Group Meeting

Ilaria Bonavita

March 28, 2017

MAX-PLANCK-GESELLSCHAFT

Max Planck Institute
of Psychiatry
German Research Institute of Psychiatry

The method

- Canonical Correlation Analysis
- Nonparametric Canonic Correlation Analysis
- NCCA practical implementation

First experiments

- Replication of GWAS of CHD
- eQTL analysis based approach

CCA

Canonical Correlation Analysis [Hotelling, 1936] - in words

Classical technique to identify and quantify the association between two sets of variables (views). It searches for the linear combination of the original variables having maximal correlation.

Canonical Correlation Analysis [Hotelling, 1936] - in words

Classical technique to identify and quantify the association between two sets of variables (views). It searches for the linear combination of the original variables having maximal correlation.

Further pairs of maximally correlated linear combinations are chosen such that they are orthogonal to those already identified.

Canonical Correlation Analysis [Hotelling, 1936] - in words

Classical technique to identify and quantify the association between two sets of variables (views). It searches for the linear combination of the original variables having maximal correlation.

Further pairs of maximally correlated linear combinations are chosen such that they are orthogonal to those already identified.

The pairs of linear combinations are called canonical variables and their correlations canonical correlations.

Motivation for CCA

Motivation for CCA

- Reveal common variability in multiple views of the same phenomenon.

Motivation for CCA

- Reveal common variability in multiple views of the same phenomenon.
- Can provide insight into the data.

Motivation for CCA

- Reveal common variability in multiple views of the same phenomenon.
- Can provide insight into the data.
- Suppress view-specific noise factors.

Motivation for CCA

- Reveal common variability in multiple views of the same phenomenon.
- Can provide insight into the data.
- Suppress view-specific noise factors.
- Induce features that capture some of the information of the other view.

Motivation for CCA

- Reveal common variability in multiple views of the same phenomenon.
- Can provide insight into the data.
- Suppress view-specific noise factors.
- Induce features that capture some of the information of the other view.
- Has been applied to problems in computer vision, natural language processing, speech recognition, genomic etc.

Motivation for CCA

- Reveal common variability in multiple views of the same phenomenon.
- Can provide insight into the data.
- Suppress view-specific noise factors.
- Induce features that capture some of the information of the other view.
- Has been applied to problems in computer vision, natural language processing, speech recognition, genomic etc.
- Several extensions (nonlinear, nonparametric, generalized...) of CCA proposed.

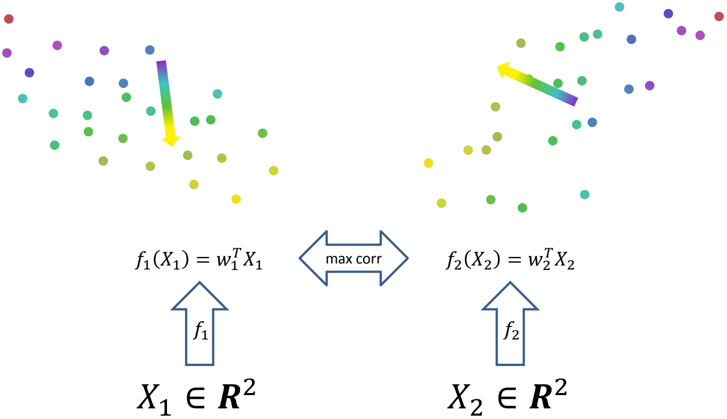
Problem formulation

$\mathbf{X} \in \mathbb{R}^{D_x}, \quad \mathbf{Y} \in \mathbb{R}^{D_y}$ two multi-dimensional vectors (views)

$$\begin{aligned} \max_{\mathbf{W}_1,\mathbf{W}_2} \quad & \mathbb{E}\big[(\mathbf{W}_1^T\mathbf{X})^T(\mathbf{W}_2^T\mathbf{Y})\big] \\ \text{s.t.} \quad & \mathbb{E}\big[(\mathbf{W}_1^T\mathbf{X})(\mathbf{W}_1^T\mathbf{X})^T\big] = \mathbb{E}\big[(\mathbf{W}_2^T\mathbf{Y})(\mathbf{W}_2^T\mathbf{Y})^T\big] = \mathbf{I}. \end{aligned} \tag{1}$$

$\mathbf{W}_1 \in \mathbb{R}^{D_x \times L}, \quad \mathbf{W}_2 \in \mathbb{R}^{D_y \times L}$ projection matrices.
$L \leq min\{D_x, D_y\}$ dimension of the transformed features.

