Johnson&Johnson PHARMACEUTICAL RESEARCH & DEVELOPMENT

BIOINF

FABIA: Factor Analysis for Bicluster Aquisition

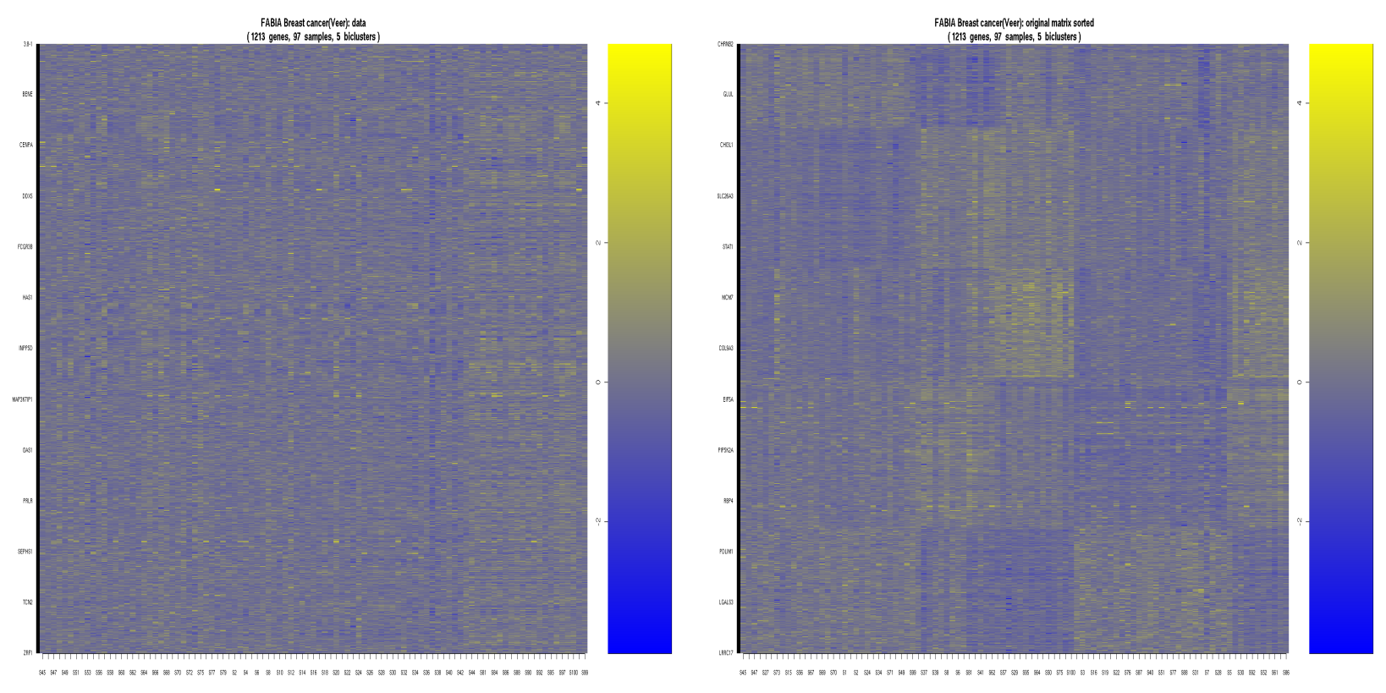
Sepp Hochreiter[1], Ulrich Bodenhofer[1], Martin Heusel[1], Andreas Mayr[1], Andreas Mitterecker[1], Adetayo Kasim[2], Tatsiana Khamiakova[2], Suzy Van Sanden[2], Dan Lin[2], Willem Talloen[3], Luc Bijnens[3], Hinrich W.H. Göhlmann[3], Ziv Shkedy[2] and Djork-Arné Clevert[1,4]

[1]Institute of Bioinformatics, Johannes Kepler University, Linz, Austria, [2]Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Hasselt, [3]Johnson & Johnson Pharmaceutical Research & Development, Division of Janssen Pharmaceutica, Beerse, Belgium and [4]Department of Nephrology and Internal Intensive Care, Charité, Berlin, Germany

Motivation: Biclustering of transcriptomic data groups genes and samples simultaneously. It is emerging as a standard tool for extracting knowledge from gene expression measurements. We propose a novel generative approach for biclustering called "FABIA: Factor Analysis for Bicluster Acquisition". FABIA is based on a multiplicative model, which accounts for linear dependencies between gene expression and conditions, and also captures heavy-tailed distributions as observed in real-world transcriptomic data. The generative framework allows to utilize well-founded model selection methods and to apply Bayesian techniques.

