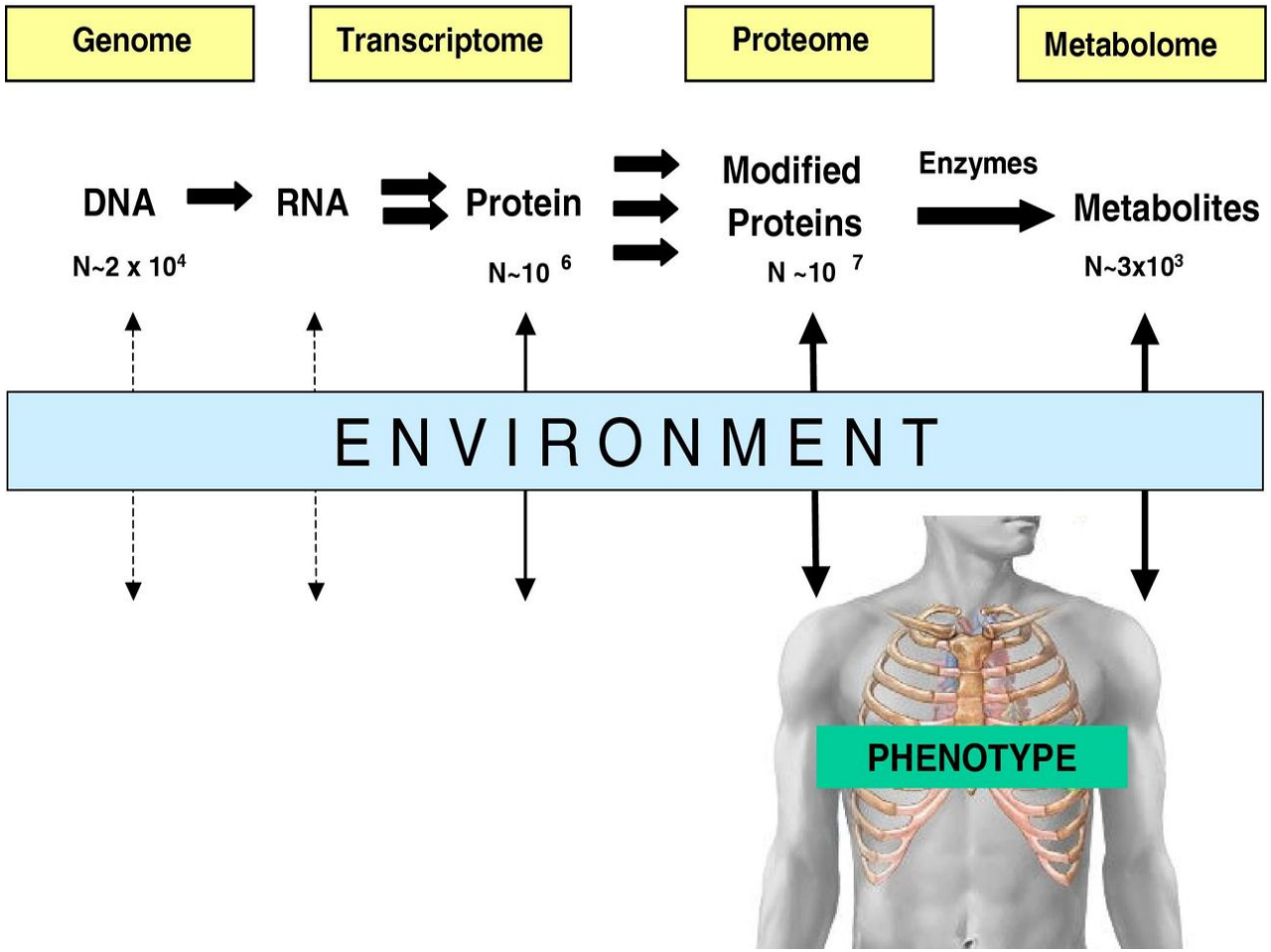
Bayesian Canonical Correlation Analysis for Multiple Groups

Lin Qiu, Lin Lin, Vernon M. Chinchilli

Department of Public Health Sciences and Statistics
Pennsylvania State University

July 29, 2019

Background



(Newgard,2017)

Omics data

- GWAS: Genome-wide association studies (GWAS) have thus far explained only a small fraction of polygenic diseases like, diabetes and cardiovascular diseases.
- Interest: genomics,epigenomics,transcriptomics,proteomics,metabolomics and micobiomics, each of these data types provides a different snapshot of the underlying biological system, and combining multiple data types has shown to be very valuable for the improvement of medical and health care services and for the implementation of preventive and precision medicine (Pan et al., 2010; Chen et al.,2013);
- Aims: Integrative statistical and computational analysis tools are needed;

Omics data

Issues to deal with:

- small sample size and large dimension (small n and large p);
- data types are different (continuous vs count);
- data structures are sparse;

Outline for section 1

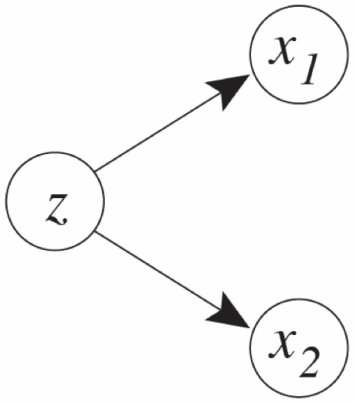
Canonical Correlation Analysis

We study the task by first modeling dependencies between two data sets of co-occurring or paired samples (x,y).The underlying assumption is that variation within either data set alone is more noisy, or at least less interesting than variation that is in common.This task has been classically solved by Cannonical Correlation Analysis (CCA) (Hotelling, 1936).

Suppose we have two random vectors $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_p})^\mathbf{T}$ and $\mathbf{y} = (\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_q})^\mathbf{T}$.CCA aims to find two projection directions, $\boldsymbol{u} \in \mathbb{R}^p$ and $\boldsymbol{v} \in \mathbb{R}^q$ so that

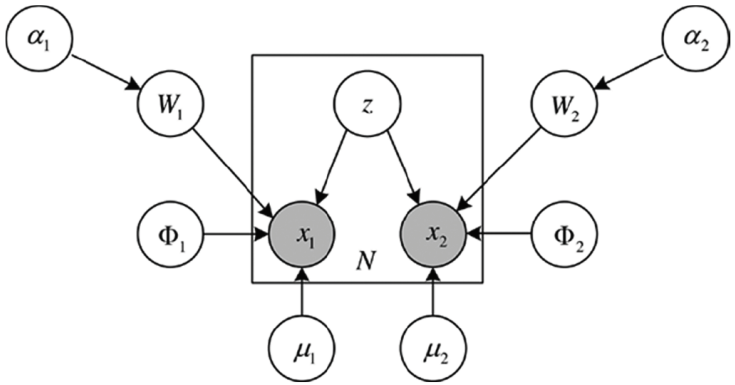
- $\rho = \arg\max_{\boldsymbol{u},\boldsymbol{v}} Corr(\boldsymbol{u}^T\boldsymbol{x}, \boldsymbol{v}^T\boldsymbol{y}) = \arg\max_{\boldsymbol{u},\boldsymbol{v}} \frac{\boldsymbol{u}^T\Sigma_{\boldsymbol{xy}}\boldsymbol{v}}{\sqrt{(\boldsymbol{u}^T\Sigma_{\boldsymbol{xx}}\boldsymbol{u})(\boldsymbol{v}^T\Sigma_{\boldsymbol{yy}}\boldsymbol{v})}}$, where$\Sigma_{\boldsymbol{xx}}, \Sigma_{\boldsymbol{yy}}, \Sigma_{\boldsymbol{xy}}$ are covariance and cross-covariance matrices.
- This maximization is equivalent to $\max_{\boldsymbol{u},\boldsymbol{v}} \boldsymbol{u}^T\Sigma_{\boldsymbol{xy}}\boldsymbol{v}$ subject to $\boldsymbol{u}^T\Sigma_{\boldsymbol{xx}}\boldsymbol{u} = 1$ and $\boldsymbol{v}^T\Sigma_{\boldsymbol{yy}}\boldsymbol{v} = 1$.
- The new variables $\eta_1 = \boldsymbol{u_1}^T\boldsymbol{x}, \xi_1 = \boldsymbol{v_1}^T\boldsymbol{y}$ are called the first pair of canonical variables or latent variables and $\rho_1 = Corr(\eta_1, \xi_1)$ is the first cannonical correlation.

