Drug sensitivity prediction algorithms

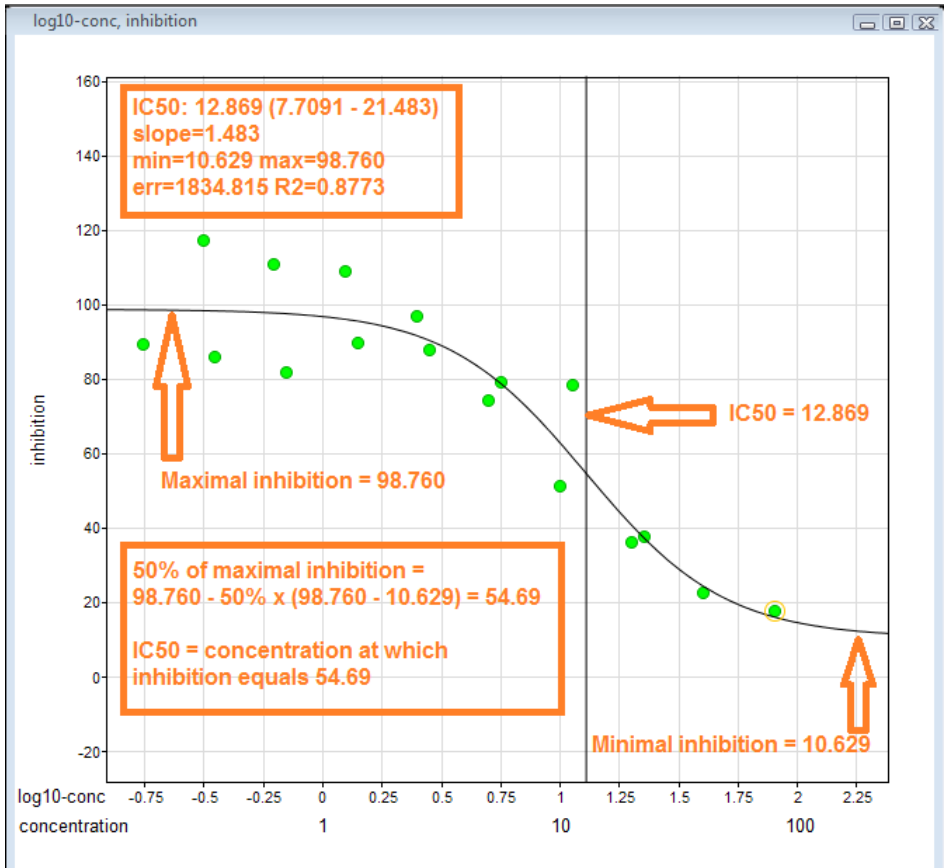
LIJING WANG

May 24th, 2016

Outline

Drug response: IC50

- Predicting the best treatment strategy from genomic information is a core goal of precision medicine.
- IC50 represents the concentration of a drug that is required for 50% inhibition. Ex.: IC50_24hr IC50_48hr IC50_72hr
- IC50 vs drug sensitivity



Variables

Nature biotechnology: A community effort to assess and improve drug sensitivity prediction algorithms, James C Costello, et al, 2014.
Total of 149 cancer cell lines (from different tissues)

- Gene expression values
- Copy Number Variation
- Gene set collections★
- *Tissue types
- *Mutations

Data Size Description

VARIABLE:

- Gene expression values: 18875*149
- Copy Number Variation: 17771*149
- Gene set collections★: 4726, 1454
- *Tissue types: 18 kinds, 5 kinds for prediction, others sample number is quite small
- *Mutations

