Mixture modelling in a fully Bayesian framework with reference to microarray gene expression data

Natalia Bochkina, Alex Lewin, Sylvia Richardson

Imperial College London, UK

N.Bochkina@imperial.ac.uk
www.bgx.org.uk

Differential expression in microarrays

We consider the problem of differential expression between two groups in microarray gene expression data y_{gcr} observed for gene $g = 1, \dots, G$, condition $c = 1, 2$ and replicate $r = 1, \dots, k_c$. We assume that the data have been appropriately normalised to account for technical noise and transformed to follow a normal distribution:

$$\begin{aligned} y_{g1r}|\alpha_g, d_g, \sigma^2_{g1} &\sim N(\alpha_g - d_g/2, \sigma^2_g), \quad r = 1, \dots, k_1, \\ y_{g2r}|\alpha_g, d_g, \sigma^2_{g2} &\sim N(\alpha_g + d_g/2, \sigma^2_g), \quad r = 1, \dots, k_2, \end{aligned}$$

where α_g is the average value of the gene, d_g is the difference between conditions and σ^2_g is the variance of gene g.
Testing hypothesis of differential expression:

$$H_0 : d_g = 0 \quad vs \quad H_1 : d_g \neq 0. \tag{1}$$

Bayesian hierarchical model

A natural Bayesian way to test hypothesis H_0 is to consider a mixture prior for the difference d_g:

$$d_g \sim \pi_0 \delta_0(d_g) + (1 - \pi_0) h(d_g|\eta),$$

where δ_0 is the delta function and h is the density of d_g under the alternative with a vector of parameters η (Lonnstedt and Speed 2003, Smyth 2004, Scott and Berger 2003).

What is the best way to specify the prior distribution h?

We propose to use a heavy tail asymmetric formulation:

$$d_g \sim \pi_0 \delta_0(x) + \pi_1 \Gamma(-x \mid 1.5, \lambda_1) + \pi_2 \Gamma(x \mid 1.5, \lambda_2), \tag{2}$$

which is well suited to the prior knowledge about d_g:

- most of d_g under H_1 have small non-zero values - peaks of the gamma density at $(-1)^c \frac{1}{2\lambda_c}$, $c = 1, 2$;
- there may be a few genes with large values - heavy tailed distribution
- possible asymmetry of positive and negative values of d_g.

We consider non-informative prior for the mean α_g and exchangeable model for the variances

$$\begin{aligned} \alpha_g &\sim 1, \\ \sigma_g^{-2}|a, b &\sim \Gamma(a, b), \end{aligned}$$

where hyperparameters a and b have non-informative priors priors $f(x) = 1/x$, $x > 0$.
Under the specified model, the posterior estimates of mean, conditional on the non-zero component, and odds are

$$\begin{aligned} E(d_g|I_g \neq 0, y_{gcr}, a, b, \eta) &= \frac{E_X X G_g^{a+(k_1+k_2)/2-1}(X)}{E_X G_g^{a+(k_1+k_2)/2-1}(X)}, \\ w_g = Odds(I_g \neq 0|y_{gcr}, a, b, \eta) &= \frac{1 - \pi_0}{\pi_0} \frac{E_X G_g^{(k_1+k_2)/2+a-1}(X)}{G_g^{(k_1+k_2)/2+a-1}(0)}. \end{aligned} \tag{3}$$

where $G_g(x) = (b + (k_1+k_2)s_g^2/2 + k(x - \bar{y}_g)^2/2)^{-1}$, $k = \frac{k_1 k_2}{k_1+k_2}$, X has density $h(x|\eta)$, E_X refers to expectation with respect to X, $\bar{y}_g$ and s_g^2 are sample mean and sample variance, and $I_g = \text{sign}(d_g)$ is the indicator function of the sign of the true value of d_g.
These formulae show how distribution h and the data enter the posterior distributions.

References

1. Berger, J. and Pericchi, L. (2001) "Objective Bayesian methods for model selection: introduction and comparison" (with discussion). In 'Model Selection' (P.Lahiri, editor), Institute of Mathematical Statistics Lecture Notes – Monograph Series volume 38.
2. Bochkina N., Richardson S. (2006) "Tail posterior probability for inference in pairwise and multiclass gene expression data", Biometrics (under revision).
3. Bochkina N.A., Sapatinas T. (2005) "On the Posterior Median Estimators of Possibly Sparse Sequences", Annals of Institute of Statistical Mathematics, Vol. 57, 315-351.

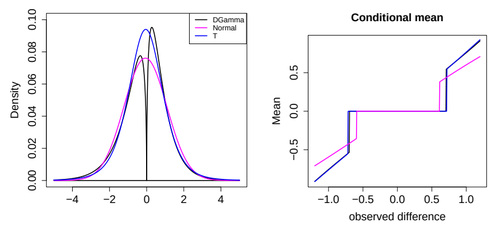
Comparison to other mixture models

We compare the asymmetric gamma model to two other mixture models:

- with normal $h(x|\eta) = N(\mu, \tau)$,
- with Student t $h(x|\eta) = t(\nu, \mu, \tau)$.

Do not consider conjugate model $h(x, \eta) = N(0, c\sigma_g^2)$ (Berger & Pericchi 2001), since $BF(H_0/H_1) \to \left(1 + \frac{k_1 k_2 c}{k_1+k_2}\right)^{-a-(k_1+k_2)/2+1} \neq 0$ as $\bar{y}_g^2/s_g^2 \to \infty$, i.e. with overwhelming evidence in favour of H_1.