Results: On 100 simulated data sets with known true, artificially implanted biclusters, FABIA clearly outperformed all 11 competitors. On these data sets, FABIA was able to separate spurious biclusters from true biclusters by ranking biclusters according to their information content. FABIA was tested on three microarray data sets with known sub-clusters, where it was two times the best and once the second best method among the compared biclustering approaches.

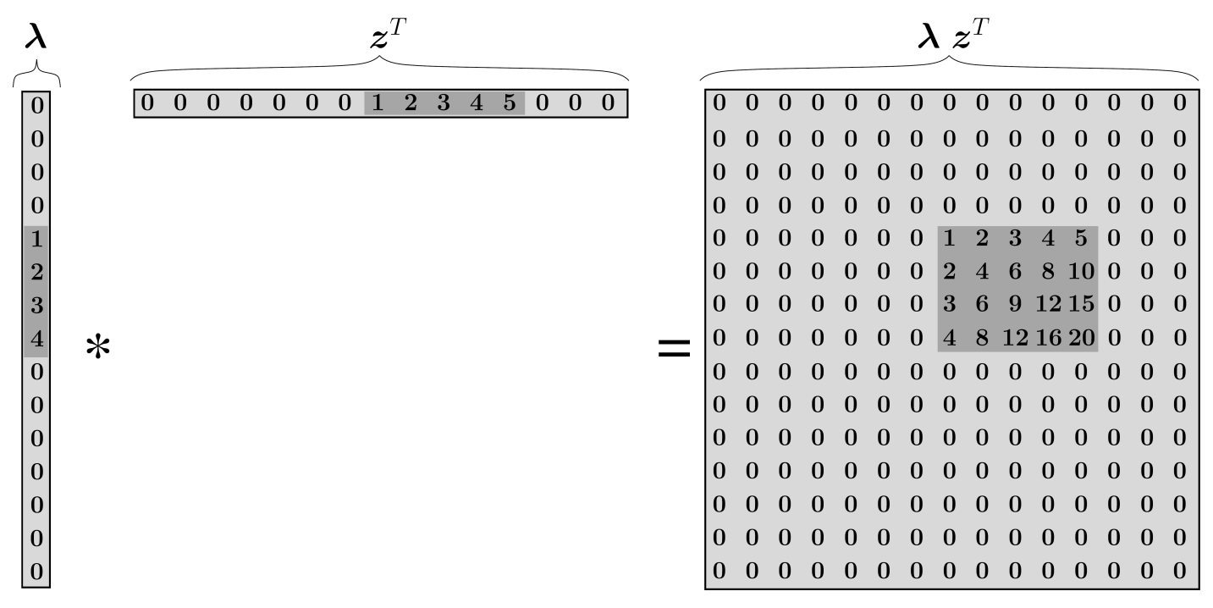
Biclusters



Biclusters are subsets of rows and columns, rows behave similar on a column subset and columns behave similar on a row subset.

Breast cancer data (van't Veer et al., 2002) and biclusters found by reordering

The FABIA Model



A subset of rows and columns can be represented as an outer product $\boldsymbol{\lambda}\, \boldsymbol{z}^T$ of two sparse vectors $\boldsymbol{\lambda}$ and $\boldsymbol{z}$.

