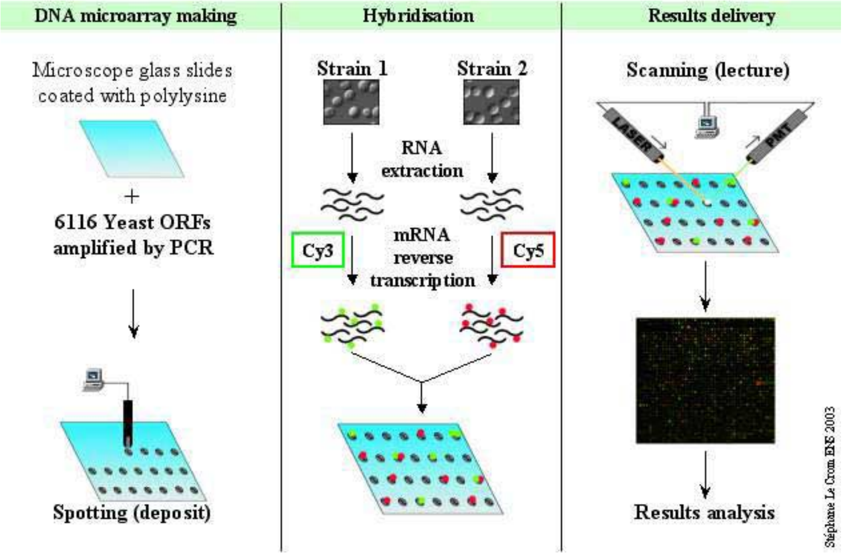
CSci 8980: Advanced Topics in Graphical Models

Application: Gene Expression Analysis

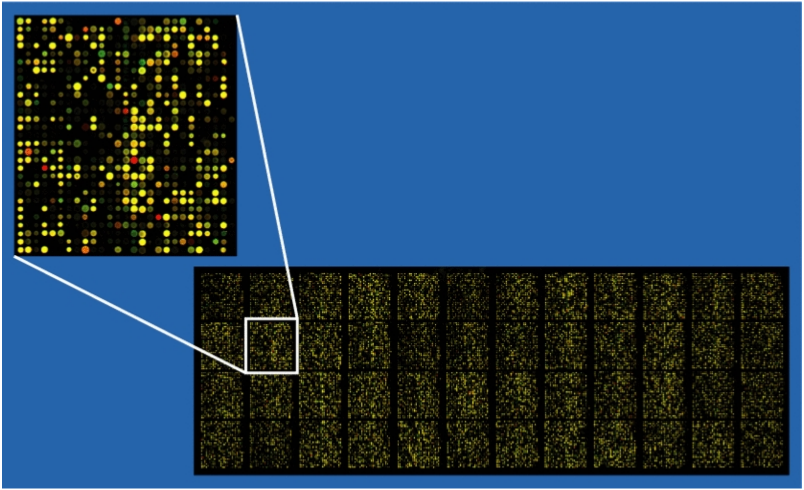
Instructor: Arindam Banerjee

November 20, 2007

Microarray Technology



Microarray Data



Microarray Data

- T gene expression profiles under M conditions

Microarray Data

- T gene expression profiles under M conditions
- x_i is the profile for the i^{th} gene

Microarray Data

- T gene expression profiles under M conditions
- x_i is the profile for the i^{th} gene
- x_{im} is the value for gene i, condition m

Microarray Data

- T gene expression profiles under M conditions
- x_i is the profile for the i^{th} gene
- x_{im} is the value for gene i, condition m
- Each profile is assumed to be generated by one cluster

Microarray Data

- T gene expression profiles under M conditions
- x_i is the profile for the i^{th} gene
- x_{im} is the value for gene i, condition m
- Each profile is assumed to be generated by one cluster
- Total Q clusters, later $Q \to \infty$

Microarray Data

- T gene expression profiles under M conditions
- x_i is the profile for the i^{th} gene
- x_{im} is the value for gene i, condition m
- Each profile is assumed to be generated by one cluster
- Total Q clusters, later $Q \to \infty$
- Clustering is an assignment $(c_1, \ldots, c_T)$