Probabilistic Interpretation of CCA



$$x_1 = W_1 z + \mu_1 + \varepsilon_1$$
$$x_2 = W_2 z + \mu_2 + \varepsilon_2$$

(a) 1a:PCCA



(b) 1b:BCCA

- 1a treats CCA as a generative model (Bach and Jordan,2005).
- 1b shows the hierarchical Bayesian CCA model (Wang,2007).

Inter-battery factor analysis(IBFA)

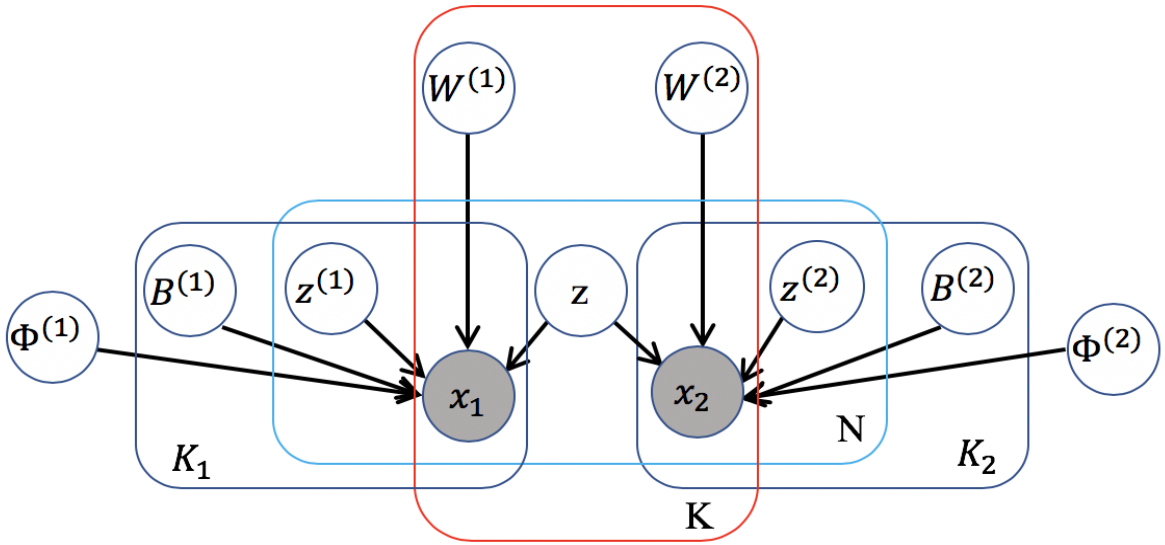


Figure: 1c:IBFA

$$\boldsymbol{x}^{(m)} \sim N(\boldsymbol{W}^{(m)}\boldsymbol{z} + \boldsymbol{B}^{(m)}\boldsymbol{z}^{(m)}, \boldsymbol{\Sigma}^{(m)}).$$

- 1c, IBFA model has view-specific latent variable (Browne,1979;Klami and Kaski,2006,2008).

Literature review

Automatic relevance determination(ARD;Neal,1996) prior

ARD is a Normal-Gamma prior for the projection weights.
$ARD(\boldsymbol{W}^{(m)}|\alpha_0,\beta_0) = \prod_{k=1}^{K} p(\boldsymbol{w}_k^{(m)}|\alpha_k^{(m)})p(\alpha_k^{(m)}|\alpha_0,\beta_0)$,
$\alpha_k^{(m)} \sim Gamma(\alpha_0,\beta_0)$,
$\boldsymbol{W}_k^{(m)} \sim N(\boldsymbol{0},(\alpha_k^{(m)})^{-1}\boldsymbol{I})$.
The number of components is automatically selected by pushing $\alpha_k^{(m)}$ of unnecessary towards infinity (Klami and Kaski, 2007;Nagarajan,2008).

Group-wise Sparsity(Klami et al.,2013)

$y = [\boldsymbol{z};\boldsymbol{z^{(1)}};\boldsymbol{z^{(2)}}] \in \mathbb{R}^{K_c \times 1}$, where $K_c = K + K_1 + K_2$,
IBFA model becomes to
y$\sim N(\boldsymbol{0},\boldsymbol{I})$, $\boldsymbol{x} \sim N(\boldsymbol{W}\boldsymbol{y},\boldsymbol{\Sigma}), \boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}, D = D_1 + D_2, \boldsymbol{W} \in \mathbb{R}^{D\times K_c}$.

Group Bayesian CCA(GBCCA)

A group may correspond to one view of the same set of objects, one of many data sets tied by co-occurrence, or a set of alternative variables collected from statistics table to measure one property of interest.

- Aims: Find components that capture joint variability between data sets instead of individual variables. Given a collection of $\boldsymbol{X}_1, ..., \boldsymbol{X}_M$ of M data sets of dimensionalities $D_1, ..., D_M$, the task is to find $K < \Sigma_{m=1}^{M} D_m$ components that describe the dependencies between data sets $\boldsymbol{X}_m$.

Outline for section 2

Model:GBCCA, DGBCCA(Dynamic GBCCA)

Assume a collection of obervations $y_i \in \mathbb{R}^D$ for $i = 1, ..., N, \boldsymbol{Y} \in \mathbb{R}^{N \times D}$, and the D variables are split into M groups or subsets. For notational simplicity, assume the first D_1 variables correspond to the first group G_1, D_M variables to G_M. Let $\boldsymbol{Y} = [\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(M)}]$, $\boldsymbol{x}^{(1)} \in \mathbb{R}^{N \times D_1}$ and $\boldsymbol{x}^{(M)} \in \mathbb{R}^{N \times D_M}$ for the N observations, $\boldsymbol{W} \in \mathbb{R}^{D \times K}$, $\boldsymbol{Z} \in \mathbb{R}^{N \times K}$ The model for the m^{th} group of the i^{th} sample is

$$\boldsymbol{x}_i^{(m)} \sim N(\boldsymbol{Z}_i \boldsymbol{W}^{(m)^T}, \tau_m^{-1} \boldsymbol{I}),$$