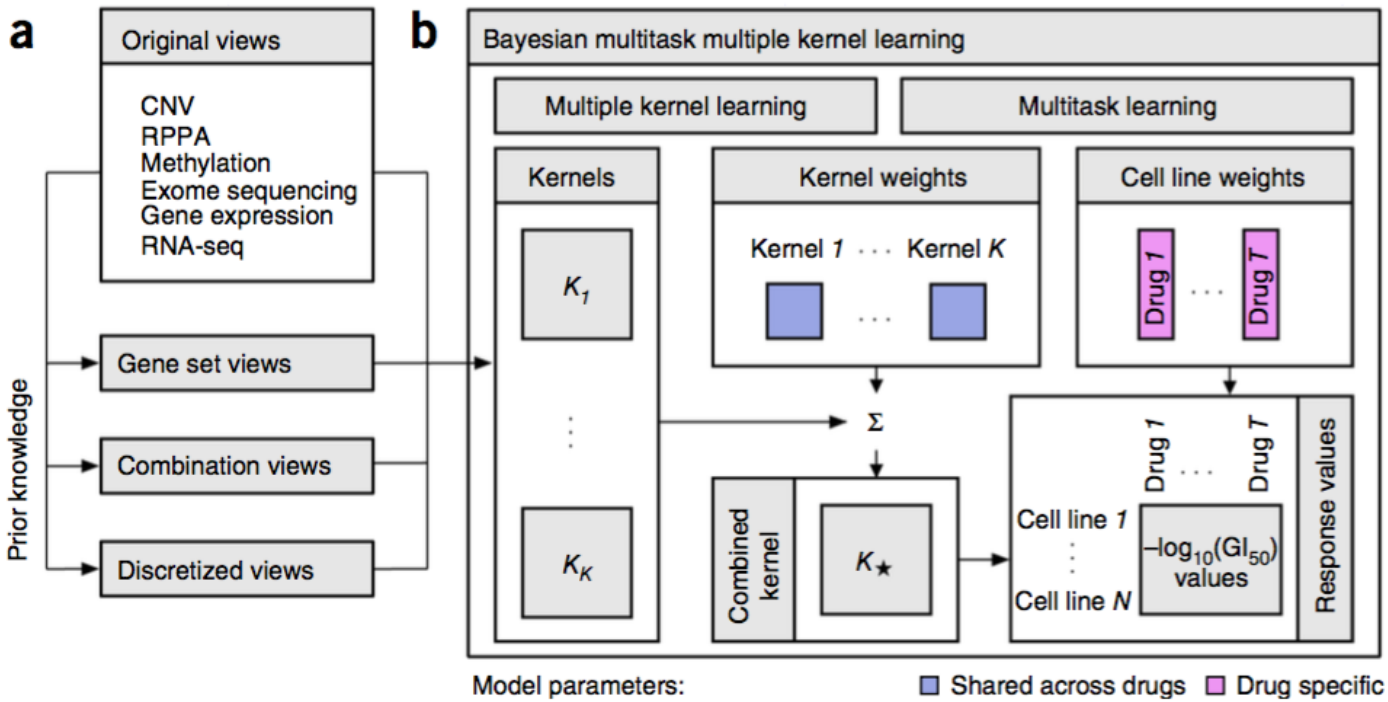
PREDICT: $IC50\_72hr$
Criterion: Nature Paper: 28 compounds ranking on drug sensitivity.
Our goal: Minimize Mean Square Error.

Bayesian multitask multiple kernel learning

Kernelized regression: a regression approach that computes outputs from similarities between cell lines.

