \cap H_1 \) is also an extreme point of \( C \).

If \( x^* \) does not belong to the relative interior of \( C \cap H_1 \), there exists a hyperplane \( H_2 \) which properly separates \( x^* \) and \( C \cap H_1 \). Since \( x^* \in C \cap H_1 \), \( x^* \in H_2 \). Since \( H_2 \) properly separates \( x^* \) and \( C \cap H_1 \), \( C \cap H_1 \not
\subset H_2 \). So, \( \text{dim}(C \cap H_1 \cap H_2) \) is less than \( \text{dim}(C \cap H_1) \).

With each new hyperplane, the dimension of the intersection of \( C \) with the hyperplanes is reduced, and this is repeated at most \( n \) times, until we have \( x^* \) in the relative interior of \( C \cap H_1 \cap \cdots \cap H_k \), for \( k \leq n \). By applying lemma 6.5.8, \( x^* \) occurs at an extreme point of \( C \cap H_1 \cap \cdots \cap H_k \), and by lemma 6.5.7, this is also an extreme point of \( C \). \( \Box \)

Thus, from Theorem 6.5.9, we obtain Corollary 6.5.5
6.5.8 The Number of Extreme Points

As in Section 6.5.5, it follows from (6.3.1) that the inequalities (6.3.2) take the form

\[ \alpha' p_i + \beta_i \geq 0, \quad i = 1, \ldots, k + 1, \]

where \( \alpha = (k_1, \ldots, k_{k+1})' \), \( p_i = (p_{ii}, \ldots, p_{(k+1)i})' \) and \( \beta_i = -dp_{ii} \). Since we consider PRAM matrices that are diagonally dominant, there also exist a class of inequalities of the form \( \frac{1}{2} \leq p_i \leq 1 \) for \( i = 1, \ldots, k + 1 \). Extreme points can be found by taking the intersection of \( k + 1 \) of the inequalities:

\[ (I_1) \text{ All } k + 1 \text{ of the inequalities } \alpha' p_i + \beta_i \geq 0; \text{ or} \]
\[ (I_2) \text{ Any } k \text{ of the inequalities } \alpha' p_i + \beta_i \geq 0, \text{ with any one of } \frac{1}{2} \leq p_i \leq 1; \text{ or} \]
\[ (I_3) \text{ Any } k - 1 \text{ of the inequalities } \alpha' p_i + \beta_i \geq 0, \text{ with any two of } \frac{1}{2} \leq p_i \leq 1; \text{ or} \]
\[ \vdots \]
\[ (I_{k+2}) \text{ All } k + 1 \text{ of } \frac{1}{2} \leq x_i \leq 1 \]

This produces:

\[ (E_1) \binom{k+1}{k+1}\binom{k+1}{0} 2^0 \equiv 1 \text{ extreme point}; \]
\[ (E_2) \binom{k+1}{k}\binom{k+1}{1} 2^1 \text{ extreme points}; \]
\[ (E_3) \binom{k+1}{k-1}\binom{k+1}{2} 2^2 \text{ extreme points}; \]
\[ \vdots \]
\[ (E_{k+2}) \binom{k+1}{0}\binom{k+1}{k+1} 2^{k+1} \text{ extreme points}. \]
Therefore, the total number of extreme points is

\[
\sum_{i=0}^{k+1} \binom{k+1}{i}^2 = \sum_{i=0}^{k+1} \left[ \frac{(k+1)k(k-1)...(k-i+2)}{i!} \right]^2.
\] (6.5.15)

As \( k \to \infty \),

\[(k+1)k(k-1)...(k-i+2) \approx k^i,
\]

so expression (6.5.15) is asymptotic to

\[
\sum_{i=0}^{\infty} \frac{(k+1)^{2i}}{(i!)^2} = \sum_{i=0}^{\infty} \frac{(k+1)^{2i}}{(i!)^2}.
\]

The modified Bessel function of order 0 is defined as

\[I_0(z) = \sum_{i=0}^{\infty} \frac{(\frac{1}{4}z^2)^i}{(i!)^2},\]

\( z \in \mathbb{C} \). Taking \( z = 2\sqrt{2}(k+1) \), we see that (6.5.15), the number of extreme points, is asymptotic to \( I_0(2\sqrt{2}(k+1)) \), as \( k \to \infty \). From the Digital Library of Mathematical Functions (2013),

\[I_0(z) \approx (2\pi z)^{-1/2}e^z,
\]

as \( z \to \infty \); therefore

\[I_0(2\sqrt{2}(k+1)) \approx (4\pi \sqrt{2}(k+1))^{-1/2}e^{2\sqrt{2}(k+1)} \approx c k^{-1/2}e^{2\sqrt{2}k}
\]

where \( c \) is a constant. The implication here is that ideally, the dimension of the variable that is subject to PRAM should not be too large.
6.5.9 Algorithm

From Corollary 6.5.5, the minimum of the measure EBIL (6.3.3) occurs at an extreme point of the region $C$, as defined in (6.5.13), and the planes $\frac{1}{2} \leq p_i \leq 1$ for $i, ..., k + 1$.

We now describe an algorithm for obtaining the entries $p_i$ that result in an optimal PRAM matrix:

1. For a given $T = (T_1, ..., T_{k+1})$ and given $d$, verify whether or not Proposition 6.5.1 is satisfied.
2. Locate all extreme points; the number of extreme points is given by (6.5.15).
3. Retain all extreme points that are diagonally dominant, i.e., that satisfy $\frac{1}{2} \leq p_i \leq 1$, for $i = 1, ..., k + 1$.
4. Evaluate (6.3.3) at the diagonally dominant extreme points.
5. Find the diagonally dominant extreme point which gives the minimum value of EBIL over all such points.

The R code is in Appendix C.3.

6.5.10 A Numerical Experiment

We demonstrate our proposed methodology with a numerical experiment. We set $d = 100$, $T = (T_1, 80, 90)$, where $T_1$ varies from 20 to 90 in increments of 10. Using our methodology and the built-in constrOptim routine in the statistical package R (see R Documentation), we seek to find optimal $p_i$ such that the measure EBIL (6.3.3) is minimized subject to the inequalities (6.3.2). The results are displayed in Table 6.1. Our proposed methodology obtained $p_i$’s that lead to smaller values of the
measure \( EBIL \) (6.3.3) as compared to the built-in \( \text{constrOptim} \) routine in R, in all cases. In two cases, \((T_1 = 80, 90)\), our methodology obtained two optimal points.

### Table 6.1. Results comparing our methodology with built-in \( \text{constrOptim} \) routine in R.

<table>
<thead>
<tr>
<th>( T_1 )</th>
<th>Proposed Methodology</th>
<th>( \text{constrOptim} ) routine in R</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>(0.5, 1, 1)</td>
<td>38.6043</td>
</tr>
<tr>
<td>30</td>
<td>(0.5, 1, 1)</td>
<td>52.0355</td>
</tr>
<tr>
<td>40</td>
<td>(1, 0.75, 1)</td>
<td>57.5284</td>
</tr>
<tr>
<td>50</td>
<td>(1, 0.75, 1)</td>
<td>59.5420</td>
</tr>
<tr>
<td>60</td>
<td>(1, 0.75, 1)</td>
<td>61.2164</td>
</tr>
<tr>
<td>70</td>
<td>(1, 0.75, 1)</td>
<td>62.6499</td>
</tr>
<tr>
<td>80</td>
<td>(1, 0.75, 1)</td>
<td>63.9032</td>
</tr>
<tr>
<td></td>
<td>(0.75, 1, 1)</td>
<td>63.9032</td>
</tr>
<tr>
<td>90</td>
<td>(1, 1, 0.7778)</td>
<td>63.9032</td>
</tr>
<tr>
<td></td>
<td>(0.7778, 1, 1)</td>
<td>63.9032</td>
</tr>
</tbody>
</table>

### 6.6 A New Class of PRAM Matrices

Thus far, we have established results for a class of PRAM matrices, as defined in (6.5.1). We proceed to see if similar results can be established for another class of PRAM matrices, which we define next.

We consider a class of \((k + 1) \times (k + 1)\) PRAM matrices, \( P = (p_{ij}) \), where

\[
p_{ij} = \begin{cases} 
p_i, & i = j \\
k^{-1}(1 - p_i), & i \neq j
\end{cases}
\]  

(6.6.1)

for \( i, j = 2, \ldots, k + 1 \) and allow the entries in the first row to be free to vary. Hence,
$P$ takes the form

$$
P = \begin{pmatrix}
p_{1,1} & p_{1,2} & \cdots & p_{1,k} & 1 - \sum_{i=1}^{k} p_{1,i} \\
k^{-1}(1-p_2) & p_2 & k^{-1}(1-p_2) & \cdots & k^{-1}(1-p_2) \\
\vdots \\
k^{-1}(1-p_{k+1}) & k^{-1}(1-p_{k+1}) & \cdots & k^{-1}(1-p_{k+1}) & p_{k+1}
\end{pmatrix}.
$$

It needs to be noted that in the application of PRAM, it is typical to consider PRAM matrices that are diagonally dominant, i.e. $\frac{1}{2} \leq p_i \leq 1$ for $i = 2, \ldots, k+1$, and $\frac{1}{2} \leq p_{1,1} \leq 1$. We have proven that PRAM matrices as defined in (6.6.1) display similar monotonic properties as PRAM matrices defined in (6.5.1) only under very strict conditions, when all $T_i$ are equal and $k \leq 6$, and when $k \geq 7$, the monotonicity property does not exist all the time. We prove these next.

### 6.6.1 Monotonicity Properties of Entropy-Based Information Loss Function

**Proposition 6.6.1.** When all $T_i$ equal, EBIL is monotonic decreasing in $p_i$, for $2 \leq i \leq k+1$, where $2 \leq k \leq 6$.