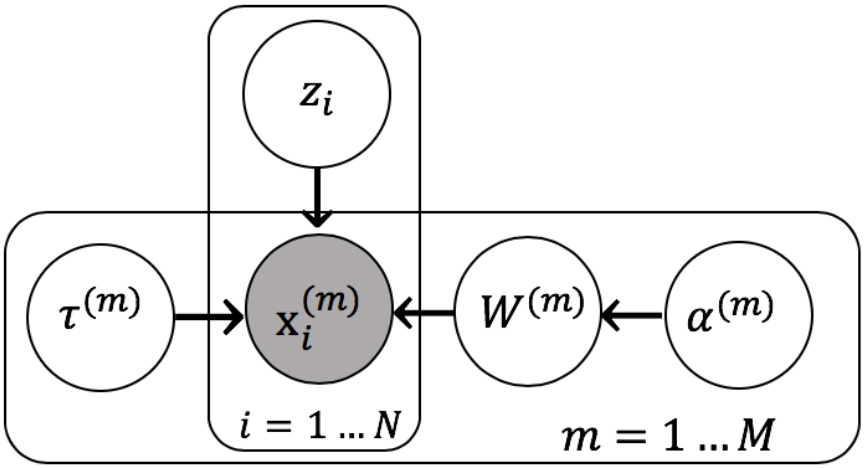
$$\boldsymbol{W}_{:,k}^{(m)^T} \sim N(\boldsymbol{0}, \boldsymbol{\alpha}_k^{(m)-1} \boldsymbol{I}),$$

$$\boldsymbol{\alpha}_k^{(m)} \sim Gamma(\alpha_0, \beta_0),$$

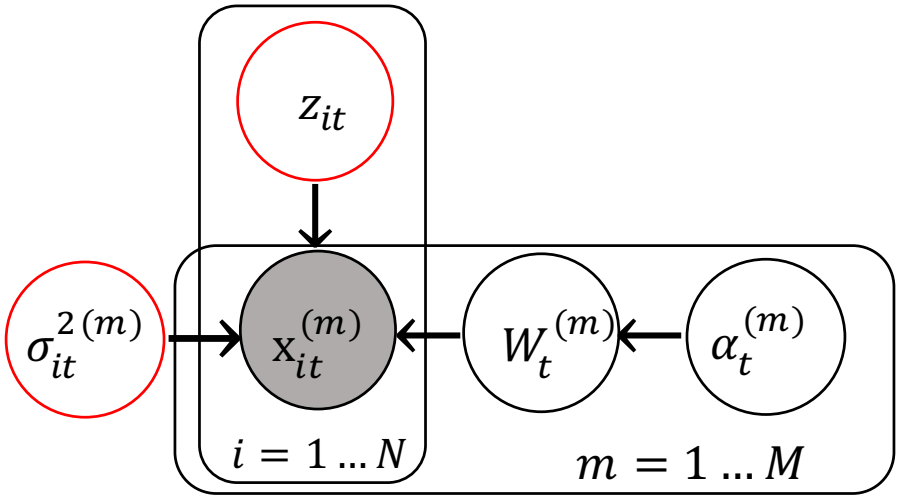
$$\boldsymbol{\tau}_m \sim Gamma(\alpha_0, \beta_0),$$

$$\boldsymbol{Z_i} \sim N(\boldsymbol{0}, \boldsymbol{I}).$$

GBCCA,DGBCCA model



(a) GBCCA



(b) DGBCCA

Outline for section 3

GBCCA model:Inference

Full joint likelihood

$$p(\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{\tau}, \boldsymbol{Z}) = p(\boldsymbol{W})p(\boldsymbol{\tau})p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{\tau})$$

$$= p(\boldsymbol{W}) \prod_{m=1}^{M} Gamma(\tau_m|\alpha_\tau, \beta_\tau) \prod_{i=1}^{N} (N(\boldsymbol{z}_i|0, \boldsymbol{I}) N(\boldsymbol{x}_i^{(m)}|Z_i \boldsymbol{W}^{(m)}, \boldsymbol{\tau}_m))$$

We assume the weight matrix $\boldsymbol{W}$ is made sparse by a group-wise ARD prior,

$$p(\boldsymbol{W}|\alpha) = \prod_{m=1}^{M} \prod_{k=1}^{K} \prod_{d=1}^{D_m} N(w_{k,d}^{(m)}|0, \alpha_k^{(m)^{-1}})$$

The ARD makes groups of variables inactive for specific components by driving $\alpha_{m,k}^{-1}$ to zero.

GBCCA model:Variational updates

The variational distribution $Q(\mathbf{\Theta})$ is assumed to factorize over the component variables θ_i in $\boldsymbol{\theta}$,

$$Q(\boldsymbol{\theta}) = \prod_{i=1} q_i(\theta_i)$$

In order to minimize the Kullback-Leibler (KL) divergence between the approximating distribution $Q(\boldsymbol{\theta})$ and the true posterior $P(\boldsymbol{\theta}|D)$.The best distribution q_j^* for each of the factors q_j can be expressed as

$$q_j^*(\boldsymbol{\theta}_j|D) = \frac{e^{E_{i\neq j}[lnp(\boldsymbol{\theta},\boldsymbol{D})]}}{\int e^{E_{i\neq j}[lnp(\boldsymbol{\theta},\boldsymbol{D})]} d\boldsymbol{\theta}_j}$$

$$lnq_j^*(\boldsymbol{\theta}_j|D) = E_{i\neq j}[lnp(\boldsymbol{\theta},\boldsymbol{D})] + constant$$

(Bishop,1999)

GBCCA model:Variational updates

latent variables $\boldsymbol{Z}$

$$q(\boldsymbol{Z}) = \prod_{i=1}^{N} N(z_i|\mu_i^{(z)}, \Sigma^{(z)}),$$

$$\mu_i^{(z)} = \Sigma_{m=1}^{M}\Sigma^{(z)}\langle W^{(m)}\rangle\langle\tau_{(m)}\rangle x_i^{(m)}, \Sigma^{(z)} = (I_k + \Sigma_{m=1}^{M}\langle\tau_m\rangle\langle W^{(m)}W^{(m)^T})^{-1}.$$

projection matrices $\boldsymbol{W}$

$$q(\boldsymbol{W}) = \prod_{m=1}^{M}\prod_{j=1}^{D_m} N(w_{:,j}^{(m)}|\mu_{m,j}^{(w)}, \Sigma_m^{(w)}),$$

$$\mu_{m,j}^{(w)} = \Sigma_m^{(w)}\langle\tau_m\rangle(\Sigma_{i=1}^{N} x_{ij}^{(m)})\langle\boldsymbol{Z}_i\rangle,$$

$$\Sigma_m^{(w)} = (\langle\tau_m\rangle\Sigma_{i=1}^{N}\langle\boldsymbol{Z}_i\boldsymbol{Z}_i^T + \langle\boldsymbol{\alpha_m}\rangle^{-1})^{-1}.$$

GBCCA model:Variational updates

ARD parameters

$$q(\boldsymbol{\alpha}) = \prod_{m=1}^{M} \prod_{k=1}^{K} Gamma(\alpha_m^{\alpha}, \beta_{mk}^{\alpha}),$$

$$\alpha_m^{\alpha} = \alpha^{\alpha} + D_m/2, \beta_{mk}^{\alpha} = \beta^{\beta} + w_k^{(m)^T} w_k^{(m)}/2.$$

noise precision parameters

$$q(\boldsymbol{\tau}) = \prod_{m=1}^{M} Gamma(\tau_m | \alpha_m^{\tau}, \beta_m^{\tau})$$

$$\alpha_m^{\tau} = \alpha^{\tau} + D_m N/2, \beta_m^{\tau} = \beta^{\tau} + 1/2\langle (x_i^{(m)} - Z_i \boldsymbol{W}^{(\boldsymbol{m})^T})^2 \rangle.$$