Model for p biclusters and additive noise

$$\boldsymbol{X} = \sum_{i=1}^{p} \boldsymbol{\lambda}_i \boldsymbol{z}_i^T + \boldsymbol{\Upsilon} = \boldsymbol{\Lambda}\, \boldsymbol{Z} + \boldsymbol{\Upsilon}$$

Generative interpretation by a factor analysis model with p factors

$$\boldsymbol{x} = \sum_{i=1}^{p} \boldsymbol{\lambda}_i \tilde{z}_i + \boldsymbol{\epsilon} = \boldsymbol{\Lambda}\, \tilde{\boldsymbol{z}} + \boldsymbol{\epsilon}$$

Sparseness by a Laplace distribution as priors

$$p(\tilde{\boldsymbol{z}}) = \left(\frac{1}{\sqrt{2}}\right)^p \prod_{i=1}^{p} e^{-\sqrt{2}|\tilde{z}_i|}, \quad p(\boldsymbol{\lambda}_j) = \left(\frac{1}{\sqrt{2}}\right)^p \prod_{k=1}^{p} e^{-\sqrt{2}|\lambda_{kj}|}$$

with $\boldsymbol{X}, \boldsymbol{\Upsilon} \in \mathbb{R}^{n \times l}, \boldsymbol{\Lambda} \in \mathbb{R}^{n \times p}, \boldsymbol{Z} \in \mathbb{R}^{p \times l}, \boldsymbol{\lambda}_i \in \mathbb{R}^n, \boldsymbol{z}_i \in \mathbb{R}^l, \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{\Lambda} \in \mathbb{R}^{n \times p}, \tilde{\boldsymbol{z}} \in \mathbb{R}^p, \boldsymbol{\epsilon} \in \mathbb{R}^n, \boldsymbol{\epsilon}\ is\ N(\boldsymbol{0}, \boldsymbol{\Psi}), \boldsymbol{\Psi} \in \mathbb{R}^{n \times n}$

and n genes, l samples

Model Selection

Likelihood of the model parameters $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ is analytically untractable → variational EM algorithm for maximization the posterior → maximum of a parametrized Gaussian model family

$$p(\boldsymbol{x}|\boldsymbol{\Lambda}, \boldsymbol{\Psi}) = \int p(\boldsymbol{x}|\tilde{\boldsymbol{z}}_j, \boldsymbol{\Lambda}, \boldsymbol{\Psi})\, p(\tilde{\boldsymbol{z}})\, d\tilde{\boldsymbol{z}} \approx \operatorname*{argmax}_{\boldsymbol{\xi}} p(\boldsymbol{x}|\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\xi})$$

E-Step:

$$E(\tilde{\boldsymbol{z}}_j|\boldsymbol{x}_j) = (\boldsymbol{\Lambda}^T \boldsymbol{\Psi}^{-1} \boldsymbol{\Lambda} + \boldsymbol{\Xi}_j^{-1})^{-1} \boldsymbol{\Lambda}^T \boldsymbol{\Psi}^{-1} \boldsymbol{x}_j$$

$$E(\tilde{\boldsymbol{z}}_j \tilde{\boldsymbol{z}}_j^T|\boldsymbol{x}_j) = (\boldsymbol{\Lambda}^T \boldsymbol{\Psi}^{-1} \boldsymbol{\Lambda} + \boldsymbol{\Xi}_j^{-1})^{-1} + E(\tilde{\boldsymbol{z}}_j|\boldsymbol{x}_j)\, E(\tilde{\boldsymbol{z}}_j|\boldsymbol{x}_j)^T$$

$$\boldsymbol{\xi}_j = \operatorname{diag}\left(\sqrt{E(\tilde{\boldsymbol{z}}_j \tilde{\boldsymbol{z}}_j^T|\boldsymbol{x}_j)}\right)$$

M-Step:

