ADAPTING MASKING TECHNIQUES FOR ESTIMATION PROBLEMS INVOLVING NON-MONOTONIC RELATIONSHIPS IN PRIVACY-PRESERVING DATA MINING

By

MOHAMMAD SAAD AL-AHMADI

Bachelor of Science in Computer Science
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia
1996

Master of Science in Telecommunications Management
Oklahoma State University
Stillwater, OK
2001

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
July, 2006

ADAPTING MASKING TECHNIQUES FOR

ESTIMATION PROBLEMS INVOLVING

NON-MONOTONIC RELATIONSHIPS

IN PRIVACY-PRESERVING

DATA MINING

Dissertation Approved:

Dr. Rathindra Sarathy

Dr. Ramesh Sharda

Dr. Dursun Delen

Dr. Goutam Chakraborty

Dr. A. Gordon Emslie
Dean of the Graduate College

ACKNOWLEDGMENTS

In the name of Allah, the Most Merciful, the Most Compassionate

Thanks, glorification and praises to Almighty God (Allah) for His uncountable favors and bounties upon me and my family. Among them are His assistance in finishing my doctoral program successfully and the blessing of a good, supportive, motivational Ph.D. committee chaired by Dr. Rathindra Sarathy to whom I owe all respect and admiration. Dr. Sarathy was always there for me when I needed him the most. In our fruitful meetings and discussions, I learned from him what I would not have learned from any other source or setting. My high appreciation and thanks are extended to all other members on my Ph.D. committee: Drs. Ramesh Sharda, Goutam Chakraborty and Dursun Delen. The courses they taught and ideas they raised were the best toolkit for me to finish my Ph.D. journey in an enjoyable and successful manner.

I would also like to thank and show the deepest appreciation for the dearest and closest souls to my soul - my parents, Saad and Hamida, who taught me that there is no limit for success, and that bigger ideas and goals work the best in this life. They never stopped supporting and encouraging me and always kept me in their prayers. If I live the rest of my life as a servant at their feet, then I repay them nothing for what they have done for me.

My thanks are extended to my beloved family: Maha, Asim, Shymaa, Ammar, Yasser, and Saad. Without their unconditional and unlimited patience, love, understanding and support, none of what I have done would have been possible. This is

not my work alone, but it is the fruit of our collaborative work as a family. Congratulations to all of us on *our* graduation.

I am also grateful and thankful to my brothers, sisters, and all of my relatives and friends in the USA and in Saudi Arabia who supported me and did not forget me in their good prayers. Many of them shared the joy of graduation with me either face-to-face, by phone or by e-mail. Knowing how long the list would be prevents me from listing individual names.

Additionally, I would like to thank Dr. Ali Amiri, my professor and friend, for his support and encouragement during my doctoral studies, as well as my research partners for the fruitful, cooperative effort that resulted in some publications: Drs. Ramesh Sharda, Rick Wilson, and Peter Rosen. In addition, I would like to thank all faculty, staff, my peers and colleagues, especially my friends Ashish Gupta and Han Li, for the professional academic environment they created in the Department of Management Science and Information Systems (MSIS) in the William S. Spears School of Business at Oklahoma State University.

I would like also to acknowledge and thank Dr. Andrew Robinson for developing and providing the equivalence package for running (model-validation) equivalence tests. Last but not least, I would like to thank the lovely secretary of our department, Tane Kester, for the tremendous amount of time and effort she put in editing this manuscript.

TABLE OF CONTENTS

Chapter Page

Chapter Page

Chapter Page

Chapter Page

LIST OF TABLES

Table Page

Table Page

Table Page

LIST OF FIGURES

ACRONYMS AND ABBREVIATIONS

ANN	Artificial Neural Networks
CADP	Correlated-Noise Additive Data Perturbation (known also as Kim's method)
CC	Canonical Correlation
CDF	Cumulative Distribution Function
C-GADP	General Additive Data Perturbation: the Copula Approach
CI	Confidence Interval
CNM-EGADP	Non-Monotonic Exact General Additive Perturbation: the Copula Approach
DCA	Data-Centric Approach
DM	Data Mining
DV	Dependent variable
EC	European Community
EGADP	Enhanced (or Exact) General Additive Data Perturbation
ET	Equivalence Tests

FDA	Food and Drug Administration
GADP	General Additive Data Perturbation
IPSO	Information P**r**eserving Statistical Obfuscation
IV	Independent variable
LS-SVM	Least Square Support Vector Machines
MDA	Multiple Discriminant Analysis
MLP	Multilayer Perceptron Neural Networks
MR	Multiple Regression
MSE	Mean Square(d) Error or (Mean) Sum of Square(d) Errors
NM-EGADP	Non-Monotonic EGADP
PDF	Probability Distribution Function
PPDM	Privacy-Preserving Data Mining
PPE	Privacy-Preserving Estimation (Regression)
PTTE	The Paired t-Test for Equivalence
RBM	Relationship-Based Masking
RBT	Rotation-Based Transformation

S2-MLP	Secure 2-party Multivariate Linear Regression
S2-MSA	Secure 2-party Multivariate Statistical Analysis
SADP	Simple Additive Data Perturbation
SDC	Statistical Disclosure Control
SDL	Statistical Disclosure Limitation
SVM	Support Vector Machines
TOST	The Two One-Sided t-Test

MATHEMATICAL SYMBOLS AND NOTATIONS

S The set of non-confidential *numeric* and *categorical* attributes

X The set of confidential *numeric* attributes

Y The set of masked *numeric* attributes

V The set of random variate, which is used to generate independent un-scaled noise set **b** for masking confidential attributes X (after scaling: **e**) and generating **Y**

S_i An non-confidential *numeric* and *categorical* attribute i

X_i A confidential *numeric* or *categorical* attribute i

Y_i A masked *numeric* and *categorical* attribute i

V_i A column-vector random variate, which is used to generate independent un-scaled noise b_i for masking attribute X_i (after scaling: e_i) and generating Y_i

p The number of non-confidential attributes **S**

q The number of confidential attributes **X** (or masked attributes **Y**)

d The total number of attributes in a dataset (equals to $p + q$)

e The set of column-vector independent noise resulted from scaling **b** to have the same covariance matrix of **r** (i.e. $\sum_{\mathbf{e}} = \sum_{\mathbf{r}}$)

e_i A column-vector independent noise corresponding to r_i resulted from scaling **b** to have the same covariance matrix of **r** (i.e. $\sum_{\mathbf{e}} = \sum_{\mathbf{r}}$)

b The set of column-vector residuals resulted from regressing random variate **V** on both non-confidential and confidential attributes (**S** and **X**).

b_i A column-vector residuals resulted from regressing a random variate V_i on both non-confidential and confidential attributes (**S** and **X**).

r The set of column-vector residuals resulted from regressing confidential attributes **X** on non-confidential attributes **S**.

r_i A column-vector residuals resulted from regressing one confidential attribute X_i on non-confidential attribute **S**.

$E(.)$ Expected value

$E(.|.)$ Conditional expectation

ε Independent noise or residuals resulted usually from a regression of a random variable on a set of independent variables

ε^{int} The initial prediction error a snooper gets by learning the conditional expectations $E(\mathbf{X}|\mathbf{S})$

ε^{new} The new prediction error a snooper gets by trying to improve his prediction and condition on both $E(\mathbf{X}|\mathbf{S})$ and **Y**

$\sum_{\mathbf{r}}$ The covariance matrix of the residuals resulted from regressing **X** on **S**

$\sum_{\mathbf{b}}$ The covariance matrix of the un-scaled independent noise terms resulted from regressing **V** on **S** and **X**

$\sum_{\mathbf{e}}$ The covariance matrix of the scaled independent noise terms and specified to be equal to $\sum_{\mathbf{r}}$

$f(x)$ Strictly a real (deterministic) mathematical function that represents the exact component of the relationship between an independent variable x a dependent variable y and with no random element involved

x An independent random variable (IV)

y A dependent random variable (DV)

$Corr(.,.)$ The Pearson correlation measures between two random variables. It is also used as a possible data-utility measure for linear relationships.

$\mathbf{I}$ The input to Artificial Neural Networks ANN

$\mathbf{O}$ The output to Artificial Neural Networks ANN

A a dependent random variable

a_i The actual values of a dependent random variable A

$\hat{a}_i$ The corresponding predicted (or fitted) values using a specific nonlinear regression model

$\overline{a}$ The mean of the dependent variable A

R^2 The coefficient of determination

H_0 The null hypothesis

H_a The alternative hypothesis

μ_o The population mean of observations

μ_p The population mean of models predications

δ The equivalence margin of significance tests

δ_0 The equivalence margin of the regression *intercept* using the regression-based validation test for model validation

δ_1 The equivalence margin the regression *slope* using the regression-based validation test for model validation

α The test size (alpha-level) for significance tests or equivalence tests

CHAPTER

I. INTRODUCTION

The maturity of current information technology, especially telecommunications, storage and database technology, facilitates the collection, transmission and storage of huge amounts of raw data, unimagined until a few years ago. For the raw data to be utilized, they must be processed and transformed into information and knowledge that have added value, such as helping to accomplish the task at hand more effectively and efficiently. Data mining techniques and algorithms attempt to aid decision making by analyzing stored data to find useful patterns and to build decision-support models. These extracted patterns and models help to reduce the uncertainty in decision-making environments.

Statisticians and researchers conduct surveys and collect datasets that usually contain tens of records. These datasets are considered to be very large when they contain a few hundred records (Hand, 1998). Traditional statistical techniques are the main (and the most suitable) tools for analyzing these datasets. The main objective of the analysis is to make inference and estimate population parameters from collected samples.

Frequently, statistical agencies must release samples of datasets to external researchers. However, these datasets may have sensitive information about previously surveyed human subjects. This raises many questions about the privacy and confidentiality of individuals in released datasets. Privacy and confidentiality have recently become critical issues and a central concern for many people (Grupe et al.,

2002). Sometimes these concerns result in people refusing to respond and share personal information, or worse, providing wrong responses.

Many laws emphasize the importance of privacy and define the limits of legal uses of collected data. In the healthcare domain, for example, the U.S. Department of Health and Human Services (DHHS) added new standards and regulations to the Health Insurance Portability and Accountability Act of 1996 (HIPAA). The new standards aim to protect "*the privacy of certain individually identifiable health data*" (CDC, 2003). Grupe et al. (2002, EXHIBIT 1 pp. 65) listed a dozen privacy-related acts and legislations issued between 1970 and 2000 in the United States.

On the other hand, these acts and concerns limit, either legally and/or ethically, the releasing of datasets for reasons (sometimes legitimate) such as conducting research in the academic domain or obtaining competitive advantage in the business domain. In some cases, statistical offices face a dilemma of what can be called "war of acts." While they must protect the privacy of individuals in their datasets, they are also legally required to disseminate these datasets. The conflicting objectives of the Privacy Act of 1974 and the Freedom of Information Act is just one example of this dilemma (Fienberg, 1994). This has led to an evolution in the field of statistical disclosure limitation (SDL).

SDL methods attempt to find a balance between data utility (valid analytical results) and data security (privacy and confidentiality of individuals). In general, these methods try to either (a) limit the access to the values of sensitive attributes (mainly at the individual level), or (b) mask the values of confidential attributes in datasets while maintaining the general statistical characteristics of the datasets (such as mean, standard deviation, and covariance matrix). Data perturbation methods for microdata are one class

of masking methods. Since most statistical analysis methods are based on linear models, data perturbation methods generally aim to maintain linear relationships (Muralidhar et al., 1999).

When the size of datasets are large, traditional statistical analysis techniques may not be the appropriate tools to use for two reasons (Hand, 1998; 2000; Hand et al., 2000). First, traditional statistical tools become unsuitable for making sense of the data and for making inferences about the population for large datasets; for instance, almost any small difference in a large dataset becomes significant. Second, large datasets may suggest that data was not collected for inference (parameter estimation) about the population, and that another type of analysis might be more appropriate. In most cases, a significant amount of collected data is generated as a consequence of some unplanned activities (e.g., transactional databases) vs. planned activities (e.g., experiment or survey designs). Therefore, as the size of datasets grows exponentially, the use of other (size-matched) analytical tools such as data mining becomes more appropriate.

Examples of large datasets are abundant. MarketTouch, a company located in Georgia, supports direct marketers with data and analytical tools (DMReview.com, 2004). It has a six-terabyte database called Real America Database (RADBÒ), which provides information about more than 93 million households and 200 million individuals. It is updated monthly with more than 20 million records.

Statistical agencies also experience this phenomenon of rapidly growing datasets. The Census Bureau (2001) reported that the collected Census 2000 data consist of "*information about the 115.9 million housing units and 281.4 million people across the United States.*" These large sizes suggest the need for analytical tools that are suitable for

large datasets, and again, data mining tools naturally come into play. Actually, the Census Bureau (accessed 2004) has started providing programs that have data mining capabilities such as DataFerrett (Federated Electronic Research, Review, Extraction and Tabulation Tool), which can be used to analyze and extract data from TheDataWeb - a repository of datasets that can be accessed freely online or bought offline. These datasets cover more than 95 subjects.

Data mining techniques may lead to more significant threats to privacy and confidentiality compared to traditional statistical analytical techniques. Domingo-Ferrer and Torra (2003) made a connection between SDL methods and some AI (artificial intelligent) tools (note that statistics and AI fields are among the main fields contributing to the data mining field). They suggested there are two possible risks of AI (or equivalently DM) tools regarding the privacy and confidentiality of released masked datasets: disclosure and re-identification threats.

From a disclosure threat perspective, DM tools can be used to aggregate or combine masked copies of a specific original dataset (Domingo-Ferrer and Torra, 2003). The goal is to reverse the masking effect and build the original dataset, which raises a *confidentiality* issue. This may pose a great threat when simple unsophisticated SDL techniques, such as simple additive data perturbation (SADP) method (Traub et al., 1984), are used and many masked copies are released. DM tools are also used to enforce data integrity and consistency in distributed databases by re-identifying different records belonging to the same individual. Contrary to DM, SDL methods aim to avoid identity disclosure. Thus, from a re-identification threat perspective, DM tools can be used to re-identify individuals in masked datasets (raising a *privacy* issue). Domingo-Ferrer and

Torra (2003) suggested that SDL methods should consider the existence of DM tools to assess their impact on disclosure (value disclosure) and re-identification (identity disclosure) threats.

These concerns about privacy and confidentiality when DM tools are used have led to the birth of privacy-preserving data mining (PPDM). The main goal of PPDM is to find useful patterns and build accurate models from datasets without accessing the individuals' precise original values in records of datasets (Agrawal and Srikant, 2000). In many cases, PPDM algorithms employ one or more methods to protect the data. One protection method used is Simple Additive Data Perturbation Method (SADP) (Traub et al., 1984), which has undesirable characteristics in terms of data utility and data security (Muralidhar et al., 1999). Most of the newer and more sophisticated data perturbation and masking methods, such as C-GADP (Sarathy et al., 2002), IPSO (Burridge, 2003), EGADP (Muralidhar and Sarathy, 2005b) and data shuffling (Muralidhar and Sarathy, 2003a; 2005a), have not been investigated in the PPDM domain. The only exception is the GADP method (Muralidhar et al., 1999), which appears in a few privacy-preserving classification studies (Islam and Brankovic, 2004; Wilson and Rosen, 2002; 2003) . This study will discuss the possibilities of using some of these masking methods for *monotonic* relationships in privacy-preserving estimation (PPE), as well as how these methods can be adapted and extended for the more general case involving *non-monotonic* relationships.

I.1. PPE Problem – Definition and Research Scope

The goal of this study is to investigate the possibility of using and adapting some data masking (mainly data perturbation and data shuffling) methods to develop new

privacy-preserving estimation PPE algorithms for numeric variables involving non-monotonic relationships. Estimation (or regression) tries to estimate and quantify the relationships among variables in datasets. The main goal is to investigate the impact of newer data perturbation methods on different types of relationships in datasets. The required criterion is simple yet powerful: any used perturbation (or masking) method should not hurt or alter (to the extent possible) the structure and type of relationships among variables existing in original datasets. The result of testing such impact could be either *preservation* or *destruction* of the original relationships in masked datasets. Unfortunately, current data perturbation methods do not preserve all possible relationships, as will be seen later. They preserve monotonic nonlinear relationships, at best, and simpler relationships (linear ones). This study proposes four new PPE masking algorithms, based on the concepts of data perturbation and data shuffling, to preserve the more difficult non-monotonic relationships. The focus and scope of this study in the space of the privacy-preserving data mining PPDM techniques is represented in Figure 1.

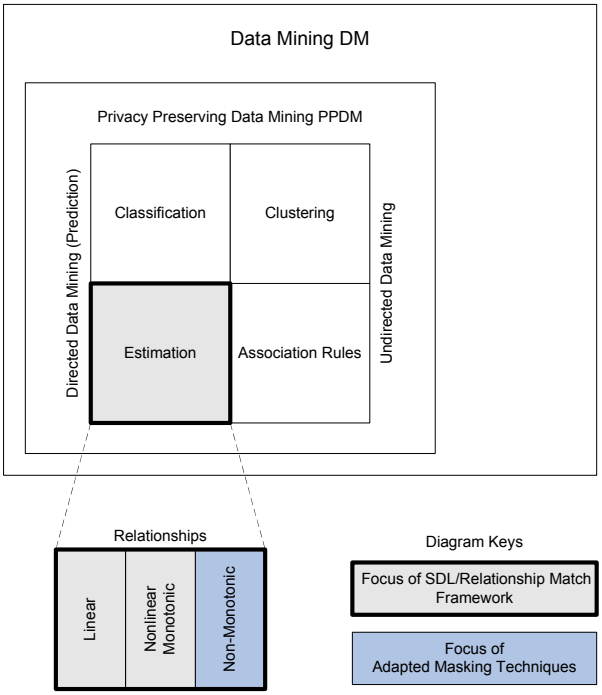


Figure 1. Research scope of the study

I.2. PPE Problem – Importance and Requirements

There are many important real-life problems that require building estimation models for continuous variables. Many of these models are related to human subjects and involve confidential attributes. In the following paragraphs, we shed some light on some aspects illustrating the importance and the need to protect the privacy and confidentiality of human-related data used in regression models. Before we start, we want to resolve some terminology issues. While the term "estimation" appears more frequently in engineering literature, other fields use the term "regression" (Ridgeway, 2003). Therefore, we may use the terms estimation and regression (technique(s)/ model(s)/ problem(s)) interchangeably.

The need to build estimation models (prediction models for continuous variables) occurs frequently in different domains. Berry and Linoff (2000; 2004) have mentioned

many typical examples including estimating the number of children in a family, the total household income of a family, real estate value, and a customer's lifetime value.

Estimation is also indirectly related to classification. Instead of classifying customers as "respondents" or "non-respondents," for example, estimation techniques can be used to assign a probability of expected level of responsiveness (Berry and Linoff, 2004). This is very useful in marketing campaigns when budgets are insufficient to target all possible respondents. Then, expected respondents with the highest probability of responsiveness can be targeted first.

In large datasets, which typify DM datasets, the frequency of each class in categorical and binary variables is usually high. On the other hand, a real-number salary figure, for example, in a released dataset can be unique to the degree that it can single out the real identity of a de-identified whole record. Categorical and binary attributes automatically have more security against disclosure risks than continuous variables. Because of this, Domingo-Ferrer et al. (2001) indicated that while non-perturbative and masking sampling methods (Skinner et al., 1994), which are based on the concept of sample uniqueness and population uniqueness, may suit categorical microdata, they could not be used for numeric microdata. Hence, continuous attributes should attract more attention for protection against disclosure risks than categorical attributes.

Regression models that involve allocating huge amounts of money based on information about human subjects are sometimes required. Ridgeway (2003) briefly mentioned a real-world case, which exemplifies the importance of such estimation models. Medicare (2004) is a federal health insurance program that covers elderly people (65 years or older), younger people with disabilities, or those suffering from End-Stage

Renal Disease (ESRD). There are many Centers for Medicare and Medicare Services (CMS) nationwide. CMS is required by the 1997 Balanced Budget Act to develop a "*prospective payment system*" to allocate a $4.3 billion budget to healthcare facilities that help Medicare-insured inpatients to rehabilitate. Part of the system is a (regression) "*cost model*" built using different features (attributes) of patients to predict the cost of rehabilitation. These attributes come from two main sources. CMS provides attributes such as age, reason for hospitalization, and cost of care. Secondary sources provide functional ability scores measuring motor and cognitive abilities of patients. The cost model is used to predict the cost of rehabilitation per patient. Accordingly, healthcare facilities are reimbursed.

Regression models are frequently used in business problems. Linear regression is the best tool when all relationships among variables in a dataset are linear. This is guaranteed when the dataset is normally distributed. However, most business datasets are non-normal (Zhang, 2004), which opens the door to all possible forms of relationships including non-linear (monotonic or non-monotonic) relationships.

Clearly, estimation and regression models for both linear and nonlinear relationships are important for many applications. In many cases, accessing sensitive data about human subjects might be necessary to build such models. Consequently, concerns about privacy and confidentiality automatically arise. However, there is a definite shortage in the amount of research related to privacy-preserving estimation technique (PPE), especially for nonlinear and non-monotonic relationships, as we will see in the PPE-related literature (Subection II.1.1) in the next chapter.

In summary, estimation (regression) is one pillar of the four main pillars of data mining (DM) techniques (see Figure 1 above) and is frequently used for different applications. In many cases, individuals' privacy and confidentiality become a greater issue for numeric values than categorical values (Domingo-Ferrer et al., 2001; 2003). Actually, converting numeric values into categorical ones is one way to protect privacy and confidentiality. Alternatively, masking methods such as data perturbation can be used. Accordingly, the first of three main requirements that masking methods need to satisfy in the context of PPE is:

- **Requirement I:** Masked datasets must allow accurate estimation models to be built, while preserving individuals' privacy and confidentiality.

Different types of relationships can exist in a dataset. For instance, multivariate normal datasets guarantee that all existing relationships among variables are linear. For this special case, some existing masking methods are readily available and can perfectly preserve linear relationships. However, most (business) datasets contain nonlinear relationships (Zhang, 2004), which can be monotonic or non-monotonic (Fisher, 1970). "*A truth about data mining not widely discussed is that the relationships in data the miner seeks are either very easy to characterize, or very, very hard,*" (Pyle, 2003). This leads us to the second requirement:

- **Requirement II:** Masking methods must preserve all possible types of relationships among variables (non-monotonic or monotonic; linear or nonlinear).

Al-Ahmadi et al. (2004) suggested an approach, called the "Data-Centric" Approach in PPDM, for developing new privacy-preserving data mining PPDM algorithms. This approach suggests, unlike many current PPDM algorithms (Agrawal and Srikant, 2000;

Thuraisingham, 2005), that any new PPDM algorithm should focus only on altering and changing original datasets without changing standard data mining algorithms. This should be done in a way that does not hurt the validity of required analyses (data utility) while maximizing data security. This approach is important for reasons that will be apparent in the Subsection II.1.2 titled "Data-Centric Approach (DCA) for Privacy-Preserving Data Mining." The third requirement is:

- **Requirement III:** PPE methods must be Data-Centric.

I.3. Motivation Example

Boyens et al. (2002) have indicated that more businesses have started to outsource data mining tasks to specialized data mining and knowledge discovery consultant companies. In other cases, some organizations such as statistical offices are required by law to release datasets to outsiders such as researchers and data miners. In addition, some businesses need to release datasets to specific parties because of an alliance, although they (i.e. the businesses) may have the needed expertise to run and build different data mining models by themselves. In all of these cases, the data utility and data security concerns are important.

Table 1. First 5 and last 5 records of the store dataset (motivation example)

NO	Non-Confidential Attributes		Confidential Attributes	
	Age (S_1)	Gender (S_2)	Expenditure \$ (X_1)	Debt \$ (X_2)
1	53.11	0	258.57	514.74
2	57.37	1	211.35	569.33
3	22.06	0	224.45	609.51
4	25.46	1	272.98	547.63
5	51.25	1	292.63	516.33
:	**:**	**:**	**:**	**:**
996	37.16	0	391.54	413.41
997	50.72	1	301.59	500.72
998	28.71	1	296.49	531.81
999	27.71	0	308.91	537
1000	29.62	1	298.04	469.97
Min	20.009	0	172.610	383.490
Max	59.969	1	444.190	632.800
Mean	40.264	0.468	296.775	504.692
STD	11.884	0.499	59.299	59.054

To motivate our discussion, we provide a *hypothetical* example. A store wants to release a 1,000-record dataset to an allied market analysis firm while protecting the privacy and confidentiality of its customers. The dataset consists of four main variables (along with other identifier fields such as Name and Address). Two of these variables are non-confidential and two are confidential. The two non-confidential attributes are the age of customers (numeric between 20-60 years) and the gender (binary – 0: male or 1: female), and they are denoted by S_1 and S_2, respectively. The two confidential attributes are the annual expenditure in \$ (numeric) and the debt in \$ (numeric), and they are denoted by X_1 and X_2, respectively. The first five and the last five records of the dataset are shown in Table 1. The main analysis required by the market analysis firm is regression and estimation modeling.

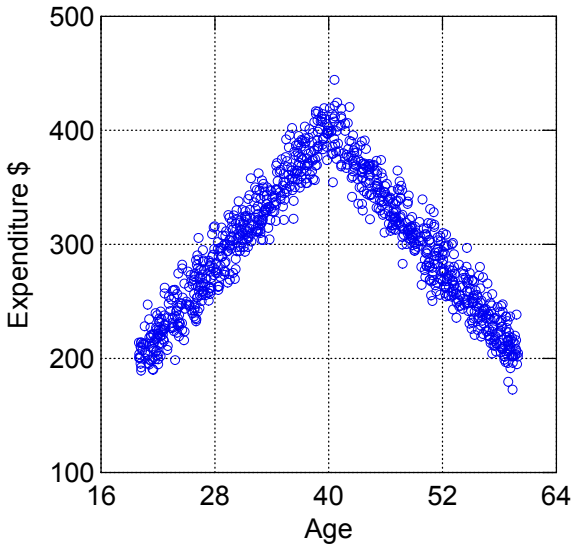


Figure 2. Motivation Example: Non-monotonic relationships between Age (S_1) and Expenditure$ (X_1)

The store is required ethically and legally to protect the privacy and confidentiality of its customers (data security). This is also very important for maintaining customer trust and loyalty, and retaining profitable customers. At the same time and from a different perspective, the store is required to enable the allied market analyst firm to obtain accurate regression models from the released masked dataset (data utility). As an initial precaution, the store de-identifies the dataset by removing identifier fields such as Name and Address from the dataset.

Figure 41 (Appendix D) shows the relationships among these four variables. The figure clearly indicates the existence of nonlinear non-monotonic relationships (e.g., the relationship between S_1: age and X_1: expenditure; see Figure 2). In Section II.3 in the next chapter and Appendix E, we show that current masking techniques are unable to maintain relationships of the type shown in Figure 2 (i.e. non-monotonic relationships).

I.4. Summary and Outline

In this chapter, we introduced the privacy-preserving estimation PPE involving non-monotonic relationships and we specified its scope. We also talked about its importance from a practical point of view. In addition, we outlined the *general* requirements of the privacy-preserving data mining PPDM and the *specific* requirements of the privacy preserving estimation PPE.

In the next chapter (Chapter II), the relevant literature related to PPDM is briefly reviewed. The focus is on the limited literature of PPE. Then, the related work and concepts in masking methods are presented. This includes the advantage of using masking methods in PPDM. Optimal masking methods are defined, and the difficulty of their practical implementation is explained. Some recent masking methods are discussed, and their limitation in dealing with non-monotonic relationships is demonstrated. This chapter concludes by listing possible research questions in PPE.

Chapter III lays the theoretical basis for our proposed masking methods and their needed tools. The chapter starts by defining a "relationship" in the context of estimation and regression problems. Then, the theoretical role of Artificial Neural Networks (ANN) in learning such relationships is presented. Finally, the roles of residuals left after removing conditional expectations $E(\mathbf{X}|\mathbf{S})$ in defining and guiding data utility and data security requirements are discussed. Chapter IV proposes four new masking methods for preserving non-monotonic relationships by adapting and extending some existing masking methods. The chapter talks about their implementation and their assumptions. Chapter V treats the subject of assessing the success of the proposed masking methods in terms of data utility and data security. Some possible new and existing data utility and

data security measures for measuring the effectiveness of PPE masking methods are listed. Chapter VI assesses the effectiveness of Relationship-Based Masking (RBM) when the relationships among confidential attributes are linear. It starts by demonstrating the use of some of the RBM data utility and data security measures on the motivation example. Additionally, it briefly examines how a snooper may try to compromise a masked dataset involving non-monotonic relationships. Then it explains how the characteristics of original datasets determine the characteristics of masked attributes. Finally, the determination of characteristics is empirically demonstrated.

Chapter VII discusses the effectiveness of the RBM approach in terms of data utility when the relationships among confidential attributes are nonlinear. Eleven simulated datasets are used. The data utility assessment is done using the measures proposed in Section V.3. Although the focus is on data utility for reasons explained there, the chapter briefly discusses the subject of data security.

Chapter VIII concludes this work by summarizing the main findings and results of this research. Additionally, the limitations of proposed methods are discussed. Further, some possible future research trends in PPE are given. In addition, the possibility of using the proposed masking methods for other PPDM techniques such as privacy-preserving classification is suggested.

CHAPTER

II. LITERATURE REVIEW

In this chapter, we start by reviewing the literature related to Privacy-Preserving Data Mining (PPDM). The focus is mainly on Privacy-Preserving Estimation (PPE). Then, we review the main related concepts in statistical disclosure control (SDC). We also review some data perturbation and data shuffling masking methods. Next, we investigate the possibility of using these masking methods for PPE and assess their impact on non-monotonic relationships. We conclude this chapter by articulating some of the possible research questions in PPE.

II.1. Related Work in Privacy-Preserving Data Mining (PPDM)

The data mining "algorithm" (or "technique" in the terminology from Berry and Linoff (2000; 2004)) is one of five different dimensions that can be used to classify privacy-preserving data mining methods (Verykios et al., 2004b). Similar to the classification of data mining (DM) techniques proposed by Berry and Linoff (2000; 2004), privacy preserving data mining (PPDM) techniques can be classified as: (a) directed PPDM techniques: privacy preserving estimation and privacy preserving classification (both can be called predication techniques), and (b) undirected PPDM techniques: privacy preserving association rules and privacy preserving clustering.

While the field of privacy-preserving data mining is new, relatively good progress has been made on privacy-preserving classification and association rules. Examples of

privacy-preserving classification are Agrawal and Srikant (2000), Du and Zhan (2002; 2003), Du et al. (2004), Islam and Brankovic (2004), Johnsten and Raghavan (2000; 2001), Kantarcioglu and Clifton (2004b), Kantarcioglu and Vaidya (2003), Lindell and Pinkas (2002), Vaidya and Clifton (2004), Vaidya et al.(2004), and Yang et al. (2005). Examples of *privacy-preserving association rules* are Ashrafi et al. (2003; 2004), Evfimievski et al. (2002), Evfimievski et al. (2004), Kantarcioglu and Clifton (2004a), Oliveira and Zaïane (2003a), Oliveira et al. (2004), Rizvi and Haritsa (2002), Saygin et al.(2002), Vaidya and Clifton (2002), Verykios et al. (2004a), and Zhang et al. (2004).

Some progress has been made in *privacy-preserving clustering*. Examples include Klusch et al. (2003), Lin et al. (2004), Merugu and Ghosh (2003a; 2003b) Oliveira and Zaïane (2003b; 2004a; 2004b), and Vaidya and Clifton (2003). However, there has been little research in *privacy-preserving estimation (regression)*. Figure 3 shows an abstract view of privacy-privacy data mining (PPDM) related literature broken down by PPDM

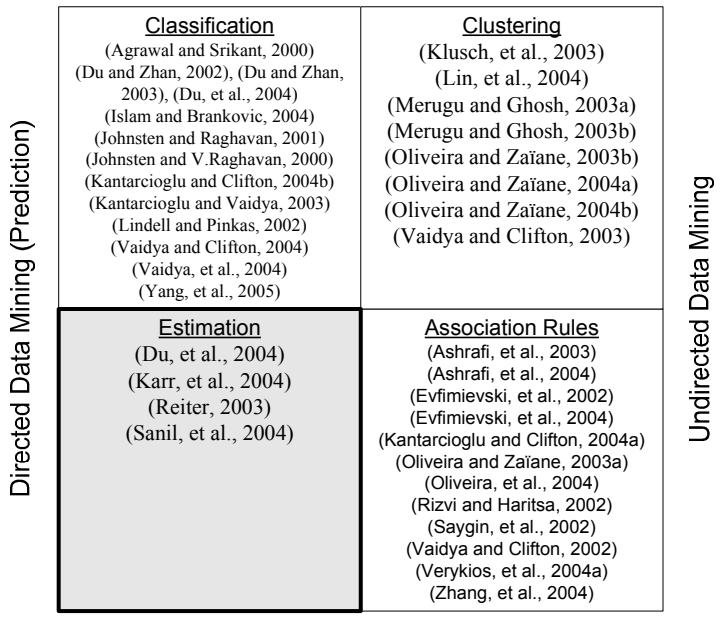


Figure 3. Privacy Preserving Data Mining PPDM Literature

technique.

II.1.1. Review of Privacy-Preserving Estimation (PPE) Literature

Sanil et al. (2004) proposed an algorithm for computing the exact coefficients of multiple linear regression on the union of a *vertically*-distributed (or partitioned) dataset without sharing original values. This algorithm is applicable when there is a single shared, non-confidential dependent variable, and the unshared confidential, independent variables are owned by more than two parties (agents) involved in the estimate process. It is based on Powell's algorithm (Brent, 2002; Powell, 1964) for finding the minimum (as a numerical solution for a series of one-dimensional minimization problems) of a multivariable quadratic function without calculating its derivative. In addition, the algorithm of Sanil et al. (2004) utilizes the secure summation algorithm (Benaloh, 1987; Clifton et al., 2002), which is considered to be a part of secure multiparty computation in the cryptography literature (Schneier, 1996, pp. 551-552). The secure summation algorithm is used to share a statistical summary (total), populated partially by every party without revealing how much each party contributes to that statistic. This total is needed for estimating the regression coefficients iteratively. By the end of the proposed regression algorithm, each party can calculate accurately, from the global regression model, the coefficients of the variables they own and share them with other parties.

Karr et al. (2004) dealt with the case of building multiple linear regression on the union of a *horizontally*-distributed dataset. They suggested two approaches. The first approach, called the *secure data integration* procedure, is used to integrate horizontally-distributed datasets from more than two parties (agents) into one dataset while protecting the identity of the data source. The integrated dataset could then be shared among

cooperative parties. Each party could locally run linear regression analysis (or any other statistical analysis) and any of its diagnostics on the integrated dataset. This approach clearly does not rise to the minimum requirements of protecting the privacy and confidentiality of surveyed human subjects because it does not mask confidential attributes and aims only to protect the identity of the data sources (i.e. the identity of the involved parties, not the identity or confidentiality of surveyed human subjects).

A second approach is based on the additive nature of the linear regression analysis. Instead of sharing and integrating the original unmasked records of the distributed datasets, statistics required to calculate the least squares estimators of linear regression coefficients are shared and integrated in a secure manner using the secure summation algorithm (Benaloh, 1987; Clifton et al., 2002; Schneier, 1996). In this approach, diagnostics could not be calculated directly. Karr et al. (2004) proposed two approaches to resolve this issue based on whether the computations of diagnostics are additive with respect to the parties. When the nature of the computations of the diagnostics are additive, such as R^2 and correlations, locally-calculated numerators and denominators of the diagnostic measures are shared using the secure summation algorithm to calculate the global diagnostic measures. When this is not possible, as in the case of sharing the residuals, simulated residuals derived from synthetic independent variables and mimicking the relationships among original residuals and independent variables are integrated and shared using the secure data integration procedure. The procedure of creating synthetic residuals is very similar to the procedure proposed by Reiter (2003), which is briefly discussed below.

Remote regression servers (cf. Duncan and Mukherjee, 2000; Keller-McNulty and Unger, 1998; Schouten and Cigrang, 2003) are access-limitation methods for protecting microdata while enabling users to build linear regression models. Instead of running the regression analysis locally, users submit a request of the regression analyses they require and the server returns the results in terms of regression coefficients and standard errors. Although this approach has the advantage of building linear regression models using original values, users do not usually have any means of checking the fit of their models. Reiter (2003) proposed a method to enable users to check the fit of their models while limiting the disclosure risks. The proposed method is based on releasing artificial, simulated (marginally-wise) dependent and independent variables, residuals and fitted values that mimic the original relationships of the built models. Then these synthetic variables could be used, similar to the use of original variables, to assess the fit of the regression models. He suggested that this approach could be used to provide diagnostics for other possible remotely built, generalized linear models.

The computations of many multivariate methods, including multivariate linear regression, depend on matrix computations such as matrix multiplication and matrix inverse. Based on this well-known fact, Du et al. (2004) proposed some protocols called secure two-party matrix computations protocols, which enable two agents to collaboratively run matrix computations without knowing about or accessing the other party's original, sensitive values and without the involvement of a third party. They suggest that secure versions of some multivariate statistical analysis could be formulated using these secure matrix computations building blocks. This approach is generally called the *Secure 2-party Multivariate Statistical Analysis (S2-MSA)*. These protocols are used

to reformulate some specific multivariate statistical analysis problems including the *Secure 2-party Multivariate Linear Regression (S2-MLP)* problem, when the dependent variable is known to both agents.

As we have seen, not much has been done in privacy-preserving estimation and regression when we need to release a dataset for general analysis. The focus of current PPE methods are linear relationships and linear regression methods. Our work deals with a more difficult problem that involves non-monotonic relationships. Most prior work deals with partitioned (vertically or horizontally) datasets while our work is about centralized datasets, whose masked copies will be released in whole for regression tasks. The above methods require the involvement of all parties every time they must run a regression model, while our proposed methods release the whole masked datasets and do not require timely cooperation between the data owner and the data analyst. Many of the above methods assume and restrict applicability to the existence of a single shared, dependent variable, while our methods allow building regression models using any numeric variable as a dependent variable. The mentioned methods are proposed mainly for regression tasks while there is initial evidence that our approach can be used for other analyses (see Section VIII.2 titled “Possible Opportunities and Limitations”).

II.1.2. Data-Centric Approach (DCA) for Privacy-Preserving Data Mining

Many PPDM approaches modify some existing DM algorithm(s) while masking the data (Thuraisingham, 2005); see, for example, Agrawal and Srikant (2000). Enforcing the use of a modified algorithm with a masked dataset to ensure accurate results is not a good idea for many reasons. First, data miners usually employ more than one algorithm

to mine a dataset. Examining all data mining algorithms, as well as modifying them, is not feasible. Second, once a dataset is released, there is no guarantee as to which algorithm might be applied. Using a non-prescribed, standard algorithm may lead to incorrect conclusions and actions.

Instead, as suggested by Al-Ahmadi et al. (2004), datasets should be protected or masked without reference to a specific DM algorithm. More recently, Oliveira and Zaïane (2004c) supported the concept of Data-Centric Approach (DCA) in their standardization suggestions to PPDM researchers and developers. Oliveira and Zaïane (2004a) applied the DCA concept practically in developing a new PPDM clustering algorithm called Rotation-Based Transformation (RBT), which tries to modify only the data such that applying standard distance-based clustering algorithms on both the normalized original datasets and the transformed datasets produce the same results. Hence, data perturbation and masking methods can be a good starting point for implementing the Data-Centric Approach (DCA) in PPDM.

Next, we review the latest developments in data masking methods and see how we can use some of them in estimation problems. In Appendix A, we also provide a simple framework called the SDL/Relationship Match Framework for choosing a specific data perturbation method (or, in general, any masking method) to mask a specific dataset for building estimation models. The match is based on the type of existing, most difficult relationships in original datasets that the chosen masking method can preserve and reproduce in masked datasets.

II.2. Related Work in Masking Methods

Statistical Disclosure Limitation (SDL), known also as Statistical Disclosure Control (SDC), techniques are a set of techniques that aim to control the amount of disclosed information about sensitive attributes at the level of individual records in disseminated datasets while providing valid overall statically analyzable datasets. The main goals of SDL techniques are twofold: (a) to minimize disclosure risks by protecting the privacy (identity disclosure) and confidentiality (value disclosure) of individuals surveyed or included in released datasets, and simultaneously (b) to maximize data utility (preserving original (statistical) characteristics of datasets) at the aggregate level (Willenborg and Waal, 2001).

The phrase "disclosure limitation" in SDL (or "disclosure control" in SDC) indicates two things. First, there is a specific amount of information, usually at an aggregate level, in the original dataset one wants to reveal. Second, one wants to make sure that no further information (mainly sensitive or confidential) can be learned. In fact, when a dataset is released, some sort of disclosure automatically occurs, and total elimination of disclosure is impossible (Fienberg, 1994). Therefore, there are always possible disclosure risks associated with any released dataset. Hence, methods to assess different possible disclosure risks for any masking method are needed. Similarly, methods are also needed to assess data utility. Thus, when a masking method is proposed, existing or new measures that quantify data utility and disclosure risk are used to prove the effectiveness of the masking method.

Two natural types of disclosure one wants to prevent are: identities of individuals (identity disclosure) and exact values of their confidential attributes (exact value

disclosure). The problem of disclosure control is complicated by the concept of the second type of value disclosure: the partial value disclosure. In this case, the exact values of confidential attributes are protected, but good statistical estimates can be gained by accessing masked attributes (Adam and Wortmann, 1989; Dalenius, 1977; Muralidhar et al., 1999; Muralidhar and Sarathy, 1999; Sarathy and Muralidhar, 2002).

The only way to completely eliminate disclosure risks is to avoid releasing datasets (given that other physical security measures are implemented). In this situation, unfortunately, no data utility is achieved (Fienberg, 1994). Datasets need to be disseminated on many occasions for many good, possible reasons. In the case of statistical databases, permissible queries should get responses. Two important reasons, among others, are that dissemination is either (a) legally required, as in the case of the Bureau of Census, or (b) research-motivated. For example, a hospital releases information about patients with a specific disease to a medical research institute developing a cure for the disease. In the data mining field, datasets can be released to an external DM and knowledge discovery consultant company in an effort to gain some competitive advantage. There is an increasing trend in many businesses to outsource data mining tasks (Boyens et al., 2002).

There are basically two general approaches for disclosure control: limiting access to sensitive attributes or masking original confidential attributes (Adam and Wortmann, 1989; Willenborg and Waal, 2001). Query-restrictions used in statistical databases (Adam and Wortmann, 1989) are one example of access limitation approaches. In this approach, a query returns only an aggregate statistic such as the SUM. The number of records used to build the aggregate statistic should be more than a specific threshold for the result to be

sent to the user. In addition, successive queries should not have a large overlap. However, in many cases, this is not sufficient to prevent disclosure.

When the results of permissible queries are restricted to aggregate information, *inferential* disclosure can occur when a snooper can learn unrevealed information by combining the results of more than one query. A snooper is a legitimate user who misuses access privilege to obtain unauthorized content (Adam and Wortmann, 1989; Muralidhar et al., 1999; Muralidhar et al., 2001). For example, a snooper can combine the results of the following two queries, which can be issued by different users, to learn the salary of a company's president: query 1 - "total salaries the company pays," and query 2 - "total salaries the company pays except for the president" (Clifton, 2003). Although the result of each query by itself does not represent a privacy or confidentiality threat, combining the two leads to exact disclosure. One proposed solution (which may not be practical) is to track all replied queries of all users to avoid answering a query that can increase the amount of information released and lead to disclosure. Malvestuto and Moscarini (2003) discussed the inference problem of confidential values using repeated queries in multidimensional databases, as well as a graphical auditing solution (the answer map) for the problem.

Willenborg and Waal (2001) devoted a book to the discussion of the principles related to masking techniques. They differentiated between two types of datasets that need to be protected: tabular data (aggregated data) and microdata (individual records). Perturbative and non-perturbative masking methods are reviewed for each dataset type. The former replaces original values with fabricated ones, while the latter utilizes original

values. Willenborg and Waal also reviewed data utility and data security measures for each dataset type.

Duncan and Pearson (1991) suggested that masked microdata should be released instead of aggregated data (such as aggregated statistic responses to restricted queries, or tabular data) to maximize data utility and to allow for different types of analyses. Muralidhar and Sarathy (2005b) adopted this viewpoint and pointed out that accessing microdata is a requirement for data mining analysis tasks. Clifton (2003) also supported this opinion. One important category of perturbative SDL methods for microdata is data perturbation methods.

In data perturbation, a noise term is used to change original confidential attributes and generate masked ones. There are mainly two types of perturbation methods: multiplicative and additive data perturbation methods. Multiplicative data perturbation methods generate masked values by multiplying original unmasked confidential values by an error term with a mean equal to 1 and with a small variance (Kim and Winkler, 2003; Muralidhar et al., 1995). On the other hand, additive data perturbation methods add an error term with mean 0 and a specific variance or covariance matrix to the confidential attributes. Examples of the latest additive data perturbation methods are GADP (Muralidhar et al., 1999), C-GADP (Sarathy et al., 2002), IPSO (Burridge, 2003), and EGADP (Muralidhar and Sarathy, 2005b).

One of the advantages of data perturbation methods is that they automatically safeguard against *exact* value disclosure (Adam and Wortmann, 1989; Muralidhar and Sarathy, 1999). However, their ability to protect against *partial* value disclosure varies from one method to another. Dalenius (1977) defined disclosure risk by the increment in

the snooper knowledge after accessing the masked attributes. To account for the worst-case scenario, the assumption should be that the snooper has maximum knowledge about confidential attributes represented in the form of their distributions (Muralidhar and Sarathy, 2003c).

The remainder of this section is divided into three subsections. The first subsection discusses the advantages of using masking methods in developing PPE and PPDM methods. The second subsection talks about the Conditional Independence Theory for developing optimal (in terms of data utility and data security) masking methods and the practical limitation of this theory. The third discusses briefly the latest related masking methods and their optimality status. Then, in the main section to follow, we test the performance of some of these methods on the store dataset in the motivation example to see whether they can be used for PPE involving non-monotonic relationships. We then list some difficulties in developing a new PPE method for non-monotonic relationships. Finally, we conclude this chapter by compiling a list of possible PPE research questions.

II.2.1. Advantages of Using Masking Methods for PPE

Data can be classified as aggregate data (tabular summaries) and microdata (full data). While access limitation methods for microdata can achieve some security goals, full access to microdata is important for DM. Actually, one of the biggest barriers facing data mining projects today is the "*inability to release data*" due to privacy concerns (Clifton, 2003). Masking methods are a sound choice since they allow the release of microdata without any access restriction. They have rich literature and are built on a solid theoretical basis (cf. Adam and Wortmann, 1989; Willenborg and Waal, 2001). Equally important, they are rigorous techniques for maintaining privacy and confidentiality while

maximizing data utility (mainly for statistical analysis) (Muralidhar et al., 1999; Muralidhar and Sarathy, 2003a; 2005a; Sarathy et al., 2002). Actually, masking techniques automatically provide protection against exact value disclosure (Adam and Wortmann, 1989; Muralidhar and Sarathy, 1999). From another perspective, they are good starting points for practically implementing the Data-Centric Approach (DCA) in PPDM (Al-Ahmadi et al., 2004), as discussed earlier.

II.2.2. Optimal Masking Methods

In this section, we will talk about the optimality requirements of masking methods based on the *Conditional Independence Theory*. Then we will briefly discuss its practical limitations.

II.2.1.2 Conditional Independence Theory

The ultimate goals for any masking method are: (a) to maximize data security of the masked dataset (minimize disclosure risks: both value and identity risks), and (b) to maximize the data utility (minimize information loss or maximize data accuracy) of datasets after the protection. Many prior perturbation methods were often developed based on the concept that there is always a *trade-off* between these two goals. Examples include simple additive data perturbation SADP method (Traub et al., 1984) and correlated-noise additive data perturbation CADP known as Kim's method (Fuller, 1993; Kim, 1986; Tendick, 1991).

Muralidhar and Sarathy (2003c) developed a theoretical framework for perturbation methods, called the Conditional Independence Theory, which eliminates the need for a trade-off between data utility and disclosure risks (beyond a specific level defined by the characteristics of the data). The framework suggests that the perturbed

values **Y** should be generated (a) only from the conditional distribution $f(\mathbf{X}|\mathbf{S})$ and (b) *independently* from the original confidential attributes **X** given the non-confidential attribute **S**. The importance of these conditions, once they are met, lies in the fact that both maximum data utility and data security (minimum disclosure risk) requirements will be automatically and simultaneously satisfied. Muralidhar and Sarathy (2003c) detailed these requirements of the conditional independence theory and their consequences in terms of marginal, joint and conditional distributions. The use of the conditional distribution $f(\mathbf{X}|\mathbf{S})$ in the data masking literature is not new, and it has been examined in different contexts: multiple imputation (Little, 1993; Rubin, 1993), categorical data (Fienberg et al., 1998), and disclosure risk measures (Willenborg and Waal, 2001).

Muralidhar and Sarathy (2003c) proved that the two required conditions of the conditional independence theory, once met, satisfy both data utility and disclosure risk requirements as follows. The first condition is that perturbed values **Y** should be generated from the conditional distribution $f(\mathbf{X}|\mathbf{S})$:

$$\mathbf{Y}: \quad f_{\mathbf{X}|\mathbf{S}}(\mathbf{X}|\mathbf{S}). \tag{2.1}$$

Second, **Y** should be independent of **X** given **S**:

$$f_{\mathbf{X},\mathbf{Y}|\mathbf{S}}(\mathbf{X},\mathbf{Y}|\mathbf{S}) = f_{\mathbf{X}|\mathbf{S}}(\mathbf{X}|\mathbf{S}) f_{\mathbf{Y}|\mathbf{S}}(\mathbf{Y}|\mathbf{S}). \tag{2.2}$$

In terms of *data utility requirements*, two characteristics from the original dataset should be preserved in the perturbed dataset. First, the joint distribution between the non-confidential attributes **S** and the perturbed attributes **Y** should equal the joint distribution between the non-confidential attributes **S** and the confidential attributes **X**: $f(\mathbf{Y},\mathbf{S}) = f(\mathbf{X},\mathbf{S})$. The theory of conditional independence assures this since **Y** is generated from $f(\mathbf{X}|\mathbf{S})$, $f(\mathbf{Y}|\mathbf{S}) = f(\mathbf{X}|\mathbf{S})$, and therefore:

$$f(\mathbf{Y},\mathbf{S}) = f(\mathbf{Y}|\mathbf{S})f(\mathbf{S}) = f(\mathbf{X}|\mathbf{S})f(\mathbf{S}) = f(\mathbf{X},\mathbf{S}). \quad (2.3)$$

Second, the marginal distribution of the perturbed attributes **Y** should be the same as the marginal distribution of the confidential attributes **X**: $f(\mathbf{Y}) = f(\mathbf{X})$. Again, the theory of conditional independence satisfies this requirement. This can be seen using (2.3):

$$f(\mathbf{Y}) = \int_{\mathbf{S}} f(\mathbf{Y},\mathbf{S})d\mathbf{s} = \int_{\mathbf{S}} f(\mathbf{X},\mathbf{S})d\mathbf{s} = f(\mathbf{X}). \quad (2.4)$$

Preserving the joint distribution of the perturbed dataset to be the same as that of the original dataset (i.e. $f(\mathbf{Y},\mathbf{S}) = f(\mathbf{X},\mathbf{S})$) maintains all relationships among variables. This suggests that applying any analysis that depends on relationships among variables in the perturbed dataset should provide similar (if not exact) results when applying the same analysis method on the original dataset. In addition, maintaining identical marginal distributions of the perturbed and the original datasets produces the same results for univariate analysis of the original and masked datasets for each corresponding variable.

In terms of *data security requirements*, there are two assumptions regarding intruders: (a) they have full access to the non-confidential attributes **S**, and (b) they have the maximum knowledge about confidential attributes: the conditional distribution of confidential attributes **X** conditioned on non-confidential attributes **S**: $f(\mathbf{X}|\mathbf{S})$. Both assumptions collectively account for the worst-case scenario: that the snooper does a good job in trying to breach the confidentiality and privacy of individuals in masked datasets even before their release. In this context, disclosure risk is defined by the increase in the snooper ability (or reduction of his uncertainty) to predict confidential attributes once the masked dataset is released (Dalenius, 1977; Duncan and Lambert, 1986). Thus, accessing the released masked dataset (i.e. **S** and **Y**) might cause an increase

in snooper prediction power since there is now more information (i.e. **Y**) to use. However, the theory of conditional independence requires that **Y** is independent of **X** given **S,** and this makes the following relationship true: $f(\mathbf{X}|\mathbf{S},\mathbf{Y}) = f(\mathbf{X}|\mathbf{S})$.

This can be proven by using (2.2):

$$LHS: f(\mathbf{X}|\mathbf{S},\mathbf{Y}) = \frac{f(\mathbf{S},\mathbf{X},\mathbf{Y})}{f(\mathbf{S},\mathbf{Y})} = \frac{f(\mathbf{S},\mathbf{X},\mathbf{Y})}{f(\mathbf{Y}|\mathbf{S})f(\mathbf{S})} =$$

$$\frac{f(\mathbf{X},\mathbf{Y}|\mathbf{S})}{f(\mathbf{Y}|\mathbf{S})} = \frac{f(\mathbf{X}|\mathbf{S})f(\mathbf{Y}|\mathbf{S})}{f(\mathbf{Y}|\mathbf{S})} = f(\mathbf{X}|\mathbf{S}): RHS.$$

Therefore, snoopers do not gain any incremental knowledge (beyond what they already know about confidential attributes **X** from non-confidential attributes **S**) by accessing masked attributes **Y**.

As we have seen, the Conditional Independence Theory, once met, guarantees maximum data utility and data security. For an interesting discussion about whether the Conditional Independence Theory is sufficient to reach the optimal level of data utility and data security, refer to comments made by Polettini and Stander (2003) and the rejoinder made by Muralidhar and Sarathy (2003b).

II.2.2.2 Practical Limitation of the Optimal Procedure

Utilizing the concept of conditional independence to develop masking methods for every possible dataset (including non-normal ones) proves to be a difficult problem (Muralidhar and Sarathy, 2003c). The main reason is the difficulty in learning the true multivariate conditional density function *f*(**X**|**S**). In practice, *f*(**X**|**S**) is usually unknowable and difficult to estimate.

This limits the practical applicability of the Conditional Independence Theory in developing optimal masking methods. Nevertheless, some special cases (such as the case of multivariate normal data) have been addressed in the SDL literature in which optimal masking methods have been built based on this theory. In the next subsection, we will discuss some of the latest (PPE-related) masking methods. In addition, we will explain which masking methods (and in which settings) satisfy the Conditional Independence Theory.

II.2.3. Recent Masking Methods

Muralidhar et al. (1999) proposed a new perturbation method called the **G**eneral **A**dditive **D**ata **P**erturbation (GADP) method that avoids the problems in early perturbation methods. GADP is mainly designed for maintaining linear relationships. The ideal situation is when datasets are normally distributed, which assures that all relationships among variables are linear (Kotz et al., 2000). Burridge (2003) and Muralidhar and Sarathy (2005b) reported that GADP experiences *sampling error* that affects its performance when it is applied to small datasets. Burridge (2003) suggested a new perturbation method called **I**nformation **P**reserving **S**tatistical **O**bfuscation (IPSO) method, which does not suffer from the same problem. The idea behind IPSO lies in the concept of capturing the *sufficient statistics* (Anderson, 2003; Johnson and Wichern, 1998; Lehmann and Casella, 1998) from original and normally distributed datasets, and utilizing them to produce perturbed values in masked datasets. Muralidhar and Sarathy (2005b) recognized that although IPSO does a good job when dealing with small datasets, it has a problem in data security. They proposed a variant from GADP method that uses sufficient statistics to avoid sampling error in small datasets. The new approach is called

Enhanced (or **E**xact) **G**eneral **A**dditive **D**ata **P**erturbation (EGADP) method. EGADP maintains *exact* linear relationships, even in small data.

All the above three masking methods are proposed for linear relationships. Sarathy et al. (2002) proposed a new method called the **G**eneral **A**dditive **D**ata **P**erturbation method: the **C**opula approach (C-GADP), which can maintain monotonic (linear or nonlinear) pairwise relationships. A copula is a function that joins a group of functions of marginals into one multivariate (copula) distribution (Jouini and Clemen, 1996; Nelsen, 1999; Schweizer, 1991; Sklar, 1959). C-GADP utilizes a multivariate normal copula (Clemen and Reilly, 1999; Joe, 1997) to transform non-normal distributions into normal ones, which capture the monotonic dependence structure (rank-order correlation) of original datasets. Then GADP or EGADP can be applied on these transformed normal datasets to produce (normally distributed) masked attributes. Finally, masked attributes are transformed to their original marginals.

Muralidhar and Sarathy (2003a; 2005a) proposed a new masking method called data shuffling, which combines of the advantages of perturbation methods and data-swapping methods. Like data perturbation methods, the new approach maximizes data security and data utility in the case of linear or monotonic nonlinear relationships. Like data-swapping methods, it maintains the marginal distributions of the masked confidential attributes *exactly* as the marginal distributions of the original confidential attributes. In addition, data shuffling has the advantage of efficient implementation by utilizing "*only rank order data,*" (Muralidhar and Sarathy, 2005a, pp.1).

There are two approaches to implementing data shuffling: parametric and non-parametric. The parametric approach is similar to the C-GADP perturbation method

(Sarathy et al., 2002) with an additional step: rank ordering the original values of the confidential attributes **X** according to the rank-order of the perturbed values **Y** to get the final shuffled values. The non-parametric data shuffling approach has the advantage of bypassing the problem of identifying unknown marginal distributions. It utilizes the empirical CDF distribution as an estimation mechanism of unknown marginal distributions.

From an *optimality* perspective, GADP and EGADP applied to multivariate normal data satisfy the Conditional Independence Theory. Therefore, they are *optimal* in terms of maximizing data utility and data security. Although IPSO provides *ideal* data utility by preserving exactly all (linear) relationships in normal data, it has a problem in the data security aspect since masked attributes **Y** are not generated independently from $f(\mathbf{X}|\mathbf{S})$ (Muralidhar and Sarathy, 2005b).

When applied to non-normal data, GADP and EGADP provide *ideal* security but not ideal data utility. They only reproduce linear relationships in masked data. This is generally adequate for a majority of statistical analyses of masked data, but not for data mining. Many studies (Fuller, 1993; Sarathy et al., 2002; Sullivan and Fuller, 1989) concluded that applying additive perturbation methods (including the above sophisticated methods) on non-normal datasets reduces the accuracy and utility of masked datasets because nonlinear relationships are not preserved. Hence, GADP and EGADP may not be suitable for general data mining.

Because of the known difficulty of estimating the true conditional density function $f(\mathbf{X}|\mathbf{S})$ in non-normal datasets, developing *optimal* masking methods based on the Conditional Independence Theory is not feasible in this case (Muralidhar and Sarathy,

2003c). Nevertheless, preserving some characteristics of non-normal datasets is possible. For example, C-GADP and data shuffling can maintain monotonic (linear or nonlinear) pairwise relationships. Both C-GADP and data shuffling provide *optimal* data security but not optimal data utility because they do "*not utilize the true conditional density* $f(\mathbf{X}|\mathbf{S})$," (Muralidhar and Sarathy, 2003c). Instead, they approximate $f(\mathbf{X}|\mathbf{S})$ with a multi-variate normal copula distribution. Nevertheless, the level of provided data utility may be sufficient for some data mining applications.

II.3. Impact of Current Masking Methods on Non-Monotonic Relationships

Lechner and Pohlmeier (2004) investigated how some *disclosure limitation masking* methods affect estimating *nonlinear* data models using econometric estimation techniques. Two disclosure limitation methods were studied: blanking and noise addition. The researchers used three econometric estimation techniques: the SIMEX method, the calibration method, and a semi-parametric sample selectivity estimator. Lechner and Pohlmeier (2004) concluded that the disclosure limitations techniques, at varying degrees, make it difficult to get nonlinear models and other estimates from masked datasets similar to those obtainable from original datasets. Lechner and Pohlmeier (2004) further pointed out that while the effects of masking methods on the estimators of linear (regression) models have been well-understood, "*nonlinear regression techniques coping with implications of data masking are still at their infancy.*"

We picked two advanced data masking methods that have been proven to provide maximum data utility and data security: EGADP (Muralidhar and Sarathy, 2005b) for linear relationships, and (C-EGADP-based) data shuffling (Muralidhar and Sarathy,

2003a; 2005a) for monotonic nonlinear relationships. Neither has been used in the PPDM arena. These methods were applied to the store dataset (see Table 1I.3). While they provide good security measures for the store dataset, these methods do not produce useful masked datasets for the required regression tasks because of the existence of non-monotonic relationships. For example, the relationship between age and expenditure variables (see Figure 2 in Section I.3) is not preserved in the masked data. The destructive effect of these methods on this type of relationship is demonstrated in Figure 4. Sarathy et al. (2002) pointed out the limitations of these methods in the case of non-monotonic relationships.

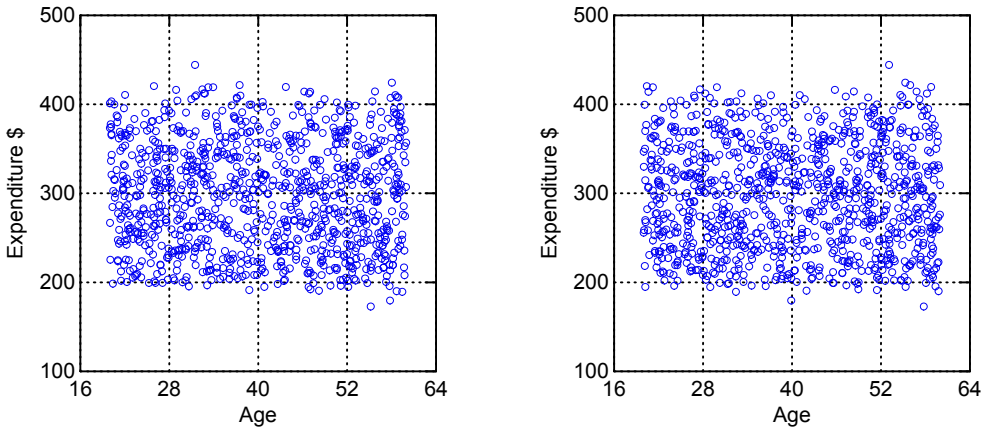


Figure 4. Impact of current masking methods (EGADP and data shuffling) on non-monotonic relationships

Clearly, there are no masking methods capable of preserving non-monotonic relationships in masked datasets as they exist in original datasets. However, the literature on SDL and data perturbation methods provides sophisticated masking methods for maintaining monotonic relationships (linear and nonlinear) and achieving high levels (sometimes optimal) of data utility and data security. Our goal is to adapt and extend some of these methods for the case of non-monotonic relationships. In addition, we want

to develop or adopt some measures for data utility and data security. These measures will be used to investigate the effectiveness of our proposed methods. Next, we will discuss briefly some challenges in developing a PPE masking method for non-monotonic relationships.

II.3.1. Challenges in Developing Practical PPE Masking Methods for Non-Monotonic Relationships

SDL masking methods were developed based on the concept that maintaining some aggregate correlation-based measures of original datasets in masked datasets would preserve the corresponding relationships captured by the aggregate measures. For example, GADP (Muralidhar et al., 1999), IPSO (Burridge, 2003) and EGADP (Muralidhar and Sarathy, 2005b) try to maintain the Pearson correlation matrix (along mean and covariance matrix). This ensures that all original *linear* relationships are reproduced in the masked datasets. On the other hand, C-GADP (Sarathy et al., 2002) and data shuffling (Muralidhar and Sarathy, 2003a; 2005a) try to maintain Spearman rank-order correlation matrix. This ensures that all original *monotonic nonlinear* (and *linear,* in the case of multivariate normal datasets) relationships are reproduced in the masked datasets.

One challenge in developing a new PPE-masking approach for non-monotonic relationships is that there is no aggregate measure capable of capturing non-monotonic relationships that we can use to reproduce this type of relationship in masked datasets. Thus, a different strategy is needed for developing the new approach. Since PPE, and regression methods in general, are about quantifying the relationships among variables (Rud, 2001), “relationships” will be the basis for developing our new PPE approach. In

the first section in Chapter III, we will provide a formal regression-related definition of relationships in PPE.

Another challenge that arises is the need to test the effectiveness of the newly developed PPE method in terms of data utility and data security. Data security measures are based on general concepts and, therefore, existing security measures can be used. However, data utility measures are specific to the task at hand (Domingo-Ferrer et al., 2001; 2003). For PPE, it is necessary to maintain relationships among the variables in masked datasets in the manner they exist in original datasets. For monotonic relationships, data utility (in masking techniques) can be easily checked by comparing the relative aggregate measures of the original and masked datasets. Similarity between these aggregate measures declares the success of the masking methods in terms of data utility. Again, there is no aggregate measure for non-monotonic relationships; therefore, we need to take a different path in developing suitable data utility measures. Once more, we will use the concept of *relationships* as the basis for our data utility measures. The new proposed data utility measures, along with adopted security measures, are discussed in Chapter V.

II.4. Research Questions

This study raises and addresses the following questions:

- **Question 1**: Since there is no aggregate measure for non-monotonic relationships, how should relationships be defined (and later measured) in PPE and PPDM?
- **Question 2:** Is it possible to develop/adapt new masking methods in PPE to maintain non-monotonic relationships while maximizing data utility and data security?

- **Question 3:** What is the theoretical basis for these masking methods?
- **Question 4:** What is data utility in PPE? What is data security in PPE? How can we quantify data utility and data security in PPE (success measures)?
- **Question 5:** Can the new methods preserve other types of relationships (linear or monotonic nonlinear)?
- **Question 6:** What are the assumptions of these new methods? What will happen if these assumptions are violated?
- **Question 7:** How can we establish the validity of the new methods (experiment design)?
- **Question 8:** How well do these methods compare with existing methods?

CHAPTER

III. RELATIONSHIP-BASED MASKING: THEORETICAL BASIS

The prvious chapter discussed the need for maintaining and reproducing the different types of possible relationships in masked datasets, as they exist in original datasets. In addition, it was seen that existing masking methods do not preserve all types of relationships in masked datasets. This chapter introduces four interrelated Relationship-Based Masking (RBM) methods for reproducing different types of relationships in masked datasets. The RBM approach is a three-stage approach: identifying relationships, analyzing residuals, and applying masking. These three stages are interrelated. The theoretical basis for the RBM methods lies mainly in the first two stages (relationships and residuals). Hence, the focus of this chapter is on the first two stages (and the tools related to them) while the subject of the following chapter is mainly the third stage (along with the first two stages).

First, a formal definition for a relationship in the estimation and regression context is needed. In addition, we need an effective, proven way to learn and capture different possible types of relationships that may exist in a dataset. It is also critical to understand the roles that residuals play in defining and guiding data utility and data security requirements to develop effective masking methods.

Section III.1 draws attention to the formal statistical definitions of a relationship and discusses the one that is usually used in estimation and regression modeling. In

addition, this section specifies the class[1] of relationships that the RBM approach tries to learn from original datasets, and whether it is data-utility or data-security driven. Section III.2 investigates the capability of Artificial Neural Networks (ANN) in capturing relationships including non-monotonic ones. Section III.3 concludes this chapter by investigating the roles of residuals **r** left after fitting confidential attributes **X** (as dependent variables) to non-confidential attributes **S** (as independent variables) in defining and guiding data utility and data security requirements.

III.1. Conditional Expectation: The Formal Definition of Relationships in PPE

Macnaughton (2002) compiles and discusses *seven* definitions of a relationship between two random variables. One of these definitions that is particularly suited to this study is that there is a relationship between variables x and y, if the values of y can be expressed as:

$$y = f(x) + e \tag{3.1}$$

"*where e is usually viewed as being independent of x*" and represents the random component of the relationship. $f(x)$ is strictly a real (deterministic) mathematical function that represents the exact component of the relationship with no random element involved. A strict mathematical function means that it maps from *one or more* values of x to *only* one value of y (*but not* to more than one value of y). For example, the function $y = f(x) = x^2$ is a mathematical function because it maps multiple values of x such as (-2,2) to

[1] We may use the term "*class*" when we refer to a relationship based on the confidentiality level of its involved dependent and independent variables. Examples include the *class* of relationships between **X** and **S** and the *class* of relationships among confidential attributes **X** or $E(X_i|X_j)$. In addition, we may use the term "*type*" or "*shape*" when we talk about the shape of a relationship: linear or monotonic nonlinear, or non-monotonic.

exactly one value of *y* (4). Thus, a mathematical function is either one-to-one (1-1) mapping or many-to-one (M-1) mapping, but not one-to-many (1-M) mapping. The former two cases are called *single-valued* mapping while the latter case is called a *multi-valued* mapping (Bishop, 1995). The existence of multi-valued mapping in data, which violates the definition of a mathematical function, impacts the effectiveness of our proposed masking methods in terms of data utility, as we will discuss later.

Muralidhar and Sarathy (2005a) suggest that perturbed variables **Y** satisfy the minimum security requirements when they are generated using a function $g(\mathbf{S},e)$ of only the non-confidential attributes **S,** and a noise term *e* that is independent of the original confidential attributes **X** given the non-confidential attributes **S**. Notice the similarity of formula (3.1) with the function $g(\mathbf{S},e)$ when one chooses function *g* to be the addition of independent/orthogonal noise *e* to a random function *f* of non-confidential attributes **S**:

$$Y = g(S,e) = f(S) + e \tag{3.2}$$

where **e** is independent of **X** (or $\mathbf{e} \perp \mathbf{X}$) given **S** as a security requirement, $\perp$ denotes "orthogonal to", and $f(\mathbf{S})$ is any random function of **S**. Equation (3.2) only suggests the requirements of the data security of perturbation methods and it does not consider or guarantee data utility.