GBCCA model:predictive inference

$p(X^{(m)}|Y^{-(m)})$

The task is to predict unobserved groups based on observed data.Given $\boldsymbol{Y}^{-(m)}$,$q(\boldsymbol{W})$ and $q(\tau)$, we can approximate the posterior distribution for the latent variables $q(\boldsymbol{Z})$ corresponding to $\boldsymbol{Y}^{-(m)}$ and approximate the mean of the predictive distribution as

$$\begin{aligned}\langle \boldsymbol{X}^{(m)}|\boldsymbol{Y}^{-(m)}\rangle &= \langle \boldsymbol{Z}\boldsymbol{W}^{(m)}\rangle_{q(\boldsymbol{W}^{(m)}),q(\boldsymbol{Z})} \\ &= \boldsymbol{Y}^{-(m)}\boldsymbol{\tau}\langle \boldsymbol{W}^{(-m)^T}\rangle \boldsymbol{\Sigma}^{-1}\langle \boldsymbol{W}^{(m)}\rangle\end{aligned}$$

$$\boldsymbol{\tau} = diag(\langle \tau_j\rangle \boldsymbol{I}_{\boldsymbol{D}_j\, j\neq m}), \boldsymbol{\Sigma} = \boldsymbol{I}_K + \Sigma_{j\neq m}\langle \tau_j\rangle \langle \boldsymbol{W}^{(j)}\boldsymbol{W}^{(j)^T}\rangle.$$

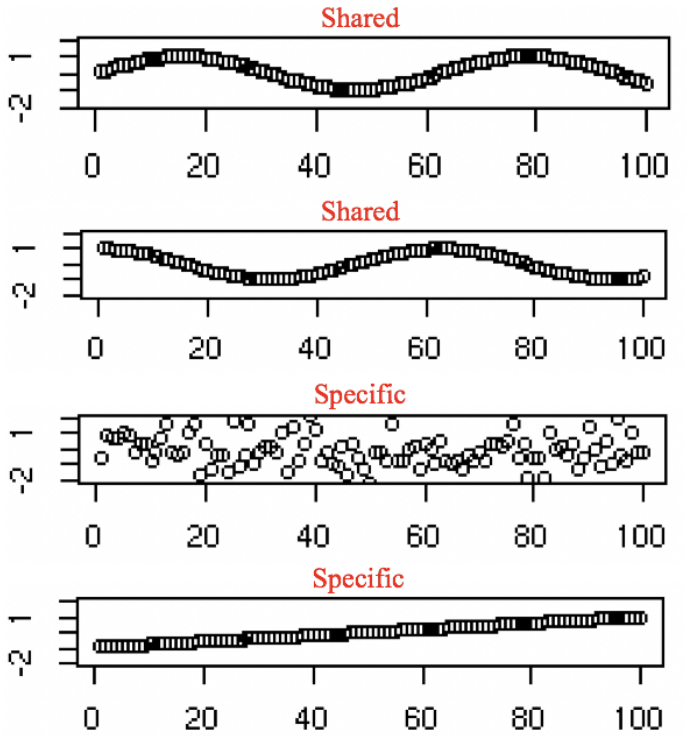
Outline for section 4

Strategy: M=2, GBCCA $\rightarrow$ BIBFA

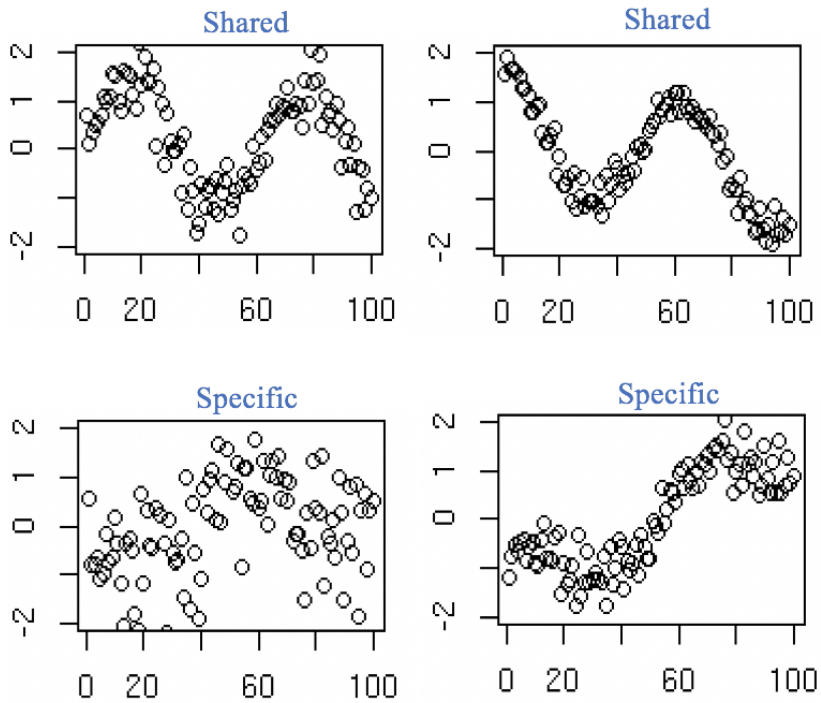
We conducted simulations on the BIBFA model to show that the inference procedure converges to the correct solution.

- We generated two collections of data sets,$\{\boldsymbol{Y}^1, \boldsymbol{Y}^2\}$, with $N = 100$ samples and dimensions $(50, 40)$.
- The parameters are set to contain all types of components (view-specific and shared components).$K = 4$ true components, $\alpha = 1$,$\tau_1 = 3$ and $\tau_2 = 6$.
- We fit the BIBFA model with $K = 6$, other parameters of this model are: $z_n \sim N(\mathbf{0}, \mathbf{1})$;initial $\tau = 1000$; $\alpha_k^{(m)} \sim Gamma(10^{-14}, 10^{-14})$;$\tau_m \sim Gamma(10^{-14}, 10^{-14})$.
- The maximum number of iterations is 10000 to train the model parameters. We use the convergence as a relative change of $L(Q)$ falling below 10^{-6}.

