BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–3

Suppl. Material to Robust multi-scale clustering of large DNA microarray datasets with the consensus algorithm

Thomas Grotkjær [a,*], Ole Winther [b], Birgitte Regenberg [a], Jens Nielsen [a] and Lars Kai Hansen [b]

[a]Center for Microbial Biotechnology, BioCentrum-DTU, Building 223 and
[b]Informatics and Mathematical Modelling, Building 321,
Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

MIXTURE OF GAUSSIANS

We cluster the data by modelling the data vectors $\mathbf{x}$ as a Mixture of K Gaussians (MoG)

$$p(\mathbf{x}|\boldsymbol{\theta},\mathcal{M}) = \sum_{k=1}^{K} p(k|\boldsymbol{\theta},\mathcal{M})p(\mathbf{x}|k,\boldsymbol{\theta},\mathcal{M}) \ , \qquad (1)$$

where $\boldsymbol{\theta}$ is the model parameters, $\mathcal{M}$ is the model, $p(k|\boldsymbol{\theta},\mathcal{M})$ is the mixing proportions (which sum to one: $\sum_k p(k|\boldsymbol{\theta},\mathcal{M}) = 1$) and $p(\mathbf{x}|k,\boldsymbol{\theta},\mathcal{M})$ is Gaussian with mean vector $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$:

$$p(\mathbf{x}|k,\boldsymbol{\theta},\mathcal{M}) = \frac{1}{\sqrt{(2\pi)^M \det \boldsymbol{\Sigma}_k}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right) \ . \qquad (2)$$

The parameters of the MoG are thus $\boldsymbol{\theta} = \{(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)|k = 1,\ldots,K\}$ with $\pi_k \equiv p(k|\boldsymbol{\theta},\mathcal{M})$. In maximum likelihood (ML) MoG the parameters are estimated as $\boldsymbol{\theta}^{\mathrm{ML}} = \mathrm{argmax}_{\boldsymbol{\theta}}\, L(\boldsymbol{\theta})$ where the objective is the log-Likelihood $L(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\mathbf{x}_n|\boldsymbol{\theta},\mathcal{M})$. This naturally leads to a set of iterative expectation maximisation (EM) updates which are guaranteed to converge to a local maximum of the likelihood. In the E-step, the responsibility γ_{kn} which is the probability of cluster k given gene n is updated for fixed parameters as follows

$$\gamma_{kn} \equiv p(k|\mathbf{x}_n,\boldsymbol{\theta}) = \frac{p(k)p(\mathbf{x}_n|k,\boldsymbol{\theta})}{\sum_{k'} p(k')p(\mathbf{x}_n|k',\boldsymbol{\theta})} \ . \qquad (3)$$

In the M-step, the parameters are estimated for fixed responsibilities (defining $N_k = \sum_n \gamma_{kn}$):

$$\pi_k = \frac{1}{N} N_k \qquad (4)$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_n \gamma_{kn} \mathbf{x}_n \qquad (5)$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_n \gamma_{kn} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \ . \qquad (6)$$

After convergence the responsibilities can be used to make clustering by assigning gene n to the cluster which is the most probable.

BAYESIAN APPROACH

The ML MoG has a fatal flaw, namely trivial solutions exist with diverging likelihood when $\boldsymbol{\mu}_k = \mathbf{x}_n$ and $\boldsymbol{\Sigma}_k \to 0$ (MacKay, 2003). This problem can be understood and cured by a Bayesian approach where a prior $p(\boldsymbol{\theta})$ is introduced. We can now argue that for any reasonable prior this singular solution has a vanishing posterior probability $p(\boldsymbol{\theta}|\mathcal{D}_N)d\boldsymbol{\theta} \propto \prod_n p(\mathbf{x}_n|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ although it's probability density diverges, i.e. the peak is narrow, but the probability volume (the integral of the density in a local neighbourhood around the maximum) is negligible. In the Bayesian approach we aim at calculating the quantities of interest by averaging out uncertainty, e.g. the marginal likelihood

$$p(\mathcal{D}_N|\mathcal{M}) = \int d\boldsymbol{\theta} \prod_n p(\mathbf{x}_n|\boldsymbol{\theta},\mathcal{M})p(\boldsymbol{\theta}|\mathcal{M}) \qquad (7)$$

can be used for model order selection, i.e. with no prior expectations about what K should be, we will choose the K that maximizes the probability of the data.

