Computational Statistics and Data Analysis 72 (2014) 77–91

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Inference for longitudinal data with nonignorable nonmonotone missing responses

Sanjoy K. Sinha [a,*], Amit Kaushal [b], Wenzhong Xiao [c]

[a] School of Mathematics and Statistics, Carleton University, Ottawa, ON, K1S 5B6, Canada
[b] Stanford Genome Technology Center, Stanford, CA 94305, USA
[c] Harvard Medical School, Boston, MA 02114, USA

ARTICLE INFO

Article history:
Received 6 April 2012
Received in revised form 31 October 2013
Accepted 31 October 2013
Available online 11 November 2013

Keywords:
False discovery rate
Importance sampling
Incomplete data
Linear mixed model
Longitudinal study
Maximum likelihood
Proteomics experiment

ABSTRACT

For the analysis of longitudinal data with nonignorable and nonmonotone missing responses, a full likelihood method often requires intensive computation, especially when there are many follow-up times. The authors propose and explore a Monte Carlo method, based on importance sampling, for approximating the maximum likelihood estimators. The finite-sample properties of the proposed estimators are studied using simulations. An application of the proposed method is also provided using longitudinal data on peptide intensities obtained from a proteomics experiment of trauma patients.

1. Introduction

In many clinical experiments, we often investigate changes in a specific characteristic in the participating individuals being observed repeatedly over time. Longitudinal studies are commonly performed for the investigation of individual changes over time and for exploring the effects of aging and other factors that are likely to influence the change. The repeated measurements from individuals in a longitudinal study are correlated by nature. To model the correlations among observations and also to investigate the individual (or group) effects on the responses, linear and generalized linear mixed models are often used. For an overview of mixed models and their applications to longitudinal data analysis, see Diggle et al. (2002).

We often encounter missing data in longitudinal experiments. Missing data occur whenever one or more of the sequences of measurements from experimental units or individuals are incomplete, in the sense that the desired measurements are not available, are lost due to technical problems, or otherwise not taken. The missingness in the longitudinal data often depends on the unobserved value of the outcome at a given assessment time, that is, the missing data are often nonignorable. When data are nonignorably missing, it is necessary to model the missing data mechanism for valid statistical inferences.

Analysis of missing data has been considered by many authors in the literature (e.g., (Diggle and Kenward, 1994; Ibrahim et al., 1999, 2001; Molenberghs and Verbeke, 2001; Sinha et al., 2010, 2011; Verbeke and Molenberghs, 2005; Wu et al., 2009; Xie, 2008; Yi and Cook, 2002); and many others). Little (1995) discusses techniques for modeling the data and the missing data mechanism simultaneously, and presents a number of examples to describe likelihood-based inferences via maximum

* Corresponding author.
E-mail address: sinha@math.carleton.ca (S.K. Sinha).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.10.027