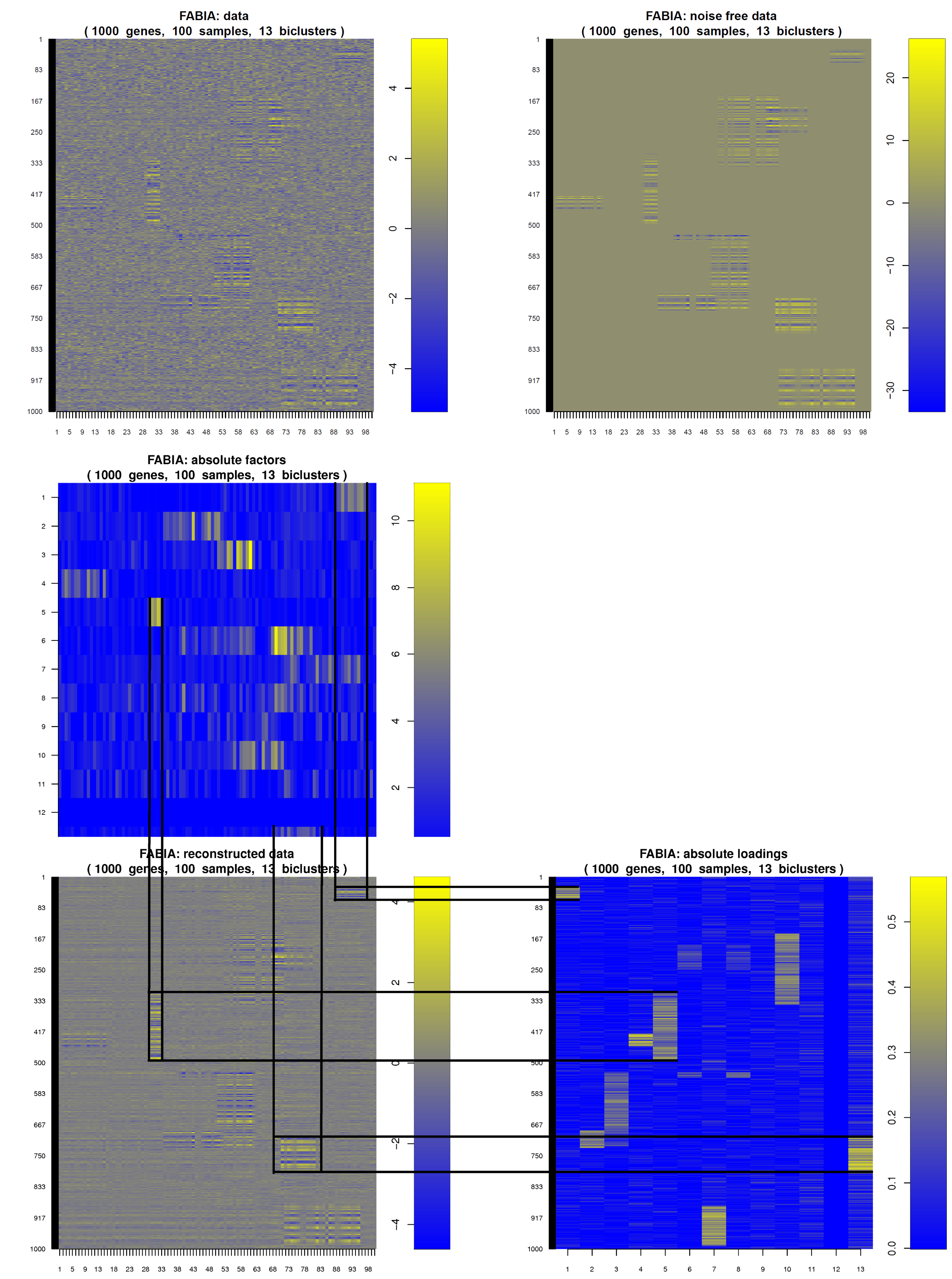
$$\boldsymbol{\Lambda}^{\text{new}} = \frac{\frac{1}{l}\sum_{j=1}^{l} \boldsymbol{x}_j E(\tilde{\boldsymbol{z}}_j|\boldsymbol{x}_j)^T - \frac{\alpha}{l} \boldsymbol{\Psi} \operatorname{sign}(\boldsymbol{\Lambda})}{\frac{1}{l}\sum_{j=1}^{l} E(\tilde{\boldsymbol{z}}_j \tilde{\boldsymbol{z}}_j^T|\boldsymbol{x}_j)}$$

$$\operatorname{diag}(\boldsymbol{\Psi}^{\text{new}}) = \boldsymbol{\Psi}^{\text{EM}} + \operatorname{diag}\left(\frac{\alpha}{l} \boldsymbol{\Psi} \operatorname{sign}(\boldsymbol{\Lambda}) (\boldsymbol{\Lambda}^{\text{new}})^T\right)$$