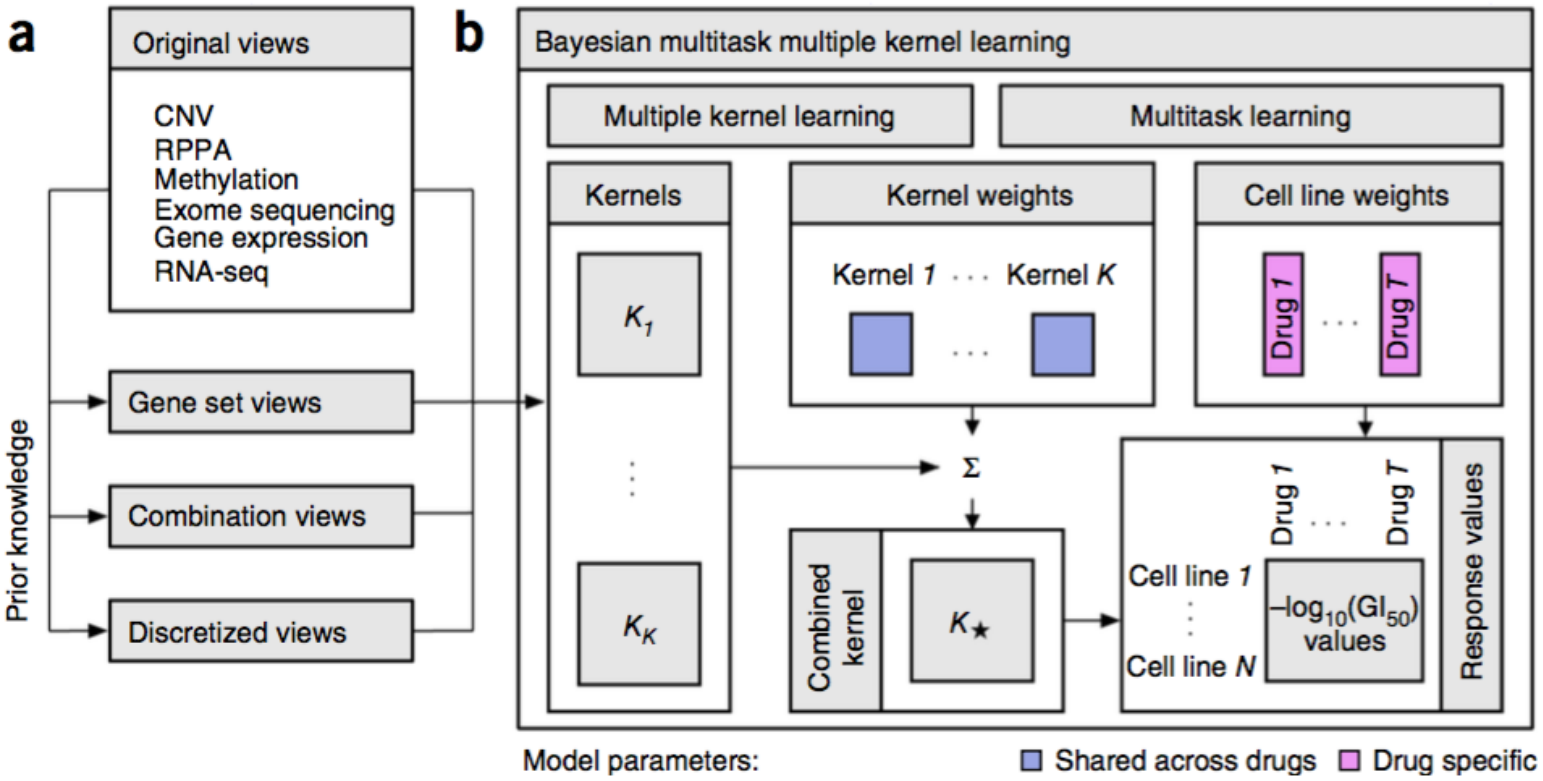
$$k_{t,k}(x_{t,k,i}, x_{t,k,j}) = \frac{x_{t,k,i}^T x_{t,k,j}}{x_{t,k,i}^T x_{t,k,i} + x_{t,k,j}^T x_{t,k,j} - x_{t,k,i}^T x_{t,k,j}} \quad \forall(t,k,i,j)$$

$$k_{t,k}(x_{t,k,i}, x_{t,k,j}) = \exp\left(-\frac{||x_{t,k,i} - x_{t,k,j}||^2}{2\sigma_{t,k}^2}\right) \quad \forall(t,k,i,j)$$



