

A Bayesian calibration model for combining different pre-processing methods in Affymetrix chips

Measurement error and model criticism

Marta Blangiardo
joint work with Sylvia Richardson

Imperial College Centre for Biostatistics
m.blangiardo@imperial.ac.uk

Bayesian Biostatistics Conference, Houston TX

26-28 January 2009

Outline

- 1 Introduction to gene expression data
- 2 Importance of pre-processing methods in gene expression data analysis
- 3 Bayesian calibration model to combine several pre-processing methods
- 4 Strategy of model criticism
- 5 Performance on simulated data and on a biological experiment
- 6 General Remarks

Outline

- 1 Introduction to gene expression data
- 2 Importance of pre-processing methods in gene expression data analysis
- 3 Bayesian calibration model to combine several pre-processing methods
- 4 Strategy of model criticism
- 5 Performance on simulated data and on a biological experiment
- 6 General Remarks

Outline

- 1 Introduction to gene expression data
- 2 Importance of pre-processing methods in gene expression data analysis
- 3 Bayesian calibration model to combine several pre-processing methods
- 4 Strategy of model criticism
- 5 Performance on simulated data and on a biological experiment
- 6 General Remarks

Outline

- 1 Introduction to gene expression data
- 2 Importance of pre-processing methods in gene expression data analysis
- 3 Bayesian calibration model to combine several pre-processing methods
- 4 Strategy of model criticism
- 5 Performance on simulated data and on a biological experiment
- 6 General Remarks

Outline

- 1 Introduction to gene expression data
- 2 Importance of pre-processing methods in gene expression data analysis
- 3 Bayesian calibration model to combine several pre-processing methods
- 4 Strategy of model criticism
- 5 Performance on simulated data and on a biological experiment
- 6 General Remarks

Outline

- 1 Introduction to gene expression data
- 2 Importance of pre-processing methods in gene expression data analysis
- 3 Bayesian calibration model to combine several pre-processing methods
- 4 Strategy of model criticism
- 5 Performance on simulated data and on a biological experiment
- 6 General Remarks

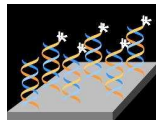
Statistics for gene expression data

- Extracting the message from microarray data needs statistical as well as biological understanding.
- Statistical modelling, in contrast to data analysis, gives a framework for formally organising assumptions about signal and noise.
- Need of structured models that reflect data generation process:
 - Bayesian hierarchical modelling approach
 - Inference based on posterior distribution of quantities of interest
 - Predictive check and model refinement

What are gene expression data?

What are gene expression data?

DNA Microarrays are used to measure the relative abundance of mRNA, providing information on gene expression in a particular cell type, under specific conditions



What are gene expression data?

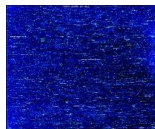
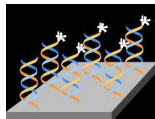
DNA Microarrays are used to measure the relative abundance of mRNA, providing information on gene expression in a particular cell type, under specific conditions

Gene expression data (e.g. Affymetrix™) results from the scanning of arrays where hybridisation between a sample and a large number of probes has taken place:

- gene expression measure for each gene.

The expression level of ten of thousands of probes are measured on a single microarray:

- gene expression profile



What are gene expression data?

DNA Microarrays are used to measure the relative abundance of mRNA, providing information on gene expression in a particular cell type, under specific conditions

Gene expression data (e.g. Affymetrix™) results from the scanning of arrays where hybridisation between a sample and a large number of probes has taken place:

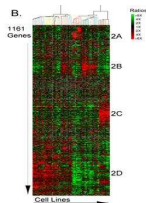
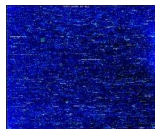
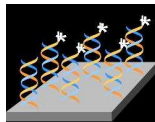
- gene expression measure for each gene.

The expression level of ten of thousands of probes are measured on a single microarray:

- gene expression profile

Typically, gene expression profiles are obtained for several samples, in a single or related experiments:

- gene expression data matrix



What are gene expression data?

DNA Microarrays are used to measure the relative abundance of mRNA, providing information on gene expression in a particular cell type, under specific conditions

Gene expression data (e.g. Affymetrix™) results from the scanning of arrays where hybridisation between a sample and a large number of probes has taken place:

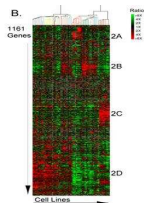
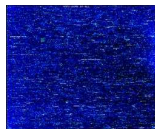
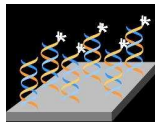
- gene expression measure for each gene.

The expression level of ten of thousands of probes are measured on a single microarray:

- gene expression profile

Typically, gene expression profiles are obtained for several samples, in a single or related experiments:

- gene expression data matrix



Different questions are exploited: gene discovery (supervised and unsupervised analysis), gene comparison (differential expression)

Pre-processing: open issue in the analysis of gene expression data

- Some of the variability between the conditions is due to the technology used (technical variability)
- “Pre-processing” involves a series of steps to extract the signal and to account for technical variability:
 - Background correction
 - Normalisation
 - Summarisation
- Different pre-processing techniques lead to different results.
- There is no agreed gold standard

Combining several pre-processing methods: an example of Bayesian synthesis for differential expression

- We propose to combine the measures of differential expression coming from several pre-processing methods
- Measurement error perspective: each pre-processing gives a surrogate measure for the true (latent) differential expression
- General measurement error model for differential expression

$$Z_{gj} = X_g + U_j$$

- X_g is the latent true differential expression for the gene g
- Z_{gj} is the differential expression measured by the pre-processing j on the gene g
- U_j is the bias for the pre-processing j

Bayesian calibration model: first level modelling

Condition 1

$$y_{gj1r} \sim N\left(\alpha_{gj} - \frac{1}{2} \times \delta_g \times \phi_j, \sigma_{gj1}^2\right)$$

Condition 2

$$y_{gj2r} \sim N\left(\alpha_{gj} + \frac{1}{2} \times \delta_g \times \phi_j, \sigma_{gj2}^2\right)$$

- $g = 1, \dots, G$ is the gene index
- $j = 1, \dots, J$ is the pre-processing method index
- $r = 1, \dots, R$ is the replicate index

The parameter α_{gj} represents the global gene expression, whereas δ_g is the true (**latent**) differential expression that we would like to capture for gene g . The method-specific coefficient ϕ_j quantifies the **bias** of the method j .

