Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Coordinate ascent for penalized semiparametric regression on high-dimensional panel count data

Tong Tong Wu*, Xin He

Department of Epidemiology and Biostatistics, University of Maryland, College Park, MD 20742, United States

ARTICLE INFO

Article history:
Received 1 April 2011
Received in revised form 2 July 2011
Accepted 2 July 2011
Available online 14 July 2011

Keywords:
Estimating function
Generalized cross-validation
Lasso
Pseudo-objective function
Recurrent events
Semiparametric models
Survival data

ABSTRACT

This paper explores a fast algorithm to select relevant predictors for the response process with panel count data. Based on the lasso penalized pseudo-objective function derived from an estimating equation, the coordinate ascent accelerates the estimation of regression coefficients. The coordinate ascent algorithm is capable of selecting relevant predictors for underdetermined problems where the number of predictors far exceeds the number of cases. It relies on a tuning constant that can be chosen by generalized cross-validation. Our tests on simulated and real data demonstrate the virtue of penalized regression in model building and prediction for panel count data in ultrahigh-dimensional settings.

Published by Elsevier B.V.

1. Introduction

As a special type of longitudinal data, panel count data are frequently observed in prospective studies involving recurrent events (Thall, 1988; Thall and Lachin, 1988; Sun and Kalbfleisch, 1995; Sun and Wei, 2000; Sun, 2006). The primary interest of these studies is the time to a certain event. Subjects in the studies may experience several such events over time. In panel count data, instead of observing the actual recurrent event times, only the number of events that have occurred between consecutive observation/assessment times is available. The following are three important characteristics of panel count data. First, subjects can only be observed at finite discrete time points. Second, only the number of events occurring prior to each observation time is known. Third, the observation times may vary from subject to subject. It is clear that panel count data are a general form of recurrent event data. Researchers may observe this kind of data often in demographical studies, epidemiological studies, medical periodic follow-up studies, etc., since it is either impossible or impractical to maintain continuous observations of subjects in these studies. A typical example is the question on the number of kids in the US Census survey. Only the number of kids born between the consecutive censuses rather than birth time is asked. Another example is the bladder cancer study conducted by the Veterans Administration Cooperative Urological Research Group (Byar, 1980; Andrews and Herzberg, 1985; Wei et al., 1989), which we will analyze in this paper. In this study, subjects who were diagnosed with superficial bladder tumors were randomized into three treatment groups and followed for 53 months. At the beginning of the study, all the tumors were removed. Many patients had multiple recurrences of tumors during the study. At each clinical follow-up visit, the recurrent tumors were removed and the number of those recurrent tumors between visits was recorded. Sun and Wei (2000) displayed in Table 1 the data of 36 patients from the placebo group.

As pointed out by Sun and Wei (2000), if the event time is available, we will have recurrent event data in the examples mentioned above and can use the existing regression methods for recurrent event data (Prentice et al., 1981; Andersen and

* Corresponding author. Tel.: +1 301 405 3085.
E-mail address: ttwu@umd.edu (T.T. Wu).

0167-9473/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.csda.2011.07.003

Table 1
Results of simulation example for different covariate distributions and sample sizes with true parameter $\beta = (1, 1, 1, 0, \ldots, 0)$ based on 50 replicates. Standard errors of estimates appear in parentheses. The fifth column lists the prediction error ratio of the lasso penalized estimates to the unpenalized estimates.

Marginal dist.	(p, n)	ρ	λ	*PE* ratio	N_{nonzero}	N_{true}
$N(0, 1/2^2)$	(10, 100)	0	26.52 (0.52)	52.60% (8.80%)	3.32 (0.11)	2.76 (0.07)
$N(0, 1/2^2)$	(10, 100)	0.5	27.52 (0.50)	50.07% (7.23%)	3.34 (0.10)	2.84 (0.05)
$N(0, 1/2^2)$	(10, 200)	0	46.84 (0.88)	43.73% (4.46%)	3.22 (0.06)	3 (0)
$N(0, 1/2^2)$	(10, 200)	0.5	43.84 (1.09)	59.13% (6.76%)	3.56 (0.09)	3 (0)
$N(0, 1/2^2)$	(50, 200)	0	35.64 (0.92)	4.25% (1.36%)	5.04 (0.19)	3 (0)
$N(0, 1/2^2)$	(50, 200)	0.5	44.28 (0.86)	5.66% (4.41%)	4.06 (0.12)	3 (0)
$N(0, 1/2^2)$	(500, 200)	0	43.16 (0.80)	NA NA	6.82 (0.33)	2.94 (0.03)
$N(0, 1/2^2)$	(500, 200)	0.5	43.00 (1.20)	NA NA	6.78 (0.41)	3 (0)
Bernoulli (0.5)	(50, 200)	0	299.9 (0.10)	NA NA	3 (0)	3 (0)
Bernoulli (0.5)	(500, 200)	0	478.2 (4.70)	NA NA	2.98 (0.02)	2.98 (0.02)