CCA illustration



$f_1(X_1) = w_1^T X_1$

max corr

$f_2(X_2) = w_2^T X_2$

f_1

f_2

$X_1 \in \boldsymbol{R}^2$

$X_2 \in \boldsymbol{R}^2$

Two views of each instance have the same color

CCA solution

The solution can be expressed in terms of SVD of the matrix $\mathbf{T} = \mathbf{\Sigma}_{xx}^{-1/2}\mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1/2}$:

$$(\mathbf{W}_1, \mathbf{W}_2) = (\mathbf{\Sigma}_{xx}^{-1/2}\mathbf{U}, \mathbf{\Sigma}_{yy}^{-1/2}\mathbf{V}) \tag{2}$$

$\mathbf{\Sigma}_{xy} = \mathbb{E}[\mathbf{X}\mathbf{Y}^T] \approx \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{y}_i^T$, $\mathbf{\Sigma}_{xx}$, $\mathbf{\Sigma}_{yy}$,

$\mathbf{U} \in \mathbb{R}^{D_x \times L}$, $\quad \mathbf{V} \in \mathbb{R}^{D_y \times L}$ top L left and right singular vectors of $\mathbf{T}$.

CCA shortcoming: restriction to linear mapping

But many real-world multi-view datasets show highly nonlinear reletionship. Need to generalise the method...

CCA shortcoming: restriction to linear mapping

But many real-world multi-view datasets show highly nonlinear reletionship. Need to generalise the method...

Find $\mathbf{f} : \mathbb{R}^{D_x} \to \mathbb{R}^L$, $\mathbf{g} : \mathbb{R}^{D_y} \to \mathbb{R}^L$ solving

$$\begin{aligned} \max_{\mathbf{f}\in\mathcal{A},\mathbf{g}\in\mathcal{B}} \quad & \mathbb{E}\big[(\mathbf{f}(X)^T\mathbf{g}(Y))\big] \\ \text{s.t.} \quad & \mathbb{E}\big[(\mathbf{f}(X)\mathbf{f}(X)^T\big] = \mathbb{E}\big[(\mathbf{g}(Y)\mathbf{g}(Y)^T\big] = \mathbf{I}. \end{aligned} \tag{3}$$

$\mathcal{A}$ and $\mathcal{B}$ two families of measurable functions.

Nonparametric CCA

Nonparametric CCA

[Michaeli et al. 2016]

$\mathcal{A}$, $\mathcal{B}$ set of all (nonparametric) measurable functions of X and Y. We can rewrite (3) as an optimization problem over the Hilbert spaces:

$$\mathcal{H}_x = \{q : \mathbb{R}^{D_x} \to \mathbb{R} \mid \mathbb{E}[q^2(X)] < \infty\},$$
$$\mathcal{H}_y = \{u : \mathbb{R}^{D_y} \to \mathbb{R} \mid \mathbb{E}[u^2(Y)] < \infty\},$$

endowed with the inner products $\langle q, r\rangle_{\mathcal{H}_x} = \mathbb{E}[q(X)r(X)]$ and $\langle u, v\rangle_{\mathcal{H}_y} = \mathbb{E}[u(X)v(X)]$.

Nonparametric CCA

Then,

$$\mathbb{E}[f_i(X)g_i(Y)] = \int f_i(\mathbf{x})\Big(\int g_i(\mathbf{y})s(\mathbf{x},\mathbf{y})p(\mathbf{y})d\mathbf{y}\Big)p(\mathbf{x})d\mathbf{x} = \langle f_i, \mathcal{S}g_i\rangle_{\mathcal{H}_x},$$

where

$$s(\mathbf{x},\mathbf{y}) = \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}, \tag{4}$$

and $\mathcal{S} : \mathcal{H}_y \to \mathcal{H}_x$ operator defined by $(\mathcal{S}u)(\mathbf{x}) = \int u(\mathbf{y})s(\mathbf{x},\mathbf{y})p(\mathbf{y})d\mathbf{y}$.