Microarray Data

- T gene expression profiles under M conditions
- x_i is the profile for the i^{th} gene
- x_{im} is the value for gene i, condition m
- Each profile is assumed to be generated by one cluster
- Total Q clusters, later $Q \to \infty$
- Clustering is an assignment $(c_1, \ldots, c_T)$
    - Each $c_i \in \{1, \ldots, Q\}$

Microarray Data

- T gene expression profiles under M conditions
- x_i is the profile for the i^{th} gene
- x_{im} is the value for gene i, condition m
- Each profile is assumed to be generated by one cluster
- Total Q clusters, later $Q \to \infty$
- Clustering is an assignment $(c_1, \ldots, c_T)$
  - Each $c_i \in \{1, \ldots, Q\}$
  - Want posterior probability over assignments

Microarray Data

- T gene expression profiles under M conditions
- x_i is the profile for the i^{th} gene
- x_{im} is the value for gene i, condition m
- Each profile is assumed to be generated by one cluster
- Total Q clusters, later $Q \to \infty$
- Clustering is an assignment $(c_1, \ldots, c_T)$
    - Each $c_i \in \{1, \ldots, Q\}$
    - Want posterior probability over assignments
- Idea: Use a Bayesian infinite mixture model

Infinite Mixture Model for Gene Expression

- Level 1: Data generation

$$p(x_i|c_i = j, \mu_h, \sigma_h^2, [h]_1^Q) = \mathcal{N}(x; \mu_j, \sigma_j^2 \mathbb{I})$$

Infinite Mixture Model for Gene Expression

- Level 1: Data generation

$$p(x_i|c_i = j, \mu_h, \sigma_h^2, [h]_1^Q) = \mathcal{N}(x; \mu_j, \sigma_j^2 \mathbb{I})$$

- Level 2: Priors for parameters

$$\begin{aligned} p(\mu_j|\lambda, r) &= \mathcal{N}(\mu; \lambda, 1/r\mathbb{I}) \\ p(\sigma_j^{-2}|\beta, w) &= f_G(\sigma^{-2}; \beta/2, \beta w/2) \end{aligned}$$

Infinite Mixture Model for Gene Expression

- Level 1: Data generation

$$p(x_i|c_i = j, \mu_h, \sigma_h^2, [h]_1^Q) = \mathcal{N}(x; \mu_j, \sigma_j^2 \mathbb{I})$$

- Level 2: Priors for parameters

$$\begin{aligned} p(\mu_j|\lambda, r) &= \mathcal{N}(\mu; \lambda, 1/r\mathbb{I}) \\ p(\sigma_j^{-2}|\beta, w) &= f_G(\sigma^{-2}; \beta/2, \beta w/2) \end{aligned}$$

- Level 2: Prior for clustering $c_i \sim Discrete(\pi)$

Infinite Mixture Model for Gene Expression (Contd.)

- Prior for hyper-parameters

$$
\begin{aligned}
p(w|\sigma_x^2) &= f_G(w; 1/2, 1/(2\sigma_x^2)) \\
p(\beta) &= f_G(\beta; 1/2, 1/2) \\
p(r|\sigma_x^2) &= f_G(r; 1/2, 1/(2\sigma_x^2)) \\
p(\lambda|\mu_x, \sigma_x^2) &= f_N(\lambda|\mu_x, \sigma_x^2\mathbb{I})
\end{aligned}
$$

Infinite Mixture Model for Gene Expression (Contd.)

- Prior for hyper-parameters

$$
\begin{aligned}
p(w|\sigma_x^2) &= f_G(w; 1/2, 1/(2\sigma_x^2)) \\
p(\beta) &= f_G(\beta; 1/2, 1/2) \\
p(r|\sigma_x^2) &= f_G(r; 1/2, 1/(2\sigma_x^2)) \\
p(\lambda|\mu_x, \sigma_x^2) &= f_N(\lambda|\mu_x, \sigma_x^2 \mathbb{I})
\end{aligned}
$$

- (μ_x, σ_x^2) are empirical mean, variance

Infinite Mixture Model for Gene Expression (Contd.)