Data utility can be ensured if the function $f(\mathbf{S})$ is carefully chosen. Macnaughton (2002) suggested that the mathematical form of *f* is usually determined by data analysis, although theoretical considerations can also be taken into account. He further suggested that such an approach usually leads to choosing the form of *f* as the best estimate of the conditional expectation $E(\mathbf{Y}|\mathbf{S})$ in Equation (3.2) (note that we are using the terminology of data perturbation: **S**, **X** and **Y**). Actually, $E(\mathbf{Y}|\mathbf{S})$ cannot be calculated

initially since the masked values **Y** are not available. Nevertheless, the conditional expectation $E(\mathbf{Y}|\mathbf{S})$ is specified to be $E(\mathbf{X}|\mathbf{S})$ to maximize data utility. Thus, in our context, formula (3.2) can be expressed as:

$$\mathbf{Y} = E(\mathbf{X}|\mathbf{S}) + \mathbf{e} \tag{3.3}$$

where **e** is independent of **X** given **S** (or $\mathbf{e} \perp \mathbf{S},\mathbf{X}$) and **e** resembles the characteristics of **r** obtained from:

$$\mathbf{X} = E(\mathbf{X}|\mathbf{S}) + \mathbf{r} \tag{3.4}$$

where **r** is independent of **S** (or $\mathbf{r} \perp \mathbf{S}$) by definition. Equation (3.3) is the essence of EGADP (Muralidhar and Sarathy, 2005b). EGADP can preserve *linear* characteristics of original datasets perfectly, even for very small datasets, while providing maximum data security (Section II.3). However, it cannot, as we saw earlier, preserve *non-linear* relationships. Estimating the conditional expectations $E(\mathbf{X}|\mathbf{S})$ accurately, and accordingly the "*relationships*" between **X** and **S** by the regression definition, is critical for maintaining data utility.

We conclude this section by discussing the difference between independence and orthogonality (uncorrelated). Independence implies orthogonality; the reverse is not always true (Hunter, 1972). When we learn relationships $E(\mathbf{X}|\mathbf{S})$ from multivariate normal data, residuals **r** are always guaranteed to be independent of $E(\mathbf{X}|\mathbf{S})$ or any function of **S** (see Rhodes (1971), Property 8, pp. 692). When distributions of data are non-normal, the only condition guaranteed at all times is that residuals **r** are orthogonal (uncorrelated) to $E(\mathbf{X}|\mathbf{S})$ or any function of **S** (see Rhodes (1971), Proposition 2b, pp. 690 and Bickel and Doksum (2001), Proposition 1.4.1 (a) and (b), pp. 34). Nevertheless, "*(w)hen two variables are uncorrelated, they may be (and **often** are)*

independent,"(Schield, 1995). This means that residuals **r** are ***often*** *independent* of $E(\mathbf{X}|\mathbf{S})$ or any function of **S** regardless of the data distribution. In our discussion, we do not make a distinction between these two concepts and we use independence and orthogonality interchangeably.

III.2. Artificial Neural Networks (ANN) Approaches for Estimating Conditional Expectations

There are many approaches that can be used to learn relationships (conditional expectations) in datasets. For example, one can try to fit a polynomial of an appropriate degree to the data. However, this approach suffers from two limitations. First, it is a parametric approach that requires a pre-judgment of the relationship before trying to fit it to the data. This proves difficult especially if there is no theoretical guidance (Fisher, 1970). Second, it is affected by the curse of dimensionality (Bishop, 1995). Similarly, nonlinear regression requires the specification of the functional form before the estimation of parameters can begin. On the other hand, other approaches such as artificial neural networks (ANN) do not assume any pre-specified form of the relationship. In addition, although ANNs also face the problem of the curse of dimensionality, they are less affected compared to the polynomial approach (Bishop, 1995).

There are many characteristics of neural networks architectures that make them appealing in estimating nonlinear and non-monotonic conditional expectations. First, ANN approaches have been proven to be global function estimators of nonlinear and non-monotonic continuous functions (Bishop, 1995; Hagan et al., 1996). Second and more interestingly, they can be used to approximate the conditional expectation of the

network output **O** given the network input **I** (i.e. $E(\mathbf{O}|\mathbf{I})$) when the minimized error function is the *(mean) sum of squared errors* function (MSE) and the network is used to map input **I** to output **O** (Bishop, 1995; Saerens, 1996; 2000). Saerens (1996) suggests that this is a fundamental mathematical statistics result based on estimation theory and can be found in many books such as Deutsch (1965), Meditch (1969) and Shao (1999).

Bishop (1995) proves the above fact mathematically for the case of multilayer perceptron neural networks (MLP). Saerens (1996) does the same and emphasizes the importance of the assumption that the minimum error is indeed reached after the training. He also points out that many researchers arrive at the same conclusion. A few examples for the continuous case are White (1989) and Wan (1990). Examples for the binary case are Bourlard and Wellekens (1989), Ruck et al. (1990), Gish (1990), and Shoemaker (1991). For a general review, refer to Richard and Lippmann (1991).

There are two more interesting facts about using the MSE function for estimating the conditional expectations. First, the ability of the neural networks to learn the conditional expectations is not affected by the characteristics of the noise (e.g. normal vs. non-normal) in the data (Saerens, 1996). Second, the ability to learn the conditional expectation when using MSE is not specific to the multilayer perceptron neural networks MLP and it is independent of the neural network architecture (Bishop, 1995). Actually, this fact is even applicable in architectures that are not classified as neural networks as long as they try to minimize MSE and they succeed in approaching the real minimum (Bishop, 1995; Saerens, 1996).

(Multiple) Linear regression, which utilizes the concept of least squared errors, does a good job in estimating linear conditional expectations in normally distributed

datasets (where all relationships are linear). However, when applied to datasets containing nonlinear relationships, it does a poor job. This is because it assumes that relationships are linear and can be discovered by fitting lines that minimize the squared error. In this sense, the global minimum of the squared errors, which might be obtained by fitting some curves to the data, is not reached. As a result, (multiple) linear regression cannot be used as a mechanism to estimate conditional expectations when the relationships are not linear. This shows the importance of the assumption that the minimum error is indeed achieved after the estimation process stops. ANN algorithms do not assume any functional form that may hamper learning the true conditional expectation.

The above discussion is important to our context. We can use any neural networks architecture based on MSE to estimate the required conditional expectation of the confidential attributes **X** given the non-confidential attributes **S** (i.e. $E(\mathbf{X}|\mathbf{S})$) by mapping the input **S** to the output **X** regardless of the form of the conditional expectation (monotonic or (single-valued mapped) non-monotonic) or the characteristics of the error term. The main assumption is that the training procedure actually reaches, or at least approaches, the global error minimum.

MLP neural networks suffer from the trap of local minimums that may limit their ability to reach the global minimum of the minimized error function. Consequently, the real conditional expectation may not be learned accurately. However, there are other neural networks architectures that can be used instead and do not suffer from the same problem.

Support Vector Machine (SVM) (Vapnik, 2000) and Least Squares Support Vector Machine (LS-SVM) (Suykens et al., 2002) are different neural networks architectures (Schèolkopf and Smola, 2003). SVM and LS-SVM are kernel-based methods (Schèolkopf et al., 1999; Shawe-Taylor and Cristianini, 2004). They utilize the kernel trick to implicitly map an input space into a higher dimensional space called the feature space. Then, well-known linear (and non-linear) learning algorithms can be applied on the feature space to learn relationships and patterns. The main two tasks, among others, that can be run in the feature space are regression and classification.

SVM and LS- SVM do not suffer from local minimums that hinder optimization algorithms from reaching the global minimum of the cost function in the error space. This is because the cost function in the dual space (the main optimization space for (LS-) SVM cost functions) has a quadratic form with a unique global minimum (Suykens et al., 2002). In this study, we use LS-SVM (Suykens et al., 2002) and its Matlab toolbox implementation, called LS-SVMlab1.5 (2003), to learn conditional expectations. However, any learning mechanism that is theoretically capable of capturing and learning non-monotonic conditional expectations (relationships) can be used in our proposed Relationship-Based NM-EGADP masking approach (RBM).

III.3. The Roles of the Residuals (r)

III.3.1. Residuals Role in Defining Relationships among Attributes

The residuals **r** obtained from estimating $E(\mathbf{X}|\mathbf{S})$ (see Equation (3.4) in Section III.1) play important roles in defining two main groups (classes) of relationships: (a) relationships between non-confidential attributes **S** and confidential attributes **X** (i.e. $E(\mathbf{X}|\mathbf{S})$), and (b) relationships among confidential attributes **X** (i.e. $E(X_i|X_j)$ or $E(X_j|X_i)$).

For simplicity and without losing generality, we limit our discussion to two confidential attributes $\mathbf{X}$ (X_i and X_j, where $i, j = 1\text{K } q \text{ and } i \neq j$) and the non-confidential attributes **S**. The residuals **r** (r_i and r_j) are obtained from the following two equations:

$$X_i = E(X_i|\mathbf{S}) + r_i \tag{3.5}$$

and

$$X_j = E(X_j|\mathbf{S}) + r_j . \tag{3.6}$$

In our proposed masking methods, masked variables **Y** (Y_i and Y_j) are generated as follows:

$$\begin{aligned} Y_i &= E(Y_i|\mathbf{S}) + e_i \\ &= E(X_i|\mathbf{S}) + e_i \end{aligned} \tag{3.7}$$

$$\begin{aligned} Y_j &= E(Y_j|\mathbf{S}) + e_j \\ &= E(X_j|\mathbf{S}) + e_j \end{aligned} \tag{3.8}$$

where **e** (e_i and e_j) are noise terms that try to preserve certain characteristics of the residuals **r** (r_i and r_j) to maximize data utility. In addition, the noise terms **e** are generated to satisfy security requirements and to avoid providing extra information about confidential attributes **X** beyond what is known about them from the non-confidential attributes **S**.

In the proposed masking methods, we specify $E(Y_i|\mathbf{S}) = E(X_i|\mathbf{S})$ and $E(Y_j|\mathbf{S}) = E(X_j|\mathbf{S})$ (see Equations (3.7) and (3.8)) so that relationships between confidential attributes **X** (X_i and X_j) and non-confidential attributes **S** are automatically reproduced in masked datasets between masked attribute Y (Y_i and Y_j) and non-confidential attributes **S**. The only required condition is that **e** should be orthogonal to **S** or any function of **S**, similar to residuals **r**. Except for the orthogonality, the added noise terms **e**

are not required to mimic every aspect or characteristic of the residuals **r** to preserve this class of relationships.

However, relationships *among* confidential attributes **X** ($E(X_i|X_j)$ and $E(X_j|X_i)$)[2] are not directly reproduced in masked datasets by only fixing conditional expectations $E(\mathbf{X}|\mathbf{S})$ while masking. In addition to the role of conditional expectations $E(\mathbf{X}|\mathbf{S})$, $E(X_i|X_j)$ depends heavily on the characteristics of the added noise (e_i and e_j). Ideally, we want the characteristics of the added noise **e** (e_i and e_j) to be exactly the same as the characteristics of original residuals **r** (r_i and r_j) in terms of two related requirements: (a) orthogonality (mainly for maintaining relationships between **X** and **S** in masked datasets), and (b) joint distribution (mainly for maintaining relationships among confidential attributes **X** in masked datasets). Although the former is achievable, the latter is not always feasible and achievable. Therefore, we want to get as close as possible to the ideal case to maximize the data utility.

But before we discuss these data utility requirements further, we explore whether other possible types of residuals that result from different forms of estimation or (conditional) expectations can be used in masking methods to maintain relationships (especially among confidential attributes **X**) without violating other requirements. First, let us consider the set of residuals resulted from subtracting the expectations of the confidential attributes $E(\mathbf{X})$ from the confidential attributes **X**. In this case, all relationships among **X** are completely captured by this set of residuals. However, this set

[2] For simplicity of discussion, we shall only talk about $E(X_i|X_j)$ initially, with the assumption that it is a single-valued mapping.

of residuals captures *none* of the relationships between **X** and **S,** and there is no easy way to relate them to conditional expectations $E(\mathbf{X}|\mathbf{S})$.

Another idea is to use the residuals from fitting $E(X_i|\mathbf{S}X_j)$. One problem with these residuals is that they do not satisfy an important security requirement: we cannot use any function of confidential attributes **X** directly in generating masked attributes **Y** (Muralidhar and Sarathy, 2003c; 2006b). Clearly, residuals obtained from $E(X_i| X_j)$ are not suitable either since they inherit the problems of the last two types of residuals: they do not carry any information about non-confidential attributes **S**, and they violate the security requirement of avoiding conditioning on confidential attributes **X**.

Therefore, we can only use functions of non-confidential attributes **S** as specified in Equations (3.5) and (3.6). In this case, $E(X_i|X_j)$ is captured by the two (orthogonal) components in Equations (3.5) and (3.6): $E(\mathbf{X}|\mathbf{S})$ and **r**. Since $E(\mathbf{X}|\mathbf{S})$ is the fixed part during masking (refer to Equations (3.7) and (3.8)), the added noise set **e** (i.e. the dynamic or changed part) should mimic the characteristics of the residual set **r** to maintain the relationships among masked attributes **Y** as they exist among confidential attributes **X**. There are two dimensions for similarity: orthogonality and joint distributions.

In Equations (3.5) and (3.6), residuals **r** (r_i and r_j) are orthogonal to **S** and any function of **S** when the conditional expectations $E(\mathbf{X}|\mathbf{S})$ are correctly estimated (Bickel and Doksum, 2001, Proposition 1.4.1 (a) and (b), pp. 34; Rhodes, 1971, Proposition 2b, pp. 690). Notice that both $E(X_i|\mathbf{S})$ and $E(X_j|\mathbf{S})$ are functions of **S**. This is the main concept in mean square estimation. It is called "the orthogonality principle" (Gray and Davisson, 2004; Papoulis and Pillai, 2002). Interestingly, the orthogonality principle generally holds

true when the conditional expectations (e.g. $E(X_i|\mathbf{S})$ and $E(X_j|\mathbf{S})$) are nonlinear functions of the data (here it is **S**) and it is called "*Nonlinear Orthogonality Rule*" (Papoulis and Pillai, 2002).

From a data utility perspective, for **e** to resemble the characteristics of **r**, **e** should be orthogonal to **S** or any function of **S**. Otherwise, relationships between **X** and **S** will not be reproduced in masked datasets between **Y** and **S**. The orthogonality principle also has a role to play in determining $E(X_i|X_j)$. Since the residuals **r** (r_i and r_j) are orthogonal to the conditional expectations $E(\mathbf{X}|\mathbf{S})$ (both $E(X_i|\mathbf{S})$ and $E(X_j|\mathbf{S})$), $E(X_i|X_j)$ is defined *separately* by the relationships between $E(X_i|\mathbf{S})$ and $E(X_i|\mathbf{S})$, and by the relationships between r_i and r_j. Notice also that the conditional expectations $E(X_i|\mathbf{S})$ and $E(X_j|\mathbf{S})$ are used in generating the masked variables **Y**. In addition, the added noise terms **e** are required to be orthogonal to these conditional expectations, similar to **r**. Therefore, the similarity of the characteristics of relationships between e_i and e_j and the characteristics of relationships between r_i and r_j determine how well the relationships between masked attributes **Y** (Y_i and Y_j) are maintained as they exist between original attributes **X** (X_i and X_j).

Ideally, we want the marginal and joint distributions of added orthogonal noise terms **e** to be the same as the marginal and joint distributions of residuals **r**. Statistically speaking, this means $f(e_i) = f(r_i)$, $f(e_j) = f(r_j)$ and $f(e_i, e_j) = f(r_i, r_j)$. Although marginal distributions can be preserved even when they are not distributed according to well-known standard distributions using the concept of rank-based replacement "shuffling" (Muralidhar and Sarathy, 2003a; 2006b), joint distributions are more difficult, if not impossible, to preserve since they are usually unknowable in

practice. The only exception might be when residuals **r** is multivariate normally distributed. Hence, some approximation mechanisms for joint distributions are needed.

As discussed earlier, the relationships among confidential attributes are explained by two orthogonal components: $E(\mathbf{X}|\mathbf{S})$ and **r**. The greater the variation and pattern among confidential attributes **X** explained by $E(\mathbf{X}|\mathbf{S})$, the less variation and pattern among confidential attributes **X** explained by **r**. In other words, the relationships among confidential attributes **X** can be very difficult (i.e. non-monotonic relationships) to capture directly at the attributes level (especially since we cannot condition on confidential attributes for security reasons) and to reproduce using existing masking methods as we saw earlier. Nevertheless, dividing the relationships between **X** and **S** into conditional exceptions and residuals can simplify the situation. First, this helps us to automatically reproduce relationships between non-confidential attributes **S** and confidential attributes **X** in masked datasets. Second, based on the variation and pattern explained by $E(\mathbf{X}|\mathbf{S})$, what is left in **r** can be simpler relationships (monotonic nonlinear, linear, or even no relationships) even when the relationships among confidential attributes **X** are non-monotonic.

When the relationships in **r** are simple relationships, joint distributions of residuals **r** can be approximated using some approximation mechanisms used by the adapted methods at the attributes level. Most of the current masking methods are correlation-based. They try to maintain and reproduce covariance matrices of original datasets in masked dataset. In our adaptation of these masking methods, we try to reproduce the covariance matrices at the level of residuals **r**. In this context, we derive a very important result:

$$Cov(X_i, X_j) = Cov(E(X_i|\mathbf{S}), E(X_j|\mathbf{S})) + Cov(r_i, r_j). \quad (3.9)$$

This result shows that the RBM approach always works when the relationships among **X** are linear, regardless of the types of relationships between **X** and **S**: linear, monotonic nonlinear, or non-monotonic. It might also suggest that relationships among **X** can be preserved regardless of their types (monotonic or non-monotonic) as long as the patterns and relationships among residuals **r** are simple (linear or no relationships). This is because simple patterns among residuals **r** may suggest that the more complex pattern among **X** is captured by the relationships among $E(\mathbf{X}|\mathrm{S})$ (i.e. $E(X_i|\mathrm{S})$ and $E(X_j|\mathrm{S})$). Moreover, these simple patterns among **r** can be preserved by just replicating their covariance matrix among the independent noise. The proof of this important result (i.e. Equation (3.9)) is presented in Appendix B for both the case of multivariate normal data and the general case.

To recap, there are four possible types for the relationships among residuals **r**: (a) no relationships (i.e. residuals are orthogonal to each other), (b) linear relationships, (c) monotonic nonlinear relationships, and (d) non-monotonic relationships. Regardless of the type of relationships among confidential attributes **X** (including non-monotonic relationships), when relationships between conditional expectations $E(\mathbf{X}|\mathbf{S})$ ($E(X_i|\mathbf{S})$ and $E(X_j|\mathbf{S})$) explain most of the variation and pattern in the relationships among confidential attributes **X**, the remaining variation and pattern explained by relationships among residuals **r** can be simpler (monotonic linear or no relationships). In this case, when we add orthogonal noise terms **e** with the same (Pearson-based) covariance matrix of residuals **r** to conditional expectations $E(\mathbf{X}|\mathbf{S})$, we approximate the joint distribution of **r** and replicate it in **e**. This will cover cases (a) and (b), and approximate (c) if it slightly

deviates from linearity. Our proposed masking methods do not cover the last case (i.e. (d)).

III.3.2. Role of Residuals in Guiding Security Requirements

In original datasets, the conditional expectations $E(\mathbf{X}|\mathbf{S})$ are the deterministic component of the relationships between the non-confidential attributes **S** and confidential attributes **X** while the set of residuals **r** is the random component. In addition, the variance of a confidential attribute X_i can be written as (see Proposition 1.4.1 (c), pp. 34 in (Bickel and Doksum, 2001)):

$$Var(X_i) = Var(E(X_i|\mathbf{S})) + Var(r_i). \tag{3.10}$$

In the RBM approach, the set of residuals **r** is the main factor in defining the characteristics of original datasets in terms of security. For example, when there are no residuals (i.e. $Var(\mathbf{r}) = 0$) in original datasets, there is no security in the data with which to begin. In this case, releasing any data should be avoided.

When there is a random component (i.e. residual set **r** with $Var(\mathbf{r}) > 0$), a possible secure random noise **e** can be generated and added to $E(\mathbf{X}|\mathbf{S})$ to generate masked datasets. This secure noise set **e** is generated to be orthogonal to the residuals set **r** given **S**. In other words, **e** should be orthogonal to **X** given **S**. The orthogonality condition ensures that **e** does not provide more information on confidential attributes beyond what is intended originally and specified by the characteristics of original datasets. Notice that the greater the variation residuals set **r** has, the greater the security.

Sometimes, the amount of variation in the residuals **r** might not be enough to allow the RBM masking methods to work effectively. In other words, the variation in residuals **r** may explain a very small portion of the variation in the confidential attributes

X compared to the variation explained by the deterministic part $E(\mathbf{X}|\mathbf{S})$. In this case, the NM-EGADP masking methods do *not* provide effective protection. We suggest the security index (*SI*) measure for variable X_i to assess whether residuals r_i explain enough of the variation in X_i for effective masking:

$$SI(X_i) = \frac{Var(r_i)}{Var(X_i)} \times 100 \tag{3.11}$$

where 0 percent represents no security at all to begin with since X_i is completely a deterministic function of **S** (i.e. $X_i = E(X_i|\mathbf{S})$), and 100 percent represents a complete random relationship where X_i and **S** are independent and $E(X_i|\mathbf{S})$ reduces to $E(X_i)$.

Therefore, after we estimate relationships between **X** and **S**, we need to calculate this index for every confidential attribute. The acceptable level of *SI* can be different for every X_i. Nevertheless, the *SI* measure should represent a good balance between a useful relationship (between X_i and **S**) that we want to preserve and sufficient variation in the residuals to enable the RBM methods to work effectively.

Nevertheless, the sensitivity of confidential attributes along with the amount of variation in every confidential attribute X_i also plays an important role in determining the applicability of the RBM approach. For example, a variance of 50 (i.e. $Var(X_i)$) might be considered *large* when X_i represents *age*, while a variance of 500 might be considered *small* when X_i represents *annual salary* in an enterprise. In the latter, the variation in the confidential attribute *annual salary* is not enough to mask *effectively* using the RBM approach regardless of what *SI* returns. Hence, the characteristics of datasets and their confidential attributes should be evaluated case by case based on both the sensitivity and variation of confidential attributes **X** as well as the security index (*SI*) before applying the RBM approach.

When the variation in confidential attributes is enough for their sensitivity level to be masked effectively, but the *SI* measures are low, original datasets should not be masked and released. Nevertheless, some compromises are possible. The following are just two examples of possible compromises. We may add independent noise **e** with more variance than the variance of original residuals **r**. This may affect the relationships among confidential attributes. However, this approach will not affect the relationships between **X** and **S**. In addition, removing one or a few non-confidential attributes **S** may allow for more variance in the residuals **r**. In this case, one may try to remove the least important non-confidential attribute **S** and assess its impact on **r**.

III.3.3. Summary

In summary, we want **e** to mimic the characteristics of residuals **r** to maximize *data utility*, including the joint distribution of **r** and their orthogonality to **S** and any function of **S**. Although satisfying orthogonality is achievable, as will be seen when we discuss masking methods implementation, reproducing the joint distribution of **r** in **e** is more difficult, if not impossible, except for some special cases. One special case is that **r** is a multivariate normal. Hence, some approximation mechanisms are needed. Regardless of the relationships among confidential attributes **X**, relationships among the residuals **r** can be independent, linear, monotonic nonlinear or non-monotonic. We use a covariance-based approximation mechanism to cover at least the first two situations. From a *data security* perspective and in addition to the above data utility requirements, the added noise **e** should be orthogonal to **X** given **S**.

CHAPTER

IV. IMPLEMENTATION OF RELATIONSHIP-BASED MASKING

This chapter builds upon the last two chapters and introduces the algorithmic framework of the *four* Relationship-Based NM-EGADP[3] (Non-Monotonic EGADP) masking methods by discussing the implementation of two of them. We also briefly explain the main difference between the four Relationship-Based NM-EGADP masking methods. We conclude this chapter by discussing the assumptions behind these masking methods.

IV.1. NM-EGADP Perturbation and NM-EGADP Shuffling Masking Methods

In this study, we propose four interrelated, adapted masking methods that have the ability to capture and maintain non-monotonic relationships besides other relationships (i.e. monotonic relationships: linear or nonlinear) in masked datasets once the assumptions are met. To the best of our knowledge, no PPE or masking method (perturbation or shuffling) has been developed to tackle this issue. In addition, the new proposed methods have the advantage of satisfying data security while providing data utility for PPE.

---

[3] We may use RBM, NM-EGADP, or RBM NM-EGADP (abbreviated as shown or the full name) to refer to the approach we suggest in this study. These terms are interchangeable.

The four masking methods are numbered in ascending order. We abbreviate them as Method 1 to Method 4. Since they are very similar, we only present and discuss two of them in this subsection: NM-EGADP Perturbation (Method 1) and its variant NM-EGADP Shuffling[4] (Method 2). We briefly explain the main difference between the four Relationship-Based NM-EGADP masking methods at the end of this subsection. More discussion and a list of all four algorithms for proposed masking approaches can be found in Appendix C.

"NM-EGADP" stands for nonlinear **N**on-**M**onotonic **EGADP** approach. The "NM" part of the methods' names is used to indicate ability of these new adapted methods to preserve *nonlinear non-monotonic relationships.* The "EGADP" part in the name is used to point to the similarity of the general methodological framework among these methods and the EGADP masking method (Muralidhar and Sarathy, 2005b). Although it is desirable that masking methods maintain linear measures such as correlation and covariance matrices *exactly* in masked data, it is not required for PPE applications. Nevertheless, these masking methods usually generate masked data with similar linear measures.

"Shuffling" indicates the use of rank-based ordering of original values of either residuals or confidential attributes using the rank order of orthogonal (or independent) scaled noise and/or perturbed values, respectively. The shuffling approach has many advantages (Muralidhar and Sarathy, 2003a; 2006b) already mentioned in Section II.2.3. This topic will be revisited briefly at the end of this subsection. In addition, we provide a step-by-step procedure for shuffling in Appendix C, in which we also demonstrate the

[4] The term "Shuffling" is first introduced in the masking literature by Muralidhar and Sarathy (2003a; 2006).

shuffling procedure using a hypothetical example. We refer to the shuffling procedure by the operator "Shuffle *A* by *B*" in the proposed algorithms.

As mentioned earlier, the Relationship-Based NM-EGADP Masking (RBM) approach can be generally thought of as a three-stage approach: relationships, residuals, and masking. In the first stage, we estimate and learn the *relationships* between **X** and **S**. In the second stage, we calculate the *residuals* from the previous stage and we analyze *them* for relationships and patterns. In addition, we check the residuals using the security index (*SI*) to see whether they have enough variation for effective application of the RBM masking. At this stage, we also generate orthogonal noise with characteristics that resemble the characteristics of the *residuals*. In the final stage, we *mask* the confidential attributes **X** based on the characteristics of the *residuals*.

Assume that we have *p* non-confidential (numeric and categorical) attributes **S** and *q* confidential (numeric) attributes **X**. The relationships between these *p*+*q* variables involve non-monotonic relationships. We want to mask the confidential attributes **X** without altering or destroying the relationships in the original dataset. The NM-EGADP perturbation and the NM-EGADP shuffling algorithms are (note that the presentation of these algorithms is different than the presentation in Appendix C):

Stage I: Relationships:

1. Regress **X** on **S** by training *q* Least Squares Support Vector Machines (LS-SVM) neural networks N_{1j} (one for each individual confidential attribute **X**):

$$N_{1j} \sim E(f(X_j \mid \mathbf{S})), \quad j = 1,\mathrm{K},q. \tag{4.1}$$

N_1 (=[N_{11},…, N_{1q}]) learns the function of the expected value of the conditional distribution $f(\mathbf{X}|\mathbf{S})$ as discussed by Bishop (1995) and others. Refer to Section III.2 for more details.

2. Use the set of trained neural networks N_1 to calculate the following:

   a. The set of expected values μ (= $[\mu_1,\mathrm{K}\,,\mu_q]$) of the conditional distribution $f(\mathbf{X}|\mathbf{S})$ evaluated at **S** values where:

$$\mu_j = E(f(X_j|\mathbf{S}))|_{\mathbf{S}}, \quad j = 1,\mathrm{K}\,,q\,. \tag{4.2}$$

Stage II: Residuals:

   b. The orthogonal residuals set **r** (=[r_1,…, r_q]) where:

$$r_j = X_j - E(f(X_j|\mathbf{S}))|_{\mathbf{S}}, \quad j = 1,\mathrm{K}\,,q\,. \tag{4.3}$$

3. Compute the covariance matrix of the first residuals set $\sum_{\mathbf{r}}$. This covariance matrix will be used later to scale another orthogonal set of residuals and make its covariance matrix the same as $\sum_{\mathbf{r}}$.

Stage III: Masking:

4. Generate q independent random variates **V** (=[V_1,..., V_q]).

5. Regress **V** on both **S** and **X** by training another set of q LS-SVM neural networks N_2 (=[N_{21},…, N_{2q}]) where:

$$N_{2j} \sim E(f(V_j|\mathbf{S},\mathbf{X})), \quad j = 1,\mathrm{K}\,,q\,. \tag{4.4}$$

6. Use the set of the trained neural networks N_2 to calculate a second orthogonal residuals set **b** (=[b_1,…, b_q]) where:

$$b_j = V_j - E(f(V_j|\mathbf{S},\mathbf{X}))\big|_{\mathbf{S},\mathbf{X}}, \quad j = 1,\mathrm{K}\,,q\,. \tag{4.5}$$

7. Compute the covariance matrix of the second residuals set $\sum_{\mathbf{b}}$. Note that although the new set of residuals **b** is orthogonal to **S**, **X**, and **r**, the covariance matrix $\sum_{\mathbf{b}}$ is different than $\sum_{\mathbf{r}}$.
8. Compute a new residuals set **e** by scaling the (normalized) set of the orthogonal residuals **b** to have the same covariance matrix as the covariance matrix $\sum_{\mathbf{r}}$ of original dataset:

$$\mathbf{e} = (\sum_{\mathbf{r}})^{0.5} (\sum_{\mathbf{b}})^{-0.5} \mathbf{b}. \tag{4.6}$$

9. Calculate the new perturbed attributes **Y**:

$$\mathbf{Y} = \mu + \mathbf{e} = E(\mathbf{X}|\mathbf{S}) + \mathbf{e}. \tag{4.7}$$

Therefore:

$$\mathbf{Y} \sim E(f(\mathbf{X}|\mathbf{S})) + \mathbf{e}. \tag{4.8}$$

Step 9 ends the NM-EGADP perturbation algorithm. The NM-EGADP shuffling algorithm has the following extra step:

10. Shuffle **X** by **Y** to compute shuffled attributes $\mathbf{Y}_{\mathbf{shf}}$. This is done by ordering original values of individual confidential attributes **X** according to the rank-order of the corresponding perturbed attributes **Y** (see Appendix C for more details). This is possible because we utilize the expectation of the conditional distribution $f(\mathbf{X}|\mathbf{S})$ plus an orthogonal noise term **e** that resamples the covariance matrix of the residuals of the original dataset while the assumption is that the noise terms have *constant variance* and linear (or simple) patterns left among them (these assumptions are discussed later).

The perturbed attributes **Y** or the shuffled attributes $\mathbf{Y}_{\mathbf{shf}}$ then can be released (along the non-confidential attributes **S**) instead of the original confidential attributes **X**.

In Step 1, the algorithm tries to learn the expectation of the conditional distribution $f(\mathbf{X}|\mathbf{S})$ using (4.1) (i.e. $E(f(\mathbf{X}|\mathbf{S}))$ or simply $E(\mathbf{X}|\mathbf{S})$). The estimated conditional expectation is next evaluated at **S** values in Step 2-a using (4.2). This evaluation (the calculated u) is used in Step 9 to generate the perturbed values **Y** by adding a zero-mean noise term **e,** which is orthogonal to **X** given **S**, to u. This automatically minimizes disclosure risk in the perturbed values and, accordingly, in the shuffled values (Step 10). As discussed earlier in Section III.1, masking methods satisfy security requirements when they generate masked attributes **Y** using only a random function of non-confidential attributes **S** and independent noise (Muralidhar and Sarathy, 2006b). However, this does not imply that the perturbed (and accordingly, the shuffled) variables necessarily maximize data utility unless the random function is derived from the (estimated) conditional distribution $f(\mathbf{X}|\mathbf{S})$. Nevertheless, since NM-EGADP employs the conditional expectations $E(\mathbf{X}|\mathbf{S})$, the resulting perturbed and shuffled variables should have a better data utility than any random function of **S**. The better the indirect estimation of the conditional distribution $f(\mathbf{X}|\mathbf{S})$ in terms of conditional expectations, the better the data utility. If the conditional expectation is calculated accurately, it would represent the best mean squared error estimator $f(\mathbf{S})$ for **X** according to the theory of estimation and detection (Bickel and Doksum, 2001; Graybill, 1976; Jelenkovic, 2001; Shao, 1999):

$$E[\mathbf{X} - f(\mathbf{S})]^2 \geq E[\mathbf{X} - E[\mathbf{X}|\mathbf{S}]]^2 . \tag{4.9}$$

Bishop (1995) points out and proves the possibility of utilizing neural networks as a framework to model the function of expected values of conditional probability density function $f(\mathbf{X}|\mathbf{S})$ (or simply the conditional expectations $E(\mathbf{X}|\mathbf{S})$) by mapping form **S** to **X**. Please refer to Section III.2 in this study and to Section 6.1.3 (pp. 201-206, Equation

(6.46)) and its discussion in Bishop (1995) for more details. "*For many regression problems, the form of network mapping given by the conditional average … can be regarded as optimal,*" Bishop (1995; pp. 205) said. The above and other ANN characteristics discussed in Section III.2 make the use of ANN and more specifically LS-SVM in our masking methods a sound choice.

Steps 2-b and 3 of the NM-EGADP shuffling algorithm are used to compute the covariance matrix $\sum_{\mathbf{r}}$ of the residuals set **r**. $\sum_{\mathbf{r}}$ is used for rescaling in Step 8. The goal of Steps 4 to 6 is to create a new residuals set **b** that is orthogonal to **X** given **S**. Because of this orthogonality, the use of **b** in Step 9 to create perturbed attributes **Y** is possible and meets the minimum disclosure risk criterion. However, the covariance matrix $\sum_{\mathbf{b}}$ (calculated in Step 7 and used in Step 8) is different than the covariance matrix of the original residuals set $\sum_{\mathbf{r}}$, and this may lead to a poor data utility mainly among confidential attributes **X**. The goal of Step 8 is to create new orthogonal residuals set **e** that has a covariance matrix equal to $\sum_{\mathbf{r}}$. As shown in Burridge (2003), this is done by rescaling the residuals set **b** in two sub-steps. First, **b** is normalized by multiplying it by $(\sum_{\mathbf{b}})^{-0.5}$. Second, the result of the previous sub-step is scaled by multiplying it by $(\sum_{\mathbf{r}})^{0.5}$. Note that $(.)^{0.5}$ denotes a square-root matrix (Johnson and Wichern, 1998, pp. 67).

The new residuals set **e** is used in Step 9 to create new perturbed attributes **Y**. These masked values can be released, but some differences in the marginal distributions usually occur (such as discrepancies in the range and the variance of the original values). Shuffling overcomes such problems. Perturbed attributes **Y** are used as a basis to shuffle the original confidential attributes **X** to generate the shuffled attributes $\mathbf{Y}_{\mathbf{shf}}$. $\mathbf{Y}_{\mathbf{shf}}$

attributes satisfy both minimum disclosure risk and maximum (but not ideal) data utility, and $\mathbf{Y}_{\mathbf{shf}}$ along with **S** are ready to be released. As Muralidhar and Sarathy (2006b) point out, shuffling has three advantages. First, the shuffling approach maintains marginal distributions. This leads automatically to the second advantage: shuffling avoids changes in the variance. Third, shuffling can be used even if the residuals are not normal.

We conclude our discussion of the NM-EGADP shuffling algorithm with a few remarks. Note that in Steps 1 and 5, the assumption is that good training parameters have been chosen in a way that balances the trade-off between bias and variance in the context of training neural networks. Because of this dependency on training parameters (and the nature of the nonlinear monotonic or non-monotonic relationships) and the tendency of ANN approaches to over-fit the data, we should not expect linear measures, such as covariance and linear correlation measures, of original datasets to be maintained *exactly* in perturbed (or shuffled) datasets, as what we usually get using the original EGADP algorithm. On the other hand, although EGADP maintains these linear measures exactly, it destroys any nonlinear relationships (monotonic or non-monotonic) that may exist in original datasets by linearizing them. This causes biases related to relationships and leads to a poor data utility for PPE. Further, if the nature of relationships among variables is nonlinear, then linear measures may be meaningless.

We conclude this subsection by mentioning the main difference between the two RBM methods (Method 1 and Method 2) and the other two masking methods (Method 3 Method 4: Residuals-Shuffled NM-EGADP Perturbation and Residuals-Shuffled NM-EGADP Shuffling). In Method 3, we shuffle the residuals based on the scaled orthogonal noise instead of shuffling at the variables level as in Method 2. In Method 4, we shuffle at

both levels: residuals level and variables level. Refer to Appendix C for more information.

IV.2. Assumptions of the Relationship-Based Masking Methods

The assumptions underlying RBM methods are critical for defining the applicability of the RBM methods and drawing their limitation boundaries. As usual, the assumptions should be theoretically driven or practically guided to be sound.

The proposed RBM methods depend on two components to generate masked attributes: conditional expectations $E(\mathbf{X}|\mathbf{S})$ and residuals $\mathbf{r}$. There are some assumptions related to these two components. First, the estimation mechanism, LS-SVM in our case, should (be able to) learn the true conditional expectations $E(\mathbf{X}|\mathbf{S})$. This is critical from two perspectives. Incorrectly learned relationships (conditional expectations $E(\mathbf{X}|\mathbf{S})$) from original data lead to correspondingly incorrectly generated relationships $E(\mathbf{Y}|\mathbf{S})$ in masked data. Hence, data utility is affected. In addition, residuals $\mathbf{r}$ will not be orthogonal to the true conditional expectations if it can be estimated later. Hence, data security is adversely affected.

This leads us to the first assumption; namely, all relationships $E(\mathbf{X}|\mathbf{S})$ among non-confidential attributes $\mathbf{S}$ and confidential attributes $\mathbf{X}$ should be monotonic or (*single-valued mapping*) non-monotonic. As discussed in Section III.1, multi-valued mapping hinders the estimation mechanisms from learning the true conditional expectations.

In Section III.3, we talked about the roles residuals $\mathbf{r}$ plays in guiding and determining the data utility and data security requirements. We also mention the important rule that residuals $\mathbf{r}$ plays in reproducing relationships among confidential attributes $\mathbf{X}$ in masked data. In addition, we reached an interesting result relating the

covariance matrices of $E(\mathbf{X}|\mathbf{S})$ and **r** to the covariance matrix of confidential attributes **X**. We also provide its proofs in Appendix B. Hence, the proposed masking methods try to maintain the covariance of the independent added noise terms **e** as the covariance of residuals **r** to approximate the distributions of residuals.

When the residuals are independent, reproducing the covariance of **r** in **e** reduces to reproducing variances of residuals **r**. However, maintaining covariance matrices can result in maintaining linear relationships regardless of the joint distributions of **r** although it does not maintain original joint distributions. Therefore, we assume *linearity* (linear relationships and patterns) among residuals **r** for RBM methods to work.

Nevertheless, even when the relationships among confidential attributes **X** are non-monotonic, patterns and relationships among **r** can be simple when the patterns and relationships among conditional expectations $E(\mathbf{X}|\mathbf{S})$ shape (or account for) most of the patterns and relationships among **X**. In this case, even monotonic nonlinear relationships among **r**, especially when they do not deviate largely from linearity, may work. Note that in two masking methods (Method 3 and Method 4 in Appendix C), the residuals are shuffled and this may compensate for any moderate deviation from linearity given that the residuals have constant variance.

Thus, the second assumption is *linearity* among residuals **r**, which means that the patterns left among residuals must be simple patterns (no relationships or linear ones). Nevertheless, our expermintal results (Chapter VII) suggest that a little violation in this assumption, such as monotonic nonlinear relationships with moderate deviation from linearity or most of the variation among **X** accounted for by $E(\mathbf{X}|\mathbf{S})$, might not degrade

significantly the performance of the RBM methods, especially the ones employing shuffling refinement.

Finally, since the RBM approach utilizes independent normal noise terms with *constant* variance in masking the confidential attributes **X**, the third assumption for RBM to work is that residuals **r** should have *constant variance*. This assumption is basically required for maintaining the relationships among **X** and not for maintaining the relationships between **X** and **S**.

The assumption of single-valued mapping can be assessed using scatter plots. The *linearity* and *constant variance* assumptions can also be checked graphically. Actually, Hair et al. (1998) suggest that the constant variance (homoscedasticity) assumption "*is best examined graphically*" (pp. 74), and mention that scatter plots are "*the most common way to assess linearity*" (pp. 75). In addition, they discuss some statistical tests for assessing some assumptions. For more details about assumptions assessment and other related topics, see Chapter 2 titled "Examining Your Data" and Chapter 4 in (Hair et al., 1998).

CHAPTER

V. ASSESSMENT MEASURES FOR THE RBM APPROACH

For any masking method to be practical, it should prove its usefulness and effectiveness in terms of *data security* and *data utility*. The first step in this process is to define the possible security threats one wants to minimize and the data utility one wants to maximize. We already discussed these issues in Section II.2. The second step is to find or develop a suitable set of objective measures to assess the effectiveness and usefulness of the masked methods based on the provided data security and utility requirements.

V.1. Data Security Measures in PPE

The MSE security measures, proposed by Muralidhar and Sarathy (2005a; 2006a) based on Graybill (1976), are the main security measures we will use for evaluating *value disclosure (confidentiality)*. They naturally fit in the context of estimated relationships and conditional expectations using MSE. To recall, masked attributes **Y** should be independent of confidential attributes **X** given non-confidential attributes **S** (Muralidhar and Sarathy, 2003c). In addition, **S** should always be a better predictor for **X** than **Y**. Both mean that a snooper will always use **S** to obtain more accurate prediction results. However, if (s)he tries to combine **Y** with **S** to improve the prediction of **X**, (s)he gains nothing (avoid partial and inferential disclosure). These conditions can be translated in terms of the MSE in the following inequalities and equalities (Graybill, 1976; Muralidhar and Sarathy, 2005a; 2006a):

$$E[\mathbf{X} - E(\mathbf{X}|\mathbf{S})]^2 = E[\mathbf{X} - E(\mathbf{X}|\mathbf{S},\mathbf{Y})]^2 \quad (5.1)$$

and

$$E[\mathbf{X} - E(\mathbf{X}|\mathbf{S})]^2 \leq E[\mathbf{X} - E(\mathbf{X}|\mathbf{Y})]^2. \quad (5.2)$$

The above measures assume that real conditional expectations are correctly estimated.

From another perspective, when all relationships are *linear* as in the case of multivariate normal datasets, there are also corresponding security measures, proposed by (Sarathy and Muralidhar, 2002), for the above mentioned MSE security measures. These measures are presented in terms of canonical correlation *CC*:

$$CC(\mathbf{X}|\mathbf{S}) = CC(\mathbf{X}|\mathbf{S},\mathbf{Y}) \quad (5.3)$$

$$CC(\mathbf{X}|\mathbf{S}) \geq CC(\mathbf{X}|\mathbf{Y}). \quad (5.4)$$

Although these *CC* security measures are mainly for *linear* relationships, they can be adapted for *nonlinear monotonic* or *non-monotonic* relationships that completely consist of piecewise linear relationships. In this case, we call them *piecewise CC*. These piecewise measures will not be used in this dissertation since they are only applicable in the above special case. The only exception is when we talk about the performance of our methods on the motivation example in which non-monotonic relationships completely consist of piecewise linear relationships.

For assessing *identity disclosure* or *re-identification risk* (*privacy*), we use probabilistic *record linkage* (Winkler, 2004a; b) for re-identifying individuals from their masked values.

V.2. Data Utility Measures in PPE

Unlike data security measures, it is difficult to provide a global measure for data utility for every possible use or analysis because just enumerating and accounting for all

possible uses of a released dataset is not possible (Domingo-Ferrer et al., 2001; 2003). Therefore, data utility measures are usually derived based on a specific anticipated use or analysis. Privacy-Preserving Estimation (PPE) is no exception. Regression and estimation tasks are about learning conditional expectations and quantifying relationships (Rud, 2001) between a dependent variable and one or more independent variables. A PPE algorithm succeeds when regression models obtained from masked datasets are the "same" as (or very similar to) the corresponding regression models obtained from original datasets. This defines the data utility in PPE informally. However, we need to quantify this definition in terms of similarity measures of maintained relationships.

As seen earlier in Subsection II.3.1, existing masking methods attempt to preserve some aggregate (usually correlation-based) measures in masked datasets to be the "same" as the corresponding aggregate measures of original datasets. These aggregate measures are also used as data utility measures. Unfortunately, there is no aggregate measure that is able to capture non-monotonic relationships. Hence, we propose two types of possible general data utility measures for privacy-preserving regression: *parameter-based measures* and *prediction-based measures*. Although the main focus of these measures is non-monotonic relationships, they can be used for all three types of relationships.

Parameter-based data utility measures are based on the concept that when the characteristics and the relationships of the masked dataset resemble the characteristics and the relationships of the original dataset, fitting a specific regression model on both datasets should generate very similar estimated parameters. This is regardless of whether the model is mis-specified or well-specified. Nonlinear (and linear) regression estimation

procedures found in standard statistical and mathematical packages such as SAS, SPSS, and Matlab can be used for this purpose.

When one wants to estimate a regression model using non-parametric estimation techniques such as ANN (also known as black-box techniques), parameter-based data utility measures cannot be used because it is harder to extract and compare parameters from built models. Instead, we propose *prediction-based data utility measures*. Relationships are maintained in masked datasets when two regression models, one estimated from the original dataset and the other from the masked dataset, using the same estimation technique have the "same" prediction power when both models are evaluated on any specific dataset. Similar to Muralidhar and Sarathy (2005a), MSE can be used for measuring the similarity of the predication power of the two models. Hence, prediction-based data utility measures can be based on MSE. Notice the applicability of these measures for regression problems in general, even in the case of monotonic (linear or nonlinear) relationships.

A normalized form of the MSE and SSE measures for the predictive power of regression models involving nonlinear relationships is the following form of coefficient of determination R^2 (Kirkup, 2003, pp. 74):

$$R^2 = 1 - \frac{\sum (a_i - \hat{a}_i)^2}{\sum (\overline{a} - \hat{a}_i)^2} \tag{5.5}$$

where a_i is the actual value of a dependent random variable A, $\hat{a}_i$ is the corresponding predicted (or fitted) value using a specific nonlinear regression model, and $\overline{a}$ is the mean of the dependent variable A. The values of this measure lie between 0 and 1. It can easily be converted into a percentage, which facilitates the comparison process. Thus, we can

use the above R^2 measure to cast our prediction-based MSE data utility measures (and even MSE security measures) into a normalized easy-to-compare form.

The number of all possible combinations of regression models that can be built from any dataset is typically large and grows rapidly with the number of variables. Consider the motivation example in Section I.3. There are four variables: two non-confidential attributes (S_1 and S_2) and two confidential attributes (X_1 and X_2). If we take any variable as a dependent variable and calculate all combinations of the other three variables as independent variables, then we will have seven possible models for just a single dependent variable (DV). Since any variable from the four variables can be a dependent variable, there are 28 models in total. If we ignore calculating the measures for the two cases we are sure will not change (i.e. relationships between non-confidential attributes: $S_1|S_2$ and $S_2|S_1$), then we will have 26 models we need to build from original datasets and compare with the corresponding 26 models we built from masked datasets. Table 2 below shows the 14 possible combinations of IV for the two confidential DV (X_1 and X_2).

One possible approach is to calculate the data utility measures for every

Table 2. Possible different combinations of relationships to be maintained when a confidential attribute is a dependent variable in the store motivation example (2 X and 2 S). *k* is the model number.

Type of IVs	No	Confidential Dependent Variable (DV)	
		X_1	X_2
Mixed **X & S**	1	$E_k(X_1 \mid S_1, S_2, X_2) = E_k(Y_1 \mid S_1, S_2, Y_2)$	$E_k(X_2 \mid S_1, S_2, X_1) = E_k(Y_2 \mid S_1, S_2, Y_1)$
	2	$E_k(X_1 \mid S_1, X_2) = E_k(Y_1 \mid S_1, Y_2)$	$E_k(X_2 \mid S_1, X_1) = E_k(Y_2 \mid S_1, Y_1)$
	3	$E_k(X_1 \mid S_2, X_2) = E_k(Y_1 \mid S_2, Y_2)$	$E_k(X_2 \mid S_2, X_1) = E_k(Y_2 \mid S_2, Y_1)$
Pure **X**	4	$E_k(X_1 \mid X_2) = E_k(Y_1 \mid Y_2)$	$E_k(X_2 \mid X_1) = E_k(Y_2 \mid Y_1)$
Pure **S**	5	$E_k(X_1 \mid S_1, S_2) = E_k(Y_1 \mid S_1, S_2)$	$E_k(X_2 \mid S_1, S_2) = E_k(Y_2 \mid S_1, S_2)$
	6	$E_k(X_1 \mid S_1) = E_k(Y_1 \mid S_1)$	$E_k(X_2 \mid S_1) = E_k(Y_2 \mid S_1)$
	7	$E_k(X_1 \mid S_2) = E_k(Y_1 \mid S_2)$	$E_k(X_2 \mid S_2) = E_k(Y_2 \mid S_2)$

individual model (every possible relationship) and then find the average of all these individual models. However, this approach will hide the importance of relationships at the individual level and may not be the best approach. Alternatively, we can take a pragmatic approach that shows how well the relationships are preserved at the individual level for *some* specific relationships. Thus, we can calculate MSE or the R^2 of the prediction-based (MSE) data utility at the following "*representative*" settings, for example:

- $R^2(X_1|\mathbf{S}) = R^2(Y_1|\mathbf{S})$ represents the relationship between confidential and non-confidential attributes,
- $R^2(X_1|X_2) = R^2(Y_1|Y_2)$ represents the relationship among confidential attributes, and
- $R^2(X_1|\mathbf{S},X_2) = R^2(Y_1|\mathbf{S},Y_2)$ represents the relationship between confidential, and a mixture of confidential and non-confidential attributes.

The calculations and comparisons of the prediction-based data utility measures assume that the models from the original and masked datasets are estimated using the same estimation mechanism and are evaluated at the same data points.

"*Data mining tools should model nonlinearity very well, so the predicted/actual values relationship should be pretty much linear with all of the nonlinearity accounted for in the model*," Pyle (2003, pp. 443). This linearity, which indicates similarity, between predicted/actual values can be measured in different ways. Witten and Frank (2005) suggest using the correlation coefficients as a measure for the similarity between predicted and actual values (refer to Section 5.8 "Evaluating numeric prediction" and Table 5.8 in (Witten and Frank, 2005)).

Similarly, correlation can be used as a prediction-based data utility measure. When the characteristics of masked datasets resemble the characteristics of original datasets, fitted values of models built based on original datasets and fitted values of corresponding models built based on masked datasets should be similar. This assumes that fitted values are calculated using the two models with the same set of values for the independent variables, and the models, which can take different possible formats such as estimated functions or trained ANNs, are built using the same estimation mechanisms. This similarity between the two sets of fitted values can be measured by correlation between the sets. Ideally, the correlation should be 1.0 (i.e. the two sets of fitted values should be identical). The closer the correlation is to one, the more similar the masked dataset is to the original dataset.

For example, we can estimate the following two corresponding models: $E(X_i|\mathbf{S})$ from the original dataset and $E(Y_i|\mathbf{S})$ from the masked dataset. Then we calculate the fitted values for both models using the values of the independent variables **S**: $E(X_i|\mathbf{S})$|s and $E(Y_i|\mathbf{S})$|s. When the models $E(X_i|\mathbf{S})$ and $E(Y_i|\mathbf{S})$ are similar, the correlation among fitted values (i.e. $Corr(E(X_i|\mathbf{S})|s, E(Y_i|\mathbf{S})|s)$) should be very high, ideally *one*.

Similarly, we can calculate the slope of the regression of the two sets of fitted values. When they are similar, the slope is close to *one*. Equivalently, the scatter plot of the two sets of fitted values should show a strong linear pattern (close to a sharp line) with a slope equal to one.

However, one problem with the above approach is that correlation and slope only consider the *direction* of relationships between the fitted values; it does not consider their *magnitudes*. In other words, when the set of fitted values we get from the original model

and the masked model are different in the magnitude but one set is perfectly a linear transformation of the other set, we still get a correlation of *one* and a slope of *one*, which can be mistakenly interpreted as an indication of "similar models". To account for the difference in magnitudes, the sum of the differences between the two vectors can be calculated. This sum should be, or at least approach, *zero*. Similarly, the mean of the square of the differences can be used instead of the sum of the differences.

In the next subsection, we discuss statistical equivalence tests for validating models and how they can be used as data utility measures. One advantage of some of these tests is that they consider both the *direction* and *magnitude* of the relationships in establishing the similarity between the fitted values from original and masked models.

V.3. Equivalence Tests for Validating Models as Data Utility Measures

Model validation is an important stage in the process of building models. It establishes the usefulness of built models for their proposed practical goals (Rykiel, 1996). There are different possible tests for validating models (Yang et al., 2004). Many of these tests are statistical significance tests (for testing a significant *difference* between two hypotheses). They test the null hypothesis H_0 of valid models (there is *no significant difference* between *observations* and *model predictions*) against the alternative hypothesis H_a of invalid models (such *significant difference exists*) (Robinson and Froese, 2004):

$$\begin{aligned} H_0 &: \mu_o - \mu_p = 0 \ \text{; valid model} \\ H_a &: \mu_o - \mu_p \neq 0 \ \text{; invalid model} \end{aligned} \tag{5.6}$$

where μ_o is the population mean of the observations and μ_p is the population mean of the model predictions.

Unfortunately, many practitioners misinterpret the failure to reject the null hypothesis as evidence of the *trueness* of the null hypothesis (Cohen, 1990; 2003; Parkhurst, 1985; 2001). In the context of validating models, this may lead to the wrong conclusion of validated models although low power could, and more likely would, be the main reason behind this rejection failure (Robinson and Froese, 2004). Hence, statistical significance tests are unsuitable for validating models (Loehle, 1997; Mayer and Butler, 1993).

There are two other problems with using significance tests as a tool for validating models (Robinson et al., 2005). First, as the size of the sample, and thus the power, increases, any small difference between the population *observations* mean and the population *predictions* mean may become significant and invalidate tested models unless we encounter an exactly correct null hypothesis. Second, statistical significance is unassociated with practical significance. Put another way, significance tests can reject the null hypothesis of valid models due to a significant statistical difference, which is practically negligible. Model validation using equivalence tests (ET) (Robinson et al., 2005; Robinson and Froese, 2004) can avoid these three problems.

Equivalence tests are hypothesis tests that are more appropriate than standard significance tests when the research goal is to demonstrate *similarity* rather than *difference* (Parkhurst, 2001). The intended tested similarity can be between two means or two proportions (Rogers et al., 1993; Streiner, 2003). Equivalence tests are popular in the field of biomedicine where they are known as bioequivalence tests. Many drug regulatory entities in the USA (e.g. Food and Drug Administration (FDA)), Europe (e.g. European Community (EC)), and Japan require *generic drugs* to prove its equivalence to brand

names using bioequivalence tests before they are permitted to the market (Berger and Hsu, 1996; Yanagawa, 2005). (Bio-)Equivalence tests have started to find their way to other fields including psychology (Rogers et al., 1993), environmental sciences (McBride, 1999), and the military (Warner, 2002), to name a few.

There are several equivalence tests (Wellek, 2003). Among the most popular are the two one-sided t-test (TOST) and the paired t-test for equivalence (PTTE). Parkhurst (2001) pointed out that TOST was proposed independently and slightly differently by Schuirmann (1981) and Westlake (1981). Westlake (1981) proposed the test during his response to the comments made by Kirkwood (1981) on his early work in equivalence tests (Westlake, 1976; 1979). Robinson et al. (2005, pp. 905) presented the TOST as a four-step procedure. More on PTTE can be found in (Wellek, 2003, Section 5.3, pp. 77-82).

Bartko (1991) points to some early work in (bio-)equivalence tests fields, known earlier as "*proving the null hypothesis*," such as (Blackwelder, 1981; 1982), (Blackwelder and Chang, 1984), (Detsky and Sackett, 1985), (Dunnett and Gent, 1977), and (Makuch and Simon, 1978). Two good tutorials of equivalence tests are Streiner (2003) and Tamayo-Sarver et al. (2005). Wellek (2003) has written a book about equivalence tests. For more information about this book, refer to the review written by Yanagawa (2005).

Many commercial statistical programs in the market today support equivalence tests partially or fully. SAS (2003) performs power analysis for some equivalence tests. NCSS and PASS program (Hintze, 2004) support some equivalence tests besides many of their power analysis procedures. EquivTest (2006) is a program dedicated to equivalence tests analysis.

There are different possible uses for equivalence tests. Parkhurst (2001) suggests that the use of equivalence and reverse tests can reduce the misinterpretation of negative results in statistical significance tests (hypotheses testing). Equivalence tests are more appropriate than significance (hypothesis) tests when the research goal is to demonstrate *similarity* rather than *difference*.

Equivalence tests can be also used for validating models (Robinson et al., 2005; Robinson and Froese, 2004). They have the advantage of eliminating the *three* problems mentioned earlier that are associated with using statistical significance tests when validating models (Robinson et al., 2005; Robinson and Froese, 2004). First, to avoid the problem of misinterpreting the results of negative-results model-validation hypothesis tests, Loehle (1997) suggests that the null hypothesis H_0 should be that the model is invalid. This is precisely how equivalence tests test hypotheses, and why Robinson and Froese (2004) advocate the use of equivalence tests for validating models.

Second, as we have more data and the power of the test increases, the model has the advantage of proving more its validity instead of its invalidity in the case of significance tests (Robinson and Froese, 2004). Although it is not always true, equivalence tests usually need a larger sample size than significance tests (Streiner, 2003). The sample size depends mainly on the equivalence margin δ (discussed shortly): as the equivalence margin gets smaller, the required sample size increases (Streiner, 2003). Berger and Hsu (1996), Wellek (2003), Streiner (2003), and Tamayo-Sarver et al. (2005), among others, deal with the subject of sample size and power analysis for equivalence tests. Third, practical negligible discrepancy between the mean of

observations and the mean of predictions can be easily incorporated into statistical equivalence tests using the equivalence margin δ as we will see next.

Robinson and Froese (2004) suggested using equivalence tests, or more specifically TOST and PTTE, for validating models:

$$\begin{aligned} H_0 &: \mu_o - \mu_p \neq 0 \quad \text{; invalid (or different) model(s)} \\ H_a &: \mu_o - \mu_p = 0 \quad \text{; valid (or similar) model(s)} \end{aligned} \tag{5.7}$$

In model validation, we begin the test by defining a metric as the mean of the differences between observations and model predictions. Then, the equivalence margin δ, which is used around the metric to create a range of practical negligible difference or what is called *indifference or equivalence region*, should be specified. This margin can be relative, such as a percentage of the standard deviation, or absolute. Although some may consider the subjective choice of the equivalence margin as a disadvantage, or different applications may require different equivalence margins, it is easy to re-calculate these tests using different equivalence margins when the mean, the standard deviation of the differences, and the sample size are provided (Robinson and Froese, 2004).

The hypotheses in Equation (5.7) can be represented more explicitly using the equivalence margin δ as:

$$\begin{aligned} H_0 &: \left|\mu_o - \mu_p\right| > \delta \qquad \text{; invalid (or different) model(s)} \\ &\text{or: } \mu_o - \mu_p > \delta \quad \text{OR} \quad \mu_o - \mu_p < -\delta \\ H_a &: -\delta \leq \mu_o - \mu_p \leq \delta \quad \text{; valid (or similar) model(s)} \end{aligned} \tag{5.8}$$

This proposed ET tests the null hypothesis of an invalid model (model predictions are different than observations) against the alternative hypothesis of a valid model (observations and predictions are similar) (Robinson and Froese, 2004). We can adapt the usage of this ET as a data utility measure to test the null hypothesis of different models

(original datasets models and their corresponding masked datasets models are different) against the alternative hypothesis of similar models (original datasets models and their corresponding masked datasets models are similar) in terms of their predictions. The descriptions of the hypotheses shown between parenthesis in Equations (hypothesis settings) (5.7) and (5.8) indicate this usage adaptation (i.e. "different models" vs. "similar models"). The vectors of values that are compared in our case are the fitted (predicted) values from original datasets models and the fitted (predicted) values from the corresponding masked datasets models (e.g. $E(X_i|\mathbf{S})|\mathbf{s}$ vs. $E(Y_i|\mathbf{S})|\mathbf{s}$).

In testing these ET hypotheses, a special confidence interval for the metric (the mean of the differences between the fitted values from original and masked datasets) is calculated. When this confidence interval is totally contained within the range of the indifference region, the null hypothesis of different models is rejected and the practical similarity of these models is statistically established. Otherwise, we fail to reject the null hypothesis of different models (different predictions or fitted values from original and their corresponding masked models). This rejection failure is due to either a low power test (as in the case of a small sample size) or a real difference in the population.

These model-validation equivalence tests (ET) consider only the *magnitude* of the relationship unlike data-utility correlation and slope measures proposed earlier, which consider only the *direction* of the relationship. Robinson et al. (2005) extended the work of Robinson and Froese (2004) by providing a deeper regression-based equivalence test using TOST for model validation. The new test compares: (a) the similarity of the means (of predicted and of observed, or what is called "*population-level agreement*"), and (b) the closeness of the slope of regression to one (similarity between individual pairs of

observations and predictions, or what is called "*point-to-point agreement*"). Notice the similarity of (a) to the test provided by (Robinson and Froese, 2004), which focuses on the *magnitude* of the relationship, and the similarity of (b) to the proposed data-utility correlation and slope measures, which focus on the *direction* of the relationship. Hence, this test considers both the *magnitude* and the *direction* of the relationship.

The regression in this test (Robinson et al., 2005) is done between the observations vector as a dependent variable (DV) and the predictions vector (after subtracting its mean) as an independent variable (IV). Although the subtraction of the predictions mean does not change the regression slope, it makes the equivalence tests of the *regression intercept* and the *regression slope* independent. For the intercept, the ET investigates the similarity of the intercept to the mean of observations. For the slope, the ET investigates the closeness of the regression slope to *one*. Hence, the observations mean (the intercept) and the slope of one represent the metrics for the equivalence tests.

Similar to the first test (Robinson and Froese, 2004), we need also to specify an alpha level α (a test size) and equivalence (indifference) region for the two equivalence tests. Although the equivalence tests for the intercept and the slope are independent, the joint test size α should be corrected to allow explaining the results of the two tests jointly (Robinson et al., 2005). For example, when we use $\alpha = 0.05$ for each test, the joint α level will be more (0.0975). When correction is done, each test should be executed at $\alpha = 0.02532$ (or equivalently, two one-sided 97.468% CI) for the intercept test and for the slope test.

For the intercept testing, the equivalence region is the *observations mean* $\pm$ *equivalence margin* (δ_0). For the slope testing, the equivalence region is $1 \pm$

equivalence margin (δ_1). Again, equivalence margins δ_0 and δ_1 can be absolute or relative. The next step is to calculate the special (two one-sided) confidence interval for the intercept and the slope (based on their standard error for example). Finally, we reject the dissimilarity null hypothesis of both/either test(s) when its confidence interval is contained completely within the equivalence region.

The confidence intervals for these tests of equivalence assume: (1) the model is correct and (2) the residuals: (a) are independent, (b) have constant variance, and (c) are normally distributed. To avoid these assumptions, Robinson et al. (2005) suggested a non-parametric bootstrap method to construct the test confidence interval. Further, the choice of equivalence intervals in the proposed test is subjective in nature. To avoid possible resistance to the test use or results, Robinson et al. (2005) suggested a method to reverse the results of these tests and calculate the smallest TOST equivalence regions that can lead to the rejection of the null hypothesis of dissimilarity at a specific α level.

We adapt this test (Robinson et al., 2005) as our main data utility measure in the next chapter. In this adapted test, we replace *observations* with the fitted values (model predictions) from original datasets models and *predictions* with the fitted values (model predictions) from their corresponding masked datasets models.

The logic is simple, yet powerful. Models built from masked data should be very similar, in terms of predictions, to models built from their corresponding original data. This similarity means two things. First, the mean of the fitted values from an original data model and the mean of the fitted values from its corresponding masked data model (both models evaluated using the same data) should be very similar, if not equal. Second, the slope of the regression of one of these two sets of fitted values on the other should be

(close to) one. We used the equivalence package[5] developed by Robinson (2005) in the statistical programming environment R (R, 2006; Venables et al., 2002) for running these equivalence tests.

[5] We would like to acknowledge the assistance of Dr. Andrew Robinson in developing and providing the equivalence package as a result of some personal communications.

CHAPTER

VI. ASSESSING THE RBM APPROACH WHEN THE RELATIONSHIPS AMONG CONFIDENTIAL ATTRIBUTES ARE LINEAR

In this chapter, we assess the performance of the RBM approach in terms of *data utility* and *data security* when relationships among confidential attributes are *linear*. Section VI.1 demonstrates the effectiveness of some of the proposed masking methods using the motivation example (see Section I.3) and some of the measures in Chapter V. Section VI.2 discusses briefly how a snooper might try to learn more from the release of RBM-masked datasets involving non-monotonic relationship. Section VI.3 discusses how the characteristics of original datasets drive and define the characteristics of masked attributes. This includes the possible level of security RBM can provide while maximizing data utility. Section VI.4 investigates *empirically* the concepts Section VI.2 and Section VI.3 discuss.

VI.1. Illustration of the Effectiveness of NM-EGADP Approach using the Motivation Example

In this section, we want to test the effectiveness of the Relationship-Based NM-EGADP approach on the store dataset in the motivation example (Section I.3) using the adopted data security measures and some of the data utility measures developed in the previous chapter. We begin by masking the confidential attributes in the store dataset using the two NM-EGADP methods (Perturbation and Shuffling). Table 42 (Appendix D)

shows the first ten and last ten masked values of the two confidential attributes along with their original values. Figure 41 (Appendix D) shows the relationship between all variables in the original unmasked store dataset. The similarity of this figure with Figure 44 and Figure 45 (Appendix D) for masked datasets provides visual evidence that the Relationship-Based NM-EGADP approach can maintain original relationships in masked datasets.

VI.1.1. Data Utility

By comparing the (linear and nonlinear) parameter-based data utility measures, Table 49 to Table 54 (Appendix D), the parameters obtained from the masked datasets are similar to the ones obtained from the original dataset, indicating that original relationships (including non-monotonic relationships) are reproduced well in masked datasets. Prediction-based mean squared errors (MSE) data utility and data security measures using (non-monotonic) LS-SVM regression are presented in Table 3 for NM-EGADP perturbed store dataset. For comparison purposes, we also calculate MSE measure using piecewise linear regression because the non-monotonic relationships between S_1 and confidential attributes X consist of two lines (one for $S_1<40$ and another for $S_1\geq 40$, as you can see from Figure 2 (Section I.3) and Figure 41 (Appendix D)). The difference between the two LS-SVM regression models of confidential attributes **X** given the non-confidential attributes **S** is 0.65953, which is approaching zero indicating the predictive similarity of the model obtained from the original data with the one obtained from the masked data. This measure gets even better (0.090638) when we utilize the information about the actual shape of the relationship (piecewise linear) and use liner regression to calculate MSE. We reach a similar conclusion about the similarity of

Table 3. Prediction-based (MSE) data utility and data security measures for the NM-EGADP perturbed store dataset

	Measure	LS-SVM Regression	Piecewise Linear Regression
Data Security	$E[X-E(X\|SY)]^2 = E[X-E(X\|S)]^2$	232.40	229.54
		227.00	229.65
	$E[X-E(X\|S)]^2 <= E[X-E(X\|Y)]^2$	227.00	229.65
		343.49	338.73
Data Utility	$E[E(X\|S)\|_s-E(Y\|S)\|_s]^2=0$	0.65953	0.090638
	$E[E(X_1\|\ X_2)\|_{x2}-E(Y_1\|\ Y_2)\|_{x2}]^2=0$	1.0868	0.014977

models obtained from confidential attributes with models obtained from masked attributes (refer to the last row in Table 3).

The results of piecewise linear regression are better relative to the results of LS-SVM regression (even in the case of security measures). This is an indication of the difficulty in learning nonlinear relationships. Another reason is that learning mechanisms, such as ANN approaches, that mainly depend on the data to learn relationships and assume nothing about the relationships, tend to over fit the data (Rud, 2001). The normalized *data utility* measures for relationships using R^2 for X_1 for three different relationships are:

- $R^2(X_1|\mathbf{S})$: 92.85% = $R^2(Y_1|\mathbf{S})$: 92.81% (confidential with non-confidential)
- $R^2(X_1|X_2)$: 85.73% = $R^2(Y_1|Y_2)$: 85.65% (among confidential)
- $R^2(X_1|\mathbf{S},X_2)$: 93.95%= $R^2(Y_1|\mathbf{S},Y_2)$: 92.20% (confidential with non-confidential and confidential)

The results show that, in a *predictive* sense, original relationships are preserved although there is a little discrepancy in the last result, which can be attributed to the difficulty of

learning nonlinear relationships using non-parametric approach and its tendency to over fit the data (Rud, 2001).

