Fusing Heterogeneous Data Sets
Y. Song

Multiple high dimensional measurements from different platforms on the same biological system are becoming increasingly common in biological research. These different sources of measurements not only provide us with the opportunity of a deeper understanding of the studied system, but they also introduce some new statistical challenges. All these challenges are related to the heterogeneity of the data sets. The first type of heterogeneity is the type of data, such as metabolomics, proteomics and RNAseq data in genomics. These different omics data reflect the properties of the studied biological system from different perspectives. The second type of heterogeneity is the type of scale, which indicates the measurements are obtained at different scales, such as binary, ordinal, interval and ratio-scaled variables. Within this thesis, various data fusion approaches are developed to tackle either one or two types of heterogeneity that exist in multiple data sets.

In Chapter 2, we reviewed and compared various parametric and nonparametric extensions of principal component analysis (PCA) specifically geared for binary data. The special mathematical characteristics of binary data are taken into account from different perspectives in these different extensions of PCA. We explored their performance with respect to finding the correct number of components, overfitting, retrieving the correct low dimensional structure, variable importance, etc, using both realistic simulations of binary data as well as mutation, copy number aberrations (CNA) and methylation data of the GDSC1000 project. Our results indicate that if a low dimensional structure exists in the data, most of the methods can find it. We recommend to use the parametric logistic PCA model (projection based approach) if the probabilistic generating process can be assumed underlying the data, and to use the nonparametric Gifi model if such an assumption is not valid and the data is considered as given.

In Chapter 3, we developed a robust logistic PCA model via non-convex singular value thresholding. The promising logistic PCA model for binary data has an overfitting issue because of the used exact low rank constraint. We proposed to fit a logistic PCA model via non-convex singular value thresholding to alleviate the overfitting issue. An efficient majorization-minimization (MM) algorithm is implemented to fit the model and a missing value based cross validation (CV) procedure is introduced for the model selection. Furthermore, we re-expressed the logistic PCA model based on the latent variable interpretation of the generalized linear models (GLMs) on binary data. The latent variable interpretation of the logistic PCA model not only makes the assumption of low rank structure easier to understand, but also provides us a way to define signal to noise ratio (SNR) in the simulation of multivariate binary data. Our experiments on realistic simulations of imbalanced binary data and low SNR show that the CV error based model selection procedure is successful in selecting the proposed model. And the selected model demonstrates superior performance in recovering the underlying low rank structure compared to models with exact low rank constraint and convex nuclear norm penalty.

In the Chapter 4, we developed a generalized simultaneous component analysis (GSCA) model for the data fusion of binary and quantitative data sets. Simultaneous component analysis (SCA) model is one of the standard tools for exploring the underlying dependence structure present in multiple quantitative data sets measured on the same objects. However, it does not have any provisions when a part of the data are binary. To this end, we propose the GSCA model, which takes into account the distinct mathematical properties of binary and quantitative measurements in the maximum likelihood framework. In the same way as in the SCA model, a common low dimensional subspace is assumed to represent the shared information between these two distinct types of measurements. However, the GSCA model can easily be overfitted when a rank larger than one is used, which can lead to the problem that some of the estimated parameters can become very large. To achieve a low rank solution

and combat overfitting, we propose to use non-convex singular value thresholding. An efficient majorization algorithm is developed to fit this model with different concave penalties. Realistic simulations (low SNR and highly imbalanced binary data) are used to evaluate the performance of the proposed model in recovering the underlying structure. Also, a missing value based CV procedure is implemented for the model selection. We illustrate the usefulness of the GSCA model for exploratory data analysis of quantitative gene expression and binary CNA measurements obtained from the GDSC1000 data sets.

In Chapter 5, we proposed a penalized exponential family SCA (P-ESCA) model for the data fusion of multiple data sets with two types of heterogeneity. Multiple sets of measurements on the same objects obtained from different platforms may reflect partially complementary information of the studied system. However, the heterogeneity of such data sets introduces some new statistical challenges for their data fusion. First, the separation of information that is common across all or some of the data sets, and the information that is specific to each data set is problematic. Furthermore, these data sets are often a mix of quantitative and discrete (binary or categorical) data types, while commonly used data fusion methods require all data sets to be quantitative. Therefore, we proposed an exponential family simultaneous component analysis (ESCA) model to tackle the potential mixed data types problem of multiple data sets. In addition, a structured sparse pattern of the loading matrix is induced through a nearly unbiased group concave penalty to disentangle the global, local common and distinct information of the multiple data sets. An efficient MM algorithm is derived to fit the proposed model. Analytic solutions are derived for updating all the parameters of the model in each iteration, and the algorithm will decrease the objective function in each iteration monotonically. For model selection, a missing value based CV procedure is implemented. The advantages of the proposed method in comparison with other approaches are assessed using comprehensive simulations as well as the analysis of real data from a chronic lymphocytic leukaemia (CLL) study.

In Chapter 6, we considered various extensions of the developed P-ESCA model with respect to new penalties (element-wise, group-wise and their composition) and new model selection approach. Also, we remarked the potential of the semi-parametric XPCA model and non-parametric representation matrices approach in tackling the data sets of heterogeneous measurement scales. Furthermore, it is also interesting to generalize the P-ESCA model for prediction tasks or to tacking into account the experimental design underlining the used multiple data sets.