M=2, GBCCA → BIBFA



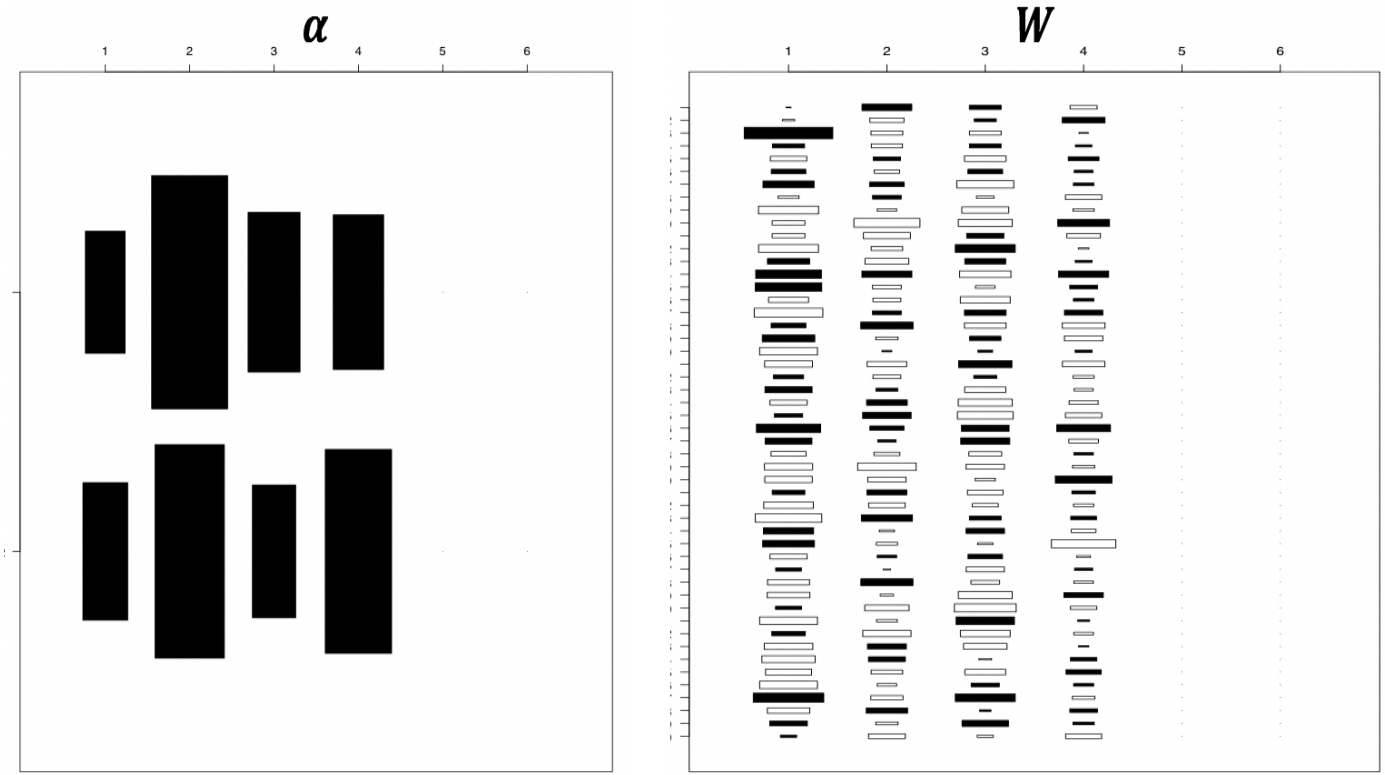
(a) 1a:True latent components z



(b) 1b:Learned latent components z

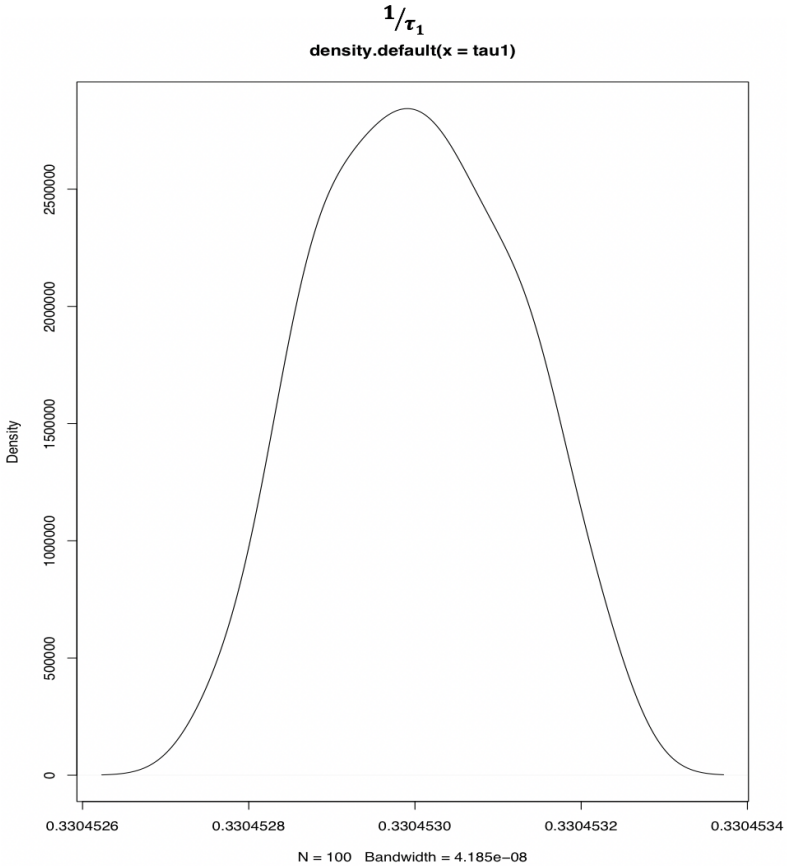
BIBFA learns the correct latent components and discards the access ones

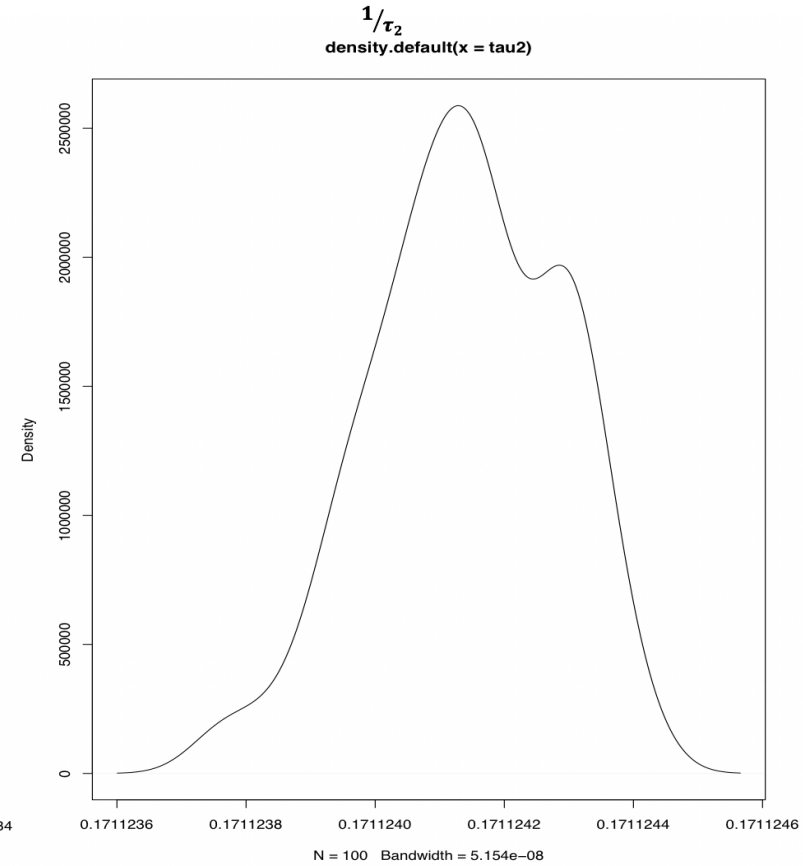
- α is the variance of components, $\boldsymbol{W}$ denotes the elements



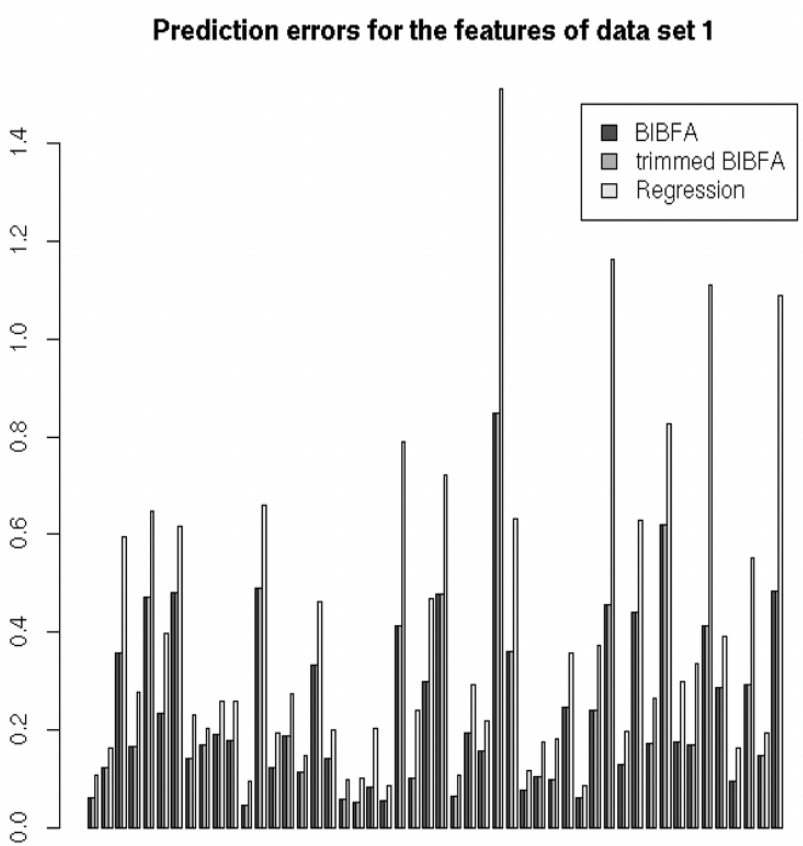
BIBFA finds the correct posterior distributions for model parameters $q(\tau_m)$