The Pearson correlation matrices of the original dataset (Table 43, Appendix D), the perturbed dataset (Table 44, Appendix D), and the shuffled dataset (Table 45, Appendix D) are similar, indicating that linear relationships are well maintained in masked datasets. This holds true although many correlations are weak to begin with because of the existence of non-monotonic relationships. However, maintaining the (weak) product-moment correlation matrices of original and masked datasets to be similar can be useful. For example, if a statistician has access to the original dataset and decides that a specific transformation is needed before running a standard statistic method, then (s)he can obtain similar results from masked and original datasets in case (s)he is not allowed to access the original dataset. Similarly, the Spearman rank-order correlation matrices of the original dataset (Table 46, Appendix D), the perturbed dataset (Table 47, Appendix D), and the shuffled dataset (Table 48, Appendix D) are very similar, indicating that monotonic nonlinear relationships are also well reproduced in masked datasets, which is an advantage for PPE.

VI.1.2. Data Security

For data security (identity disclosure), there are only *seven* records (0.7 percent) re-identified using probabilistic record linkage programs. This is a good result and represents a good first wall of defense. Combined with the fact that masking methods provide automatic protection against exact value disclosure (a second wall of defense), it gets better. The two requirements for data security (partial or inferential value disclosure) using MSE measures are met (as shown in Table 3 above). The results for both data

security requirements are almost perfect in the case of piecewise regression (229.54=229.65 and 229.65≤ 338.73). On the other hand, while the measure of one data security requirement using LS-SVM regression is very good (227 ≤ 343.49), there is a slight difference in the other requirement (227= 232.40), and this can be attributed again to the tendency of ANN approaches to over fit the data (Rud, 2001). Nonetheless, this result (i.e. 227= 232.40) still suggests that **S** is a better predictor for **X** than **Y**. The normalized form using R^2 of the two *data security* requirements for X_1 is:

- $R^2(X_1|\mathbf{S})$: 92.85% $= R^2(X_1|\mathbf{S},\mathbf{Y})$: 92.69%
- $R^2(X_1|\mathbf{S})$: 92.85% $\geq R^2(X_1|\mathbf{Y})$: 89.10%,

and for X_2 is:

- $R^2(X_2|\mathbf{S})$: 93.25% $= R^2(X_2|\mathbf{S},\mathbf{Y})$: 93.04%
- $R^2(X_2|\mathbf{S})$: 93.25% $\geq R^2(X_2|\mathbf{Y})$: 89.91%.

Clearly, the above results represent effective protection against partial value disclosure.

For the motivation example, we find that the *CC* between the masked variables **Y** and the original confidential variables **X** is 0.96662, which is very high and represents a security threat. This is in agreement with the high predictive scores in the above measures. In addition, the *CC* between the non-confidential attributes and the confidential attributes is 0.082266, which is very low. This may mean that **Y** is a better predictor of the confidential attributes than the non-confidential attributes **S** (instead of the reverse, which is the desirable). In other words, this seems to suggest that the Relationship-Based NM-EGADP approach does not satisfy the data security requirements. Thus, releasing the masked dataset in this case will increase the snoopers' prediction ability. To recall, the data security requirements state that the best predictor of the confidential attributes **X**

are the non-confidential attributes **S**. Therefore, a snooper always will use the non-confidential attributes **S**. When the snooper tries to improve his/her prediction power by using both the non-confidential **S** and the masked **Y** attributes to predict the confidential attributes **X**, the masked variables **Y** do not increase the accuracy of prediction. This is because **X** and **Y** are independent given **S**.

However, before we jump to the conclusion that RBM NM-EGADP approach does not satisfy the data security requirements, let us remember that the canonical correlation analysis fails to detect nonlinear relationships. By referring to Figure 41 (Appendix D), we notice there is a very strong non-monotonic relationship between S_1 and X_1. This relationship consists of two lines. Therefore, *piecewise CC* security measures are applicable. By calculating the piecewise *CC* for each line (S_1<40 and S_2≥40), we find very strong *CC*: 0.98498 and 0.98126, respectively. Both are higher than the one we get from using **Y** to predict **X** (0.96662). When we use both **S** and **Y** to predict **X**, we get: 0.98498 and 0.98126 (piecewise). This indicates that the Relationship-Based NM-EGADP approach satisfies the data security requirements. For comparison purposes, when **Y** alone is used to predict **X**, the piecewise *CC* is 0.96771 and 0.96533, respectively. In other words, the problem lies with the dataset itself and not the masking methods. This becomes clearer by calculating the security index (*SI*) for X_1 and X_2 (see Subsection III.3.2): $SI(X_1)$ = 6.67% and $SI(X_2)$ = 6.31%. Obliviously, the residuals **r** explains a very small portion of the variation in the confidential attributes **X** and most of the variation in **X** explained by (a function of) **S** (i.e. $E(\mathbf{X}|\mathbf{S})$) with which to begin. This can also be seen from another perspective. The *piecewise CC* between the non-

confidential attributes **S** and the confidential attributes **X** is very high (0.98) indicating a strong relationship, which explains most of the variability in **X** with which to start.

Therefore, the store should not release this specific dataset. The store should be advised that they must always evaluate the strength of the relationship between the non-confidential attributes **S** and the confidential attributes **X** based on *CC* and MSE security requirements before they decide to release a specific dataset. Releasing datasets with very strong relationship (e.g. 0.80 or higher *CC* or piecewise *CC* correlation) may not be a good idea. Nevertheless, this depends on the sensitivity and the variance of confidential attributes. Another option is to treat those non-confidential attributes that have very strong relationships with confidential attributes as confidential attributes and mask them, unless the outside world already has access to those non-confidential attributes. In this case, it is safer not to release such datasets.

To assess the security of our method(s), we simulate a new 1000-record dataset, which is similar to the store dataset. However, it only contains two variables: one non-confidential **S** and one confidential **X**, which correspond to S_1 and X_1 in the store dataset, with less *piecewise CC* than the one we encounter in the store dataset (0.98). The goal is to obtain some evidence that our NM-EGADP methods (both perturbation and shuffling) really satisfy the MSE security requirements. Table 4 shows the first five and last five records of the original attributes (non-confidential **S** and confidential **X**) along with the corresponding NM-EGADP perturbed attribute **Y** and NM-EGADP shuffled attribute **Y_SHF**. Figure 5 is divided into three subfigures showing the relationship between the non-confidential and the confidential attributes along with the corresponding relationships between the non-confidential and the perturbed attributes, and the non-

Table 4. First 5 and last 5 records of the two-variable dataset, to check the *piecewise CC* security of NM-EGADP masking procedures

NO	S	X	Masked	
			Perturbed Y	Shuffled Y_SHF
1	53.12	45.89	62.94	62.63
2	57.37	32.81	78.142	78.52
3	22.06	51.14	52.149	51.51
4	25.46	72.10	44.255	42.91
5	51.25	65.57	64.215	64.20
:	**:**	**:**	**:**	**:**
996	37.16	96.33	92.00	92.69
997	50.72	70.32	65.18	65.18
998	28.71	69.37	24.53	28.05
999	27.71	89.99	56.87	56.31
1000	29.62	63.57	76.55	76.54
Min	20.01	9.7333	0.49153	9.7333
Max	59.97	132.84	123.13	132.84
Mean	40.264	62.722	62.722	62.722
STD	11.884	19.201	19.195	19.201

confidential and shuffled attributes, respectively. By comparing the three figures, they are similar and there is a strong graphical indication that our methods preserve conditional expectations (relationships).

The *CC* between the confidential attributes **X** and the masked attribute **Y** is 0.35875. The *CC* between the confidential attributes **X** and the non-confidential attributes **S** is 0.017245. However, the low canonical correlation can be, as we saw earlier, an indication of either a weak *linear* relationship between the two variables or the existence of *nonlinear* relationship. It is clear from Figure 5 that the relationship is non-monotonic. Using the *piecewise CC* (two lines: S<40 and S≥40), we get the following measures: 0.6035 and 0.58826. Both indicate a stronger relationship with **X** than **Y**. In addition, when we measure the *piecewise CC* between **X** on one side and **S** and corresponding **Y** values on the other side (i.e. *piecewise CC*(**X**|**SY**)), we get 0.6035 and 0.58832. This means that the snooper will definitely use **S** to predict **X** values, and adding **Y** to the equation will not enhance his/her ability to predict the confidential values. We got similar

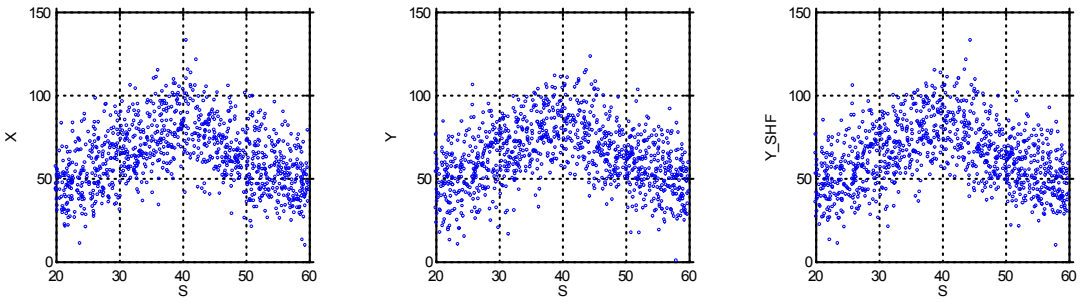


Figure 5. Diagrams of the two-variable dataset (aimed to check the piecewise *CC* security of NM-EGADP masking procedures) and its masked datasets.
(a) non-confidential attribute S vs. confidential attribute X, (b) non-confidential attribute S vs. NM-EGADP perturbed attribute Y, and (c) non-confidential attribute S vs. NM-EGADP shuffled attribute Y_SHF.

results for the NM-EGADP shuffling. Hence, NM-EGADP approach (using perturbation and shuffling) satisfies the R^2 and MSE security requirements and provides optimal security.

In addition, we present some graphical evidence in "Appendix E – Graphical Pilot Study – Comparisons for PPE Masking Methods" that the NM-EGADP shuffling procedure works well and preserves all types of relationships (including non-monotonic ones) while EGADP (shuffling) does not preserve nonlinear relationships and (C-GADP based) data shuffling only maintains monotonic relationships. .

VI.2. How a Snooper May Compromise RBM Masked Data

This section starts by discussing the snooper model presented by Fuller (1993). Then, it uses this model as a basis for discussing how a snooper may deal with RBM-masked datasets. The assumption in the (two-stage) Fuller's snooper model is that original datasets are normally distributed, which ensures all relationships are linear. In addition, *all* attributes in the original datasets are masked before the release of the data. The assumption is also that the snooper has a target individual with some known values for few attributes. The snooper will try to take advantage of the released masked data to

learn more about the original values of the other masked attributes for the target. This process consists of two stages: re-identification of the target record (identity disclosure) and predicting the original values of the masked attributes more accurately (value disclosure).

At the first stage, the snooper tries to re-identity the masked record belonging to the target individual. This is needed because *all* attributes are masked and no direct match is possible. Fuller (1993) discussed two probabilistic approaches based on whether the snooper is sure that the masked record for the target is among the released data. Both approaches use the known values as a basis for computing the probability that a specific masked record is the target's record. The masked record with the highest probabilistic match is assumed to be the target record. Refer to Equations 2.10 to 2.12 (pp. 388) and the discussion of them in Fuller (1993) for more information.

After re-identification, the snooper will try to enhance his prediction of the unknown values in the target record (than what was possible before) by using both the known pre-release values and the masked values from the released record. Enhancing prediction means less variance in the prediction error. To facilitate our discussion and the comparison with the RBM approach, we shall call the unknown values of the target **X**, the known values **S,** and the values in the masked record **Y**.

It is known from the estimation theory that the conditional expectation $E(\mathbf{X}|\mathbf{S})$ is the best predictor for **X** using **S**. This means that $E(\mathbf{X}|\mathbf{S})$ has the minimum variance of prediction error. When the data is normally distributed, the conditional expectation $E(\mathbf{X}|\mathbf{S})$ is linear in the form of:

$$E(\mathbf{X}|\mathbf{S}) = b_0 + b_1\mathbf{S}. \qquad (6.1)$$

Moreover, this conditional expectation can be calculated before the release of the masked data, as long as we know the mean and the covariance of the original data in addition to the few known values **S** for the target record (refer to Equation 2.3 in Fuller (1993)). Once the masked data is released and the target record is successfully re-identified, the snooper will try to use the masked data to enhance his prediction and obtain tighter prediction interval than the one (s)he obtains from (6.1). Thus, (s)he will try to fit a prediction model for unknown values in the re-identified target record in the form of:

$$E(\mathbf{X}|\mathbf{S}\,\mathbf{Y}) = b_0 + b_1\mathbf{S} + b_2\mathbf{Y}. \qquad (6.2)$$

Refer to Equations 2.5 to 2.9 (pp. 387) and the discussion of them in Fuller (1993) for more information.

As said earlier, the Fuller's snooper model assumes that all attributes in original datasets are masked and the non-confidential attributes **S** are null. This assumption emphasizes the importance of the re-identification step. Conversely, the RBM approach assumes the existence of **S** attributes, which can be numeric, categorical or both. When **S** is categorical, the re-identification step is still required as long as there is enough frequency of each category and *unique*, or a *small* number of, combinations of the categorical attributes *do not* exist, especially when they include the target record(s).

When **S** attributes are numeric or combinations of numeric and categorical, 100 percent re-identification rate is most likely to occur automatically (for all or most records) once the masked data is released. This is because most numeric values tend to be distinctive, and their uniqueness easily leads to *exact* record identification. Thus, the snooper does not usually need to go through the re-identification stage when (s)he deals

with RBM-masked datasets. Therefore, we *intentionally* ignored the **S** variables when we tried to assess the performance of the RBM approach in terms of re-identification in other sections in this study.

In Fuller's snooper model, all relationships are linear due to normality. When this is the case, sophisticated masking methods, such as GADP (Muralidhar et al., 1999) and EGADP (Muralidhar and Sarathy, 2005b), can be used without increasing the risk of value disclosure beyond what the (linear) conditional expectation in (6.1) explains before the release of the data. Hence, these masking methods reduce Equation (6.2) to Equation (6.1) (i.e. $b_2 = 0$).

As seen earlier, when the relationships between **S** and **X** are linear, the covariance matrix along the mean of original data and the few known values **S** can be used to estimate the conditional expectations $E(\mathbf{X}|\mathbf{S})$ before even the release of the masked data (see Equation (6.1)). However, when the relationships between **S** and **X** are non-monotonic as in RBM-masked datasets, the covariance and the mean are not useful for estimating the conditional expectations $E(\mathbf{X}|\mathbf{S})$ since the relationships are not linear any more as in Equation (6.1).

The snooper may or may not know about the existence of *non-monotonic* relationships in original datasets before the release of the RBM-masked data. Nonetheless, the *non-monotonic* relationships are automatically disclosed once the masked data is released. However, their disclosure is not considered a threat or a drawback. Actually, it is one of the main goals of the RBM approach; namely, to preserve and reproduce different types of relationships in masked data as they exist in original datasets for enhancing data utility in PPE. Moreover, the conditional expectations

characterize the whole dataset at the *attributes level* and not the confidential values of each individual at the *record level*.

As usual, the snooper is better off using all available information to improve her/his prediction. Therefore, since the conditional expectation $E(\mathbf{X}|\mathbf{S})$ is the best predictor for **X** using **S** (i.e. it has the minimum variance of prediction error) and the values of **S** are known for all records in the released datasets, the snooper will first try to estimate the conditional expectations $E(\mathbf{X}|\mathbf{S})$ by estimating $E(\mathbf{Y}|\mathbf{S})$. This is because the RBM approach specifies that $E(\mathbf{Y}|\mathbf{S}) = E(\mathbf{X}|\mathbf{S})$. $E(\mathbf{X}|\mathbf{S})$ is usually estimated from the following model:

$$\mathbf{X} = E(\mathbf{X}|\mathbf{S}) + \varepsilon^{\text{int}} \tag{6.3}$$

where $Var(\varepsilon^{\text{int}})$ is the variance of the *initial* prediction error that the snooper plans to reduce. Note the similarity between (6.3) and (6.1) when the conditional expectations are linear.

Next, the snooper will try to combine **Y** and **S** (or the evaluation of the best function of **S** that explains **X** (i.e. the fitted values $E(\mathbf{X}|\mathbf{S})$|s)) to improve her/his prediction accuracy of **X** and obtain less variance in the prediction error. The new prediction model might be written as:

$$\mathbf{X} = E(\mathbf{X}|\mathbf{Y}, E(\mathbf{X}|\mathbf{S})|_{\mathbf{S}}) + \varepsilon^{new} \tag{6.4}$$

where $Var(\varepsilon^{new})$ is the variance of the *new* prediction error that the snooper hopes to be less than the variance of the *initial* prediction error in (6.3) (i.e. $Var(\varepsilon^{new}) < Var(\varepsilon^{\text{int}})$). When the relationships between **X** and **Y** and the relationships between **X** and $E(\mathbf{X}|\mathbf{S})|_{\mathbf{S}}$ are linear (which is reasonable to assume, as we will discuss shortly in this

section and in the following section), the relationship in (6.4) might take the following linear form:

$$E(\mathbf{X}|\mathbf{Y}, E(\mathbf{X}|\mathbf{S})|_{\mathbf{S}}) = b_0 + b_1 E(\mathbf{X}|\mathbf{S})|_{\mathbf{S}} + b_2 \mathbf{Y}. \quad (6.5)$$

However, the snooper gains nothing because **X** and **Y** are independent (or orthogonal) given (the evaluation of) the best predictor $E(\mathbf{X}|\mathbf{S})$, and the relationship in (6.4) should reduce to:

$$\begin{aligned} E(\mathbf{X}|\mathbf{Y}, E(\mathbf{X}|\mathbf{S})|_{\mathbf{S}}) &= E(\mathbf{X}|E(\mathbf{X}|\mathbf{S})|_{\mathbf{S}}) \\ &= E(\mathbf{X}|\mathbf{S}) \end{aligned} \quad (6.6)$$

and $b_2 = 0$ in (6.5):

$$E(\mathbf{X}|\mathbf{Y}, E(\mathbf{X}|\mathbf{S})|_{\mathbf{S}}) = b_0 + b_1 E(\mathbf{X}|\mathbf{S})|_{\mathbf{S}}. \quad (6.7)$$

In terms of the variance of the prediction error, Equations (6.6) and (6.7) mean there is no improvement or reduction in the variance of the prediction error (i.e. $Var(\varepsilon^{new}) = Var(\varepsilon^{\text{int}})$).

Nevertheless, when the relationship between $E(\mathbf{X}|\mathbf{S})$ (or $E(\mathbf{X}|\mathbf{S})|\mathbf{s}$) and **X** is very strong in original datasets to begin with, there is no need to reduce the variance of the prediction error $Var(\varepsilon^{\text{int}})$ because it is already low. In this case, the security of masked data can be easily compromised in two ways: using what $E(\mathbf{X}|\mathbf{S})$ reveals about **X** or using what **Y** reveals about **X**.

In the first situation, the snooper gains considerable knowledge about the confidential attributes just from the release of the masked data since $E(\mathbf{Y}|\mathbf{S})$ is strongly correlated with **X**. This is because the RBM approach maintains $E(\mathbf{Y}|\mathbf{S})$ to be the same as $E(\mathbf{X}|\mathbf{S})$ for enhancing data utility, and $E(\mathbf{X}|\mathbf{S})$ explains most of the variation in **X**. We talked in Subsection III.3.2 about the Security Index (*SI*) measure, which is calculated as

1-$Var(r_i)/Var(X_i)$ or $Var(E(X_i|\mathbf{S})/Var(X_i)$, and how it can be used to assess whether there is enough variance in residuals (based on the characteristics of original datasets) for effective use of the RBM masking approach.

In the second situation, the snooper depends on the strength of the relationships between **X** and **Y**. The relationships between **X** and **Y** are *linear*, as mentioned earlier. The reason for this linearity is again that the RBM approach specifies $E(\mathbf{Y}|\mathbf{S}) = E(\mathbf{X}|\mathbf{S})$ for maximizing data utility. Hence, both are functions of **S** and we want them to be the same. Note that X and Y are not independent. They are *only* independent given the best predictor using S (i.e. $E(\mathbf{X}|\mathbf{S})$). If **S** is null, it is possible, at least hypothetically, to generate **Y** independently from **X**. In some special cases, such as the multivariate normal data, this is practically possible. Examples of this linearity between **X** and **Y** are presented in Figure 15 to Figure 22 in Section VI.4. We will discuss the *linearity* of relationships between **X** and **Y** more formally in the next section.

Since the cause of linearity between **X** and **Y** is the specification $E(\mathbf{Y}|\mathbf{S}) = E(\mathbf{X}|\mathbf{S})$, the strength of relationships between **X** and $E(\mathbf{X}|\mathbf{S})$ mentioned in the first point also has a direct impact on the strength of the *linear* relationships between **X** and **Y**. The stronger the (non-monotonic) relationships between **S** and **X**, the stronger the *linear* relationships between **X** and **Y**. Note that **Y** does not explain more about **X** than $E(\mathbf{X}|\mathbf{S})$ (more on this in the next section). Nevertheless, the snooper may try to fit a *simple* linear model between **Y** and **X** instead of fitting a more complicated model for the non-monotonic relationships between **S** and **X** knowing that (s)he may scarify a *little accuracy* for *simplicity*. As it is clear by now, the characteristics of original datasets play a main role in the effectiveness of applying the RBM approach.

VI.3. How the Characteristics of Original Datasets Determine the Characteristics of Masked Attributes

From the discussion in the previous section and in Section III.3, it is clear that the characteristics of original datasets shape the characteristics of masked attributes and dictate the maximum level of possible security using RBM approach. In this section, we express the characteristics of masked attributes **Y** in terms of the characteristics of original datasets (both non-confidential **S** and confidential attributes **X**).

More specifically, we derive some useful relationships and results (under the assumptions of normal residuals with constant variance) that relate every masked attribute Y_i to its original unmasked confidential attribute X_i based *solely on the characteristics of original datasets*. This allows owners of original datasets to assess the characteristics of masked attributes, including the strength of their association with confidential attributes after just estimating $E(X_i|\mathbf{S})$ and before masking. Thus, they can make an informed decision about whether to mask and release the data.

When $Y_i = E(X_i|\mathbf{S}) + e_i$, the *covariance* between a confidential attribute X_i and its masked copy Y_i is:

$$Cov(X_i, Y_i) = Var(E(X_i|\mathbf{S})) \tag{6.8}$$

where $i = 1\text{K } q$.

By using (6.8), we derive the following important result for measuring the *association* (*correlation*) strength between X_i and Y_i when the RBM NM-EGADP approach is used for masking:

$$Corr(X_i, Y_i) = 1 - \eta \tag{6.9}$$

where

$$\eta = \frac{Var(r_i)}{Var(X_i)} \tag{6.10}$$

where r_i is from Equation (3.5) in Section III.3.1.

Note that η links the *correlation security measure* in (6.9) and the *security index measure* (*SI*) discussed in Subsection III.3.2. Equation (6.9) helps data owners to calculate the association (correlation) between X_i and Y_i, and evaluate whether that association level is acceptable for the sensitivity level of confidential attributes (from a *security* point of view). Refer to Subsection III.3.2 for more information about what we mean by the "*sensitivity level of confidential attributes*".

We can also write Equation (6.9) differently using Proposition 1.4.1 (c) pp. 34 in (Bickel and Doksum, 2001):

$$Corr(X_i, Y_i) = \frac{Var(E(X_i|\mathbf{S}))}{Var(X_i)} \tag{6.11}$$

Equation (6.11) leads us to (see pp. 36-37 in (Bickel and Doksum, 2001)):

$$\begin{aligned} Corr(X_i, Y_i) &= R^2(X_i|\mathbf{S}) \\ &= [Corr(X_i, E(X_i|\mathbf{S}))]^2 \end{aligned} \tag{6.12}$$

We also argue that the range for both correlations (i.e. $Corr(X_i, Y_i)$ and $Corr(X_i, E(X_i|\mathbf{S}))$) in (6.12) is between 0 and 1. For $Corr(X_i, Y_i)$, this is clear from (6.11) since it equals the division of two positive quantities (i.e. the variances).

In addition, the relationships between **X** (*observations*) and $E(\mathbf{X}|\mathbf{S})$ or the fitted values $E(\mathbf{X}|\mathbf{S})$|s (*predicted*) tend to be linear when the estimation mechanism, such as ANN, learns the conditional expectations well. When the conditional expectations are well estimated, the model (i.e. the estimated function of $E(\mathbf{X}|\mathbf{S})$) accounts for all

nonlinearity that may exist between **S** and **X** (see pp. 443 in (Pyle, 2003)). The linear relationship between **X** and $E(\mathbf{X}|\mathbf{S})$ is also clear from:

$$\mathbf{X} = E(\mathbf{X}|\mathbf{S}) + \mathbf{r}. \tag{6.13}$$

Based on the above argument and its validity, we also derive an *upper bound* for the correlation in (6.9) and (6.11):

$$Corr(X_i, Y_i) \leq Corr(X_i, E(X_i|\mathbf{S})) \tag{6.14}$$

where the *equality* only holds when $Corr(X_i, E(X_i|\mathbf{S})) = 1$ or $Corr(X_i, E(X_i|\mathbf{S})) = 0$. The former means that X_i is a complete deterministic function of **S** and $E(X_i|\mathbf{S})$ explains all variation in X_i (i.e. $Var(E(X_i|\mathbf{S})) = Var(X_i)$ and $Var(r_i) = 0$). The latter means that the relationship between X_i and **S** is completely random (i.e.$Var(E(X_i|\mathbf{S})) = 0$ and $Var(r_i) = Var(X_i)$). This becomes more obvious when we consider the values that equate both correlations in Equation (6.12).

As known, any linear relationship between two random variables can be represented in the form of regression line:

$$X_i = b_0 + b_1 Y_i. \tag{6.15}$$

We derive analytically the coefficients of the regression line (b_0 and b_1) between X_i and Y_i before even generating Y_i. The intercept b_0 can be calculated as:

$$b_0 = \eta E(X_i) = \frac{Var(r_i)}{Var(X_i)} E(X_i) \tag{6.16}$$

and the slope b_1 can be calculated as:

$$\begin{aligned} b_1 &= Corr(X_i, Y_i) = 1 - \eta = 1 - \frac{Var(r_i)}{Var(X_i)} \\ &= \frac{Var(E(X_i|\mathbf{S}))}{Var(X_i)} \end{aligned} \tag{6.17}$$

For more discussion about the results in this section (and others), their derivations and proofs, and their connections to other security measures, refer to Appendix P.

VI.4. Assessing the Impact of the Characteristics of Original Datasets on the Effectiveness of the RBM Approach

As discussed in the previous two sections and in Section III.3, the characteristics of original datasets play an important role in the effectiveness of the RBM approach. One of the main characteristics of original datasets is the variance of the residuals that affects the level of protection that the RBM approach can provide while maximizing data utility. We want to *empirically* investigate this impact using three levels of variance of confidential attributes and three simulated datasets including the motivation example. In the first subsection, we discuss the three simulated datasets and discuss how the differences in their characteristics affect the security index (*SI*) measure and, accordingly, the generated masked attributes. In the second subsection, we present the data utility measures. In the third section, we talk about data security measures. For convenience, we may repeat some equations that appear in other parts of this study.

VI.4.1. Original Datasets and their Characteristics

The two new datasets have the same two non-confidential attributes **S** as in the motivation example (ME.L) dataset. They differ from the motivation example in that the confidential attributes **X** are generated with more variance than the motivation example. This reflects on the variance of the residuals r_i obtained by subtracting the conditional expectations $E(X_i|\mathbf{S})$ from X_i because we tried to fix the variance of the $E(X_i|\mathbf{S})$ as Table 5 and Table 6 may suggest. We call these two datasets "ME.L.MV1" and "ME.L.MV2"

while we call the motivation example dataset "ME.L". "ME" stands for motivation example. We use "L" in the datasets' names to point to the *linear* relationships between the confidential attributes X_1 and X_2. "MV" means more variance. The two numbers (1 and 2) are to distinguish between the two new levels of variance.

In this section, we present the calculations of some measures based on two estimation mechanisms for the conditional expectations: piecewise linear (regression) and LS-SVM. The reason we decide to use the piecewise linear estimation mechanism is that the conditional expectations for the three datasets are piecewise linear. Thus, we would like to utilize this extra information and see the impact of using (precise) conditional expectations versus using close (but not precise) LS-SVM estimated conditional expectations on some of the measures.

Figure 6, Figure 9 and Figure 12 show the scatter-plots matrices for ME.L, ME.L.MV1 and ME.L.MV2 datasets, respectively. The patterns among the fitted values **u**, when estimated using LS-SVM, and among their residuals **r** for each dataset are presented in Figure 7, Figure 10 and Figure 13, respectively. When the estimation mechanism used is piecewise linear regression, the corresponding figures are Figure 8, Figure 11 and Figure 14. Both sets of figures show (*general*) *linear* patterns among **u** corresponding to the linear patterns among **X**. Nevertheless, the linear patterns are sharper and more obvious when the piecewise linear regression estimation mechanisms is used.

Bickel and Doksum (2001) suggest that (Proposition 1.4.1 (c), pp. 34):

$$Var(X_i) = Var(E(X_i|\mathbf{S})) + Var(r_i). \qquad (6.18)$$

Equation (6.18) can be used as a measure to check the accuracy of learned conditional expectations. We present the calculation of this measure in Table 5 for X_1 and in Table 6 for X_2. When we use the piecewise linear estimation mechanism, the results *precisely* validate Equation (6.18). However, there are some differences between the RHS and LHS calculations of Equation (6.18) when the LS-SVM estimation is used. In addition, we provide the calculations of the security index measure (*SI*) for variables X_1 and X_2 for every dataset. *SI* is calculated as $1 - Var(r_i)/Var(X_i)$ or $Var(u_i)/Var(X_i)$. Note that $u_i = E(X_i|\mathbf{S})$. The lower the *SI* measure, the less security RBM can provide. Thus, ME.L is the least secure dataset and ME.L.MV2 is the most secure. However, the security has an inverse relationship with the data utility. The stronger the relationships (conditional expectations) between **X** and **S**, the more insecure the dataset.

We derived the following important result earlier and provided its proofs in Appendix B:

$$Cov(X_i, X_j) = Cov(E(X_i|\mathbf{S}), E(X_j|\mathbf{S})) + Cov(r_i, r_j). \quad (6.19)$$

Table 7 shows the calculations for this measure using the two estimation mechanisms (piecewise linear and LS-SVM). Since the conditional expectations for these datasets are piecewise linear, the measures precisely match the relation presented in Equation (6.19) when we use the piecewise-linearly estimated conditional expectations. When we use the LS-SVM estimation, slight discrepancies occur.

In addition, Equation (6.8) (i.e. $Cov(X_i,Y_i) = Var(u_i)$) expresses the covariance between X_i and Y_i in terms of one of the original datasets' characteristics (i.e. $Var(u_i)$). We calculated this measure for all four masked datasets for each dataset. Table 8 shows

the results for X_1 and Y_1 while Table 9 shows the results for X_2 and Y_2. The measures (almost) hold. We went further and masked the piecewise linear parts of the conditional expectations of the attributes separately. Then we calculated the same measures and listed them in the same two tables. We obtain comparable results to the previous (LS-SVM estimated) measures with one exception. The calculations of the measure in Equation (6.8) exactly hold for all datasets masked using masking method 1. Hence, this masking method adds an independent normally distributed noise (with the same covariance as with the residuals covariance) and does not employ any shuffling refinement.

We earlier argued that the relationships between **X** and **Y** are linear because RBM specifies that $E(\mathbf{Y}|\mathbf{S}) = E(\mathbf{X}|\mathbf{S})$. Figure 15 to Figure 20 provide graphical evidence for this claim. The more variance the residuals have, the weaker the linear relationships between **X** and **Y**. The correlation measures between **X** and **Y** show this trend. To make the comparison between the *linear* relationships between X_i and Y_i in the three datasets easier, we plot them next to each other using a unified scale (the scale of the data with the largest residuals variance). Figure 21 shows the three plots for comparing X_1 vs. Y_1 while Figure 22 shows the three plots for comparing X_2 vs. Y_2.

Since the relationship between X_i and Y_i is *linear*, we derived the coefficients of a regression line in the form of "$X_i = b_0 + b_1 Y_i$", as presented in Equations (6.16) and (6.17) in the previous section. These coefficients are calculated based on the characteristics of original datasets and before generating masking attributes. We calculated these coefficients for both X_1 (Table 10) and X_2 (Table 11) in two ways. First, we run linear regression models for each masked dataset and report the results in the left half of the two tables. On the right side of these tables, we present the regression

coefficients based on the calculations of Equations (6.16) and (6.17). Clearly, they are similar.

These regression coefficients can have a role to play in security since they relate the masked attributes **Y** to the confidential attributes **X**. We developed some correlation-based measures to assess the security that RBM can provide in generating **Y** based solely on the characteristics of original datasets in the previous section (more on this in the data security subsection).

Finally, we present the scatter-plots matrices of the masked datasets for the ME.L (Motivation Example) dataset in Appendix D. For the other two datasets, we present the scatter plots for their masked datasets at the end of this chapter.

Table 5. $Var(X_1) = Var(u_1) + Var(r_1)$

Method	Dataset	Var					SI
		(u1)	(r1)	(u1)+(r1)	(X1)	Diff%	
Piecewise Linear	ME.L	3278.720	237.615	3516.335	3516.335	0.00%	6.76%
	ME.L.MV1	3205.861	2523.685	5729.547	5729.547	0.00%	44.05%
	ME.L.MV2	3136.636	10094.741	13231.377	13231.377	0.00%	76.29%
LS-SVM	ME.L	3280.863	234.462	3515.326	3516.335	0.03%	6.67%
	ME.L.MV1	3234.839	2487.245	5722.084	5729.547	0.13%	43.41%
	ME.L.MV2	3254.133	9948.647	13202.780	13231.377	0.22%	75.19%

Table 6. $Var(X_2) = Var(u_2) + Var(r_2)$

Method	Dataset	Var					SI
		(u2)	(r2)	(u2)+(r2)	(X2)	Diff%	
Piecewise Linear	ME.L	3265.245	222.144	3487.389	3487.389	0.00%	6.37%
	ME.L.MV1	3086.605	1479.537	4566.142	4566.142	0.00%	32.40%
	ME.L.MV2	2923.988	5918.147	8842.135	8842.135	0.00%	66.93%
LS-SVM	ME.L	3256.945	219.989	3476.934	3487.389	0.30%	6.31%
	ME.L.MV1	3076.161	1469.744	4545.905	4566.142	0.44%	32.19%
	ME.L.MV2	2940.068	5860.834	8800.902	8842.135	0.47%	66.28%

Table 7. $Cov(X_1,X_2) = Cov(u_1,u_2) + Cov(r_1,r_2)$

Method	Dataset	Cov				
		(u1,u2)	(r1,r2)	(u1,u2)+(r1,r2)	(X1,X2)	Diff%
Piecewise Linear	ME.L	-3270.774	3.904	-3266.871	-3266.871	0.00%
	ME.L.MV1	-3144.049	23.621	-3120.428	-3120.428	0.00%
	ME.L.MV2	-3020.134	94.484	-2925.649	-2925.649	0.00%
LS-SVM	ME.L	-3262.102	1.306	-3260.796	-3266.871	0.19%
	ME.L.MV1	-3129.142	17.527	-3111.614	-3120.428	0.28%
	ME.L.MV2	-2993.171	73.047	-2920.124	-2925.649	0.19%

Table 8. Covariance between X_1 and Y_1 and its relationship to the variance of u_1

Method	Dataset	Var(u_1)	Cov(X_1, Y_1) for Masking Method			
			1	2	3	4
Peicewise Linear	ME.L	3278.72	**3278.72**	3281.77	3276.56	3282.03
	ME.L.MV1	3205.86	**3205.86**	3201.56	3202.97	3187.59
	ME.L.MV2	3136.64	**3136.64**	3051.19	3121.63	3049.08
LS-SVM	ME.L	3280.86	3281.41	3277.31	3286.91	3276.03
	ME.L.MV1	3234.84	3239.07	3221.72	3273.40	3237.25
	ME.L.MV2	3254.13	3270.15	3659.79	3306.64	3294.33

Table 9. Covariance between X_2 and Y_2 and its relationship to the variance of u_2

Method	Dataset	Var(u_2)	Cov(X_2, Y_2) for Masking Method			
			1	2	3	4
Peicewise Linear	ME.L	3265.24	**3265.24**	3269.02	3269.18	3269.30
	ME.L.MV1	3086.61	**3086.61**	3071.76	3099.30	3075.25
	ME.L.MV2	2923.99	**2923.99**	2915.84	2953.70	2917.46
LS-SVM	ME.L	3256.94	3261.43	3263.47	3262.40	3263.21
	ME.L.MV1	3076.16	3083.28	3061.68	3092.25	3073.27
	ME.L.MV2	2940.07	2954.36	2953.25	2983.78	2958.27

Table 10. Calculating the regression coefficients of $X_1 = b_0 + b_1 Y_1$ based on the characteristics of original data

$X_1 = b_0 + b_1 Y_1$					
Masked Data				Original Data	
Dataset	Masking Method	b_0	b_1	$b_0 = Var(r_1)/Var(X_1) * E(X_1)$	$b_1 = Var(u_1)/Var(X_1)$ [$= Corr(X_1, Y_1)$]
ME.L	1	19.747	0.933	19.884	0.933
	2	20.173	0.932		
	3	20.282	0.932		
	4	20.282	0.932		
ME.L.MV1	1	128.673	0.566	128.800	0.566
	2	129.746	0.562		
	3	128.405	0.567		
	4	128.942	0.565		
ME.L.MV2	1	222.844	0.247	223.175	0.248
	2	214.181	0.277		
	3	221.699	0.251		
	4	222.359	0.249		

Table 11. Calculating the regression coefficients of $X_2 = b_0 + b_1 Y_2$ based on the characteristics of original data

$X_2 = b_0 + b_2 Y_2$					
Masked Data				Original Data	
Dataset	Masking Method	b_0	b_1	$b_0 = Var(r_2)/Var(X_2) * E(X_2)$	$b_1 = Var(u_2)/Var(X_2)$ [$= Corr(X_2, Y_2)$]
ME.L	1	31.250	0.938	31.846	0.937
	2	32.406	0.936		
	3	31.359	0.938		
	4	32.443	0.936		
ME.L.MV1	1	162.061	0.679	162.460	0.678
	2	166.296	0.671		
	3	162.291	0.678		
	4	165.015	0.673		
ME.L.MV2	1	335.535	0.335	334.510	0.337
	2	336.163	0.334		
	3	334.842	0.337		
	4	335.876	0.335		

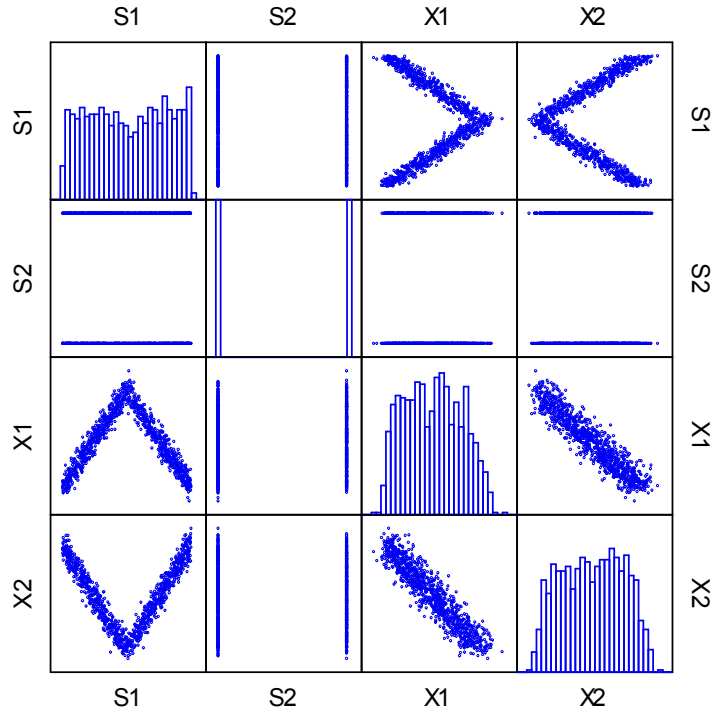


Figure 6. Motivation Example (ME.L) Dataset

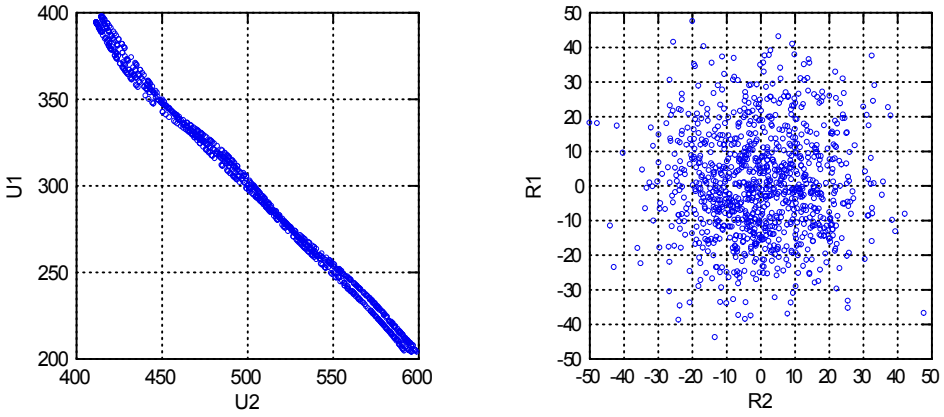


Figure 7. ME.L dataset: u_1 vs. u_2 and r_1 vs. r_2 scatter plots (LS-SVM)

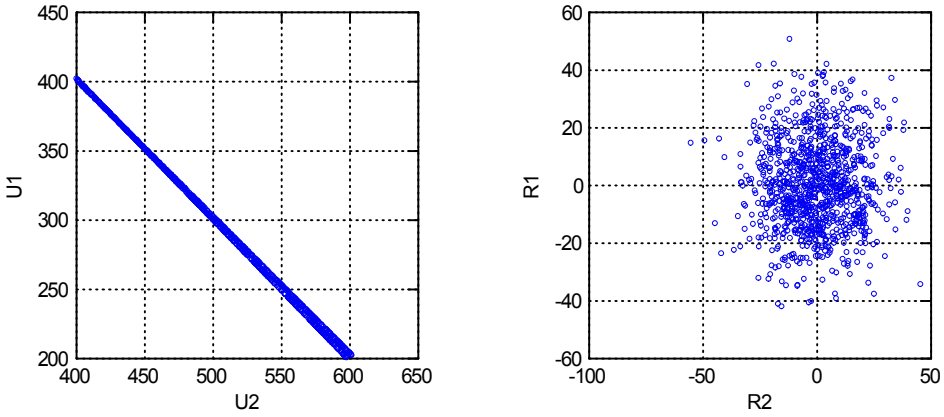


Figure 8. ME.L dataset: u_1 vs. u_2 and r_1 vs. r_2 scatter plots (piecewise linear)

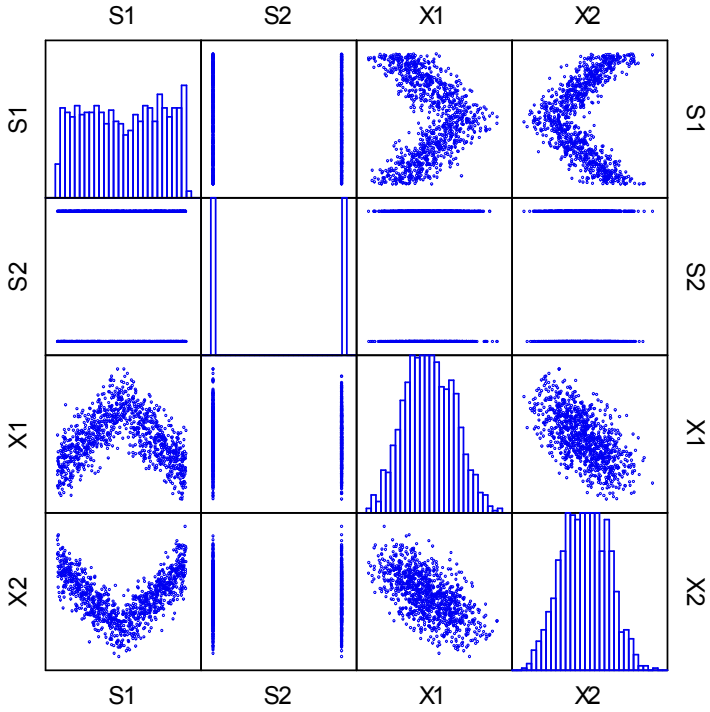


Figure 9. ME.L.MV1 Dataset

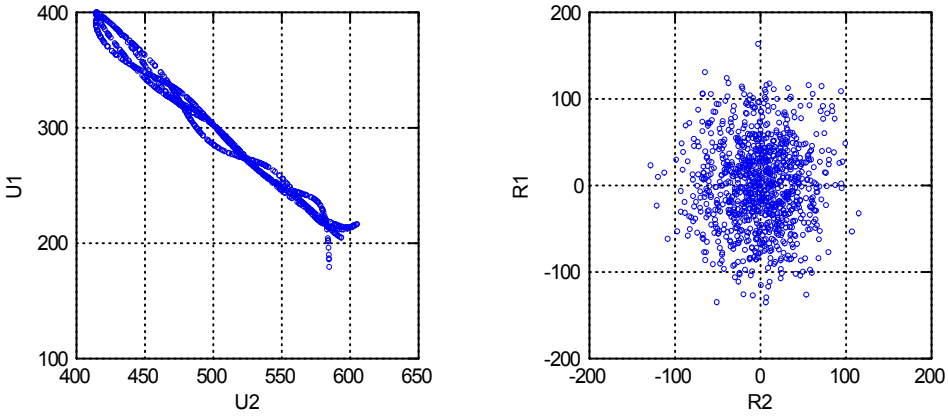


Figure 10. ME.L.MV1 dataset: u_1 vs. u_2 and r_1 vs. r_2 scatter plots (LS-SVM)

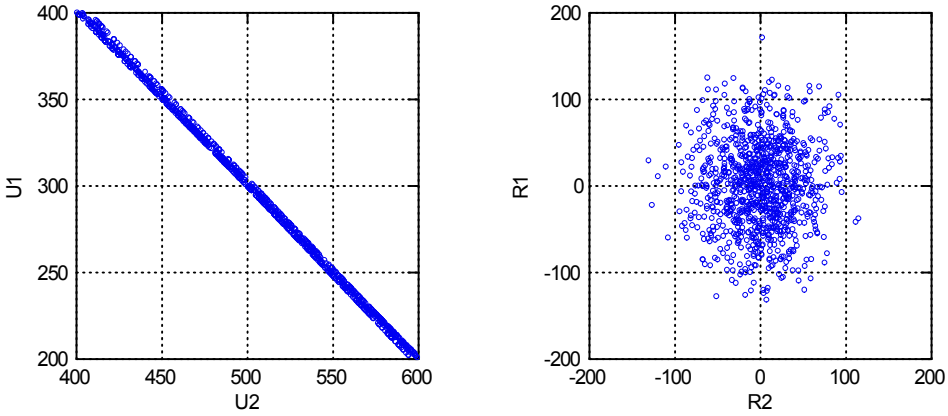


Figure 11. ME.L.MV1 dataset: u_1 vs. u_2 and r_1 vs. r_2 scatter plots (piecewise linear)

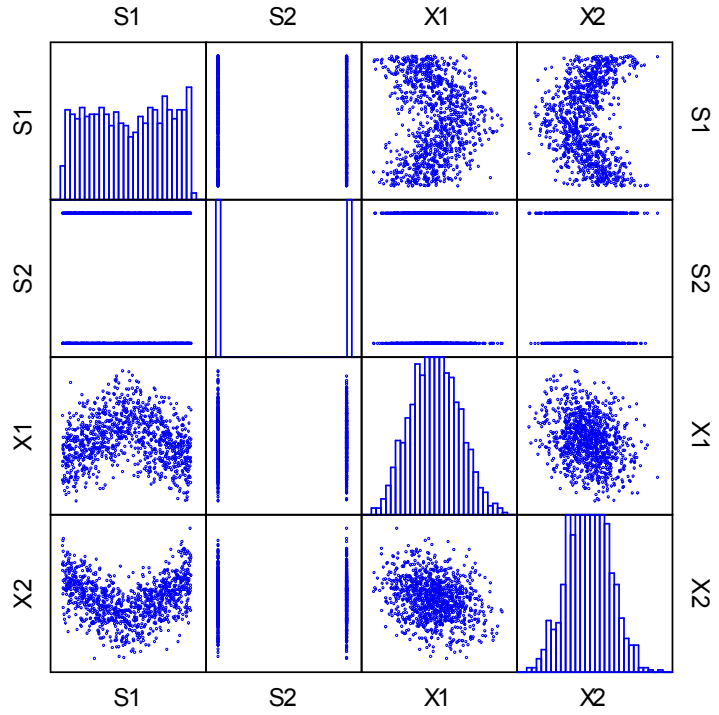


Figure 12. ME.L.MV2 Dataset

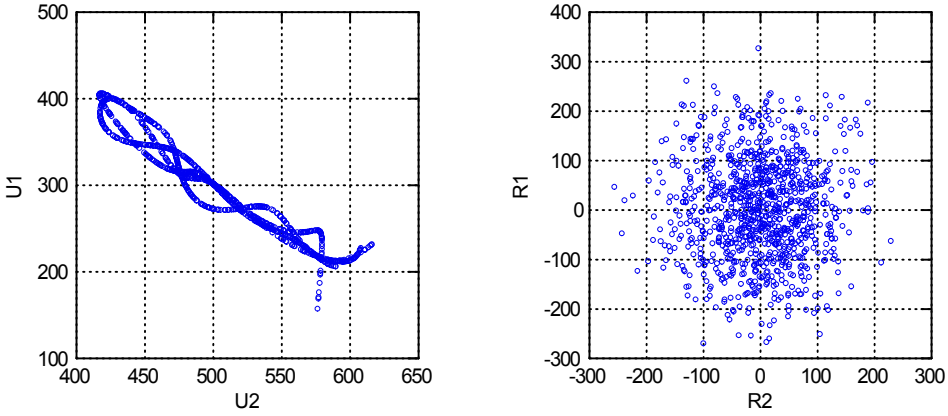


Figure 13. ME.L.MV2 dataset: u_1 vs. u_2 and r_1 vs. r_2 scatter plots (LS-SVM)

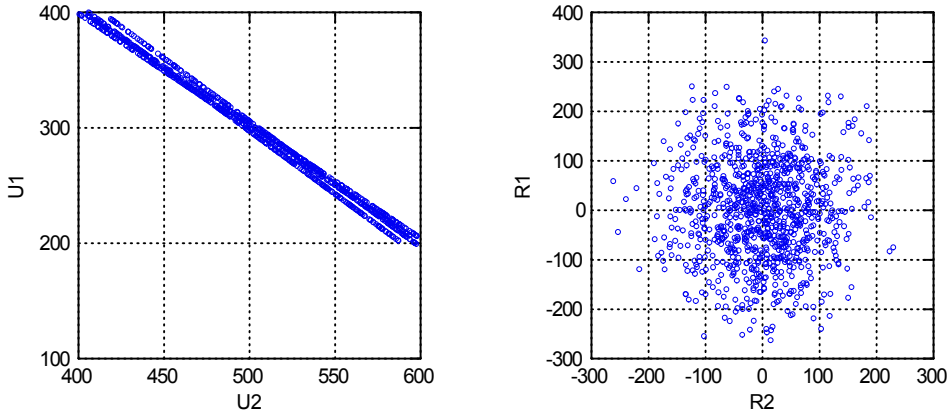


Figure 14. ME.L.MV2 dataset: u_1 vs. u_2 and r_1 vs. r_2 scatter plots (piecewise linear)

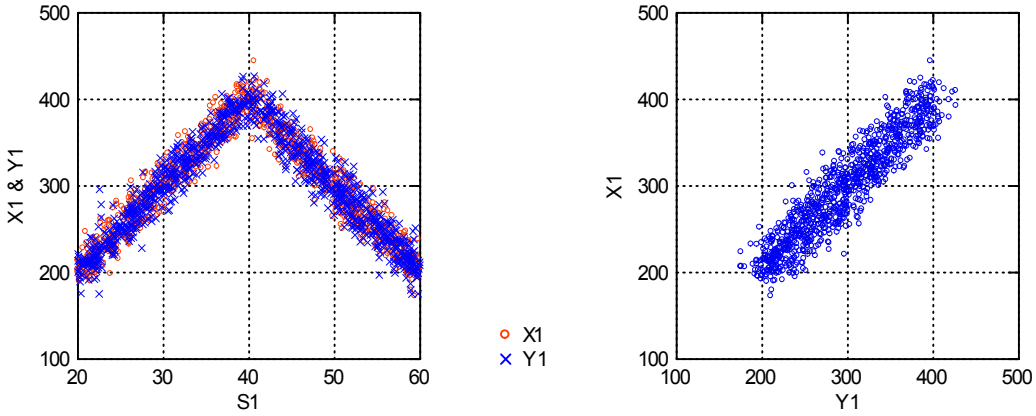


Figure 15. ME.L: X_1 & Y_1 vs. S_1 scatter plot and the linearity of relationships between X_1 & Y_1

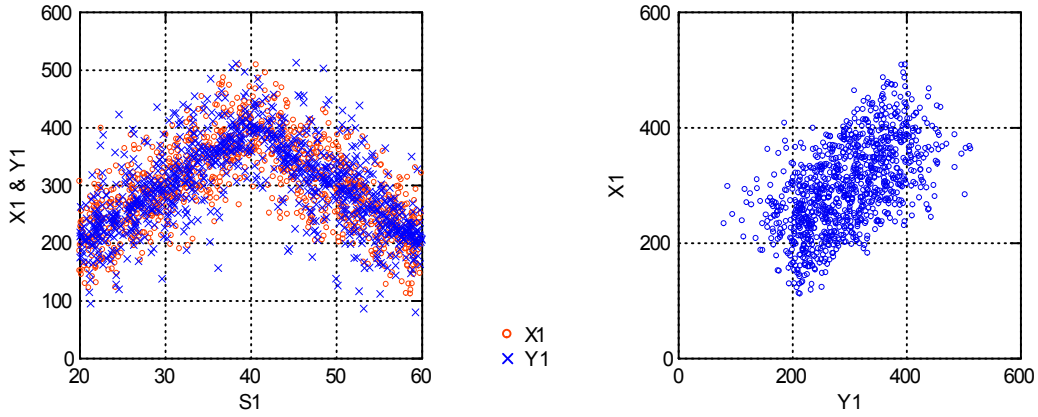


Figure 16. ME.L.MV1: X_1 & Y_1 vs. S_1 scatter plot and the linearity of relationships between X_1 & Y_1

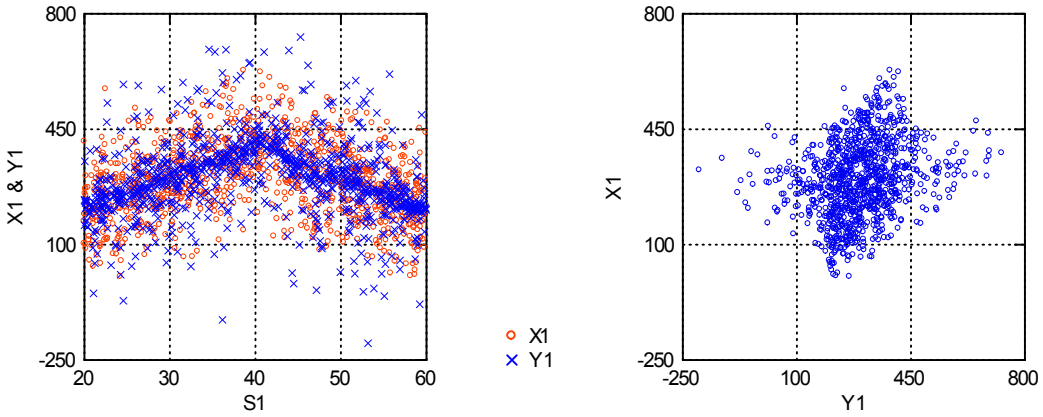


Figure 17. ME.L.MV2: X_1 & Y_1 vs. S_1 scatter plot and the linearity of relationships between X_1 & Y_1

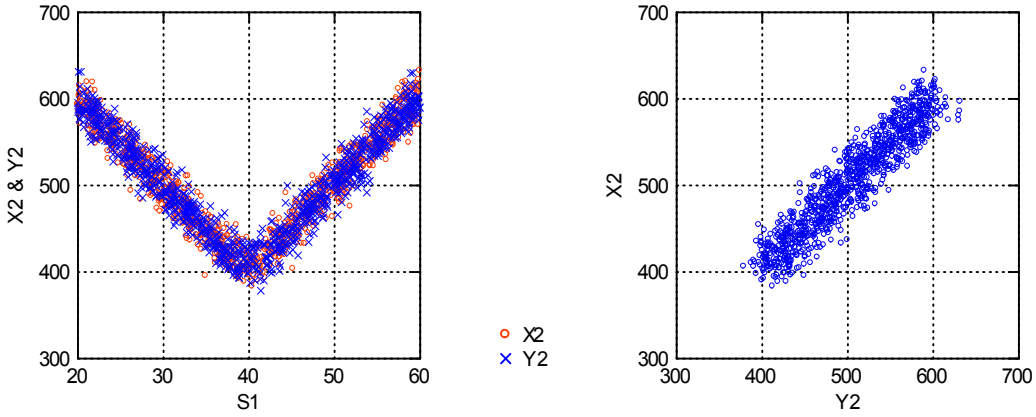


Figure 18. ME.L: X_2 & Y_2 vs. S_1 scatter plot and the linearity of relationships between X_2 & Y_2

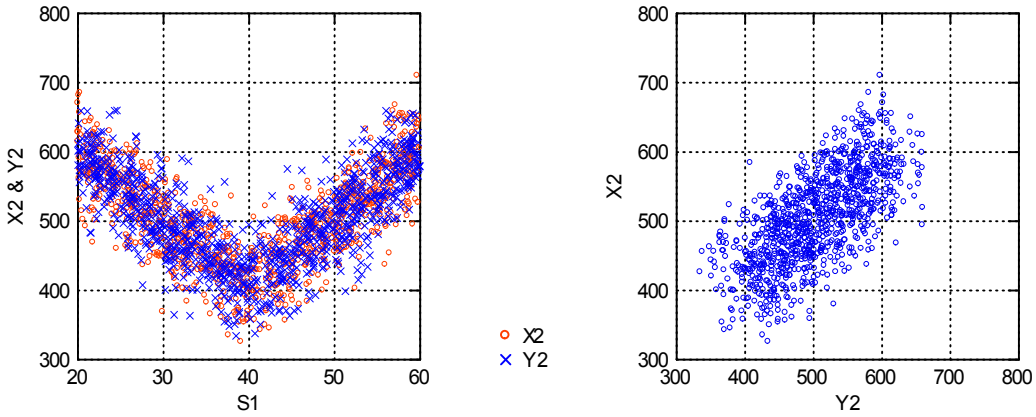


Figure 19. ME.LMV1: X_2 & Y_2 vs. S_1 scatter plot and the linearity of relationships between X_2 & Y_2

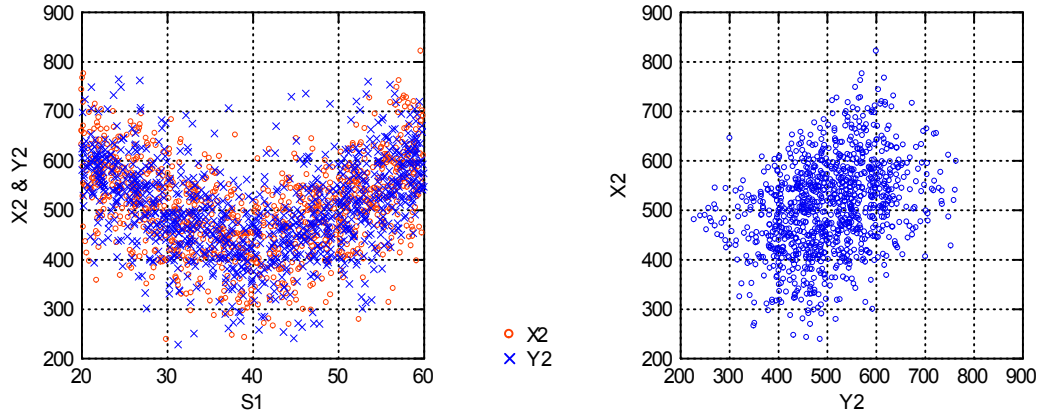


Figure 20. ME.LMV2: X_2 & Y_2 vs. S_1 scatter plot and the linearity of relationships between X_2 & Y_2

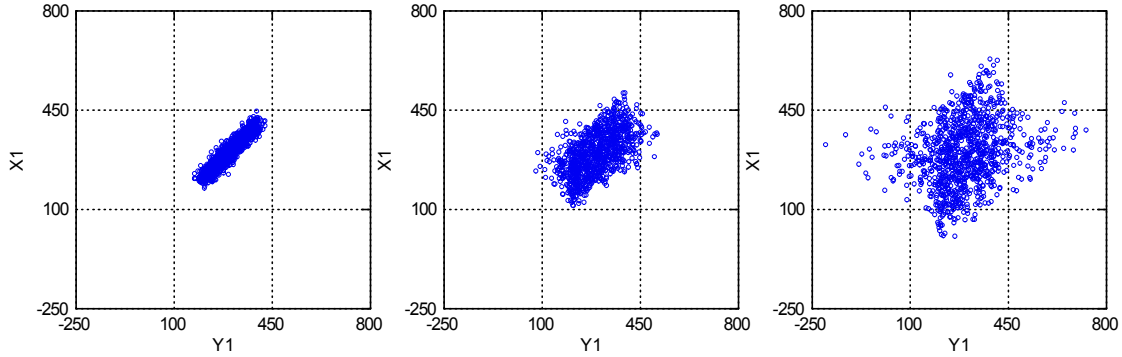


Figure 21. Comparing the linearity of relationships between X_1 & Y_1 for the three datasets on a unified scale

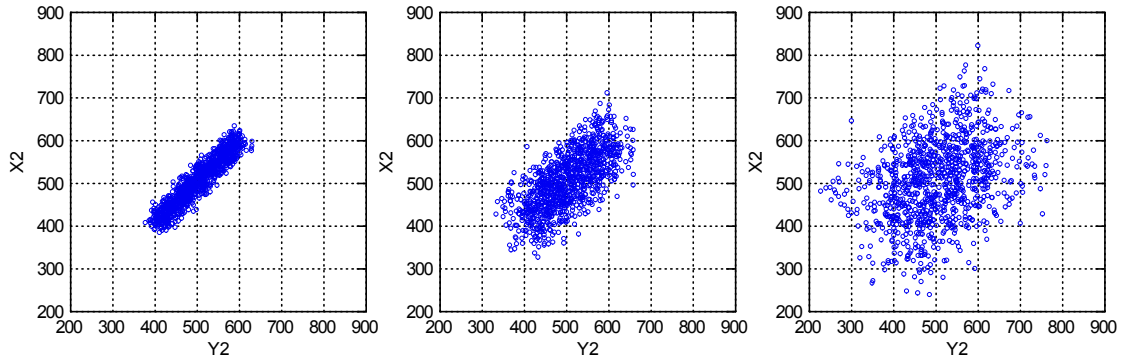


Figure 22 Comparing the linearity of relationships between X_2 & Y_2 for the three datasets on a unified scale

VI.4.2. Data Utility

In this section, we want to assess the data utility of the three masked datasets using the four RBM masking methods. We will use the equivalence tests measures discussed in Section V.3. For these tests, we need to calculate the fitted values from two models of which we want to test their equivalency. One model is built from original data and the other is built from the corresponding masked data. Each model represents the estimated conditional expectations (for specific sets of attributes divided into one dependent variable (DV) and other independent variables (IV)).

We estimate the conditional expectations and calculate the fitted values using two estimation approaches: *linear regression* and *LS-SVM*. Linear regression assumes that conditional expectations are linear. When this is the case, this estimation approach should outperform the LS-SVM estimation approach in the quality of estimated conditional expectations. Otherwise, LS-SVM approach is more flexible. Once the required conditional expectations are modeled, we evaluate the two models at the same data points from original datasets to generate the fitted values to test.

We are interested in six relationships listed in Table 12 and in Table 13. They represent three important classes of relationships. Table 12 shows the equivalence test results when the linear regression is used as an estimation mechanism while Table 13 shows the equivalence test results when LS-SVM is used as an estimation mechanism. Although we tested at 25 percent equivalence region, the listed numbers are the minimum equivalence regions required to pass the equivalence tests. These numbers represent percentages either of the mean of original data for the mean equivalence test or of a slope of one for the slope equivalence test.

For the *mean* equivalence test, both estimation mechanisms work well and all the results are significant, which leads to the rejection of the null hypotheses of dissimilar means. The minimum equivalence regions required to pass the test are very small especially in the case of testing linear conditional expectations. *RBM shows its effectiveness in preserving the mean of the predictions (fitted values) from masked data to be similar to the predictions of original data. Hence, RBM is effective in reproducing the mean of confidential attributes in masked attributes.*

When the linear regression estimation is used, the *slope* equivalence test could be used to assess the ability of the RBM approach to maintain *linear* relationships (as they are represented by the covariance matrix) regardless of the type of actual existing relationships (liner or nonlinear). Table 12 shows that all relationships in all datasets masked by the four RBM masking methods pass the slope equivalence tests except for two cases. ME.L.MV2 masked using masking methods 3 or 4 are the two exceptions. Otherwise, there is strong evidence that *RBM maintains linear relationships well*.

Moreover, masking method 1, which simply adds normal noise to **u**, slightly outperforms all other masking methods across (almost) all relationships and datasets.

Note that ME.L.MV2 is the dataset with most of the variation in **X** explained by **r** rather than **u** (r_1 explains 75.19 percent of the variation of X_1, for example). The source of the problem seems to be a combination of the amount of variation explained by **r** and the shuffling refinement, which tries to maintain marginal distributions. While masking method 3 applies shuffling refinement at the residuals level, masking method 4 applies it at two levels: residuals level and variables level. The common step is the shuffling at the residuals level. The shuffling at the residuals level for this case seems to destroy the orthogonality of the added noise, which biases the (weak to begin with) conditional expectations, especially with the large amount of variation in **r**. We do not face this problem when no shuffling refinement is used (i.e. masking method 1) or the shuffling refinement is done only at the variables level (i.e. masking method 2).

When LS-SVM estimation mechanism is used, all relationships between **S** and **X** (relationships 1 and 2 in Table 13) pass the slope equivalence tests including the one that did not pass the slope test using linear regression. Note that the relationships between **S** and **X** are non-monotonic. Hence, LS-SVM estimation is more appropriate than linear regression estimation in this case.

When we tested for relationships among **X**, two datasets did not pass the equivalence tests. In the case of $X_1|X_2$ vs. $Y_1|Y_2$, ME.L.MV1 did not pass the test when masked using method 1 or 3. Similarly, ME.L.MV2 did not pass the test when masked using method 1, 2 or 3. However, we know that the relationships between X_1 and X_2 and the relationships between Y_1 and Y_2 are linear for all three datasets (see Figure 21 and

Figure 22). Thus, the tests based on the linear regression (Table 12) might be more appropriate. Hence, all these cases pass the equivalence test, and the minimum required regions to pass the test are small. In addition, the parameter-based data utility measures for $E(X_1|X_2)$ vs. $E(Y_1|Y_2)$ (see Table 14) support the similarity of the predicted values of original and masked data.

This demonstrates an important point: the estimation mechanism should be able to learn the conditional expectations as well as possible. While the assumption of the linearity of conditional expectations when linear regression estimation mechanism is used hinders its ability to be used for testing the equivalence of nonlinear relationships, the problem of over-fitting in the LS-SVM estimation mechanism may affect the effectiveness and the accuracy of the results one obtains from the equivalence tests, especially as the variance of residual increases.

For the relationships $X_1|\mathbf{S}X_2$ vs. $Y_1|\mathbf{S}Y_2$ and $X_2|\mathbf{S}X_1$ vs. $Y_2|\mathbf{S}Y_1$, ME.L.MV1 and ME.L.MV2 datasets did not pass the slope equivalence test except for ME.L.MV1 dataset when masked using method 2 for relationship $X_2|\mathbf{S}X_1$ vs. $Y_2|\mathbf{S}Y_1$.

Finally, a general trend (with few exceptions) is that as the variance of the residual increases, maintaining the data utility becomes harder especially in the case of nonlinear relationships as estimated using the LS-SVM estimation mechanisms. This can be seen from the magnitude of the minimum required region to pass the equivalence test even when the test failed.