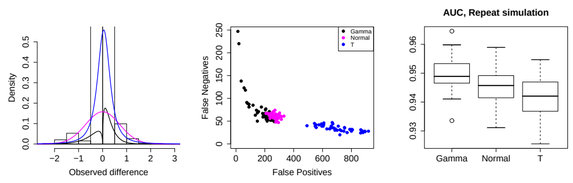
1. We use formula (3) to compare the posterior mean $E(d_g|I_g, \bar{y}_g, s_g^2)$, conditional on the component with higher posterior odds, for the three mixture models, with fixed values of the hyperparameters and s_g^2 ($s_g^2 = 1.63$).



Density $(1 - \pi_0)h(x|\eta)$, conditional posterior mean of d_g.

- Similar performance of different mixture models;
- model with larger π_0, has a smaller threshold for $\bar{y}_g$ to accept the null hypothesis;
- for the heavy tailed distributions, asymmetric double gamma and t distributions, estimated value of d_g is less shrunk towards zero compared to that under the normal model (see Bochkina & Sapatinas 2005 for details).

2. Now we consider a simulation with 3000 variables, 6 replicates, repeated 34 times, where the difference is slightly asymmetric:
$d_g \sim 0.05g(-x; \theta = 1.2) + 0.1g(x; \theta = 0.5) + 0.85N(0, 0.01)$,
$\sigma_g^2 \sim 0.03 + \text{LogN}(-3.85, 0.82)$,
$\alpha_g \sim N(7, 25)$,
where $g(x; \theta) = 0.2U[0, 2.5] + 0.4U[0.07, \theta] + 0.4(\theta + |N(0, 0.7)|)$.
We fit each of 3 mixture models in a fully Bayesian way (with vague hyperpriors).



Estimated densities of the difference $(1 - \pi_0)h(x|\eta)$ (single simulation), number of false positives and false negatives for the Bayesian rule, area under ROC curve.

- Due to the heavy tails of t distribution, it can "absorb" the differentially expressed genes in its tails, resulting in a higher rate of false positives.
- Gamma and normal models perform similarly on average, with the gamma model being marginally better in the asymmetric case.
- Gamma model can result in lower number of false positives than other models, and the model with t distribution has the lower number of false negatives.

4. Scott, J., and Berger, J. (2003). "An exploration of aspects of Bayesian multiple testing". Technical Report, Duke University.
5. Smyth, G. (2004) "Linear models and empirical Bayes methods for as- sessing differential expression in microarray experiments", Statistical Applications in Genetics and Molecular Biology 3: No. 1, Article 3. http://www.bepress.com/sagmb/vol3/iss1/art3.
6. Storey, J. (2002) "A direct approach to false discovery rates". Journal of the Royal Statistical Society, Series B 64, 479-498.

Comparison with 'objective prior' for d_g.

Consider (improper) flat prior for $d_g \sim 1$, and take the tail posterior probability as a classification rule:

$$p(t_g, \theta) = P\left\{|t_g| > \theta | y_{gcr}\right\},$$

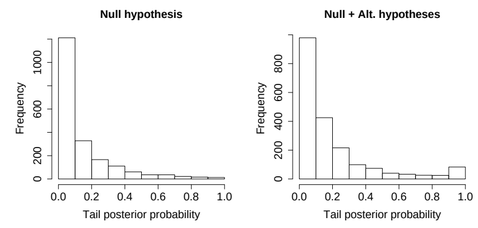
where $t_g = \frac{\sqrt{k} d_g}{\sigma_g}$ - standardised difference, $\theta > 0$ - threshold (see Bochkina and Richardson 2006 for details).

Distribution of t_g: $t_g|\bar{y}_g, \sigma_g^2 \sim N(\bar{y}_g \sigma_g/\sqrt{k}, 1)$

Threshold θ: a quantile of the posterior distribution of t_g given $\bar{y}_g = 0$, i.e. for data with the sufficient statistic for d_g equal to its value under the null hypothesis, and with the same s_g^2.

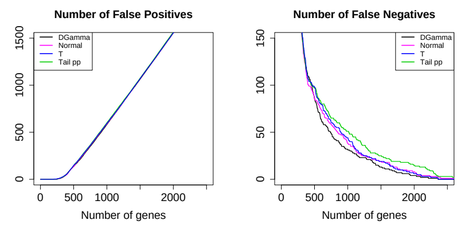
In this case, $\theta = t^{(\alpha)} = \Phi^{-1}(1 - \alpha/2)$ is gene independent!

Histogram under the null and alternative hypotheses.



Under H_0, distribution of $P\{|t_g| > t_g^\alpha | y_{gcr}\}$ is gene-independent which can be used to estimate the false discovery rate using approach of Storey 2002. Numerical integration is required to calculate the distribution function of $P\{|t_g| > t_g^\alpha | y_{gcr}\}$ (as a test statistic) under H_0.

Number of false discoveries of the mixture models and tail posterior probability in a gene list.



Advantages of the approaches:
objective: do not model the alternative distribution
mixture: less challenging to estimate π_0 and FDR.

Conclusions

- Mixture models show good performance, and π_0 - proportion of variables under H_0 is well estimated
- Asymmetric double Gamma and Normal models show better classification properties than the model with t distribution in the context of the gene expression data.
- The best model for penalty ratio $R = \frac{l(H_0, \text{reject } H_0)}{l(H_1, \text{reject } H_1)}$ for the losses of false positives to false negatives is
  - $R > 1$ - asymmetric double gamma model;
  - $R \approx 1$ - normal model;
  - $R < 1$ - model with t distribution;
- Estimate of d_g under the normal model is more shrunk towards 0 than under the models with heavy tails (gamma and t).
- Similar performance as that of the model with the objective prior, however, in a mixture model, π_0 and FDR are better estimated.

For model checks for mixture models, see poster by A.Lewin and S.Richardson.

This work was supported by Wellcome Trust.