HDBIG-S2CCA Documentation

Release 1.0.0, 7/17/2016

© Copyright 2016, ShenLab at Indiana University School of Medicine
Contact: Li Shen (shenli@iu.edu) and/or Lei Du (dulei@nwpu.edu.cn)
Question or bug reporting: The HDBIG team (hdbig@iu.edu)

1. Introduction

Recent advances in brain imaging and high throughput genotyping and sequencing techniques enable new approaches to study the influence of genetic variation on brain structure and function. HDBIG is a collection of software tools for high dimensional brain imaging genomics. These tools are designed to perform comprehensive joint analysis of heterogeneous imaging genomics data. HDBIG-S2CCA is an HDBIG toolkit focusing on Structured Sparse Canonical Correlation Analysis (S2CCA). The current version includes matlab implementations of the structure-aware SCCA model (S2CCA), the GraphNet SCCA model (GN-SCCA), the Graph OSCAR SCCA (GOSC-SCCA) model, and the Absolute value based GraphNet SCCA model (AGN-SCCA). It can be applied to examine the association between genetic variations and imaging phenotypes. See below for a list of relevant papers.

- Du L*, Yan J*, Kim S, Risacher SL, Huang H, Inlow M, Moore JH, Saykin AJ, Shen L, for the ADNI (2014) A novel structure-aware sparse learning algorithm for brain imaging genetics. MICCAI'14: Med Image Comput Comput Assist Interv, Lecture Notes in Computer Science, 8675:329-336, Boston, MA, September 14-18, 2014. (*equal contribution).
- Du L, Yan J, Kim S, Risacher SL, Huang H, Inlow M, Moore JH, Saykin AJ, Shen L, for the ADNI. (2015) GN-SCCA: GraphNet sparse canonical correlation analysis for brain imaging genetics. BIH 2015 Special Session on Neuroimaging Data Analysis and Applications, Lecture Notes in Artificial Intelligence, 9250: 275-284, London, UK, 30 August - 2 September 2015.
- Du L, Huang H, Yan J, Kim S, Risacher SL, Inlow M, Moore JH, Saykin AJ, Shen L, for the Alzheimer's Disease Neuroimaging Initiative. (2016) Structured sparse CCA for brain imaging genetics via graph OSCAR. BMC Systems Biology. In press.
- Du L, Huang H, Yan J, Kim S, Risacher SL, Inlow M, Moore JH, Saykin AJ, Shen L, for the Alzheimer's Disease Neuroimaging Initiative. (2016) Structured Sparse Canonical Correlation Analysis for Brain Imaging Genetics: An Improved GraphNet Method. Bioinformatics. 2016 Jan 21,10.1093/bioinformatics/btw033.

2. License

HDBIG-S2CCA uses GNU General Public License (GPL). The license description is included in the software package. Please review and accept the license before installing HDBIG-S2CCA via any source.

3. Download

Software

- Available at http://www.iu.edu/~hdbig/S2CCA/

Documentation

- HTML: http://www.iu.edu/~hdbig/S2CCA/HDBIG-S2CCA-v1.0.0.html
- PDF: http://www.iu.edu/~hdbig/S2CCA/HDBIG-S2CCA-v1.0.0.pdf

4. Folder Structure and Demo Examples

The package "HDBIG-S2CCA-v1.0.0.zip" consists of five subfolders.

- data: Synthetic X, Y
- example: Example functions for demonstration
- data_preprocessing: Functions for data preprocessing
- scca_code: the Matlab function(s) for the four SCCA models (Please see "Methods" and references in "Introduction" for more details)
- license: The license description.

All the functions described in the following "Methods" section are located in "scca_code". The current version only supports Matlab. For each of these functions, we have a corresponding example function for demonstration. These examples can be found under "example". Within each example, we perform the following steps

- Load synthetic data
- Data quality control (such as removing empty entries)
- Data Normalization (Let mean = 0 and standard deviation = 1)
- Running the corresponding SCCA model and return three outputs: two canonical loadings for X, Y respectively and the correlations coefficients between them

5. Methods

In this package, four state-of-the-art SCCA models are included.

- S2CCA: Structure-aware Sparse Canonical Correlation Analysis

- GN-SCCA: GraphNet Sparse Canonical Correlation Analysis

- GOSC-SCCA: Sparse Canonical Correlation Analysis via Graph OSCAR

- AGN-SCCA: Absolute value based GN-SCCA

Sparse learning using CCA has received substantial attention during the past few years. Using different penalty functions, these SCCA models can identify different structures, including meaningful structures underlying human genome and brain.

Example Usage:

- [u,v,ecorr] = s2cca(X, Y, group_Info, paras);

- [u,v,ecorr] = gn_scca(X, Y, paras);
- [u,v,ecorr] = gosc_scca(X, Y, paras);
- [u,v,ecorr] = agn_scca(X, Y, paras);

X is n*p matrix and Y is n*q matrix. For the S2CCA, the "group_Info" contains the group information (prior knowledge) of X and Y respectively. "paras" is the regularization parameters "paras" control the strength of the penalty terms.