Statistica Sinica **33** (2023), 1-30
doi:https://doi.org/10.5705/ss.202020.0401

PENALIZED REGRESSION FOR MULTIPLE TYPES OF MANY FEATURES WITH MISSING DATA

Kin Yau Wong[1], Donglin Zeng[2] and D. Y. Lin[2]

[1] *The Hong Kong Polytechnic University*
and [2] *The University of North Carolina at Chapel Hill*

Abstract: Recent technological advances have made it possible to measure multiple types of many features in biomedical studies. However, some data types or features may not be measured for all study subjects because of cost or other constraints. We use a latent variable model to characterize the relationships across and within data types and to infer missing values from observed data. We develop a penalized-likelihood approach for variable selection and parameter estimation and devise an efficient expectation-maximization algorithm to implement our approach. We establish the asymptotic properties of the proposed estimators when the number of features increases at a polynomial rate of the sample size. Finally, we demonstrate the usefulness of the proposed methods using extensive simulation studies and provide an application to a motivating multi-platform genomics study.

Key words and phrases: Adaptive lasso, factor models, integrative analysis, multi-modality data, multi-platform genomics studies, penalized regression.

1. Introduction

Modern biomedical studies often collect multiple types of data, or multi-modality data, on a large number of subjects. It is desirable to integrate such data because different modalities play unique roles in complex biological systems. For example, in the study of Alzheimer's disease, the integration of data on magnetic resonance imaging, positron emission tomography, and cerebrospinal fluid can yield more accurate disease classification (Zhang, Shen and Alzheimer's Disease Neuroimaging Initiative (2012)). In cancer research, different types of genomics data, such as DNA alterations, RNA expressions, and protein expressions, can be integrated to identify disease subtypes and predict patient survival (Shen, Olshen and Ladanyi (2009); Wang et al. (2012); Hoadley et al. (2014); Wong et al. (2019)).

Owing to cost or other constraints, certain features may not be measured on

Corresponding author: Kin Yau Wong, Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. E-mail: kin-yau.wong@polyu.edu.hk.