*to whom correspondence should be addressed

© Oxford University Press 2005.

We can find the cluster assignment probability for a data point $\mathbf{x}$ by averaging out the parameters using the posterior distribution

$$p(k|\mathbf{x}, \mathcal{D}_N, \mathcal{M}) = \frac{p(k, \mathbf{x}|\mathcal{D}_N, \mathcal{M})}{p(\mathbf{x}|\mathcal{D}_N, \mathcal{M})} \ . \tag{8}$$

where

$$p(k, \mathbf{x}|\mathcal{D}_N, \mathcal{M}) = \int d\boldsymbol{\theta}\, p(k|\theta, \mathcal{M}) p(\mathbf{x}|k, \theta, \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{D}_N, \mathcal{M}) \tag{9}$$

and $p(\mathbf{x}|\mathcal{D}_N, \mathcal{M}) = \sum_k^K p(k, \mathbf{x}|\mathcal{D}_N, \mathcal{M})$.

VARIATIONAL BAYES

Unfortunately, the Bayesian approach is not analytically tractable and we have to resort to approximate approaches. In the variational Bayes (VB) approach (Attias, 2000), the priors are chosen in the exponential conjugate family (i.e. a Dirichlet for $\boldsymbol{\pi}$ and a coupled Gaussian-Wishart for $\boldsymbol{\mu}_k$ and $\boldsymbol{\Gamma}_k = \boldsymbol{\Sigma}_k^{-1}$) and the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D}_N)$ is approximated by a simpler tractable partly factorized distribution $q(\boldsymbol{\theta}) = q(\boldsymbol{\pi}) \prod_k q(\boldsymbol{\mu}_k, \boldsymbol{\Gamma}_k)$. $q(\boldsymbol{\theta})$ is found by minimising the Kullback-Leibler (KL) divergence between q and p

$$KL(q||p) = \int d\boldsymbol{\theta}\ q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D}_N)} \ . \tag{10}$$

The hyperparameters of the priors are chosen such that the singularities are removed while at the same time the prior will have vanishing influence on the clustering. The reason for choosing the prior in the exponential conjugate family is that for this choice solving for $q(\boldsymbol{\theta})$ simply amounts to using an iterative update of the parameters of the q-distribution in the same fashion as with the EM algorithm. As a final benefit of the VB approach, we get guaranteed convergence to a maximum of the following lower bound of the marginal likelihood

$$p_{\mathrm{app}}(\mathcal{D}_N|\mathcal{M}) = \int d\boldsymbol{\theta} q(\boldsymbol{\theta}) \log \frac{p(\mathcal{D}_N, \boldsymbol{\theta}|\mathcal{M})}{q(\boldsymbol{\theta})} \leq p(\mathcal{D}_N|\mathcal{M}) \ . \tag{11}$$

VBMoG has previously been applied to DNA microarrays by (Muro *et al.*, 2003). Note that ML methods can also be used for model selection (determine the number of clusters) by subtracting complexity penalizing terms from the log-likelihood. These criteria known as Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) have been also applied in context of MoG and DNA microarrays (McLachlan *et al.*, 2002; Ghosh & Chinnaiyan, 2002; Pan *et al.*, 2002). The lower bound of the marginal likelihood above reduces to BIC in the asymptotic limit of a large number of samples (Attias, 2000).

The clustering results are not strongly dependent upon the initialization of the parameters and the choice of hyperparameters. The hyperparameters were set to the following values (using the same notation as (Attias, 2000)): $\lambda_0 = 1$, $\beta_0 = 10^{-12}$, $\boldsymbol{\rho}_0 = \mathbf{0}$, $\boldsymbol{\Phi}_0 = 10^{-3}\mathbf{K}$ and $\nu_0 = M + 1$, where $\mathbf{K}$ is the empirical $K \times K$ covariance matrix of the data. The parameters were initialized to the following values: $\lambda_k = \lambda_0$, $\beta_k = 1 + \beta_0$, $\boldsymbol{\Phi}_k = \mathbf{K}$, $\boldsymbol{\rho} \sim \mathcal{N}(\mathbf{m}, \mathbf{K})$, where $\mathbf{m}$ is the empirical mean of the data. For example, scaling $\boldsymbol{\Phi}_0$ and $\boldsymbol{\Phi}_k$ with factors $\in [10^{-5}; 10^1]$ and $\in [10^{-2}; 10^4]$, respectively, did not change the results significantly.