Nonparametric CCA

The *nonlinear* CCA problem (3) can be expressed as:

$$\max_{\substack{\langle f_i, f_j \rangle_{\mathcal{H}_x} = \delta_{ij}, \\ \langle g_i, g_j \rangle_{\mathcal{H}_y} = \delta_{ij}}} \quad \sum_{i=1}^{L} \langle \mathcal{S} g_i, f_i \rangle_{\mathcal{H}_x}. \tag{5}$$

When $\mathcal{S}$ is compact solution can be expressed in terms of its SVD, then the optimal projections are the left and right singular functions $\psi_i \in \mathcal{H}_x$, $\phi_i \in \mathcal{H}_y$:

$$f_i(\mathbf{x}) = \psi_i(\mathbf{x}), \quad g_i(\mathbf{y}) = \phi_i(\mathbf{y}), \tag{6}$$

$\sigma_1 + \cdots + \sigma_L$ is the maximal objective value, with $\sigma_1 \geq \sigma_2 \ldots$ singular values of $\mathcal{S}$.

Nonparametric CCA: interesting interpretations and keypoints

1. $\log s(\mathbf{x}, \mathbf{y})$ is the *pointwise mutual information* PMI between X and Y.
2. $\mathcal{S}$ corresponds to the *optimal prediction* (in *MSE* sense) of one view based on the other, as $(\mathcal{S}g_i)(\mathbf{x}) = \mathbb{E}[g_i(Y)|X = \mathbf{x}]$ and $(\mathcal{S}^* f_i)(\mathbf{x} = \mathbf{y}) = \mathbb{E}[f_i(X)|Y = \mathbf{y}]$.
3. NCCA solution can be expressed in terms of the *eigen-decomposition* of a certain operator, defined via the population density.
4. Similar to *kernel* CCA but do not require computing the inverse of any kernel matrices and solves a sparse eigenvalue systems.

NCCA practical implementation

$\langle \mathcal{S}g_i, f_i \rangle = \mathbb{E}[(\mathcal{S}g_i)(X) f_i(X)] \approx \frac{1}{N}\sum_{l=1}^{N} (\mathcal{S}g_i)(\mathbf{x}_l) f_i(\mathbf{x}_l)$.
$(\mathcal{S}g_i)(\mathbf{x}_l) = \mathbb{E}[s(\mathbf{x}_l, Y) g_i(Y)] \approx \frac{1}{N}\sum_{m=1}^{N} s(\mathbf{x}_l, \mathbf{y}_m) g(\mathbf{y}_m)$.
$\mathbf{S} = [s(\mathbf{x}_l, \mathbf{y}_m)]$, $\mathbf{f}_i = \frac{1}{\sqrt{N}}(f_i(\mathbf{x}_1), \ldots, f_i(\mathbf{x}_N))^T$ (simil. $\mathbf{g}_i$), then, NCCA problem (5) become:

$$\max_{\substack{\mathbf{f}_i^T \mathbf{f}_j = \delta_{ij}, \\ \mathbf{g}_i^T \mathbf{g}_j = \delta_{ij}}} \quad \frac{1}{N}\sum_{i=1}^{L} \mathbf{f}_i^T \mathbf{S} \mathbf{g}_i. \tag{7}$$

The solution is obtained computing the SVD of $\mathbf{S}$. The optimal $\mathbf{f}_i$ and $\mathbf{g}_i$ are the top L singular vectors of $\mathbf{S}$.

NCCA practical implementation

$\mathbf{S} \approx \left[\frac{\hat{p}(\mathbf{x},\mathbf{y})}{\hat{p}(\mathbf{x})\hat{p}(\mathbf{y})}\right]$, $\mathbf{f}_i = \sqrt{N}\mathbf{U}_i$, $\mathbf{g}_i = \sqrt{N}\mathbf{V}_i$.