**Proof.** Again, consider NEBIL = $-EBIL$, which is given in (6.5.5) as

$$
NEBIL(P) = \sum_{l=1}^{k+1} \sum_{m=1}^{k+1} p_{ml} T_m \log \left( \frac{p_{ml} T_m}{\sum_{j=1}^{k+1} p_{lj} T_j} \right)
$$

Using the expression for $p_{ij}$ as defined in (6.6.1), $NEBIL$ (6.5.5) becomes
\[ NEBIL = p_{1,1} T_1 \log \left( \frac{p_{1,1} T_1}{p_{1,1} T_1 + k^{-1} \sum_{j=2}^{k+1} (1 - p_j) T_j} \right) \]

\[ + k^{-1} \sum_{h=2}^{k+1} (1 - p_h) T_h \log \left( \frac{k^{-1} (1 - p_h) T_h}{p_{1,1} T_1 + k^{-1} \sum_{j=2}^{k+1} (1 - p_j) T_j} \right) \]

\[ + \sum_{i=2}^{k+1} \left[ p_{1,i} T_1 \log \left( \frac{p_{1,i} T_1}{p_{1,i} T_1 + p_i T_i + k^{-1} \sum_{j=2}^{k+1} (1 - p_j) T_j} \right) \right] \]

\[ + p_i T_i \log \left( \frac{p_i T_i}{p_{1,i} T_1 + p_i T_i + k^{-1} \sum_{j=2}^{k+1} (1 - p_j) T_j} \right) \]

\[ + k^{-1} \sum_{i=2}^{k+1} \sum_{h=2}^{k+1} \sum_{h \neq i} (1 - p_h) T_h \log \left( \frac{k^{-1} (1 - p_h) T_h}{p_{1,i} T_1 + p_i T_i + k^{-1} \sum_{j=2}^{k+1} (1 - p_j) T_j} \right). \]

Therefore, the derivative of NEBIL with respect to \( p_i \), for \( 2 \leq i \leq k + 1 \), is

\[ \frac{\partial NEBIL}{\partial p_i} = T_i \log \left( \frac{k p_i T_i}{k p_{1,i} T_1 + k p_i T_i + k^{-1} \sum_{j=2}^{k+1} (1 - p_j) T_j} \right) \]

\[ -k^{-1} T_i \sum_{h=1}^{k+1} \sum_{h \neq i} \log \left( \frac{(1 - p_i) T_i}{k p_{1,h} T_1 + \sum_{j=2}^{k+1} (1 - p_j) T_j} \right). \]
\[-k^{-1} T_i \log \left( \frac{(1 - p_i) T_i}{k p_{1,k+1} T_1 + \sum_{j=2}^{k} (1 - p_j) T_j + k p_{k+1} T_{k+1}} \right). \tag{6.6.2} \]

For more technical details on derivation in (6.6.2), see Appendix C.2. Let \( \rho_{i,1} \equiv p_i \) and \( \rho_{i,2} \equiv k^{-1}(1 - p_i) \). Expression (6.6.2) becomes

\[
T_i \log \left( \frac{\rho_{i,1} T_i}{\rho_{1,i} T_1 + \rho_{i,1} T_i + \sum_{j=2 \atop j \neq i}^{k+1} \rho_{j,2} T_j} \right)
\]

\[-k^{-1} T_i \sum_{h=1 \atop h \neq i}^{k+1} \log \left( \frac{\rho_{i,2} T_i}{\rho_{1,h} T_1 + \sum_{j=2}^{k+1} \rho_{j,2} T_j} \right) \]

\[-k^{-1} T_i \log \left( \frac{\rho_{i,2} T_i}{\rho_{1,k+1} T_1 + \sum_{j=2}^{k} \rho_{j,2} T_j + \rho_{k+1,1} T_{k+1}} \right). \tag{6.6.3} \]

Combining the log terms in expression (6.6.3), we obtain

\[
T_i \log \left[ \frac{\rho_{i,1} T_i}{\rho_{1,i} T_1 + \rho_{i,1} T_i + \sum_{j=2 \atop j \neq i}^{k+1} \rho_{j,2} T_j} \right]
\]

\[
\times \left( \left( \rho_{1,k+1} T_1 + \sum_{j=2}^{k} \rho_{j,2} T_j + \rho_{k+1,1} T_{k+1} \right) \prod_{h=1 \atop h \neq i}^{k+1} \left( \rho_{1,h} T_1 + \sum_{j=2}^{k+1} \rho_{j,2} T_j \right) \right)^{1/k} \]

\[
\frac{\rho_{i,2} T_i}{\rho_{1,2} T_1 + \sum_{j=2}^{k} \rho_{j,2} T_j + \rho_{k+1,1} T_{k+1}} \right]. \tag{6.6.4} \]

If expression (6.6.4) is greater than 0 then NEBIL is strictly increasing. This is equivalent to having the terms inside the log in expression (6.6.4) to being greater
than 1, which in turn is equivalent to

\[
\left(\rho_{i,1}T_i\right)^k \left(\rho_{1,k+1}T_1 + \sum_{j=2}^{k} \rho_{j,2}T_j + \rho_{k+1,1}T_{k+1}\right) \prod_{\substack{h=1 \atop h \neq i}}^{k+1} \left(\rho_{1,h}T_1 + \sum_{j=2}^{k+1} \rho_{j,2}T_j\right) > \left(\rho_{i,2}T_i\right)^k \left(\rho_{1,i}T_1 + \sum_{j=2}^{k+1} \rho_{j,2}T_j\right)^k. \tag{6.6.5}
\]

Consider setting the right hand side of expression (6.6.5) to be as large as possible, with \(p_{1i} = p_{1,1} = \frac{1}{2}\), and set all \(T_i\) equal. Then showing expression (6.6.5) holds is equivalent to showing the following holds:

\[
\rho_{i,1}^k \left(\rho_{i,2} + \sum_{\substack{j=2 \atop j \neq i}}^{k} \rho_{j,2} + \rho_{k+1,1}\right) \left(\frac{1}{2} + \rho_{i,2} + \sum_{\substack{j=2 \atop j \neq i}}^{k} \rho_{j,2} + \rho_{k+1,2}\right) \left(\rho_{i,2} + \sum_{\substack{j=2 \atop j \neq i}}^{k} \rho_{j,2} + \rho_{k+1,2}\right)^{k-2} > \rho_{i,2}^k \left(\frac{1}{2} + \rho_{i,1} + \sum_{\substack{j=2 \atop j \neq i}}^{k+1} \rho_{j,2}\right)^k. \tag{6.6.6}
\]

Let \(b_1 \equiv \rho_{i,1}, b_2 \equiv \rho_{i,2}, C_i \equiv \sum_{\substack{j=2 \atop j \neq i}}^{k} \rho_{j,2}, z_1 \equiv \rho_{k+1,1}, z_2 \equiv \rho_{k+1,2}\). The inequality (6.6.6) becomes

\[
b_1^k \left(\frac{1}{2} + b_2 + C_i + z_2\right)(b_2 + C_i + z_1)(b_2 + C_i + z_2)^{k-2} > b_2^k \left(\frac{1}{2} + b_1 + C_i + z_2\right)^k, \tag{6.6.7}
\]

for \(i = 1, \ldots, k+1\). Proposition 6.6.1 implies that expression (6.6.7) holds for \(k = 2, \ldots, 6\).

When \(k = 2\), \(C_i = 0\). Hence expression (6.6.7) becomes

\[
b_1^2 \left(\frac{1}{2} + b_2 + z_2\right)(b_2 + z_1) > b_2^2 \left(\frac{1}{2} + b_1 + z_2\right)^2. \tag{6.6.8}
\]
Expanding both sides of (6.6.8), we obtain

\[ b_1^2 \left( \frac{1}{2} b_2 + \frac{1}{2} z_1 + b_2^2 + b_2 z_1 + b_2 z_2 + z_1 z_2 \right) > b_2^2 \left( \frac{1}{4} + b_1 + b_1^2 + z_2 + b_1 z_2 + z_2^2 \right). \]  (6.6.9)

Using the inequalities \( \frac{1}{2} \leq b_1, z_1 \leq 1 \), \( 0 \leq b_2, z_2 \leq \frac{1}{4} \) and comparing the terms in the left hand side of expression (6.6.9) with the terms in the right hand side of expression (6.6.9), expression (6.6.8) holds, and hence expression (6.6.7) holds for \( k = 2 \).

For \( k = 3 \), expression (6.6.7) becomes

\[ b_1^3 \left( \frac{1}{2} + b_2 + C_i + z_2 \right) \left( b_2 + C_i + z_1 \right) \left( b_2 + C_i + z_2 \right) > b_2^3 \left( \frac{1}{2} + b_1 + C_i + z_2 \right)^3. \]  (6.6.10)

Expanding both sides of expression (6.6.10), using the inequalities \( \frac{1}{2} \leq b_1, z_1 \leq 1 \), \( 0 \leq b_2, z_2 \leq \frac{1}{6} \) and comparing coefficients of \( C_i^l \) for \( l = 0, \ldots, 3 \), we see again that the terms in the left hand side of expression (6.6.10) is greater than the terms in the right hand side of (6.6.10). Indeed, the same method can be used for \( k = 4, 5 \) also, by expanding expression (6.6.7) and using the inequalities \( \frac{1}{2} \leq b_1, z_1 \leq 1 \), \( 0 \leq b_2, z_2 \leq \frac{1}{2k} \).