ASSESSING THE NUMBER OF REPETITIONS

It is in general difficult to assess when a sampling based method has converged (MacKay, 2003), i.e. are the samples actually coming from the distribution we want to draw samples from? A similar problem arises in consensus clustering: When will the final consensus clustering not change significantly if we make more repetitions? One way to answer this is through resampling (or cross-validation) methods. One could for example measure the difference (using some suitable norm) between the co-occurence matrix calculated using all repetions (the full cluster ensemble) and using cluster ensembles which are generated from the full cluster ensemble by sub-sampling (cross-validation) or sampling with replacement (bootstrap).

Here we will assess the problem in a different computationally cheaper way that uses the fact that the averaged mutual information indicates how diverse the cluster ensemble is. We make a simplistic model for the way the cluster assignment $a(\mathbf{x})$ is generated from the true unknown assignment $a^{\mathrm{true}}(\mathbf{x})$: Introducing a 'flip probability' ϵ, we assume that the assignment is 'flipped' to belong to cluster k with probability $\epsilon p^{\mathrm{true}}(k)$, where $p^{\mathrm{true}}(k)$ is (the true and thus unknown) fraction of examples belonging to cluster k. With probability $(1 - \epsilon)$ the assignment is unchanged. This 'noise process' introduces false positive entries in the co-occurrence matrix, i.e. pairs that should not be in the same cluster will get a non-zero entry. We should choose the number of repetitions such that true positive (TP) entries (pairs of transcripts belonging together) should be larger than false positive (FP) entries with a safe margin.

We can calculate the expected mutual information as a function of ϵ and $\{p^{\mathrm{true}}(k)\}_{k=1}^K$:

$$\begin{aligned} \langle M_{rr'} \rangle \ &= \ \sum_k \left[(1-\epsilon)^2 p_k + (2\epsilon - \epsilon^2) p_k^2 \right] \\ &\quad \times \log \left[(1-\epsilon)^2 / p_k + (2\epsilon - \epsilon^2) \right] \\ &\quad + \left(1 - \sum_k p_k^2 \right) (2\epsilon - \epsilon^2) \log(2\epsilon - \epsilon^2) \ , \end{aligned} \tag{12}$$

where $p_k \equiv p^{\mathrm{true}}(k)$. The entropy of the marginal distributions $M_r = -\sum_k p_r(k) \log p_r(k)$ becomes on average equal to the entropy of the true marginals $\langle M_r \rangle = -\sum_k p_k \log p_k$. For large number of transcripts we can ignore fluctuations and set $\langle \overline{M}^{\mathrm{norm}} \rangle = \langle M_{rr'} \rangle / \langle M_r \rangle$. This connects the

observed averaged normalised mutual information with the unknown quantities. Next we can calculate the statistics of the entries in the co-occurrence model under this noise model ($a_n \equiv a(\mathbf{x}_n)$) for $n \neq n'$:

$$\begin{aligned} \langle C_{nn'} \rangle &= \langle a_n a_{n'} \rangle = \langle a_n \rangle \langle a_{n'} \rangle \\ &= \left[(1-\epsilon) a_n^{\mathrm{true}} + \epsilon \sum_k p_k \delta_{a_n,k} \right] \\ &\quad \times \left[(1-\epsilon) a_{n'}^{\mathrm{true}} + \epsilon \sum_k p_k \delta_{a_{n'},k} \right] \end{aligned} \tag{13}$$

If $a_n^{\mathrm{true}} = a_{n'}^{\mathrm{true}}$ this becomes:

$$C^{\mathrm{TP}} = (1-\epsilon)^2 + 2\epsilon(1-\epsilon) p_{a_n} + \epsilon^2 \sum_k p_k^2 \tag{14}$$

and if $a_n^{\mathrm{true}} \neq a_{n'}^{\mathrm{true}}$:

$$C^{\mathrm{FP}} = \epsilon(1-\epsilon)(p_{a_n} + p_{a_{n'}}) + \epsilon^2 \sum_k p_k^2 \,. \tag{15}$$