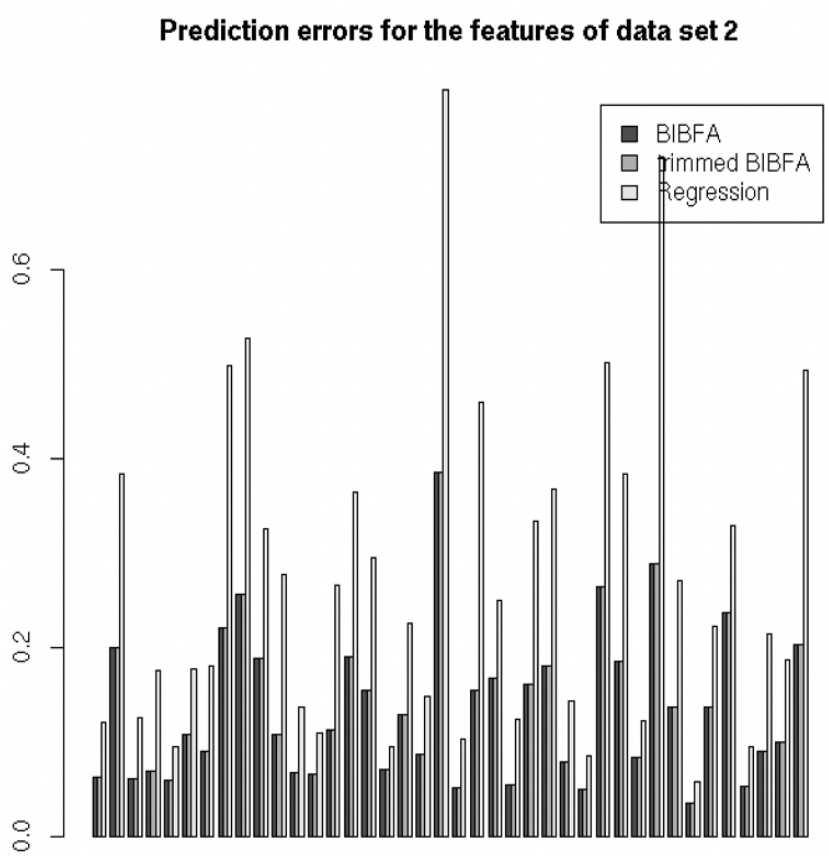
- The true $1/\tau_1 = 0.33$ and $1/\tau_2 = 0.17$





BIBFA prediction



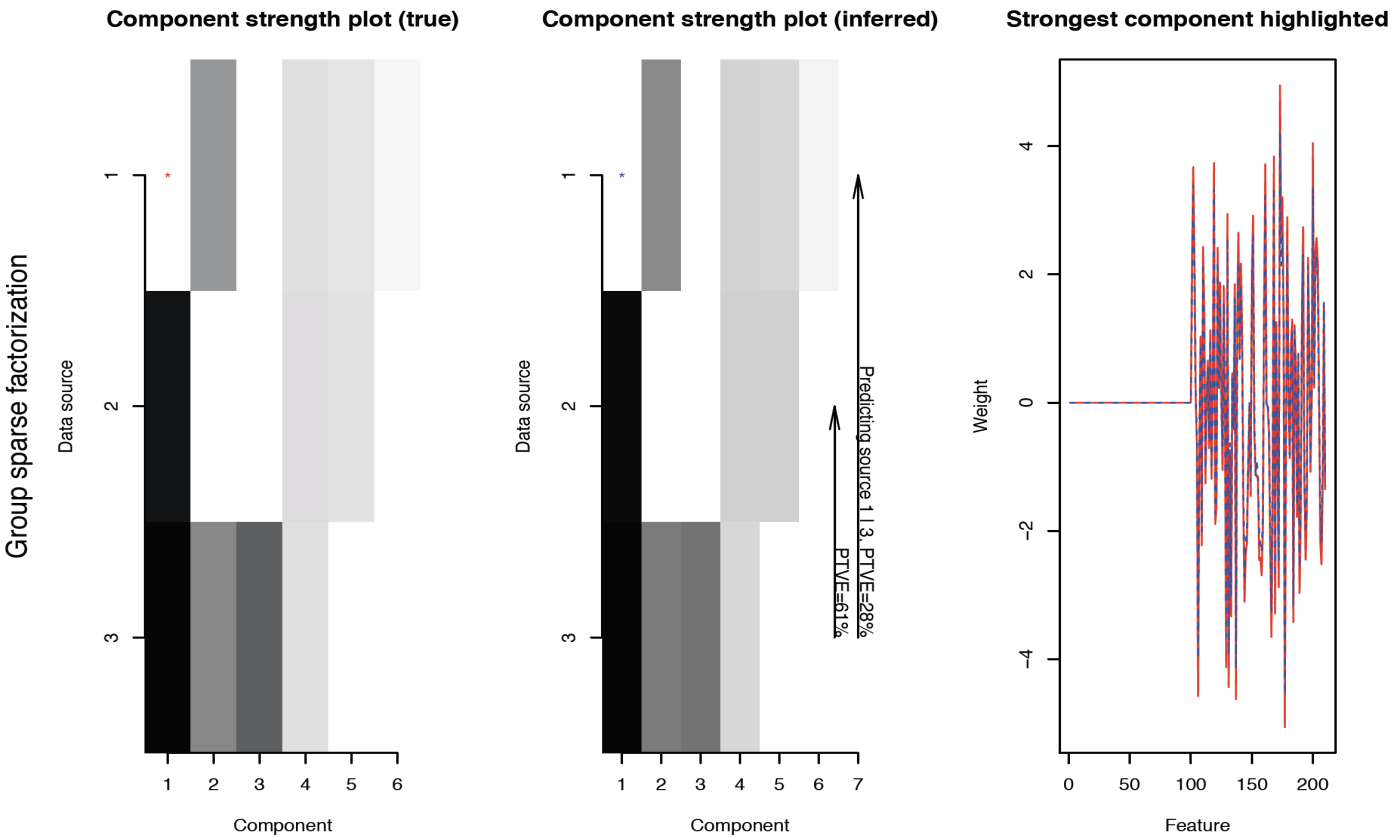


$$rmse = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2}{\frac{1}{n}\sum_{i=1}^{n}y_i^2}$$

