Journal of Multivariate Analysis 167 (2018) 378–394

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Leveraging mixed and incomplete outcomes via reduced-rank modeling

Chongliang Luo [a], Jian Liang [b], Gen Li [c], Fei Wang [d], Changshui Zhang [d], Dipak K. Dey [a], Kun Chen [a,*]

[a] Department of Statistics, University of Connecticut, Storrs, CT, United States
[b] Department of Automation, Tsinghua University, Beijing, China
[c] Department of Biostatistics, Columbia University, New York, United States
[d] Department of Healthcare Policy and Research, Weill Cornell Medical School, Cornell University, New York, United States

ARTICLE INFO

Article history:
Received 29 August 2017

AMS subject classifications:
62H20
62H25

Keywords:
Generalized linear model
Integrative learning
Missing data
Multivariate regression

ABSTRACT

Multivariate outcomes with multivariate features of possibly high dimension are routinely produced in various fields. In many real-world problems, the collected outcomes are of mixed types, including continuous measurements, binary indicators and counts, and a substantial proportion of values may also be missing. Regardless of their types, these mixed outcomes are often interrelated, representing diverse reflections or views of the same underlying data generation mechanism. As such, an integrative multivariate model can be beneficial. We develop a mixed-outcome reduced-rank regression, which effectively enables information sharing among different prediction tasks. Our approach integrates mixed and partially observed outcomes belonging to the exponential dispersion family, by assuming that all the outcomes are associated through a shared low-dimensional subspace spanned by the features. A general singular value regularized criterion is proposed, and we establish a non-asymptotic performance bound for the proposed estimators in the context of supervised learning with mixed outcomes from an exponential family and under a general sampling scheme of missing data. An iterative singular value thresholding algorithm is developed for optimization with convergence guarantee. The effectiveness of our approach is demonstrated by simulation studies and an application on predicting health-related outcomes in longitudinal studies of aging.

1. Introduction

Multivariate outcomes/responses, or measurements of diverse and yet interrelated characteristics pertaining to a single set of subjects, together with multivariate features/predictors of possibly high dimension, are routinely produced in various fields of scientific research as well as in our daily lives. Many associated statistical learning problems fall into the domain of multivariate regression analysis, whose main objective is to build an accurate and interpretable predictive model for the outcomes of interest. In a human lung study of asthmatics, for example, the goal is to understand how overall functions of lung are influenced by microscopic lung airway structure [12], which amounts to associating discrete clinically-determined asthma severity status or continuous asthma quality of life scores from questionnaires with high-dimensional measurements of lung airway tree from a computed tomography scan. In an adolescent health study, annual hospitalization counts due

* Corresponding author.
E-mail address: kun.chen@uconn.edu (K. Chen).

https://doi.org/10.1016/j.jmva.2018.04.011
0047-259X/

to various causes such as disease, accidental injury, self-inflicted injury, etc., were collected for each school district in a state [8]; the interest was to understand how the various types of health-related risks, proxied by the hospitalization counts, were related to demographics, social-economic factors, academic performances, etc. In studies of aging on elderly subjects [45], continuous measurements of health, memory and sensation scores, dichotomous measurements of various medical conditions may be well predicted by subject demographics and records of medical history, life style and social behavior.

In the aforementioned examples, several types of outcomes, e.g., continuous, binary, and count data, may all be collected from the same cohort in the same study. We refer to such a collection of outcomes as *mixed outcomes* or *outcomes of mixed types*. In general, regardless of their types of measurements, such outcomes are expected to be related, as they commonly represent diverse reflections or different views of the underlying data generation mechanism. Therefore, an integrative learning of the mixed outcomes could be highly preferable in order to enable information sharing among different prediction tasks.