The expected entries (over repetitions) for the true positive entries will approximately be normal distributed with mean C^{TP} and variance $V^{\mathrm{TP}} \equiv C^{\mathrm{TP}}(1-C^{\mathrm{TP}})/R$, where R is the number of repetitions. Demanding that the probability that a true positive entry is smaller than a false positive should be small, say below $r = 5$ standard deviations we have the following condition:

$$C^{\mathrm{TP}} - r\sqrt{V^{\mathrm{TP}}} \geq C^{\mathrm{FP}} + r\sqrt{V^{\mathrm{FP}}} \,. \tag{16}$$

This condition determines R in terms of the unknowns. If we know $p^{\mathrm{true}}(k)$, we can use the observed value of $\overline{M}^{\mathrm{norm}}$ to estimate the remaining unknown ϵ. In our program we simply set $p^{\mathrm{true}}(k) = 1/K$. It turns out that the expression for the mutual information is quite insensitive to the choice of K and we use $K = 10$. A flat distribution of $p^{\mathrm{true}}(k)$ is the ideal situation because in the non-flat case small clusters can 'drown' in the noise from the larger clusters. However by choosing the rather conservative margin $r = 5$ we hope to compensate for that and so far our program has given sensible estimates. See Figure 1 for a plot of required repetitions versus $\overline{M}^{\mathrm{norm}}$. The result is found by solving numerically for ϵ in terms of $\overline{M}^{\mathrm{norm}}$ and inserting $\epsilon(\overline{M}^{\mathrm{norm}})$ in the condition above.

SENSITIVITY ANALYSIS OF GENERATIVE MODEL

The performance of the selected clustering algorithms was investigated for different variations of the generative model. Generation of new Gaussian noise to the model parameters for each clustering run did not alter the simulation results significantly (Figure 2a–c). Randomization of the 'signal values' (Table 1 in main text) resulted in a general decreased classification error rate for all algorithms. This is also expected since the simulated dataset was constructed in such a way that there was a high resemblance between clusters 1 and 3 as well as clusters 4 and 6 (Figure 2d–f).

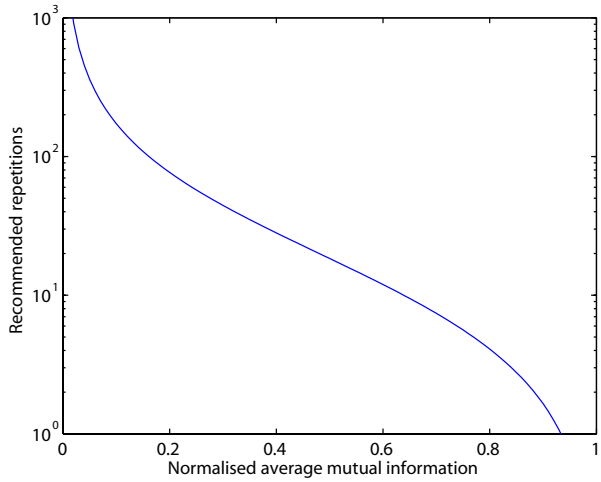

Fig. 1. The required number of repetitions versus the average normalised mutual information $\overline{M}^{\mathrm{norm}}$ for $p^{\mathrm{true}}(k) = 1/K$ with $K = 10$. Because the correct K is not known, we use the predicted number of repetitions for each K in the tested interval. Example values are $\overline{M}^{\mathrm{norm}} = \{0.1, 0.5, 0.8\}$ give 173, 18 and 4.

REFERENCES

Attias, H. (2000) A variational Bayesian framework for graphical models. Adv. Neur. Info. Proc. Sys. MIT Press.

Ghosh, D. & Chinnaiyan, A. M. (2002) Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, **18** (2), 275–286.

MacKay, D. J. C. (2003) *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge.

McLachlan, G. J., Bean, R. W. & Peel, D. (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18** (3), 413–422.

Muro, S., Takemasa, I., Oba, S., Matoba, R., Ueno, N., Maruyama, C., Yamashita, R., Sekimoto, M., Yamamoto, H., Nakamori, S., Monden, M., Ishii, S. & Kato, K. (2003) Identification of expressed genes linked to malignancy of human colorectal carcinoma by parametric clustering of quantitative expression data. *Genome Biol.*, **4** (3), 1–10.

