INTEGRATIVE ANALYSIS OF GENOMIC DATA WITH MULTI-VIEW PRINCIPAL COMPONENT ANALYSIS

JONATAN KALLUS and REBECKA JÖRNSTEN
University of Gothenburg, Gothenburg, Sweden
Chalmers University of Technology, Gothenburg, Sweden
e-mail: kallus@chalmers.se

Integrative analysis of several related high-dimensional data sets is increasingly relevant in molecular biology [3]. Multi-omics data calls for simultaneous integrative analysis both vertically (among groups of observations) and horizontally (among groups of variables). Two existing methods address general simultaneous vertical and horizontal integration: bi-modal OnPLS [1] and linked matrix factorization [2]. These methods aim at separating a signal that is global across all variable/observation groups from individual signals for each group. We address the problem of finding which subsets of groups that have a joint signal. Thus we address the harder problem of finding the jointness structure, as opposed to merely extracting the globally joint signal.

We define an objective function based on the singular value decomposition for each group. The optimum corresponds to low rank orthonormal bases for each group’s row and column space. A reparametrization to angles reduces the number of parameters and avoids optimization constraints. We define a fusing penalty term that forces different groups to share some basis vectors. A sparsity penalty is added to increase interpretability of results. The proposed method can be used for explorative analysis, e.g. integrative biclustering and finding driving combinations of variables across groups. Furthermore, uncovered jointness structure can be used for improving the separation of signal from noise. Joint signal between different types of genomic data corresponds to connections between mutations, epigenetics and phenotype that may be mechanistic, and thus have relevance for development of new drugs.

References