$p(\mathbf{x},\mathbf{y})$ estimated from the set of training data $\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^N$ with *kernel density estimates* (KDEs):

$$\hat{p}(\mathbf{x},\mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N} w(||\mathbf{x}-\mathbf{x}_i||^2/\sigma_x^2 + ||\mathbf{y}-\mathbf{y}_i||^2/\sigma_y^2),$$

$w(\cdot)$ Gaussian kernel and σ_x and σ_y kernel widths.

Experiments

Genome-wide association study of multiple congenital heart disease phenotypes with NCCA

1504 cases + 3553 controls
432097 variants
Binary phenotype

Genome-wide association study of multiple congenital heart disease phenotypes with NCCA

1504 cases + 3553 controls
432097 variants
Binary phenotype

NCCA is able to identify association between significant SNP and phenotype.

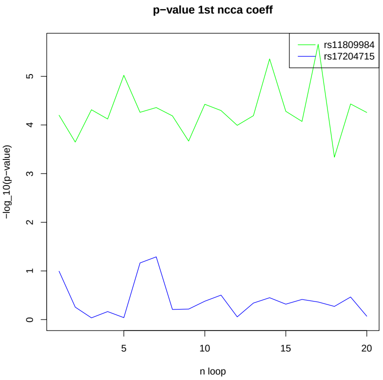
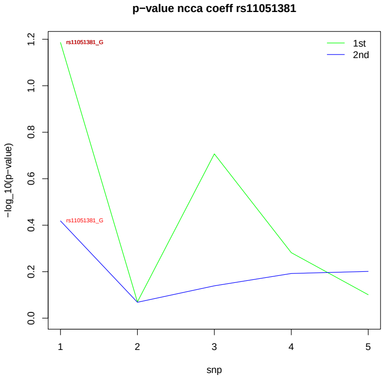
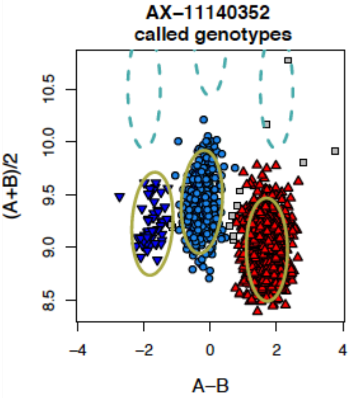


Figure: *p-value for the fist nonparametric canonical correlation coefficient. Green: significant SNP, blue: non significant SNP*

Genome-wide association study of multiple congenital heart disease phenotypes with NCCA

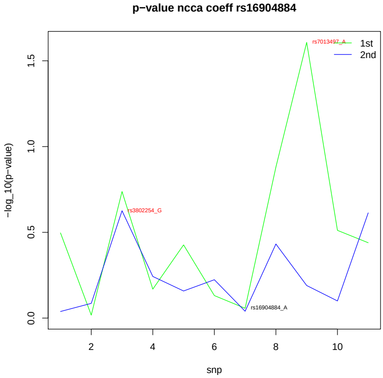
NCCA and GWAS associations are consistent.

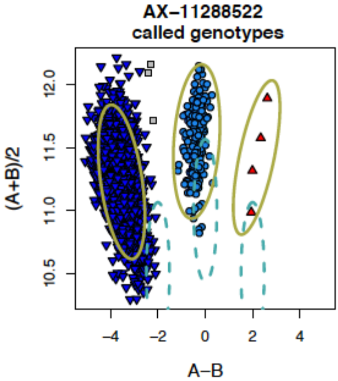




Genome-wide association study of multiple congenital heart disease phenotypes with NCCA

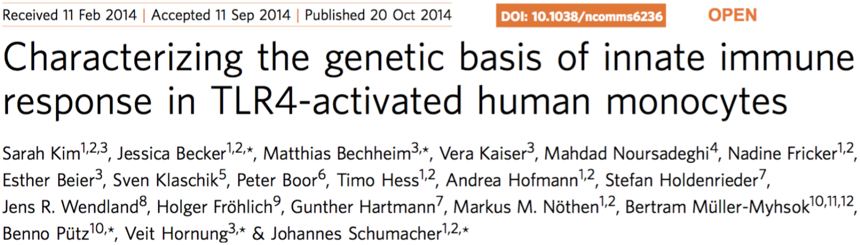
NCCA and GWAS associations are not consistent...