The same method holds for \( k = 6 \) for coefficients of \( C_i^l \), \( l = 1, \ldots, 6 \), but not for \( l = 0 \). For \( k = 6 \) and \( C_i = 0 \), expression (6.6.7) becomes

\[ b_1^6 \left( \frac{1}{2} + b_2 + z_2 \right) \left( b_2 + z_1 \right) \left( b_2 + z_2 \right)^4 > b_2^6 \left( \frac{1}{2} + b_1 + z_2 \right)^6. \]  (6.6.11)

Let \( b_2 \equiv \frac{1-b_1}{6} \) and \( z_2 \equiv \frac{1-z_1}{6} \), expression (6.6.11) becomes

\[ b_1^6 \left( \frac{5}{6} - \frac{b_1}{2} - \frac{z_1}{2} \right) \left( \frac{1}{6} - \frac{b_1}{6} + z_1 \right) \left( \frac{1}{6} - \frac{b_1}{6} - \frac{z_1}{6} \right)^4 > \left( \frac{1-b_1}{6} \right)^6 \left( \frac{2}{3} + b_1 + \frac{z_1}{6} \right)^6. \]  (6.6.12)

Showing expression (6.6.12) is true is equivalent to showing the log of the left-hand side of expression (6.6.12) is greater than the log of the right hand side of expression.
Define the function $f(b_1, z_1)$ as

$$f(b_1, z_1) = \log \left[b_1^{(5/6)} - \frac{b_1}{6} - \frac{z_1}{6}(\frac{1}{6} - \frac{b_1}{6} + z_1)(\frac{1}{3} - \frac{b_1}{6} - \frac{z_1}{6})^4\right] - \log \left[\left(\frac{1-b_1}{6}\right)^6(\frac{2}{3} + b_1 + \frac{z_1}{6})^6\right].$$

(6.6.13)

Showing $f(b_1, z_1)$ (6.6.13) is greater than 0 will prove expression (6.6.7) holds for $k = 6$. Taking derivative of $f(b_1, z_1)$ with respect to $z_1$, we obtain

$$\frac{\partial f}{\partial z_1} = \frac{6}{4 + 6b_1 - z_1} + \frac{1}{-5 + b_1 + z_1} + \frac{4}{-2 + b_1 + z_1} + \frac{6}{1 - b_1 + 6z_1}.$$ 

Since $\frac{1}{2} \leq b_1, z_1 \leq 1$, we have the following inequalities

$$\frac{12}{19} < \frac{6}{4 + 6b_1 - z_1} < 1,$$

$$-\frac{1}{3} < \frac{1}{-5 + b_1 + z_1} < -\frac{1}{4},$$

$$-\infty < \frac{4}{-2 + b_1 + z_1} < -4,$$

$$\frac{12}{13} < \frac{6}{1 - b_1 + 6z_1} < 2.$$ 

Hence the derivative of $f(b_1, z_1)$ with respect to $z_1$ is less than 0. Thus, to show that $f(b_1, z_1)$ is greater than 0, all we are left with is to show $f(b_1, 1) > 0$. Let
\[ g(b_1) = f(b_1, 1), \] so

\[ g(b_1) = 6 \log b_1 + \log \left( \frac{2}{3} - \frac{b_1}{6} \right) + \log \left( \frac{7}{6} - \frac{b_1}{6} \right) + 4 \log \left( \frac{1}{6} - \frac{b_1}{6} \right) - 6 \log \left( \frac{1 - b_1}{6} \right) - 6 \log \left( \frac{1}{2} + b_1 \right). \]

(6.6.14)

Taking the derivative of \( g(b_1) \) (6.6.14) with respect to \( b_1 \), we obtain

\[
\frac{\partial g(b_1)}{\partial b_1} = \frac{1}{b_1 - 7} + \frac{1}{b_1 - 4} + \frac{2}{1 - b_1} + \frac{6}{b_1} - \frac{12}{1 + 2b_1}. 
\]

Since \( \frac{1}{2} \leq b_1, z_1 \leq 1 \), we have the following inequalities

\[ -\frac{1}{6} < \frac{1}{b_1 - 7} < -\frac{2}{13}, \]

\[ -\frac{1}{3} < \frac{1}{b_1 - 4} < -\frac{2}{7}, \]

\[ 4 < \frac{2}{1 - b_1} < \infty, \]

\[ 6 < \frac{6}{b_1} < 12, \]

\[ -6 < -\frac{12}{1 + 2b_1} < -4. \]

Thus the derivative of \( g(b_1) \) is greater than 0. We are left to show \( g(\frac{1}{2}) \) is greater than 0.

\[
g(\frac{1}{2}) = \left( \frac{1}{2} \right)^6 \left( \frac{2}{3} - \frac{1}{12} \right) \left( \frac{7}{6} - \frac{1}{12} \right) \left( \frac{1}{6} - \frac{1}{12} \right)^4 - \left( \frac{1}{12} \right)^6
\]

\[ = \left( \frac{1}{2} \right)^6 \left( \frac{7}{12} \right) \left( \frac{13}{12} \right) \left( \frac{1}{6} \right)^4 - \left( \frac{1}{12} \right)^6
\]

\[ = \left( \frac{1}{12} \right)^6 \left( \frac{7 \times 13}{4} - 1 \right) > 0. \]

Thus \( g(b_1) > 0 \) for \( b_1 \in [\frac{1}{2}, 1] \), which in turn implies \( f(b_1, 1) > 0 \), which in turn
implies $f(b_1, z_1) > 0$ for $z_1 \in [\frac{1}{2}, 1]$, since we showed $f(b_1, z_1)$ is decreasing in $z_1$.

We have shown that expression (6.6.7) holds for $k = 2, \ldots, 6$. This means that the measure $EBIL$ is monotonic decreasing in $p_i$ for $i = 2, \ldots, k + 1$ for $k = 2, \ldots, 6$ and when all $T_i$ are equal, for the class of PRAM matrices as defined in Section 6.6. □

**Proposition 6.6.2.** $EBIL$ is not monotonic decreasing in $p_i$, for $2 \leq i \leq k + 1$, when $k \geq 7$.

To show the measure EBIL (6.3.3) is not monotonic decreasing in $p_i$, for $2 \leq i \leq k + 1$, when $k \geq 7$, we need to show there exists some values of $b_1, C_i$, and $z_1$ such that (6.6.7) is not true. Define the function $h(b_1, C_i, z_1)$ as

$$h(b_1, C_i, z_1) = b_1^k \left( \frac{1}{2} + b_2 + C_i + z_2 \right) (b_2 + C_i + z_1) (b_2 + C_i + z_2)^{k-2} - b_2^k \left( \frac{1}{2} + b_1 + C_i + z_2 \right)^k.$$  \hfill (6.6.15)

In other words, we need to show $h(b_1, C_i, z_1)$ (6.6.15) is less than 0. This is proved by induction on $k$.

**Proof.** Consider $h(\frac{1}{2}, 0, 1)$, i.e.,

$$h(\frac{1}{2}, 0, 1) = 2^{-k} (\frac{1}{2} + \frac{1}{2k}) (\frac{1 + 2k}{2k}) (\frac{1}{2k})^{k-2} - (\frac{1}{2k})^k.$$

We proceed to show $h(\frac{1}{2}, 0, 1)$ is less than 0, i.e.,

$$2^{-k} (\frac{1}{2} + \frac{1}{2k}) (\frac{1 + 2k}{2k}) (\frac{1}{2k})^{k-2} - (\frac{1}{2k})^k < 0$$

or equivalently, we show

$$(k + 1)(2k + 1) - 2^k < 0$$
for \( k \geq 7 \). Setting \( k = n + 6 \), this is also equivalent to showing

\[
(n + 7)(2n + 13) - 2^{n+6} < 0.
\]

for \( n \geq 1 \). Let \( P_n : (n + 7)(2n + 13) - 2^{n+6} < 0 \). Therefore, \( P_1 : (8)(15) - 2^7 < (8)(16) - 2^7 = 2^3 \cdot 2^4 - 2^7 = 0 \). So \( P_1 \) is true. Assume \( P_m \) is true, i.e.,

\[
P_m : (m + 7)(2m + 13) - 2^{m+6} < 0.
\]

Next, we show \( P_{m+1} \) is true, i.e.,

\[
P_{m+1} : (m + 8)(2m + 15) - 2^{m+7} < 0.
\]

We have

\[
2^{m+7} = 2(2)^{m+6} > 2(m + 7)(2m + 13) = 4m^2 + 54m + 182 > 2m^2 + 31m + 120 = (m + 8)(2m + 15)
\]

thus \( P_{m+1} \) is true. Thus \( h(\frac{1}{2}, 0, 1) \) is less than 0, and hence, the measure EBIL (6.3.3) is not monotonic decreasing in \( p_i \) for all values.

It needs to be noted that there exist some values where the measure EBIL (6.3.3)
is monotonic decreasing in $p_i$, for $2 \leq i \leq k + 1$, when $k \geq 7$. For example,

$$h\left(\frac{1}{2}, 0, \frac{1}{2}\right) = 2^{-k} \left(\frac{k + 2}{2k}\right) \left(\frac{k + 1}{2k}\right) k^{2-k} - (2k)^{-k} \left(1 + \frac{1}{2k}\right)^k$$

is non-negative. Showing $h\left(\frac{1}{2}, 0, \frac{1}{2}\right)$ is non-negative is equivalent to showing

$$\left(\frac{k + 2}{k}\right) \left(\frac{k + 1}{2}\right) - (1 + \frac{1}{2k})^k \geq 0. \quad (6.6.16)$$

Consider the function

$$(1 + \frac{x}{n})^n = \sum_{j=0}^{n} \binom{n}{j} \left(\frac{x}{n}\right)^j$$

for $n \in \mathbb{Z}^+$ and $x \geq 0$. Note that for any $0 \leq j \leq n$,

$$\binom{n}{j} \left(\frac{x}{n}\right)^j = \frac{n!}{j!(n-j)!} \left(\frac{x}{n}\right)^j$$

$$= \frac{n(n-1)\ldots(n-j+1)}{j!} \left(\frac{x}{n}\right)^j$$

$$\leq \frac{n(n-1)\ldots(n-j+1)}{j!} \frac{x^j}{n(n-1)\ldots(n-j+1)}$$

$$= \frac{x^j}{j!}.$$ 

since we consider $x \geq 0$. Therefore,

$$\left(1 + \frac{x}{n}\right)^n = \sum_{j=0}^{n} \binom{n}{j} \left(\frac{x}{n}\right)^j$$

$$\leq \sum_{j=0}^{n} \frac{x^j}{j!}$$

$$\leq \sum_{j=0}^{\infty} \frac{x^j}{j!}$$

$$= e^x. \quad (6.6.17)$$
For $x = \frac{1}{2}$ in (6.6.17), we obtain

$$\left(1 + \frac{1}{2k}\right)^k \leq e^{1/2} \approx 1.6487.$$ 

For $k = 1$, $(\frac{k+2}{k})^2 = 3$, and since $(\frac{k+2}{k})^2$ increases in $k$, inequality (6.6.16) holds.