- Prior for hyper-parameters

$$
\begin{aligned}
p(w|\sigma_x^2) &= f_G(w; 1/2, 1/(2\sigma_x^2)) \\
p(\beta) &= f_G(\beta; 1/2, 1/2) \\
p(r|\sigma_x^2) &= f_G(r; 1/2, 1/(2\sigma_x^2)) \\
p(\lambda|\mu_x, \sigma_x^2) &= f_N(\lambda|\mu_x, \sigma_x^2 \mathbb{I})
\end{aligned}
$$

- (μ_x, σ_x^2) are empirical mean, variance
- Priors on cluster-prior π

$$\pi \sim Dirichlet(\alpha/Q)$$

Gibbs Sampler

- As $Q \to \infty$, we get a DPM

Gibbs Sampler

- As $Q \to \infty$, we get a DPM
- Gibbs sampler for the model

$$p(c_i = j | c_{-i}, \mathbf{x}_i, \mu_j, \sigma_j^2) \quad \propto \quad \frac{n_{-i,j}}{T - 1 + \alpha} \mathcal{N}(x_i; \mu_j, \sigma_j^2 \mathbb{I})$$

$$p(c_i \neq c_j, j \neq i | c_{-i}, x_i, \mu_x, \sigma_x^2) \quad \propto \quad \frac{\alpha}{T - 1 + \alpha} \int \mathcal{N}(x_i | \mu, \sigma^2 \mathbb{I}) p(\mu, \sigma$$

Gibbs Sampler

- As $Q \to \infty$, we get a DPM
- Gibbs sampler for the model

$$p(c_i = j | c_{-i}, \mathbf{x}_i, \mu_j, \sigma_j^2) \propto \frac{n_{-i,j}}{T - 1 + \alpha} \mathcal{N}(x_i; \mu_j, \sigma_j^2 \mathbb{I})$$

$$p(c_i \neq c_j, j \neq i | c_{-i}, x_i, \mu_x, \sigma_x^2) \propto \frac{\alpha}{T - 1 + \alpha} \int \mathcal{N}(x_i | \mu, \sigma^2 \mathbb{I}) p(\mu, \sigma$$

- Pairwise probability of being generated by the same pattern

$$P_{ij} = \frac{\#\text{ samples after `burn-in' with } c_i = c_j}{S_B}$$

Gibbs Sampler

- As $Q \to \infty$, we get a DPM
- Gibbs sampler for the model

$$p(c_i = j | c_{-i}, \mathbf{x}_i, \mu_j, \sigma_j^2) \propto \frac{n_{-i,j}}{T-1+\alpha} \mathcal{N}(x_i; \mu_j, \sigma_j^2 \mathbb{I})$$

$$p(c_i \neq c_j, j \neq i | c_{-i}, x_i, \mu_x, \sigma_x^2) \propto \frac{\alpha}{T-1+\alpha} \int \mathcal{N}(x_i | \mu, \sigma^2 \mathbb{I}) p(\mu, \sigma$$

- Pairwise probability of being generated by the same pattern

$$P_{ij} = \frac{\#\text{ samples after 'burn-in' with } c_i = c_j}{S_B}$$

- Distance $D_{ij} = 1 - P_{ij}$

Model for Replicated Data

- Generate replicates of each expression

Model for Replicated Data

- Generate replicates of each expression
- Account for variability in gene expression data

Model for Replicated Data

- Generate replicates of each expression
- Account for variability in gene expression data
- For G replicates

$$\begin{aligned} p(x_{ik}|y_i,\psi_i) &= \mathcal{N}(x_{ik};y_i,\psi_i^2) \\ p(y_i|c_i=j,\mu_j,\sigma_j) &= \mathcal{N}(y_i|\mu_j,\sigma_j^2\mathbb{I}) \end{aligned}$$

Model for Replicated Data

- Generate replicates of each expression
- Account for variability in gene expression data
- For G replicates

$$\begin{aligned} p(x_{ik}|y_i, \psi_i) &= \mathcal{N}(x_{ik}; y_i, \psi_i^2) \\ p(y_i|c_i = j, \mu_j, \sigma_j) &= \mathcal{N}(y_i|\mu_j, \sigma_j^2 \mathbb{I}) \end{aligned}$$

- Integrating out the mean expression profile

$$p(x_{i1}, \ldots, x_{ik}|c_i = j, \mu_j, \sigma_j^2, \psi_i) = \prod_k \mathcal{N}(\bar{x}_i; \mu_j, (\sigma_j^2 + \frac{\psi_i^2}{G})\mathbb{I})$$