Table 12. Equivalence Tests using Linear Regression to Calculate the Compared Fitted Values

Relationship 1: $E(X_1\|S)\|s$ vs. $E(Y_1\|S)\|s$								
Dataset	Mean Equivalence Test				Slope Equivalence Test			
	1	2	3	4	1	2	3	4
ME.L	0.00%	0.00%	0.00%	0.00%	0.02%	2.07%	2.75%	1.56%
ME.L.MV1	0.00%	0.00%	0.01%	0.01%	0.12%	5.29%	10.01%	2.94%
ME.L.MV2	0.00%	0.01%	0.02%	0.01%	0.56%	0.58%	31.49%	29.37%
Relationship 2: $E(X_2\|S)\|s$ vs. $E(Y_2\|S)\|s$								
Dataset	Mean Equivalence Test				Slope Equivalence Test			
	1	2	3	4	1	2	3	4
ME.L	0.00%	0.00%	0.00%	0.00%	0.03%	1.53%	1.04%	1.14%
ME.L.MV1	0.00%	0.00%	0.00%	0.00%	2.06%	6.95%	3.96%	3.45%
ME.L.MV2	0.00%	0.00%	0.01%	0.01%	7.38%	6.98%	24.79%	12.78%
Relationship 3: $E(X_1\|X_2)\|x_2$ vs. $E(Y_1\|Y_2)\|x_2$								
Dataset	Mean Equivalence Test				Slope Equivalence Test			
	1	2	3	4	1	2	3	4
ME.L	0.00%	0.00%	0.00%	0.00%	0.12%	0.21%	0.28%	0.21%
ME.L.MV1	0.00%	0.00%	0.00%	0.00%	0.03%	1.66%	0.14%	1.81%
ME.L.MV2	0.00%	0.00%	0.00%	0.00%	0.94%	8.70%	1.67%	3.81%
Relationship 4: $E(X_2\|X_1)\|x_1$ vs. $E(Y_2\|Y_1)\|x_1$								
Dataset	Mean Equivalence Test				Slope Equivalence Test			
	1	2	3	4	1	2	3	4
ME.L	0.00%	0.00%	0.00%	0.00%	0.16%	0.21%	0.31%	0.21%
ME.L.MV1	0.00%	0.00%	0.00%	0.00%	0.47%	1.66%	0.84%	1.81%
ME.L.MV2	0.00%	0.00%	0.00%	0.00%	1.20%	8.70%	0.89%	3.81%
Relationship 5: $E(X_1\|SX_2)\|sx_2$ vs. $E(Y_1\|SY_2)\|sx_2$								
Dataset	Mean Equivalence Test				Slope Equivalence Test			
	1	2	3	4	1	2	3	4
ME.L	0.00%	0.00%	0.00%	0.00%	0.12%	0.22%	0.29%	0.22%
ME.L.MV1	0.00%	0.01%	0.01%	0.01%	0.03%	1.68%	0.20%	1.88%
ME.L.MV2	0.00%	0.01%	0.02%	0.01%	0.92%	8.70%	1.98%	4.07%
Relationship 6: $E(X_2\|SX_1)\|sx_1$ vs. $E(Y_2\|SY_1)\|sx_1$								
Dataset	Mean Equivalence Test				Slope Equivalence Test			
	1	2	3	4	1	2	3	4
ME.L	0.00%	0.00%	0.00%	0.00%	0.16%	0.22%	0.30%	0.22%
ME.L.MV1	0.00%	0.00%	0.00%	0.00%	0.48%	1.70%	0.84%	1.82%
ME.L.MV2	0.00%	0.00%	0.00%	0.00%	1.25%	8.74%	0.97%	3.87%

Table 13. Equivalence Tests using LS-SVM to Calculate the Compared Fitted Values

Relationship 1: $E(X_1\|S)\|s$ vs. $E(Y_1\|S)\|s^*$								
Dataset	Mean Equivalence Test				Slope Equivalence Test			
	1	2	3	4	1	2	3	4
ME.L	0.01%	0.01%	0.01%	0.01%	0.04%	0.15%	0.23%	0.15%
ME.L.MV1	0.02%	0.02%	0.03%	0.04%	0.10%	0.51%	1.00%	0.26%
ME.L.MV2	0.03%	0.08%	0.10%	0.10%	0.23%	9.99%	0.83%	1.09%
Relationship 2: $E(X_2\|S)\|s$ vs. $E(Y_2\|S)\|s^*$								
Dataset	Mean Equivalence Test				Slope Equivalence Test			
	1	2	3	4	1	2	3	4
ME.L	0.02%	0.02%	0.02%	0.02%	0.28%	0.22%	0.26%	0.22%
ME.L.MV1	0.04%	0.04%	0.04%	0.04%	0.73%	1.28%	0.46%	0.95%
ME.L.MV2	0.06%	0.06%	0.06%	0.06%	0.69%	0.95%	1.33%	0.65%
Relationship 3: $E(X_1\|X_2)\|x_2$ vs. $E(Y_1\|Y_2)\|x_{2^*}$								
Dataset	Mean Equivalence Test				Slope Equivalence Test			
	1	2	3	4	1	2	3	4
ME.L	0.09%	0.07%	0.11%	0.07%	0.41%	0.44%	0.70%	0.42%
ME.L.MV1	1.25%	0.26%	1.02%	0.21%	34.32%	6.10%	32.11%	3.25%
ME.L.MV2	1.91%	0.39%	1.83%	0.43%	71.52%	29.60%	75.81%	22.63%
Relationship 4: $E(X_2\|X_1)\|x_1$ vs. $E(Y_2\|Y_1)\|x_{1^*}$								
Dataset	Mean Equivalence Test				Slope Equivalence Test			
	1	2	3	4	1	2	3	4
ME.L	0.04%	0.03%	0.05%	0.03%	1.17%	0.35%	1.20%	0.38%
ME.L.MV1	0.37%	0.16%	0.15%	0.11%	7.23%	10.27%	1.32%	3.86%
ME.L.MV2	0.61%	0.24%	0.24%	0.23%	55.22%	47.66%	27.61%	32.98%
Relationship 5: $E(X_1\|SX_2)\|sx_2$ vs. $E(Y_1\|SY_2)\|sx_{2^*}$								
Dataset	Mean Equivalence Test				Slope Equivalence Test			
	1	2	3	4	1	2	3	4
ME.L	0.24%	0.19%	0.23%	0.20%	1.72%	1.40%	1.81%	1.27%
ME.L.MV1	1.33%	1.27%	1.07%	1.23%	38.44%	28.63%	32.34%	34.02%
ME.L.MV2	1.57%	1.75%	2.77%	3.57%	77.22%	71.92%	75.32%	73.91%
Relationship 6: $E(X_2\|SX_1)\|sx_1$ vs. $E(Y_2\|SY_1)\|sx_{1^*}$								
Dataset	Mean Equivalence Test				Slope Equivalence Test			
	1	2	3	4	1	2	3	4
ME.L	0.10%	0.14%	0.10%	0.13%	2.13%	3.10%	2.00%	2.80%
ME.L.MV1	0.76%	0.50%	0.43%	0.44%	24.90%	23.46%	28.12%	25.79%
ME.L.MV2	0.68%	0.66%	1.11%	0.95%	55.98%	55.47%	54.26%	52.87%

Table 14. Parameter-Based data utility measures for $E(X_1|X_2)$ vs. $E(Y_1|Y_2)$

Dataset	$Y_1 = b_0 + b_1 Y_2$			$X_1 = b_0 + b_1 X_2$	
	Masked Data			Original Datasets	
	Masking Method	b_0	b_1	b_0	b_1
ME.L	1	770.109	-0.938	769.554	-0.937
	2	768.572	-0.935		
	3	770.884	-0.939		
	4	768.554	-0.935		
ME.L.MV1	1	641.233	-0.683	641.342	-0.683
	2	635.700	-0.672		
	3	641.830	-0.684		
	4	635.194	-0.671		
ME.L.MV2	1	461.521	-0.328	463.084	-0.331
	2	479.005	-0.362		
	3	460.344	-0.325		
	4	456.953	-0.319		

VI.4.3. Data Security

As discussed in Section VI.2, the snooper knows that the best (s)he can do is to learn the conditional expectations **u** (= $E(\mathbf{X}|\mathbf{S})$) especially when (s)he sees non-monotonic relationships. Thus, (s)he will try to improve her/his prediction using other available information such as the masked attributes **Y**. As stated earlier, relationships between **u** and confidential attributes **X** are linear. In addition, relationships between **Y** and confidential attributes **X** are also linear. Knowing that, the snooper may choose to fit a linear regression model using both **u** and **Y** as predictors to reduce the prediction error (s)he obtains from using the best predictors **u** alone.

We simulate the above scenario (using the three ME-related datasets) to enhance the prediction of X_1 and X_2 separately beyond what is known about them from the conditional expectations **u** (or the fitted values $E(\mathbf{X}|\mathbf{S})|\mathbf{s}$). Table 15 and Table 16 present the results of this simulation. Instead of fitting a regression line in the form of $E(X_i|u_i) =$

$a + bu_i$, the snooper will try to fit one in the form of $E(X_i|u_iY_i) = a + bu_i + cY_i$ to enhance his prediction. Clearly, the snooper gains nothing by doing that since all regression coefficients c are negligible and non-significant for both X_1 and X_2 across the three datasets and the four masking methods. The regression R^2 and adjusted R^2 also confirm that there is *no* prediction improvement by using **Y** and **u** to predict the non-confidential attributes **X** over using **u** alone .

One may argue that this is just one linear combination among many other possible ones and another *untested* linear combination may reveal more about **X**. For this reason, we decided to run other (stronger) security tests that consider simultaneously all possible *linear* combinations of predictors (**u** and **Y**) that reveal the most about **X**: the canonical correlation security tests (*CC*). There are two conditions in the canonical correlation security tests (*CC*) that should be satisfied when the fitted values **u** (i.e. the best predictors) are involved. First, **u** should be the best predictor for **X**:

$$CC(\mathbf{X}|E(\mathbf{X}|\mathbf{S})|_S) \geq CC(\mathbf{X}|\mathbf{Y}). \tag{6.20}$$

In addition, no gain is possible when a snooper tries to combine both **u** and **Y** to predict **X**:

$$CC(\mathbf{X}|E(\mathbf{X}|\mathbf{S})|_S) = CC(\mathbf{X}|E(\mathbf{X}|\mathbf{S})|_S, \mathbf{Y}). \tag{6.21}$$

Table 17 demonstrates the use of the above *CC* measures for all three datasets masked using all four masking methods. Clearly, the measures indicate that the RMB approach did a good job in protecting the original datasets and no extra information can be learned about confidential attributes **X** beyond what the conditional expectations **u** reveal about them.

Although the snooper will not enhance his prediction using the masked values, datasets with low security index (*SI*) can reveal a lot to the snooper through the conditional expectations **u**. For example, in the case of ME.L dataset, $E(\mathbf{X}|\mathbf{S})$ already explains 93.3 percent and 93.7 percent (see Table 15 and Table 16) of the variation of X_1 and X_2, respectively, in terms of R^2. The correlations in Table 18 and Table 19 show similar levels of associations between **X** and **Y**. In terms of *CC* measures (Table 17), the threat seems to be greater. This again signifies *the importance of assessing the characteristics of confidential attributes **X** in original datasets before masking them*. As the portion of the variance of **X** explained by the residuals increases, the possible security level that the RBM approach can provide increases.

Finally, Table 18 and Table 19 demonstrate how closely the equality in the relationship presented in Equation (6.12) (i.e. $Corr(X_i, Y_i) = [Corr(X_i, u_i)]^2$) holds. Although there are *slight* discrepancies in some figures, they are understandable given that the estimation mechanism (the LS-SVM machine) approaches the conditional expectations. The tables also show that the upper bound (i.e. $Corr(X_i, u_i)$) for the correlation between X_i and Y_i presented in Equation (6.14) holds for all cases.

Table 15. Possible snooper scenario to compromise X_1

Dataset	Regression Line & Masking Methods		a	b (p-value)	c (p-value)	R^2	*Adj.* R^2
ME.L	$E(X_1\|u_1) = a + bu_1$		-0.046	1.000 (0.000)	-	0.933	0.933
	$E(X_1\|u_1 Y_1) = a + bu_1 + cY_1$	1	-0.046	1.000 (0.000)	0.000 (0.995)	0.933	0.933
		2	-0.044	1.005 (0.000)	-0.005 (0.868)	0.933	0.933
		3	-0.048	1.004 (0.000)	-0.003 (0.913)	0.933	0.933
		4	-0.043	1.010 (0.000)	-0.010 (0.752)	0.933	0.933
ME.L.MV1	$E(X_1\|u_1) = a + bu_1$		-0.342	1.001 (0.000)	-	0.566	0.565
	$E(X_1\|u_1 Y_1) = a + bu_1 + cY_1$	1	-0.342	1.001 (0.000)	-0.000 (0.997)	0.566	0.565
		2	-0.340	1.003 (0.000)	-0.002 (0.947)	0.566	0.565
		3	-0.334	0.998 (0.000)	0.003 (0.916)	0.566	0.565
		4	-0.342	1.001 (0.000)	0.000 (0.989)	0.566	0.565
ME.L.MV2	$E(X_1\|u_1) = a + bu_1$		-1.301	1.004 (0.000)	-	0.248	0.247
	$E(X_1\|u_1 Y_1) = a + bu_1 + cY_1$	1	-1.302	1.005 (0.000)	-0.001 (0.981)	0.248	0.247
		2	-1.135	0.998 (0.000)	0.005 (0.870)	0.248	0.247
		3	-1.312	0.999 (0.000)	0.006 (0.853)	0.248	0.247
		4	-1.315	0.999 (0.000)	0.005 (0.864)	0.248	0.247

Table 16. Possible snooper scenario to compromise X_2

Dataset	Regression Line & Masking Methods		a	b (p-value)	c (p-value)	R^2	*Adj.* R^2
ME.L	$E(X_2\|u_2) = a + bu_2$		-0.810	1.002 (0.000)	-	0.937	0.937
	$E(X_2\|u_2 Y_2) = a + bu_2 + cY_2$	1	-0.810	1.004 (0.000)	-0.003 (0.928)	0.937	0.937
		2	-0.811	1.004 (0.000)	-0.003 (0.933)	0.937	0.937
		3	-0.810	1.004 (0.000)	-0.003 (0.934)	0.937	0.937
		4	-0.811	1.005 (0.000)	-0.004 (0.911)	0.937	0.937
ME.L.MV1	$E(X_2\|u_2) = a + bu_2$		-1.660	1.003 (0.000)	-	0.678	0.678
	$E(X_2\|u_2 Y_2) = a + bu_2 + cY_2$	1	-1.660	1.004 (0.000)	-0.001 (0.986)	0.678	0.677
		2	-1.650	1.007 (0.000)	-0.003 (0.914)	0.678	0.677
		3	-1.660	1.003 (0.000)	0.000 (1.000)	0.678	0.677
		4	-1.657	1.005 (0.000)	-0.002 (0.949)	0.678	0.677
ME.L.MV2	$E(X_2\|u_2) = a + bu_2$		-3.539	1.007 (0.000)	-	0.337	0.337
	$E(X_2\|u_2 Y_2) = a + bu_2 + cY_2$	1	-3.542	1.009 (0.000)	-0.002 (0.947)	0.337	0.336
		2	-3.539	1.008 (0.000)	-0.001 (0.976)	0.337	0.336
		3	-3.547	1.009 (0.000)	-0.001 (0.963)	0.337	0.336
		4	-3.543	1.009 (0.000)	-0.002 (0.949)	0.337	0.336

Table 17. Canonical correlation security measures (*CC*) using the best predictor *E*(X|S) for the three ME-Related datasets

Dataset	Condition	Masking Methods			
		1	2	3	4
ME.L	*CC* (X|u) >= *CC* (X|Y)	0.983 >= 0.966	0.983 >= 0.966	0.983 >= 0.966	0.983 >= 0.966
	CC (X|u) = *CC* (X|u,Y)	0.983 = 0.983	0.983 = 0.983	0.983 = 0.983	0.983 = 0.983
ME.L.MV1	*CC* (X|u) >= *CC* (X|Y)	0.880 >= 0.774	0.880 >= 0.771	0.880 >= 0.776	0.880 >= 0.772
	CC (X|u) = *CC* (X|u,Y)	0.880 = 0.880	0.880 = 0.880	0.880 = 0.880	0.880 = 0.880
ME.L.MV2	*CC* (X|u) >= *CC* (X|Y)	0.673 >= 0.453	0.673 >= 0.469	0.673 >= 0.453	0.673 >= 0.451
	CC (X|u) = *CC* (X|u,Y)	0.673 = 0.673	0.673 = 0.673	0.673 = 0.673	0.673 = 0.673

Table 18. $Corr(X_1,Y_1)$: its upper bound and its relationship to $Corr(X_1, E(X_1|S))$ and the security index (*SI*)

Dataset	*SI*	$Var(r_1)$ / $Var(X_1)$	$Var(u_1)$ / $Var(X_1)$	$Corr(X_1,u_1)$	R^2=$Corr(X_1,u_1)^2$	$Corr(X_1,Y_1)$ for Masking Method			
						1	2	3	4
ME.L	6.67%	0.067	0.933	0.966	0.933	0.933	0.932	0.933	0.932
ME.L.MV1	43.41%	0.434	0.566	0.752	0.566	0.566	0.562	0.569	0.565
ME.L.MV2	75.19%	0.752	0.248	0.498	0.248	0.247	0.277	0.251	0.249

Table 19. $Corr(X_2,Y_2)$: its upper bound and its relationship to $Corr(X_2, E(X_2|S))$ and the security index (*SI*)

Dataset	*SI*	$Var(r_2)$ / $Var(X_2)$	$Var(u_2)$ / $Var(X_2)$	$Corr(X_2,u_2)$	R^2=$Corr(X_2,u_2)^2$	$Corr(X_2,Y_2)$ for Masking Method			
						1	2	3	4
ME.L	6.31%	0.063	0.937	0.968	0.937	0.937	0.936	0.937	0.936
ME.L.MV1	32.19%	0.322	0.678	0.823	0.677	0.677	0.671	0.678	0.673
ME.L.MV2	66.28%	0.663	0.337	0.581	0.338	0.335	0.334	0.337	0.335

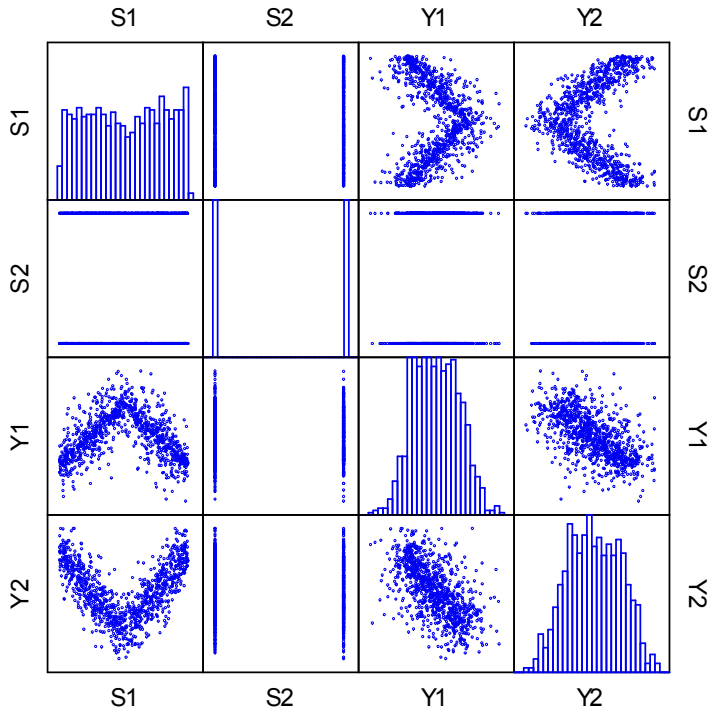


Figure 23. ME.L.MV1 – Masked using masking method 1

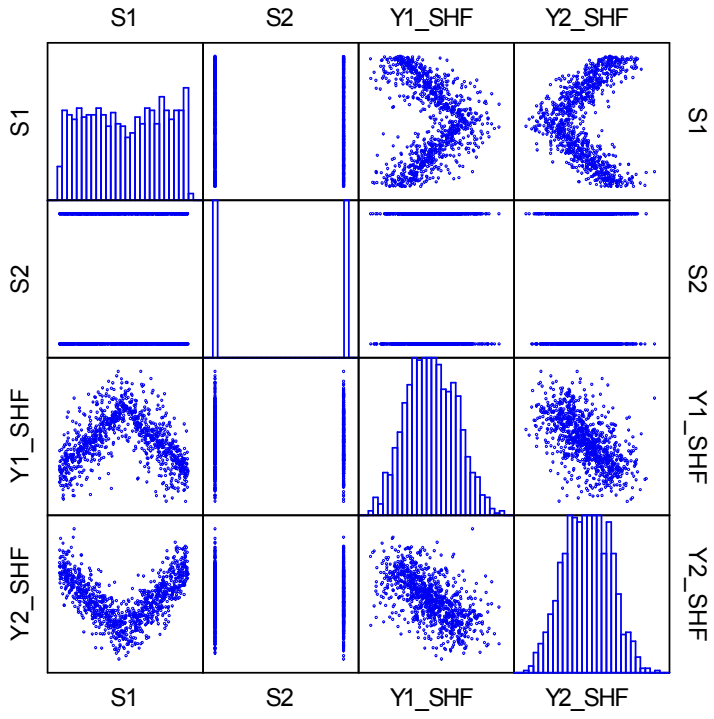


Figure 24. ME.L.MV1 – Masked using masking method 2

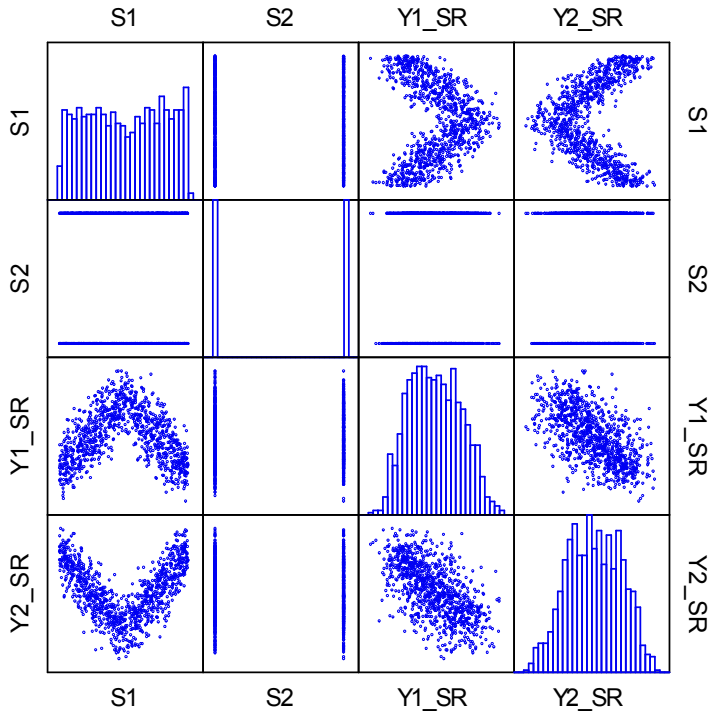


Figure 25. ME.L.MV1 – Masked using masking method 3

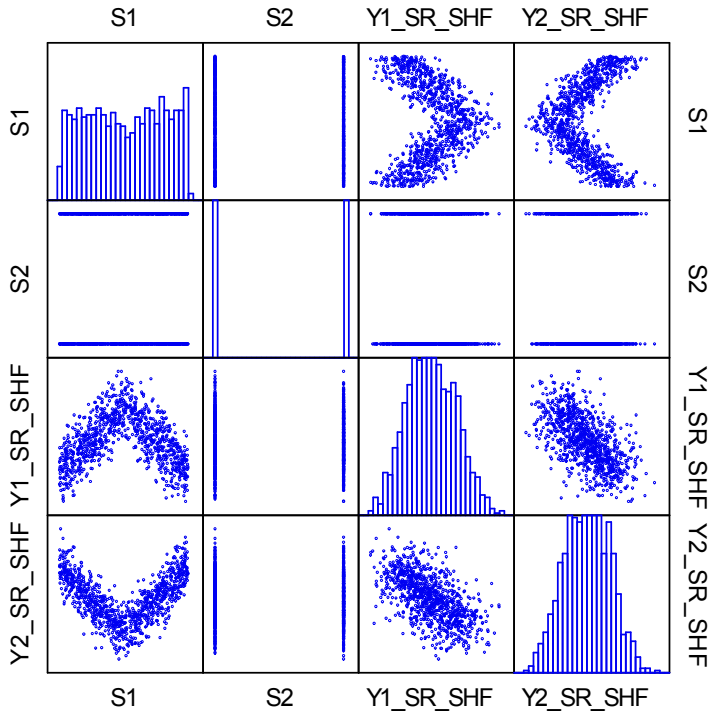


Figure 26. ME.L.MV1 – Masked using masking method 4

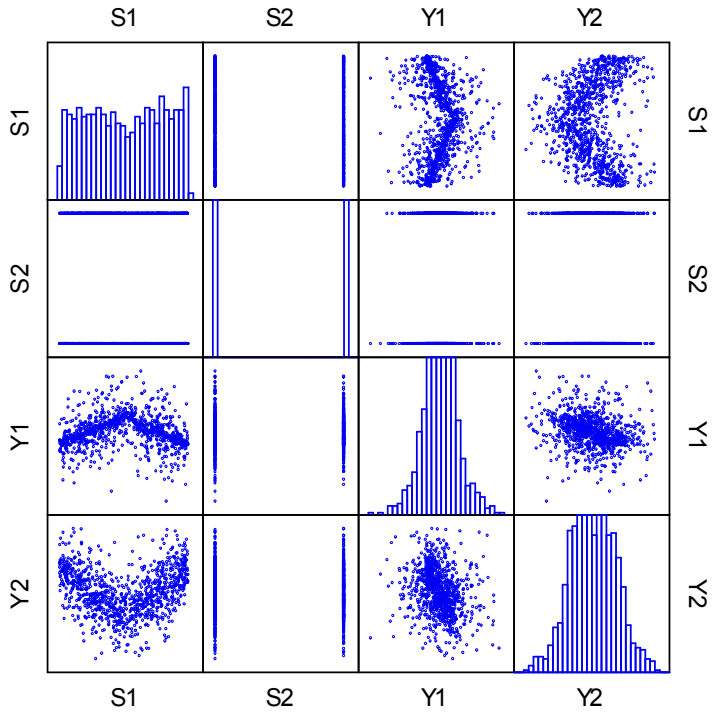


Figure 27. ME.L.MV2 – Masked using masking method 1

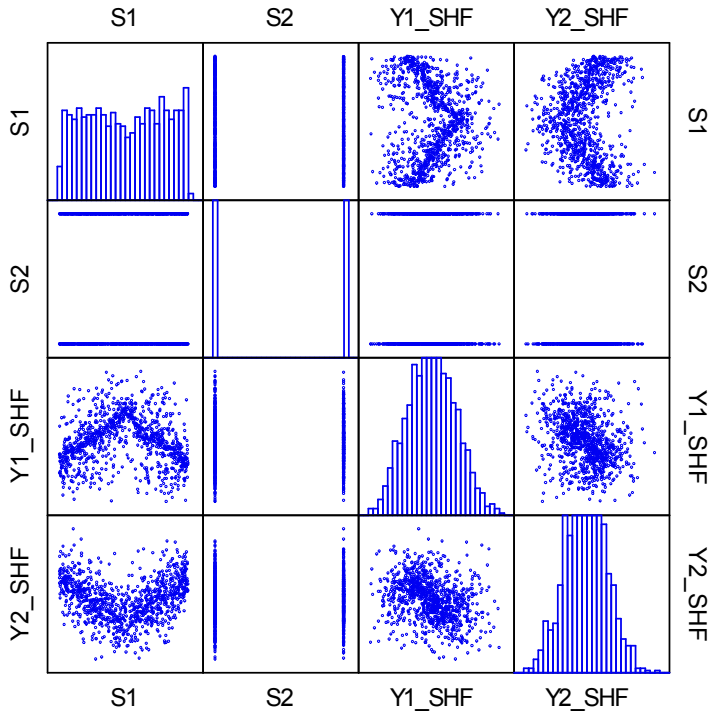


Figure 28. ME.L.MV2 – Masked using masking method 2

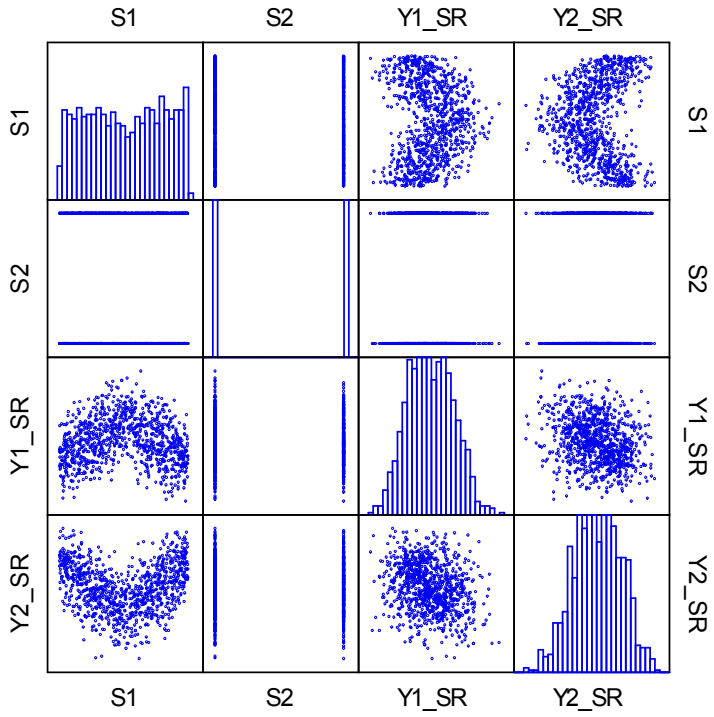


Figure 29. ME.L.MV2 – Masked using masking method 3

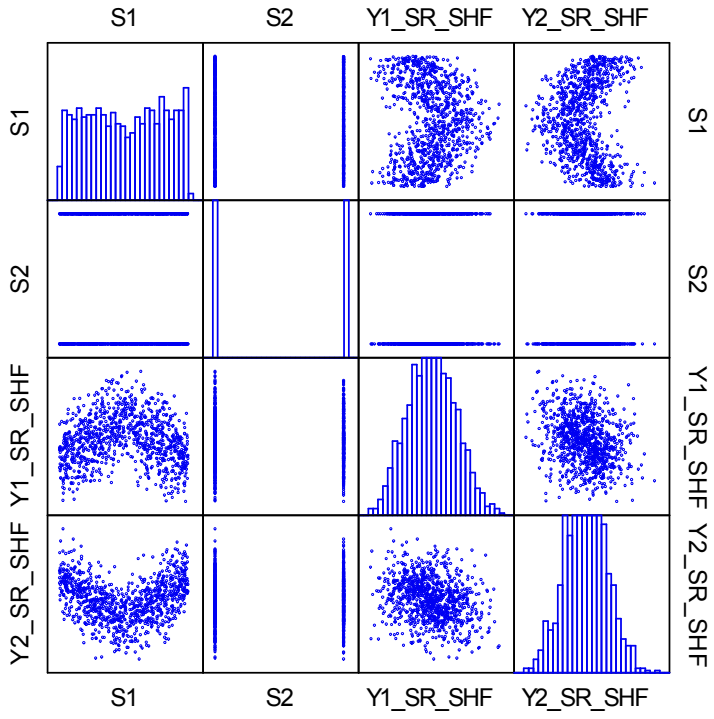


Figure 30. ME.L.MV2 – Masked using masking method 4

CHAPTER

VII. ASSESSING THE RBM DATA UTILITY WHEN THE RELATIONSHIPS AMONG CONFIDENTIAL ATTRIBUTES ARE NONLINEAR

The theoretical basis, as represented by Equation (3.9) in Subsection III.3.1 and its proofs in Appendix B, of the Relationship-Based masking (RBM) approach show that the RBM masking methods work when the relationships among **X** are linear regardless of the relationships between **X** and **S**: monotonic (linear or nonlinear) or (single-valued mapping) non-monotonic. However, we cannot theoretically establish how the RBM approach will work in the general case, or even given a specific dataset when the relationships among confidential attributes are not linear. Therefore, we conduct an experiment to assess its effectiveness for specific datasets with the aim of detecting which type of dataset/relationship class RMB is suitable for and for which it is not.

This chapter *empirically* investigates the effectiveness of the *four* proposed masking methods in Chapter IV and Appendix C in terms of data utility. We want to achieve three goals during our discussion. The first goal is to assess the effectiveness of the proposed methods in terms of data utility. The second goal is to discuss what happens when a violation in the assumptions of the masking methods occurs. Some main assumptions for the masking methods are listed in Section IV.2 such as linearity (or simple pattern) and constant variance of the residuals, and well-estimated conditional expectations. The third goal is to select the best possible masking method among the four proposed methods based on its general performance when the relationships among **X** are nonlinear.

In our experiment, we use ten simulated and derived datasets in addition to the motivation example. The first section (Section VII.1) in this chapter briefly introduces these datasets and their characteristics. The second section (Section VII.2) examines the effectiveness of the masking methods in terms of data utility using some of the measures developed in Section V.2 and, more important, in Section V.3 titled "Equivalence Tests for Validating Models as Data Utility Measures". The last section (Section VII.3) concludes briefly by trying to find one masking method among the four that does a decent job in maintaining different types and classes of relationships. It also summarizes the effects of violation of assumptions on the performance of the masking methods.

Note that we are not planning to discuss the security of the simulated datasets. The datasets with non-monotonic relationships among **X** are simulated with low security index (*SI*). We use low *SI* to enable us to generate non-monotonic relationships in two classes of relationships simultaneously. The two relationship classes are the class of relationships between **S** and **X** and the class of relationships among **X**.

RBM uses the theory of the conditional independence (Muralidhar and Sarathy, 2003c) by conditioning on the best predictor of **X** (i.e. $E(\mathbf{X}|\mathbf{S})$). When we evaluated the two security conditions using Equations (6.20) and (6.21) (security canonical-correlation measures involving the best predictor **u**) presented in Subsection VI.4.3, both security conditions held in all cases although their magnitude was high. Thus, confidentiality is satisfied to the extent the characteristics of original datasets allow. For privacy, the worst re-identification case we encountered across all datasets and masking methods is 4.80 percent re-identification rate, which is acceptable.

VII.1. Datasets

Nine of the datasets (NM.01 to NM.05, ME.L, and MNL.01 to MNL.03), including the motivation example, consist of four variables: two non-confidential attributes **S** (S_1 and S_2) and two confidential attributes **X** (X_1 and X_2). The other two datasets (NM.01.S1 and ME.L.S1) consist of three variables: one non-confidential attribute **S** (S_1) and two confidential attributes **X** (X_1 and X_2). These two datasets are derived, as their names suggest, from two of the datasets with four variables by dropping one non-confidential attribute (S_2). The goal is to study the impact of doing so on the pattern of residuals, which may affect the effectiveness of the masking methods accordingly.

All the names of datasets have an indicator as to the type of the relationship existing among confidential attributes **X** (class). We use "L" in the names of the datasets when relationships among confidential attributes **X** are *linear*. For example, the name of the motivation example dataset is ME.L and the name of its derived dataset (with one non-confidential attribute S_1) is ME.L.S1. The prefix "MNL" is used when the relationships among **X** are *monotonic nonlinear*. Examples include datasets MNL.01 to MNL.03. When the relationships among confidential attributes are non-monotonic, the names of datasets start with the prefix "NM," as in datasets NM.01 to NM.05 and NM.01.S1. This class of relationships is more difficult to reproduce in masked datasets because masking methods try to reproduce them indirectly (vs. directly in the case of relationships between non-confidential attributes **S** and confidential attributes **X**) to satisfy security requirements.

The three MNL datasets are created by randomly sampling 1,000 data points from the simulated monotonic nonlinear dataset used in Sarathy et al.(2002). For each dataset, we also pick a different pair of variables (as X_1 and X_2) from the three original confidential attributes.

We refer to a group of four datasets (the *three* MNL datasets and NM.01.S1 dataset) as *ill-behaved datasets* in this chapter. The three MNL datasets violate the assumption of constant variance of the noise (residuals) terms. This is clear from Figure 90, Figure 97 and Figure 104 in Appendix K, Appendix L and Appendix M, respectively. NM.01.S1 dataset violates the assumption that the patterns left among residuals **r** after removing the conditional expectations $E(\mathbf{X}|\mathbf{S})$ are linear or simple patterns. Clearly, Figure 111 in Appendix N shows non-monotonic pattern among **r** in the case of the NM.01.S1 dataset.

Table 20 lists all 11 datasets with their related groups and highlights the main characteristics of the datasets. In this chapter, when we present the results related to all datasets at once in the form of tables, we divide the tables into *four* horizontal sections. The first section represents the datasets with non-monotonic relationships among confidential attributes **X**. This group, which can be called the non-monotonic group, consists of *five* datasets (the NM datasets). The motivation example (ME.L) occupies the second section. The third section consists of the *three* datasets with monotonic nonlinear relationships among confidential attributes (the MNL datasets). This group is called the "monotonic nonlinear" group. The last group is called the "One S" group, which consists of only one non-confidential attribute S_1 (NM.01.S1 and ME.L.S1 datasets). Both

datasets are derived from their corrsponding two non-confidential attributes datasets as discussed earlier.

In our discussion, many of the materials related to the 11 datasets in the form of tables and figures are provided in appendices. Appendix D provides the material related to the motivation example dataset (ME.L). Appendix F to Appendix J, Appendix K to Appendix M, and Appendix N and Appendix O present the related material to the Non-Monotonic group (NM.01-NM.05 datasets), the Monotonic Nonlinear group (MNL.01-MNL.03 datasets), and the One S group (NM.01.S1 and ME.L.S1 datasets), respectively.

Table 20. Datasets characteristics

Group	Dataset	Records No	S No	X No	X vs. S	X_1 vs. X_2	$E(X_1\|S)\|s$ vs. $E(X_2\|S)\|s$	r_1 vs. r_2
Non-Monotonic	NM.01	1000	2N	2N	NM	NM	NM	NR
	NM.02	1000	2N	2N	NM	NM	NM	L
	NM.03	1000	2N	2N	NM	NM	NM	NR
	NM.04	1000	2N	2N	NM	NM	NM	NR
	NM.05	1000	2N	2N	NM	NM	NM	NR
Motivation Example	ME.L	1000	2: 1N & 1C	2N	NM	L	L	NR
Monotonic Nonlinear	MNL.01	1000	2N	2N	MNL	MNL	MNL	MNL
	MNL.02	1000	2N	2N	MNL	MNL	MNL	MNL
	MNL.03	1000	2N	2N	MNL	MNL	MNL	MNL
One S	NM.01.S1	1000	1N	2N	NM	NM	NM	NM
	ME.L.S1	1000	1N	2N	NM	L	L	NR

NR: No Relationship
L: Linear Relationships
MNL: Monotonic Nonlinear Relationships
NM: Non-Monotonic Relationships
N: Numeric
C: Categorical

VII.2. Data Utility

In any dataset, there are many possible combinations of relationships among variables. The number of relationships increases rapidly as the number of variables increases. In our simulated datasets with four variables, the number of possible relationship combinations is 28 (refer to Table 2 and the related discussion in Subsection V.2). In this section, using the same approach as in Section VI.1 and Subsection V.2, we select six relationships as a plausible representation of the most important relationships that may exist in these datasets.

In the following six subsections, we will investigate the data utility measures for the six relationships, divided equally into three classes:

1. Relationships between confidential attributes **X** and non-confidential attributes **S**:
    a. $E(X_1|\mathbf{S})|\mathbf{s}$ vs. $E(Y_1|\mathbf{S})|\mathbf{s}$, and
    b. $E(X_2|\mathbf{S})|\mathbf{s}$ vs. $E(Y_2|\mathbf{S})|\mathbf{s}$
2. Relationships among confidential attributes **X**:
    a. $E(X_1|X_2)|x_2$ vs. $E(Y_1|Y_2)|x_2$, and
    b. $E(X_2|X_1)|x_1$ vs. $E(Y_2|Y_1)|x_1$
3. Relationships between confidential attributes **X**, and a mixture of confidential attributes **X** and non-confidential attributes **S**:
    a. $E(X_1|\mathbf{S}X_2)|\mathbf{s}x_2$ vs. $E(Y_1|\mathbf{S}Y_2)|\mathbf{s}x_2$, and
    b. $E(X_2|\mathbf{S}X_1)|\mathbf{s}x_1$ vs. $E(Y_2|\mathbf{S}Y_1)|\mathbf{s}x_1$

The rest of this section is divided into six subsections, one for each relationship. In each subsection, we will discuss the data utility measures for the corresponding relationship. We begin our discussion of the data utility by presenting the measures of

equivalence tests (refer to Section V.3). First, we discuss the equivalence of the *magnitudes* of original and corresponding masked relationships (fitted values) using the equivalence tests for the mean of the fitted values obtained from the original unmasked datasets and comparing them with the mean of the fitted values obtained from the corresponding four masked datasets. When they are the same or very similar, it indicates RBM maintains the means of masked attributes similar to the corresponding means of original attributes.

Second, when the fitted values (obtained from original and their corresponding masked datasets) at the individual values, not at their means, are the same or similar, the *direction* of the relationships among them becomes *linear* and the slope of regression of one set of fitted values on the other set should be *one* or close to *one*.

Unlike ordinary regression of the sets of original and masked fitted values on each other, the test for the slope using the equivalence tests is independent of testing the mean of the fitted values (Subsection V.3). Nevertheless, ordinary regression can be also used for comparison purposes. Thus, for the similarity of the *direction* of the relationship, we also report the results of direct regression of one set of original fitted values on their corresponding masked sets of fitted values. We show the significance and the strength of the regression using R^2. In addition, we report the correlation between sets of corresponding fitted values. In our discussion, we use numbers 1 to 4 to refer to masking methods 1 to 4, respectively (see Appendix C).

VII.2.1. Relationship 1: $E(X_1|S)|s$ vs. $E(Y_1|S)|s$

Table 21 presents the results of the equivalence tests (Section V.3) of the mean of fitted values. All the equivalence tests are significant and the null hypotheses of dissimilar means are rejected (unmarked numbers indicate significant equivalence tests). Although we test the equivalence at 25 percent equivalence regions, the numbers listed represent the required minimum equivalence regions (as a percentage of the original mean) to pass the equivalence tests. All numbers of the required mean equivalence tests are very small, especially when compared to the test value (25 percent). The largest (or worst) required minimum equivalence region across all datasets and masking methods to pass the equivalence tests of mean is 0.32 percent (of the original mean) in the case of dataset NM.01.S1 and masking method 4. *This means that all masking methods perform well in maintaining the means of fitted values of masked datasets as the means of fitted values of original datasets*. Notice also that the means of fitted values are also the means of the dependent variables (either confidential or masked variables) since the means for residuals equal zero.

Table 21. Percentage of minimum equivalence regions of significant equivalence tests for mean and slope of 1 for relationships (fitted values) $E(X_1|S)|s$ vs. $E(Y_1|S)|s$

Dataset	Mean Equivalence Test				Slope Equivalence Test			
	1	2	3	4	1	2	3	4
NM.01	0.04%	0.06%	0.05%	0.06%	0.14%	0.16%	0.16%	0.23%
NM.02	0.22%	0.26%	0.18%	0.19%	0.87%	0.97%	1.01%	0.42%
NM.03	0.28%	0.29%	0.24%	0.23%	1.39%	1.47%	1.51%	0.83%
NM.04	0.19%	0.21%	0.15%	0.16%	0.69%	0.52%	0.54%	0.52%
NM.05	0.07%	0.09%	0.08%	0.09%	0.64%	1.17%	0.89%	0.99%
ME.L	0.01%	0.01%	0.01%	0.01%	0.04%	0.15%	0.23%	0.15%
MNL.01	0.07%	0.11%	0.11%	0.13%	0.76%	7.70%	1.88%	3.29%
MNL.02	0.09%	0.11%	0.12%	0.13%	1.24%	8.38%	2.82%	3.76%
MNL.03	0.10%	0.23%	0.16%	0.27%	2.93%	13.84%	3.04%	4.01%
NM.01.S1	0.06%	0.31%	0.14%	0.32%	0.33%	16.92%	2.11%	7.28%
ME.L.S1	0.00%	0.01%	0.01%	0.01%	0.03%	0.15%	0.18%	0.16%

All numbers are significant

Table 21 shows also the results of the equivalence tests of slope for the regression of fitted values obtained from original datasets on the fitted values obtained from their corresponding four masking datasets for the relationships $E(X_1|S)|s$ vs. $E(Y_1|S)|s$. All equivalence tests are significant and the null hypotheses of dissimilar slope to *one* are rejected at the range of 1±0.25 (i.e. at 25 percent of slope of *one*, or [.75± 1.25]).

However, the minimum required equivalence region to pass the equivalence test of slope of *one* is always less than 17 percent. When we exclude the group of monotonic nonlinear datasets (NML group with non-constant variance among their residuals) and NM.01.S1 datasets (the one with non-monotonic nonlinear relationships among its residuals as Figure 111 shows), the minimum required equivalence regions for significant slope equivalence tests across all datasets and masking methods are about 1 percent or less. The largest figure in this group is 1.51 percent in the case of NM.03 dataset and masking method 3. Note that the datasets we excluded are the ones we referred to earlier as the *ill-behaved datasets*.

When ill-behaved datasets are included, although all the masking methods pass the slope equivalence tests, the larger required minimum equivalence regions for masking methods 2 and 4 show that they did not perform as well as masking methods 1 and 3.

From Table 21, based on tests of both mean and slope equivalence, we can conclude that masking method 1 outperforms other masking methods in many cases. In this method, we add a normal independent noise to $E(\mathbf{X}|\mathbf{S})$ to generate $\mathbf{Y}$. This makes relearning the conditional expectations $E(\mathbf{Y}|\mathbf{S})$, which are basically $E(\mathbf{X}|\mathbf{S})$, from masked datasets easy and direct for most learning algorithms.

Another way to check the similarity of the fitted values obtained from original and their corresponding four masking methods is to draw the scatter plots of these fitted values. When the scatter plots show sharp linear relationships, with slope of one or close to one, the relationships learned from original and masked datasets are similar. Figure 31 shows the four scatter plots for the motivation example (ME.L dataset). They show strong linear relationships with slope of one indicating the similarity of relationships learned from original and masked datasets.

These figures can be quantified using different measures such as slope (Table 22), and R^2 and correlations along their significance (Table 23). These measures show similar patterns to the ones found in the slope equivalence tests. When we exclude the ill-behaved datasets, all the slopes are almost one. In addition, both R^2 and correlation measures are also significant and compatible with one another, and compatible with the slope equivalence tests. R^2 measures almost 100 percent and the correlations are almost one.

Nevertheless, masking method 1, which adds independent normal noise, outperforms other masking methods in the case of datasets with some violations in assumptions (e.g. ill-behaved datasets). In addition, it generally outperforms other masking methods across all datasets. This is similar to the conclusion we draw from the slope equivalence tests earlier.

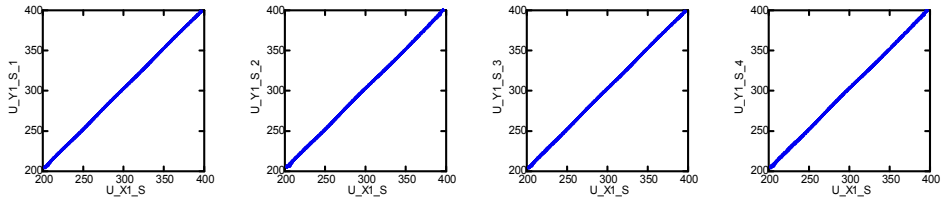


Figure 31. Motivation Example (ME.L) Dataset – $E(X_1|S)|s$ vs. $E(Y_1|S)|s$

Table 22. Slope of linear regression of relationships (fitted values) E(X1|S)|s vs. E(Y1|S)|s

Dataset	Slope			
	1	2	3	4
NM.01	0.9994	0.9998	0.9992	1.0004
NM.02	1.0003	0.9983	1.0038	0.9962
NM.03	0.9994	0.9993	1.0033	0.9971
NM.04	1.0002	0.9974	1.0005	0.9955
NM.05	0.9982	1.0011	0.9999	0.9990
ME.L	0.9999	0.9990	1.0018	0.9990
MNL.01	0.9890	1.0502	0.9895	0.9917
MNL.02	0.9919	1.0572	0.9928	0.9944
MNL.03	0.9931	0.9755	0.9180	0.8185
NM.01.S1	0.9976	1.1622	0.9786	1.0403
ME.L.S1	0.9999	0.9990	1.0014	0.9989

Table 23. R2 of linear regression and correlation of relationships (fitted values) E(X1|S)|s vs. E(Y1|S)|s

Dataset	R^2				Correlation			
	1	2	3	4	1	2	3	4
NM.01	99.98%	99.95%	99.97%	99.95%	0.9999	0.9998	0.9998	0.9998
NM.02	99.57%	99.36%	99.70%	99.68%	0.9978	0.9968	0.9985	0.9984
NM.03	99.13%	99.06%	99.32%	99.38%	0.9956	0.9953	0.9966	0.9969
NM.04	99.68%	99.61%	99.79%	99.76%	0.9984	0.9980	0.9990	0.9988
NM.05	99.58%	99.41%	99.53%	99.39%	0.9979	0.9970	0.9976	0.9970
ME.L	100.00%	99.99%	100.00%	99.99%	1.0000	0.9999	1.0000	1.0000
MNL.01	99.05%	97.84%	97.97%	96.97%	0.9952	0.9891	0.9898	0.9847
MNL.02	98.67%	97.78%	97.46%	96.79%	0.9933	0.9888	0.9872	0.9838
MNL.03	97.39%	86.19%	93.00%	81.00%	0.9869	0.9284	0.9644	0.9000
NM.01.S1	99.91%	97.52%	99.48%	97.44%	0.9996	0.9875	0.9974	0.9871
ME.L.S1	100.00%	99.99%	100.00%	99.99%	1.0000	1.0000	1.0000	1.0000

VII.2.2. Relationship 2: $E(X_2|S)|s$ vs. $E(Y_2|S)|s$

Table 24 shows the equivalence tests for the mean of the sets of fitted values obtained from original datasets versus their corresponding masked datasets, $E(X_2|S)|s$ vs. $E(Y_2|S)|s$, and for their regression slopes of one. The mean equivalence tests are examined at 25 percent of the mean of original fitted values. The slope equivalence tests are examined at slope range of 1 ± 0.25. For the mean equivalence tests, all tests are significant as in the previously discussed relationship. Moreover, all minimum required equivalence regions to pass the tests are very small. The largest figure is 0.42 percent (of the mean of original fitted values) in the case of NM.02 dataset and masking method 1. *These small values indicate that all masking methods succeeded in maintaining the magnitude of the relationships obtained from masked data to be the same as the magnitude of the relationships obtained from original data.* Consequently, *they maintain the mean of the masked attributes to be the same as the mean of the confidential attributes.* In addition, the performance of all four masking methods in maintaining the mean is comparable and no masking method is superior across all datasets. Nevertheless, in the case of the datasets with some violations in assumptions, masking method 1 performs slightly better.

For the slope equivalence tests, when we exclude the ill-behaved datasets (the three MNL datasets and NM.01.S1), the largest minimum equivalence region to pass the slope equivalence test to one is 2.11 percent (i.e. the slope in the range of 1 ± 0.0211). Other minimum equivalence regions are even less than 1.75 percent and many of them are less than 1 percent.

For the ill-behaved datasets, the required minimum equivalence regions are generally and slightly larger than other datasets. Some of them are above 11 percent (i.e. the minimum equivalence region for the slope in the range of 1±0.11) as in the case of MNL.02 and NM.01.S1 datasets when masked using masking method 2. On these datasets, masking method 1 performs better than other masking methods in most cases. On all other datasets, the performance of all four masking methods is comparable.

Figure 32 shows the scatter plots for the fitted values obtained from the original motivation example dataset against the fitted values obtained from its four masked datasets for the relationship $E(X_2|\mathbf{S})|\mathbf{s}$ vs. $E(Y_2|\mathbf{S})|\mathbf{s}$. All four scatter plots indicate very strong linear relationships with a slope equal to one. We report the slopes of all relationships in these scatter plots in Table 25. When we exclude the ill-behaved datasets, all slopes are very close to one. The minimum two slopes are 0.9885 and 0.9877 in the case of the NM.01 dataset when masked using masking methods 1 and 3, respectively.

For the ill-behaved datasets, the masking methods perform differently on these datasets based on the assumption they violate. The ill-behaved datasets can be divided into two groups based on their assumption violation: the assumption of constant variance and the assumption of simple (linear) patterns exist among residuals after learning the conditional expectations $E(\mathbf{X}|\mathbf{S})$. While the three MNL datasets mainly violate the constant variance assumption, the NM.01.S1 violates the assumption of simple linear patterns among residuals. Most masking methods perform very well on the NM.01.S1 dataset in terms of data utility. The slopes are very close to one in the case of masking methods 1 and 4 with slopes 0.9819 and 0.9979, respectively. The other two masking methods (2 and 3) are off by about ±0.08.

On the other hand, when the constant variance assumption in residuals is violated (the MNL datasets), the slopes are far from one. The closest figure to one is 0.8281in the case of the MNL.01 and masking method 1. Masking method 1 outperforms other masking methods on this group of datasets. When there is no assumption violation, the performance of all four methods is comparable. Further, all R^2 and correlations measures in Table 26 are significant. They are compatible and confirm the conclusions we draw from the slopes table (Table 25) for most cases.

Table 24. Percentage of minimum equivalence regions of significant equivalence tests for mean and slope of 1 for relationships (fitted values) E(X2|S)|s vs. E(Y2|S)|s

Dataset	Mean Equivalence Test				Slope Equivalence Test			
	1	2	3	4	1	2	3	4
NM.01	0.02%	0.02%	0.02%	0.02%	1.26%	0.15%	1.34%	0.15%
NM.02	0.22%	0.20%	0.20%	0.18%	0.47%	0.49%	0.61%	0.40%
NM.03	0.42%	0.41%	0.32%	0.32%	1.45%	1.44%	0.60%	0.95%
NM.04	0.35%	0.32%	0.28%	0.27%	1.73%	1.57%	0.86%	1.02%
NM.05	0.17%	0.18%	0.16%	0.16%	1.50%	2.11%	1.74%	1.61%
ME.L	0.02%	0.02%	0.02%	0.02%	0.28%	0.22%	0.26%	0.22%
MNL.01	0.25%	0.31%	0.29%	0.34%	3.21%	5.77%	5.56%	6.70%
MNL.02	0.34%	0.38%	0.36%	0.40%	7.63%	11.25%	7.75%	8.93%
MNL.03	0.33%	0.38%	0.36%	0.41%	6.18%	8.68%	5.48%	7.04%
NM.01.S1	0.05%	0.11%	0.06%	0.12%	1.65%	11.06%	8.30%	4.45%
ME.L.S1	0.01%	0.01%	0.01%	0.01%	0.17%	0.16%	0.20%	0.16%

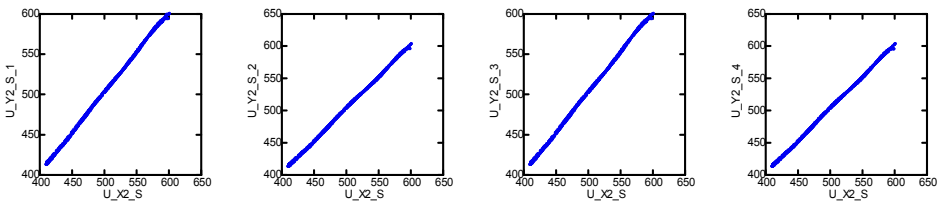


Figure 32. Motivation Example (ME.L) Dataset – $E(X_2|S)|s$ vs. $E(Y_2|S)|s$

Table 25. Slope of linear regression of relationships (fitted values) E(X2|S)|s vs. E(Y2|S)|s

Dataset	Slope			
	1	2	3	4
NM.01	0.9885	0.9994	0.9877	0.9994
NM.02	0.9969	0.9980	0.9942	0.9979
NM.03	0.9985	0.9992	0.9953	0.9989
NM.04	0.9970	0.9976	0.9945	0.9969
NM.05	0.9922	0.9971	0.9976	0.9963
ME.L	0.9980	0.9987	0.9983	0.9987
MNL.01	0.8281	0.7571	0.7672	0.6760
MNL.02	0.7727	0.7311	0.7356	0.6643
MNL.03	0.7734	0.6983	0.7087	0.6262
NM.01.S1	0.9819	1.0799	0.9198	0.9979
ME.L.S1	0.9990	0.9991	0.9987	0.9991

Table 26. R2 of linear regression and correlation of relationships (fitted values) E(X2|S)|s vs. E(Y2|S)|s

Dataset	R2				Correlation			
	1	2	3	4	1	2	3	4
NM.01	99.92%	99.94%	99.92%	99.94%	0.9996	0.9997	0.9996	0.9997
NM.02	99.61%	99.67%	99.68%	99.71%	0.9980	0.9984	0.9984	0.9986
NM.03	99.01%	99.08%	99.41%	99.41%	0.9951	0.9954	0.9970	0.9971
NM.04	98.68%	98.86%	99.16%	99.22%	0.9934	0.9943	0.9958	0.9961
NM.05	98.49%	98.39%	98.72%	98.72%	0.9924	0.9919	0.9936	0.9936
ME.L	99.95%	99.95%	99.95%	99.95%	0.9998	0.9998	0.9998	0.9997
MNL.01	83.14%	74.06%	78.43%	69.26%	0.9118	0.8606	0.8856	0.8322
MNL.02	74.10%	67.79%	70.69%	63.49%	0.8608	0.8233	0.8408	0.7968
MNL.03	75.24%	66.70%	69.84%	61.24%	0.8674	0.8167	0.8357	0.7825
NM.01.S1	99.30%	97.09%	99.00%	96.49%	0.9965	0.9853	0.9950	0.9823
ME.L.S1	99.98%	99.98%	99.98%	99.98%	0.9999	0.9999	0.9999	0.9999

VII.2.3. Relationship 3: $E(X_1|X_2)|x_2$ vs. $E(Y_1|Y_2)|x_2$

As discussed in Section III.1 and Subsection II.2.1.2, the class of this relationship (i.e. the class of relationships among confidential attributes $E(X_i|X_j)$) is more difficult to maintain than the class of the earlier two relationships (i.e. the class of relationships among non-confidential attributes and confidential attributes $E(\mathbf{X}|\mathbf{S})$) for security reasons.

Table 27. Percentage of minimum equivalence regions of significant equivalence tests for mean and slope of 1 for relationships (fitted values) E(X1|X2)|x2 vs. E(Y1|Y2)|x2

Dataset	Mean Equivalence Test				Slope Equivalence Test			
	1	2	3	4	1	2	3	4
NM.01	0.76%	0.11%	0.75%	0.11%	3.37%	0.68%	3.36%	0.65%
NM.02	2.05%	0.32%	1.22%	0.27%	4.45%	1.78%	3.13%	2.30%
NM.03	1.10%	0.36%	0.38%	0.17%	4.86%	3.83%	1.84%	0.52%
NM.04	1.46%	0.23%	1.27%	0.25%	3.27%	1.98%	3.22%	2.63%
NM.05	0.84%	0.25%	0.40%	0.16%	24.83%*	13.19%	10.17%	3.54%
ME.L	0.09%	0.07%	0.11%	0.07%	0.41%	0.44%	0.70%	0.42%
MNL.01	3.61%	0.17%	0.80%	0.17%	79.06%*	7.86%	8.45%	5.05%
MNL.02	1.53%	0.16%	0.49%	0.17%	63.20%*	6.28%	14.13%	4.10%
MNL.03	2.12%	0.28%	1.38%	0.29%	52.01%*	2.88%	60.43%*	5.39%
NM.01.S1	3.48%	1.71%	3.11%	1.64%	11.98%	11.26%	12.35%	8.15%
ME.L.S1	0.12%	0.08%	0.11%	0.08%	0.74%	0.43%	0.54%	0.44%

* indicates non-significant slope equivalence tests. Null hypotheses of dissimilar slope of one cannot be rejected.

The independent mean equivalence tests and the slope equivalence tests for the fitted values obtained from original datasets and their corresponding four masked datasets are reported in Table 27. All mean equivalence tests are significant at 25 percent equivalence regions. All masking methods maintain the mean of the fitted values of masked values close to the mean of the original fitted values. When there is no violation in assumptions, all of the minimum required equivalence regions are less than 1.50 percent except for the NM.02 dataset when masked using method 1 (it is 2.05 percent). Actually, many of them are less than 1 percent. In the case of ill-behaved datasets, masking method 1 did not perform as well as the other three masking methods, especially methods 2 and 4. Moreover, method 4 seemed to perform slightly better than method 2 in many cases.

The slope equivalence tests of slope equal to one indicate similar results: masking methods 2 and 4 outperform masking methods 1 and 3. Although not all of the slope equivalence tests are significant, they are always significant in the case of these two masking methods (i.e. 2 and 4). In addition, and similar to what we obtained in the mean equivalence tests, masking method 4 performs better than masking method 2 in most

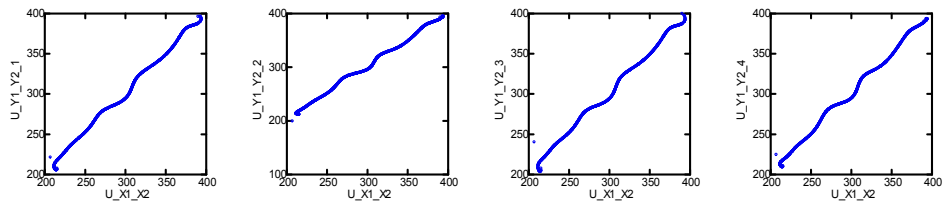


Figure 33. Motivation Example (ME.L) Dataset – $E(X_1|X_2)|x_2$ vs. $E(Y_1|Y_2)|x_2$

cases. Further, Table 27 also provides some evidence that ill-behaved datasets usually require larger minimum equivalence regions than the required regions by other datasets to pass the slope equivalence tests.

Figure 33 shows the scatter plots of the original fitted values (i.e. $E(X_1|\ X_2)|x_2$) against each of the four masked fitted values (i.e. $E(Y_1|\ Y_2)|x_2$) for the motivation example. Although they are not exactly straight lines, the existence of sharp linear pattern with slope close to one is clear in all four scatter plots. Table 28 shows the slopes for all datasets' scatter plots. For easy comparison with the slope equivalence tests, the cells that correspond to the slope equivalence table's cells with non-significant figures are also marked in the slope table. Similarly, the table of R^2 and correlation measures (Table 29), which will be discussed shortly, is also marked. We use this procedure wherever it is applicable in the rest of this chapter.

When we exclude the ill-behaved datasets, the performance of all masking methods in terms of regression slope is comparable. However, in some cases, as in the case of masking method 1 applied on NM.01 and NM.02 datasets and masking method 3 applied on NM.01 dataset, masking methods 2 and 4 perform better than masking methods 1 and 3. In addition, when the relationships among confidential attributes are

Table 28. Slope of linear regression of relationships (fitted values) E(X1|X2)|x2 vs. E(Y1|Y2)|x2

Dataset	Slope			
	1	2	3	4
NM.01	0.9447	0.9941	0.9432	0.9945
NM.02	0.9438	0.9793	0.9647	0.9760
NM.03	1.0216	1.0130	1.0070	0.9947
NM.04	0.9637	0.9799	0.9639	0.9731
NM.05	1.0666*	1.0543	0.9695	0.9980
ME.L	0.9959	0.9958	0.9981	0.9958
MNL.01	0.4462*	0.9013	0.8041	0.9263
MNL.02	1.0676*	0.9243	0.8333	0.9408
MNL.03	0.6298*	0.9088	0.7346*	0.9246
NM.01.S1	0.7963	0.6861	0.7875	0.6454
ME.L.S1	0.9926	0.9956	0.9946	0.9954

* indicates non-significant slope equivalence tests. Null hypotheses of dissimilar slope of one cannot be rejected. Provided here for easy comparison between regular slope and slope equivalence tests.

non-monotonic, the slopes in many cases are not as close to one as when the relationships are monotonic. For examples, compare NM datasets with ME.L and ME.L.S1 datasets.

In the case of the ill-behaved datasets, all slopes (corresponding to significant or non-significant slope equivalence tests) are smaller compared to the slopes of other datasets. When the assumption violated is the constant variance, masking methods 2 and (especially) 4 perform better than masking methods 1 and 3. When the assumption violated is simple (monotonic) patterns among residuals, the reverse happens: masking methods 1 and 3 perform better than masking methods 2 and 4.

Table 29 presents R^2 and correlation measures for all datasets. They show the same level of masking methods performance and similar patterns to the performance level and patterns shown by the slopes table (Table 28). Although all listed figures are significant, they are low in the cases when we encounter non-significant slope equivalence tests figures.

Table 29. R^2 of linear regression and correlation of relationships (fitted values) $E(X_1|X_2)|x_2$ vs. $E(Y_1|Y_2)|x_2$

Dataset	R^2				Correlation			
	1	2	3	4	1	2	3	4
NM.01	96.51%	99.86%	96.32%	99.85%	0.9824	0.9993	0.9814	0.9993
NM.02	97.64%	99.08%	98.83%	99.35%	0.9881	0.9954	0.9941	0.9967
NM.03	98.06%	98.24%	99.35%	99.59%	0.9902	0.9912	0.9968	0.9980
NM.04	98.86%	99.49%	98.82%	99.39%	0.9943	0.9974	0.9941	0.9969
NM.05	82.54%*	93.10%	89.04%	97.28%	0.9085*	0.9649	0.9436	0.9863
ME.L	99.61%	99.66%	99.53%	99.60%	0.9981	0.9983	0.9976	0.9980
MNL.01	11.31%*	95.99%	76.26%	96.10%	0.3363*	0.9798	0.8733	0.9803
MNL.02	42.36%*	97.21%	93.59%	96.85%	0.6508*	0.9859	0.9674	0.9841
MNL.03	33.15%*	90.12%	31.96%*	89.39%	0.5757*	0.9493	0.5654*	0.9454
NM.01.S1	72.85%	63.87%	71.81%	66.88%	0.8535	0.7992	0.8474	0.8178
ME.L.S1	99.60%	99.55%	99.61%	99.57%	0.9980	0.9977	0.9981	0.9979

* indicates non-significant slope equivalence tests (null hypotheses of dissimilar slope of one cannot be rejected). Provided to facilitate comparison with slope measures.

VII.2.4. Relationship 4: $E(X_2|X_1)|x_1$ vs. $E(Y_2|Y1)|x_1$

The class of this relationship is the same as the class of the previous relationship: the class of relationships among confidential attributes **X**. Hence, they are harder to maintain in masked datasets because they are reproduced indirectly to satisfy security requirements. For the simulated datasets we test, the group of NM datasets (NM.01-NM.05) poses another challenge: the mapping from X_1 to X_2 in original datasets and similarly the mapping from Y_1 to Y_2 in masked datasets is a multi-valued mapping. As we discussed in Section III.1, this type of mapping makes it impossible to learn the real conditional expectations using artificial neural networks (ANN) (refer also to Section 6.1.5, pp. 207-208 in Bishop (1995)). Shortly, we will discuss the impact of multi-valued mapping using the slope equivalence tests.

Table 30 presents the results of the (independent) equivalence tests of means of fitted values (original vs. masked) and their slopes (of one). For the mean equivalence tests, all the results are significant at the test value 25 percent and the null hypotheses of dissimilar means of fitted values obtained from original datasets to the means of the

Table 30. Percentage of minimum equivalence regions of significant equivalence tests for mean and slope of 1 for relationships (fitted values) E(X2|X1)|x1 vs. E(Y2|Y1)|x1

Dataset	Mean Equivalence Test				Slope Equivalence Test			
	1	2	3	4	1	2	3	4
NM.01	0.30%	0.09%	0.22%	0.07%	3.97%	2.53%	3.17%	1.86%
NM.02	4.13%	0.36%	2.75%	0.37%	97.15%*	96.77%*	96.74%*	99.39%*
NM.03	5.90%	1.01%	4.43%	1.00%	72.40%*	74.86%*	71.23%*	71.50%*
NM.04	4.19%	0.32%	3.11%	0.31%	101.14%*	103.12%*	102.55%*	106.99%*
NM.05	1.86%	0.39%	1.62%	0.37%	51.97%*	47.67%*	49.96%*	46.66%*
ME.L	0.04%	0.03%	0.05%	0.03%	1.17%	0.35%	1.20%	0.38%
MNL.01	0.32%	0.18%	0.21%	0.18%	5.56%	10.25%	3.58%	8.15%
MNL.02	0.58%	0.16%	0.92%	0.17%	2.33%	8.74%	21.58%*	9.09%
MNL.03	1.56%	0.38%	0.78%	0.35%	9.80%	2.52%	6.96%	2.91%
NM.01.S1	0.48%	0.46%	0.53%	0.46%	5.75%	15.13%	6.47%	11.24%
ME.L.S1	0.07%	0.04%	0.08%	0.03%	0.95%	0.35%	1.09%	0.41%

* indicates non-significant slope equivalence tests. Null hypotheses of dissimilar slope of one cannot be rejected.

corresponding four masked datasets are rejected. The largest minimum required equivalence region to pass the mean equivalence test across all datasets and masking methods is 5.90 percent in the case of the NM.03 dataset and masking method 1. *This means that all masking methods do a very good job in maintaining the mean of the fitted values.* This also indicates they maintain the mean of the masked attributes similar to the mean of original confidential attributes. As a final remark, the highest values we encounter are the ones associated with the multi-valued mapping, especially for NM.02 to NM.04 masked using masking methods 1 and 3. Masking methods 2 and 4 also outperform masking methods 1 and 3 in all cases although the difference is sometimes small.