### 6.7 Conclusion and Future Work

One widely-used measure of data utility for PRAM is the entropy-based information loss, $EBIL$ (Domingo-Ferrer and Torra, 2001a), which is a variant of Shannon’s entropy. Minimizing $EBIL$ has been done computationally, which is slow and expensive (Mares and Torra, 2010). The mathematical properties of $EBIL$ have received little attention in the literature, possibly because it consists of many variables which make mathematical derivations complicated. When we first started on deriving mathematical properties of $EBIL$, we had not much of an idea on what to expect. After many mathematical derivations, we were able to prove certain properties.

We have shown that for a certain class of PRAM matrices as defined by (6.5.1), $EBIL$ displays monotonic properties, which implies the minimum of $EBIL$ occurs at an extreme point of the region that satisfies a pre-determined rule for safe files. A key here is that this region is a convex set. This implies that the optimal PRAM matrix can be found at an extreme point of the region. Using these properties, we derived Corollary 6.5.5 and have developed an algorithm (Section 6.5.9) to find the optimal PRAM matrix by: 1) evaluating the value of $EBIL$ at the extreme points of the region that satisfies a pre-determined rule for safe files and, 2) then finding the
minimum of EBIL at these extreme points.

It should be noted that in Section 6.5.4, we proved that EBIL is monotonic, and hence quasi-convex, for the class of PRAM matrices as defined by (6.5.1). We have also proved that EBIL is convex for a $2 \times 2$ PRAM matrix, and are working on the proof of convexity in a general $(k+1) \times (k+1)$ PRAM matrix.

We compared our method of finding the optimal PRAM matrix with the built-in constrOptim routine in the statistical package R (see R Documentation) in Chapter 6.5.10. Our method performed better as it resulted in PRAM matrices that gave lower values of EBIL when compared to the PRAM matrices from constrOptim. This is in spite of the fact that we only used PRAM matrices of the form in (6.5.1) when using our method, whereas no such restriction was placed when using the constrOptim routine in R.

As noted in Section 6.5.8, the number of extreme points that need to be evaluated increases exponentially as the dimension of the categorical variable increases. This potentially causes an issue with regards to computation time for our algorithm, and the algorithm has a limit with regards to the dimension of the variable PRAM is applied to. A potential way to get around this issue is using Dantzig’s simplex method (Dantzig, 1951), which moves from one extreme point to another with lower cost, thus avoiding extreme points that have higher cost. Another manner to make the computations more efficient could be scalability in parallel programming.

In Section 6.6, we loosened some of the restrictions that was placed on the PRAM matrices in Section 6.5. The monotonicity properties for EBIL were only present under a certain set of conditions, when all $T_i$ are equal, and for $k \leq 6$. This suggests that our algorithm proposed in Section 6.5.9 will no longer be feasible for other class of PRAM matrices.

In this Chapter, we viewed finding the optimal PRAM matrix as a constrained
optimization problem. Another way to view the problem may be to obtain a “good” PRAM matrix, one that produces a safe file upon application to the data, with a tolerable amount of information loss, rather than minimizing information loss.
Chapter 7

Conclusions and Future Work

Statistical agencies face the challenge of providing high-quality data products to aid statistical research, while also guarding against the risk of disclosing confidential information on individual respondents. The field of Statistical Disclosure Control (SDC) has been developed to balance the objectives of providing data for valid inference and safeguarding confidential information.

Microdata are sets of records containing detailed information on individual respondents and many Statistical Disclosure Control (SDC) methods have been developed for microdata. One of the SDC methods for categorical variables is the Post Randomization Method (PRAM), which was first introduced by Gouweleeuw et al. (1998). The basic idea underlying PRAM is to misclassify values of the categorical variables, via a known probability mechanism. This mechanism is expressed in the form of a transition matrix, and in the context of PRAM, also called a PRAM matrix.

There are a number of issues which have led to limited use of PRAM for census data. A commonly-discussed issue is the adjusting of estimators of parameters arising statistical analyses. Another issue deals with dependence between variables
that are subject to PRAM. My dissertation research has focused on three primary methodological developments. Another issue is the problem of obtaining an optimal PRAM matrix that produces a safe file which maximizes data utility simultaneously.

In Chapter 4, we develop and implement EM-type algorithms to obtain asymptotically unbiased estimators of parameters in generalized linear models (GLMs), when chosen variables are subject to PRAM. The basic ideas are based on the “EM by method of weights” developed by Ibrahim (1990) for GLMs with covariates missing at random, and on the approach proposed by van den Hout and Kooiman (2006) for linear regression with covariates subject to PRAM. We build on these ideas by developing an EM-type algorithm that obtains unbiased estimates of GLMs when both covariate and response variables are subject to PRAM. Our results show that my EM-type algorithms perform well, in terms of obtaining unbiased parameter estimates, in various settings such as logistic, Poisson, gamma, and exponential regression.

Next, we extend the methodology developed in Chapter 4 by relaxing the assumption of independence of covariates when estimating parameters in GLMs when variables are subject to PRAM, leading to the need for a more complex EM algorithm to take into account the dependence of covariates. This is done by modeling the distribution of the covariates subject to PRAM as a product of univariate conditional distributions. This approach advances the methodology by making it more applicable in practice by allowing dependence between the covariates, and results in more accurate estimates of the regression parameters, especially in smaller sample sizes. The methodology was implemented successfully on the 1993 Current Population Survey.

Since the distribution for the covariates subject to PRAM is specified as a series of univariate conditional distributions (or series of multinomial logistic regressions), the
question of finding the best fit for each of the multinomial logistic regression arises. The effect of PRAM on model selection techniques is a potential avenue for future work.

Lastly, in Chapter 6, we studied a widely-used measure of data utility for PRAM: entropy-based information loss, $EBIL$ (Domingo-Ferrer and Torra, 2001a), which is a variant of Shannon’s entropy. The mathematical properties of $EBIL$ have received little attention in the literature, possibly because it consists of many variables which make mathematical derivations complicated. Minimizing $EBIL$ has been done computationally, which has been noted to be slow and expensive (Mares and Torra, 2010).

We have shown that for a certain class of PRAM matrices as defined by (6.5.1), $EBIL$ displays monotonic properties, which implies the minimum of $EBIL$ occurs at an extreme point of the region that satisfies a pre-determined rule for safe files. A key here is that this region is a convex set. This implies that the optimal PRAM matrix can be found at an extreme point of the region, and hence can be applied to the microdata more efficiently and in a less computationally expensive manner. I have developed a method (Chapter 6.5.9) to find the optimal PRAM matrix by: 1) evaluating the value of $EBIL$ at the extreme points of the region that satisfies a pre-determined rule for safe files and, 2) then finding the minimum of $EBIL$ at these extreme points. We compared our method of finding the optimal PRAM matrix with the built-in constrOptim routine in the statistical package R. Our method performed better as it resulted in PRAM matrices that gave lower values of $EBIL$ when compared to the PRAM matrices from constrOptim.

Future work includes using Dantzig’s simplex method (Dantzig, 1951), which moves from one extreme point to another with lower cost, thus avoiding extreme points that have higher cost. This will save even more computing time, thus
allowing a more efficient manner to derive an optimal PRAM matrix. Another manner to make the computations more efficient could be scalability in parallel programming.

Thus far, we viewed finding the optimal PRAM matrix as a constrained optimization problem. Another way to view the problem may be to obtain a “good” PRAM matrix, one that produces a safe file upon application to the data, with a tolerable amount of information loss, rather than minimizing information loss.
Appendix A

Derivation of Weights for EM Algorithms

A.1 Weights for EM Algorithm I

\[
q_j = P(W = w_j | W^* = w_k, Y, Z, \phi^{(v)}) \\
= \frac{P(Y, W = w_j, W^* = w_k | Z, \phi^{(v)})}{P(Y, W^* = w_k | Z, \phi^{(v)})} \\
= \frac{P(Y|W = w_j, W^* = w_k, Z, \phi^{(v)}) P(W = w_j, W^* = w_k, Z, \phi^{(v)})}{\sum_{l=1}^{J} P(Y, W^* = w_k, W = w_l | Z, \phi^{(v)})} \\
= \frac{P(Y|W = w_j, Z, \phi^{(v)}) P(W^* = w_k|W = w_j) P(W = w_j)}{\sum_{l=1}^{J} P(Y|W = w_l, W^* = w_k, Z, \phi^{(v)}) P(W^* = w_k|W = w_l) P(W = w_l)} \\
= \frac{P(Y|w_j, Z, \phi^{(v)}) p_{Wj k} \pi_j^{(v)}}{\sum_{l=1}^{J} P(Y|w_l, Z, \phi^{(v)}) p_{Wl k} \pi_l^{(v)}}
\]

Note:

\[P(Y|W, W^*, Z, \phi) = P(Y|W, Z, \phi)\] since when \(W\) is known, \(Y\) does not depend on \(W^*\).
\( P(W^*|W, Z, \phi) = P(W^*|W) \) because of PRAM, \( W^* \) only depends on \( W \).

Also assume \( W, Z \) independent.