Model for Replicated Data

- Generate replicates of each expression
- Account for variability in gene expression data
- For G replicates

$$
\begin{aligned}
p(x_{ik}|y_i,\psi_i) &= \mathcal{N}(x_{ik};y_i,\psi_i^2) \\
p(y_i|c_i=j,\mu_j,\sigma_j) &= \mathcal{N}(y_i|\mu_j,\sigma_j^2\mathbb{I})
\end{aligned}
$$

- Integrating out the mean expression profile

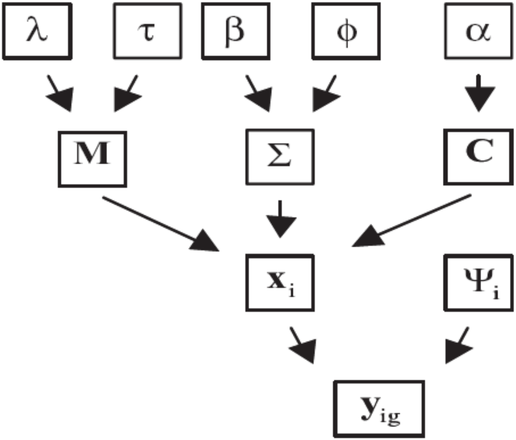
$$
p(x_{i1},\ldots,x_{ik}|c_i=j,\mu_j,\sigma_j^2,\psi_i) = \prod_k \mathcal{N}(\bar{x}_i;\mu_j,(\sigma_j^2+\frac{\psi_i^2}{G})\mathbb{I})
$$

- Gibbs sampler used for inference

Experimental Results

- Move to paper for results

Results



Gibbs Sampler

- For the replicated model

$$p(c_i = q | C_{-i}, x_{i1}, \ldots, x_{ik}, \mu_q, \sigma_q^2, \psi_i, \alpha)$$
$$\propto \frac{n_{-i,q}}{T - 1 + \alpha} \mathcal{N}(\bar{x}_i; \mu_q, (\sigma_q^2 + \psi_i^2 / G)\mathbb{I})$$

$$p(c_i \neq c_j, j \neq i | C_{-i}, x_{ik}, \psi_i, \alpha)$$
$$\propto \frac{\alpha}{T - 1 + \alpha} \int \mathcal{N}(\bar{x}_i; \mu, (\sigma^2 + \psi_i^2 / G)\mathbb{I}) p(\mu, \sigma^2 | \lambda, \tau,$$

Gibbs Sampler

- For the replicated model

$$p(c_i = q | C_{-i}, x_{i1}, \ldots, x_{ik}, \mu_q, \sigma_q^2, \psi_i, \alpha)$$
$$\propto \frac{n_{-i,q}}{T - 1 + \alpha} \mathcal{N}(\bar{x}_i; \mu_q, (\sigma_q^2 + \psi_i^2 / G)\mathbb{I})$$
$$p(c_i \neq c_j, j \neq i | C_{-i}, x_{ik}, \psi_i, \alpha)$$
$$\propto \frac{\alpha}{T - 1 + \alpha} \int \mathcal{N}(\bar{x}_i; \mu, (\sigma^2 + \psi_i^2 / G)\mathbb{I}) p(\mu, \sigma^2 | \lambda, \tau,$$

- The integral is implemented as follows

Gibbs Sampler

- For the replicated model

$$p(c_i = q | C_{-i}, x_{i1}, \ldots, x_{ik}, \mu_q, \sigma_q^2, \psi_i, \alpha)$$
$$\propto \frac{n_{-i,q}}{T - 1 + \alpha} \mathcal{N}(\bar{x}_i; \mu_q, (\sigma_q^2 + \psi_i^2 / G)\mathbb{I})$$
$$p(c_i \neq c_j, j \neq i | C_{-i}, x_{ik}, \psi_i, \alpha)$$
$$\propto \frac{\alpha}{T - 1 + \alpha} \int \mathcal{N}(\bar{x}_i; \mu, (\sigma^2 + \psi_i^2 / G)\mathbb{I}) p(\mu, \sigma^2 | \lambda, \tau,$$