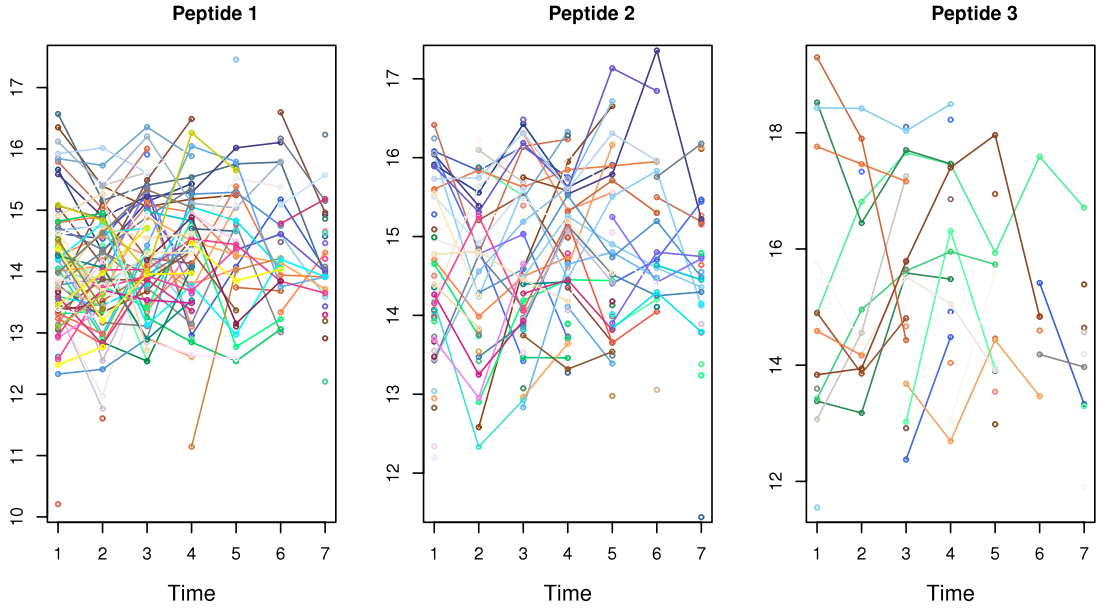

Fig. 1. Logarithm of peptide intensities at seven time points.

likelihood or Bayesian approaches. Little and Rubin (2002) review methods for analyzing data with various types of missing data mechanisms. Note that for analyzing longitudinal data with nonignorable and nonmonotone missing responses, a full likelihood method often requires intensive computation, especially when there are many follow-up times. Our goal is to find a suitable method that is computationally feasible and is also almost as efficient as the exact maximum likelihood method for analyzing the data.

This research was motivated by a time course proteomics experiment in which longitudinal measurements on peptide intensities were obtained from monocytes of a group of trauma patients being studied at Stanford Genome Technology Center. A monocyte is a type of white blood cell and is a part of the human body's immune system. Trauma patients experience monocyte dysfunctions during the development of multiple organ dysfunction syndrome (MODS). The focus of the analysis is on the expression levels of proteins in monocytes of trauma patients, where the abundance of protein molecules is determined by peptide intensities measured repeatedly over a certain period after the injury to a trauma patient.

Note that peptides are short chains of amino acids, and polymers of amino acids make up protein molecules. Peptide intensities were measured by mass spectrometry (MS), an analytical technique for the determination of the elemental composition of a sample or molecule. Mass spectrometry can be useful for determining what proteins are expressed in cancer cells that are not expressed in healthy cells, possibly leading to further understanding of the disease and to the development of drugs targeting these proteins.

A goal of the aforementioned proteomics experiment is to identify the peptides and corresponding proteins that are differentially expressed over time. Fig. 1 exhibits plots of natural logarithms of peptide intensities for three representative peptides measured at seven observation times 0.5, 1, 4, 7, 14, 21, and 28 days after the injury to a patient, labeled as times $1, \ldots, 7$, respectively. Measurements on peptide 1 were obtained from 97 patients, whereas measurements on peptides 2 and 3 were obtained from 88 and 30 patients, respectively. The peptide intensities for a given patient are shown with a distinct color in the three plots. The log-intensities for peptide 2 indicate a positive trend over time, whereas they show a curvilinear trend for peptide 3.

A feature of the proteomics experiment is that the repeated measurements from the patients are correlated. To model the correlations among these repeated observations, we consider using a linear mixed effects model with a block-diagonal covariance structure, as introduced in the next section.

Another important and also challenging feature of the proteomics study is that the data contain nonmonotone missing values, as the measurements on the peptide intensity were intermittently missing at the seven observation times. For example, for peptide 1, although measurements were obtained from 97 patients, only a subset of 75 patients had measurements at the first time-point; a different subset of 74 patients had measurements at the second time-point; a further different subset of 69 patients had measurements at the third time-point; and so on. Even though the overall proportion of observed responses decreases over time, the missing data pattern is nonmonotone, i.e., some patients' responses are missing at one occasion and observed at the next occasion.

The missing data mechanism in the proteomics experiment is considered nonignorable or not missing at random (NMAR) in that the missingness is due to the low abundance of the peptide intensities. Fig. 2 presents boxplots of peptide intensities (on the log2 scale) against five replicates (repeated measurements) from the trauma patients. Boxplot 1 shows the peptide