### A.2 Weights for EM Algorithm II

\[
\begin{align*}
    r_j &= P(Y = j|Y^* = k, X, \beta^{(\nu)}) \\
    &= \frac{P(Y = j, Y^* = k|X, \beta^{(\nu)})}{P(Y^* = k|X, \beta^{(\nu)})} \\
    &= \frac{P(Y = j|X, \beta^{(\nu)}) P(Y^* = k|Y = j)}{\sum_l P(Y^* = k, Y = l|X, \beta^{(\nu)})} \\
    &= \frac{P(Y = j|X, \beta^{(\nu)}) P_{Y_{jk}}}{\sum_l P_{Y_{lk}} P(Y = l|X, \beta^{(\nu)})} \\
\end{align*}
\]

### A.3 Weights for EM Algorithm III

\[
\begin{align*}
    s_{ml} &= P(Y = m, W = w_l|Y^* = k, W^* = w_j, Z, \phi^{(\nu)}) \\
    &= P(Y = m|W = w_l, Y^* = k, W^* = w_j, Z, \phi^{(\nu)}) \cdot P(W = w_l|Y^* = k, W^* = w_j, Z, \phi^{(\nu)}) .
\end{align*}
\]

The first part is

\[
\begin{align*}
    P(Y = m|W = w_l, Y^* = k, W^* = w_j, Z, \phi^{(\nu)}) \\
    &= \frac{P(Y = m, W = w_l, Y^* = k, W^* = w_j|Z, \phi^{(\nu)})}{P(W = w_l, Y^* = k, W^* = w_j|Z, \phi^{(\nu)})} \\
    &= \frac{P(Y^* = k|Y = m) P(Y = m|W = w_l, Z, \phi^{(\nu)}) P(W^* = w_j|W = w_l) P(W = w_l)}{\sum_a P(Y = a, W = w_l, Y^* = k, W^* = w_j|Z, \phi^{(\nu)})} \\
    &= \frac{P_{Y_{mk}} P(Y = m|W = w_l, Z, \phi^{(\nu)}) P(W^* = w_j|W = w_l) P(W = w_l)}{\sum_a P(Y^* = k|Y = a) P(Y = a|W = w_l, Z, \phi^{(\nu)}) P(W^* = w_j|W = w_l) P(W = w_l)}
\end{align*}
\]
\[
= P_{Ymk} P(Y = m| W = w_l, Z, \phi^{(v)}) \\
\sum_a P_{Yak} P(Y = a| W = w_l, Z, \phi^{(v)}) .
\]

The second part is

\[
P(W = w_l| Y^* = k, W^* = w_j, Z, \phi^{(v)})
\]

\[
= \frac{P(W = w_l, Y^* = k, W^* = w_j| Z, \phi^{(v)})}{P(Y^* = k, W^* = w_j| Z, \phi^{(v)})}
\]

\[
= \frac{\sum_b P(Y = b, W = w_l, Y^* = k, W^* = w_j| Z, \phi^{(v)})}{\sum_c \sum_d P(Y = d, W = w_c, Y^* = k, W^* = w_j| Z, \phi^{(v)})}
\]

\[
= \frac{\sum_b P(Y^* = k| Y = b) P(Y = b| W = w_l, Z, \phi^{(v)}) P(W^* = w_j| W = w_l) P(W = w_l)}{\sum_c \sum_d P(Y^* = k| Y = d) P(Y = d| W = w_c, Z, \phi^{(v)}) P(W^* = w_j| W = w_c) P(W = w_c)}
\]

\[
= \frac{p_{Wlj} \pi(l) \sum_b p_{Ybk} P(Y = b| W = w_l, Z, \phi^{(v)})}{\sum_c p_{Wcj} \pi(c) \sum_d p_{Ydk} P(Y = d| W = w_c, Z, \phi^{(v)})} .
\]

Hence

\[
P(Y = m, W = w_l| Y^* = k, W^* = w_j, Z, \phi^{(v)})
\]

\[
= \frac{p_{Ymk} P(Y = m| W = w_l, Z, \phi^{(v)})}{\sum_a p_{Yak} P(Y = a| W = w_l, Z, \phi^{(v)})} \cdot \frac{p_{Wlj} \pi(l) \sum_b p_{Ybk} P(Y = b| W = w_l, Z, \phi^{(v)})}{\sum_c p_{Wcj} \pi(c) \sum_d p_{Ydk} P(Y = d| W = w_c, Z, \phi^{(v)})} .
\]
Appendix B

Reduction in Number of Parameters for Distribution of Covariates

From (5.2.2), the distribution of the covariates subject to PRAM, $W$, is expressed as a series of conditional distributions $(W_h|W_1, \ldots, W_{h-1})$, $(W_{h-1}|W_1, \ldots, W_{h-2})$, $\ldots$, $W_1$.

Each conditional distribution $(W_c|W_1, \ldots, W_{c-1})$ for $c = 2, \ldots, h$ can be modeled as a multinomial logistic regression model, while $W_1$ can be modeled as a multinomial distribution. Let the number of levels for each covariate $W_c$ be $J_c$, for $c = 1, \ldots, h$. Thus, the number of parameters needed to be estimated for $(W_c|W_1, \ldots, W_{c-1})$ is

$$(J_c - 1) \times \left[ 1 + \sum_{j=1}^{c-1} (J_j - 1) \right]. \quad (B.0.1)$$

Note that since $J_c$ is the number of levels for covariate $W_c$, $J_c$ takes on positive integers greater than or equal to 2, i.e. $J_c \geq 2$ for $c = 1, \ldots, h$. 
In Ibrahim (1990) and van den Hout and Kooiman (2006), the distribution of the covariates subject to PRAM, $W$, was expressed as a joint multinomial distribution. Thus the number of parameters needed to be estimated for the joint multinomial distribution is

$$\left( \prod_{i=1}^{h} J_i \right) - 1.$$  \hfill (B.0.2)

By specifying the distribution of the covariates subject to PRAM, $W$, as a series of conditional distributions, we reduce the number of parameters to be estimated when compared to specifying the distribution as a joint multinomial distribution. From (B.0.1) and (B.0.2), we want to prove the following proposition:

**Proposition B.0.1.** \[ h \sum_{i=1}^{h} (J_i - 1) \times \left[ 1 + \sum_{j=1}^{i-1} (J_j - 1) \right] \leq \left( \prod_{i=1}^{h} J_i \right) - 1, \quad J_i \geq 2 \text{ for all } i. \]

Before proving Proposition B.0.1, we establish the following lemmas:

**Lemma B.0.2.** \[ \frac{J_{k+1}}{J_{k+1} - 1} \leq \prod_{i=1}^{k} J_i, \quad J_i \geq 2 \text{ for all } i. \]

**Lemma B.0.3.** \[ \sum_{i=1}^{k} J_i \leq \prod_{i=1}^{k} J_i, \quad J_i \geq 2 \text{ for all } i. \]

**Lemma B.0.4.** \[ \left( \prod_{i=1}^{k} J_i \right) + (J_{k+1} - 1) \times \left[ 1 + \sum_{j=1}^{k} (J_j - 1) \right] \leq \prod_{i=1}^{k+1} J_i, \quad J_i \geq 2 \text{ for all } i. \]

**Proof.** (Lemma B.0.2) \[ \frac{J_{k+1}}{J_{k+1} - 1} \] is a decreasing function in $J_{k+1}$. Since $J_{k+1} \geq 2$ for $k \geq 1$, \[ \frac{J_{k+1}}{J_{k+1} - 1} \leq 2. \] Note the smallest possible value of $\prod_{i=1}^{k} J_i$ is $2^k$, in other words, $2^k \leq \prod_{i=1}^{k} J_i$. Therefore we have the following

\[ \frac{J_{k+1}}{J_{k+1} - 1} \leq 2 \leq 2^k \leq \prod_{i=1}^{k} J_i, \]

for $k \geq 1$. \hfill $\square$
Proof. (Lemma B.0.3) We use induction on $k$ to prove Lemma B.0.3.

Let $P_k : \sum_{i=1}^{k} J_i \leq \prod_{i=1}^{k} J_i$. $P_1$ is true, since $J_1 \leq J_1$. Assume $P_n$ is true, i.e.,

$$\sum_{i=1}^{n} J_i \leq \prod_{i=1}^{n} J_i.$$ \hfill (B.0.3)

We need to show that $P_{n+1}$ is true, i.e.,

$$\sum_{i=1}^{n+1} J_i \leq \prod_{i=1}^{n+1} J_i.$$ \hfill (B.0.4)

From Lemma B.0.2, we have

$$\prod_{i=1}^{n} J_i \geq \frac{J_{n+1}}{J_{n+1} - 1}.$$ 

Moving terms around, we have

$$\prod_{i=1}^{n+1} J_i \geq \prod_{i=1}^{n} J_i + J_{n+1}.$$ 

From the inductive step (B.0.3), we have

$$\prod_{i=1}^{n+1} J_i \geq \sum_{i=1}^{n} J_i + J_{n+1} = \sum_{i=1}^{n+1} J_i.$$ 

Thus $P_{n+1}$ (B.0.4) holds. \hfill \Box

Proof. (Lemma B.0.4) Since $k \geq 1$ and using Lemma B.0.3, we have

$$\left( \sum_{i=1}^{k} J_i \right) - k + 1 \leq \sum_{i=1}^{k} J_i \leq \prod_{i=1}^{k} J_i,$$

and therefore

$$1 + \sum_{i=1}^{k} (J_i - 1) \leq \prod_{i=1}^{k} J_i.$$
Since \( J_{k+1} - 1 \geq 1 \),
\[
(J_{k+1} - 1) \left\{ \prod_{i=1}^{k} J_i - \left[ 1 + \sum_{j=1}^{k} (J_j - 1) \right] \right\} \geq 0.
\]

Moving terms around, we obtain the inequality
\[
\left( \prod_{i=1}^{k+1} J_i \right) \geq \left( \prod_{i=1}^{k} J_i \right) + (J_{k+1} - 1) \times \left[ 1 + \sum_{j=1}^{k} (J_j - 1) \right]
\]

We proceed to prove Proposition B.0.1. This is proved by induction on \( h \).