For the slope equivalence, all the results for the NM.02 to NM.05 datasets are non-significant, and the null hypotheses of mean not equal to one cannot be rejected. These datasets suffer from the multi-valued mapping. It must be noted that each masking method of the four masking methods performs as badly as the other ones, and there is no superior method in this case. Nevertheless, although NM.01 (the check mark) dataset also

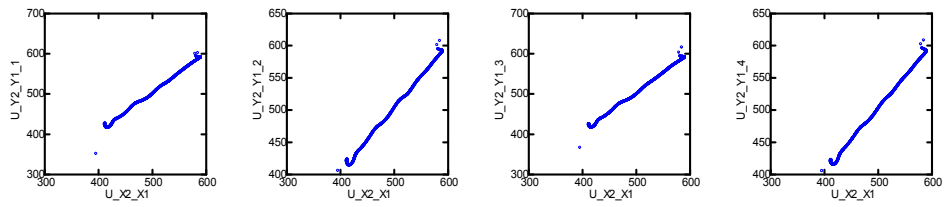


Figure 34. Motivation Example (ME.L) Dataset – $E(X_2|X_1)|\mathbf{x}_1$ vs. $E(Y_2|Y_1)|\mathbf{x}_1$

suffers from the multi-valued mapping problem, it does not suffer from it on all its range of X_1 values as the other datasets do (compare the relationships between the two confidential attributes **X** using the figures in the related appendices for datasets NM.01-NM.05: Appendix F-Appendix H). The multi-valued mapping in this dataset happens only near the head of the check mark. Therefore, this dataset passes the slope equivalence tests. Masking method 4 performs better than other masking methods on this dataset.

Both ME.L and ME.L.S1 pass the slope equivalence tests. *Masking methods 2 and 4 perform better than masking methods 1 and 3* as they measured by the minimum required equivalence regions to pass the test. The smaller the required regions are, the more similar the compared measures. All the slopes for the ill-behaved datasets pass the slope equivalence tests except in the case of the MNL.02 dataset masked using masking method 3. The required minimum equivalence regions for the ill-behaved datasets are generally larger than the ones required by other datasets given they do not suffer from the multi-valued mapping problem.

For the motivation example, Figure 34 shows the scatter plots of the fitted values obtained from the original dataset versus the fitted values obtained from the four corresponding masked datasets. Clearly, all four scatter plots demonstrate very strong

Table 31. Slope of linear regression of relationships (fitted values) E(X2|X1)|x1 vs. E(Y2|Y1)|x1

Dataset	Slope			
	1	2	3	4
NM.01	0.9550	0.9693	0.9656	0.9785
NM.02	0.7379*	0.6729*	0.5945*	0.2531*
NM.03	1.1186*	0.6714*	1.0558*	0.6805*
NM.04	0.0605**	-0.1604**	-0.0858**	-0.4524*
NM.05	1.1847*	0.8324*	1.2002*	0.8755*
ME.L	0.9887	0.9972	0.9886	0.9970
MNL.01	0.9645	0.8740	0.9201	0.8871
MNL.02	0.9380	0.9099	1.0886*	0.9054
MNL.03	0.7725	0.8914	0.8721	0.9125
NM.01.S1	0.6872	0.7088	0.6252	0.6703
ME.L.S1	0.9911	0.9970	0.9897	0.9966

* indicates non-significant slope equivalence tests. Null hypotheses of dissimilar slope of one cannot be rejected. Provided here for easy comparison between regular slope and slope equivalence tests.

linear relationships with slopes close to one. Table 31 lists the slopes of all the scatter plots for the relationship $E(X_2|X_1)|x_1$ vs. $E(Y_2|Y_1)|x_1$.

In Table 31, the slopes in the case of the motivation example and its derived dataset (i.e. ME.L and ME.L.S1) are very close to one. For NM.01 (with a little multi-valued mapping problem at the head), the smallest slope is 0.9550. Although it is less than the slopes of the ME.L datasets, it is better than the slopes shown by other multi-valued mapping datasets (NM.02 to NM.05).

In Table 32, the R^2 of the regression of the fitted values and the correlation among them are non-significant in the case of dataset NM.04 masked using methods 1, 2 and 3. When the slope equivalence tests are non-significant (i.e. in the case of the four datasets with the multi-valued mapping problem: NM.02-NM.05), their values are less than other values.

Table 32. R2 of linear regression and correlation of relationships (fitted values) E(X2|X1)|x1 vs. E(Y2|Y1)|x1

Dataset	R^2				Correlation			
	1	2	3	4	1	2	3	4
NM.01	98.55%	98.66%	99.00%	99.08%	0.9927	0.9933	0.9950	0.9954
NM.02	3.20%*	3.28%*	3.00%*	0.65%*	0.1788*	0.1811*	0.1731*	0.0809*
NM.03	33.81%*	19.33%*	33.30%*	21.97%*	0.5815*	0.4397*	0.5771*	0.4687*
NM.04	0.02%**	0.21%**	0.06%**	2.24%*	0.0145**	-0.0463**	-0.0251**	-0.1498*
NM.05	59.95%*	46.66%*	63.06%*	49.81%*	0.7743*	0.6831*	0.7941*	0.7057*
ME.L	99.68%	99.78%	99.72%	99.80%	0.9984	0.9989	0.9986	0.9990
MNL.01	92.71%	95.01%	93.80%	94.52%	0.9629	0.9747	0.9685	0.9722
MNL.02	93.19%	98.09%	87.42%*	97.87%	0.9653	0.9904	0.9350*	0.9893
MNL.03	82.46%	88.84%	91.55%	90.42%	0.9081	0.9426	0.9568	0.9509
NM.01.S1	67.67%	63.15%	61.50%	62.50%	0.8226	0.7947	0.7842	0.7906
ME.L.S1	99.74%	99.74%	99.72%	99.75%	0.9987	0.9987	0.9986	0.9988

* indicates non-significant slope equivalence tests (null hypotheses of dissimilar slope of one cannot be rejected). Provided to facilitate comparison with slope measures.
** indicates, in addition to non-significant slope equivalence tests, non-significant R^2 and correlation measures.

Because of that, all the slope equivalence tests for NM.02 to NM.05 datasets are non-significant; notice how the corresponding R^2 and correlations measures are low in these cases.

VII.2.5. Relationship 5: $E(X_1|SX_2)|sx_2$ vs. $E(Y_1|SY_2)|sx_2$

The data utility measures in this subsection try to verify whether the relationships among the confidential attribute X_1 and a mixture of non-confidential attributes **S** and other confidential attributes X_2 in original datasets are correctly reproduced during masking among the corresponding masked attribute Y_1 and a set of non-confidential attributes **S** and other masked attributes Y_2 in masked datasets. Table 33 shows the equivalence tests for both the mean and the slope. All mean equivalence tests are significant, and the null hypotheses of dissimilar means of fitted values obtained from original datasets and corresponding means of fitted values obtained from their four masked datasets are rejected at test value 25 percent.

When the ill-behaved datasets are excluded, the largest minimum required equivalence region to pass the mean equivalence tests is 1.91 percent in the case of the NM.05 dataset masked using method 2. Many other figures are even less than 1 percent. For the ill-behaved datasets, the largest minimum required equivalence region is 6.31 percent in the case of masking method 1 applied on the NM.01.S1 dataset. Nevertheless, many figures are still less than 2 percent or even 1 percent. These results provide evidence that all masking methods are able to maintain the means of masked attributes as the means of confidential attributes, as they are here measured by fitted values. The performance of the four masking methods is comparable across all datasets. Nevertheless, masking method 4 performed slightly better than other masking methods on NM.01-NM.05 datasets.

All slope equivalence tests for the ill-behaved datasets are non-significant, and we cannot reject the null hypotheses of dissimilar means to one. Although masking method 1 applied on NM.05 dataset did not pass the test, its required minimum equivalence region to pass the slope equivalence test (26.11 percent) is less than the ones required by the ill-

Table 33. Percentage of minimum equivalence regions of significant equivalence tests for mean and slope of 1 for relationships (fitted values) E(X1|SX2)|sx2 vs. E(Y1|SY2)|sx2

Dataset	Mean Equivalence Test				Slope Equivalence Test			
	1	2	3	4	1	2	3	4
NM.01	0.79%	0.64%	0.76%	0.62%	3.89%	1.44%	3.74%	1.29%
NM.02	1.36%	0.90%	1.13%	0.50%	3.22%	1.74%	2.27%	1.33%
NM.03	1.35%	1.47%	1.02%	0.94%	14.15%	12.21%	8.63%	7.41%
NM.04	1.10%	1.18%	1.26%	0.94%	7.88%	5.93%	4.00%	3.34%
NM.05	1.58%	1.91%	0.60%	0.39%	26.11%*	24.44%	15.85%	12.21%
ME.L	0.24%	0.19%	0.23%	0.20%	1.72%	1.40%	1.81%	1.27%
MNL.01	1.62%	0.90%	1.12%	0.96%	60.16%*	54.95%*	44.27%*	52.25%*
MNL.02	1.28%	1.66%	1.81%	1.93%	34.63%*	32.99%*	29.50%*	33.59%*
MNL.03	3.33%	1.44%	2.01%	1.93%	67.23%*	62.08%*	56.47%*	63.88%*
NM.01.S1	6.31%	4.70%	3.64%	5.43%	48.14%*	30.81%*	39.57%*	46.74%*
ME.L.S1	0.12%	0.13%	0.12%	0.13%	1.02%	0.65%	1.12%	0.60%

* indicates non-significant slope equivalence tests. Null hypotheses of dissimilar slope of one cannot be rejected.

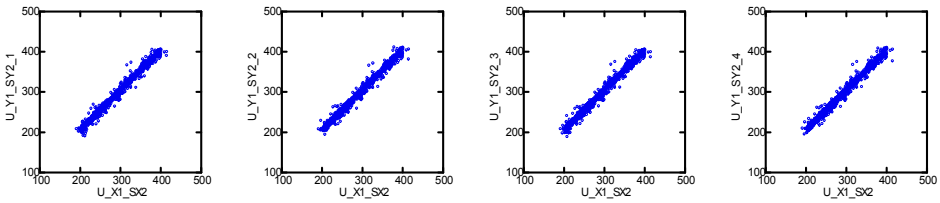


Figure 35. Motivation Example (ME.L) Dataset – $E(X_1|SX_2)|sx_2$ vs. $E(Y_1|SY_2)|sx_2$

behaved datasets and it is close to the test value (25 percent). All other masking methods applied on other datasets pass the slope equivalence test of slope equal to one. Their required minimum equivalence regions to pass the slope equivalence tests tend to be larger than those required by early relationships. Masking method 4 shows superior performance when compared to other masking methods, as demonstrated by its smaller required minimum equivalence regions when we exclude the ill-behaved datasets.

Figure 35 shows the scatter plots for the set of fitted values obtained from the original motivation example dataset (ME.L) versus each set of fitted values obtained from each of its four masked copies. Clearly, they show strong linear patterns although they may not form very sharp lines. This indicates that all four masking methods succeeded in reproducing the relationship $E(X_1|\mathbf{S}X_2)|x_2$ in masked datasets represented by the relationship $E(Y_1|\mathbf{S}Y_2)|x_2$.

Table 34 quantifies these scatter plots in terms of regression slopes for all datasets. For the ill-behaved datasets, the slopes are far from one. This confirms the results we got from the slope equivalence tests. The slope of the fitted values of the NM.05 dataset masked using method 1 is the worst slope among the slopes of the four masking methods for this dataset. Notice this corresponds to the case that did not pass the slope equivalence tests besides the ill-behaved datasets. Nevertheless, this slope (0.9949)

Table 34. Slope of linear regression of relationships (fitted values) E(X1|SX2)|sx2 vs. E(Y1|SY2)|sx2

Dataset	Slope			
	1	2	3	4
NM.01	0.9487	0.9763	0.9486	0.9769
NM.02	0.9781	0.9756	0.9783	0.9747
NM.03	0.9996	0.9883	0.9755	0.9782
NM.04	0.9940	0.9898	0.9843	0.9869
NM.05	0.9949*	1.0047	1.0035	0.9981
ME.L	0.9926	0.9900	0.9934	0.9884
MNL.01	0.5446*	0.7703*	0.7195*	0.8270*
MNL.02	0.6480*	0.7953*	0.7644*	0.8487*
MNL.03	0.4543*	0.6771*	0.6032*	0.6472*
NM.01.S1	0.4637*	0.5497*	0.4876*	0.4164*
ME.L.S1	0.9963	0.9935	0.9969	0.9929

is close to one. By checking the R^2 and correlation measures, which we will discuss shortly, the difference in masking methods' performance becomes clearer. For other datasets, the slopes are always about or more than 0.95. In some cases, masking methods 2 and 4 outperform masking methods 1 and 3. In others, this distinction dissolves.

Table 35 presents R^2 and correlation measures for the linear patterns in the scatter plots. All these measures are significant. Nonetheless, notice how these measures are low

Table 35. R2 of linear regression and correlation of relationships (fitted values) E(X1|SX2)|sx2 vs. E(Y1|SY2)|sx2

Dataset	R^2				Correlation			
	1	2	3	4	1	2	3	4
NM.01	97.61%	98.22%	97.43%	98.10%	0.9880	0.9911	0.9871	0.9905
NM.02	95.89%	96.93%	96.71%	97.87%	0.9792	0.9845	0.9834	0.9893
NM.03	87.85%	88.72%	90.92%	92.23%	0.9373	0.9419	0.9535	0.9604
NM.04	93.13%	94.52%	95.75%	96.52%	0.9651	0.9722	0.9785	0.9825
NM.05	76.16%*	78.47%	86.56%	89.53%	0.8727*	0.8858	0.9304	0.9462
ME.L	98.35%	98.39%	98.34%	98.37%	0.9917	0.9919	0.9917	0.9918
MNL.01	24.36%*	37.71%*	43.17%*	42.57%*	0.4936*	0.6141*	0.6571*	0.6524*
MNL.02	45.45%*	56.37%*	56.97%*	59.41%*	0.6742*	0.7508*	0.7548*	0.7708*
MNL.03	17.23%*	28.48%*	29.08%*	26.10%*	0.4151*	0.5337*	0.5392*	0.5109*
NM.01.S1	26.80%*	41.09%*	32.37%*	24.86%*	0.5177*	0.6410*	0.5690*	0.4986*
ME.L.S1	99.17%	99.24%	99.15%	99.23%	0.9959	0.9962	0.9957	0.9962

compared to others in the table cells that correspond to the non-significant slope equivalence tests in Table 33 (i.e. all the ill-behaved datasets and NM.05 masked using method 1).

VII.2.6. Relationship 6: $E(X_2|\mathbf{S}X_1)|\mathbf{s}x_1$ vs. $E(Y_2|\mathbf{S}Y_1)|\mathbf{s}x_1$

Table 36 shows the results of the equivalence tests for the relationship $E(X_2|\mathbf{S}X_1)|x_1$ obtained from original datasets to its corresponding relationships $E(Y_2|\mathbf{S}Y_1)|x_1$ obtained from their four masked copies in terms of mean and slope. For the mean, all masking methods pass the equivalence tests indicating the ability of these methods to preserve the mean of confidential attributes in the masked attributes. The largest required minimum equivalence region to pass the equivalence test is 3.06 percent in the case of the ill-behaved dataset (MNL.03) when masked using method 1. This number drops to 1.21 percent in the case of masking method 3 applied to the NM.03 dataset when we excluded the ill-behaved datasets. Most of the other numbers are less

Table 36. Percentage of minimum equivalence regions of significant equivalence tests for mean and slope of 1 for relationships (fitted values) E(X2|SX1)|sx1 vs. E(Y2|SY1)|sx1

Dataset	Mean Equivalence Test				Slope Equivalence Test			
	1	2	3	4	1	2	3	4
NM.01	0.09%	0.08%	0.09%	0.08%	1.08%	0.78%	1.19%	0.61%
NM.02	0.31%	0.36%	0.28%	0.37%	0.66%	0.56%	0.75%	0.87%
NM.03	0.87%	0.69%	1.21%	0.65%	1.14%	1.02%	1.30%	1.08%
NM.04	0.60%	0.67%	0.70%	0.79%	2.06%	2.17%	2.42%	2.40%
NM.05	0.64%	0.50%	0.44%	0.40%	2.56%	3.88%	3.38%	4.33%
ME.L	0.10%	0.14%	0.10%	0.13%	2.13%	3.10%	2.00%	2.80%
MNL.01	1.19%	2.27%	1.11%	1.25%	72.27%*	72.90%*	73.39%*	72.00%*
MNL.02	2.19%	3.01%	1.49%	1.66%	65.84%*	61.47%*	56.43%*	54.48%*
MNL.03	3.06%	1.25%	1.28%	1.22%	55.75%*	62.61%*	52.26%*	60.17%*
NM.01.S1	1.29%	0.81%	1.02%	0.85%	30.81%*	42.93%*	52.19%*	41.41%*
ME.L.S1	0.11%	0.07%	0.10%	0.07%	1.28%	1.11%	1.15%	1.07%

* indicates non-significant slope equivalence tests. Null hypotheses of dissimilar slope of one cannot be rejected.

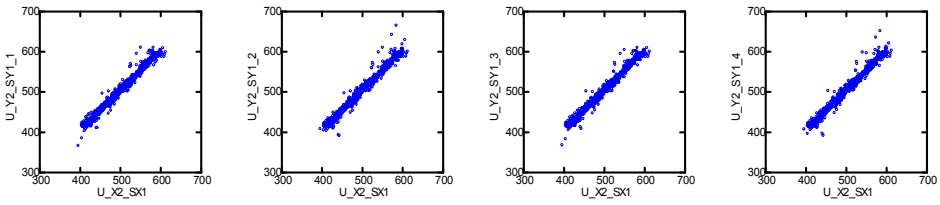


Figure 36. Motivation Example (ME_L) Dataset – $E(X_2|SX_1)|sx_1$ vs. $E(Y_2|SY_1)|sx_1$

than 1 percent.

For the slope equivalence tests, masking methods applied on the ill-behaved datasets did not pass the tests and the null hypotheses of dissimilar mean to one cannot be rejected. The slope equivalence tests for masking methods applied on all other datasets were significant. The null hypotheses of dissimilar slope to one are rejected. The largest required minimum equivalence region to pass the slope equivalence tests is 4.33 percent in the case of masking method 4 applied on the NM.05 dataset. In general, the performance of the four masking methods on each dataset is comparable regardless of the category of the dataset (i.e. whether it is an ill-behaved dataset) and there is no superior masking method in all cases.

Figure 36 shows the scatter plots of the fitted values obtained from the original motivation example dataset (ME.L) versus each of the fitted values obtained from each one of its four masked versions. Strong or sharp lines with slope of one indicate identical fitted values. Linear pattern with slope close to one point to similar fitted values. The linear pattern in Figure 36 is the latter one indicating similar fitted values.

Table 37 shows the slopes of the regression lines of the linear patterns for all datasets. The magnitude levels of the reported slopes are compatible with the slope significance tests although all slopes (or linear regression models) for the fitted values are significant (refer to R^2 measures in Table 38). This means the slopes are high when the slope equivalence tests are significant and low when the tests are non-significant. When the slope equivalence tests are non-significant in the case of the ill-behaved datasets, the slopes are far off. The closest slope to one is 0.8458 in the case of the MNL.02 dataset masked using masking method 2. When the ill-behaved datasets are excluded, all slopes are very high and, in many cases, close to one. The minimum slope is 0.9705 in the case of NM.03 masked by method 3.

Table 38 reports the R2 and correlation measures for the fitted values. They are also high when the slope equivalence tests are significant and low when the slope equivalence tests are non-significant. The performance of the four masking methods is comparable to one another.

Table 37. Slope of linear regression of relationships (fitted values) E(X2|SX1)|sx1 vs. E(Y2|SY1)|sx1

Dataset	Slope			
	1	2	3	4
NM.01	0.9887	0.9994	0.9881	0.9985
NM.02	0.9945	0.9961	0.9969	0.9997
NM.03	0.9746	0.9805	0.9705	0.9801
NM.04	0.9867	0.9941	0.9945	0.9954
NM.05	0.9745	0.9847	0.9840	0.9871
ME.L	0.9948	1.0005	0.9945	0.9992
MNL.01	0.7986*	0.8341*	0.6354*	0.7072*
MNL.02	0.7702*	0.8458*	0.7644*	0.7342*
MNL.03	0.5625*	0.5950*	0.6046*	0.6460*
NM.01.S1	0.6176*	0.8191*	0.6192*	0.6785*
ME.L.S1	0.9944	0.9959	0.9944	0.9957

Table 38. R^2 of linear regression and correlation of relationships (fitted values) E(X2|SX1)|sx1 vs. E(Y2|SY1)|sx1

Dataset	R2				Correlation			
	1	2	3	4	1	2	3	4
NM.01	99.50%	99.57%	99.59%	99.62%	0.9975	0.9978	0.9980	0.9981
NM.02	99.31%	99.50%	99.42%	99.53%	0.9966	0.9975	0.9971	0.9976
NM.03	97.63%	97.93%	97.31%	97.86%	0.9881	0.9896	0.9864	0.9892
NM.04	97.59%	98.10%	97.93%	98.01%	0.9879	0.9905	0.9896	0.9900
NM.05	96.16%	95.89%	96.25%	95.70%	0.9806	0.9792	0.9811	0.9783
ME.L	98.19%	97.85%	98.27%	97.99%	0.9909	0.9892	0.9913	0.9899
MNL.01	24.83%*	25.31%*	19.36%*	22.39%*	0.4983*	0.5031*	0.4400*	0.4732*
MNL.02	29.13%*	35.56%*	36.29%*	36.41%*	0.5397*	0.5963*	0.6024*	0.6034*
MNL.03	27.67%*	24.93%*	31.75%*	28.54%*	0.5260*	0.4993*	0.5635*	0.5342*
NM.01.S1	45.83%*	49.85%*	32.51%*	42.82%*	0.6770*	0.7060*	0.5702*	0.6544*
ME.L.S1	98.83%	99.08%	98.94%	99.10%	0.9941	0.9954	0.9947	0.9955

VII.3. Summary

From the results in the previous section, it is clear that no one masking method is superior to all other methods in every class of relationships. Nevertheless, masking method 4 seems to do a reasonable job across all the different relationships and datasets. This is because masking method 4 employs the shuffling refinement, which corrects the marginal distributions at two levels: the residuals level and the variables level. Because of this refinement, it has also the advantage of maintaining the marginal distributions of masked variables as the marginal distributions of confidential attributes.

It is also interesting to compare the difference in performance between the motivation example dataset (ME.L) and its derived dataset (ML.L.S1) with the check-mark dataset (NM.01) and its derived dataset (NM.01.S1). Note that ML.L.S1 does not violate any assumption while NM.01.S1 violates the

assumption of simple linear patterns among residuals. There is no difference in performance between motivation example dataset (ME.L) and its derived dataset (ML.L.S1). On the other hand, there is a difference in performance between check-mark dataset (NM.01) and its derived dataset (NM.01.S1). This difference is not just in the magnitude of data utility measures (especially, slope-related ones) but even in the significance of the measures. In the next chapter, we will summarize the general conclusions and findings of this chapter and the whole study.

CHAPTER

VIII. CONCLUSIONS

VIII.1. Main Findings and Conclusions

In this section, we summarize the main finding and conclusions of this study in terms of *data utility* and *data security*.

VIII.1.1. Data Utility

We earlier picked three important classes of relationships to preserve in masked datasets: relationships between **S** and **X**, relationships among **X**, and relationships between **X** and a mixture of **S** and **X**. The types of these relationships can be linear, monotonic nonlinear, or non-monotonic. The main results on data utility are:

1. The RBM approach allows the relationships between non-confidential attributes **S** and confidential attributes **X** (i.e. $E(\mathbf{X}|\mathbf{S})$ to be estimated and reproduced in masked datasets between non-confidential attributes **S** and masked attributes **Y**. There are only two conditions. First, the estimation mechanism used should be able to learn and estimate conditional expectations well. The ANN estimation approach employed by RBM learns conditional expectations well when the minimum of the mean squared error function is reached regardless of the type of relationship. Second, the added noise **e** should be independent of (or orthogonal to) **S** or any function of **S** similar to the residuals **r**.

2. When the purpose of the masking is only to preserve relationships between non-confidential attributes **S** and confidential attributes **X**, or at least they represent the most important class of relationships to preserve first, simple addition of a normal noise to the estimated conditional expectations $E(\mathbf{X}|\mathbf{S})$ works well for RBM. This is because most estimation mechanisms find it easier to learn conditional expectations when the residuals (noise terms) are normally distributed (with constant variance). Masking method 1 achieves exactly that. Refer to the discussion of relationships $E(X_1|\mathbf{S})$ and $E(X_2|\mathbf{S})$ in the previous chapter. However, masking method 1 does not seem to be the best among other masking methods in preserving other classes of relationships.
3. When relationships among confidential attributes **X** are linear, all proposed masking methods preserve data utlity regardless of the relationships between non-confidential attributes **S** and confidential attributes **X** (i.e. monotonic linear or nonlinear, or (single-valued mapping) non-monotonic relationships). This is clear from Equation (3.9) in Subsection III.3.1 and their proofs in Appendix B. We need only to add independent noise **e** with the same covariance of **r** to $E(\mathbf{X}|\mathbf{S})$ to produce masked data **Y** with the same linear relationships among **Y**, as the linear relationships among **X**.
4. When the relationships among confidential attributes **X** are nonlinear (monotonic nonlinear or non-monotonic), the RBM approach preserves data utility in our experiment as long as the patterns and relationships

among the fitted values (i.e. the evaluation of the conditional expectations $E(X_i|\mathbf{S})|\mathbf{s}$ and $E(X_j|\mathbf{S})|\mathbf{s}$) corresponds and accounts for (most of) the nonlinear patterns among **X**, as checked by scatter plots.

5. The initial results suggest that the RBM approach works for the class of relationships between **X** and a mixture of **S** and **X** (i.e. $E(X_i|\mathbf{SX}_i)$) as long as relationships among **X** are linear. This happens because the two subcomponents or (sub-)classes of relationships contributing to this class of relationships (i.e. the class of relationships among **S** and **X**, and the class of relationships among **X**) are well maintained and preserved in the masked data.
6. The RBM approach may work for the class of relationships between **X** and a mixture of **S** and **X** (i.e. $E(X_i|\mathbf{SX}_j)$) when relationships among **X** are nonlinear as long as the patterns and relationships among the conditional expectations $E(X_i|\mathbf{SX}_j)$ correspond to the patterns and relationships among **X**. This correspondences leaves simple patterns (linear or no relationships) in residuals **r** that can be reproduced easily in the independent noise **e** by specifying only $\text{Cov}(e_i,e_j) = \text{Cov}(r_i,r_j)$.
7. Except for the two cases mentioned in the previous point, the RBM approach does not seem to work well in many cases for the class of relationships between **X** and a mixture of **S** and **X** (i.e. $E(X_i|\mathbf{SX}_j)$) when relationships among **X** are nonlinear and the patterns among **r** are not linear. The non-simplicity of the patterns among residuals **r** indicate that the residuals account for at least some, if not all, of the non-monotonic

patterns in $X_i|\mathbf{S}X_j$ instead of patterns among conditional expectations accounting for them. Two reasons might be behind this behavior. First, the tendency of the ANN approach is to over-fit the data, especially as the number of independent variables increase and when no mechanism to reduce the over-fitting is applied. Second, there may be some interactions between **S** and **X** when they are used together as independent variables. The RBM approach does not explicitly consider such interactions or relationships in generating masked data.

8. Masking methods that employ the shuffling refinement at the level of variables (methods 2 and 4) perform relatively better in the class of relationships that involve non-confidential attributes **X** as independent variables than other masking methods. Hence, the RBM approach tries to reproduce these relationships indirectly for security reasons. Actually, masking method 4 outperforms masking method 2 on many occasions. This is because masking method 4 implements shuffling refinement at two levels: the residuals level and the variables level.
9. The RBM approach utilizes the concepts of *conditional expectations* and the manipulation of *residuals* to achieve *data utility* and *data security*. It has a good theoretical basis and utilizes the concept of conditional independence theory (Muralidhar and Sarathy, 2003c). However, although the RBM approach is promising and covers many classes of relationships well, it is clear from the above discussion that the RBM approach is not generalizable to all possible classes of relationships and all existing

datasets. Nonetheless, it is one step forward in the field (i.e. masking datasets involving nonlinear and non-monotonic relationships for PPE), which is still in its infancy.

VIII.1.2. Data security

The RBM approach is developed based on the concepts of conditional independence theory for masking (perturbation) methods (Muralidhar and Sarathy, 2003c). The RBM approach generates masked variables **Y** in a way that makes the non-confidential attributes **S** (or functions of **S** such as $E(\mathbf{X}|\mathbf{S})$) the best predictors for the confidential attributes **X**. In addition, combining **S** and **Y** will not improve the prediction of **X** beyond what is known about them from **S** (or the best predictor $E(\mathbf{X}|\mathbf{S})$).

Notice that **X** and **Y** are not independent unless we condition on **S** (or functions of **S**). In addition, the relationships between **X** and **Y** will tend to be linear (with positive slope) since we want to maximize data utility and **S** links the two. When the variance of residuals r_i is a small portion of the variance of confidential attributes X_i and most of this variance is explained by the variance of the conditional expectations (i.e. $Var(E(X_i|\mathbf{S}))$), there is not enough security in original datasets with which to begin. In other words, generating masked datasets from original data with *very small variance* of residuals **r** *while maximizing data utility* would result in *insecure* masked datasets, in which a snooper can use **Y** to predict **X** with high confidence without even using **S**. This is not to say that **Y** is a better predictor for **X** than **S** (or $E(\mathbf{X}|\mathbf{S})$). **S** is still the best predictor. However, it is easier for the snooper to fit a simple linear regression model between **Y** and **X** rather than to fit a nonlinear model (or an ANN model), especially when the gain from the latter approach is very small.

The security index measure (*SI*) in Section III.3.2 can be used to assess if the confidential attributes **X** (based on their characteristics in original datasets) have enough variance in the residuals to allow for effective masking using the RBM approach. We also draw attention to the importance of evaluating each confidential attribute based on the sensitivity of the variables to see whether there is enough variance at the attributes level to with which to begin. In addition, we developed the theoretical connection between the security measure (*SI*) for a confidential attribute X_i and the correlation of the confidential attribute with its masked version (i.e. $Corr(X_i, Y_i)$). This helps the owner of data to pre-assess the possible disclosure risk associated with a specific confidential attribute (based on its original characteristics) before releasing the masked data or even before masking takes place.

When a dataset with a low *SI* measure has to be released, some compromises on the side of data utility can be made. For example, we can add a noise term with more variance than we have in the residuals. Although this approach may not affect the class of relationships between **S** and **X**, it will affect the class of relationships among **X**, for example. We discussed some possible compromises in Section III.3.2. Fuller (1993) also discussed some possible compromises in Section 2 of his paper. Some of Fuller's suggestions might be applicable to our context.

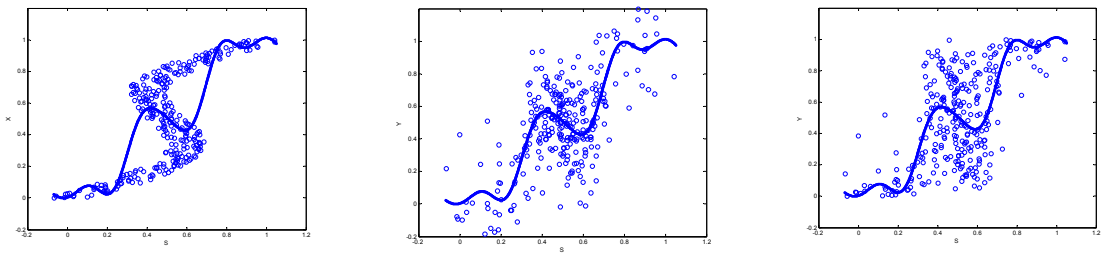
Finally, the empirical work we have done shows that the RBM approach satisfies the requirements of protection against *value disclosure* although the problem of over-fitting affects one of the measures in some cases. However, we showed that the measure holds true using more controlled datasets (i.e. datasets with their non-monotonic relationships consist of piecewise linear relationships; the motivation-example related

datasets). In addition and surprisingly, although **X** and **Y** are not independent unless conditioned on **S**, the worst re-identification rate we encountered was about 5 percent, which is acceptable.

VIII.2. Possible Opportunities and Limitations

The definition of a relationship between variables A and B adopted in our proposed methods (NM-EGADP masking methods) is the mathematical definition (Macnaughton, 2002): $A=g(B)+e$ where $g(B)$ is a strict, mathematical function. A strict mathematical function is a single-valued mapping (one-to-one or many-to-one mapping) but not multi-valued mapping (one-to-many). Usually the mathematical function $g(B)$ is replaced by the conditional expectation $E(A|B)$ since the best estimator of $g(B)$ ends to be $E(A|B)$ (Bickel and Doksum, 2001; Shao, 1999). Again, $E(A|B)$ is a mathematical relationship in the strict sense we mentioned.

When the data contains multi-valued mapping areas as in inverse problems (Bishop, 1995), the definition of mathematical functions is violated. This leads to the following serious consequence: the conditional expectation in this case is wrongly learned because the average of two or more solutions is not necessarily a (valid) solution (Bishop, 1995). This impact of the multi-valued mapping on learning the conditional expectation is automatically inherited by the RBM approach since they adopt the mathematical and conditional expectation definition for a relationship. Figure 37 demonstrates what can be called "by-product" limitation of our proposed methods. For a good treatment of the subject of the function mapping (single-valued vs. multi-valued mapping) on the performance of learning conditional expectation using ANN, refer to (Bishop, 1995, pp. 207-208 specifically and Chapter 6 in general).



(a) Original **(b) NM-EGADP Perturbed** **(c) NM-EGADP Shuffling**

Figure 37. Impact of multi-valued data on learning the conditional expectation (a) wrongly learned conditional expectation from original multi-valued dataset (b) impact of that on producing perturbed values (c) impact of that on producing shuffled values

From another perspective, we think (and there is initial evidence) that our proposed methods can be used for other privacy-preserving data mining techniques such as classification and clustering. To recall, our proposed methods are about maintaining and reproducing relationships and "*relationships between variables then give life to MR*" (Multiple Regression), "*and indeed to all multivariate statistical techniques,*" Rud (2001, pp. 106) said. Although this possibility will not be investigated in this study, we will provide a simple classification example using Multiple Discriminant Analysis (MDA). We run MDA on our original motivation example and its two masked versions (using NM-EGADP perturbation and its shuffling variant). S_2 is considered the target class variable in this analysis. The confusion matrices of the three models are shown in Table 39. The similarity among the three listed confusion matrices in this example suggests that using our masking methods for other PPDM techniques is a promising direction of research that we plan to pursue later. Figure 52 shows an example for possible application for privacy-preserving clustering.

VIII.3. Contributions of this Study

In this section, we summarize the main contributions of this study. *First*, it adapts some masking techniques for privacy-preserving estimation (PPE) for the more difficult types of relationships (non-monotonic). *Second*, an appropriate theoretical definition of relationships (i.e. conditional expectation) in the context of regression applications is specified and utilized as a basis for the RBM approach. *Third* and based on this definition, existing effective theoretically based masking methods for monotonic

(a) Original Dataset

			Predicted S2		
			.00	1.00	Total
S2	.00	Count	282	250	532
		% of Total	28.2%	25.0%	53.2%
	1.00	Count	242	226	468
		% of Total	24.2%	22.6%	46.8%
Total		Count	524	476	1000
		% of Total	52.4%	47.6%	100.0%

(b) NM-EGADP Perturbed Dataset

			Predicted S2		
			.00	1.00	Total
S2	.00	Count	281	251	532
		% of Total	28.1%	25.1%	53.2%
	1.00	Count	238	230	468
		% of Total	23.8%	23.0%	46.8%
Total		Count	519	481	1000
		% of Total	51.9%	48.1%	100.0%

(c) NM-EGADP Shuffled Dataset

			Predicted S2		
			.00	1.00	Total
S2	.00	Count	275	257	532
		% of Total	27.5%	25.7%	53.2%
	1.00	Count	230	238	468
		% of Total	23.0%	23.8%	46.8%
Total		Count	505	495	1000
		% of Total	50.5%	49.5%	100.0%

Table 39. MDA classification example using the motivation example dataset and its masked copies

relationships are adapted and extended for non-monotonic relationships. *Fourth*, this adaptation employs some learning and estimation mechanisms for learning the (monotonic and, more importantly, non-monotonic) conditional expectations and relationships. Although other learning approaches might be also appropriate, we adopt Artificial Neural Networks (ANN) approaches as our main learning mechanism based on the presented ample theoretical evidences of their ability to learn different types of conditional expectations. We also mention the main theoretical criterion of this ability; namely, reaching the global minimum of the mean squared errors function after training. *Fifth*, while we adopt existing security measures for value disclosure (MSE measures) and identity disclosure (percent of re-identified records using record linkage), we propose two new types of data utility measures: parameter-based measures and (MSE) prediction-based measures. In addition, we adapt the use of a model validation test using regression-based equivalence tests as a data utility measure. *Sixth*, we use a form of nonlinear regression coefficient of determination R^2 measures to obtain normalized forms of our proposed MSE data utility and existing MSE data security measures. *Seventh*, we adapt and propose an extension for Canonical Correlation Analysis security measures (piecewise *CC*) for special cases of nonlinear relationships (when they completely consist of piecewise linear relationships). *Eighth*, we propose a simple, yet insightful, framework (the SDL/Relationships Match Framework) in Appendix A (pp. 189) to guide the use of any specific existing or new masking method on a dataset based on the relationships existing in the dataset that the masking method can preserve. *Last but not least*, we express the characteristics of masked attributes in terms of the characteristics of original data. Thus, these characteristics including security-related ones can be calculated before

even the masking takes place. In addition, we derive an *upper bound* for the correlation between confidential attributes and their masked copies. The proofs for many results are also provided.

References

The Centers for Disease Control and Prevention (CDC) "HIPAA privacy rule and public health - guidance from CDC and the U.S. Department of health and human services," (accessed: May, 2004, published April 11, 2003), 2003, http://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm.

Adam, N.R., and Wortmann, J.C. "Security-control methods for statistical databases - a comparative-study," *Computing Surveys* (21:4), 1989, pp. 515-556.

Agrawal, R., and Srikant, R. "Privacy-preserving data mining," *Sigmod Record* (29:2), 2000, pp. 439-450.

Al-Ahmadi, M., Sarathy, R., and Delen, D. "Privacy preserving data mining: Issues and opportunities," *Proceedings of the Workshop on Data Mining Research in Oklahoma (PowerPoint presentation format)*, Tulsa, OK, USA, February 6, 2004, pp. 30 (slides).

Anderson, T.W. *An introduction to multivariate statistical analysis*, Wiley-Interscience, Hoboken, N.J., 2003.

Ashrafi, M.Z., Taniar, D., and Smith, K. "Towards privacy preserving distributed association rule mining," *Lecture notes in computer science*:2918), 2003, pp. 279-289.

Ashrafi, M.Z., Taniar, D., and Smith, K. "Reducing communication cost in a privacy preserving distributed association rule mining," *Lecture Notes In Computer Science 0302-9743 2004; [No] 2973*, 2004, pp. Pages 381-392.

Bartko, J.J. "Proving the null hypothesis," *American Psychologist* (46:10), 1991, pp. 1089-1089.

Benaloh, J.C. "Secret sharing homomorphisms: Keeping shares of a secret secret (extended abstract)," (accessed 2005: Jan 7), 1987, pp. 251-260, http://research.microsoft.com/copyright/accept.asp?path=/crypto/papers/ssh.ps&pub=15.

Berger, R.L., and Hsu, J.C. "Bioequivalence trials, intersection-union tests and equivalence confidence sets," *Statistical Science* (11:4), 1996, pp. 283-302.

Berry, M.J.A., and Linoff, G. *Mastering data mining : The art and science of customer relationship management*, Wiley Computer Pub., New York, 2000.

Berry, M.J.A., and Linoff, G. *Data mining techniques : For marketing, sales, and customer relationship management*, Wiley, Indianapolis, Ind., 2004.

Bickel, P.J., and Doksum, K.A. *Mathematical statistics: Basic ideas and selected topics*, Prentice Hall, Upper Saddle River, N.J., 2001.

Bishop, C.M. *Neural networks for pattern recognition*, Clarendon Press ; Oxford University Press, Oxford, New York, 1995.

Blackwelder, W.C. "Proving the null hypothesis in clinical-trials," *Controlled Clinical Trials* (2:1), 1981, pp. 67-67.

Blackwelder, W.C. "Proving the null hypothesis in clinical-trials," *Controlled Clinical Trials* (3:4), 1982, pp. 345-353.

Blackwelder, W.C., and Chang, M.A. "Sample-size graphs for proving the null hypothesis," *Controlled Clinical Trials* (5:2), 1984, pp. 97-105.

Bourlard, H., and Wellekens, C. "Links between markov models and multilayer perceptrons," In *Advances in neural information processing systems 1*, Touretzky (ed.) Morgan Kaufmann Publishers, San Mateo, CA, 1989, pp. 502-510.

Boyens, C., Günther, O., and Teltzrow, M. "Privacy conflicts in crm services for online shops: A case study," *Proceedings of the IEEE international conference on Privacy, security and data mining*, 2002, pp. 27-35.

Brent, R.P. *Algorithms for minimization without derivatives*, Dover Publications, Mineola, N.Y., 2002.

Census Bureau "Introduction to census 2000 data products," (accessed: Aug 2004), 2001, http://www.census.gov/prod/2001pubs/mso-01icdp.pdf.

Census Bureau "Dataferrett: For thedataweb," (accessed: Aug 2004), 2004, http://dataferrett.census.gov/TheDataWeb/index.html.

Burridge, J. "Information preserving statistical obfuscation," *Statistics and Computing* (13:4), 2003, pp. 321-327.

Clemen, R.T., and Reilly, T. "Correlations and copulas for decision and risk analysis," *Management Science* (45:2), 1999, pp. 208-224.

Clifton, C. "Security and privacy," In *The handbook of data mining*, N. Ye (ed.) Lawrence Erlbaum Associates, Publishers, Mahwah, N.J., 2003, pp. 441-452.

Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., and Zhu, M.Y. "Tools for privacy preserving distributed data mining," *ACM SIGKDD Explorations Newsletter* (4:2), 2002, pp. 28-34.

Cohen, J. "Things I have learned (so far)," *American Psychologist* (45:12), 1990, pp. 1304-1312.

Cohen, J. "Things I have learned (so far) - chapter 16 " In *Methodological issues & strategies in clinical research*, A. E. Kazdin (ed.) American Psychological Association, Washington, DC, 2003, pp. xix, 913 p.

Dalenius, T. "Towards a methodology for statistical disclosure control," *Statistisk Tidskrift* (5, 1977, pp. 429-444.

Detsky, A.S., and Sackett, D.L. "When was a "Negative" Clinical trial big enough? How many patients you needed depends on what you found," *Archives of Internal Medicine* (145:4), 1985, pp. 709-712.

Deutsch, R. *Estimation theory*, Prentice-Hall, Englewood Cliffs, N.J.,, 1965.

Domingo-Ferrer, J., Mateo, J.M., and Torres, A. "Comparing SDC methods for microdata on the basis of information loss and disclosure risk," *Proceedings of the New Techniques and Technologies for Statistics – Exchange of Technology and Know-How Conference ETK-NTTS'2001 (Pre-proceedings)*, Luxembourg: Eurostat, 2001, pp. 807-826.

Domingo-Ferrer, J., Mateo, J.M., and Torres, A. "Concepts for the evaluation of anonymized data," (accessed 2004: Dec 12), 2003, pp. 1-13, http://vneumann.etse.urv.es/publications/bcpi/awe03.pdf.

Domingo-Ferrer, J., and Torra, V. "On the connections between statistical disclosure control for microdata and some artificial intelligence tools," *Information Sciences* (151:1), 2003, pp. 153-170.

Du, W., Han, Y.S., and Chen, S. "Privacy-preserving multivariate statistical analysis: Linear regression and classification," (accessed 2004: Nov 21), 2004, http://www.cis.syr.edu/~wedu/Research/paper/sdm2004_privacy.pdf.

Du, W., and Zhan, Z. "Building decision tree classifier on private data," *Proceedings of the IEEE international conference on Privacy, security and data mining - Volume 14*, Maebashi City, Japan 2002, pp. 1-8.

Du, W., and Zhan, Z. "Using randomized response techniques for privacy-preserving data mining " *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* Washington, D.C. , USA, 2003, pp. 505-510

Duncan, G.T., and Lambert, D. "Disclosure-limited data dissemination," *Journal of the American Statistical Association* (81:393), 1986, pp. 10-18.

Duncan, G.T., and Mukherjee, S. "Optimal disclosure limitation strategy in statistical databases: Deterring tracker attacks through additive noise," *Journal of the American Statistical Association* (95:451), 2000, pp. 720-729.

Duncan, G.T., and Pearson, R.W. "Enhancing access to microdata while protecting confidentiality: Prospects for the future," *STATISTICAL SCIENCE* (6, 1991, pp. 219-239.

Dunnett, C.W., and Gent, M. "Significance testing to establish equivalence between treatments, with special reference to data in form of 2 x 2 tables," *Biometrics* (33:4), 1977, pp. 593-602.

Durrett, R. *Probability: Theory and examples*, Thomson Brooks/Cole, Belmont, CA, 2005.

EquivTest "Equivtest 2," (accessed: Feb 28th, 2006), 2006, Statistical Solutions, http://www.statsol.ie/equivtest/equivtest.htm.

Evfimievski, A., Srikant, R., Agrawal, R., and Gehrke, J. "Privacy preserving mining of association rules," *Acm Sigkdd International Conference On Knowledge Discovery And Data Mining 2002; 8th*, 2002, pp. Pages 217-228.

Evfimievski, A., Srikant, R., Agrawal, R., and Gehrke, J. "Privacy preserving mining of association rules," *Information systems* (29:4), 2004, pp. 343-364.

Fienberg, S.E. "Conflicts between the needs for access to statistical information and demands for confidentiality," *Journal of Official Statistics* (10:2), 1994, pp. 115-132.

Fienberg, S.E., Makov, U.E., and Steele, R.J. "Disclosure limitation using perturbation and related methods for categorical data," *Journal of Official Statistics* (14:4), 1998, pp. 485-502.

Fisher, R.A. *Statistical methods for research workers*, Hafner Pub. Co., Darien, Conn.,, 1970.

Fuller, W.A. "Masking procedures for microdata disclosure limitation," *Journal of Officail Statistics* (9, 1993, pp. 383-406.

Gish, H. "A probabilistic approach to the understanding and training of neural network classifiers," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1990, pp. 1361-1364.

Gray, R.M., and Davisson, L.D. *An introduction to statistical signal processing*, Cambridge University Press, Cambridge, UK ; New York, 2004.

Graybill, F.A. *Theory and application of the linear model*, Duxbury Press, North Scituate, Mass., 1976.

Grupe, F.H., Kuechler, W., and Sweeney, S. "Dealing with data privacy protection: An issue for the 21st century," *Information Systems Management* (19:4), 2002, pp. 61-70.

Hagan, M.T., Demuth, H.B., and Beale, M.H. *Neural network design*, PWS Pub., Boston, 1996.

Hair, J.F., Anderson, R.E., Tatham, R.L., and Black, W.C. *Multivariate data analysis*, Prentice Hall, Upper Saddle River, N.J., 1998.

Hand, D.J. "Data mining: Statistics and more?," *AMERICAN STATISTICIAN* (52:2), 1998, pp. 112-118.

Hand, D.J. "Data mining: New challenges for statisticians," *Social Science Computer Review* (18:4), 2000, pp. 442-449.

Hand, D.J., Blunt, G., Kelly, M., and Adams, N. "Data mining for fun and profit," *STATISTICAL SCIENCE* (15:2), 2000, pp. 111-126.

Hintze, J.L. *Ncss and pass. Number cruncher statistical systems*, WWW.NCSS.COM, Kaysville, Utah, 2004.

Hunter, J.J. "Independence, conditional expectation, and zero covariance," *American Statistician* (26:5), 1972, pp. 22-24.

Islam, M.Z., and Brankovic, L. "A framework for privacy preserving classification in data mining " *Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation* Dunedin, New Zealand 2004, pp. 163-168
Jelenkovic, P. "Expectation and conditional expectation," (accessed 2004: July 24th), 2001, http://comet.ctr.columbia.edu/~kobj/e6711/lecture2.pdf.

Joe, H. *Multivariate models and dependence concepts*, Chapman & Hall, London ; New York, 1997.

Johnson, R.A., and Wichern, D.W. *Applied multivariate statistical analysis*, Prentice Hall, Upper Saddle River, N.J., 1998.

Johnsten, T., and Raghavan, V.V. "Impact of decision-region based classification mining algorithms on database security," *Proceedings of the IFIP WG 11.3 Thirteenth International Conference on Database Security: Research Advances in Database and Information Systems Security*, 2000, pp. 177-191.

Johnsten, T., and Raghavan, V.V. "Security procedures for classification mining algorithms," *Proceedings of the fifteenth annual working conference on Database and application security (Das'01)*, Niagara, Ontario, Canada 2001, pp. 285-297.

Jouini, M., and Clemen, R. "Copula models for aggregating expert opinions," *OPERATIONS RESEARCH* (44:3), 1996, pp. 444-457.

Kantarcioglu, M., and Clifton, C. "Privacy-preserving distributed mining of association rules on horizontally partitioned data," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING* (16:9), 2004a, pp. 1026-1037.

Kantarcioglu, M., and Clifton, C. "Privately computing a distributed k-nn classifier," *KNOWLEDGE DISCOVERY IN DATABASES: PKDD 2004, PROCEEDINGS* (3202, 2004b, pp. 279-290.

Kantarcioglu, M., and Vaidya, J. "Privacy preserving naive bayes classifier for horizontally partitioned data," *Proceedings of the Workshop on Privacy Preserving Data Mining held in association with The Third IEEE International Conference on Data Mining*, Melbourne, Florida, USA, 2003, pp. 19-22.

Karr, A.F., Lin, X., Sanil, A.P., and Reiter, J.P. "Secure regression on distributed databases," (accessed 2004: Nov 10), 2004, pp. 1-18, http://www.niss.org/technicalreports/tr141.pdf.

Keller-McNulty, S., and Unger, E.A. "A database system prototype for remote access to information based on confidential data," *Journal of Official Statistics* (14:4), 1998, pp. 347-360.

Kim, J. "A method for limiting disclosure in microdata based on random noise and transformation," *ASA Proc. Survey Res. Section*, 1986, pp. 370–374.

Kim, J.J., and Winkler, W.E. "Multiplicative noise for masking continuous data," (accessed Aug 2004), 2003, http://www.census.gov/srd/papers/pdf/rrs2003-01.pdf.

Kirkup, L. "Principles and applications of non-linear least squares: An introduction for physical scientists using Excel's solver," (accessed 2004: Nov 22), 2003, http://www.science.uts.edu.au/physics/Nonlin2003v5.pdf.

Kirkwood, T.B.L. "Bioequivalence testing - a need to rethink," *Biometrics* (37:3), 1981, pp. 589-591.

Klusch, M., Lodi, S., and Moro, G. "Distributed clustering based on sampling local density estimates," *Proceedings of the Intl. Joint Conference on Artificial Intelligence (IJCAI 2003), Mexico*, 2003, pp. 485-490.

Kotz, S., Johnson, N.L., Balakrishnan, N., and Johnson, N.L. *Continuous multivariate distributions*, Wiley, New York, 2000.

Lechner, S., and Pohlmeier, W. "To blank or not to blank? A comparison of the effects of disclosure limitation methods on nonlinear regression estimates," *Lecture Notes In Computer Science* (3050, 2004, pp. 187-200.

Lehmann, E.L., and Casella, G. *Theory of point estimation*, Springer, New York, 1998.

Lin, X., Clifton, C., and Zhu, M. "Privacy-preserving clustering with distributed EM mixture modeling," *Knowledge and Information Systems* 2004, pp. 1-14 (online).

Lindell, Y., and Pinkas, B. "Privacy preserving data mining," *Journal of cryptology : the journal of the International Association for Cryptologic Research* (15:Part 3), 2002, pp. 177-206.

Little, R.J.A. "Statistical analysis of masked data," *Journal of Official Statistics* (9:2), 1993, pp. 407-426.

Loehle, C. "A hypothesis testing framework for evaluating ecosystem model performance," *Ecological Modelling* (97:3), 1997, pp. 153-165.

LS-SVMlab1.5 "LS-SVMlab toolbox - a matlab toolbox," (accessed 2004: Mar 10), 2003, http://www.esat.kuleuven.ac.be/sista/lssvmlab/.

Macnaughton, D.B. "Definition of relationship between variables," (accessed 2004: August 25), 2002, http://www.matstat.com/teach/p0045.htm.

Makuch, R., and Simon, R. "Sample-size requirements for evaluating a conservative therapy," *Cancer Treatment Reports* (62:7), 1978, pp. 1037-1040.

Malvestuto, F., and Moscarini, M. "Privacy in multidimensional databases," In *Multidimensional databases : Problems and solutions*, M. Rafanelli (ed.) Idea Group Pub., Hershey, PA, 2003, pp. 310-360.

Mayer, D.G., and Butler, D.G. "Statistical validation," *Ecological Modelling* (68:1-2), 1993, pp. 21-32.

McBride, G.B. "Equivalence tests can enhance environmental science and management," *Australian & New Zealand Journal of Statistics* (41:1), 1999, pp. 19-29.

Medicare "Medicare official website," (accessed July, 2004), 2004, http://www.medicare.gov.

Meditch, J.S. *Stochastic optimal linear estimation and control*, McGraw-Hill, New York,, 1969.

Merugu, S., and Ghosh, J. "Privacy-preserving distributed clustering using generative models," *Proceedings of the Third IEEE International Conference on Data Mining ICDM'03*, 2003a, pp. 211-218.

Merugu, S., and Ghosh, J. "A probabilistic approach to privacy-sensitive distributed data mining," *Proceedings of the 6th International Conference on Information Technology (CIT)*, 2003b, pp. 405-410.

Muralidhar, K., Batra, D., and Kirs, P.J. "Accessibility, security, and accuracy in statistical databases - the case for the multiplicative fixed data perturbation approach," *Management Science* (41:9), 1995, pp. 1549-1564.

Muralidhar, K., Parsa, R., and Sarathy, R. "A general additive data perturbation method for database security," *Management Science* (45:10), 1999, pp. 1399-1415.

Muralidhar, K., and Sarathy, R. "Security of random data perturbation methods," *ACM Transactions on Database Systems* (24:4), 1999, pp. 487-493.

Muralidhar, K., and Sarathy, R. "The data shuffle: A new masking procedure for numerical data," *Proceedings of the 8th INFORMS Computing Society*, Chandler, AZ, January, 2003a, pp. (unknown??).

Muralidhar, K., and Sarathy, R. "A rejoinder to the comments by polettini and Stander," *Statistics and Computing* (13:4), 2003b, pp. 339-342.

Muralidhar, K., and Sarathy, R. "A theoretical basis for perturbation methods," *Statistics and Computing* (13:4), 2003c, pp. 329-335.

Muralidhar, K., and Sarathy, R. "Data shuffling - a new masking approach for numerical data," *Management Science (forthcoming)*, 2005a, pp. 1-35 (first draft).

Muralidhar, K., and Sarathy, R. "An enhanced data perturbation approach for small data sets," *Decision Sciences* (36:3), 2005b, pp. 513-529.

Muralidhar, K., and Sarathy, R. "Data shuffling - a new masking approach for numerical data," *Management Science* (52:5), 2006a, pp. 658-670.

Muralidhar, K., and Sarathy, R. "Data shuffling - a new masking approach for numerical data," *Management Science (forthcoming)*, 2006b, pp. 1-35 (draft).

Muralidhar, K., Sarathy, R., and Parsa, R. "An improved security requirement for data perturbation with implications for e-commerce," *Decision Sciences* (32:4), 2001, pp. 683-698.

Nelsen, R.B. *An introduction to copulas*, Springer, New York, 1999.

Oliveira, S.R.M., and Zaïane, O.R. "Algorithms for balancing privacy and knowledge discovery in association rule mining," *International Database Engineering and Applications Symposium 1098-8068 2003; 7th*, 2003a, pp. Pages 54-65.

Oliveira, S.R.M., and Zaïane, O.R. "Privacy preserving clustering by data transformation," *Proceedings of the 18th Brazilian Symposium on Databases (SBBD 2003)*, Manaus, Brazil, 6-8 October, 2003b, pp. 304-318.

Oliveira, S.R.M., and Zaïane, O.R. "Achieving privacy preservation when sharing data for clustering," *Proceedings of the International Workshop on Secure Data Management in a Connected World (SDM'04) in conjunction with VLDB 2004*, Toronto, Canada, August, 2004a, pp. 67-82.

Oliveira, S.R.M., and Zaïane, O.R. "Privacy-preserving clustering by object similarity-based representation and dimensionality reduction transformation," *Proceedings of the Workshop on Privacy and Security Aspects of Data Mining (PSDM'04) in conjunction with the Fourth IEEE International Conference on Data Mining (ICDM'04)*, Brighton, UK, 2004b, pp. 21-30.

Oliveira, S.R.M., and Zaïane, O.R. "Toward standardization in privacy-preserving data mining," *Proceedings of the ACM SIGKDD 3rd Workshop on Data Mining Standards (DM-SSP 2004)*, Seattle, WA, USA, August 22, 2004c, pp. 7-17.

Oliveira, S.R.M., Zaïane, O.R., and Saygin, Y. "Secure association rule sharing," *ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING, PROCEEDINGS* (3056, 2004, pp. 74-85.

Papoulis, A., and Pillai, S.U. *Probability, random variables and stochastic processes*, McGraw-Hill, Boston, 2002.

Parkhurst, D.F. "Interpreting failure to reject a null hypothesis," *Bulletin of the Ecological Society of America* (66, 1985, pp. 301–302.

Parkhurst, D.F. "Statistical significance tests: Equivalence and reverse tests should reduce misinterpretation," *Bioscience* (51:12), 2001, pp. 1051-1057.

Polettini, S., and Stander, J. "A comment on "A theoretical basis for perturbation methods" By krishnamurty muralidhar and rathindra sarathy," *Statistics and Computing* (13:4), 2003, pp. 337-338.

Powell, M. "An efficient method for finding the minimum of a function of several variables without calculating derivatives," *Computer Journal* (7, 1964, pp. 152-162.

Pyle, D. *Business modeling and data mining*, Morgan Kaufmann Publishers, Amsterdam; Boston, 2003.

R "The R project for statistical computing," (accessed 2006: Jan 14th), 2006, http://www.r-project.org/.

Reiter, J.P. "Model diagnostics for remote access regression servers," *Statistics and Computing* (13:4), 2003, pp. 371-380.

Rhodes, I.B. "A tutorial introduction to estimation and filtering," *IEEE Transactions on Automatic Control* (Ac16:6), 1971, pp. 688-706.

Richard, M.D., and Lippmann, R.P. "Neural network classifiers estimate bayesian a posteriori probabilities," *Neural Computation* (3, 1991, pp. 461-483.

Ridgeway, G. "Strategies and methods for prediction," In *The handbook of data mining*, N. Ye (ed.) Lawrence Erlbaum Associates, Publishers, Mahwah, N.J., 2003, pp. 159-191.

Rizvi, S., and Haritsa, J.R. "Maintaining data privacy in association rule mining," *Proceedings of the 28th Very Large Data Base (VLDB2002) Conference*, Hong Kong, China, 2002, pp. 1-12.

Robinson, A.P. "The equivalence package," (accessed 2005: Oct 10th), 2005, http://cran.r-project.org/doc/packages/equivalence.pdf.

Robinson, A.P., Duursma, R.A., and Marshall, J.D. "A regression-based equivalence test for model validation: Shifting the burden of proof," *Tree Physiology* (25:7), 2005, pp. 903-913.

Robinson, A.P., and Froese, R.E. "Model validation using equivalence tests," *Ecological Modelling* (176:3-4), 2004, pp. 349-358.

Rogers, J.L., Howard, K.I., and Vessey, J.T. "Using significance tests to evaluate equivalence between two experimental groups," *Psychological Bulletin* (113:3), 1993, pp. 553-565.

Rubin, D.B. "Discussion: Statistical disclosure limitation," *Journal of Official Statistics* (9:2), 1993, pp. 461-468.

Ruck, D.W., Rogers, S.K., Kabrisky, M., Oxley, M.E., and Suter, B.W. "The multilayer perceptron as an approximation to a bayes optimal discriminant function," *IEEE Transactions on Neural Networks* (1:4), 1990, pp. 296-298.

Rud, O.P. *Data mining cookbook : Modeling data for marketing, risk and customer relationship management*, Wiley, New York, 2001.

Rykiel, J.E.J. "Testing ecological models: The meaning of validation," *Ecological Modelling* (90:3), 1996, pp. 229-244.

Saerens, M. "Non mean square error criteria for the training of learning machines," *Proceedings of the IEEE International Conference on Machine Learning (ICML 1996)*, Bari, Italy, 1996, pp. 427-434.

Saerens, M. "Building cost functions minimizing to some summary statistics," *IEEE Transactions on Neural Networks* (6:11), 2000, pp. 1263-1271.

Sanil, A.P., Karr, A.F., Lin, X., and Reiter, J.P. "Privacy preserving regression modelling via distributed computation," *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data* Seattle, WA, USA, 2004, pp. 677-682
Sarathy, R., and Muralidhar, K. "The security of confidential numerical data in databases," *Information Systems Research* (13:4), 2002, pp. 389-403.

Sarathy, R., Muralidhar, K., and Parsa, R. "Perturbing non-normal confidential attributes: The copula approach," *Management Science* (48:12), 2002, pp. 1613-1627.

SAS "Power and sample size computations with SAS," (accessed Jan 20th, 2006), 2003, http://support.sas.com/rnd/app/da/new/pdf/PSS.pdf.

Saygin, Y., Verykios, V.S., and Elmagarmid, A.K. "Privacy preserving association rule mining," *International Workshop On Research Issues In Data Engineering 1066-1395 2002; 12th*, 2002, pp. Pages 151-158.

Schèolkopf, B., Burges, C.J.C., and Smola, A.J. *Advances in kernel methods : Support vector learning*, MIT Press, Cambridge, Mass., 1999.

Schèolkopf, B., and Smola, A.J. "Support vector machines," In *The handbook of brain theory and neural networks*, M. A. Arbib (ed.) MIT Press, Cambridge, Mass., 2003, pp. 1119-1125.

Schield, M. "Correlation, determination and causality in introductory statistics," *American Statistical Association (ASA), JSM (Joint Statistical Meetings), August, 1995, Section on Statistical Education*, 1995, pp. 1-6.

Schneier, B. *Applied cryptography : Protocols, algorithms, and source code in c*, Wiley, New York, 1996.

Schouten, B., and Cigrang, M. "Remote access systems for statistical analysis of microdata," *Statistics and Computing* (13:4), 2003, pp. 381-389.

Schuirmann, D.L. "On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval (abstract)," *Biometrics* (37:3), 1981, pp. 617.

Schweizer, B. "Thirty years of copulas," In *Advances in probability distributions with given marginals : Beyond the copulas*, G. Dall'Aglio, S. Kotz and G. Salinetti (eds.), Kluwer Academic Publishers, Dordrecht ; Boston, 1991, pp. 13-50.

Shao, J. *Mathematical statistics*, Springer, New York, 1999.

Shawe-Taylor, J., and Cristianini, N. *Kernel methods for pattern analysis*, Cambridge University Press, Cambridge, UK ; New York, 2004.

Shoemaker, P.A. "A note on least-squares learning procedures and classification by neural network models," *IEEE Transactions on Neural Networks* (2:1), 1991, pp. 158-160.

Skinner, C., Marsh, C., Openshaw, S., and Wymer, C. "Disclosure control for census microdata," *Journal of Official Statistics* (10:1), 1994, pp. 31-51.

Sklar, A. "Fonctions de répartition à n dimensions et leurs mages," *Pub. Inst. Statist. Univ. Paris* (8, 1959, pp. 229-231.

DMReview.com Web Editorial Staff "Industry implementations," (accessed: Aug 2004, published on June 11, 2004), 2004, http://www.dmreview.com/article_sub.cfm?articleID=1004813.

Streiner, D.L. "Unicorns do exist: A tutorial on "Proving" The null hypothesis," *Canadian Journal of Psychiatry* (48:11), 2003, pp. 756-761.

Sullivan, G., and Fuller, W.A. "The use of measurement error to avoid disclosure," *Proceedings of the Amer. Statist. Assoc., Survey Res. Methods Section*, 1989 pp. 802-807.

Suykens, J.A.K., Gestel, T.V., Brabanter, J.D., Moor, B.D., and Vandewalle, J. *Least squares support vector machines*, World Scientific, River Edge, NJ, 2002.

Tamayo-Sarver, J.H., Albert, J.M., Tamayo-Sarver, M., and Cydulka, R.K. "Advanced statistics: How to determine whether your intervention is different, at least as effective as, or equivalent: A basic introduction," *Academic Emergency Medicine* (12:6), 2005, pp. 536-542.

Tendick, P. "Optimal noise addition for preserving confidentiality in multivariate data," *Journal of Statistical Planning and Inference* (27:3), 1991, pp. 341-353.

Thuraisingham, B. "Privacy-preserving data mining: Development and directions," *Journal of Database Management* (16:1), 2005, pp. 75-87.

Traub, J.F., Yemini, Y., and Wozniakowski, H. "The statistical security of a statistical database," *ACM Transactions on Database Systems* (9:4), 1984, pp. 672-679.

Vaidya, J., and Clifton, C. "Privacy preserving association rule mining in vertically partitioned data," *Acm Sigkdd International Conference on Knowledge Discovery And Data Mining 2002; 8th*, 2002, pp. Pages 639-644.

Vaidya, J., and Clifton, C. "Privacy-preserving k-means clustering over vertically partitioned data," (accessed 2004: Nov 12), 2003, http://www.cs.purdue.edu/homes/jsvaidya/pub-papers/vaidya-kmeans.pdf.

Vaidya, J., and Clifton, C. "Privacy preserving naive bayes classifier for vertically partitioned data," *Proceedings of the 2004 SIAM International Conference on Data Mining*, Orlando, Florida, USA, 2004, pp. 1-5.

Vaidya, J., Kantarcioglu, M., and Clifton, C. "Privacy preserving naive bayes classification," *submitted (March 2004) to Data Mining and Knowledge Discovery*, 2004,
Vapnik, V.N. *The nature of statistical learning theory*, Springer, New York, 2000.

Venables, W.N., Smith, D.M., and R Development Core Team. *An introduction to R: Notes on R: A programming environment for data analysis and graphics, version 1.4.1*, Network Theory, Bristol, 2002.

Verykios, V., Elmagarmid, A., Bertino, E., Saygin, Y., and Dasseni, E. "Association rule hiding," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING* (16:4), 2004a, pp. 434-447.

Verykios, V.S., Bertino, E., Fovino, I.N., Provenza, L.P., Saygin, Y., and Theodoridis, Y. "State-of-the-art in privacy preserving data mining," *SIGMOD record; ACM* (33:1), 2004b, pp. 50-57.

Wan, E.A. "Neural network classification: A bayesian interpretation," *IEEE Transactions on Neural Networks* (1:4), 1990, pp. 303-305.

Warner, B. (accessed 2005: Dec 14th), 2002, Military Operations Research Society, MORS Workshop - Working Group 2, 15-17 October 2002, Albuquerque, New Mexico, http://www.mors.org/meetings/test_eval/presentations/C_Warner.pdf and http://www.mors.org/publications/reports/2002-T-E_M-S_VV-A.pdf.

Wellek, S. *Testing statistical hypotheses of equivalence*, Chapman & Hall/CRC, Boca Raton, Fla., 2003.

Westlake, W.J. "Symmetrical confidence-intervals for bioequivalence trials," *Biometrics* (32:4), 1976, pp. 741-744.

Westlake, W.J. "Statistical aspects of comparative bioavailability trials," *Biometrics* (35:1), 1979, pp. 273-280.

Westlake, W.J. "Bioequivalence testing - a need to rethink - reply," *Biometrics* (37:3), 1981, pp. 591-593.

White, H. "Learning in artificial neural networks: A statistical perspective," *Neural Computation* (1, 1989, pp. 425-464.

Willenborg, L.C.R.J., and Waal, T.d. *Elements of statistical disclosure control*, Springer, New York, 2001.

Wilson, R.L., and Rosen, P.A. "The impact of data perturbation techniques on data mining accuracy," *Proceedings of the 33rd Annual Meeting of the Decision Sciences Institute*, 2002, pp. 181-185.

Wilson, R.L., and Rosen, P.A. "Protecting data through perturbation techniques: The impact on knowledge discovery in databases," *Journal of Database Management* (14:2), 2003, pp. 14-26.

Winkler, W. "Masking and re-identification methods for public-use microdata: Overview and research problems," *PRIVACY IN STATISTICAL DATABASES, PROCEEDINGS* (3050, 2004a, pp. 231-246.

Winkler, W. "Re-identification methods for masked microdata," *PRIVACY IN STATISTICAL DATABASES, PROCEEDINGS* (3050, 2004b, pp. 216-230.

Witten, I.H., and Frank, E. *Data mining : Practical machine learning tools and techniques*, Morgan Kaufman, Amsterdam ; Boston, MA, 2005.

Yanagawa, T. "Book reviews: 12," *Biometrics* (61:1), 2005, pp. 320-321.

Yang, Y.Q., Monserud, R.A., and Huang, S.M. "An evaluation of diagnostic tests and their roles in validating forest biometric models," *Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere* (34:3), 2004, pp. 619-629.