- The integral is implemented as follows
  - Sample μ_p, σ_p from the prior

Gibbs Sampler

- For the replicated model

$$p(c_i = q | C_{-i}, x_{i1}, \ldots, x_{ik}, \mu_q, \sigma_q^2, \psi_i, \alpha)$$
$$\propto \frac{n_{-i,q}}{T - 1 + \alpha} \mathcal{N}(\bar{x}_i; \mu_q, (\sigma_q^2 + \psi_i^2 / G)\mathbb{I})$$
$$p(c_i \neq c_j, j \neq i | C_{-i}, x_{ik}, \psi_i, \alpha)$$
$$\propto \frac{\alpha}{T - 1 + \alpha} \int \mathcal{N}(\bar{x}_i; \mu, (\sigma^2 + \psi_i^2 / G)\mathbb{I}) p(\mu, \sigma^2 | \lambda, \tau,$$

- The integral is implemented as follows
  - Sample μ_p, σ_p from the prior
  - Approximate integral with $\mathcal{N}(\bar{x}_i; \mu_p, (\sigma_p + \psi_i^2 / G)\mathbb{I})$

Gibbs Sample with ‘Reverse Annealing’

- Heuristic to handle mixing problems of Gibbs sampler

Gibbs Sample with ‘Reverse Annealing’

- Heuristic to handle mixing problems of Gibbs sampler
- If π is the target posterior, use flattened target

$$\pi^{(\xi)}(x) = \frac{\pi^{\xi}(x)}{K(\xi)}, \quad \xi < 1$$

Gibbs Sample with 'Reverse Annealing'

- Heuristic to handle mixing problems of Gibbs sampler
- If π is the target posterior, use flattened target

$$\pi^{(\xi)}(x) = \frac{\pi^{\xi}(x)}{K(\xi)}, \quad \xi < 1$$

- The conditional probability can be flattened

$$p(c_i = j | C_{-i}, \Theta) = \frac{p(c_i = j | C_{-i}, \Theta)^{\xi}}{K(\xi)}, \quad \xi < 1$$

Gibbs Sample with 'Reverse Annealing'

- Heuristic to handle mixing problems of Gibbs sampler
- If π is the target posterior, use flattened target

$$\pi^{(\xi)}(x) = \frac{\pi^{\xi}(x)}{K(\xi)}, \quad \xi < 1$$

- The conditional probability can be flattened

$$p(c_i = j | C_{-i}, \Theta) = \frac{p(c_i = j | C_{-i}, \Theta)^{\xi}}{K(\xi)}, \quad \xi < 1$$

- Let $\xi \to 1$ as iterations $n \to \infty$

Results

Table 2. (A) Average adjusted RAND indices for the 'elliptical' simulated data and (B) average number of clusters induced by different IMM algorithms with Auto cluster creation

Method	Number of replicates							
	Low Noise				High Noise			
	1	2	3	4	1	2	3	4
A								
Heuristic methods								
Average link; correlation	0.76	0.76	0.80	0.90	0.20	0.34	0.32	0.43
Average link; distance	0.86	1.00	0.96	1.00	0.00	0.00	0.00	0.00
Complete link; correlation	0.70	0.71	0.75	0.82	0.26	0.39	0.44	0.52
Complete link; distance	0.96	0.93	0.96	0.96	0.07	0.27	0.27	0.53
k-means; correlation	0.76	0.76	0.80	0.90	0.31	0.52	0.57	0.62
k-means; distance	0.86	1.00	0.96	1.00	0.00	0.21	0.16	0.48
SD-adjusted heuristic methods								
Average link; correlation	NA	0.32	0.59	0.61	NA	0.18	0.48	0.59
Average link; distance	NA	0.96	1.00	1.00	NA	0.13	0.47	0.67
Complete link; correlation	NA	0.29	0.46	0.63	NA	0.18	0.48	0.57
Complete link; distance	NA	0.96	1.00	1.00	NA	0.24	0.67	0.83
k-means; correlation	NA	0.47	0.59	0.71	NA	0.22	0.44	0.51
k-means; distance	NA	0.95	1.00	1.00	NA	0.49	0.78	0.77
Model-based methods								
IMM—elliptical	0.93	1.00	1.00	1.00	0.41	0.98	1.00	1.00
IMM—elliptical Auto	0.60	1.00	1.00	1.00	0.34	0.99	1.00	1.00
IMM—spherical	0.93	1.00	0.96	1.00	0.41	0.48	0.66	0.75
IMM—spherical Auto	0.60	0.98	0.99	1.00	0.34	0.82	0.85	0.92
IMM—averaged	0.93	0.93	0.99	1.00	0.41	0.47	0.67	0.77
IMM—averaged Auto	0.60	0.61	0.62	0.64	0.34	0.45	0.46	0.50
FMM—elliptical	0.98	1.00	1.00	1.00	0.30	0.98	1.00	1.00
FMM—spherical	0.98	1.00	1.00	1.00	0.30	0.86	0.89	0.95
FMM—averaged	0.98	0.80	0.95	1.00	0.30	0.44	0.50	0.57
MCLUST-HC	0.97	0.97	0.99	1.00	0.36	0.44	0.57	0.51
MCLUST-FITTS	0.97	0.96	0.97	1.00	0.36	0.38	0.43	0.50
B								
IMM—elliptical Auto	16.2	6.0	6.0	6.0	15.6	6.0	6.0	17.6
IMM—spherical Auto	16.2	7.6	7.8	7.0	15.6	10.4	10.2	6.0