**Proof. (Proposition B.0.1)** Let
\[
P_h : \sum_{i=1}^{h} (J_i - 1) \times \left[ 1 + \sum_{j=1}^{i-1} (J_j - 1) \right] \leq \left( \prod_{i=1}^{h} J_i \right) - 1.
\]

When \( h = 1 \), \( P_1 : J_1 - 1 \leq J_1 - 1 \). So \( P_1 \) is true. Assume \( P_k \) is true, i.e.,
\[
\sum_{i=1}^{k} (J_i - 1) \times \left[ 1 + \sum_{j=1}^{i-1} (J_j - 1) \right] \leq \left( \prod_{i=1}^{k} J_i \right) - 1. \tag{B.0.5}
\]

Next we show that \( P_{k+1} \) is true, i.e.,
\[
\sum_{i=1}^{k+1} (J_i - 1) \times \left[ 1 + \sum_{j=1}^{i-1} (J_j - 1) \right] \leq \left( \prod_{i=1}^{k+1} J_i \right) - 1. \tag{B.0.6}
\]

The LHS of expression (B.0.6) is
\[
\sum_{i=1}^{k+1} (J_i - 1) \times \left[ 1 + \sum_{j=1}^{i-1} (J_j - 1) \right]
\]
\[
= \sum_{i=1}^{k} (J_i - 1) \times \left[ 1 + \sum_{j=1}^{i-1} (J_j - 1) \right] + (J_{k+1} - 1) \times \left[ 1 + \sum_{j=1}^{k} (J_j - 1) \right],
\]

and from the inductive step (B.0.5), we have

\[
\sum_{i=1}^{k} (J_i - 1) \times \left[ 1 + \sum_{j=1}^{i-1} (J_j - 1) \right] + (J_{k+1} - 1) \times \left[ 1 + \sum_{j=1}^{k} (J_j - 1) \right] \\
\leq \left( \prod_{i=1}^{k} J_i \right) - 1 + (J_{k+1} - 1) \times \left[ 1 + \sum_{j=1}^{k} (J_j - 1) \right].
\]

Using Lemma B.0.4, we have

\[
\left( \prod_{i=1}^{k} J_i \right) - 1 + (J_{k+1} - 1) \times \left[ 1 + \sum_{j=1}^{k} (J_j - 1) \right] \leq \left( \prod_{i=1}^{k+1} J_i \right) - 1.
\]

Thus \( P_{k+1} \) (B.0.6) holds.
Appendix C

Monotonicity of Entropy-Based Information Loss

C.1 Derivative of $EBIL$ (6.5.7)

Using the expression for $p_{ij}$ as defined in (6.5.1), $NEBIL$ can be written in the following manner. For any $i \in \{1, ..., k + 1\}$,

$$NEBIL = p_i T_i \log \left( \frac{p_i T_i}{p_i T_i + \sum_{j=1 \atop j \neq i}^{k+1} \frac{1-p_i}{k} T_j} \right)$$

$$+ \sum_{h=1 \atop h \neq i}^{k+1} \frac{1-p_h}{k} T_h \log \left( \frac{\frac{1-p_h}{k} T_h}{p_i T_i + \sum_{j=1 \atop j \neq i}^{k+1} \frac{1-p_j}{k} T_j} \right)$$
+ \sum_{h=1 \atop h \neq i}^{k+1} \frac{1-p_i}{k} T_i \log \left( \frac{1}{1-p_i T_i + \sum_{m=1 \atop m \neq h, m \neq i}^{k+1} \frac{1-p_m}{k} T_m} \right) \\
+ \sum_{h=1 \atop h \neq i}^{k+1} p_h T_h \log \left( \frac{1-p_i}{k} + \sum_{m=1 \atop m \neq h}^{k+1} \frac{1-p_m}{k} T_m \right) \\
+ \sum_{h=1 \atop h \neq i}^{k+1} \sum_{j=1 \atop f \neq i, f \neq h}^{k+1} \frac{1-p_f}{k} T_f \log \left( \frac{1-p_i}{k} + \sum_{m=1 \atop m \neq h, m \neq i}^{k+1} \frac{1-p_m}{k} T_m \right)
. \quad (C.1.1)

Differentiating each term inside the summands of (C.1.1) with respect to $p_i$, we obtain

\[
\frac{\partial}{\partial p_i} p_i T_i \log \left( \frac{p_i T_i}{p_i T_i + \sum_{j=1 \atop j \neq i}^{k+1} \frac{1-p_j}{k} T_j} \right) = T_i \log \left( \frac{p_i T_i}{p_i T_i + \sum_{j=1 \atop j \neq i}^{k+1} \frac{1-p_j}{k} T_j} \right) - p_i T_i^2
+ \frac{p_i T_i}{p_i T_i + \sum_{j=1 \atop j \neq i}^{k+1} \frac{1-p_j}{k} T_j} \frac{1}{p_i T_i + \sum_{j=1 \atop j \neq i}^{k+1} \frac{1-p_j}{k} T_j}
\]

\[
= T_i \log \left( \frac{p_i T_i}{p_i T_i + \sum_{j=1 \atop j \neq i}^{k+1} \frac{1-p_j}{k} T_j} \right) - p_i T_i^2
+ \frac{p_i T_i}{p_i T_i + \sum_{j=1 \atop j \neq i}^{k+1} \frac{1-p_j}{k} T_j} \frac{1}{p_i T_i + \sum_{j=1 \atop j \neq i}^{k+1} \frac{1-p_j}{k} T_j}
\]
\[ T_i \log \left( \frac{p_i T_i}{p_i T_i + \sum_{j=1}^{k+1} \frac{1-p_i}{k} T_j} \right) + \frac{T_i \sum_{j=1}^{k+1} \frac{1-p_i}{k} T_j}{p_i T_i + \sum_{j=1}^{k+1} \frac{1-p_i}{k} T_j} =: (C.1.2) \]

\[
\frac{\partial}{\partial p_i} \frac{1-p_i}{k} T_i \log \left( \frac{1-p_i}{k} T_i \right) = \frac{1-p_i}{k} T_i \log \left( \frac{1-p_i}{k} T_i \right) \frac{1-p_i}{k} T_i + \sum_{j=1}^{k+1} \frac{1-p_j}{k} T_j \]

\[
= - \frac{T_i}{k} \frac{1-p_i}{k} T_i \log \left( \frac{1-p_i}{k} T_i + \sum_{m=1}^{k+1} \frac{1-p_m}{k} T_m \right) =: (C.1.3) \]
\[
\partial_{p_i} \frac{p_h T_h}{\left( \frac{1-p_i}{k} T_i + p_h T_h + \sum_{m=1}^{k+1} \frac{1-p_m}{k} T_m \right) \left( -\frac{1}{k} T_i - \frac{1-p_i}{k} T_i \right)^2} = \frac{p_h T_h}{\left( \frac{1-p_i}{k} T_i + p_h T_h + \sum_{m=1}^{k+1} \frac{1-p_m}{k} T_m \right)^2} \bigg/ \frac{p_h T_h}{\left( \frac{1-p_i}{k} T_i + p_h T_h + \sum_{m=1}^{k+1} \frac{1-p_m}{k} T_m \right)}.
\]

and
\[
\frac{\partial V}{\partial p_i} \frac{1 - p_f T_f}{k T_i} \log \left( \frac{\frac{1-\rho T_f}{k}}{\frac{1-\rho_i}{k} + p_h T_h + \sum_{m=1}^{k+1} \frac{1-\rho_m}{k} T_m} \right) \\
= \frac{1 - p_f T_f}{k T_i} \frac{-\frac{1-\rho T_f}{k} \left( -\frac{1}{k} \right) \left( \frac{1-\rho_i}{k} + p_h T_h + \sum_{m=1}^{k+1} \frac{1-\rho_m}{k} T_m \right)}{\left( \frac{1-\rho_i}{k} + p_h T_h + \sum_{m=1}^{k+1} \frac{1-\rho_m}{k} T_m \right)^2} \\
= \frac{1}{k T_i} \frac{1-\rho T_f}{k} \frac{T_i}{p_i T_i + \sum_{j=1}^{k+1} \frac{1-\rho_j}{k} T_j} + \frac{T_i}{p_i T_i + \sum_{j=1}^{k+1} \frac{1-\rho_j}{k} T_j} - \sum_{h=1}^{k+1} \frac{T_i}{p_i T_i + \sum_{j=1}^{k+1} \frac{1-\rho_j}{k} T_j} - \sum_{h=1}^{k+1} \frac{T_i}{p_i T_i + \sum_{j=1}^{k+1} \frac{1-\rho_j}{k} T_j} - \sum_{h=1}^{k+1} \frac{T_i}{p_i T_i + \sum_{j=1}^{k+1} \frac{1-\rho_j}{k} T_j} \\
- \sum_{m=1}^{k+1} \frac{1}{k T_i} \log \left( \frac{\frac{1-\rho T_f}{k} \frac{1-\rho_i}{k} + p_h T_h + \sum_{m=1}^{k+1} \frac{1-\rho_m}{k} T_m}{\left( \frac{1-\rho_i}{k} + p_h T_h + \sum_{m=1}^{k+1} \frac{1-\rho_m}{k} T_m \right)} \right) \frac{1-\rho T_f}{k} \frac{1-\rho_i}{k} + p_h T_h + \sum_{m=1}^{k+1} \frac{1-\rho_m}{k} T_m \\
- \sum_{m=1}^{k+1} \frac{1}{k T_i} \frac{p_h T_h + \sum_{m=1}^{k+1} \frac{1-\rho_m}{k} T_m}{\frac{1-\rho_i}{k} + p_h T_h + \sum_{m=1}^{k+1} \frac{1-\rho_m}{k} T_m} \\
- \sum_{m=1}^{k+1} \frac{1}{k T_i} \frac{1-\rho T_f}{k} \frac{1-\rho_i}{k} + p_h T_h + \sum_{m=1}^{k+1} \frac{1-\rho_m}{k} T_m \\
= (C.1.6)
\]

Using (C.1.2) to (C.1.6), the derivative of NEBIL is
\begin{align*}
&+ \sum_{h=1 \atop h \neq i}^{k+1} \frac{1}{k} T_i \frac{p_h T_h}{1-p_i T_i + p_h T_h + \sum_{m=1 \atop m \neq h, m \neq i}^{k+1} \frac{1-p_m T_m}{k}} \\
&+ \sum_{h=1 \atop h \neq i}^{k+1} \sum_{f=1 \atop f \neq i, f \neq h}^{k+1} \frac{1}{k} T_i \frac{1-p_f T_f}{k} \sum_{m=1 \atop m \neq h, m \neq i}^{k+1} \frac{1-p_m T_m}{k}.
\end{align*}