Bayesian multitask multiple kernel learning

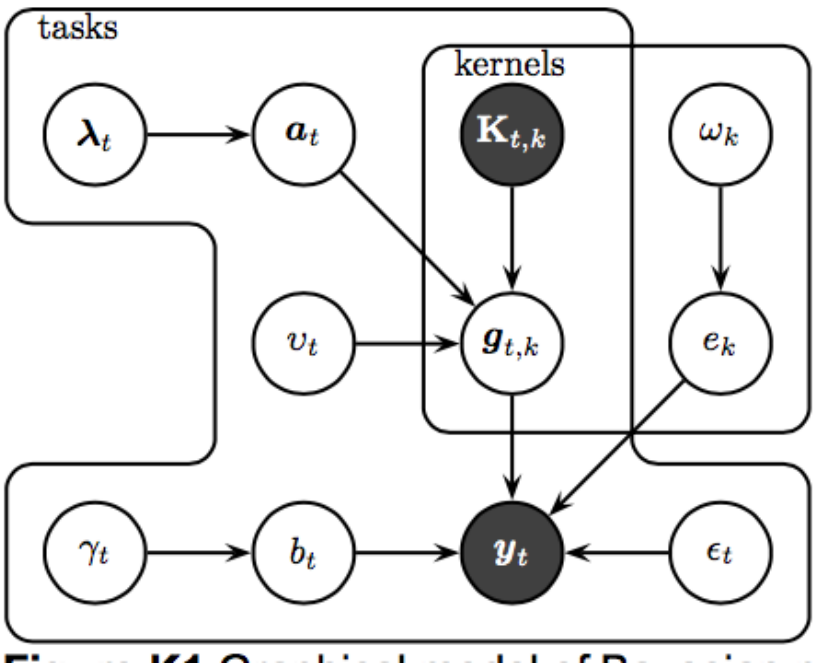
Multiview learning: Provided K kernels, involve all genomic views.
Multitask learning: 28 compounds.(In our case it could be eliminated)



Bayesian multitask multiple kernel learning

Bayesian Inference:

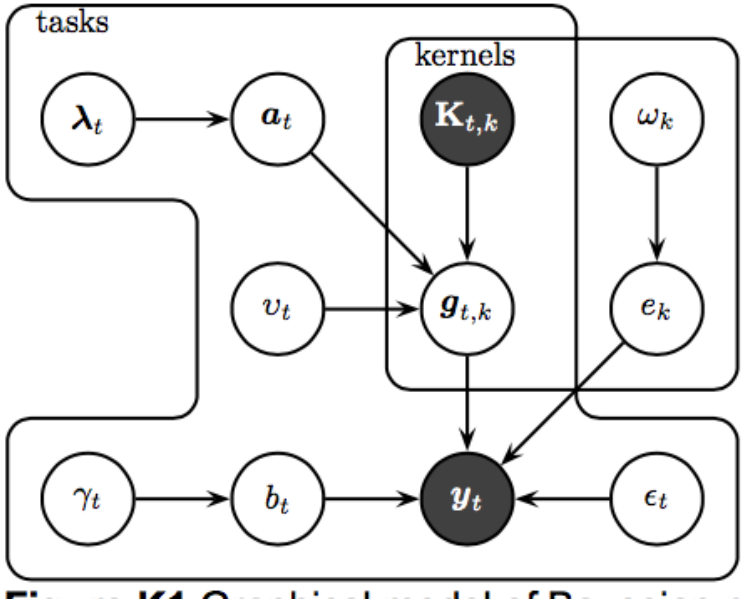
- t: the index for drugs
- k: the index for genomic views
- i: the index for cell lines
- T: the number of drugs
- K: the number of genomic views
- N: the number of cell lines in the training set



Bayesian multitask multiple kernel learning

Bayesian Inference: Learning model parameters: deterministic variational approximation.