Strategy:M>2,GBCCA

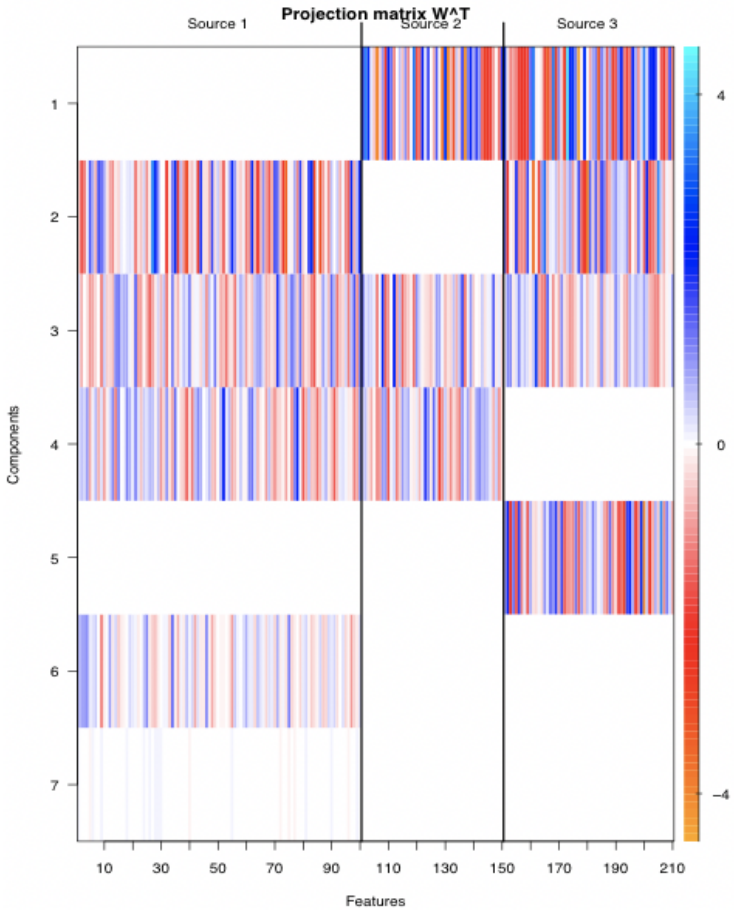
- We generated three collections of datasets,$\{\boldsymbol{Y}^1, \boldsymbol{Y}^2, \boldsymbol{Y}^3\}$, with $N = 200$ samples and dimensions $(100, 50, 60)$.
- $K = 6$,Z and W were drawn from $N(0, 1)$,$\tau = 1$.
- the first 20 samples as the test data, the model was trained under an initial $\tau = 1000$,$\alpha_\tau = 10$,$\beta_\tau = 10$,$\alpha_0 = 10$,$\beta_0 = 1$.
- The total number of Gibbs Sampling steps is 5000, burn-in samples are 2500, and $K = 15$.

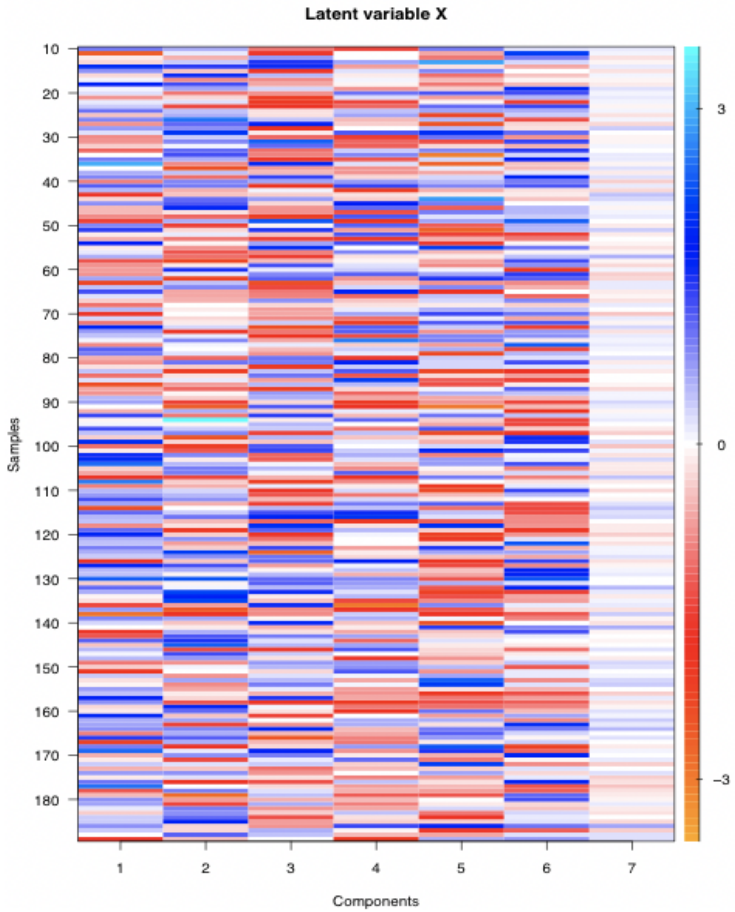
GBCCA infers the correct component affiliations



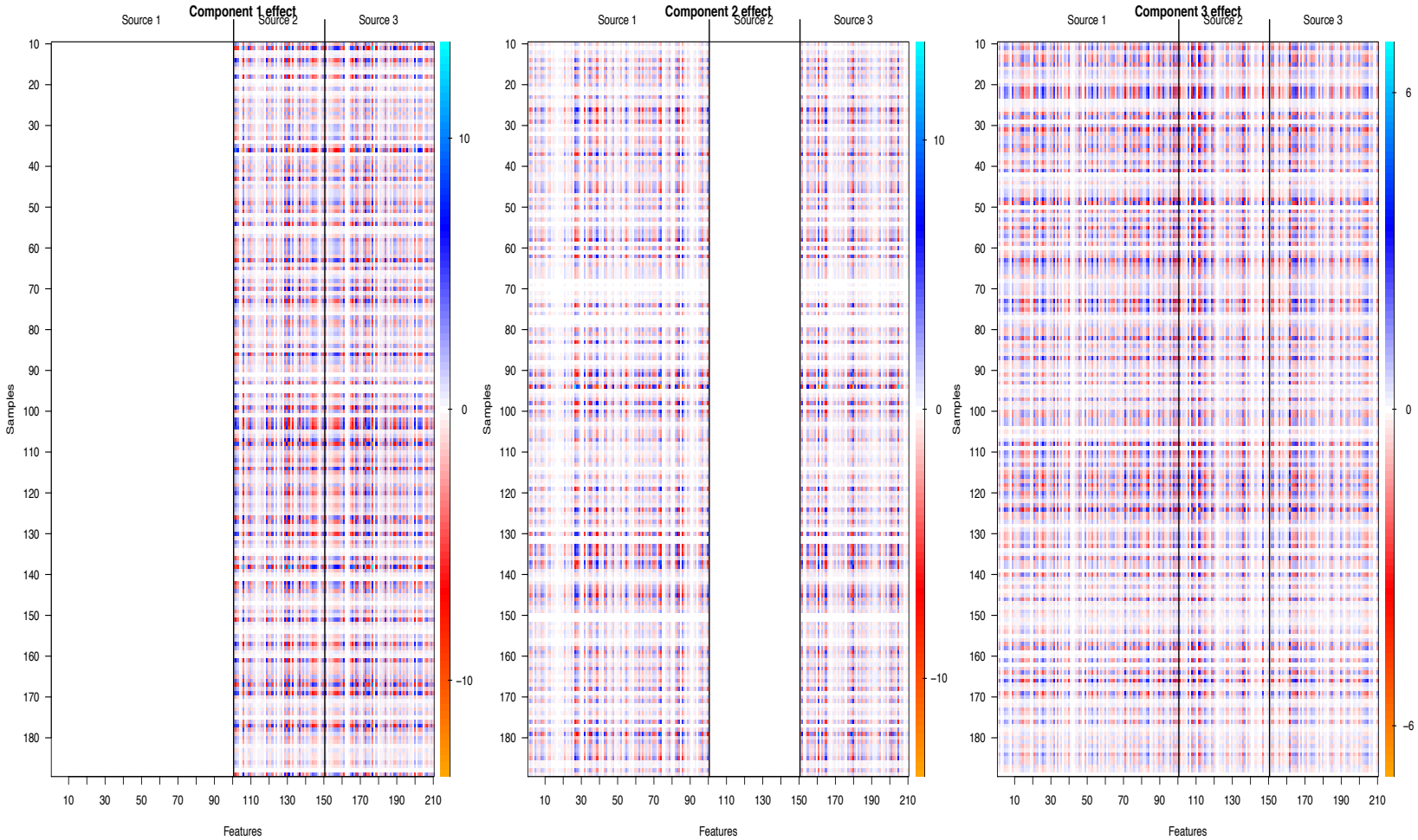
$$ptve = \frac{1 - \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2}{\frac{1}{n}\sum_{i=1}^{n} y_i^2} * 100\%$$