(C.1.7)

Simplifying (C.1.7), we obtain

\[
T_i \log \left( \frac{kp_i T_i}{kp_i T_i + \sum_{j=1 \atop j \neq i}^{k+1} (1-p_j) T_j} \right) \\
- k^{-1} T_i \sum_{h=1 \atop h \neq i}^{k+1} \log \left( \frac{(1-p_i) T_i}{(1-p_i) T_i + kp_i T_h + \sum_{m=1 \atop m \neq h, m \neq i}^{k+1} (1-p_m) T_m} \right).
\]

**C.2 Derivative of \textit{EBIL} (6.6.2)**

Using the expression for \( p_{ij} \) as defined in (6.6.1), \textit{NEBIL} can be written in the following manner.

\[
\text{NEBIL} = p_{1,1} T_1 \log \left( \frac{p_{1,1} T_1}{p_{1,1} T_1 + \sum_{j=2}^{k+1} \frac{1-p_j T_j}{k}} \right)
\]
\[ + \sum_{h=2}^{k+1} \frac{1-p_h}{k} T_h \log \left( \frac{1-p_h}{k} T_h \right) \left( \frac{1}{p_{1,1} T_1 + \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j} \right) \]

\[ + p_{1,2} T_1 \log \left( \frac{p_{1,2} T_1}{p_{1,2} T_1 + p_{2} T_2 + \sum_{j=3}^{k+1} \frac{1-p_j}{k} T_j} \right) \]

\[ + p_{2} T_2 \log \left( \frac{p_{2} T_2}{p_{1,2} T_1 + p_{2} T_2 + \sum_{j=3}^{k+1} \frac{1-p_j}{k} T_j} \right) \]

\[ + \sum_{m=3}^{k+1} \frac{1-p_m}{k} T_m \log \left( \frac{1-p_m}{k} T_m \right) \left( \frac{1}{p_{1,3} T_1 + p_{2} T_2 + \sum_{j=3}^{k+1} \frac{1-p_j}{k} T_j} \right) \]

\[ + p_{1,3} T_1 \log \left( \frac{p_{1,3} T_1}{p_{1,3} T_1 + p_{3} T_3 + \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j} \right) \]

\[ + p_{3} T_3 \log \left( \frac{p_{3} T_3}{p_{1,3} T_1 + p_{3} T_3 + \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j} \right) \]

\[ + \sum_{m=2}^{k+1} \frac{1-p_m}{k} T_m \log \left( \frac{1-p_m}{k} T_m \right) \left( \frac{1}{p_{1,3} T_1 + p_{3} T_3 + \sum_{j=3}^{k+1} \frac{1-p_j}{k} T_j} \right) \]

\[ + \vdots \]

\[ + \]
\[
+ \left( 1 - \sum_{i=1}^{k} p_{1,i} \right) T_1 \log \left( \frac{(1 - \sum_{i=1}^{k} p_{1,i}) T_1}{(1 - \sum_{i=1}^{k} p_{1,i}) T_1 + \sum_{j=2}^{k} \frac{1-p_j}{k} T_j + p_{k+1} T_{k+1}} \right) \\
+ \sum_{m=2}^{k} \frac{1-p_m}{k} T_m \log \left( \frac{\frac{1-p_m}{k} T_m}{(1 - \sum_{i=1}^{k} p_{1,i}) T_1 + \sum_{j=2}^{k} \frac{1-p_j}{k} T_j + p_{k+1} T_{k+1}} \right) \\
+ p_{k+1} T_{k+1} \log \left( \frac{p_{k+1} T_{k+1}}{(1 - \sum_{i=1}^{k} p_{1,i}) T_1 + \sum_{j=2}^{k} \frac{1-p_j}{k} T_j + p_{k+1} T_{k+1}} \right). \quad (C.2.1)
\]

Differentiating (C.2.1) with respect to \( p_2 \) term by term, we obtain

\[
\frac{\partial}{\partial p_2} p_{1,1} T_1 \log \left( \frac{p_{1,1} T_1}{p_{1,1} T_1 + \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j} \right) \\
= \frac{k^{-1} p_{1,1} T_1 T_2}{p_{1,1} T_1 + \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j}. \quad (C.2.2)
\]

\[
\frac{\partial}{\partial p_2} \frac{1-p_2}{k} T_2 \log \left( \frac{\frac{1-p_2}{k} T_2}{p_{1,1} T_1 + \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j} \right) \\
= -\frac{T_2}{k} \log \left( \frac{\frac{1-p_2}{k} T_2}{p_{1,1} T_1 + \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j} \right) - \frac{T_2}{k} \left( \frac{p_{1,1} T_1 + \sum_{j=3}^{k+1} \frac{1-p_j}{k} T_j}{p_{1,1} T_1 + \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j} \right). \quad (C.2.3)
\]
\[
\frac{\partial}{\partial p_2} \frac{1 - p_m}{k} T_m \log \left( \frac{\frac{1 - p_m}{k} T_m}{p_{1,1} T_1 + \sum_{j=2}^{k+1} \frac{1 - p_j}{k} T_j} \right)
\]
\[
= -k^{-1} \frac{(1 - p_m)T_2 T_m}{p_{1,1} T_1 + \sum_{j=2}^{k+1} \frac{1 - p_j}{k} T_j}, \quad \text{(C.2.4)}
\]
for \( m = 3, \ldots, k + 1 \).

\[
\frac{\partial}{\partial p_2} p_{1,2} T_1 \log \left( \frac{\frac{p_{1,2} T_1}{p_{1,2} T_1 + p_2 T_2 + \sum_{j=3}^{k+1} \frac{1 - p_j}{k} T_j}} \right)
\]
\[
= - \frac{p_{1,2} T_1 T_2}{p_{1,2} T_1 + p_2 T_2 + \sum_{j=3}^{k+1} \frac{1 - p_j}{k} T_j}, \quad \text{(C.2.5)}
\]

\[
\frac{\partial}{\partial p_2} p_2 T_2 \log \left( \frac{\frac{p_2 T_2}{p_{1,2} T_1 + p_2 T_2 + \sum_{j=3}^{k+1} \frac{1 - p_j}{k} T_j}} \right)
\]
\[
= T_2 \log \left( \frac{\frac{p_2 T_2}{p_{1,2} T_1 + p_2 T_2 + \sum_{j=3}^{k+1} \frac{1 - p_j}{k} T_j}} \right)
\]
\[
+ \frac{p_{1,2} T_1 T_2 + T_2 \sum_{j=3}^{k+1} \frac{1 - p_j}{k} T_j}{p_{1,2} T_1 + p_2 T_2 + \sum_{j=3}^{k+1} \frac{1 - p_j}{k} T_j}, \quad \text{(C.2.6)}
\]
\[
\frac{\partial}{\partial p_2} \frac{1 - p_m}{k} T_m \log \left( \frac{\frac{1 - p_m}{k} T_m}{p_{1,h} T_1 + p_2 T_2 + \sum_{j=3}^{k+1} \frac{1-p_j}{k} T_j} \right)
= - \frac{1 - p_m}{k} T_m T_2
\]
\[\frac{1}{p_{1,2} T_1 + p_2 T_2 + \sum_{j=3}^{k+1} \frac{1-p_j}{k} T_j}, \tag{C.2.7}\]
for \(m = 3, \ldots, k+1\).

\[
\frac{\partial}{\partial p_2} (1 - \sum_{i=1}^{k} p_{1,i}) T_1 \log \left( \frac{(1 - \sum_{i=1}^{k} p_{1,i}) T_1}{(1 - \sum_{i=1}^{k} p_{1,i}) T_1 + \sum_{j=2}^{k} \frac{1-p_j}{k} T_j + p_{k+1} T_{k+1}} \right)
= \frac{k^{-1}(1 - \sum_{i=1}^{k} p_{1,i}) T_1 T_2}{(1 - \sum_{i=1}^{k} p_{1,i}) T_1 + \sum_{j=2}^{k} \frac{1-p_j}{k} T_j + p_{k+1} T_{k+1}}, \tag{C.2.8}\]

\[
\frac{\partial}{\partial p_2} \frac{1 - p_2}{k} T_2 \log \left( \frac{\frac{1 - p_2}{k} T_2}{(1 - \sum_{i=1}^{k} p_{1,i}) T_1 + \sum_{j=2}^{k} \frac{1-p_j}{k} T_j + p_{k+1} T_{k+1}} \right)
= - \frac{T_2}{k} \log \left( \frac{\frac{1 - p_2}{k} T_2}{(1 - \sum_{i=1}^{k} p_{1,i}) T_1 + \sum_{j=2}^{k} \frac{1-p_j}{k} T_j + p_{k+1} T_{k+1}} \right)
\]
\[\frac{T_2 (1 - \sum_{i=1}^{k} p_{1,i}) T_1 + \sum_{j=3}^{k} \frac{1-p_j}{k} T_j + p_{k+1} T_{k+1}}{(1 - \sum_{i=1}^{k} p_{1,i}) T_1 + \sum_{j=2}^{k} \frac{1-p_j}{k} T_j + p_{k+1} T_{k+1}}, \tag{C.2.9}\]
\[
\frac{\partial}{\partial p_2} \frac{1 - p_m T_m \log}{p_k k T_m \log} \left( \frac{1 - p_m T_m}{(1 - \sum_{i=1}^{k} p_{1,i})T_1 + \sum_{j=2}^{k} \frac{1-p_j}{k} T_j + p_{k+1} T_{k+1}} \right) \\
= \frac{k^{-1}(1-p_m)T_m T_2}{(1 - \sum_{i=1}^{k} p_{1,i})T_1 + \sum_{j=2}^{k} \frac{1-p_j}{k} T_j + p_{k+1} T_{k+1}}, \quad (C.2.10)
\]

for \( m = 3, \ldots, k \).