Results

Table 3. (A) Average adjusted RAND indices for the 'spherical' simulated data (B) average number of clusters induced by different IMM algorithms with Auto cluster creation

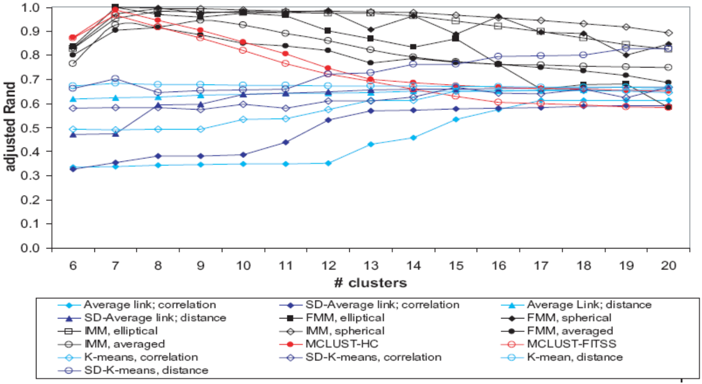
Method	Number of replicates Low Noise 1	2	3	4	High Noise 1	2	3	4
A								
Heuristic methods								
Average link; correlation	0.61	0.67	0.67	0.82	0.25	0.33	0.42	0.40
Average link; distance	0.28	0.70	0.73	0.90	0.00	0.00	0.00	0.00
Complete link; correlation	0.69	0.72	0.73	0.79	0.33	0.46	0.49	0.53
Complete link; distance	0.62	0.76	0.80	0.96	0.00	0.00	0.02	0.09
k-means; correlation	0.70	0.67	0.67	0.82	**0.38**	0.51	0.56	0.57
k-means; distance	0.44	0.70	0.73	0.90	0.00	0.00	0.00	0.00
SD-adjusted heuristic methods								
Average link; correlation	NA	0.35	0.64	0.73	NA	0.14	0.35	0.34
Average link; distance	NA	0.77	0.75	0.90	NA	0.00	0.00	0.00
Complete link; correlation	NA	0.33	0.70	0.77	NA	0.11	0.35	0.49
Complete link; distance	NA	0.77	0.85	0.96	NA	0.09	0.14	0.23
k-means; correlation	NA	0.52	0.75	0.75	NA	0.18	0.53	0.55
k-means; distance	NA	0.83	0.75	0.90	NA	0.14	0.17	0.28
Model-based methods								
IMM—elliptical	**0.90**	0.85	0.91	0.96	0.23	0.80	0.87	0.90
IMM—elliptical Auto	0.52	**1.00**	**1.00**	**1.00**	0.28	0.77	0.87	0.90
IMM—spherical	**0.90**	**1.00**	**1.00**	**1.00**	0.23	**0.84**	**0.90**	**0.93**
IMM—spherical Auto	0.52	**1.00**	**1.00**	**1.00**	0.28	**0.84**	0.89	**0.93**
IMM—averaged	**0.90**	0.94	0.91	0.92	0.23	0.42	0.46	0.64
IMM—averaged Auto	0.52	0.51	0.55	0.56	0.28	0.36	0.40	0.45
FMM—elliptical	0.71	**1.00**	**1.00**	**1.00**	0.23	0.78	0.87	0.90
FMM—spherical	0.71	**1.00**	**1.00**	**1.00**	0.23	**0.84**	**0.90**	0.92
FMM—averaged	0.71	0.89	0.93	0.96	0.23	0.32	0.39	0.48
MCLUST-HC	0.77	0.80	0.85	0.85	0.26	0.24	0.29	0.31
MCLUST-FITTS	0.77	0.90	0.95	0.95	0.26	0.27	0.38	0.35
B								
IMM—elliptical Auto	19.0	6.8	6.4	6.2	15.4	5.8	6.0	6.0
IMM—spherical Auto	19.0	6.0	6.0	6.0	15.4	6.0	6.0	6.0

