Chapter I
Introduction to Missing Data

ABSTRACT

In this chapter, the traditional missing data imputation issues such as missing data patterns and mechanisms are described. Attention is paid to the best models to deal with particular missing data mechanisms. A review of traditional missing data imputation methods, namely case deletion and prediction rules, is conducted. For case deletion, list-wise and pair-wise deletions are reviewed. In addition, for prediction rules, the imputation techniques such as mean substitution, hot-deck, regression and decision trees are also reviewed. Two missing data examples are studied, namely: the Sudoku puzzle and a mechanical system. The major conclusions drawn from these examples are that there is a need for an accurate model that describes inter-relationships and rules that define the data and that a good optimization method is required for a successful missing data estimation procedure.

INTRODUCTION

Datasets are frequently characterized by their incompleteness. There are a number of reasons why data become missing (Ljung, 1989). These include sensor failures, omitted entries in databases and non-response in questionnaires. In many situations, data collectors put in place firm measures to circumvent any incompleteness in data gathering. Nevertheless, it is unfortunate that despite all these efforts, data incompleteness remains a major problem in data analysis (Beunckens, Sotto, & Molenberghs, 2008; Schafer, 1997; Schafer & Olsen, 1998). The specific reason for the incompleteness of data is usually not known in advance, particularly in engineering problems. Consequently, methods for averting missing data are normally not successful. The absence of complete data then hampers decision-making processes because of the dependence of decisions on *full* information (Stefanakos & Athanassoulis, 2001; Marwala, Chakraverty, & Mahola, 2006).

In one way or another, most scientific, business and economic decisions are related to the information available at the time of making such decisions. For example, many business decisions are dependent on the availability of sales data and other information, while progresses in research are based on discovery of knowledge from various experiments and measured parameters. For example, in aerospace engineer-

ing, there are many fault detection mechanisms where the measured data are either partially corrupted or otherwise incomplete (Marwala & Heyns, 1998). In many applications, merely ignoring the incomplete record is not an optimal option because this may lead to biased results in statistical modeling resulting in, for example, a breakdown in machine automation or control. For this reason, it is essential to make decisions based on available data.

Most decision support systems such as the commonly used neural networks, support vector machines and many other computational intelligence techniques are predictive models that take observed data as inputs and predict an outcome (Bishop, 1995; Marwala & Chakraverty, 2006). Such models fail when one or more inputs are missing. Consequently, they cannot be used for decision-making purpose if the data variables are not complete. The end goal of the missing data estimation process is usually to make optimal decisions. To achieve this goal, appropriate approximations to the missing data need to be found. Once the missing variables values have been estimated, then pattern recognition tools for decision-making can be used.

The problem that missing data poses to a decision making process is more apparent in online applications where data have to be used nearly instantly after being obtained. In a situation where some variables are not available, it becomes difficult to carry on with the decision making process thereby stopping the application all together. In essence, the major challenge is that the standard computational intelligence techniques are not able to process input data with missing values. They cannot perform classification or regression if one of the variables is missing. Another major issue that is of concern here is that many missing data imputation techniques developed thus far are mainly suited for survey datasets. In this case, data analysts do have adequate time to study the reasons why data components are missing. However, in many engineering problems, missing data are usually required in real-time. Therefore, there is no time to understand why data components are missing. This calls for a development of robust methods that are effective for missing data estimation regardless of the cause of why the data are missing.

It is important to differentiate between missing data estimation and imputation. Missing data imputation essentially means *dealing* with missing data. This can include either by deleting that set with missing values or by using techniques such as list-wise deletion or estimating the missing values. Therefore, in this chapter missing data estimation is viewed as a sub-set of missing data imputation. It has generally been observed that in many data sets from the social sciences that missing data imputation is a valid way of dealing with missing data. This is because in social science the goal of the statistical analysis is to estimate statistical parameters such as averages and standard deviations. However, in the engineering field, where data are usually needed for manual decision support or automated decision support, the deleting of the entry is usually not an option. Therefore, in most cases, an estimation of the missing values has to be made. For that reason, in engineering problems, the term "missing data estimation" is more valid than the term "missing data imputation". Therefore, this book concentrates on the specific problem of missing data estimation rather than the general term missing data imputation.

This chapter discusses general approaches that have been used to deal with the problem of estimating missing values in an information system. Their advantages and disadvantages are discussed and some missing data imputation theory is discussed. This chapter concludes by discussing two classical missing data problems, which are the Sudoku puzzle and a problem that has been confronting engineers for some time, particularly in the aerospace sector. In discussing these two problems, key issues that are the subject of this book are identified.