\[
\frac{\partial}{\partial p_2} p_{k+1} T_{k+1} \log \left( \frac{p_{k+1} T_{k+1}}{(1 - \sum_{i=1}^{k} p_{1,i})T_1 + \sum_{j=2}^{k} \frac{1-p_j}{k} T_j + p_{k+1} T_{k+1}} \right) \\
= \frac{k^{-1}p_{k+1} T_{k+1} T_2}{(1 - \sum_{i=1}^{k} p_{1,i})T_1 + \sum_{j=2}^{k} \frac{1-p_j}{k} T_j + p_{k+1} T_{k+1}}. \quad (C.2.11)
\]

Using (C.2.2) to (C.2.11), the derivative of \( NEBIL \) with respect to \( p_i \) for \( i = 2, \ldots, k \) is

\[
\frac{T_i}{k} \sum_{h \neq i}^{k} \frac{p_{1,h} T_1}{p_{1,h} T_1 + \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j} - \frac{T_i}{k} \sum_{h \neq i}^{k} \log \left( \frac{1-p_i}{p_{1,h} T_1 + \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j} \right) \\
= \frac{T_i}{k} \sum_{h \neq i}^{k} \frac{p_{1,h} T_1 + \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j}{p_{1,h} T_1 + \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j} + \frac{T_i}{k} \sum_{h \neq i}^{k} \frac{p_{1,h} T_1}{p_{1,h} T_1 + \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j} \\
- \frac{T_i}{k} \sum_{h \neq i}^{k} \frac{1-p_i}{p_{1,h} T_1 + \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j} + \frac{T_i}{k} \sum_{h \neq i}^{k} \frac{p_{1,h} T_1}{p_{1,h} T_1 + \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j} \\
\sum_{j=2}^{k+1} \frac{(1-p_j)}{k} T_j, \sum_{j=2}^{k+1} \frac{(1-p_j)}{k} T_j
\]
\[- \frac{p_{1,i}T_1T_i}{p_{1,i}T_1 + p_iT_i + \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j} + T_i \log \left( \frac{p_iT_i}{p_{1,i}T_1 + p_iT_i + \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j} \right) \]

\[+ \frac{p_{1,i}T_1T_i + T_i \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j}{p_{1,i}T_1 + p_iT_i + \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j} - \frac{p_{1,i}T_1 + T_i \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j}{p_{1,i}T_1 + p_iT_i + \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j} \]

\[+ \frac{T_i}{k} \log \left( \frac{1 - \sum_{i=1}^{k} p_{1,i})T_1}{(1 - \sum_{i=1}^{k} p_{1,i})T_1 + \sum_{j=2}^{k} \frac{1-p_j}{k} T_j + p_{k+1}T_{k+1}} \right) \]

\[- \frac{T_i}{k} \log \left( \frac{1 - \sum_{i=1}^{k} p_{1,i})T_1 + \sum_{j=2}^{k} \frac{1-p_j}{k} T_j + p_{k+1}T_{k+1}}{(1 - \sum_{i=1}^{k} p_{1,i})T_1 + \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j + p_{k+1}T_{k+1}} \right) \]

\[- \frac{T_i}{k} \frac{(1 - \sum_{i=1}^{k} p_{1,i})T_1 + \sum_{j=2}^{k} \frac{1-p_j}{k} T_j + p_{k+1}T_{k+1}}{(1 - \sum_{i=1}^{k} p_{1,i})T_1 + \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j + p_{k+1}T_{k+1}} \]

\[+ \frac{T_i}{k} \frac{\sum_{j=3}^{k} (\frac{1-p_j}{k})T_j}{(1 - \sum_{i=1}^{k} p_{1,i})T_1 + \sum_{j=2}^{k+1} \frac{1-p_j}{k} T_j + p_{k+1}T_{k+1}} \]

\[+ \frac{T_i}{k} \frac{p_{k+1}T_{k+1}}{(1 - \sum_{i=1}^{k} p_{1,i})T_1 + \sum_{j=2}^{k} \frac{1-p_j}{k} T_j + p_{k+1}T_{k+1}} . \]  \hspace{1cm} (C.2.12)

Simplifying (C.2.12), we obtain
\[
T_i \log \left( \frac{kp_i T_i}{kp_{1,i} T_1 + kp_i T_i + \sum_{j=2}^{k+1} (1 - p_j) T_j} \right) \\
- k^{-1} T_i \sum_{h=1, h \neq i}^{k+1} \log \left( \frac{(1 - p_i) T_i}{kp_{1,h} T_1 + \sum_{j=2}^{k+1} (1 - p_j) T_j} \right) \\
- k^{-1} T_i \log \left( \frac{(1 - p_i) T_i}{kp_{1,k+1} T_1 + \sum_{j=2}^{k} (1 - p_j) T_j + kp_{k+1} T_{k+1}} \right).
\]

C.3 Algorithm from Chapter 6.5.9

```
# begin algorithm#

k1<-seq(20,90,10)
results<-NULL

for (j in 1:length(k1))
{
  k<-c(k1[j],80,90)
  d<-100
  to_do<-check(d,k)
```
if (to_do==TRUE)
{
    extreme<-extreme_p2a<-extreme_p2b<-extreme_p2c<-extreme_p3a
    <-extreme_p3b<-extreme_p3c<-extreme_p4<-NULL

    extreme_p1<-solve1(d,k)

    extreme_p2a<-rbind(solve2a1(d,k),solve2a2(d,k),solve2a3(d,k),
                        solve2a4(d,k),solve2a5(d,k),solve2a6(d,k))

    extreme_p2b<-rbind(solve2b1(d,k),solve2b2(d,k),solve2b3(d,k),
                        solve2b4(d,k),solve2b5(d,k),solve2b6(d,k))

    extreme_p2c<-rbind(solve2c1(d,k),solve2c2(d,k),solve2c3(d,k),
                        solve2c4(d,k),solve2c5(d,k),solve2c6(d,k))

    extreme_p3a<-rbind(solve3a1(d,k),solve3a2(d,k),solve3a3(d,k),
                        solve3a4(d,k),solve3a5(d,k),solve3a6(d,k),solve3a7(d,k),
                        solve3a8(d,k),solve3a9(d,k),solve3a10(d,k),solve3a11(d,k),
                        solve3a12(d,k))

    extreme_p3b<-rbind(solve3b1(d,k),solve3b2(d,k),solve3b3(d,k),
                        solve3b4(d,k),solve3b5(d,k),solve3b6(d,k),solve3b7(d,k),
                        solve3b8(d,k),solve3b9(d,k),solve3b10(d,k),solve3b11(d,k),
                        solve3b12(d,k))
extreme_p3c<-rbind(solve3c1(d,k),solve3c2(d,k),solve3c3(d,k),
solve3c4(d,k),solve3c5(d,k),solve3c6(d,k),solve3c7(d,k),
solve3c8(d,k),solve3c9(d,k),solve3c10(d,k),solve3c11(d,k),
solve3c12(d,k))

extreme_p4<-rbind(c(0.5,0.5,0.5),c(0.5,0.5,1),c(0.5,1.0,5),
c(0.5,1,1),c(1,0.5,0.5),c(1,0.5,1),c(1,1,0.5))

##list extreme points in array

extreme<-rbind(extreme_p1,extreme_p2a,extreme_p2b,extreme_p2c,
extreme_p3a,extreme_p3b,extreme_p3c,extreme_p4)

##number of extreme points

numpoints<-dim(extreme)[1]
keeppoints<-array(,c(dim(extreme)[1],dim(extreme)[2]))

for (i in 1:numpoints)
{
  if (sum(extreme[i,]<=1) + sum(extreme[i,]>=0.5) == 6)
  {
    keeppoints[i,]<-extreme[i,]
  }
}
## keep extreme points only in domain

```r
keeppoints <- na.omit(keeppoints)
numpoints_eval <- dim(keeppoints)[1]
```

## evaluate EBIL at extreme points in domain

```r
EBIL_eval <- array(, c(numpoints_eval, 1))
```

```r
store_index <- NULL
```

```r
for (i in 1:numpoints_eval) {
  EBIL_eval[i] <- EBIL3vars(c(keeppoints[i,]), k)
}
```

## location of point(s) that minimize EBIL

```r
min_points <- which(EBIL_eval == min(EBIL_eval))
```

```r
num_min <- length(min_points)
```

## store location of points that minimize EBIL

```r
points <- array(, c(num_min, 3))
```
for (i in 1:num_min)
{
    points[i,]<-keeppoints[min_points[i],]
}

for (h in 1:dim(points)[1])
{
    store<-NULL
    store<-cbind(k1[j],points[h,1],points[h,2],points[h,3],
                 EBIL3vars(points[h,],k))
    results<-rbind(results,store)
}

}
Bibliography


Vita

Yong Ming Jeffrey Woo

Research Interests:
Statistical Disclosure Control/Limitation, EM Algorithm, Generalized Linear Models, Constrained Optimization

Education:
Bachelor of Science in Statistics (Honors), Second Major in Mathematics: The University of Michigan, April 2007.

Select Presentations:
“Generalized Linear Models with Variables Subject to Post Randomization Method, with Dependent Covariates,” Joint Statistical Meetings, Montreal, QC, August 2013. Invited.

Select Publications:
Yong Ming Jeffrey Woo and Aleksandra Slavković, “Generalized Linear Models with Variables Subject to Post Randomization Method”. Submitted.
Yong Ming Jeffrey Woo and Aleksandra Slavković, “Logistic Regression with Variables Subject to Post Randomization Method”. In J. Domingo-Ferrer and I. Tinnirello, editors, Privacy in Statistical Databases, volume 7556 in Lecture Notes in Computer Science, pages 116 - 130. Springer 2012.

Funding & Awards:
U.S. Census Bureau Dissertation Fellowship, 2012-2013; $50,000.
GlaxoSmithKline (GSK) Scholar Award, 2007.

Teaching Experience:
STAT 100: Statistical Concepts and Reasoning
STAT 200: Elementary Statistics.