$$\lambda_{t,i} \sim \mathcal{G}(\lambda_{t,i};\, \alpha_\lambda, \beta_\lambda) \quad \forall (t,i)$$
$$a_{t,i}|\lambda_{t,i} \sim \mathcal{N}(a_{t,i};\, 0, \lambda_{t,i}^{-1}) \quad \forall (t,i)$$
$$\upsilon_t \sim \mathcal{G}(\upsilon_t;\, \alpha_\upsilon, \beta_\upsilon) \quad \forall t$$
$$g_{t,k}|a_t, K_{t,k}, \upsilon_t \sim \mathcal{N}(g_{t,k};\, K_{t,k}a_t, \upsilon_t^{-1} I) \quad \forall (t,k)$$
$$\gamma_t \sim \mathcal{G}(\gamma_t;\, \alpha_\gamma, \beta_\gamma) \quad \forall t$$
$$b_t|\gamma_t \sim \mathcal{N}(b_t;\, 0, \gamma_t^{-1}) \quad \forall t$$
$$\omega_k \sim \mathcal{G}(\omega_k;\, \alpha_\omega, \beta_\omega) \quad \forall k$$
$$e_k|\omega_k \sim \mathcal{N}(e_k;\, 0, \omega_k^{-1}) \quad \forall k$$
$$\varepsilon_t \sim \mathcal{G}(\varepsilon_t;\, \alpha_\varepsilon, \beta_\varepsilon) \quad \forall t$$
$$y_t|b_t, e, G_t, \varepsilon_t \sim \mathcal{G}\left(y_t;\, \sum_{k=1}^{K} e_k g_{t,k} + b_t 1, \varepsilon_t^{-1} I\right) \quad \forall t$$



Methods comparison

BMMKL: (α, β): (1, 1) default priors
Random Forest Regression: NOT DO VARIABLE SELECTION: computational cause. Default randomForest()
Elastic Net: $\alpha = 0.5$

Table: Methods comparison

Method	MSE
BMMKL	0.231187
Random Forest	0.231770
Elastic Net	0.263836
BMMKL(with gene set)	0.231729

How to deal with gene set? Brainstormmmmmmming...

Sparse PCA

For selecting promising genes in dataset. $\Sigma = X^T X$

$$\begin{aligned} \max \quad & v^T \Sigma v \\ \text{subject to} \quad & \|v\|_2 = 1 \quad \textit{Eq. 1} \\ & \|v\|_0 \leq k. \end{aligned}$$

We could easily convert SPCA problem into SDP to solve.

Sparse PCA

Goal: Construct subspaces for different dimensions.
Example:
gene_set_1: involve s genes. n*s expression data. $\rightarrow$ s*k
gene_set_2: involve t genes. n*t expression data. $\rightarrow$ t*l
Setting: nonzeros1 = 0.25*s, nonzeros2 = 0.25*t. To get k and l, percentage of explained variance is larger than 60 percent.
Where is n*k? n*l?

Consider the inclusion map $\iota_n : \mathbb{R}^n \to \mathbb{R}^{n+1}$, $\iota_n(x_1, \ldots, x_n) = (x_1, \ldots, x_n, 0)$. It is easy to see that ι_n induces an inclusion of $\mathrm{Gr}(k, n)$ into $\mathrm{Gr}(k, n+1)$ which we will call natural inclusion and, with a slight abuse of notation, also denote by ι_n. For any $m > n$, composition of successive

Distance between subspaces

Once we have $\mathbf{A} \in Gr(k, n)$, $\mathbf{B} \in Gr(l, n)$, we could use the following formula to calculate distance.

$$\delta(\mathbf{A}, \mathbf{B}) = \left(\sum_{i=1}^{min(k,l)} \theta_i(\mathbf{A}, \mathbf{B})^2 \right)^{1/2} \tag{1}$$

$\theta_i = cos^{-1}\sigma_i$, σ_i is the nonzero sigular values of $\mathbf{A}^T\mathbf{B}$. (Ke Ye, Lek-Heng Lim, 2014)