Results

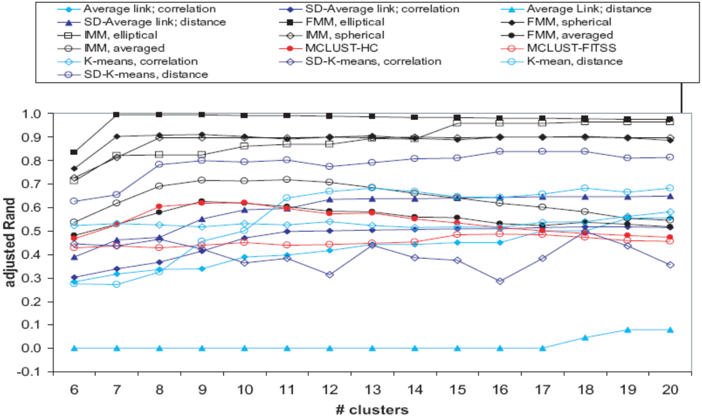
Table 4. Automatic IMM clustering results for datasets with outliers

Noise level	Variance	RAND	No. of clusters
Low	Averaged	0.78	13.4
Low	Elliptical	0.99	7.6
Low	Spherical	0.99	8
High	Averaged	0.51	16.6
High	Elliptical	0.97	13
High	Spherical	0.89	8.8

Results


adjusted Rand
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
clusters
Average link; correlation
SD-Average link; correlation
Average Link; distance
SD-Average link; distance
FMM, elliptical
FMM, spherical
IMM, elliptical
IMM, spherical
FMM, averaged
IMM, averaged
MCLUST-HC
MCLUST-FITSS
K-means, correlation
SD-K-means, correlation
K-mean, distance
SD-K-means, distance

Results



Average link; correlation
SD-Average link; correlation
Average Link; distance
SD-Average link; distance
FMM, elliptical
FMM, spherical
IMM, elliptical
IMM, spherical
FMM, averaged
IMM, averaged
MCLUST-HC
MCLUST-FITSS
K-means, correlation
SD-K-means, correlation
K-mean, distance
SD-K-means, distance
adjusted Rand
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
clusters

Results

Method	Number of replicates 1	2	3	4
A				
Heuristic methods				
Average link; correlation	**0.88**	0.87	0.91	0.87
Average link; distance	0.67	0.90	0.92	0.86
Complete link; correlation	0.69	0.78	0.70	0.68
Complete link; distance	0.62	0.85	0.88	0.96
K-means; correlation	0.75	0.77	0.77	0.87
K-means; distance	0.86	0.92	0.91	0.86
SD-adjusted heuristic methods				
Average link; correlation	NA	0.68	0.82	0.82
Average link; distance	NA	0.84	0.86	0.86
Complete link; correlation	NA	0.52	0.67	0.72
Complete link; distance	NA	0.83	0.96	0.97
K-means; correlation	NA	0.63	0.75	0.64
K-means; distance	NA	0.79	0.85	0.86
Model-based methods				
IMM—elliptical	0.85	0.92	**0.97**	**0.97**
IMM—elliptical Auto	0.78	**0.94**	0.94	0.96
IMM—averaged	0.85	0.92	0.94	**0.97**
IMM—averaged Auto	0.78	0.93	0.96	0.96
FMM—elliptical	0.71	0.75	0.93	**0.97**
IMM—averaged	0.71	0.85	0.91	**0.97**
MCLUST-HC	0.68	0.90	0.96	**0.97**
MCLUST-FITTS	0.68	0.86	0.89	**0.97**
B				
IMM—elliptical Auto	2.8	4.2	4.8	5.0
IMM—averaged Auto	2.8	4.7	4.5	5.0