Yang, Z., Zhong, S., and Wright, R.N. "Privacy-preserving classification of customer data without loss of accuracy," (accessed 2005: Jan 9), 2005, pp. 1-11, http://www.cse.buffalo.edu/~szhong/papers/mining2.pdf
Zhang, G.P. (ed.). *Neural networks in business forecasting*, Idea Group, Hershey, Pa., 2004.

Zhang, N., Wang, S., and Zhao, W. "A new scheme on privacy preserving association rule mining," *Lecture notes in computer science*:3202), 2004, pp. 484-495.

Appendices

Appendix A – SDL/Relationship Match Framework

In this section, we propose an abstract framework called the SDL/Relationship Match framework. The framework is a very simple yet informative framework that can be used to match and guide the use and application of any existing or newly developed data masking method with a specific dataset for regression tasks (and maybe other analyses) based on the type of relationships they share (the type of the relationship that the dataset contains and that the masking method can preserve).

The goal of the SDL/Relationship Match framework for PPE is threefold: (a) emphasizing the importance of relationships among variables (as a data utility measure) in estimation and privacy-preserving estimation (PPE) problems, (b) suggesting an initial mechanism (by utilizing existing SDL methods) to implement the concept of Data-Centric Approach (DCA) in the PPDM field by focusing only on altering datasets as opposed to PPE and PPDM algorithms, and (c) identifying any possible gap in the SDL literature in terms of maintained relationships in masked datasets (e.g., non-monotonic relationships).

As stated earlier, this framework suggests evaluating the effectiveness of existing or new SDL masking methods based on the relationships they maintain. Effectiveness in this context also means that SDL methods should first prove to be effective in terms of security requirements before they are tested in terms of data utility (maintained

relationships). Dalenius (1977) and Duncan and Lambert (1986) suggested that accessing masked attributes should not increase the ability of snoopers to predict confidential attributes. In other words, masked attributes should not carry information about confidential attributes beyond what is already known from accessing the non-confidential attributes.

Once the SDL masking method provides a sufficient level of security, it is evaluated for the type of relationships it can maintain. Datasets, on the other hand, are evaluated based on existing relationships between confidential and non-confidential attributes. Scatter plots and scatter matrix plots can be used for such an assessment. Then one can apply an appropriate SDL masking method on a specific dataset based on the relationship match. Refer to Figure 38 for the framework.

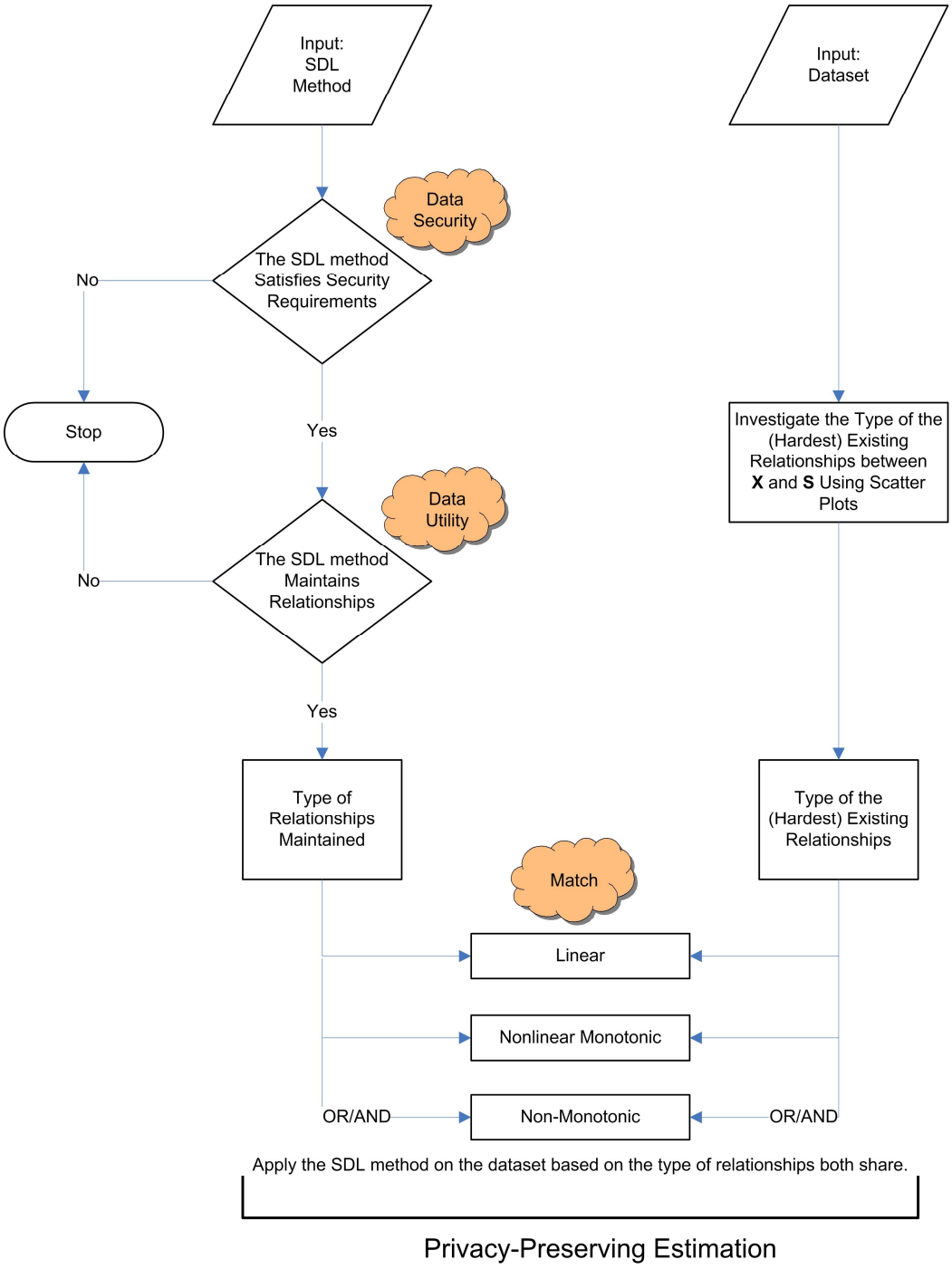


Figure 38. SDL/Relationship Match Framework

Appendix B – The Relationship between the Covariance Matrices of Confidential Attributes X, Conditional Expectations E(X|S) and Residuals r

In this section, we begin by stating a theorem (Theorem I) that explains the relationship among the covariance matrix of the confidential attributes **X**, the covariance matrices of the confidential expectations $E(\mathbf{X}|\mathbf{S})$, and of the residuals **r** in the case of multivariate normally distributed datasets. A proof for this special-case theorem is then provided. Next, we show that the same relationship generally holds true regardless of the dataset distribution (Theorem II). This is done by alternating the first theorem proof. Finally, we provide an equivalent relationship in terms of correlation matrices (Proposition I). Since, by definition, the covariance is a binary relationship between two random variables, we state these theorems and their proofs using two confidential attributes $\mathbf{X}$ (X_i and X_j, where $i, j = 1\text{K } q$ and $i \neq j$).

Theorem I

In Gaussian (normally distributed) datasets, the relationship between the covariance matrix of confidential attributes $\mathbf{X}$ (X_i and X_j) and the covariance matrix of the conditional expectations of confidential attributes $\mathbf{X}$, given non-confidential attributes $\mathbf{S}$ ($E(X_i|\mathbf{S})$ and $E(X_j|\mathbf{S})$) and the covariance matrix of their residuals **r** (r_i and r_j), can be expressed as:

$$Cov(X_i, X_j) = Cov(E(X_i|\mathbf{S}), E(X_j|\mathbf{S})) + Cov(r_i, r_j). \quad \text{(B.1)}$$

♦

Proof:

We start by expressing the covariance matrix of confidential attributes in terms of expectations (see Bickel and Doksum (2001), Equation (A.11.14), pp. 458):

$$Cov(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j). \tag{B.2}$$

By expanding the first term in the RHS $E(X_i X_j)$, we get:

$$Cov(X_i, X_j) = E((E(X_i|\mathbf{S}) + r_i) \cdot (E(X_j|\mathbf{S}) + r_j)) - E(X_i)E(X_j). \tag{B.3}$$

The following is the result of simple multiplications:

$$\begin{aligned} Cov(X_i, X_j) = {} & E(E(X_i|\mathbf{S})E(X_j|\mathbf{S}) + r_i E(X_j|\mathbf{S}) + r_j E(X_i|\mathbf{S}) + r_i r_j) \\ & - E(X_i)E(X_j) \end{aligned} \tag{B.4}$$

Then we distribute the expectations for every added term:

$$\begin{aligned} Cov(X_i, X_j) = {} & E(E(X_i|\mathbf{S})E(X_j|\mathbf{S})) + E(r_i E(X_j|\mathbf{S})) \\ & + E(r_j E(X_i|\mathbf{S})) + E(r_i r_j) - E(X_i)E(X_j) \end{aligned} \tag{B.5}$$

Since r_i is independent of $E(X_j|\mathbf{S})$ and r_j is independent of $E(X_i|\mathbf{S})$ in the case of multivariate normally distributed datasets (see Rhodes (1971), Property 8, pp. 692), and using the fact that $E(WZ) = E(W)E(Z)$ when $W \perp Z$ (see Bickel and Doksum (2001), Equation (A.11.21), pp. 459), Equation (B.5) can be written as:

$$\begin{aligned} Cov(X_1, X_2) = {} & E(E(X_i|\mathbf{S})E(X_j|\mathbf{S})) + E(r_i)E(E(X_j|\mathbf{S})) \\ & + E(r_j)E(E(X_i|\mathbf{S})) + E(r_i r_j) - E(X_i)E(X_j) \end{aligned} \tag{B.6}$$

By recalling that $E(r_i) = E(r_j) = 0$, the above expression (with some terms rearranged) becomes:

$$\begin{aligned} Cov(X_i, X_j) = {} & E(E(X_i|\mathbf{S})E(X_j|\mathbf{S})) - E(X_i)E(X_j) \\ & + E(r_i r_j) \end{aligned} \tag{B.7}$$

Using the following facts:

- $E(X_i)$ can be written as $E(E(X_i|\mathbf{S}))$ (Durrett, 2005, Property (1.1f), pp. 224)
- $E(X_j)$ can be written as $E(E(X_j|\mathbf{S}))$ (Durrett, 2005, Property (1.1f), pp. 224)
- $E(r_i)E(r_j) = 0$,

Equation (B.7) can be rewritten as:

$$\begin{aligned} Cov(X_i, X_j) = {} & E(E(X_i|\mathbf{S})E(X_j|\mathbf{S})) - E(E(X_i|\mathbf{S}))E(E(X_j|\mathbf{S})) \\ & + E(r_i r_j) - E(r_i)E(r_j) \end{aligned} \tag{B.8}$$

This simply restates and confirms Equation (B.1) that:

$$Cov(X_i, X_j) = Cov(E(X_i|\mathbf{S}), E(X_j|\mathbf{S})) + Cov(r_i, r_j).$$

This concludes our proof.

♦

Theorem II

Generally speaking and regardless of dataset distributions, the relationship between the covariance matrix of confidential attributes $\mathbf{X}$ (X_i and X_j) and the covariance matrix of the conditional expectations of confidential attributes $\mathbf{X}$, given non-confidential attributes $\mathbf{S}$ ($E(X_i|\mathbf{S})$ and $E(X_j|\mathbf{S})$) and the covariance matrix of their residuals $\boldsymbol{r}$ (r_1 and r_2), can be expressed as:

$$Cov(X_i, X_j) = Cov(E(X_i|\mathbf{S}), E(X_j|\mathbf{S})) + Cov(r_i, r_j) \tag{B.9}$$

regardless of the shape (linear, monotonic, or non-monotonic) of the conditional expectations $E(X_i|\mathbf{S})$ and $E(X_j|\mathbf{S})$.

♦

Proof:

Again, we start by expressing the covariance matrix of confidential attributes in terms of expectations (see Bickel and Doksum (2001), Equation (A.11.14), pp. 458):

$$Cov(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j). \tag{B.10}$$

By expanding the first term in the RHS $E(X_i X_j)$, we get:

$$Cov(X_i, X_j) = E((E(X_i|\mathbf{S}) + r_i)\cdot(E(X_j|\mathbf{S}) + r_j)) - E(X_i)E(X_j). \tag{B.11}$$

The following is the result of simple multiplications:

$$\begin{aligned} Cov(X_i, X_j) = {} & E(E(X_i|\mathbf{S})E(X_j|\mathbf{S}) + r_i E(X_j|\mathbf{S}) + r_j E(X_i|\mathbf{S}) + r_i r_j) \\ & - E(X_i)E(X_j) \end{aligned} \tag{B.12}$$

Then we distribute expectations:

$$\begin{aligned} Cov(X_i, X_j) = {} & E(E(X_i|\mathbf{S})E(X_j|\mathbf{S})) + E(r_i E(X_j|\mathbf{S})) \\ & + E(r_j E(X_i|\mathbf{S})) + E(r_i r_j) - E(X_i)E(X_j) \end{aligned} \tag{B.13}$$

The residuals **r** (r_1 and r_2) are uncorrelated with any and every function of **S** (see Rhodes (1971), Proposition 2b, pp. 690, and Bickel and Doksum (2001), Proposition 1.4.1 (a), pp. 34). "Any and every function of **S**" includes the conditional expectations $E(X_1|\mathbf{S})$ and $E(X_2|\mathbf{S})$ (see Bickel and Doksum (2001), Proposition 1.4.1 (b), pp. 34). Since the covariance matrix is a scaled version of the correlation matrix, we get the following:

$$Cov(r_i, E(X_j|\mathbf{S})) = 0 \tag{B.14}$$

and

$$Cov(r_j, E(X_i|\mathbf{S})) = 0. \tag{B.15}$$

Expand (B.14) and (B.15) using Bickel and Doksum (2001) Equation (A.11.14) pp. 458:

$$Cov(r_i, E(X_j|\mathbf{S})) = E(r_i \cdot E(X_j|\mathbf{S})) - E(r_i)E(E(X_j|\mathbf{S})) = 0 \qquad \text{(B.16)}$$

and

$$Cov(r_j, E(X_i|\mathbf{S})) = E(r_j \cdot E(X_i|\mathbf{S})) - E(r_j)E(E(X_i|\mathbf{S})) = 0. \qquad \text{(B.17)}$$

Combining (B.16) and (B.17) with the fact that $E(r_i) = E(r_j) = 0$ leads us to:

$$E(r_i \cdot E(X_j|\mathbf{S})) = 0 \qquad \text{(B.18)}$$

and

$$E(r_j \cdot E(X_i|\mathbf{S})) = 0. \qquad \text{(B.19)}$$

Using (B.18) and (B.19), Equation (B.13) (with some terms rearranged) becomes:

$$\begin{aligned} Cov(X_i, X_j) = {} & E(E(X_i|\mathbf{S})E(X_j|\mathbf{S})) - E(X_i)E(X_j) \\ & + E(r_i r_j) \end{aligned} \qquad \text{(B.20)}$$

Using the following facts:

- $E(X_i)$ can be written as $E(E(X_i|\mathbf{S}))$ (Durrett, 2005, Property (1.1f), pp. 224)
- $E(X_j)$ can be written as $E(E(X_j|\mathbf{S}))$ (Durrett, 2005, Property (1.1f), pp. 224)
- $E(r_i)E(r_j) = 0$,

Equation (B.20) can be rewritten as:

$$\begin{aligned} Cov(X_i, X_j) = {} & E(E(X_i|\mathbf{S})E(X_j|\mathbf{S})) - E(E(X_i|\mathbf{S}))E(E(X_j|\mathbf{S})) \\ & + E(r_i r_j) - E(r_i)E(r_j) \end{aligned} \qquad \text{(B.21)}$$

This simply restates and confirms Equation (B.9) that:

$$Cov(X_i, X_j) = Cov(E(X_i|\mathbf{S}), E(X_j|\mathbf{S})) + Cov(r_i, r_j).$$

This concludes our proof.

♦

Proposition I

The relationship among the correlation of two confidential attributes **X** (X_i and X_j), the correlation of the two conditional expectations ($E(X_i|\mathbf{S})$ and $E(X_j|\mathbf{S})$) obtained from regressing the two confidential attributes on non-confidential attributes **S**, and the correlation between the residuals **r** (r_i and r_j) left from the regression is:

$$Corr(X_i, X_j) = \sqrt{\frac{Var(E(X_i|\mathbf{S}))Var(E(X_j|\mathbf{S}))}{Var(X_i)Var(X_j)}} \cdot Corr(E(X_i|\mathbf{S}), E(X_j|\mathbf{S})) + \sqrt{\frac{Var(r_i)Var(r_j))}{Var(X_i)Var(X_j)}} \cdot Corr(r_i, r_j) \quad \text{(B.22)}$$

♦

Proof:

The relationship between the correlation matrix and the covariance matrix is (Bickel and Doksum, 2001, Equation A.11.18, pp.458):

$$Corr(X_i, X_j) = \frac{Cov(X_i, X_j)}{\sqrt{Var(X_i)Var(X_j)}} \quad \text{(B.23)}$$

We start from Equations (B.1) and (B.9):

$$Cov(X_i, X_j) = Cov(E(X_i|\mathbf{S}), E(X_j|\mathbf{S})) + Cov(r_i, r_j). \quad \text{(B.24)}$$

By using (B.23) and (B.24):

$$\sqrt{Var(X_i)Var(X_j)} \cdot Corr(X_i, X_j) = \sqrt{Var(E(X_i|\mathbf{S}))Var(E(X_j|\mathbf{S}))} \cdot Corr(E(X_i|\mathbf{S}), E(X_j|\mathbf{S})) + \sqrt{Var(r_i)Var(r_j)} \cdot Corr(r_i, r_j) \quad \text{(B.25)}$$

By dividing both sides of (B.25) by $\sqrt{Var(X_i)Var(X_j)}$, we get (B.22). This concludes our proof.

♦

Appendix C – Relationship-Based NM-EGADP Masking Algorithms

In this study, we propose four interrelated Relationship-Based PPE masking methods, numbered in ascending order, for masking datasets containing non-monotonic relationships. The goal is to release masked datasets for building estimation (regression) models. In this appendix, we list and discuss these masking algorithms.

Figure 39 shows the general schema of how these masking methods work. The general process is shown in the middle column. An outline of the corresponding mathematical process is shown in the third column. The procedure starts by learning the relationships between confidential attributes **X** and non-confidential attributes **S** (i.e. the conditional expectations $E(\mathbf{X}|\mathbf{S})$). Then, orthogonal (to both **S** and **X**) noise terms are generated. The covariance matrix of this set of orthogonal noise terms are scaled to mimic the characteristics of the original residuals set resulted from the regression process in the first step. Finally, we add the scaled orthogonal noise to the original conditional expectations $E(\mathbf{X}|\mathbf{S})$ to generate masked variables **Y**.

There are two shuffling-based dimensions differentiating between the proposed four masking methods: whether we shuffle the residuals by the added orthogonal noise, and whether we shuffle the confidential attributes by the perturbed variables. The possible combinations of these two dimensions, with two possibilities each, represent the four masking methods (as shown in Figure 40 below and the first column in Figure 39 below).

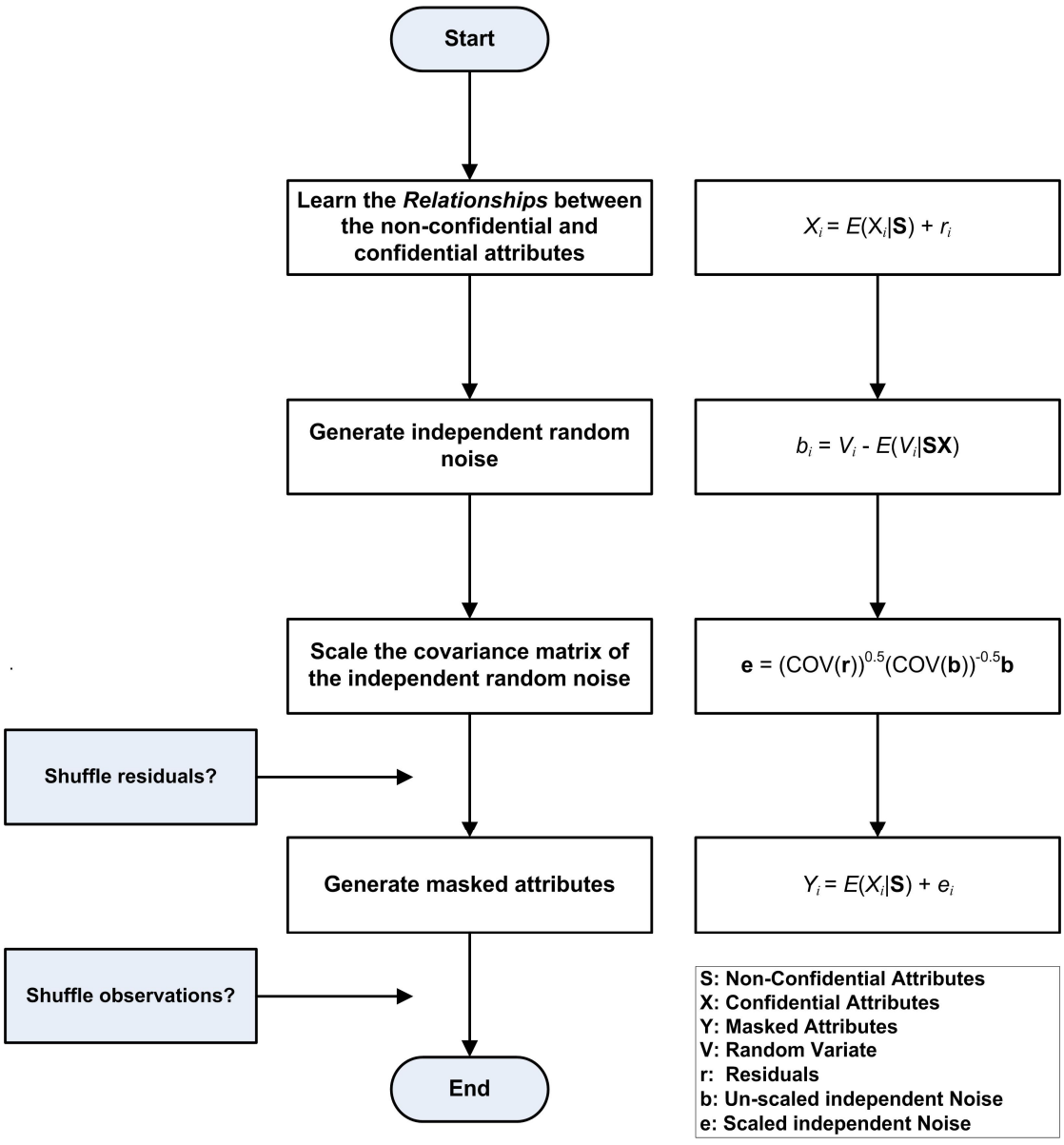


Figure 39. General schema of how Relationship-Based Masking (RBM) approach works

Shuffling is the process of ranking order observations in one random variable based on the rank-order of observations in another random variable (Muralidhar and Sarathy, 2003a; 2006b). It has the advantage of changing the rank order of the original variables while maintaining their marginal distributions. Next, we explain the shuffling process in more detail. We also demonstrate the process using a hypothetical example.

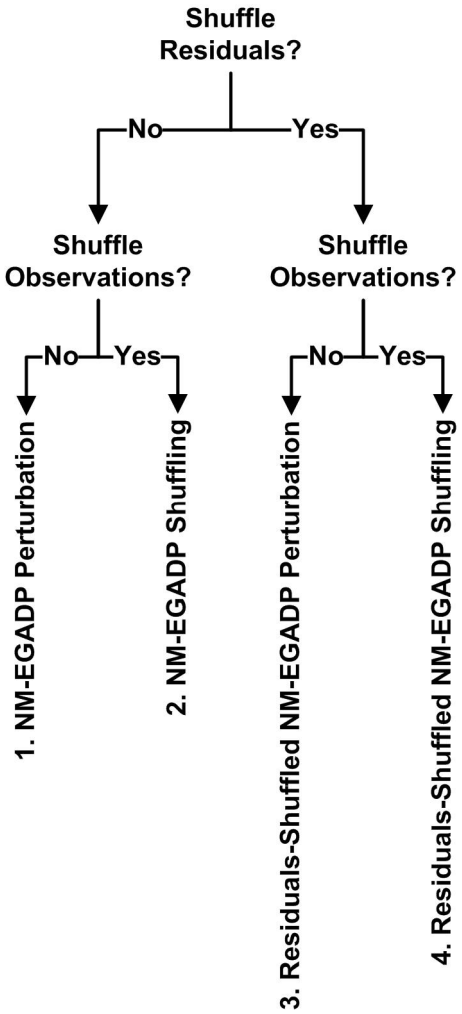


Figure 40. Classification of the NM-EGADP masking methods

How Shuffling Works

When we say "Shuffle *A* by *B*," where *A* and *B* are two random variables, we mean the following:

1. Find the rank-order R_A of all observations in the first random variable *A*.
2. Find the rank-order R_B of all observations in the second random variable *B*.

3. Create a third random variable C (the shuffled version of the random variable A) by rearranging the observations in A according to the rank-order of B, or more specifically:

$$C(Row_B(R_B)) = A(Row_A(R_A)); \text{ where } R_A = R_B.$$

By the end of this procedure, the shuffled variable C will have the *exact* marginal distribution of variable A with the *exact* rank-order of variable B.

Example:

Table 40 shows the shuffling procedure using a hypothetical example. Now, we demonstrate the shuffling procedure for a specific instance of the hypothetical example. Let us try to "Shuffle A by B" for the observations that have the rank-order of 2 (i.e. $R_A = R_B = 2$):

1. Locate two corresponding observations in A and B based on the equality of their rank-orders (here, $R_A = R_B = 2$)
2. Start from the above formula: $C(Row_B(R_B)) = A(Row_A(R_A))$; where $R_A = R_B$
3. Replace the rank-orders R_A and R_B by their values: $C(Row_B(2)) = A(Row_A(2))$
4. Use the corresponding rows/numbers of these rank-orders as indices for the variables C and A: $C(1) = A(7)$
5. Assign the observation of the corresponding index of A to the variable C at the corresponding B-derived index: $C(1) = 2.9$

Notice that to shuffle the whole random variable A by B, we just repeat the above steps for every observation in A.

♦

Table 40. Hypothetical example demonstrating "Shuffle *A* by *B*" for generating a shuffled variable *C*

A	R_A	$Row_A(R_A)$	B	R_B	$Row_B(R_B)$	C	R_C
8.8	8	1	2.3	2	1	2.9	2
4.5	4	2	7.5	7	2	7.3	7
7.3	7	3	8.4	8	3	8.8	8
9.7	9	4	6.8	6	4	6.4	6
3.2	3	5	4.2	4	5	4.5	4
10.5	10	6	3.6	3	6	3.2	3
2.9	2	7	1.1	1	7	1.8	1
5.3	5	8	5.9	5	8	5.3	5
6.4	6	9	10.9	10	9	10.5	10
1.8	1	10	9.9	9	10	9.7	9

R_A, R_B, and R_C represent the rank-order of the random variables A, B and C.

Masking Methods

Now, we present the algorithms of the two groups of masking methods: the NM-EGADP masking methods. While the four methods are very similar, the distinction between the two groups is how the orthogonal noise is generated (a normal noise vs. a normal copula noise). The first method (method 1) is the basis for all four masking methods. No shuffling is used in the basis methods. The other three masking methods use the shuffling at the level of either residuals, perturbed variables, or both. We surround the shuffling steps in the algorithms by frames when listed shortly. We use the word "perturbation" in naming the methods when no shuffling at the level of perturbed variables is involved (regardless of whether residuals are shuffled). When shuffling is done at the level of perturbed variables, we replace the word "perturbation" with the word "shuffling". In addition, when residuals are shuffled before they are added to conditional expectations, we use the phrase "residuals-shuffled" in the beginning of the methods' names. When both levels of shuffling are done, we use both "residuals-shuffled" and "shuffling" in the

Table 41. A list of the four NM-EGADP masking methods and their characteristics.

No	Method	Variable	Noise	Shuffled Residuals	Shuffled Observation
1	**NM-EGADP Perturbation**	**Y**	Normal	-	-
2	**NM-EGADP Shuffling**	**Y_Shf** or $\mathbf{Y}_{shf}$	Normal	-	⎫
3	**Residuals-Shuffled NM-EGADP Perturbation**	**Y_SR**	Normal	⎫	-
4	**Residuals-Shuffled NM-EGADP Shuffling**	**Y_SR_Shf**	Normal	⎫	⎫

method name. Table 41 lists the names of the four methods and their characteristics in terms of the type of the added noise and the involvement level of shuffling.

Relationship-Based NM-EGADP Masking Algorithms

The four masking methods use (un-shuffled/shuffled) normal noise.

Method 1: NM-EGADP Perturbation

1. Regress **X** on **S** by training q Least Squares Support Vector Machines LS-SVM neural networks N_{1j} (one for each individual confidential attribute X_j):

$$N_{1j} = E(X_j \mid \mathbf{S}), \quad j = 1,\mathrm{K}\,,q \tag{C.1}$$

   $\mathbf{N}_1$ (=[N_{11},…, N_{1q}]) learns the set of conditional expectations $E(\mathbf{X}|\mathbf{S})$.

2. Use the set of trained neural networks $\mathbf{N}_1$ to calculate the following:

   a. The set of conditional expected values μ (= $[\mu_1,\mathrm{K}\,,\mu_q]$) of the conditional expectations $E(\mathbf{X}|\mathbf{S})$ evaluated at **S** values where:

$$\mu_j = E(X_j \mid \mathbf{S})|_{\mathbf{S}}, \quad j = 1,\mathrm{K}\,,q \tag{C.2}$$

   b. The residuals set **r** (=[r_1,…, r_q]) where:

$$r_j = X_j - E(X_j \mid \mathbf{S})|_{\mathbf{S}}, \quad j = 1,\mathrm{K}\,,q\,. \tag{C.3}$$

3. Compute the covariance matrix of the first residuals set $\sum_{\mathbf{r}}$, which will be used later to scale another set of orthogonal residuals and makes it have the same covariance matrix $\sum_{\mathbf{r}}$.
4. Generate q independent random variates **V** (=[$V_1, \ldots, V_q$]).
5. Regress **V** on both **S** and **X** by training another set of q LS-SVM neural networks $\mathbf{N}_2$ (=[$N_{21}, \ldots, N_{2q}$]) where:

$$N_{2j} = E(V_j \mid \mathbf{S}, \mathbf{X}), \quad j = 1, \mathrm{K}, q. \tag{C.4}$$

6. Use the set of the trained neural networks $\mathbf{N}_2$ to calculate a second orthogonal residuals set **b** (=[$b_1, \ldots, b_q$]) where:

$$b_j = V_j - E(V_j \mid \mathbf{S}, \mathbf{X})\big|_{\mathbf{S},\mathbf{X}}, \quad j = 1, \mathrm{K}, q. \tag{C.5}$$

7. Compute the covariance matrix of the second residuals set $\sum_{\mathbf{b}}$. Note that although the new set of residuals **b** is orthogonal to **S**, **X** and **r**, the covariance matrix $\sum_{\mathbf{b}}$ is different than $\sum_{\mathbf{r}}$.
8. Generate a new orthogonal set of residuals **e** by scaling the (normalized) set of orthogonal residuals **b** by the covariance matrix $\sum_{\mathbf{r}}$ of the original residuals:

$$\mathbf{e} = (\sum_{\mathbf{r}})^{0.5} (\sum_{\mathbf{b}})^{-0.5} \mathbf{b}. \tag{C.6}$$

9. Calculate the new perturbed attributes **Y**:

$$\mathbf{Y} = \mu + \mathbf{e} = E(\mathbf{X} \mid \mathbf{S}) + \mathbf{e}. \tag{C.7}$$

Method 2: NM-EGADP Shuffling

1. Perform steps 1 to 9 of Method 1
2. Shuffle **X** by **Y** to generate the shuffled attributes $\mathbf{Y}_{\mathbf{shf}}$.

Method 3: Residuals-Shuffled NM-EGADP Perturbation

1. Perform steps 1 to 8 of Method 1
2. Shuffle **r** by **e** to generate the shuffled orthogonal residuals $\mathbf{e}_{\mathbf{shf}}$.
3. Calculate the new perturbed attributes **Y**:

$$\mathbf{Y} = \mu + \mathbf{e}_{\mathbf{shf}} = E(\mathbf{X} \mid \mathbf{S}) + \mathbf{e}_{\mathbf{shf}} . \tag{C.8}$$

Method 4: Residuals-Shuffled NM-EGADP Shuffling

1. Perform steps 1 to 8 of Method 1
2. Shuffle **r** by **e** to generate the shuffled orthogonal residuals $\mathbf{e}_{\mathbf{shf}}$.
3. Calculate the new perturbed attributes **Y**:

$$\mathbf{Y} = \mu + \mathbf{e}_{\mathbf{shf}} = E(\mathbf{X} \mid \mathbf{S}) + \mathbf{e}_{\mathbf{shf}} . \tag{C.9}$$

4. Shuffle **X** by **Y** to generate the shuffled attributes $\mathbf{Y}_{\mathbf{shf}}$.

Appendix D– Extended Results Related to the Motivation Example

Table 42. Sample of the Motivation Example Dataset

No	S1	S2	X1	X2	Y1	Y2	$Y1_{shf}$	$Y2_{shf}$
1	53.12	0	258.57	514.74	271.54	546.83	270.04	546.22
2	57.37	1	211.35	569.33	220.82	603.43	219.80	608.35
3	22.06	0	224.45	609.51	223.10	555.46	221.74	555.42
4	25.46	1	272.98	547.63	262.40	532.68	261.18	533.55
5	51.25	1	292.63	516.33	268.89	517.11	268.45	520.38
6	33.69	0	332.74	459.77	366.92	476.60	365.43	474.31
7	52.63	0	271.99	523.39	288.68	524.81	291.32	525.92
8	55.13	0	268.46	576.32	249.96	557.04	248.90	556.42
9	29.54	1	303.64	481.62	288.66	477.52	291.06	475.90
10	59.44	1	214.64	590.22	208.29	601.59	210.45	605.57
:	**:**	**:**	**:**	**:**	**:**	**:**	**:**	**:**
991	41.35	1	418.87	426.78	396.27	435.37	399.37	437.04
992	34.30	1	342.65	457.91	370.11	448.22	367.94	448.21
993	26.25	0	262.73	531.11	263.65	527.80	262.18	529.54
994	26.28	0	286.67	549.47	277.26	553.15	277.25	552.01
995	50.65	0	268.78	519.65	299.97	502.34	302.79	502.53
996	37.16	0	391.54	413.41	359.38	399.86	357.00	398.41
997	50.72	1	301.59	500.72	296.39	503.60	299.00	504.17
998	28.71	1	296.49	531.81	266.50	512.61	264.34	514.69
999	27.71	0	308.91	537.00	288.56	525.41	290.84	526.75
1000	29.62	1	298.04	469.97	303.20	501.94	306.17	502.46
Range	**39.96**	**1**	**271.58**	**249.31**	**264.83**	**244.17**	**271.58**	**249.31**
Min	**20.01**	**0**	**172.61**	**383.49**	**168.46**	**381.97**	**172.61**	**383.49**
Max	**59.97**	**1**	**444.19**	**632.80**	**433.29**	**626.14**	**444.19**	**632.80**
Mean	**40.2640**	**.4680**	**296.7751**	**504.6919**	**296.7753**	**504.6918**	**296.7751**	**504.6919**
STD	**11.88385**	**.49922**	**59.29870**	**59.05412**	**59.29040**	**58.95759**	**59.29870**	**59.05412**
VAR	**141.226**	**.249**	**3516.335**	**3487.389**	**3515.351**	**3475.997**	**3516.335**	**3487.389**

S: non-confidential attributes
X: original confidential attributes
Y: masked (perturbed) confidential attributes
Yshf: masked (shuffled) confidential attributes

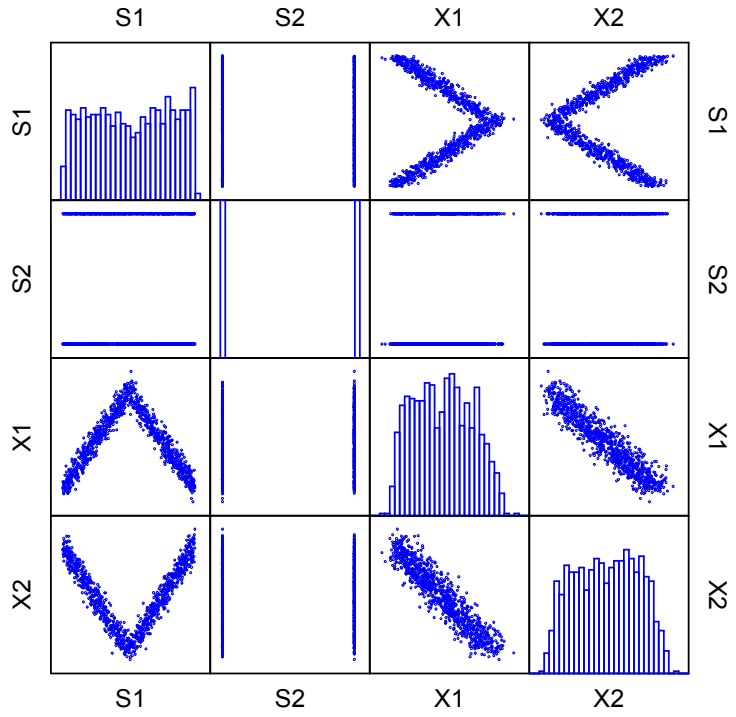


Figure 41. Motivation Example (ME.L) – Original dataset

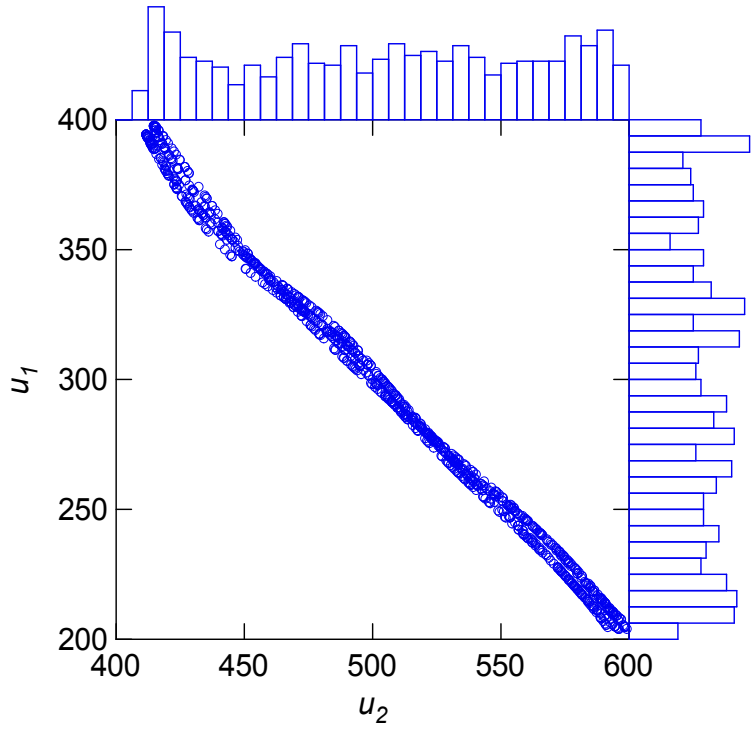


Figure 42. Motivation Example – Predicted values $E(X_1|S)$|s vs. $E(X_2|S)$|s

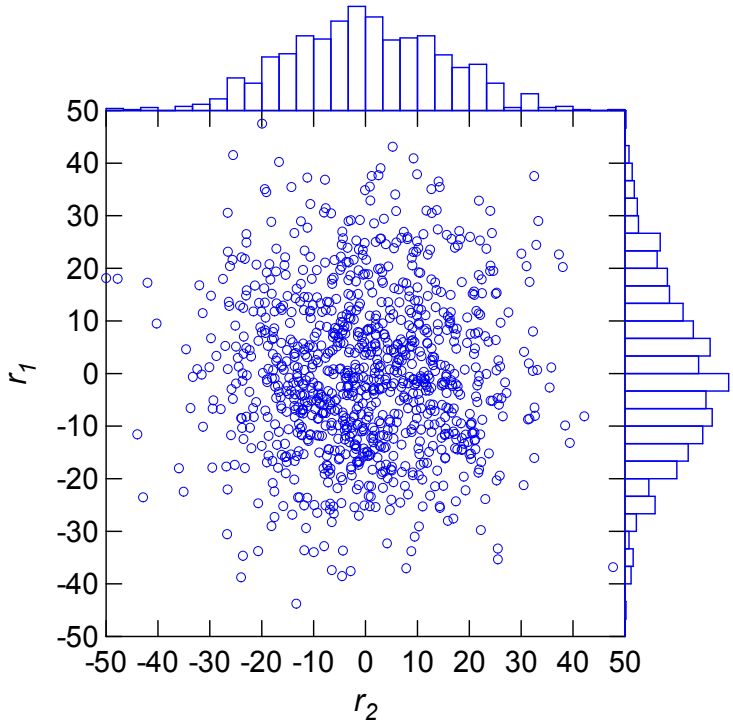


Figure 43. Motivation Example – Residuals r_1 vs. r_2

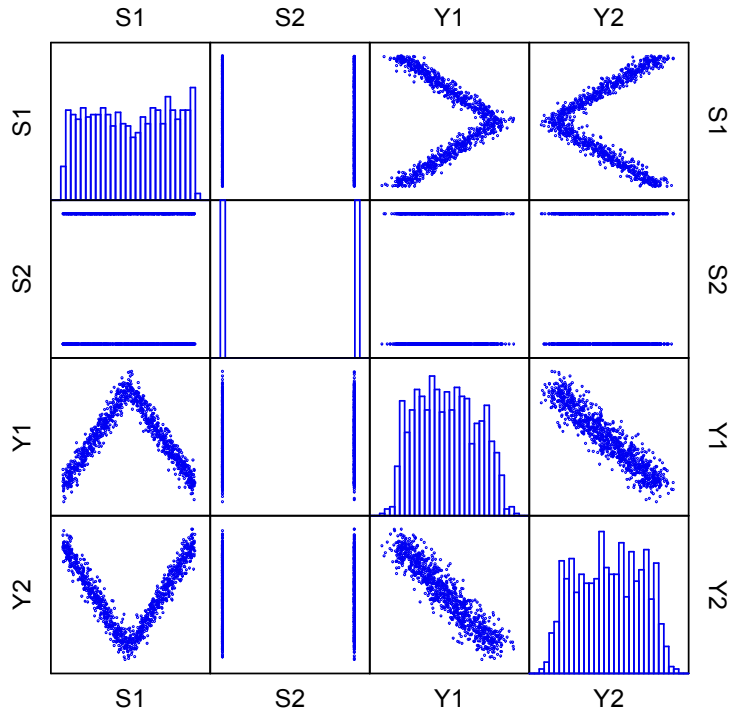


Figure 44. Motivation Example – The NM-EGADP Perturbation (masking method 1) masked dataset

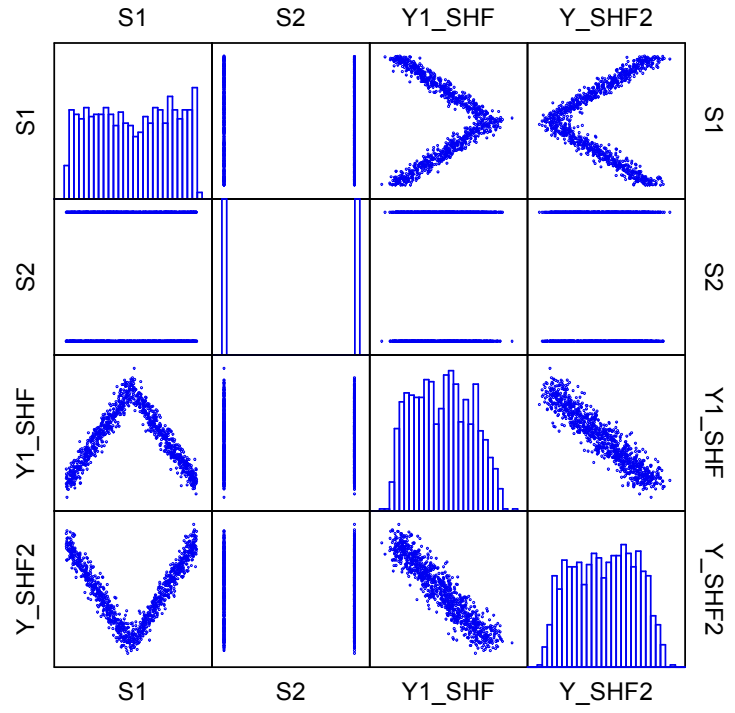


Figure 45. Motivation Example – The NM-EGADP Shuffling (masking method 2) masked dataset

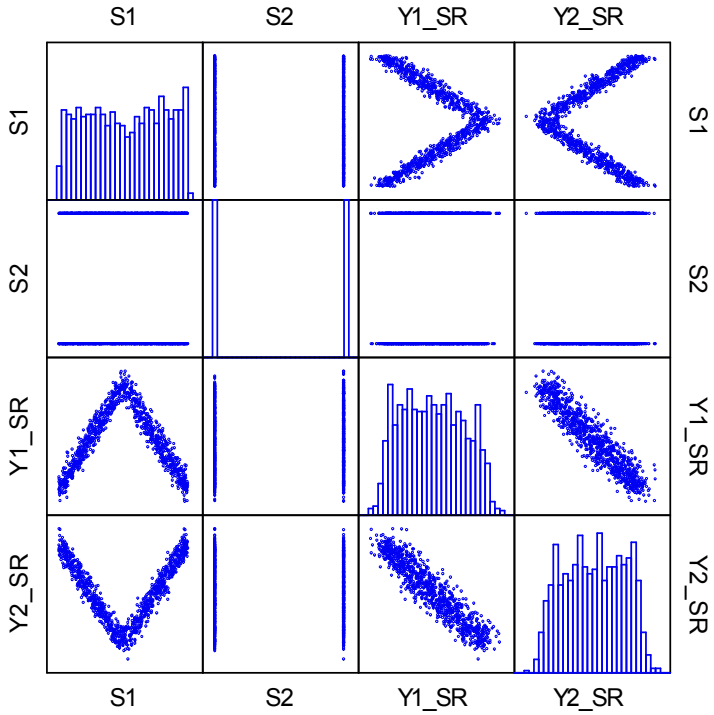


Figure 46. Motivation Example (ME.L) – Masked using masking method 3

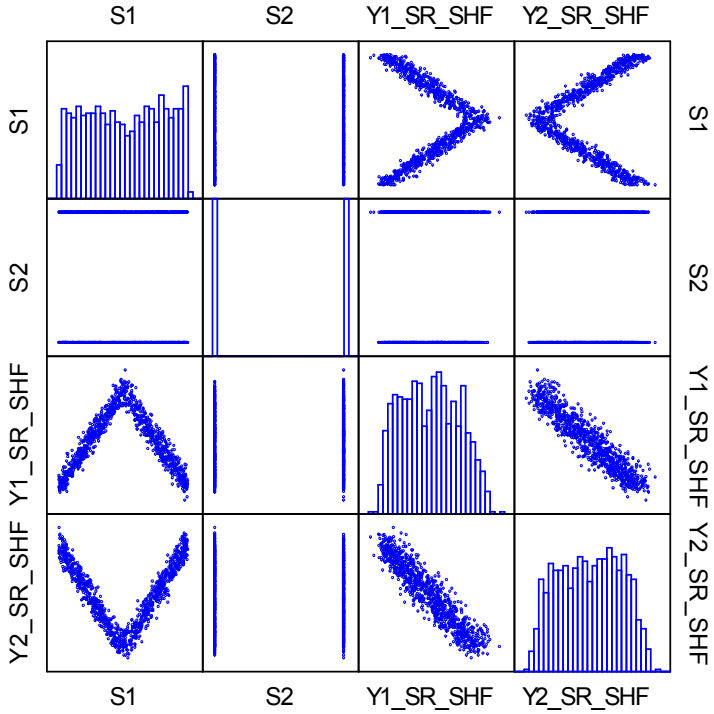


Figure 47. Motivation Example (ME.L) – Masked using masking method 4

Note:

The visual similarity between the figures of the masked datasets (both perturbated and shuffled) and the figure of the original dataset suggests that NM-EGADP masking procedures work fine in term of data utility.

Table 43. Motivation Example - Original dataset Pearson correlations

Correlations

		s1	s2	x1	x2
s1	Pearson Correlation	1	-.010	-.026	.052
	Sig. (2-tailed)		.759	.409	.100
	N	1000	1000	1000	1000
s2	Pearson Correlation	-.010	1	.032	-.031
	Sig. (2-tailed)	.759		.308	.320
	N	1000	1000	1000	1000
x1	Pearson Correlation	-.026	.032	1	-.933
	Sig. (2-tailed)	.409	.308		.000
	N	1000	1000	1000	1000
x2	Pearson Correlation	.052	-.031	-.933	1
	Sig. (2-tailed)	.100	.320	.000	
	N	1000	1000	1000	1000

Table 44. Motivation Example - NM-EGADP perturbed dataset Pearson correlations

Correlations

		s1	s2	y1	y2
s1	Pearson Correlation	1	-.010	-.026	.052
	Sig. (2-tailed)		.759	.409	.100
	N	1000	1000	1000	1000
s2	Pearson Correlation	-.010	1	.032	-.031
	Sig. (2-tailed)	.759		.308	.320
	N	1000	1000	1000	1000
y1	Pearson Correlation	-.026	.032	1	-.933
	Sig. (2-tailed)	.409	.308		.000
	N	1000	1000	1000	1000
y2	Pearson Correlation	.052	-.031	-.933	1
	Sig. (2-tailed)	.100	.320	.000	
	N	1000	1000	1000	1000

Table 45. Motivation Example - NM-EGADP shuffled dataset Pearson correlations

Correlations

		s1	s2	y1_shf	y_shf2
s1	Pearson Correlation	1	-.010	-.027	.053
	Sig. (2-tailed)		.759	.391	.096
	N	1000	1000	1000	1000
s2	Pearson Correlation	-.010	1	.032	-.032
	Sig. (2-tailed)	.759		.307	.316
	N	1000	1000	1000	1000
y1_shf	Pearson Correlation	-.027	.032	1	-.931
	Sig. (2-tailed)	.391	.307		.000
	N	1000	1000	1000	1000
y_shf2	Pearson Correlation	.053	-.032	-.931	1
	Sig. (2-tailed)	.096	.316	.000	
	N	1000	1000	1000	1000

Note:

The similarity between the correlation matrices of the masked datasets (using the NM-EGADP perturbation and NM-EGADP shuffling) and the correlation matrix of the original dataset suggests that NM-EGADP masking procedures not only maintain non-monotonic relationships but also maintain *linear* relationships (another extra advantage in term of data utility).

Table 46. Motivation Example - Original dataset rank-order (Spearman) correlations

Correlations

			s1	s2	x1	x2
Spearman's rho	s1	Correlation Coefficient	1.000	-.008	-.031	.054
		Sig. (2-tailed)	.	.794	.321	.090
		N	1000	1000	1000	1000
	s2	Correlation Coefficient	-.008	1.000	.030	-.033
		Sig. (2-tailed)	.794	.	.342	.296
		N	1000	1000	1000	1000
	x1	Correlation Coefficient	-.031	.030	1.000	-.938
		Sig. (2-tailed)	.321	.342	.	.000
		N	1000	1000	1000	1000
	x2	Correlation Coefficient	.054	-.033	-.938	1.000
		Sig. (2-tailed)	.090	.296	.000	.
		N	1000	1000	1000	1000

Table 47. Motivation Example - NM-EGADP perturbed dataset rank-order (Spearman) correlations

Correlations

			s1	s2	y1	y2
Spearman's rho	s1	Correlation Coefficient	1.000	-.008	-.033	.056
		Sig. (2-tailed)	.	.794	.298	.078
		N	1000	1000	1000	1000
	s2	Correlation Coefficient	-.008	1.000	.031	-.034
		Sig. (2-tailed)	.794	.	.332	.289
		N	1000	1000	1000	1000
	y1	Correlation Coefficient	-.033	.031	1.000	-.938
		Sig. (2-tailed)	.298	.332	.	.000
		N	1000	1000	1000	1000
	y2	Correlation Coefficient	.056	-.034	-.938	1.000
		Sig. (2-tailed)	.078	.289	.000	.
		N	1000	1000	1000	1000

Table 48. Motivation Example - NM-EGADP shuffled dataset rank-order (Spearman) correlations

Correlations

			s1	s2	y1_shf	y_shf2
Spearman's rho	s1	Correlation Coefficient	1.000	-.008	-.033	.056
		Sig. (2-tailed)	.	.794	.298	.078
		N	1000	1000	1000	1000
	s2	Correlation Coefficient	-.008	1.000	.031	-.034
		Sig. (2-tailed)	.794	.	.332	.289
		N	1000	1000	1000	1000
	y1_shf	Correlation Coefficient	-.033	.031	1.000	-.938
		Sig. (2-tailed)	.298	.332	.	.000
		N	1000	1000	1000	1000
	y_shf2	Correlation Coefficient	.056	-.034	-.938	1.000
		Sig. (2-tailed)	.078	.289	.000	.
		N	1000	1000	1000	1000

Note:

The similarity between the rank-order correlation matrices of the masked datasets (the NM-EGADP perturbation and NM-EGADP shuffling) and the rank-order correlation matrix of the original dataset suggests that NM-EGADP masking procedures not only maintain non-monotonic relationships but also maintain *monotonic nonlinear* relationships (another extra advantage in term of data utility).

Fitting a Wrong Regression Model: Linear Regression Case

Note:

We know that the relationship between X_1 and S_1 is non-monotonic. Assume the data analyst fits a wrong model (a linear regression model). If he fits the wrong model to the masked datasets and masking procedures work well, he should get similar results.

Table 49. Original dataset - Fitting a wrong regression model (Linear Regression Case: $X_1|S_1$)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	302.029	6.628		45.566	.000
	s1	-.130	.158	-.026	-.826	.409

a. Dependent Variable: x1

Table 50. NM-EGADP perturbed dataset - Fitting a wrong regression model (Linear Regression Case: $X_1|S_1$)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	302.025	6.628		45.571	.000
	s1	-.130	.158	-.026	-.826	.409

a. Dependent Variable: y1

Table 51. NM-EGADP shuffled dataset - Fitting a wrong regression model (Linear Regression Case: $X_1|S_1$)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	302.229	6.628		45.597	.000
	s1	-.135	.158	-.027	-.858	.391

a. Dependent Variable: y1_shf

Motivation Example – Data Utility Assessment by Fitting Parametric Regression Models for Data Involving Non-Monotonic Relationships

Goal:

To assess the *parametric-similarity* between the parameters of models estimated from the motivation example original unmasked dataset and the parameters of models estimated from the motivation example masked (perturbed or shuffled) datasets.

Notes:

- Nonlinear regression facility in SPSS is used to estimate the parameters of the nonlinear regression models.
- Parameter *c* always stands for the constant term (if it appears in the model).
- We choose the value 0.0001 as a starting value for all parameters (*a*, *b*, and *c*).

Table 52. Motivation Example - Data Utility assessment by fitting nonlinear Parametric Regression Models: (X|S), (Y|S), and (Y_{shf}|S)

Nonlinear Regression Fitted Model	Model Parameters Estimated From								
	Original Dataset DV: X_1 - IV: S_1S_2			Perturbed Dataset DV: Y_1 - IV: S_1S_2			Shuffled Dataset DV: Y_1_shf - IV: S_1S_2		
	a	***B***	***c***	***a***	***b***	***c***	***a***	***b***	***c***
$aS_1 + bS_2$	6.266	47.287		6.266	47.287		6.265	47.321	
$aS_1 + bS_2 + c$	-0.129	3.803	300.187	-0.129	3.803	300.182	-0.134	3.808	300.384
$aS_1^2 + bS_2^2$	0.104	117.107		0.104	117.107		0.104	117.150	
$aS_1^2 + bS_2^2 + c$	-0.009	3.585	310.551	-0.009	3.585	310.548	-0.009	3.590	310.655
$aS_1^3 + bS_2^3$	0.002	167.810		0.002	167.809		0.002	167.852	
$aS_1^3 + bS_2^3 + c$	0.000	3.326	314.220	0.000	3.326	314.218	0.000	3.330	314.291
$aS_1^{10} + bS_2^{10}$	0.000	264.570		0.000	264.568		0.000	264.587	
$aS_1^{10} + bS_2^{10} +$	0.000	2.948	313.929	0.000	2.948	313.928	0.000	2.953	313.944

Table 53. Motivation Example - Data Utility assessment by fitting nonlinear Parametric Regression Models: ($X_1|X_2$), ($Y_1|Y_2$), and ($Y_{1\_shf}|Y_{2\_shf}$)

Nonlinear Regression Fitted Model	Model Parameters Estimated From					
	Original Dataset DV: X_1 - IV: X_2		Perturbed Dataset DV: Y_1 - IV: Y_2		Shuffled Dataset DV: Y_1_shf - IV: Y_2_shf	
	a	*C*	*a*	*c*	*a*	*c*
aX_2	0.567		0.568		0.567	
$aX_2 + c$	-0.937	769.554	-0.938	770.091	-0.935	768.654
aX_2^2	0.001		0.001		0.001	
$aX_2^2 + c$	-0.001	535.702	-0.001	536.116	-0.001	535.375
aX_2^3	0.000		0.000		0.000	
$aX_2^3 + c$	0.000	457.432	0.000	457.837	0.000	457.292
aX_2^{10}	0.000		0.000		0.000	
$aX_2^{10} + c$	0.000	344.865	0.000	345.752	0.000	344.940

Table 54. Motivation Example - Data Utility assessment by fitting nonlinear Parametric Regression Models: ($X_1|SX_2$), ($Y_1|SY_2$), and ($Y_{1\_shf}|SY_{2\_shf}$)

Nonlinear Regression Fitted Model	Model Parameters Estimated From											
	Original Dataset DV: X_1 - IV: S_1S_2				Perturbed Dataset DV: Y_1 - IV: S_1S_2				Shuffled Dataset DV: Y_1_shf - IV: S_1S_2			
	a	*b*	*c*	*d*	*a*	*b*	*c*	*d*	*a*	*b*	*c*	*d*
$aS_1 + bS_2 + dX_2$	2.376	22.783		0.359	2.371	22.745		0.360	2.368	22.788		0.360
$aS_1 + bS_2 + dX_2$	0.112	0.366	765.412	-0.938	0.112	0.372	765.937	-0.939	0.109	0.353	764.620	-0.936
$aS_1^2 + bS_2^2 + dX_2^2$	0.029	52.952		0.001	0.029	52.842		0.001	0.029	52.941		0.001
$aS_1^2 + bS_2^2 + dX_2^2$	0.001	0.128	534.451	-0.001	0.001	0.151	534.866	-0.001	0.001	0.138	534.185	-0.001
$aS_1^3 + bS_2^3 + dX_2^3$	0.000	87.252		0.000	0.000	87.064		0.000	0.000	87.219		0.000
$aS_1^3 + bS_2^3 + dX_2^3$	0.000	-0.061	457.009	0.000	0.000	-0.024	457.406	0.000	0.000	-0.030	456.896	0.000
$aS_1^{10} + bS_2^{10} + dX$	0.000	230.525		0.000	0.000	229.776		0.000	0.000	230.492		0.000
$aS_1^{10} + bS_2^{10} + dX + c$	0.000	-0.012	344.976	0.000	0.000	0.083	345.814	0.000	0.000	0.138	345.011	0.000

Possible Usefulness for other Tasks: Classification using Discriminant Analysis

(Part of my research Agenda after graduation)

Table 55. Classification: Original Dataset: IV: S_2, DV: S_1 X_1 X_2

s2 * Predicted Group for Analysis 1 Crosstabulation

			Predicted Group for Analysis 1		
			.00	1.00	Total
s2	.00	Count	282	250	532
		% of Total	28.2%	25.0%	53.2%
	1.00	Count	242	226	468
		% of Total	24.2%	22.6%	46.8%
Total		Count	524	476	1000
		% of Total	52.4%	47.6%	100.0%

Table 56. Classification: NM-EGADP Perturbed Dataset: IV: S_2, DV: S_1 Y_1 Y_2

s2 * Predicted Group for Analysis 1 Crosstabulation

			Predicted Group for Analysis 1		
			.00	1.00	Total
s2	.00	Count	281	251	532
		% of Total	28.1%	25.1%	53.2%
	1.00	Count	238	230	468
		% of Total	23.8%	23.0%	46.8%
Total		Count	519	481	1000
		% of Total	51.9%	48.1%	100.0%

Table 57. Classification: NM-EGADP Shuffled Dataset: IV: S_2, DV: S_1 Y_{1shf} Y_{2shf}

s2 * Predicted Group for Analysis 1 Crosstabulation

			Predicted Group for Analysis 1		
			.00	1.00	Total
s2	.00	Count	275	257	532
		% of Total	27.5%	25.7%	53.2%
	1.00	Count	230	238	468
		% of Total	23.0%	23.8%	46.8%
Total		Count	505	495	1000
		% of Total	50.5%	49.5%	100.0%

Appendix E – Graphical Pilot Study – Comparisons for PPE Masking Methods

In this appendix, we present some *graphical* evidence that the NM-EGADP shuffling procedure work well and preserve different types of relationships including non-monotonic ones while EGADP (shuffling) doesn't preserve nonlinear relationships and (C-GADP based) data shuffling only maintains monotonic relationships.