$$\boldsymbol{\Psi}^{\text{new}} = \operatorname{diag}\left(\frac{1}{l}\sum_{j=1}^{l} \boldsymbol{x}_j \boldsymbol{x}_j^T - \boldsymbol{\Lambda}^{\text{new}} \frac{1}{l}\sum_{j=1}^{l} E(\tilde{\boldsymbol{z}}_j|\boldsymbol{x}_j) \boldsymbol{x}_j^T\right)$$

$$\boldsymbol{\Lambda}^{\text{new}} = \operatorname{proj}\left(\frac{\frac{1}{l}\sum_{j=1}^{l} \boldsymbol{x}_j E(\tilde{\boldsymbol{z}}_j|\boldsymbol{x}_j)^T}{\frac{1}{l}\sum_{j=1}^{l} E(\tilde{\boldsymbol{z}}_j \tilde{\boldsymbol{z}}_j^T|\boldsymbol{x}_j)}, spL\right)$$

spL: sparseness parameter on loadings



Results

Simulated datasets with known biclusters

100 independend data sets with 1000 genes, 100 samples and 10 implanted multiplicative biclusters.

Gene expression data sets

Breast cancer (van't Veer et al., 2002), multiple tissue types (Su et al., 2002), diffuse large-B-cell lymphoma (DLBCL) (Rosenwald et al., 2002).

Results on 100 simulated and gene expression data sets

	Simulated data	Breast cancer				Multiple tissues				DLBCL			
Method	Score	Score	#bc	#g	#s	Score	#bc	#g	#s	Score	#bc	#g	#s
FABIA	0.478 (1e-2)	**0.52**	3	92	31	0.53	5	356	29	**0.37**	2	59	62
FABIAS	**0.564 (3e-3)**	**0.52**	3	144	32	0.44	5	435	30	*0.35*	2	104	60
MFSC	0.057 (2e-3)	0.17	5	87	24	0.31	5	431	24	0.18	5	50	42
plaid_ss	0.045 (9e-4)	*0.39*	5	500	38	0.56	5	1903	35	0.30	5	339	72
plaid_ms	0.072 (4e-4)	*0.39*	5	175	38	0.50	5	571	42	0.28	5	143	63
plaid_ms_5	0.083 (6e-4)	0.29	5	56	29	0.23	5	71	26	0.21	5	68	47
plaid_a_ss	-	0.37	5	796	35	**0.65**	5	3711	31	0.28	5	389	68
plaid_a_ms	-	0.34	5	194	35	*0.58*	5	583	34	0.27	5	95	61
plaid_a_ms_5	-	0.16	5	5	26	0.20	5	11	25	0.18	5	4	68
ISA_1	0.333 (5e-2)	0.03	25	55	4	0.05	29	230	6	0.01	56	26	8
ISA_2	0.299 (6e-2)	0.25	2	466	42	0.37	3	1904	28	0.22	1	267	74
ISA_3	0.188 (4e-2)	0.22	1	742	33	0.35	3	2856	28	0.18	2	385	58
OPSM	0.012 (1e-4)	0.04	12	172	8	0.04	19	643	12	0.03	6	162	4
SAMBA	0.006 (5e-5)	0.02	38	37	7	0.03	59	53	8	0.02	38	19	15
SAMBA_01	0.002 (6e-5)	0.01	79	33	8	0.01	128	53	9	0.01	70	18	14
xMOTIF	0.004 (2e-4)	0.07	5	61	6	0.11	5	628	6	0.05	5	9	9
Bimax	0.001 (7e-6)	0.01	1	1213	97	0.10	4	35	5	0.07	5	73	5
CC	0.046 (5e-3)	0.11	5	12	12	nc	nc	nc	nc	0.05	5	10	10
plaid_t_ab	0.037 (4e-3)	0.24	2	40	23	0.38	5	255	22	0.17	1	3	44
plaid_t_a	0.006 (3e-5)	0.23	2	24	20	0.39	5	274	24	0.11	3	6	24
spec_1	0.032 (5e-4)	0.12	13	198	28	0.37	5	395	20	0.05	28	133	32
spec_2	0.011 (5e-4)	0.07	14	77	22	0.21	1	117	39	0.08	8	82	44
FLOC	-	0.04	5	343	5	nc	nc	nc	nc	0.03	5	167	5

Bioinformatics: http://bioinformatics.oxfordjournals.org/cgi/content/abstract/26/12/1520