However, most existing multivariate techniques are only applicable for analyzing one type of outcomes at a time. For continuous outcomes, multivariate linear regression and its extensions have been extensively studied, e.g., ridge regression for overcoming multicollinearity [24], sparse regression for variable selection [19,37,40,52], and reduced-rank regression for dimension reduction [1,4,39]. Besides rank-constrained estimation, reduced-rank models can be realized through singular value regularization [10,27,33,36,51,54]. Recently, several authors considered sparse and reduced-rank models [5,9,11,48]. As soon as we step into the territory of non-Gaussian and/or non-linear analysis, the modeling of multivariate dependency becomes much more complicated. Vector generalized linear models were extended from their univariate counterparts based on a multivariate analogue of dispersion model family distributions, in which the correlation of the outcomes is explicitly modeled by an association matrix; see [44] for a comprehensive review on related topics. Yee and Hastie [50] proposed reduced-rank vector generalized linear models (RR-VGLM). She [42] further studied RR-VGLM and proposed an iterative algorithm with convergence guarantee. However, neither of them considered incomplete data and studied the theoretical properties of RR-VGLM. Yuan et al. [51] studied semiparametric and nonparametric low-rank models. There is also a rich literature on using sufficient dimension reduction to explore multivariate association; see [14,30,31] and the references therein.

Simultaneous statistical modeling of mixed outcomes is under-explored thus far. To the best of our knowledge, most of the existing approaches attempt to model the correlation among mixed outcomes in an explicit way; a drawback of such, however, is that it may not be applicable in high-dimensional settings. Cox and Wermuth [16] and Fitzmaurice and Laird [22] considered likelihood based methods by factorizing the joint distribution as marginal and conditional distributions. Prentice and Zhao [38] and Zhao et al. [53] used generalized estimating equations [32] to handle mixtures of continuous and discrete outcomes. Indeed, direct joint modeling of mixed outcomes is challenging due to the lack of convenient multivariate distributions, even when the number of outcomes is small. Another strategy is to induce multivariate dependency through some shared latent variable, conditional on which the outcomes are then assumed to be independent [17,41].

Our particular interest here is on generalizing and leveraging a reduced-rank matrix structure for modeling mixed multivariate outcomes with multivariate features, both of which are possibly of high dimension. From the publication of the seminal work by Anderson [1] several decades ago to the current era of big data, reduced-rank models have been very attractive, especially for modeling continuous multivariate data, in which the low-rank assumption of certain coefficient matrices conveniently captures the dependencies among the variables and systematically mitigates the curse of dimensionality. In the regression context, the low-rank assumption translates into a latent variable model, implying that all the outcomes are associated with the same small set of latent variables that are themselves linear functions of the original high-dimensional features/predictors. This elegant idea brings a genuine multivariate flavor to the model and, in an implicit way, induces and takes advantage of the dependency among the outcomes.

Giving the prevalence of big data, it is appealing to explore the use of reduced-rank structure in an integrative analysis of mixed outcomes, especially when the main goal is on dimension reduction and prediction. Similar ideas recently appeared in Udell et al. [49], in which the authors mainly focused on unsupervised learning and computation. Here, our goal is to provide a comprehensive study on a *mixed-response reduced-rank generalized linear regression model* (mRRR). The main contributions of this paper and some key features of our proposed approach are outlined as follows:

(i) Our approach integrates multivariate outcomes of mixed types belonging to an exponential dispersion family, and is able to conveniently handle incomplete data records in the multivariate statistical analysis.
(ii) We study the theoretical properties of mRRR in a general high-dimensional non-asymptotic framework. Finite-sample performance bounds are established for mRRR under a general setup of incomplete and mixed outcomes from an exponential family.
(iii) We provide a general, practical modeling framework and computational implementation for analyzing high-dimensional mixed outcomes, taking into account offset terms, fixed effects of control variables, and differential dispersion of the mixed-type outcomes. Our model and implementation can be readily extended to enable robust estimation, variable selection, etc.

The rest of the paper is organized as follows. In Section 2, we propose the mRRR framework for jointly analyzing mixed outcomes. In Section 3, we establish oracle inequalities for mRRR. A unified iterative algorithm is presented in Section 4. The performance gain by mRRR over several alternative modeling strategies is demonstrated via simulations in Section 5. In Section 6, we apply mRRR to build a joint predictive model of health conditions with data from studies of aging. Some concluding remarks are provided in Section 7.