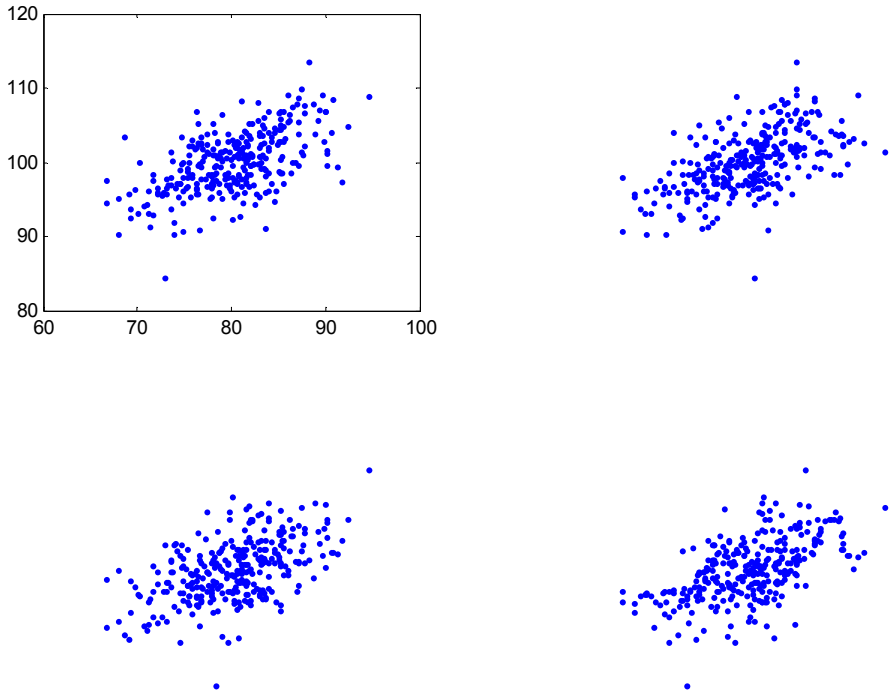


Figure 48. Graphical Pilot Study: Linear relationships (bivariate normal dataset)

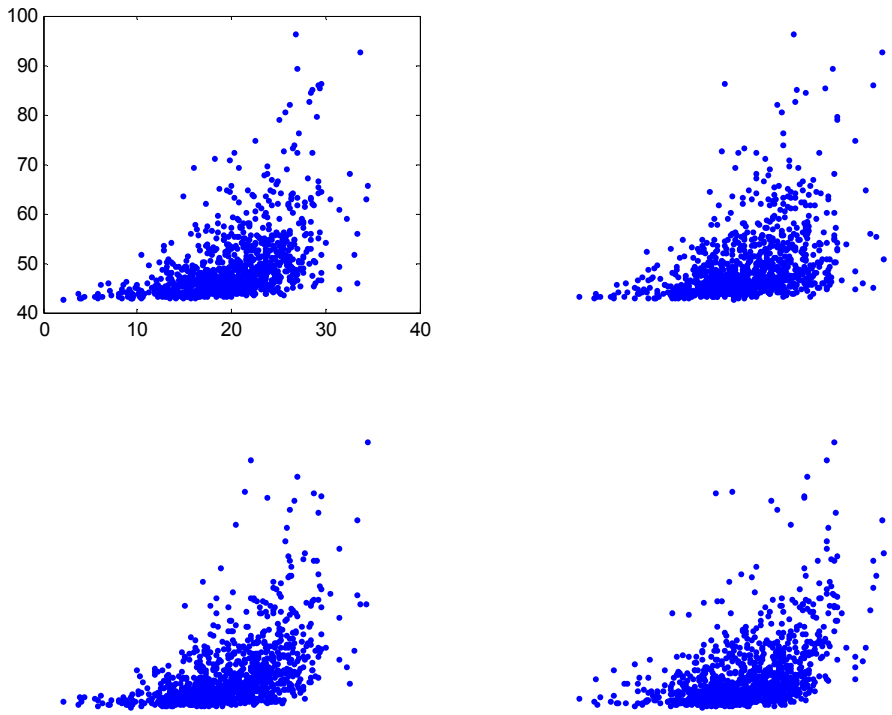


Figure 49. Graphical Pilot Study: Monotonic nonlinear relationships I

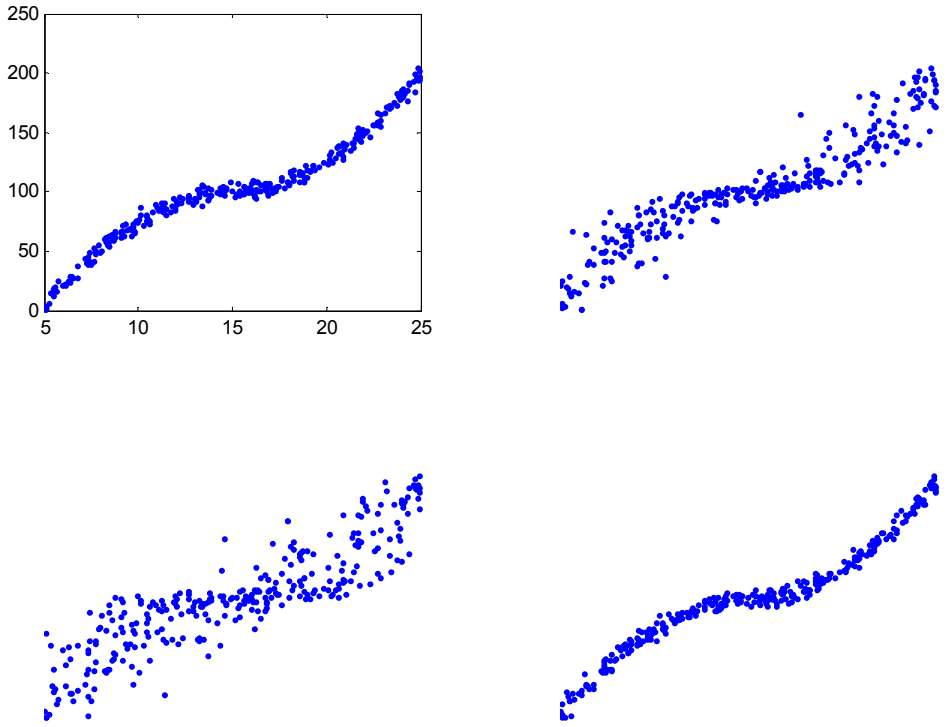


Figure 50. Graphical Pilot Study: Monotonic nonlinear relationships II

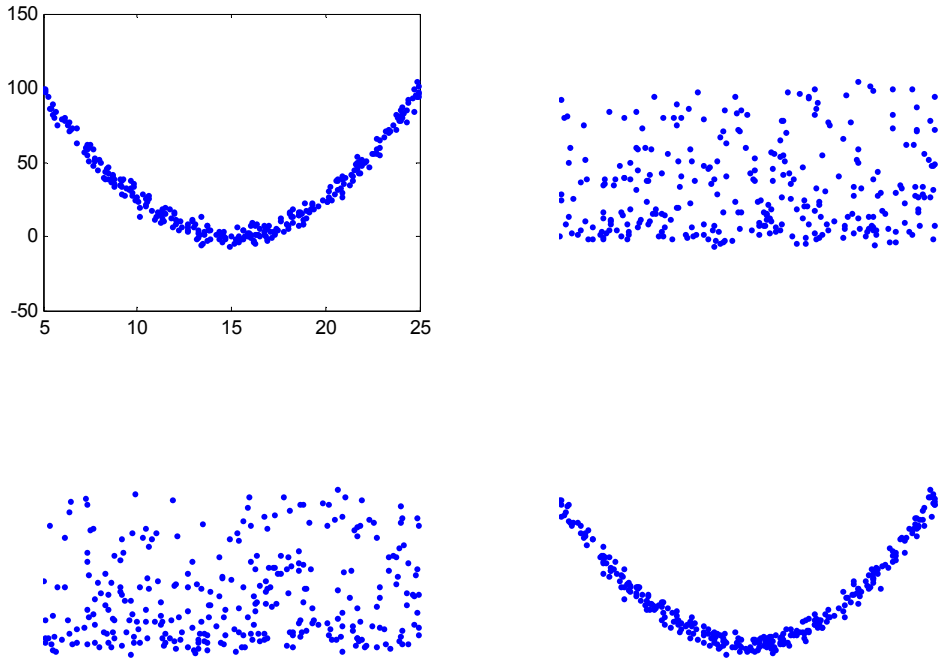


Figure 51. Graphical Pilot Study: Non-Monotonic relationships (U-shape data)

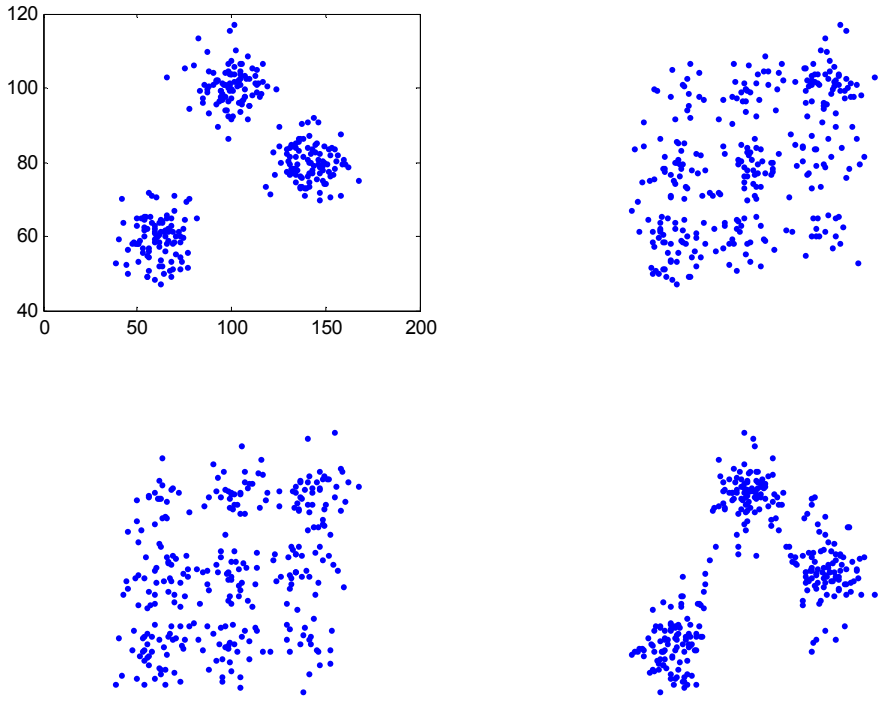


Figure 52. Graphical Pilot Study: Non-Monotonic relationships (3-cluster data)

Appendix F – Dataset: NM.01 (Check Mark)