Pan, W., Lin, J. & Le, C. T. (2002) Model-based cluster analysis of microarray gene-expression data. *Genome Biol.*, **3** (2), 1–9.

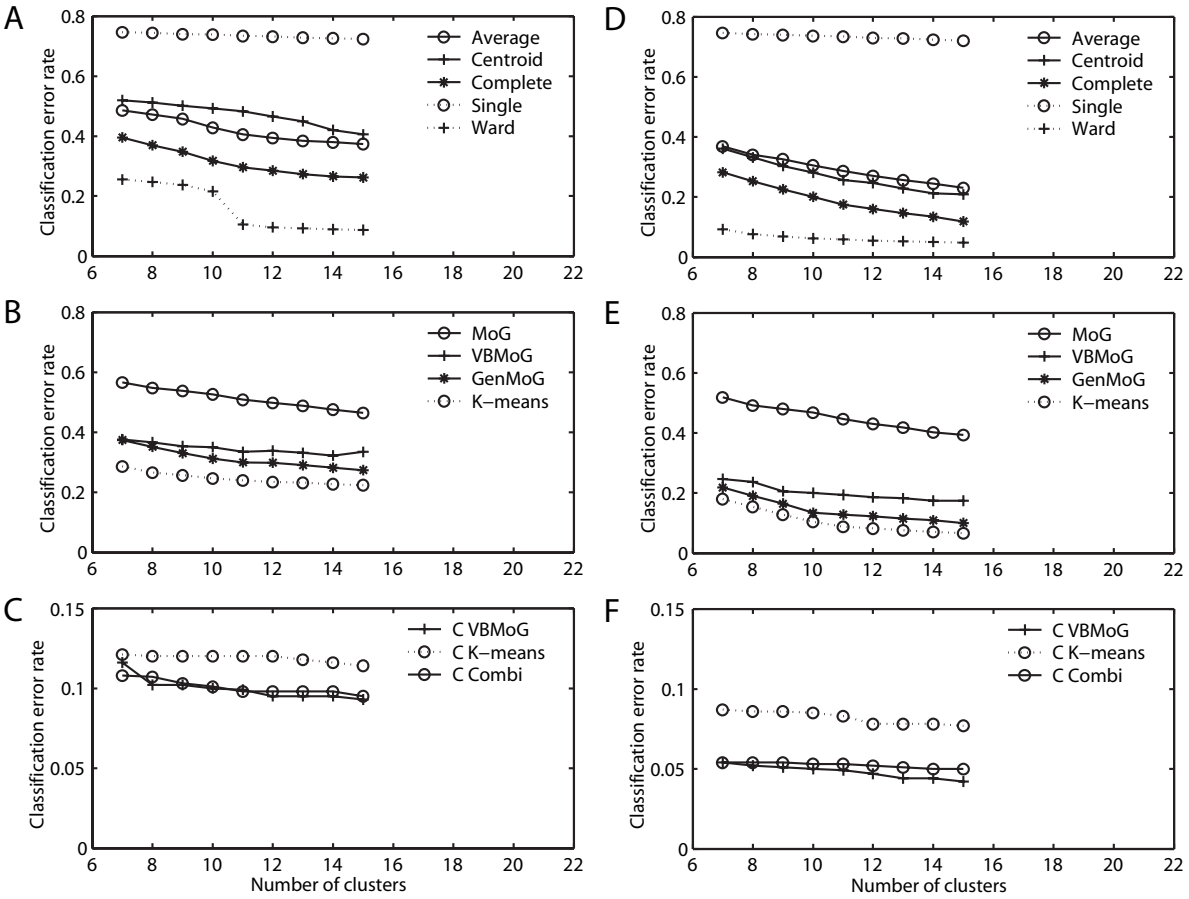

Fig. 2. Classification error rate as a function of number of clusters for selected clustering methods. In **a–c** new Gaussian (Eq. 8 in main text) noise is generated for each clustering run. In **d–f** the signal values are randomised and new Gaussian noise are generated for each clustering run. Note, the classification error rates in **c** and **f** (y-axis scale changed) are much smaller compared to the results in **a**, **b**, **d** and **e**. See Figure 4 in main text for further explanation.