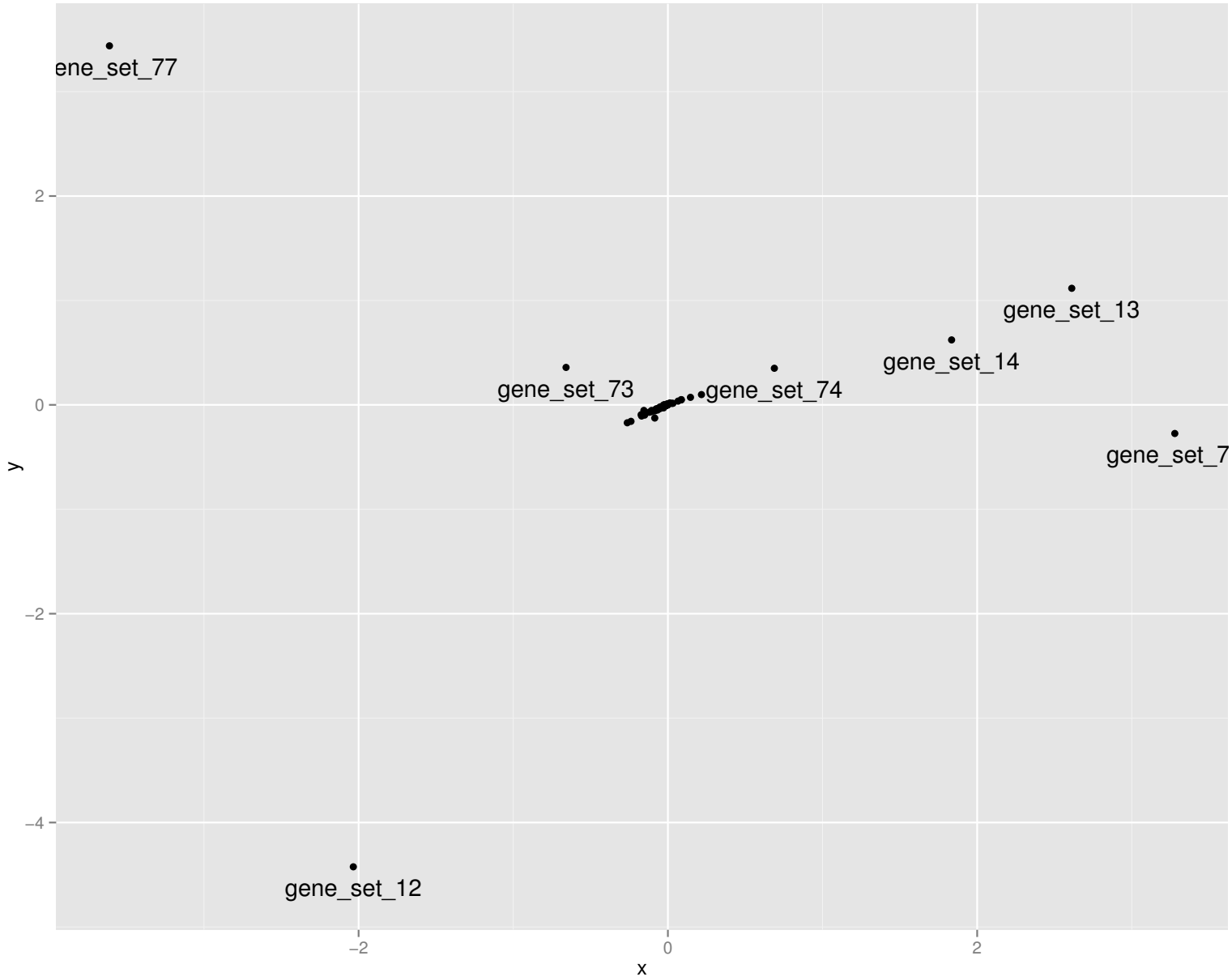
MDS plot

Distance matrix(dissimilarity matrix):

$$\min_{x1,\ldots,x_I} \sum_{i<j} (\|x_i - x_j\| - \delta_{i,j})^2.$$

$\rightarrow$ two dimension for visualization. (80 gene_set)

MDS plot


ene_set_77
gene_set_13
gene_set_14
gene_set_73
gene_set_74
gene_set_7
gene_set_12
2
0
−2
−4
y
−2
0
2
x

Questions

- Variable selection
- Gene set involving
- SPCA parameter selection, using nonzeros?
- New method?