sdfd

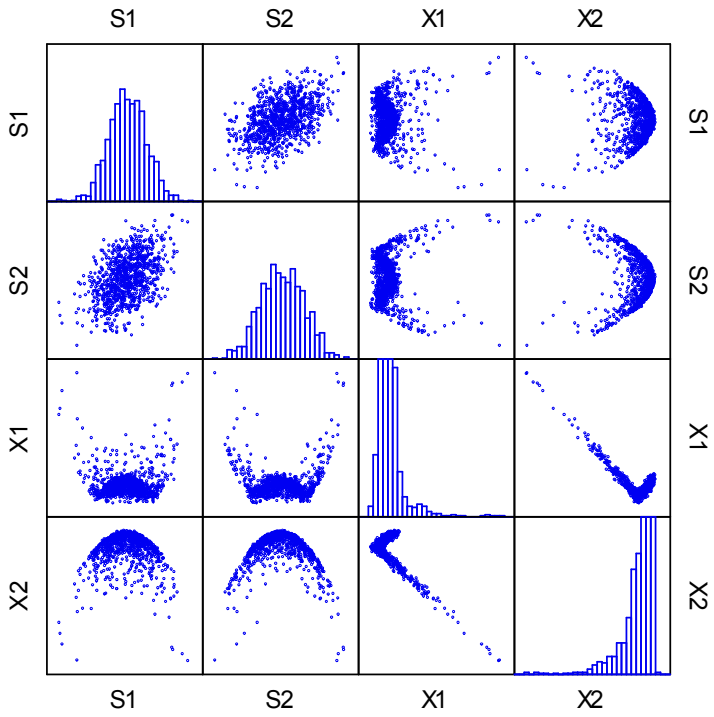


Figure 53. NM.01 Dataset

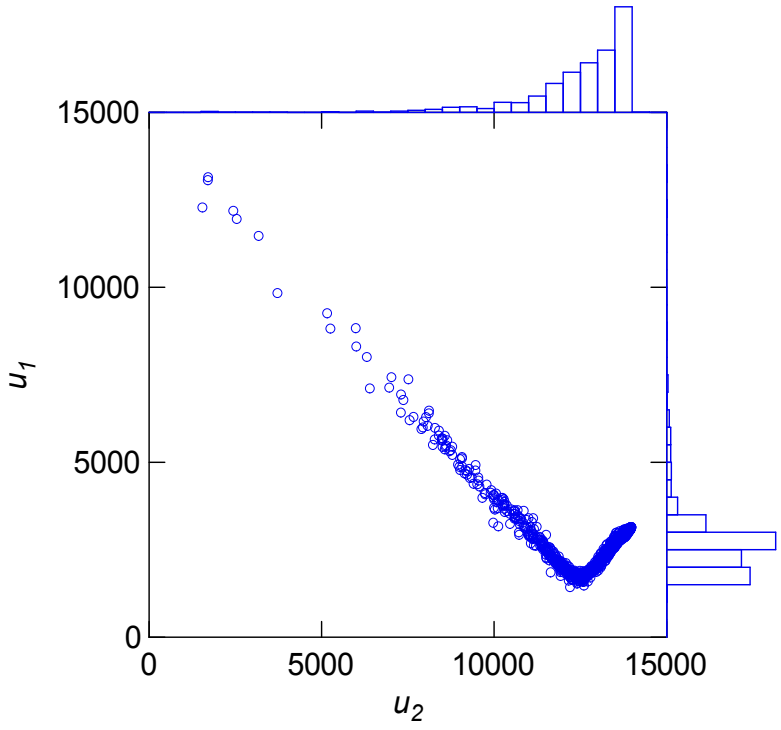


Figure 54. NM.01 Dataset – Predicted values $E(X_1|S)|s$ vs. $E(X_2|S)|s$

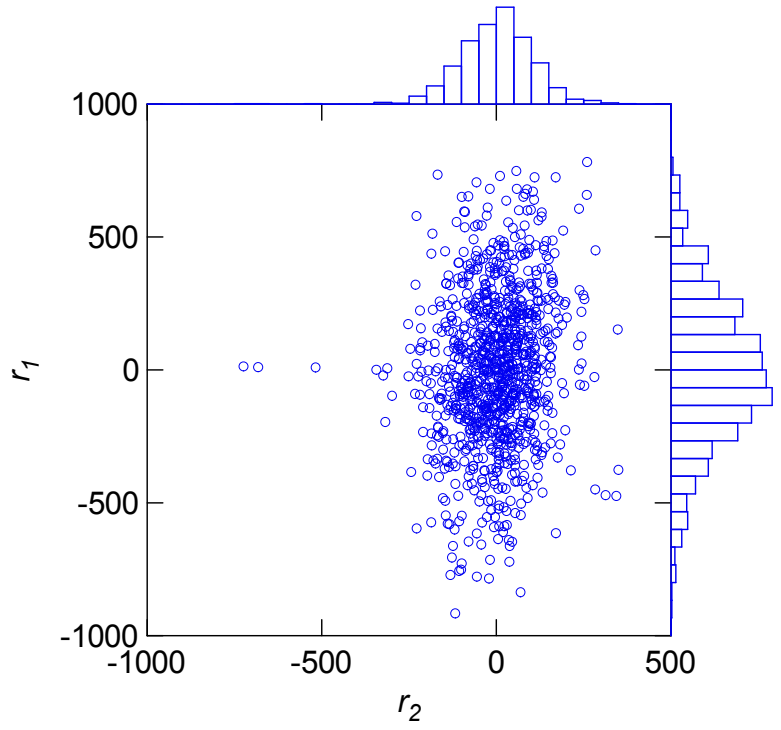


Figure 55. NM.01 Dataset – Residuals r_1 vs. r_2

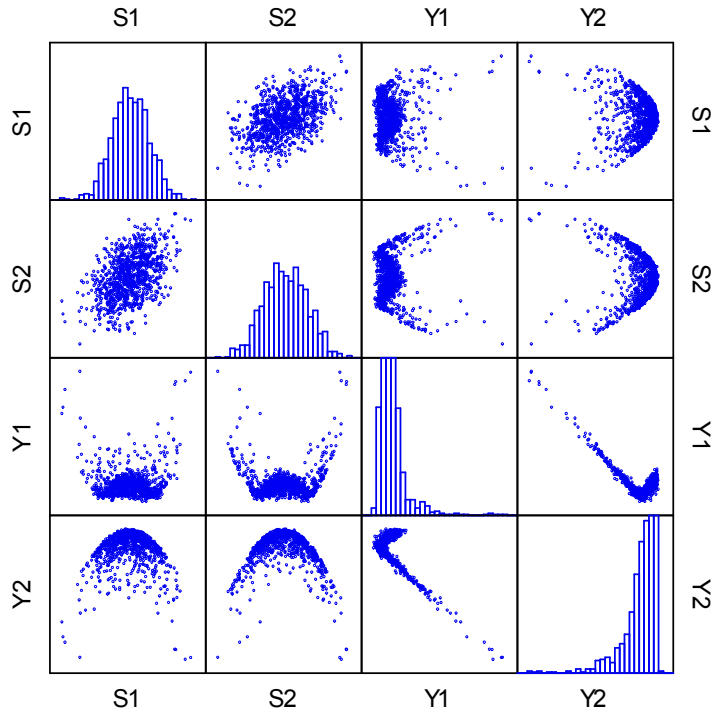


Figure 56. NM.01 Dataset – Masked using masking method 1

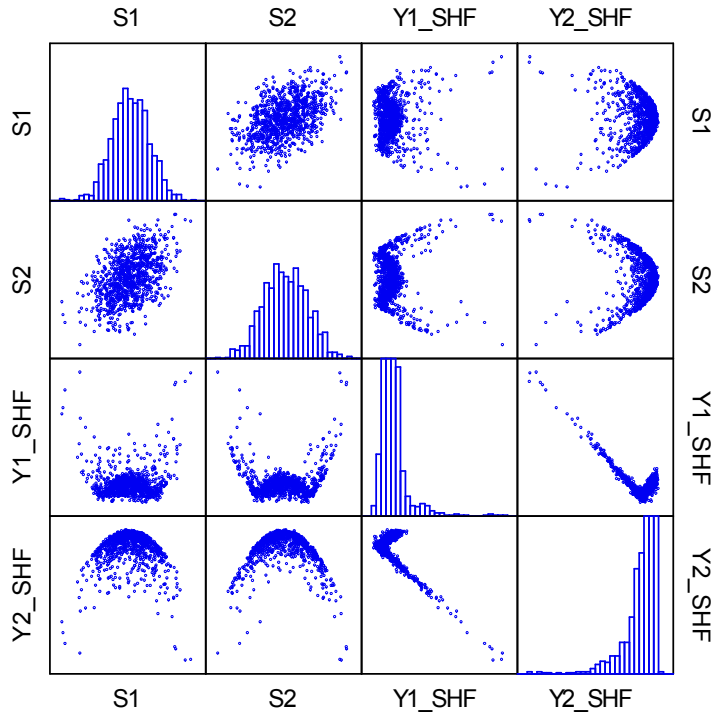


Figure 57. NM.01 Dataset – Masked using masking method 2

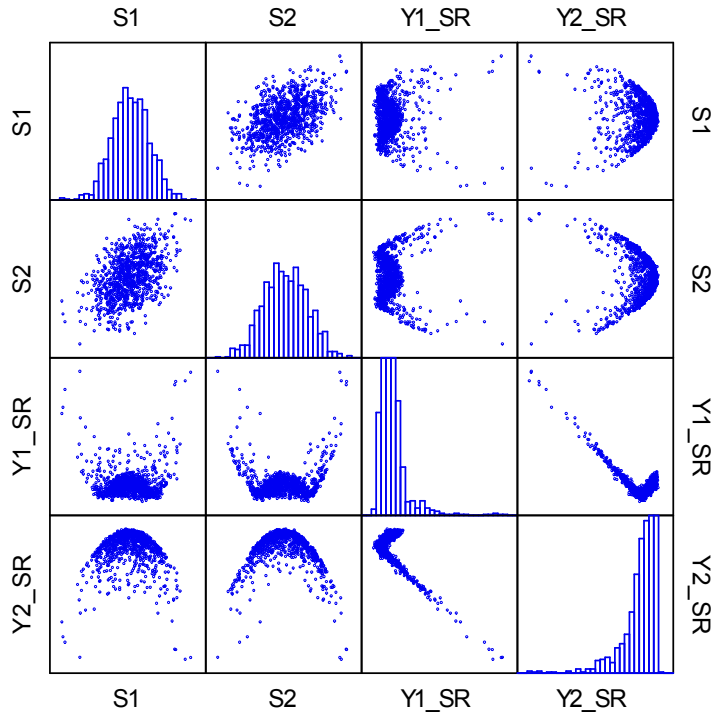


Figure 58. NM.01 Dataset – Masked using masking method 3

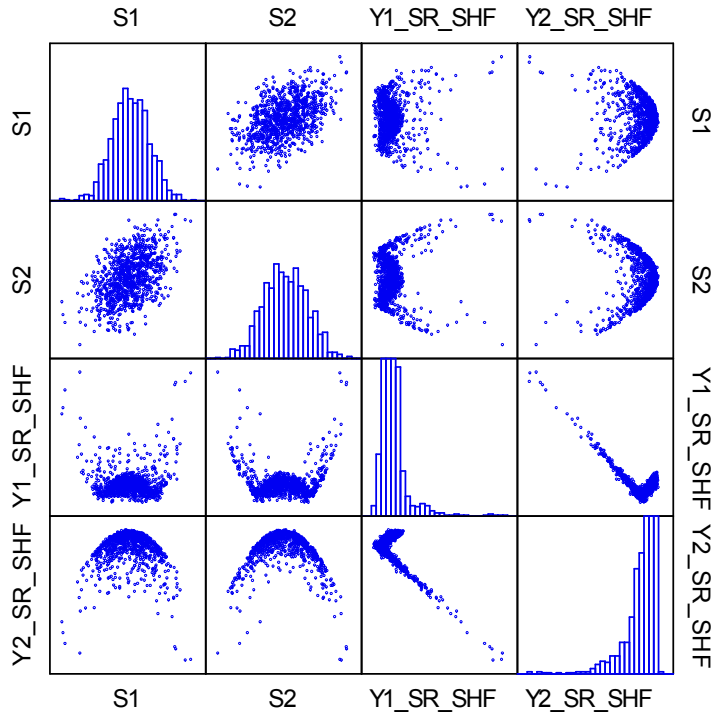


Figure 59. NM.01 Dataset – Masked using masking method 4

Appendix G – Dataset: NM.02

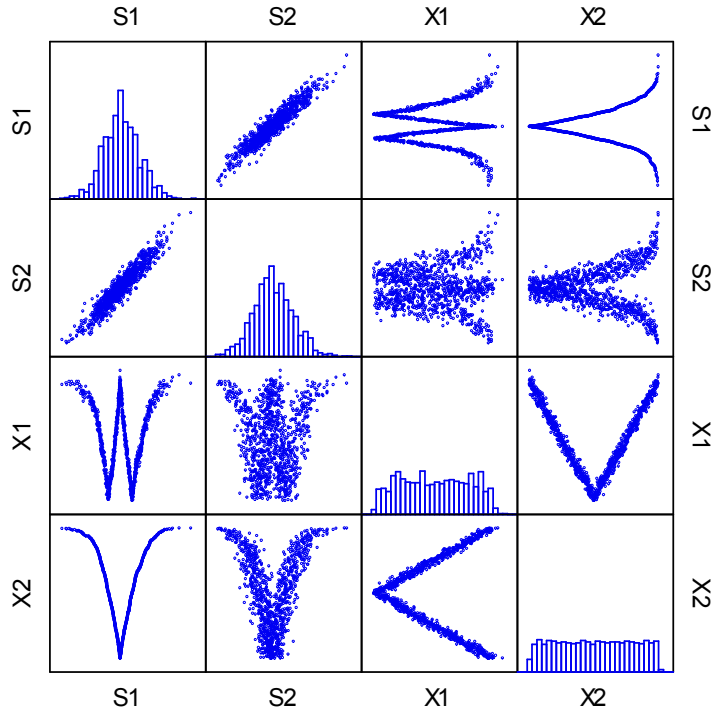


Figure 60. NM.02 Dataset

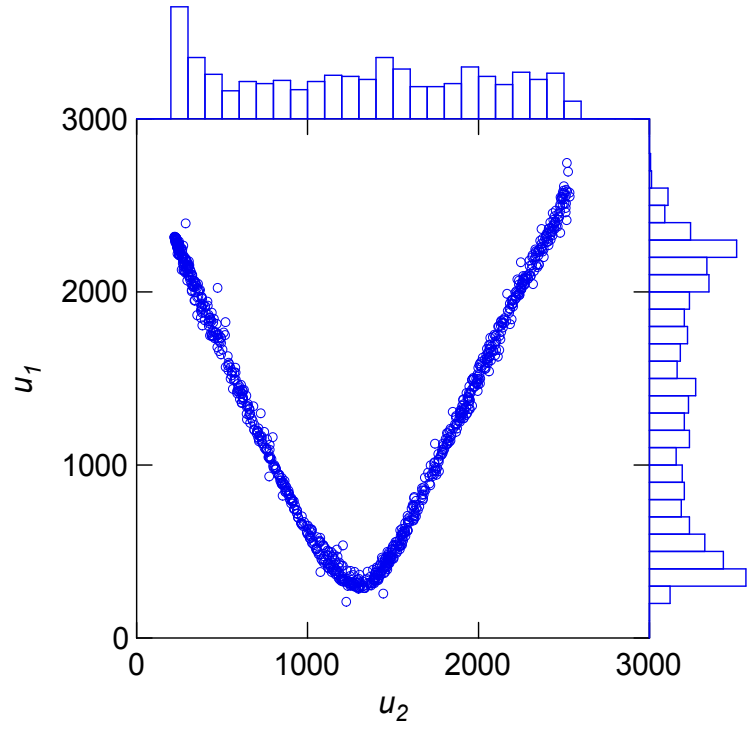


Figure 61. NM.02 Dataset – Predicted values $E(X_1|S)|s$ vs. $E(X_2|S)|s$

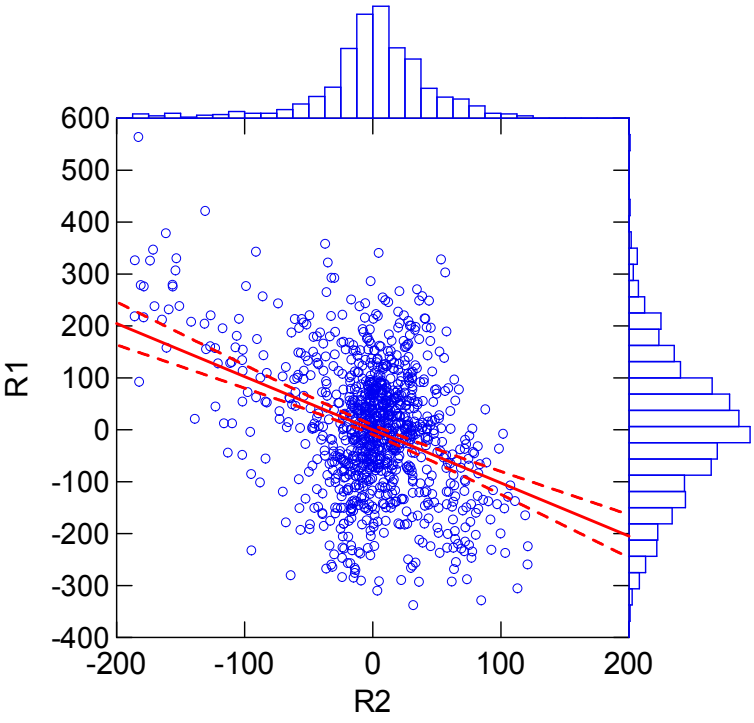


Figure 62. NM.02 Dataset – Residuals r_1 vs. r_2

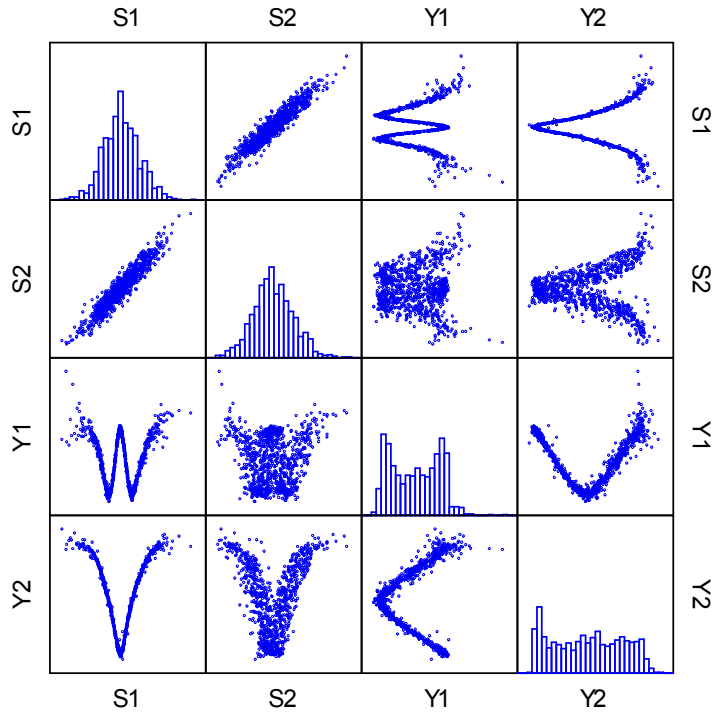


Figure 63. NM.02 Dataset – Masked using masking method 1

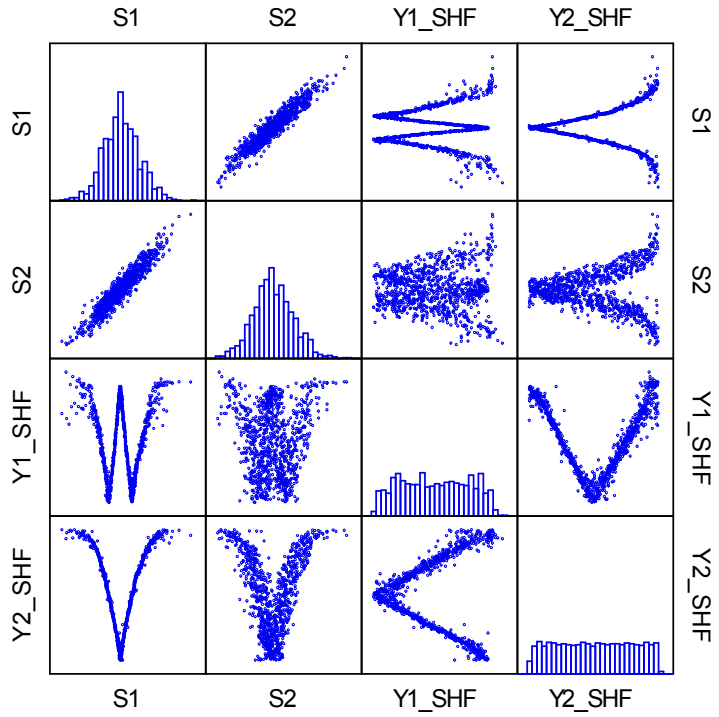


Figure 64. NM.02 Dataset – Masked using masking method 2

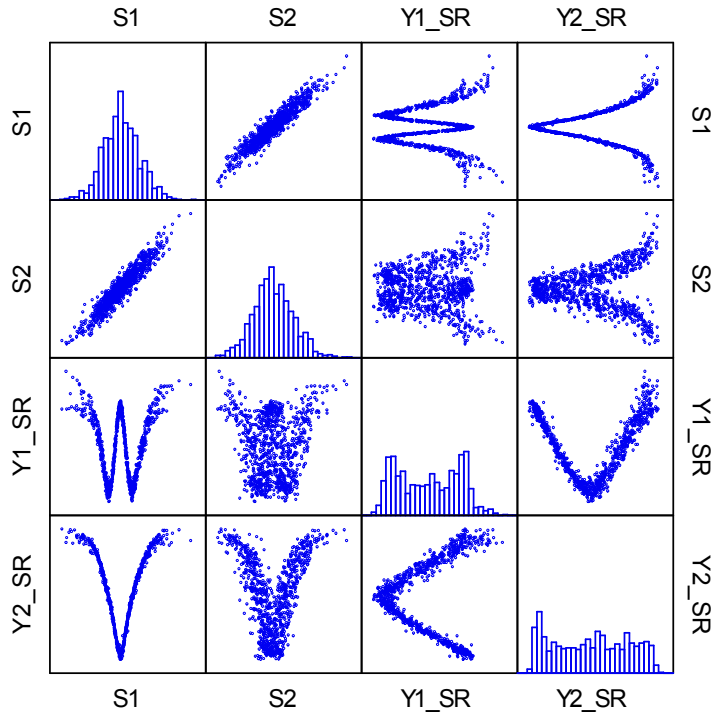


Figure 65. NM.02 Dataset – Masked using masking method 3

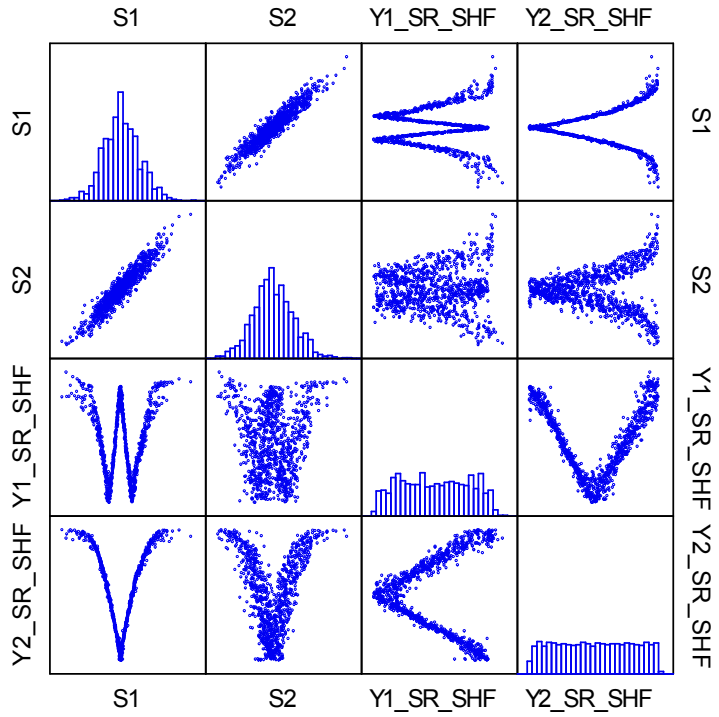


Figure 66. NM.02 Dataset – Masked using masking method 4

Appendix H – Dataset: NM.03

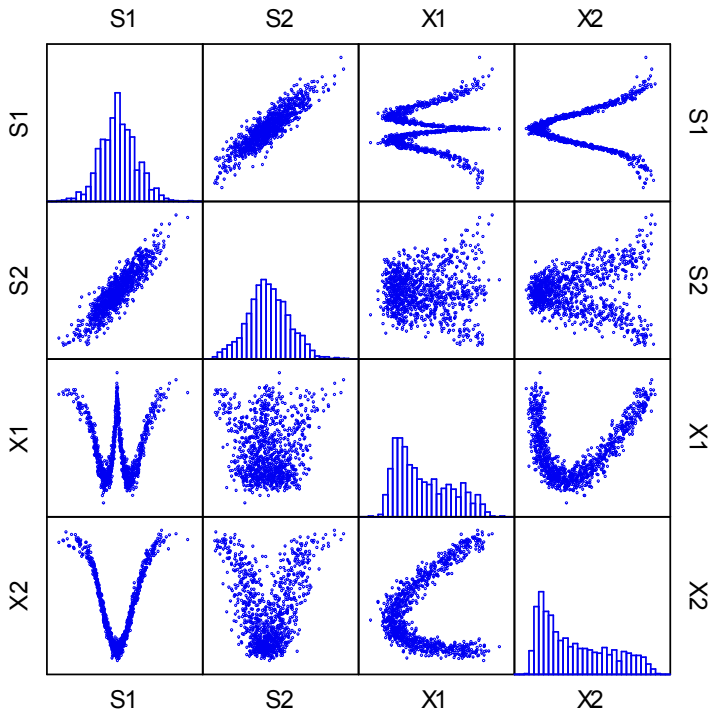


Figure 67. NM.03 Dataset

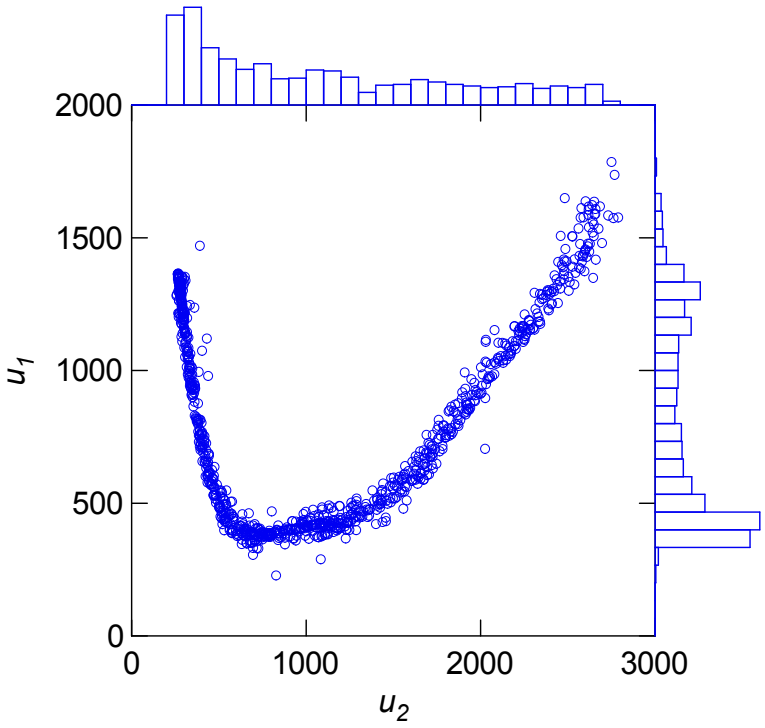


Figure 68. NM.03 Dataset – Predicted values $E(X_1|S)$|s vs. $E(X_2|S)$|s

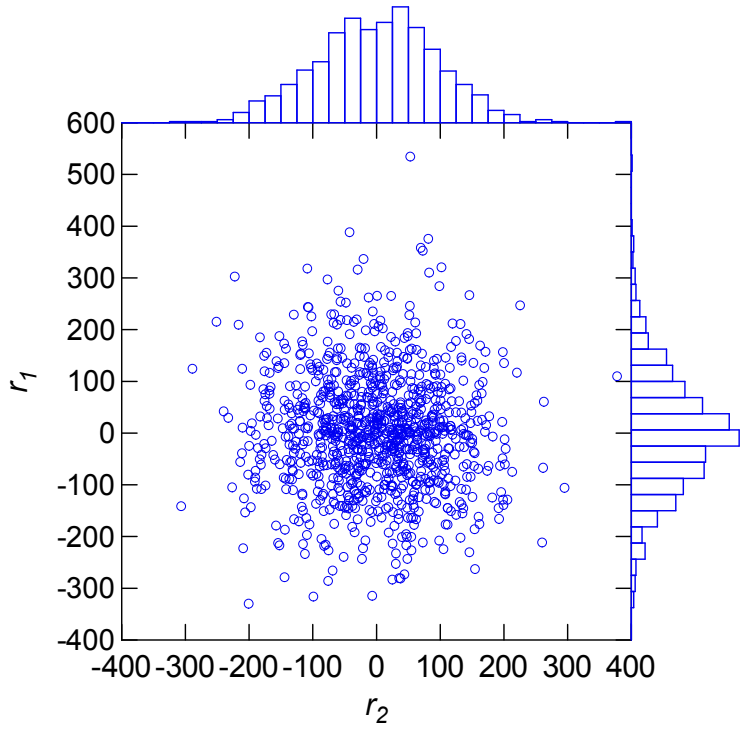


Figure 69. NM.03 Dataset – Residuals r_1 vs. r_2

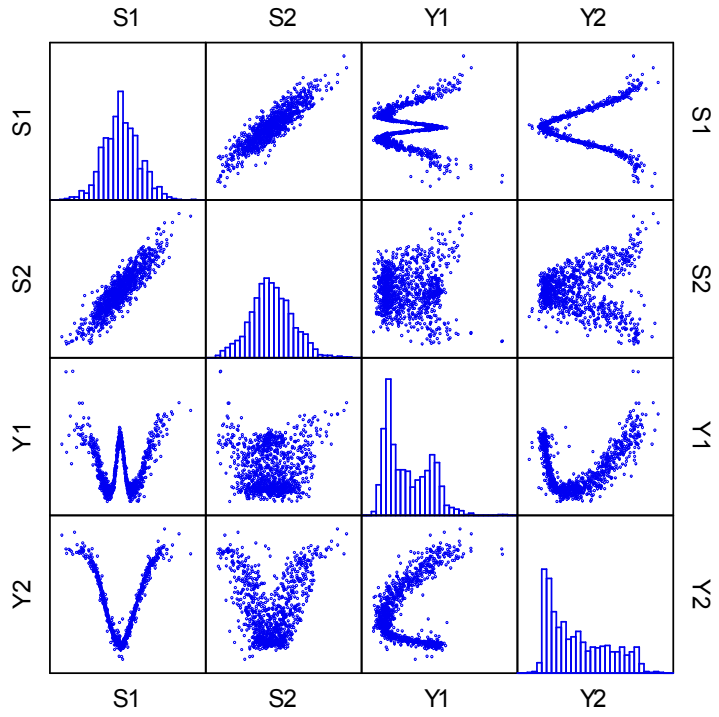


Figure 70. NM.03 Dataset – Masked using masking method 1

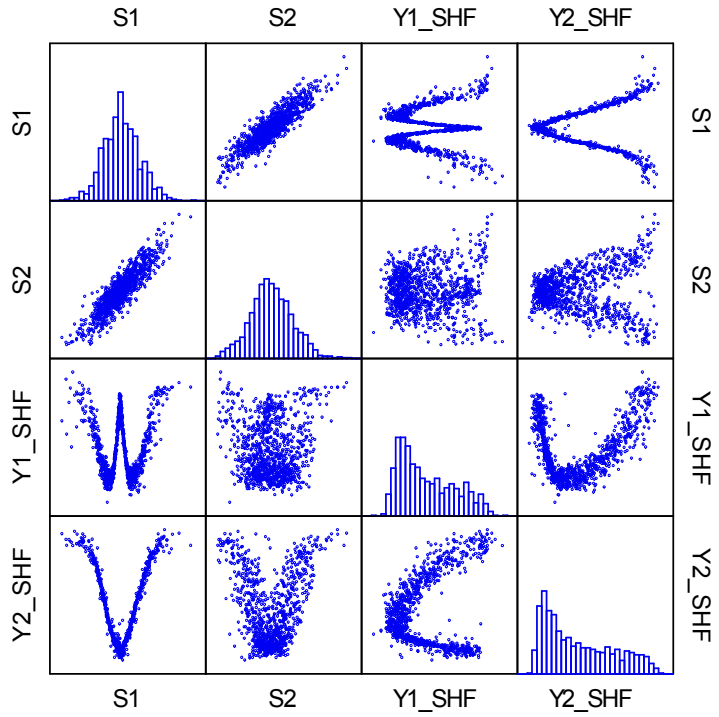


Figure 71. NM.03 Dataset – Masked using masking method 2

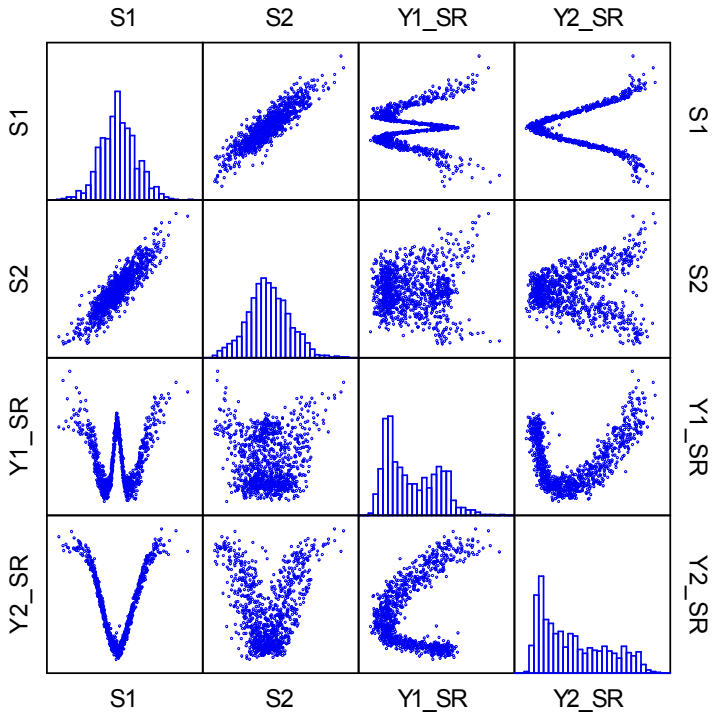


Figure 72. NM.03 Dataset – Masked using masking method 3

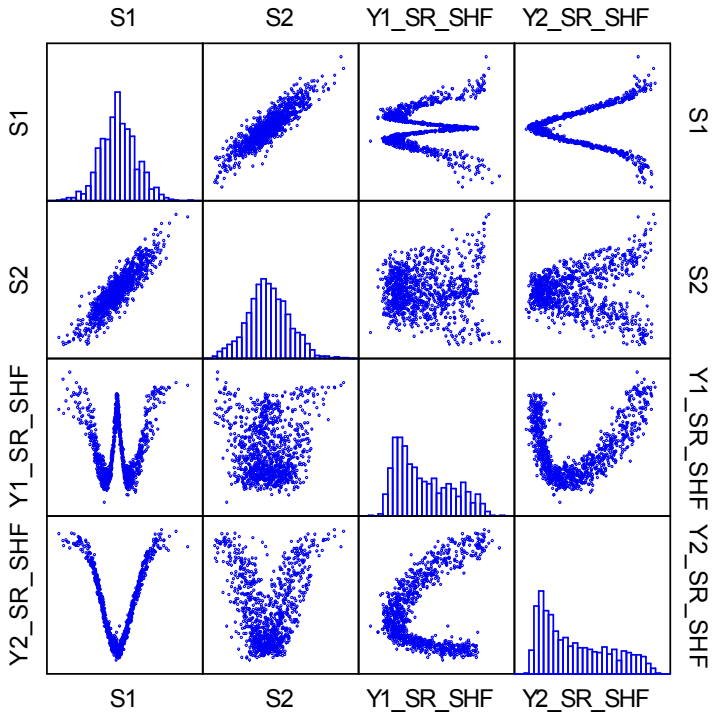


Figure 73. NM.03 Dataset – Masked using masking method 4

Appendix I – Dataset: NM.04

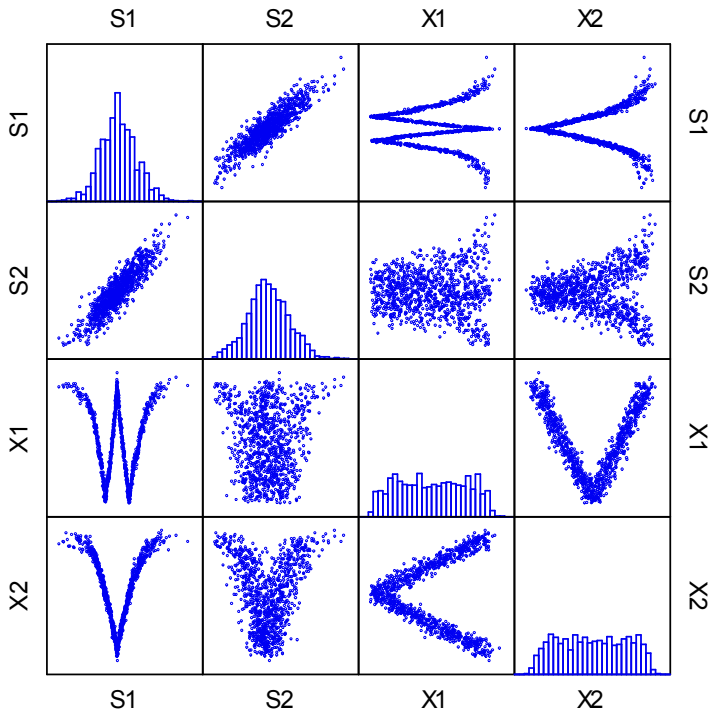


Figure 74. NM.04 Dataset

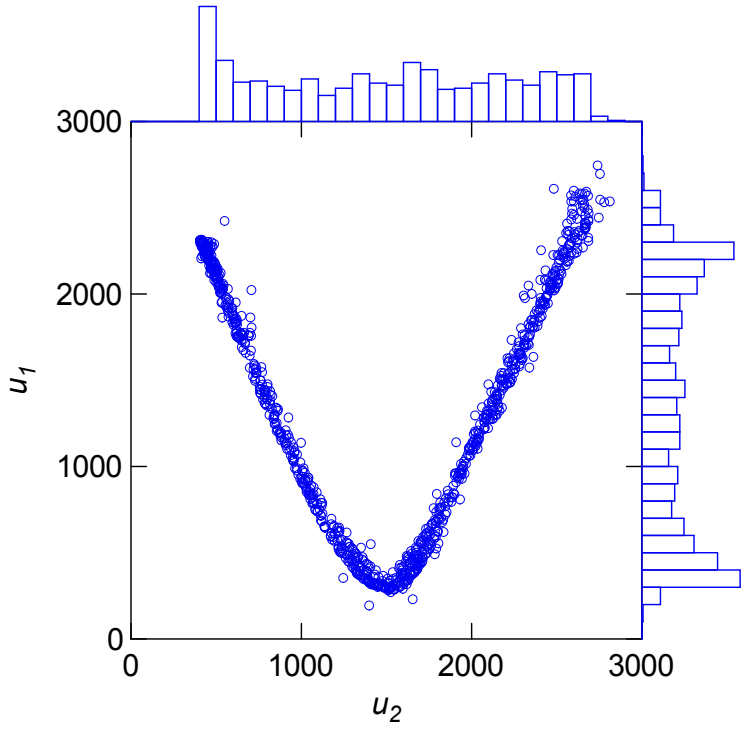


Figure 75. NM.04 Dataset – Predicted values $E(X_1|S)$|s vs. $E(X_2|S)$|s

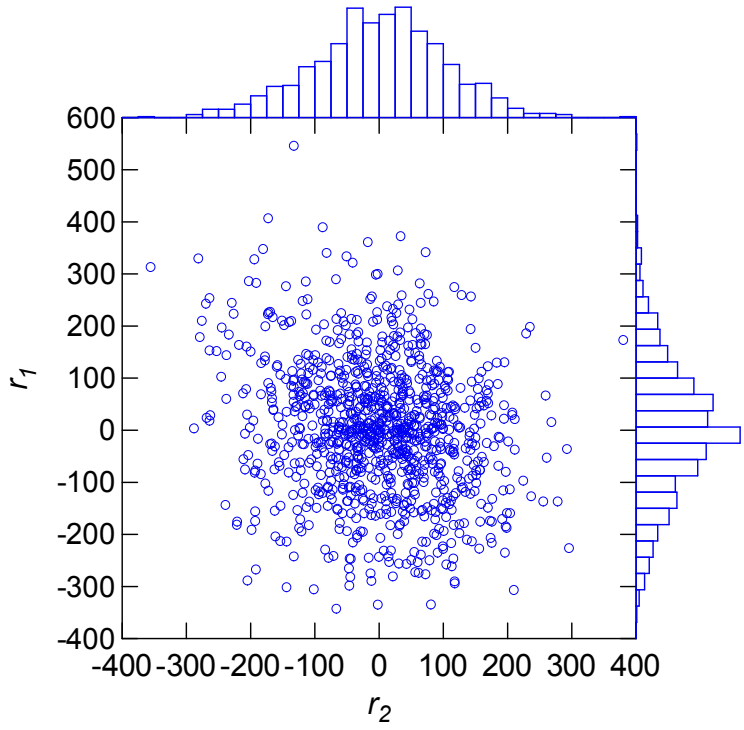


Figure 76. NM.04 Dataset – Residuals r_1 vs. r_2

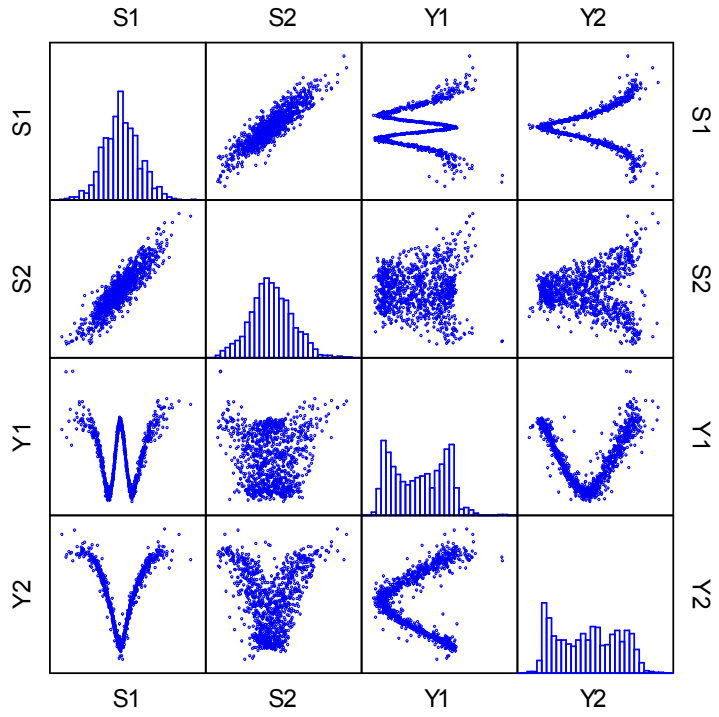


Figure 77. NM.04 Dataset – Masked using masking method 1

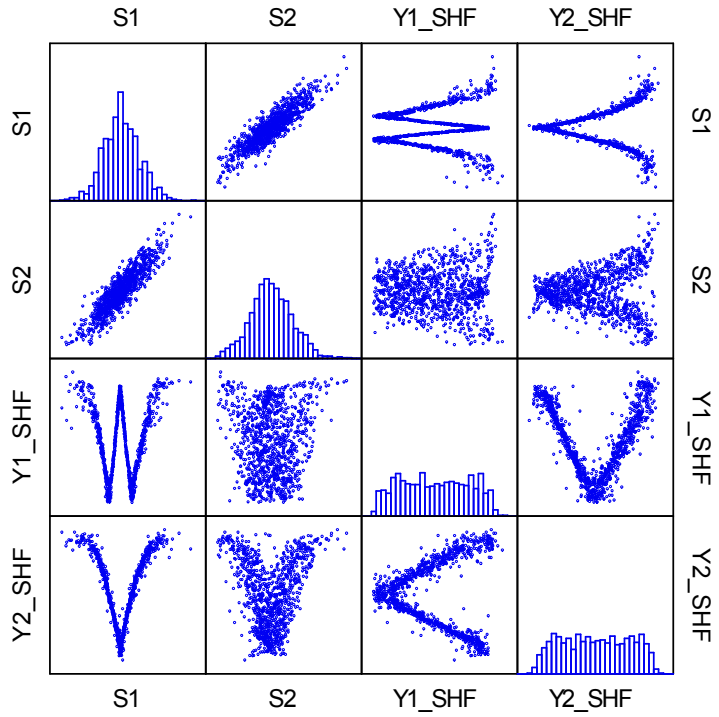


Figure 78. NM.04 Dataset – Masked using masking method 2

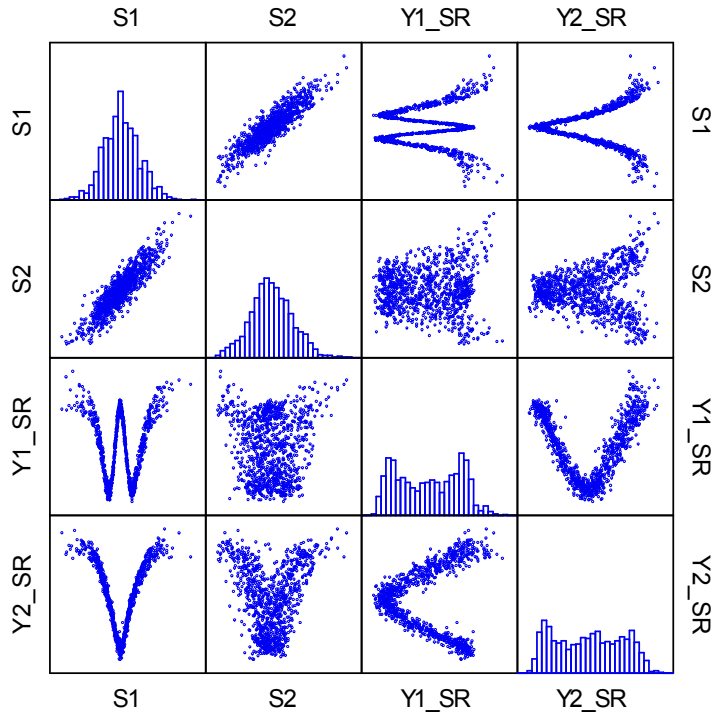


Figure 79. NM.04 Dataset – Masked using masking method 3

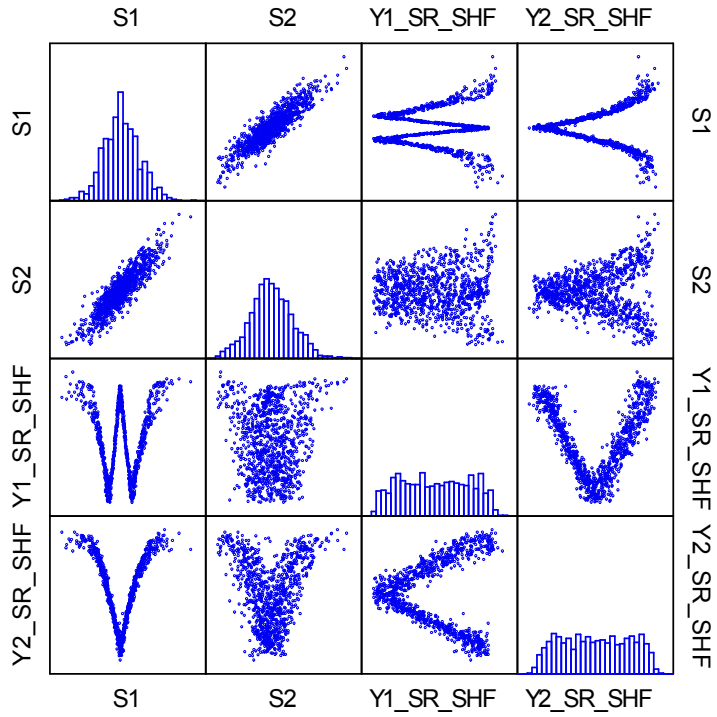


Figure 80. NM.04 Dataset – Masked using masking method 4

Appendix J – Dataset: NM.05

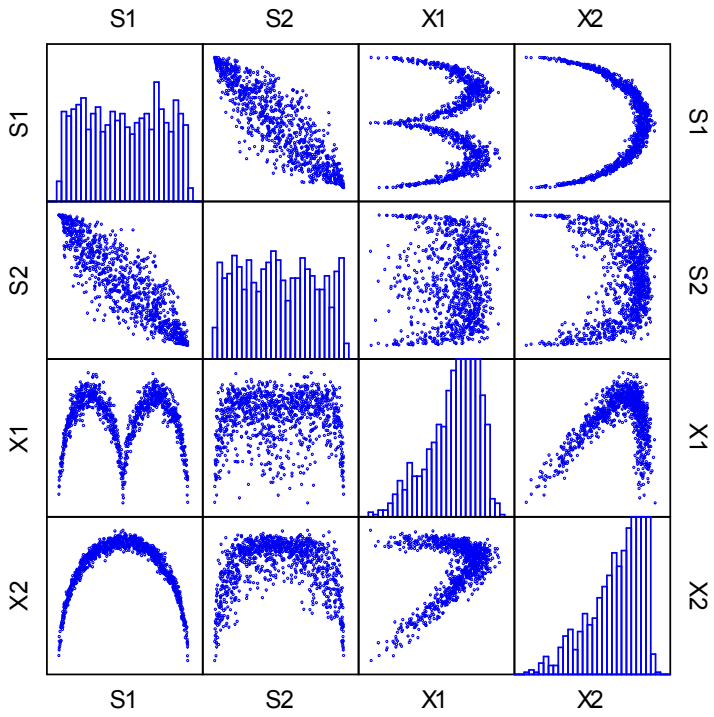


Figure 81. NM.05 Dataset

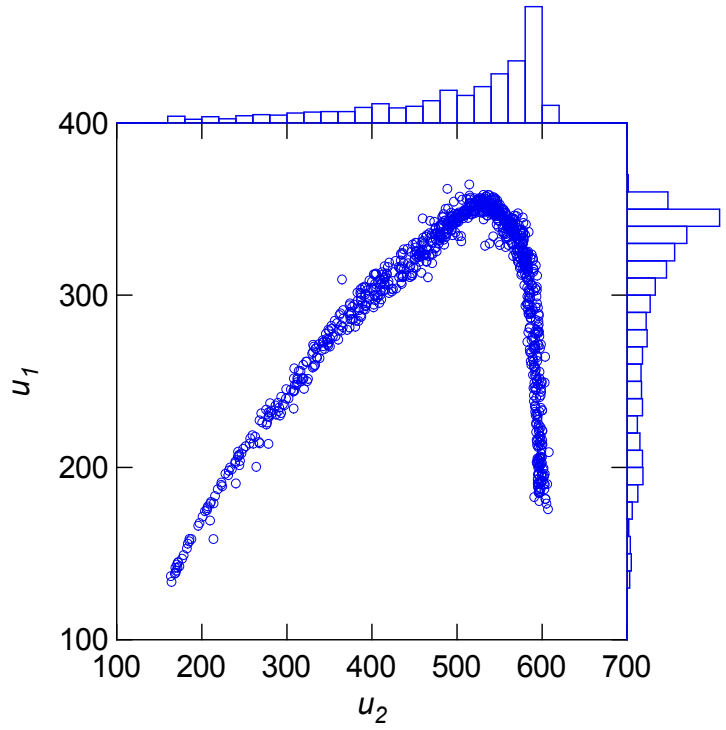


Figure 82. NM.05 Dataset – Predicted values $E(X_1|S)$|s vs. $E(X_2|S)$|s

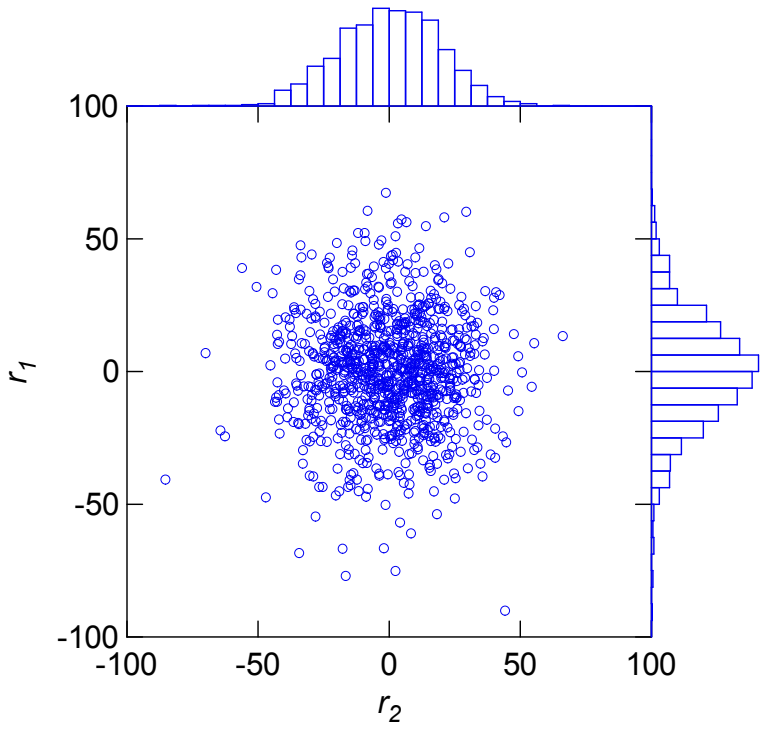


Figure 83. NM.05 Dataset – Residuals r_1 vs. r_2

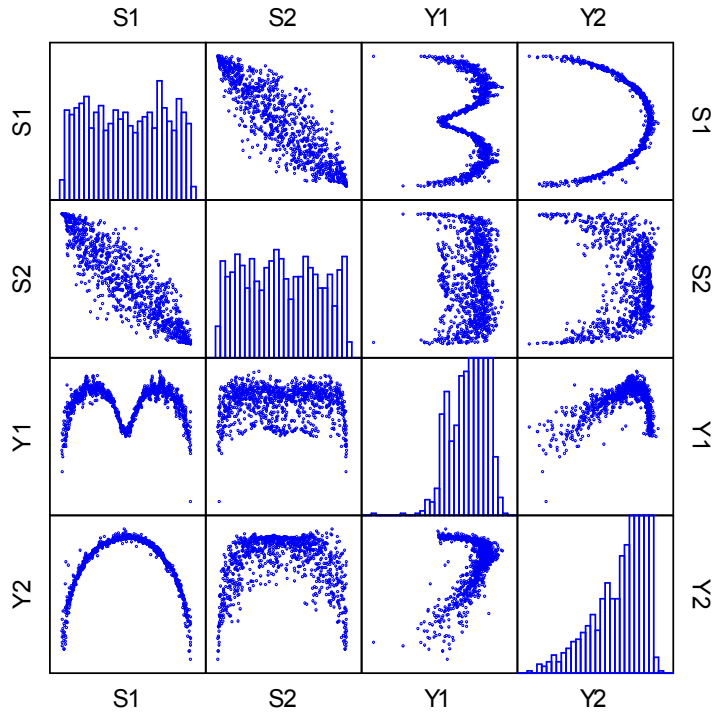


Figure 84. NM.05 Dataset – Masked using masking method 1

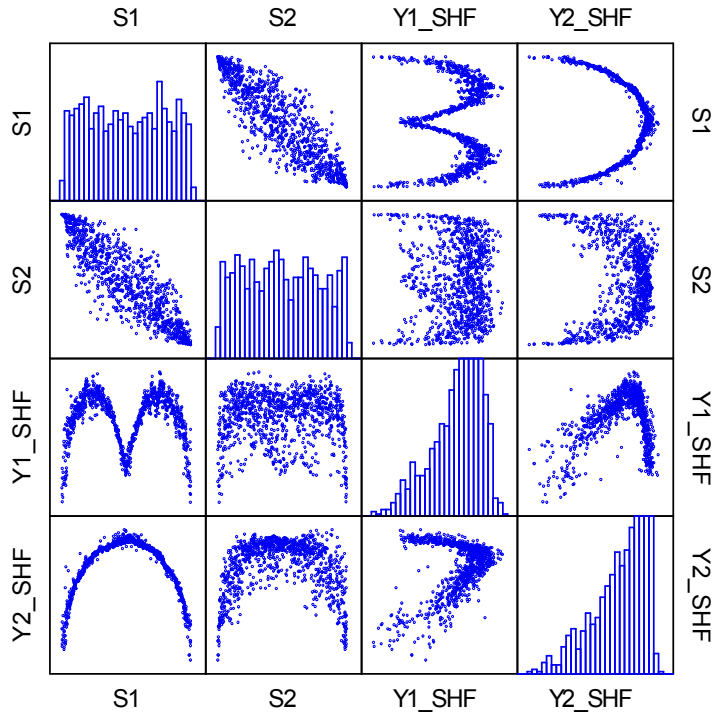


Figure 85. NM.05 Dataset – Masked using masking method 2

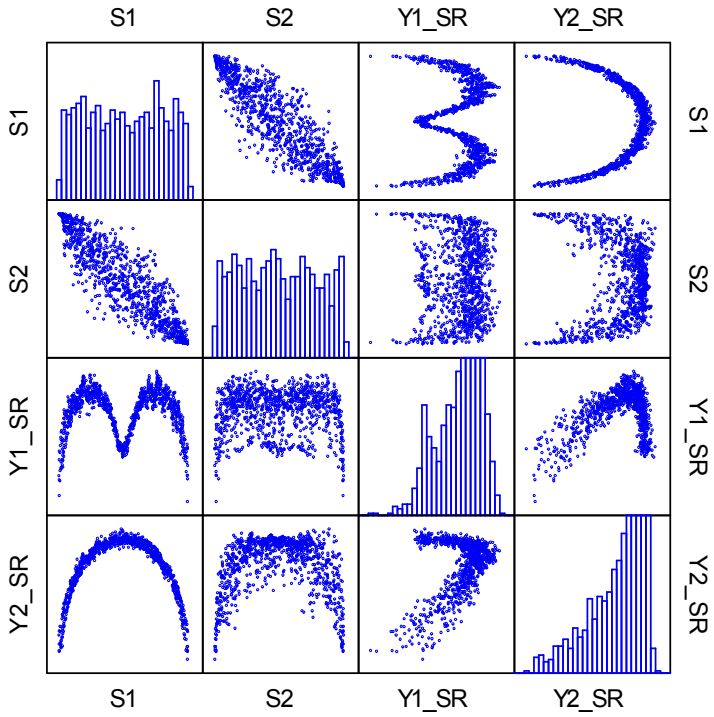


Figure 86. NM.05 Dataset – Masked using masking method 3

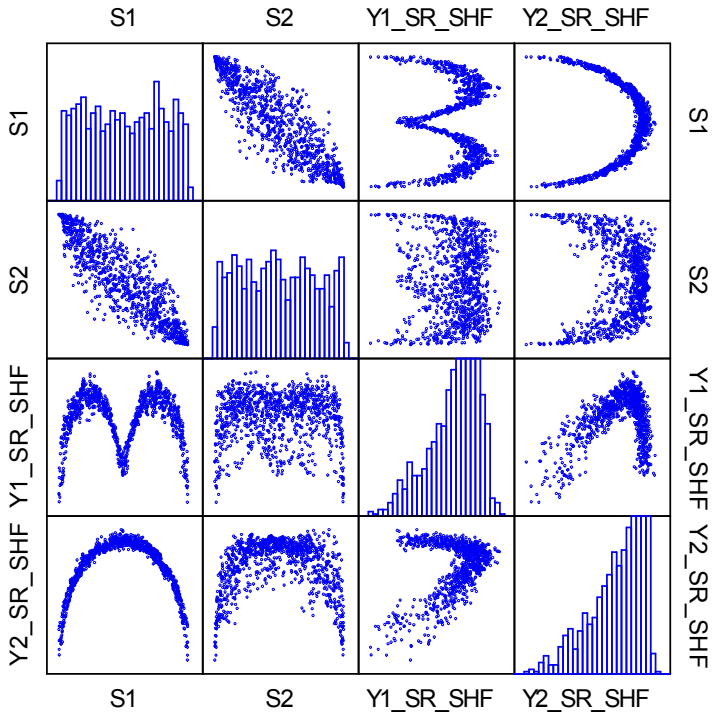


Figure 87. NM.05 Dataset – Masked using masking method 4

Appendix K – Dataset: MNL.01

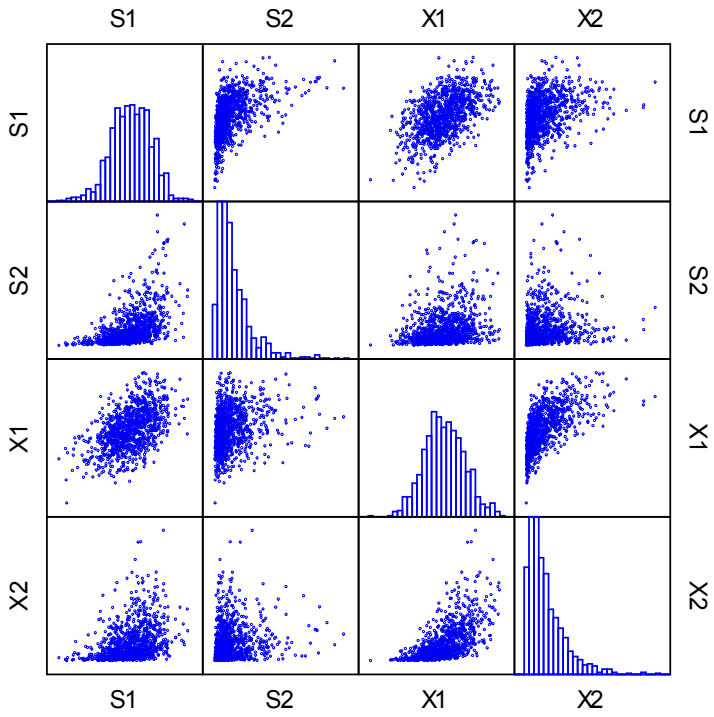


Figure 88. MNL.01 Dataset

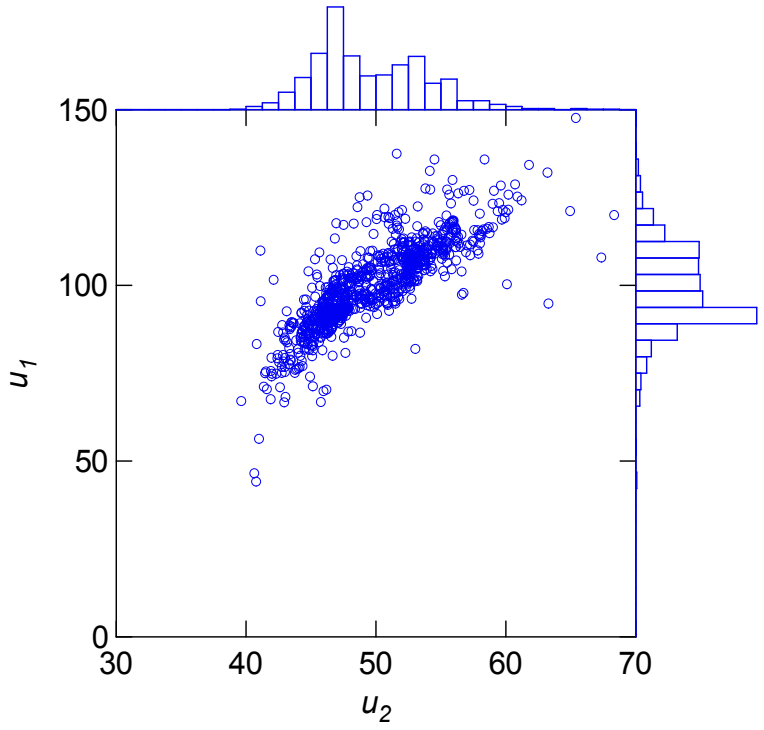


Figure 89. MNL.01 Dataset – Predicted values $E(X_1|S)|s$ vs. $E(X_2|S)|s$

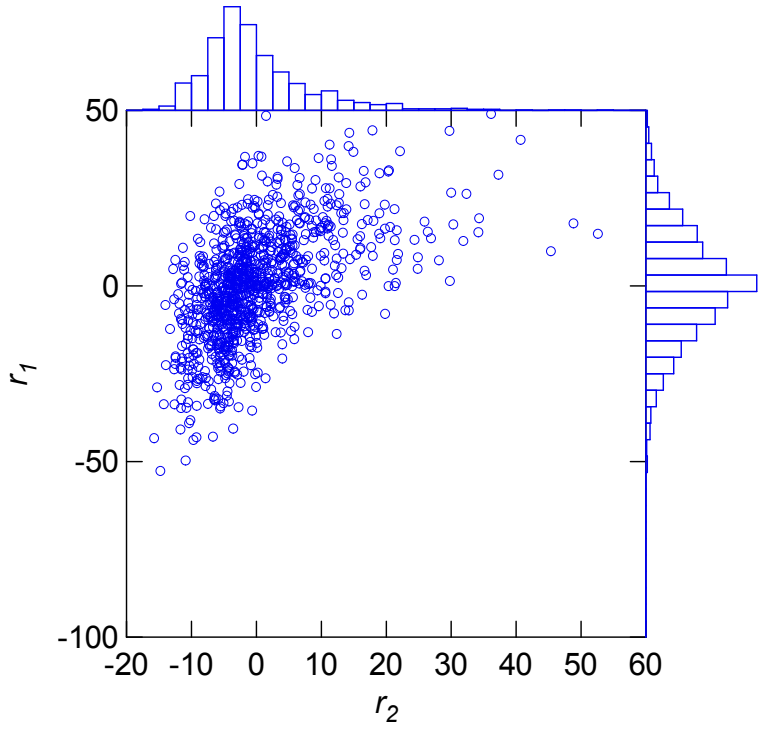


Figure 90. MNL.01 Dataset – Residuals r_1 vs. r_2

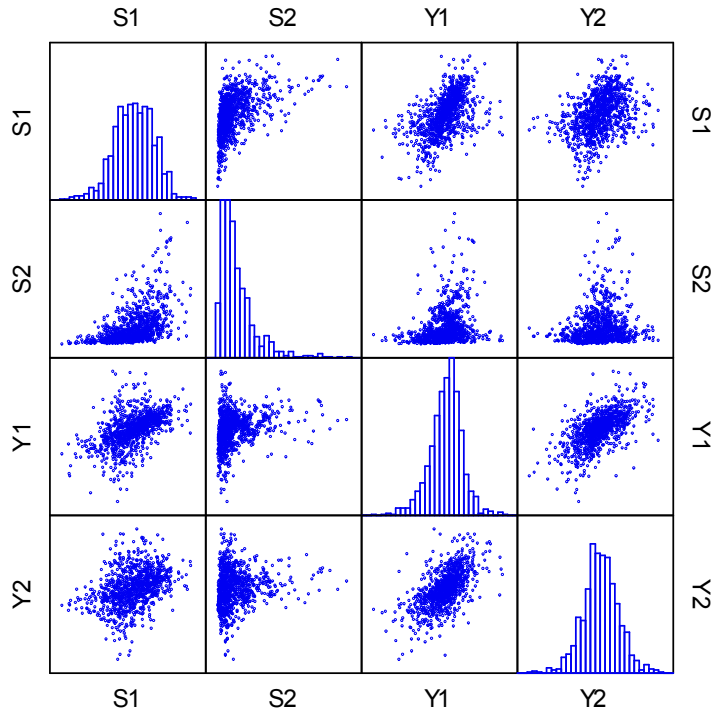


Figure 91. MNL.01 Dataset – Masked using masking method 1

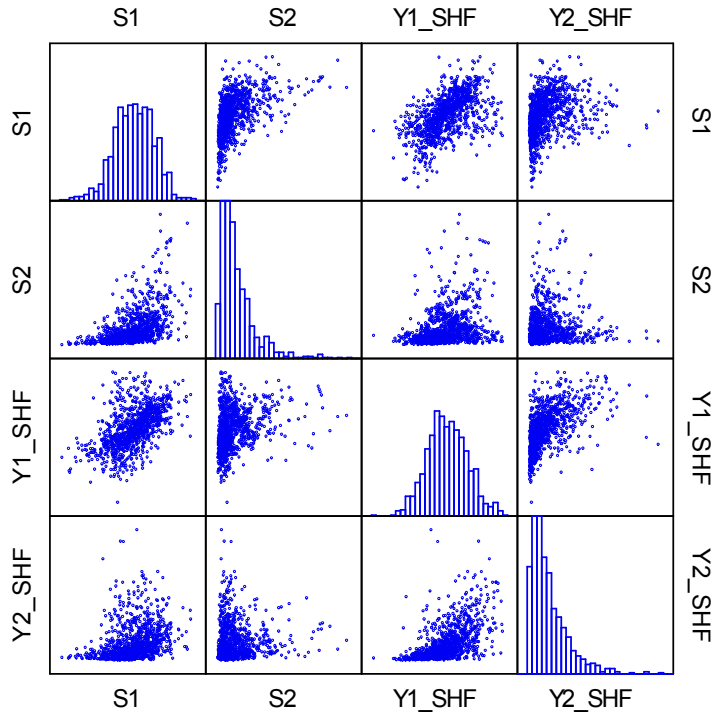


Figure 92. MNL.01 Dataset – Masked using masking method 2

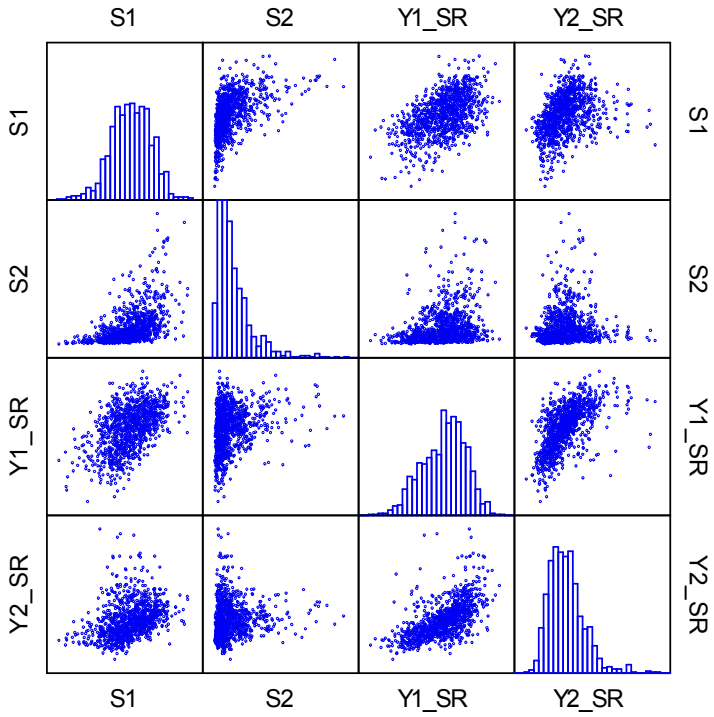


Figure 93. MNL.01 Dataset – Masked using masking method 3

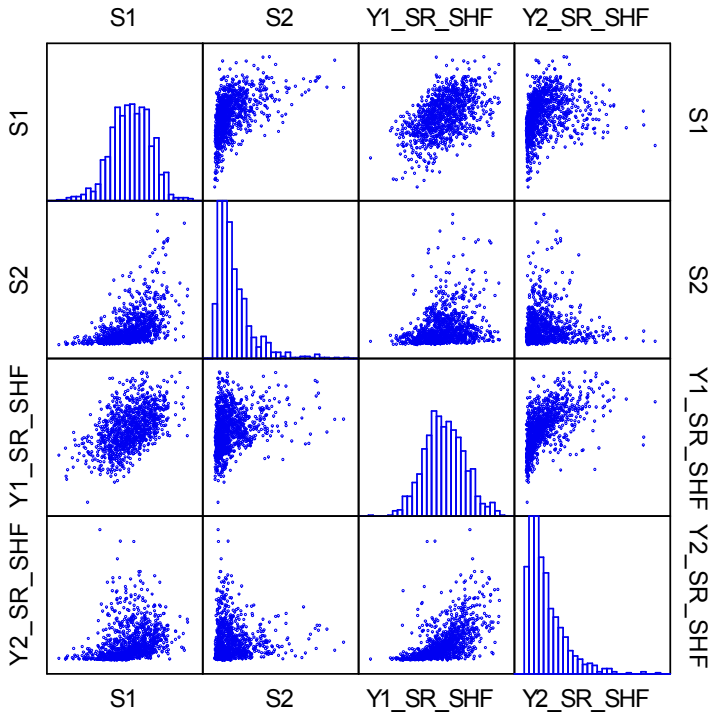


Figure 94. MNL.01 Dataset – Masked using masking method 4

Appendix L – Dataset: MNL.02

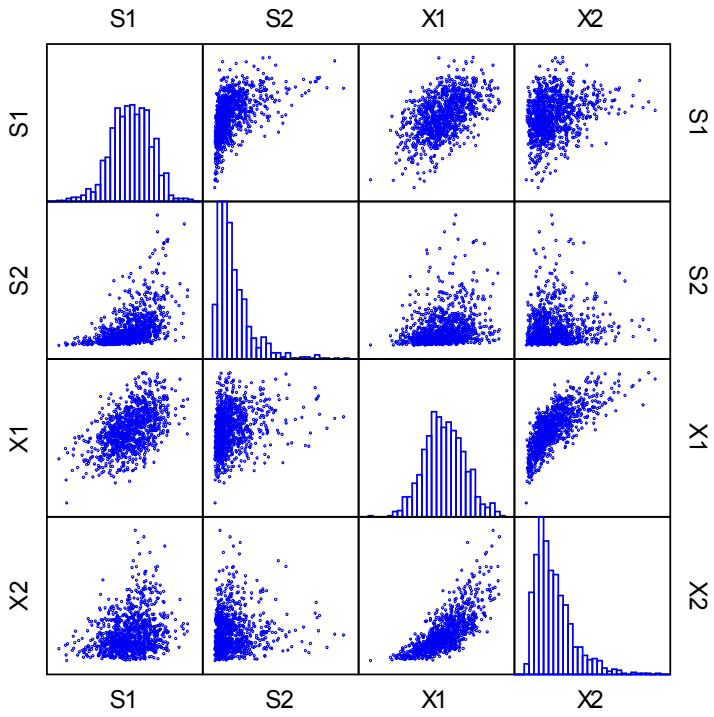


Figure 95. MNL.02 Dataset

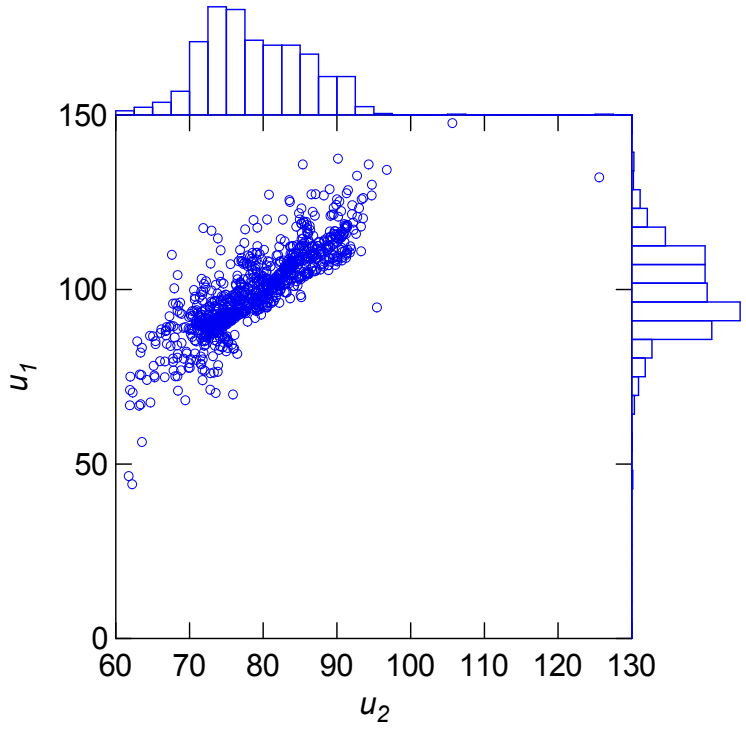


Figure 96. MNL.02 Dataset – Predicted values $E(X_1|S)|s$ vs. $E(X_2|S)|s$

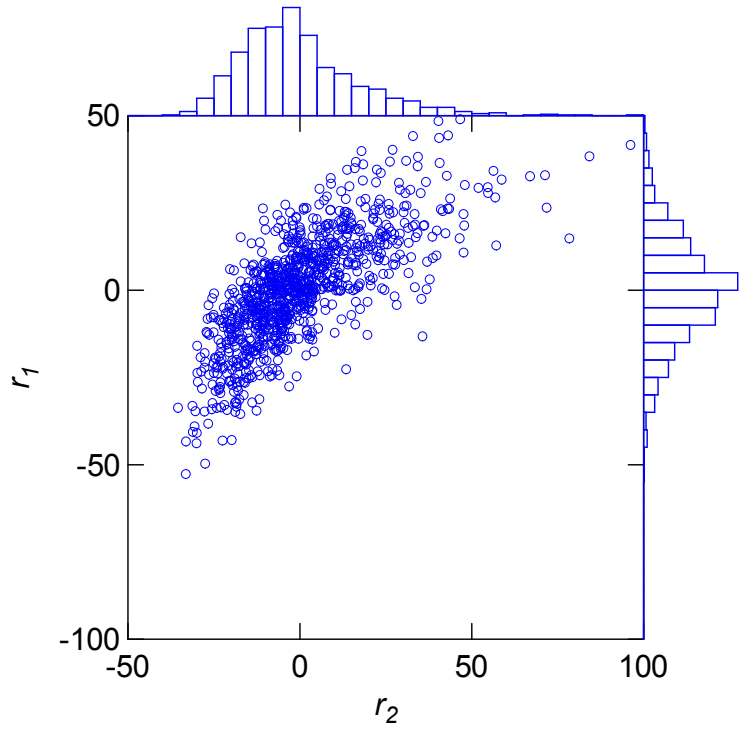


Figure 97. MNL.02 Dataset – Residuals r_1 vs. r_2

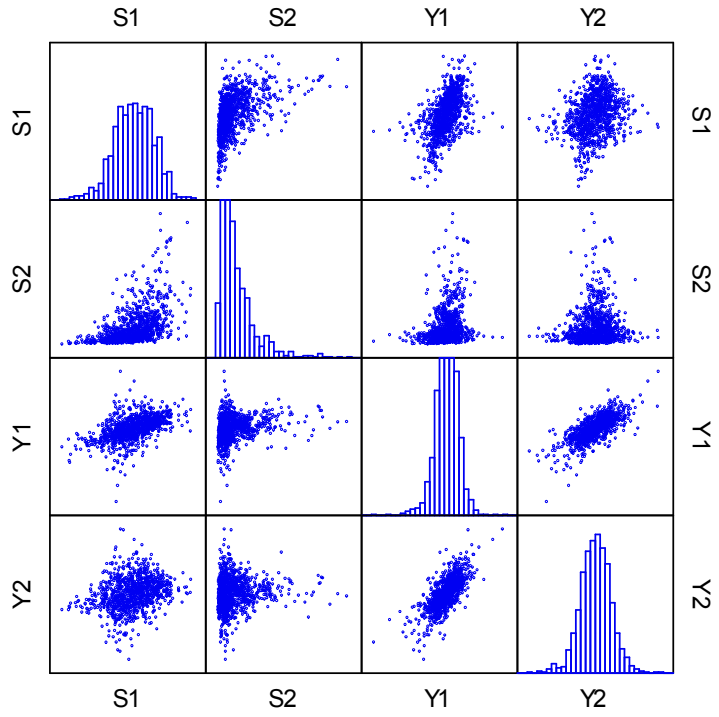


Figure 98. MNL.02 Dataset – Masked using masking method 1

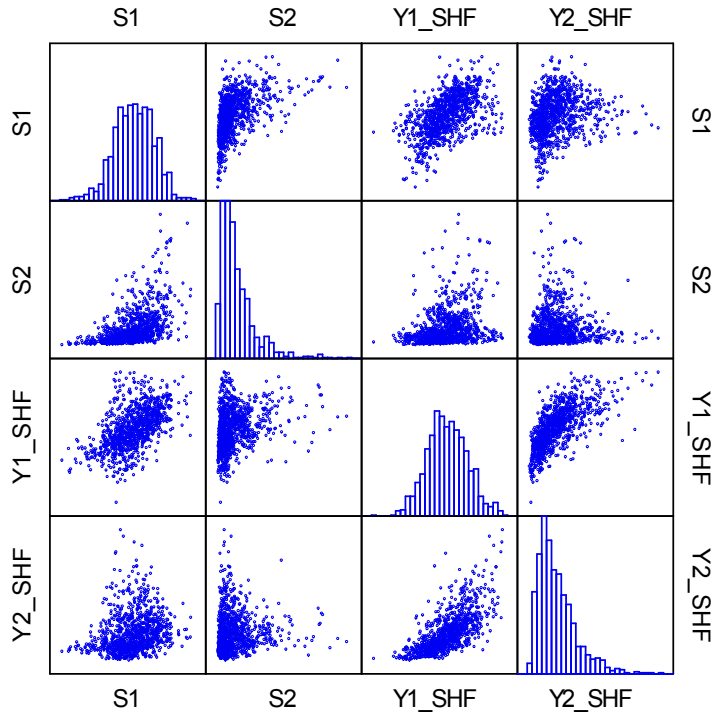


Figure 99. MNL.02 Dataset – Masked using masking method 2

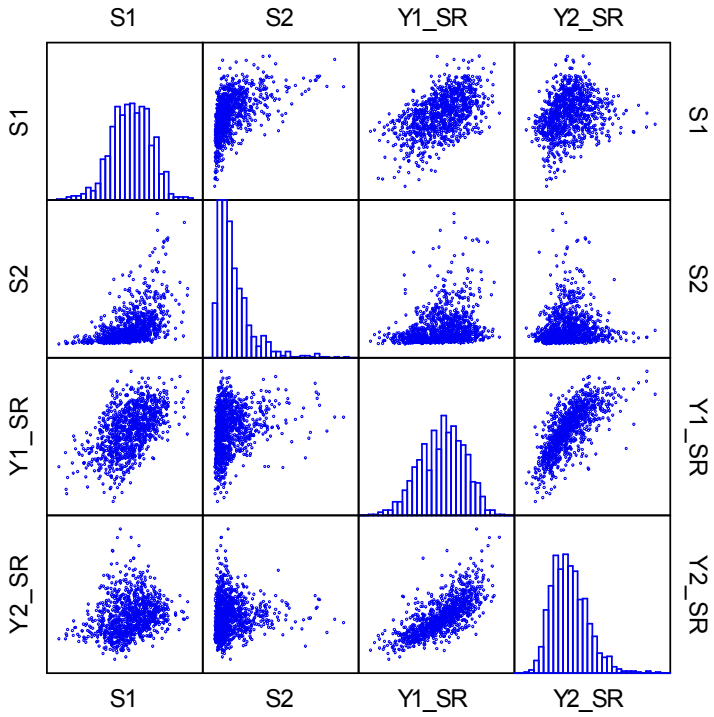


Figure 100. MNL.02 Dataset – Masked using masking method 3

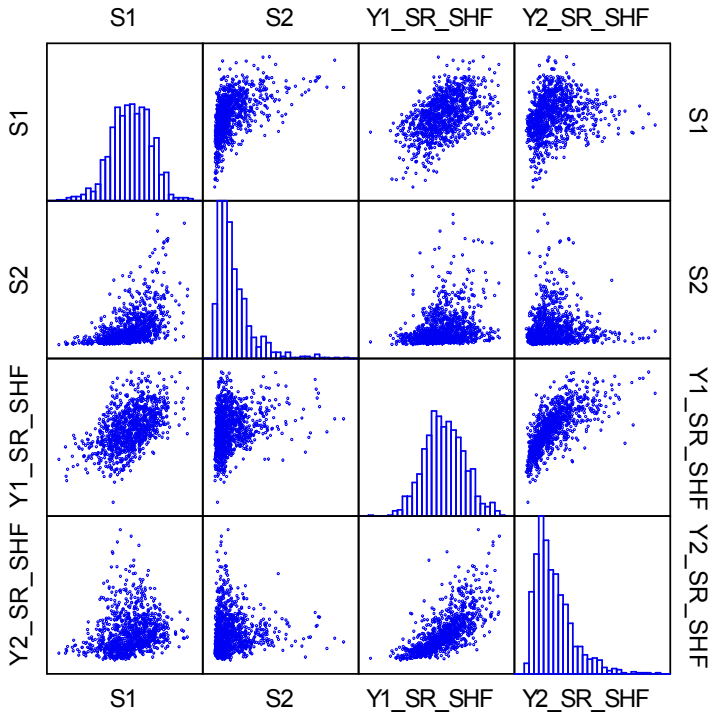


Figure 101. MNL.02 Dataset – Masked using masking method 4

Appendix M – Dataset: MNL.03

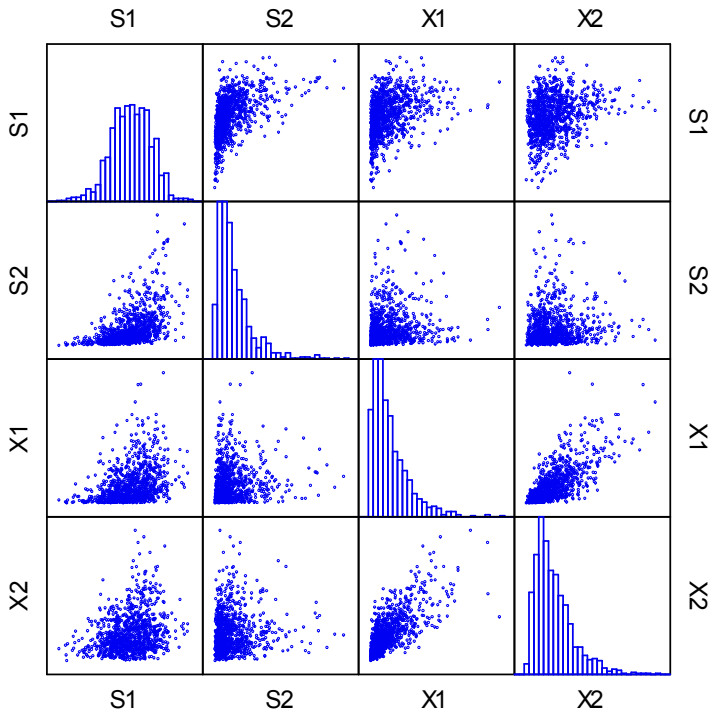


Figure 102. MNL.03 Dataset

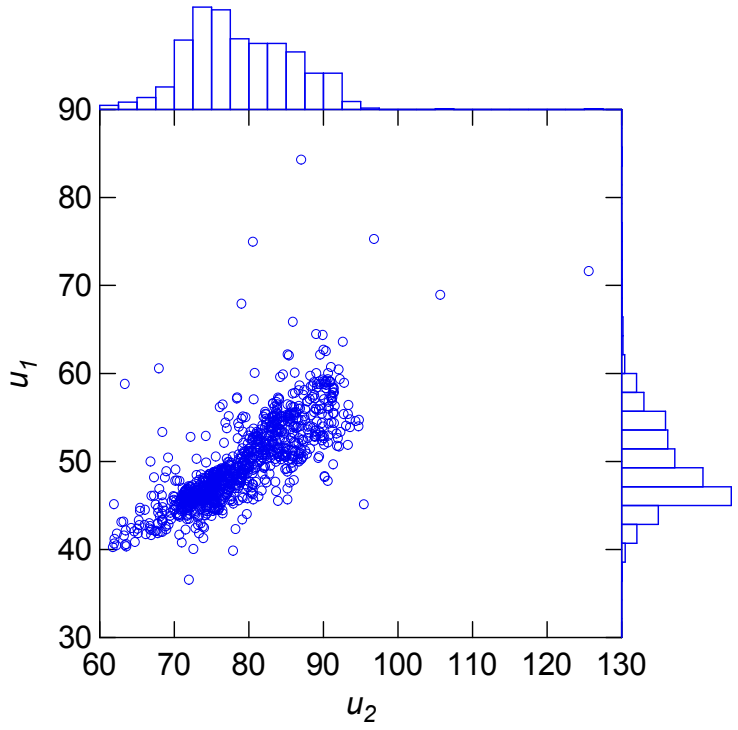


Figure 103. MNL.03 Dataset – Predicted values $E(X_1|S)$|s vs. $E(X_2|S)$|s

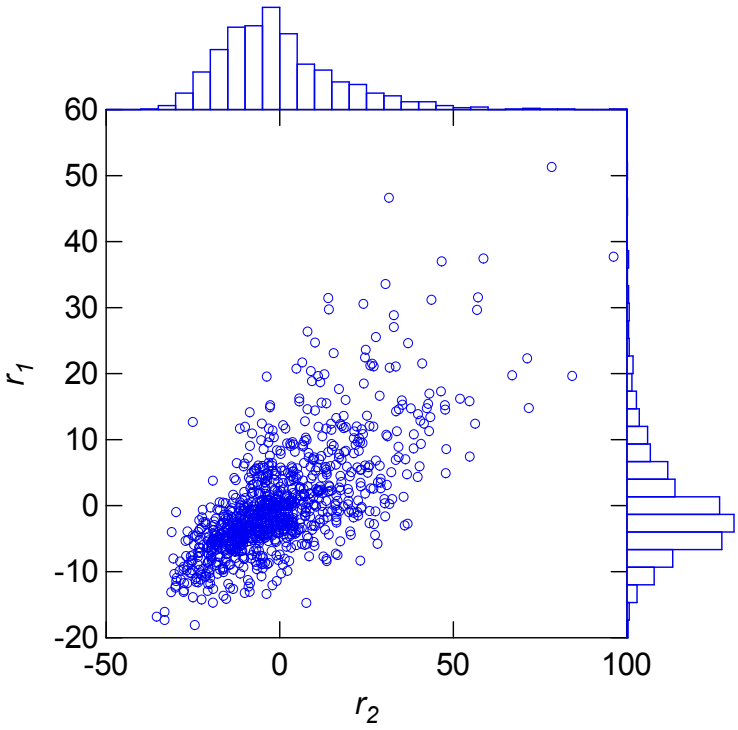


Figure 104. MNL.03 Dataset – Residuals r_1 vs. r_2

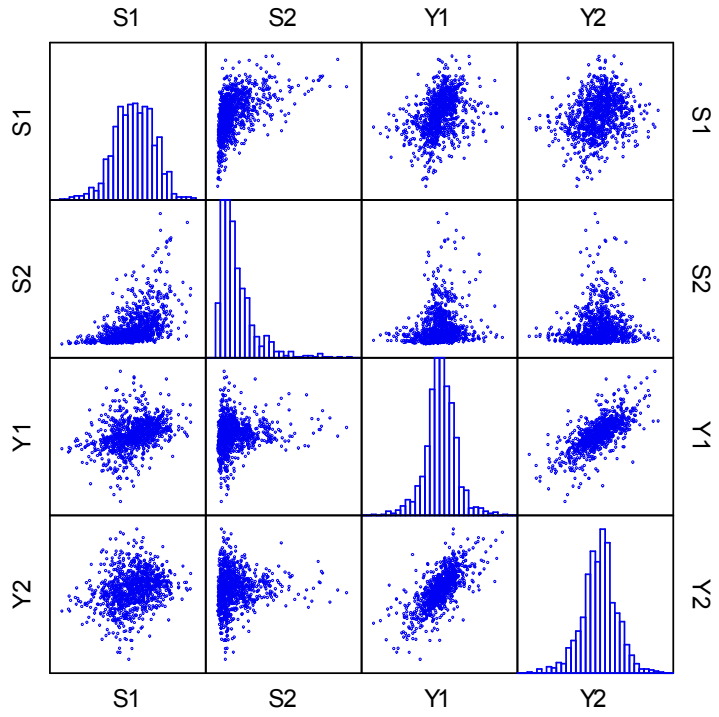


Figure 105. MNL.03 Dataset – Masked using masking method 1

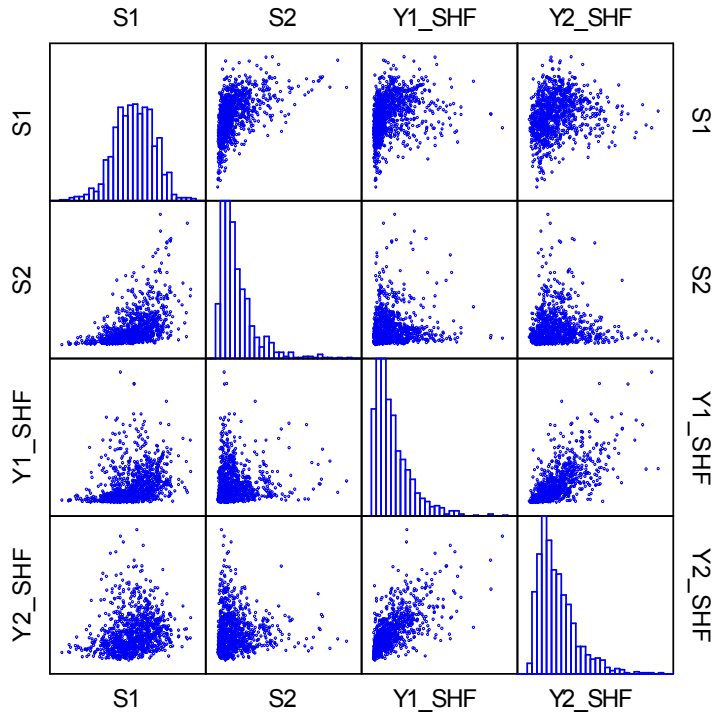


Figure 106. MNL.03 Dataset – Masked using masking method 2

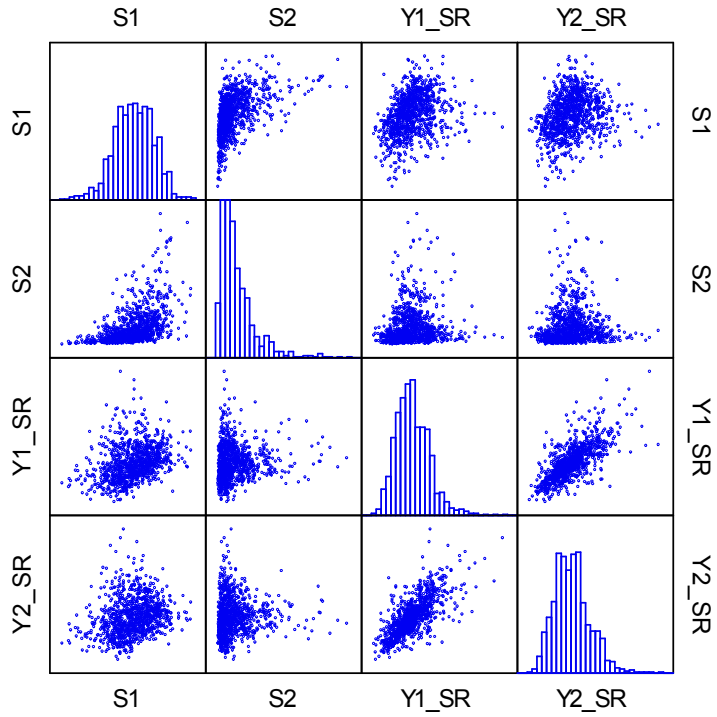


Figure 107. MNL.03 Dataset – Masked using masking method 3

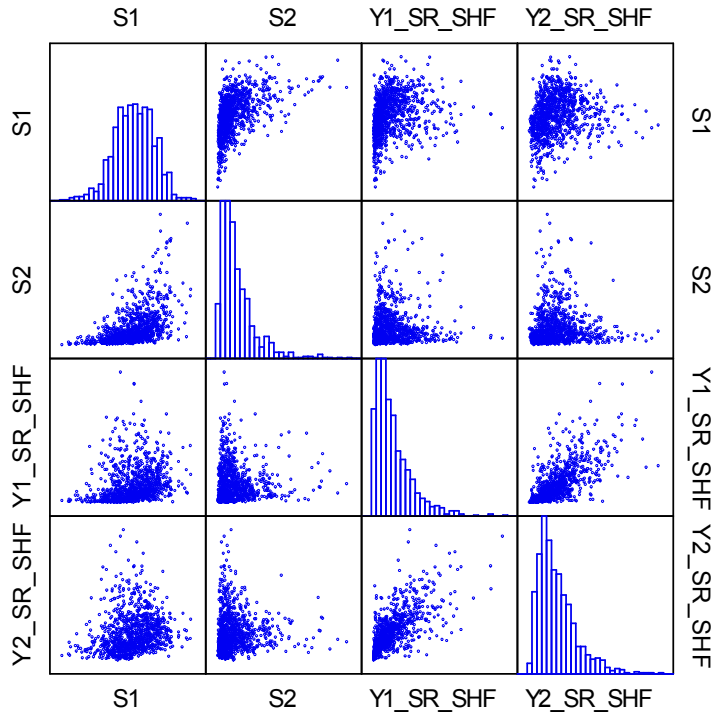


Figure 108. MNL.03 Dataset – Masked using masking method 4

Appendix N– Dataset: NM.01.S1 – Check Mark Dataset with One S with Non-Monotonic Relationships among Residuals

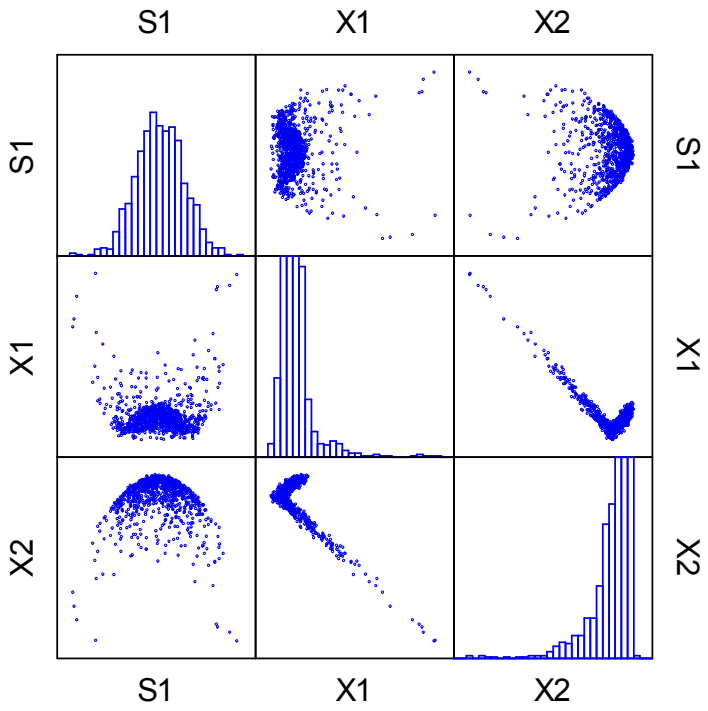


Figure 109. NM.01.S1 Dataset

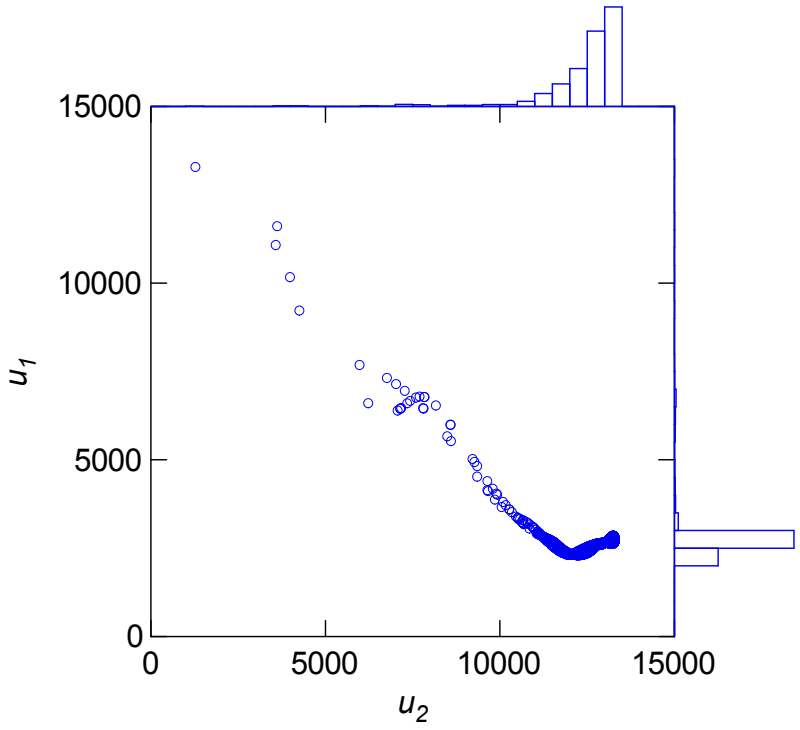


Figure 110. NM.01.S1 Dataset – Predicted values $E(X_1|S)|$s vs. $E(X_2|S)|$s

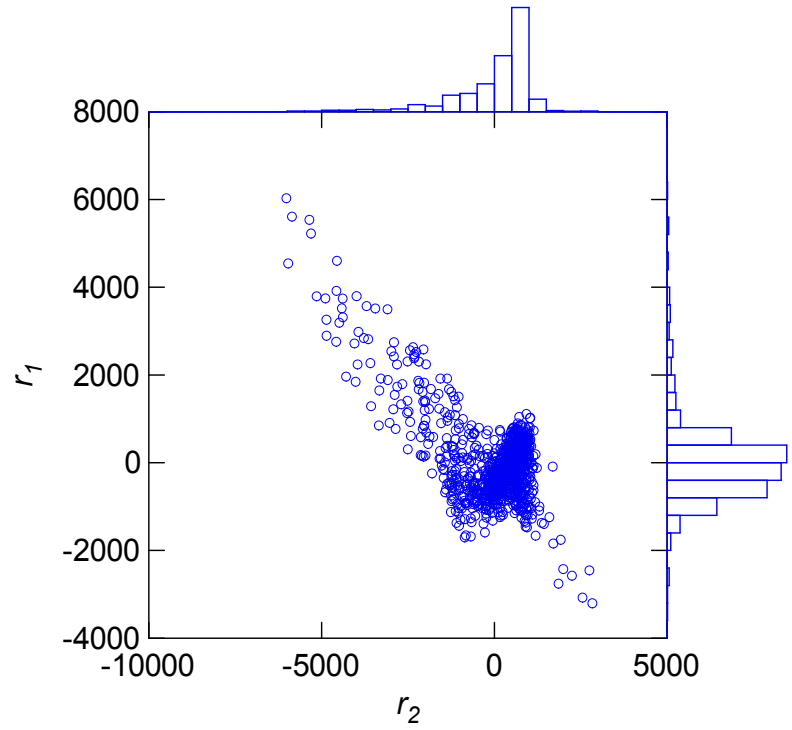


Figure 111. NM.01.S1 Dataset – Residuals r_1 vs. r_2

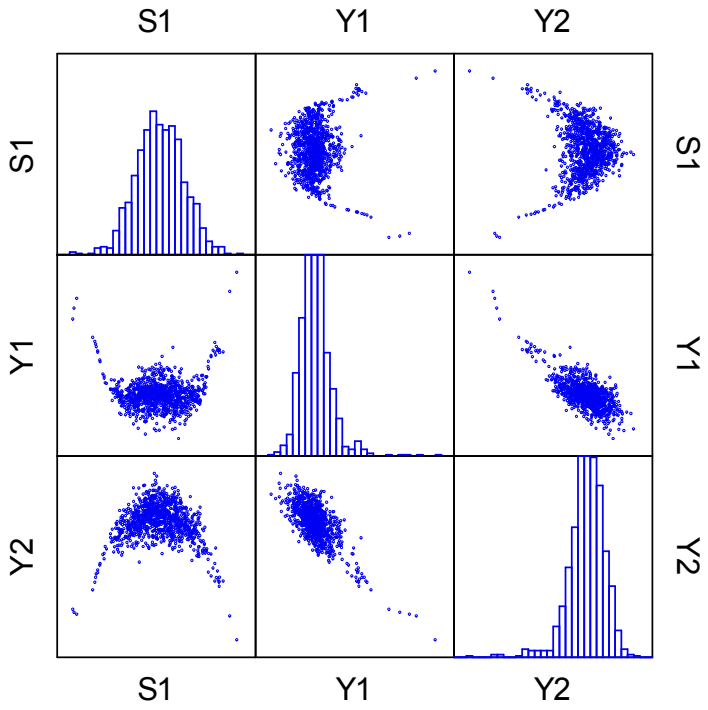


Figure 112. NM.01.S1 Dataset – Masked using masking method 1

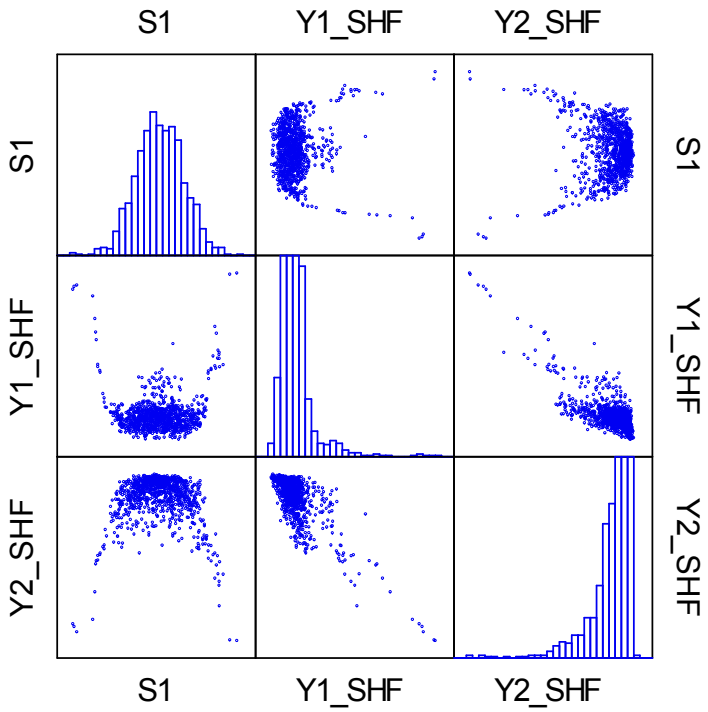


Figure 113. NM.01.S1 Dataset – Masked using masking method 2

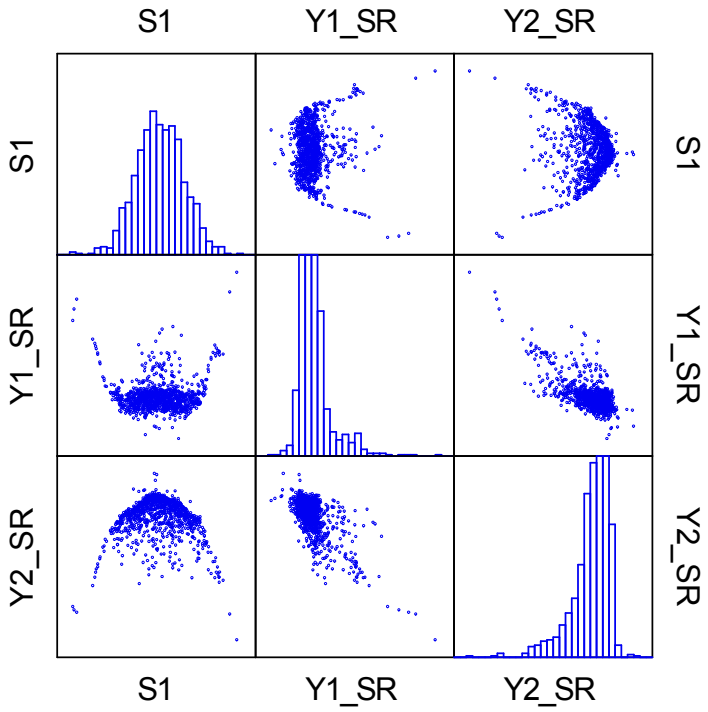


Figure 114. NM.01.S1 Dataset – Masked using masking method 3

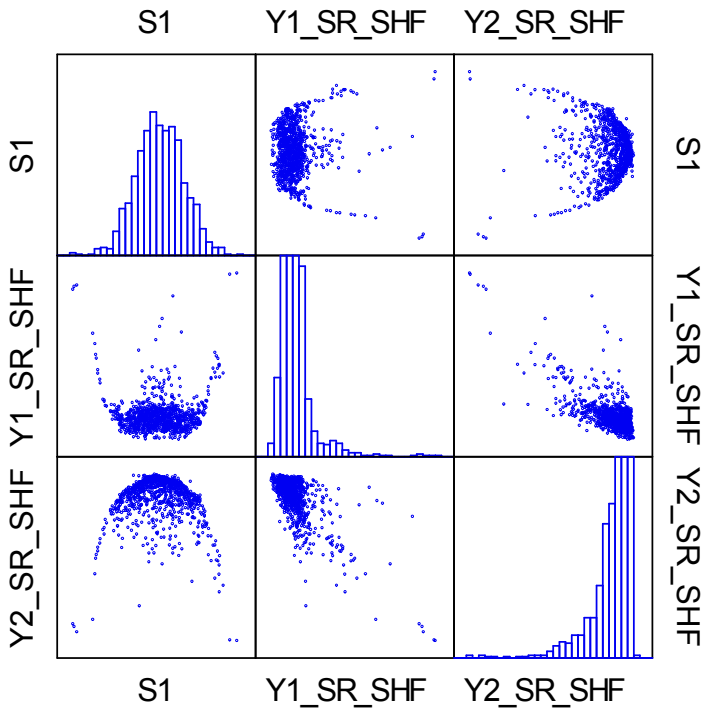


Figure 115. NM.01.S1 Dataset – Masked using masking method 4

Appendix O – Dataset: ME.L.S1 – Motivation Example with One S

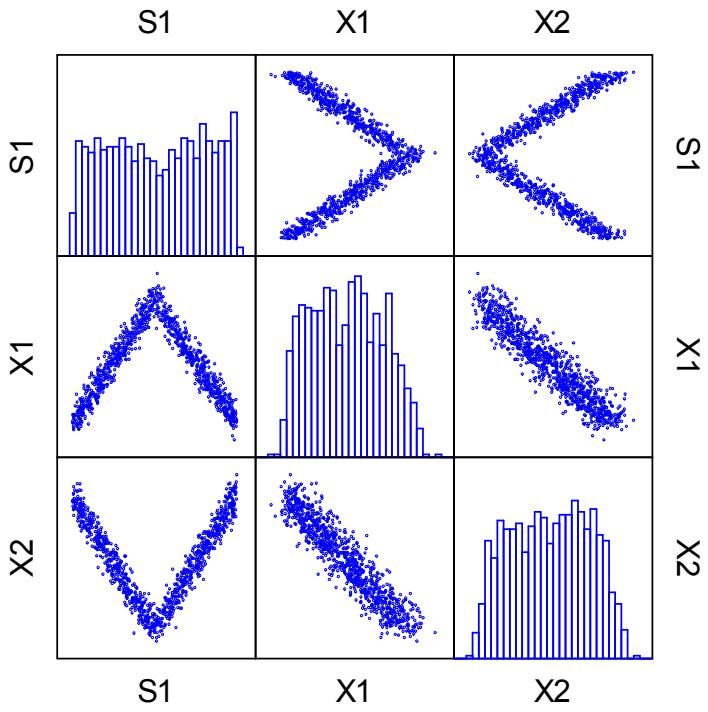


Figure 116. ME.L.S1 Dataset

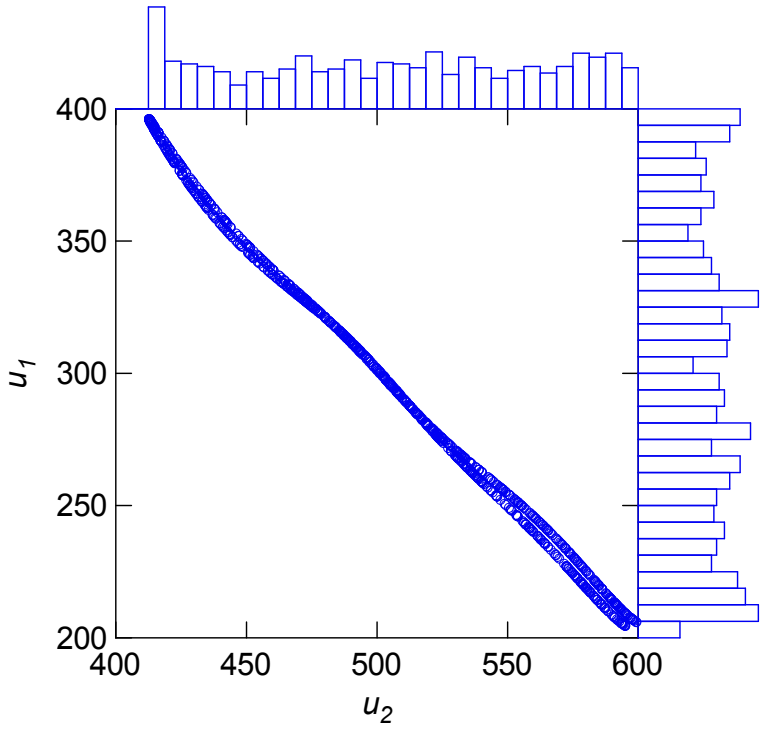


Figure 117. ME.L.S1 Dataset – Predicted values $E(X_1|S)|s$ vs. $E(X_2|S)|s$

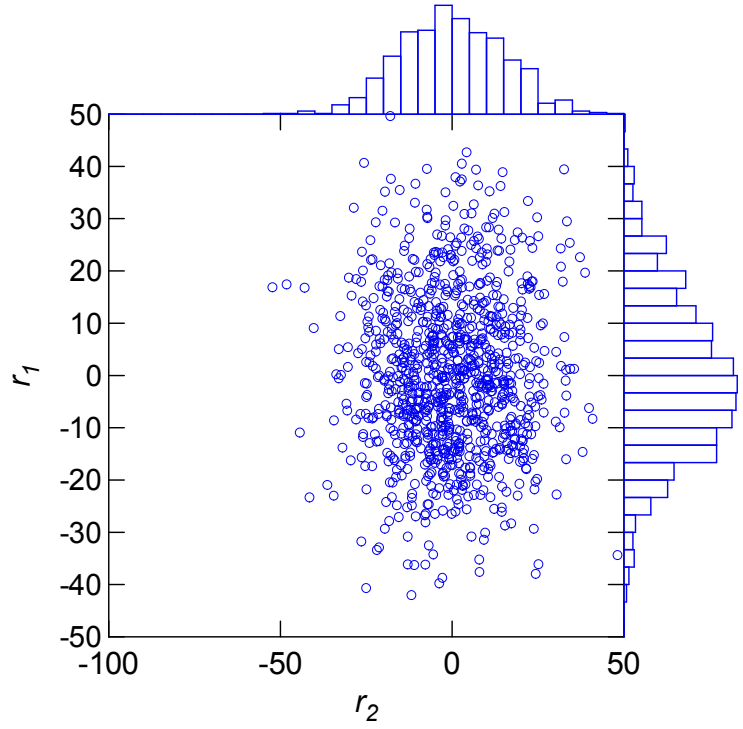


Figure 118. ME.L.S1 Dataset – Residuals r_1 vs. r_2

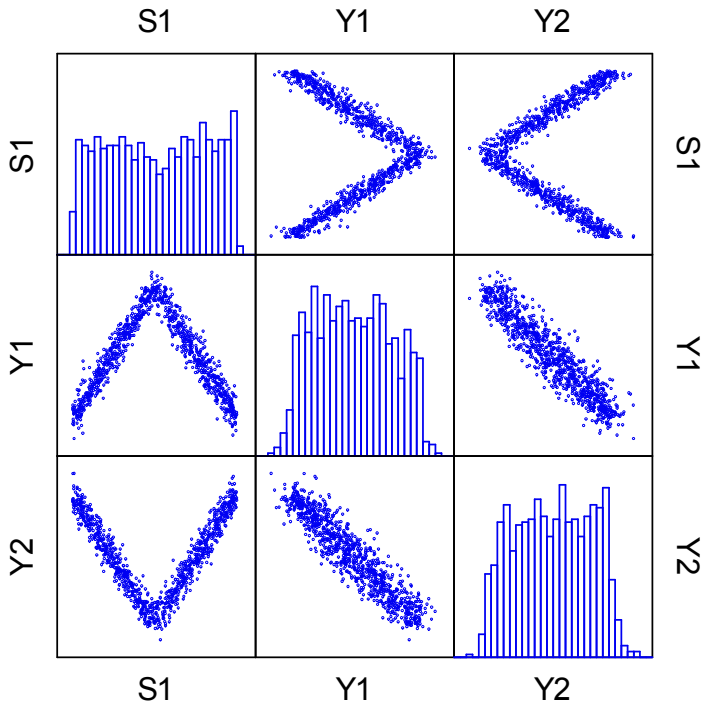


Figure 119. ME.L.S1 Dataset – Masked using masking method 1

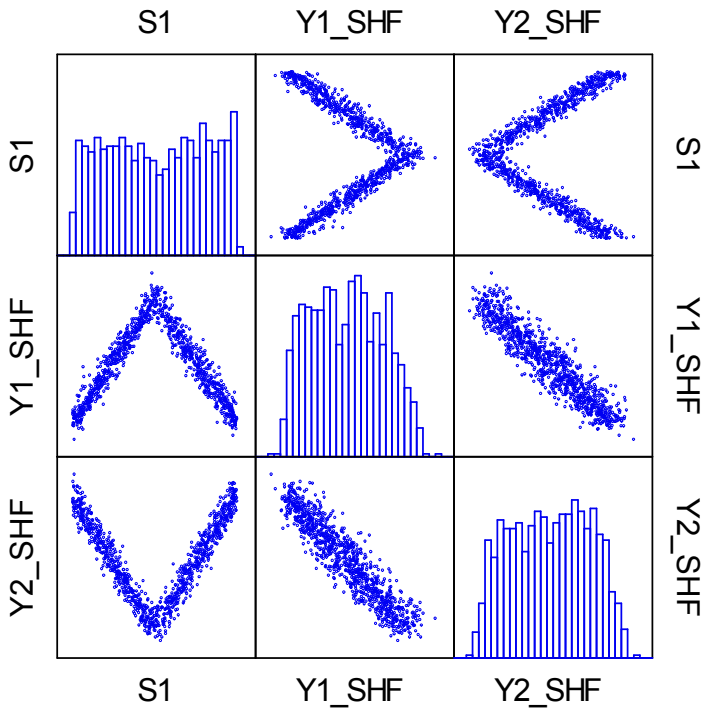


Figure 120. ME.L.S1 Dataset – Masked using masking method 2

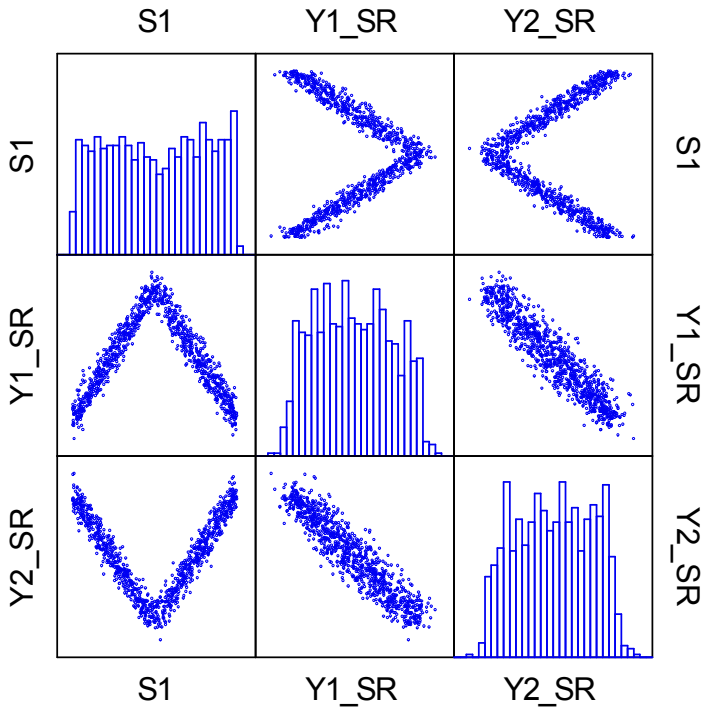


Figure 121. ME.L.S1 Dataset – Masked using masking method 3

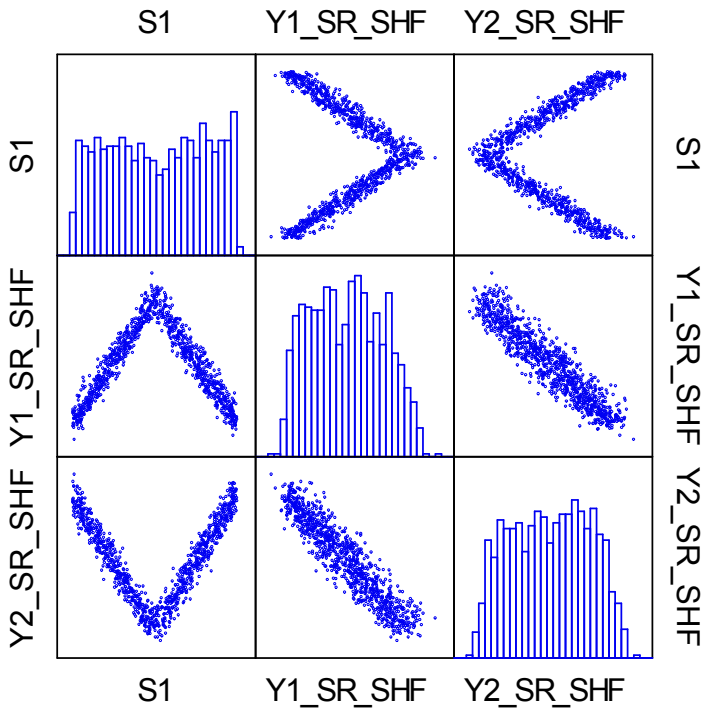


Figure 122. ME.L.S1 Dataset – Masked using masking method 4

Appendix P – Important Results Relating the Characteristics of Masked Attributes to the Characteristics of Original Data

In this appendix, we present two important results along their proofs. In addition, we provide some other results without explicit proofs.

Proposition I:

The relationship between the covariance of X_i and its masked copy Y_i masked using the RBM NM-EGADP approach and the conditional expectation of a confidential attribute X_i on the non-confidential attributes **S** is:

$$Cov(X_i, Y_i) = Var(E(X_i|\mathbf{S})) \tag{P.1}$$

where $i = 1\text{K } q$.

♦

Proof:

Using (A.11.14) pp. 458 in (Bickel and Doksum, 2001), we can write the above covariance as follows:

$$Cov(X_i, Y_i) = E(X_i Y_i) - E(X_i)E(Y_i) \tag{P.2}$$

Since $X_i = E(X_i|\mathbf{S}) + r_i$ and $Y_i = E(X_i|\mathbf{S}) + e_i$, Equation (P.2) can be written as:

$$Cov(X_i, Y_i) = E((u_i + r_i)(u_i + e_i)) - E(u_i + r_i)E(u_i + e_i) \tag{P.3}$$

where $u_i = E(X_i|\mathbf{S})$. Since $r_i, e_i \perp u_i$, we can write Equation (P.3) as:

$$Cov(X_i, Y_i) = E((u_i + r_i)(u_i + e_i)) - (E(u_i) + E(r_i))(E(u_i) + E(e_i)) \tag{P.4}$$

Since $E(r_i) = E(e_i) = 0$, we can write the following:

$$Cov(X_i, Y_i) = E((u_i + r_i)(u_i + e_i)) - [E(u_i)]^2 \tag{P.5}$$

This can be simplified to:

$$Cov(X_i, Y_i) = E(u_i^2 + u_i e_i + u_i r_i + r_i e_i) - [E(u_i)]^2 \quad \text{(P.6)}$$

By distributing the expectations, we get:

$$Cov(X_i, Y_i) = E(u_i^2) + E(u_i e_i) + E(u_i r_i) + E(r_i e_i) - [E(u_i)]^2 \quad \text{(P.7)}$$

As said, $r_i, e_i \perp u_i$ and $E(r_i) = E(e_i) = 0$. In addition, $e_i \perp r_i$ as a security requirement, which leads to:

$$Cov(X_i, Y_i) = E(u_i^2) + \cancel{E(u_i)E(e_i)} + \cancel{E(u_i)E(r_i)} + \cancel{E(r_i)E(e_i)} - [E(u_i)]^2 \quad \text{(P.8)}$$

Or simply:

$$Cov(X_i, Y_i) = E(u_i^2) - [E(u_i)]^2 \quad \text{(P.9)}$$

Using (A.11.15) pp. 458 in (Bickel and Doksum, 2001), this leads to:

$$Cov(X_i, Y_i) = Var(u_i) \quad \text{(P.10)}$$

Or:

$$Cov(X_i, Y_i) = Var(E(X_i|\mathbf{S})) \quad \text{(P.11)}$$

This ends our proof.

♦

Proposition II:

The security measure (*SI*) is related to the correlation between X_i and its masked copy Y_i, which is masked using the RBM NM-EGADP approach, as follows:

$$Corr(X_i, Y_i) = 1 - \eta \quad \text{(P.12)}$$

Where

$$\eta = \frac{Var(r_i)}{Var(X_i)} \quad \text{(P.13)}$$

♦

Proof:

By using Proposition 1.4.1 (c) pp. 34 in (Bickel and Doksum, 2001), we can write Equation (6.10) as:

$$\eta = \frac{Var(r_i)}{Var(r_i) + Var(E(X_i|\mathbf{S}))} \tag{P.14}$$

Using Equation (P.1), we can write:

$$\eta = \frac{Var(r_i)}{Var(r_i) + Cov(X_i, Y_i)} \tag{P.15}$$

We can simplify it in a series of steps:

$$\eta Var(r_i) + \eta Cov(X_i, Y_i) = Var(r_i) \tag{P.16}$$

$$\eta Cov(X_i, Y_i) = Var(r_i) - \eta Var(r_i) \tag{P.17}$$

$$\eta Cov(X_i, Y_i) = Var(r_i)(1 - \eta) \tag{P.18}$$

$$Cov(X_i, Y_i) = Var(r_i)\left(\frac{1}{\eta} - 1\right) \tag{P.19}$$

To convert the covariance in the LHS of (P.19) to correlation (see (A.11.18) pp. 458 in (Bickel and Doksum, 2001)), we must divide both sides by $\sqrt{Var(X_i)Var(Y_i)}$. However, since one of the specifications of the RBM approach is that $Var(Y_i) = Var(X_i)$, we can divide both sides by $Var(X_i)$:

$$\frac{Cov(X_i, Y_i)}{Var(X_i)} = \frac{Var(r_i)}{Var(X_i)}\left(\frac{1}{\eta} - 1\right) \tag{P.20}$$

This leads to

$$Corr(X_i, Y_i) = \eta\left(\frac{1}{\eta} - 1\right) \tag{P.21}$$

which can be simplified to:

$$Corr(X_i, Y_i) = 1 - \eta \tag{P.22}$$

This ends our proof.

♦

Notice also that Equation (6.9) (or (P.22)) can be written as:

$$Corr(X_i, Y_i) = \frac{Var(E(X_i|\mathbf{S}))}{Var(X_i)} \tag{P.23}$$

Since $\eta = Var(r_i)/Var(X_i)$ and using Proposition 1.4.1 (c) pp. 34 in (Bickel and Doksum, 2001), we have:

$$\frac{Var(E(X_i|\mathbf{S}))}{Var(X_i)} + \frac{Var(r_i)}{Var(X_i)} = 1 \tag{P.24}$$

But we prefer to present Equation (6.9) in terms of the variance of residuals **r** because *Corr*(X_i, Y_i) is used to assess *security* (rather than *utility*) and residuals **r** are what determine the possible *security* level from the characteristics of original datasets. Nonetheless, Equation (6.11) is still interesting because when $E(X_i|S)$ represents the best linear predictor of X_i, The RHS of Equation (6.11) becomes (see pp. 36-37 and Equation (1.4.11) pp. 36 in (Bickel and Doksum, 2001)):

$$R^2(X_i|\mathbf{S}) = \frac{Var(E(X_i|\mathbf{S}))}{Var(X_i)} \tag{P.25}$$

Therefore, $Corr(X_i, Y_i) = R^2(X_i|\mathbf{S})$. In addition, Formula (P.25) is equivalent to (see pp. 40 in (Bickel and Doksum, 2001)):

$$R^2(X_i|\mathbf{S}) = [Corr(X_i, E(X_i|\mathbf{S}))]^2 \left[= \frac{Var(E(X_i|\mathbf{S}))}{Var(X_i)}\right] \tag{P.26}$$

As said earlier, the assumption is that the conditional expectations $E(X_i|S)$ in (P.25) and (P.26) represent the best linear predictor for X_i. Nevertheless, when the estimation mechanism, the ANN machine in our case, does a good job in estimating the

conditional expectations, the relationships between the observations (X_i) and predicted or the fitted values ($E(X_i|\mathbf{S})|\mathbf{s}$) becomes linear since the model accounts for all *nonlinearity* that may exist between **S** and **X** (see pp. 443 in (Pyle, 2003)). In this case, formulas (P.25) and (P.26) holds true even when the relationships between **S** and **X** are nonlinear: monotonic or non-monotonic. We obtained some initial empirical evidence that support this claim.

From (P.23) and (P.26), we also can write:

$$Corr(X_i, Y_i) = [Corr(X_i, E(X_i|\mathbf{S}))^{-2} \quad \text{(P.27)}$$

Since the relationship between X_i and Y_i is always positive linear, which is clear from Equation (P.23) (i.e. a division of two positive quantities). In addition, the relationship between X_i and $E(X_i|\mathbf{S})$ is also always linear positive. This means that the range of possible values for the correlation in both sides of Equation (P.27) is between 0 and 1. This leads to the following inequality, which defines an *upper bound* for the correlation-based security measure in (6.9):

$$Corr(X_i, Y_i) \leq Corr(X_i, E(X_i|\mathbf{S})) \quad \text{(P.28)}$$

Notice the similarity of this inequality with the following security measures:

$$CC(\mathbf{X}, \mathbf{Y}) \leq CC(\mathbf{X}, \mathbf{S}) \quad \text{(P.29)}$$

and

$$\text{MSE}(X_i|\mathbf{Y}) \geq \text{MSE}(X_i|\mathbf{S}) \quad \text{(P.30)}$$

The "<" sign in (P.28) always holds except in two cases where the "<" sign turn into absolute equal sign "=". These two cases become obvious by combining the information in (P.27) with (P.28). These two cases happen when:

$$Corr(X_i, E(X_i|\mathbf{S})) = 1 \quad \text{(P.31)}$$

or

$$Corr(X_i, E(X_i|\mathbf{S})) = 0 \quad \text{(P.32)}$$

Equation (P.31) indicates 100% *deterministic* (linear) relationship between **S** and **X** while Equation (P.32) shows *no relationship* at all (i.e. complete random or independent variables) in the *linear sense* between **S** and **X**. Since we assumed the estimation mechanisms did a good job in learning the true conditional expectations and in modeling any nonlinearity may exist (so that the relationships between X_i and $E(X_i|\mathbf{S})$ is linear), Equation (P.32) suggests that X_i and $E(X_i|\mathbf{S})$ are actually independent.

Now we will present a third result that enables us to estimate the *regression line* between an original confidential attribute X_i as a dependent variable and its masked copy Y_i as independent variable (or the reverse) based on the characteristics of original datasets and before even the masking takes place. Then, we will use the first two propositions and the discussion follows them to prove the third proposition.

Proposition III:

When masking using the RBM approach, the relationship between an original confidential attribute X_i and its masked copy Y_i becomes linear (since the goal is to maximize data utility and the non-confidential attributes **S** link them together). This relationship takes the form of a line:

$$X_i = b_0 + b_1 Y_i \quad \text{(P.33)}$$

where the intercept b_0 can be calculated as:

$$b_0 = \eta E(X_i) = \frac{Var(r_i)}{Var(X_i)} E(X_i) \quad \text{(P.34)}$$

and the slope b_1 can be calculated as:

$$b_1 = Corr(X_i, Y_i) = 1 - \eta = 1 - \frac{Var(r_i)}{Var(X_i)} = \frac{Var(E(X_i|\mathbf{S}))}{Var(X_i)} \quad \text{(P.35)}$$

♦

Proof:

Bickel and Doksum (2001, Theorem 1.4.3 pp. 38) suggest that "*the unique best linear predictor*" takes the form of (using the masking terminology):

$$X_i = b_0 + b_1 Y_i \quad \text{(P.36)}$$

where the slope is:

$$b_1 = \frac{Cov(X_i, Y_i)}{Var(Y_i)} \quad \text{(P.37)}$$

and the intercept is:

$$b_0 = E(X_i) - b_1 E(Y_i) \quad \text{(P.38)}$$

We can rewrite the slope equation (Equation (P.37)) using the fact $Var(Y_i) = Var(X_i)$ and using some of the above derived results as follows:

$$b_1 = \frac{Cov(X_i, Y_i)}{\sqrt{[Var(Y_i)]^2}} \quad \text{(P.39)}$$

$$b_1 = \frac{Cov(X_i, Y_i)}{\sqrt{Var(X_i) Var(Y_i)}} \quad \text{(P.40)}$$

$$b_1 = Corr(X_i, Y_i) \quad \text{(P.41)}$$

$$b_1 = 1 - \eta = 1 - \frac{Var(r_i)}{Var(X_i)} \quad \text{(P.42)}$$

$$b_1 = \frac{Var(E(X_i|\mathbf{S}))}{Var(X_i)} \quad \text{(P.43)}$$

We can rewrite the intercept equation (Equation (P.38)) using the fact $E(Y_i) = E(X_i)$ and using some of the above derived results as follows:

$$b_0 = E(X_i) - b_1 E(X_i) \tag{P.44}$$

$$b_0 = E(X_i)(1 - b_1) \tag{P.45}$$

$$b_0 = E(X_i)(1 - Corr(X_i, Y_i)) \tag{P.46}$$

$$b_0 = E(X_i)(1 - (1 - \eta)) \tag{P.47}$$

$$b_0 = \eta E(X_i) \tag{P.48}$$

$$b_0 = \frac{Var(r_i)}{Var(X_i)} E(X_i) \tag{P.49}$$

This ends our proof.

♦

VITA

Mohammad Saad Al-Ahmadi

Candidate for the Degree of

Doctor of Philosophy

Thesis: ADAPTING MASKING TECHNIQUES FOR ESTIMATION PROBLEMS INVOLVING NON-MONOTONIC RELATIONSHIPS IN PRIVACY-PRESERVING DATA MINING

Major Field: Business Administration – Management Information Systems (MIS)

Biographical:

Education: Received Bachelor of Science degree in Computer Science from King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia in 1996; Received Master of Science degree in Telecommunications Management (MSTM) from Oklahoma State University, Stillwater, Oklahoma in 2001; Completed the requirements for the Doctor of Philosophy degree with a major in Management Information Systems at Oklahoma State University in July 2006.

Experience: Worked as Software Engineer in Saudi Iron & Steel Company (Hadeed) – Saudi Basic Industries Corporation (SABIC) in the Jubail Industrial city, Saudi Arabia from 1996 to 1998; Became a Certified Independent SAP/R3 ABAP/4 Consultant in 1997; Worked as Graduate Assistant in the Department of Accounting and Management Information Systems (ACCT & MIS), the College of Industrial Management (CIM) at King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia from 1998 to 1999; Worked as Web Programmer for iTradeFair.com, Stillwater, Oklahoma in 2001.

Professional Memberships: Association for Computing Machinery (ACM), Association for Information Systems (AIS), Decision Sciences Institute (DSI), Institute of Electrical and Electronics Engineers (IEEE), The Institute for Operations Research and the Management Sciences (INFORMS), The Honor Society of Phi Kappa Phi.