...binary phenotype might be not suitable for this kind of methods.

SNP genotype - gene expression levels association analysis



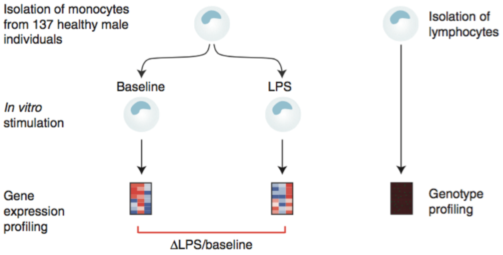
Received 11 Feb 2014 | Accepted 11 Sep 2014 | Published 20 Oct 2014 DOI: 10.1038/ncomms6236 OPEN

Characterizing the genetic basis of innate immune response in TLR4-activated human monocytes

Sarah Kim[1,2,3], Jessica Becker[1,2,*], Matthias Bechheim[3,*], Vera Kaiser[3], Mahdad Noursadeghi[4], Nadine Fricker[1,2], Esther Beier[3], Sven Klaschik[5], Peter Boor[6], Timo Hess[1,2], Andrea Hofmann[1,2], Stefan Holdenrieder[7], Jens R. Wendland[8], Holger Fröhlich[9], Gunther Hartmann[7], Markus M. Nöthen[1,2], Bertram Müller-Myhsok[10,11,12], Benno Pütz[10,*], Veit Hornung[3,*] & Johannes Schumacher[1,2,*]

- Toll-like receptors (TLRs) play a pivotal role in antimicrobial defense.
- Mutations and polymorphisms in TLR and TLR- signalling genes have been shown to confer susceptibility to many infectious and inflammatory diseases.
- Comparing unstimulated versus TLR4-stimulated monocytes revealed 1471 eQTLs unique to TLR4 stimulation.

Design of experiment and data



- 137 healthy individuals.
- Genotype profiles.
- Gene expressions level across:
  - 3 time points (baseline, 90 mins, 6 hours).
  - 3 treatments.

  only 1 tp, 1 treatment (LPS) included in the paper.

Planned steps

Using NCCA for association testing:

Planned steps

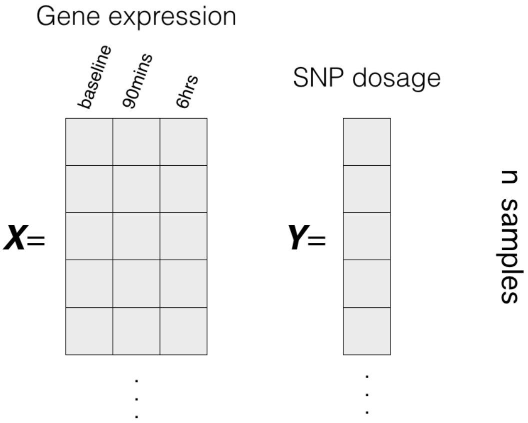
Using NCCA for association testing:

- Single SNP - single gene (across 3 tp).

Planned steps

Using NCCA for association testing:

- Single SNP - single gene (across 3 tp).



Gene expression

baseline

90mins

6hrs

SNP dosage

$\boldsymbol{X}$=

$\boldsymbol{Y}$=

n samples

Planned steps

Using NCCA for association testing:

- Single SNP - single gene (across 3 tp).
- Single SNP - multiple genes (across 3 tp).

Planned steps

Using NCCA for association testing:

- Single SNP - single gene (across 3 tp).
- Single SNP - multiple genes (across 3 tp).
- Subsets of SNP - subsets of genes, using prior biological knowoledge.

Planned steps

Using NCCA for association testing:

- Single SNP - single gene (across 3 tp).
- Single SNP - multiple genes (across 3 tp).
- Subsets of SNP - subsets of genes, using prior biological knowoledge.

Test the method on other genetics datasets (e.g. SNP - multi phenotype association on dyslexia data).

Thanks!