Gill, 1982; Wei et al., 1989; Pepe and Cai, 1993; Cook and Lawless, 2007). To analyze panel count data, several methods have been proposed in the framework of counting processes. For example, Kalbfleisch and Lawless (1985) discussed the fitting of Markov models to panel count data. Sun and Kalbfleisch (1995) and Wellner and Zhang (2000) considered estimating the mean function of the underlying point process that generates panel count data. Sun and Wei (2000) and He et al. (2008) fitted semiparametric regression models for panel count data.

With the development of data collection technologies, datasets with large numbers of predictor variables are produced in a wide variety of scientific fields. Variable selection therefore has emerged as one of the most important topics in statistics in the past decade to handle such problems by identifying the most relevant predictors to the response. A common approach is to add penalties that shrink parameter estimates to 0. Many authors have worked on variable selection in regression and classification (Candes and Tao, 2007; Chen et al., 1998; Fan and Li, 2001; Fu, 1998; Hastie et al., 2004; Knight and Fu, 2000; Tibshirani, 1996; Wang et al., 2006; Wang and Shen, 2007). A recent trend for variable selection is to investigate the case when the number of predictors p is much larger than the number of observations n. The theoretical properties of lasso have been studied by many authors (Huang et al., 2008; Jia and Yu, 2010; Kim et al., 2008; Meinshausen et al., 2006; Meinshausen and Buehlmann, 2009; Ravikumar et al., 2008; Wainwright, 2009; Zhao and Yu, 2006). How to optimize the objective function and estimate the parameters is anther important issue when $p > n$. Available computational algorithms include, but are not limited to, solution path algorithms (Efron et al., 2004), interior-point method (Kim et al., 2007), linear programming (Candes and Tao, 2007; Wang et al., 2006), $l1$-constrained quadratic programming (Wainwright, 2009), and one-step local linear approximation (Zou and Li, 2008). Software packages are also developed for ultrahigh-dimensional data in linear regression and generalized linear regression, for instance, LARS (Efron et al., 2004), $l1$-MAGIC for Dantzig selector (Candes and Tao, 2007), $l1$-logreg (Kim et al., 2007), glmnet (Friedman et al., 2010), and Mendel (Lange et al., 2001, 2011). For the Cox proportional model, one can download the R packages coxnet (Simon et al., 2011) and CoxLARS (Wu, in press). However, variable selection for panel count data has not been considered much. To our best knowledge, the only paper considering variable selection for panel count data is Tong et al. (2009). In their paper, the authors developed a penalization method based on estimating functions for overdetermined problems with $p < n$ and proved the oracle property (Fan and Li, 2001) of the lasso-type penalized estimators when $p < n$. No work has been conducted for underdetermined cases with $p > n$.

Motivated by studies with large amount of information available, for example, genetic information, where the number of genetic markers p far exceeds the number of subjects n, we study the variable selection problem for high-dimensional penal count data in the current paper. In this setting, conventional methods of regression analysis have to add an additional prescreening step to reduce the dimension before the full analysis (Dudoit et al., 2002; Li et al., 2004, 2005). Wang and Shen (2007) argue against this arbitrary step of univariate feature selection and advocate imposing a lasso penalty (Tibshirani, 1996), which encourages sparse solutions. One immediate difficulty for lasso penalized estimation with high-dimensional data is the computational complexity. This computational hurdle renders high-dimensional problems intractable. How to handle the nondifferentiability of the lasso penalty is another computational problem. These two obstacles can be solved beautifully by applying coordinate descent/ascent algorithms (Friedman et al., 2007; Wu and Lange, 2008). Compared to the existing computational algorithms mentioned in the previous paragraph, coordinate descent/ascent is easy to understand and implement. The idea is to update one parameter at a time. In updating a single parameter, one can confine attention to the intervals to the left or right of the origin and ignore the kink in the lasso. This maneuver leads to fast reliable optimization

Download English Version:

https://daneshyari.com/en/article/416498

Download Persian Version:

https://daneshyari.com/article/416498

Daneshyari.com