GBCCA:projection matrix and latent variable effect





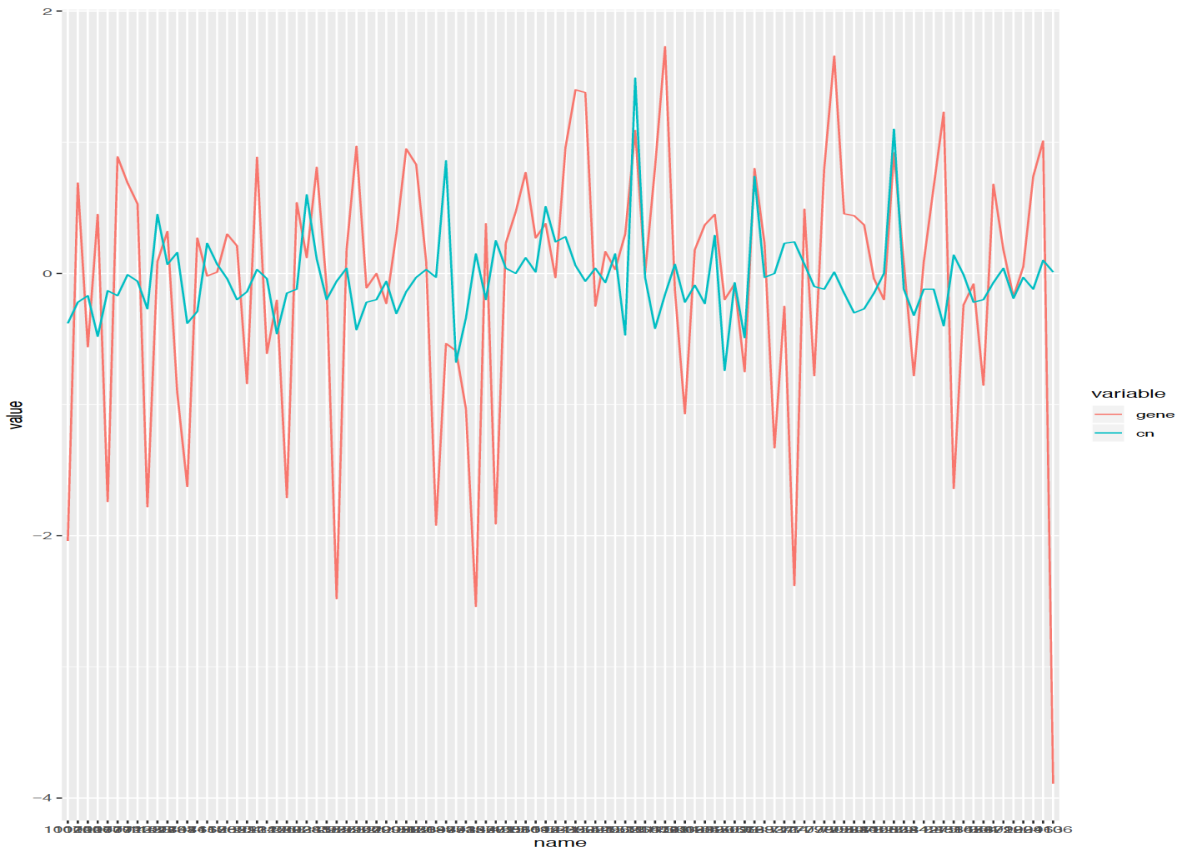
GBCCA:component effects for each feature Y=ZW



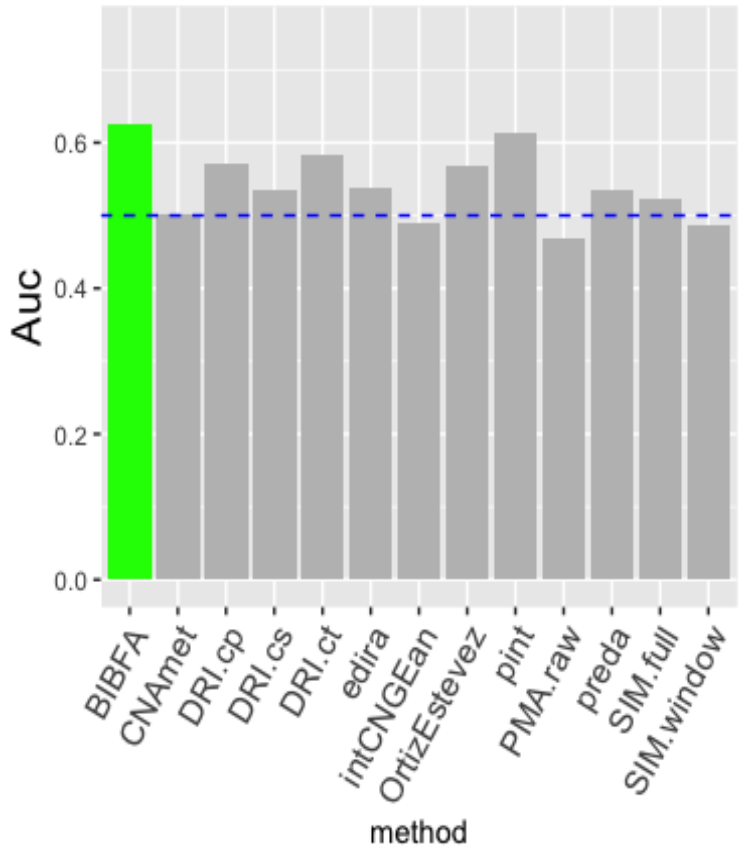
BIBFA:Pollack data set(Lahti et al.,2012)

- This data set contains 41 breast cancer samples, each sample has 4296 genes' expression value and their corresponding copy number value. The data dimension is $Y^{(m)} = 41 \times 4296$.
- There are 38 known cancer genes.
- Gene expression and copy number data can be integrated to identify DNA copy number alternations which induce the changes of expression levels of associated genes (Huang et al.,2011)
- BIBFA model was run by different K values ranging from 4 to 60 and we chose the best K according to the variational lower bound,resulting in K=41.
- Measure the $s_g = \Sigma_{k=1}^{K} |\langle W_{g,k}^{(1)} \rangle \langle W_{g,k}^{(2)} \rangle|$ for each gene to rank the genes (Klami et al.,2013), and we repeated the experimental setup to obtain results comparable with Lahti et al(2012).

Pollack data set



BIBFA(M=2,GBCCA)



Summary

- The GBCCA model can handle high dimension problems, while the previous BCCA (Wang, 2007) can only deal with dimension less than 10.
- The GBCCA model can accurately capture the true components for multiple groups, inferring the group dependencies.
- Under $M = 2$, we showed GBCCA outperfomed by supervised regression models in predictive tasks.

Thank you! ***Email***:luq7@psu.edu ***Wechat***:lqconnor