Bayesian calibration model: second level modelling

(Variance structure)

The variance σ_{gjk}^2 can be decomposed into

- a gene (g) and condition (k) specific component
- a coefficient specific to the pre-processing method (j) and to the condition (k)

$$\sigma_{gjk}^2 = \exp(\lambda_{jk}) \times \sigma_{gk}^2.$$

Alternatively, to allow more flexibility, additional dependency on the global expression of each gene can be added:

$$\exp(\lambda_{0jk} + \sum_v \lambda_{vjk} \times \bar{y}_g^v) \times \sigma_{gk}^2$$

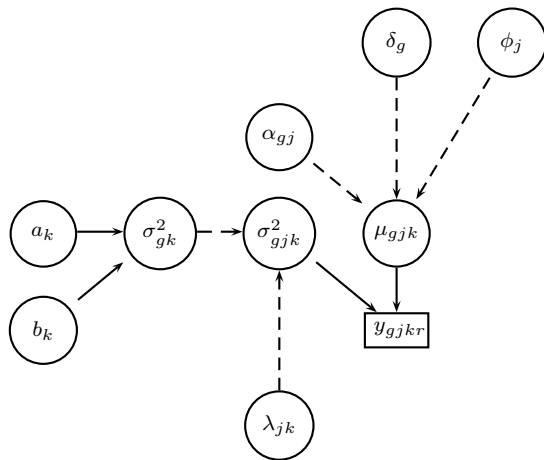
where

$$\bar{y}_g = \frac{1}{2JR} \sum_{j,k,r} y_{gjk r}$$

We found that the quadratic polynomial guarantees enough flexibility involving a limited number of parameters

$$\sigma_{gjk}^2 = \exp(\lambda_{0jk} + \lambda_{1jk} \times \bar{y}_g + \lambda_{2jk} \times \bar{y}_g^2) \times \sigma_{gk}^2.$$

Graph of the model



Model: Stochastic nodes (solid), Logical nodes (dashed)

Features of the model

- Hierarchical model on the gene specific variances

$$\sigma_{gk}^2 \sim Ga^{-1}(a_k, b_k)$$

with minimally informative priors on the hyperparameters

- Independent minimally informative normal prior for δ_g , α_{gj} , λ_{0jk} , λ_{1jk} , λ_{2jk}
- Independent minimally informative lognormal prior on the bias parameter ϕ_j

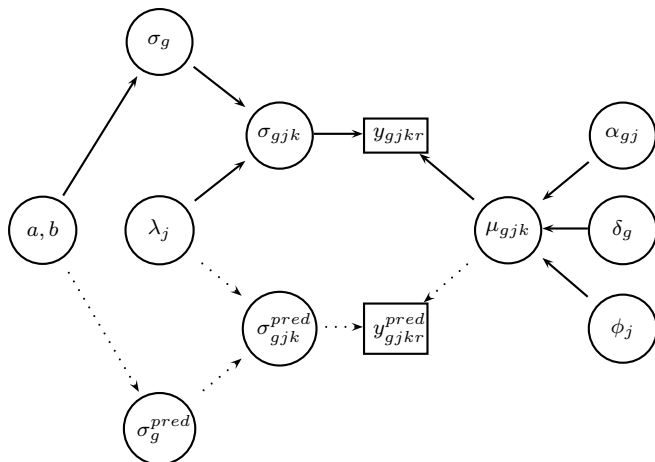
The model allows the borrowing of information across genes to estimate ϕ_j and λ_{jk} , and across methods to estimate δ_g and σ_{gk}^2 .

An MCMC algorithm coded in WinBUGS is used to simulate the posterior distribution of all unknown parameters.

Model checking

- Tension between the flexibility and the complexity of the parametrisation.
- Importance of including model checking in the analysis
 - Evaluate the role of specific parameters of the model
 - Mixed Posterior Predictive check
 - Compare the fit of different models
 - Deviance Information Criterion

Mixed Posterior Predictive check



Model (solid), Prediction (dotted)

Discrepancy function for y_{gjkr} and $y_{gjkr}^{pred} \Rightarrow$ p-value

Gelman A., Meng X., Stern H.: Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* 1996, 6: 733 - 807

Deviance Information Criterion

An extension of the Akaike Information Criterion when dealing with Bayesian hierarchical models, the Deviance Information Criterion (DIC) is defined as a function of the deviance of the model and of the effective number of parameters included:

$$DIC = E_{\theta}[D(\theta)] + p_D$$

where $E_{\theta}[D(\theta)]$ is the posterior mean of the deviance of the model and p_D is the effective number of parameters. When comparing two or more models, the one characterised by the smallest DIC shows the best fit to the data in hand.

Spiegelhalter D., Best N., Carlin B., van der Linde A.: Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B* 2002, 64(4): 583 - 639

Strategy of model checking

First level of the model

The structure of the bias attributable to each method can be multiplicative or additive

Additive bias	Multiplicative bias
$\delta_g + \xi_j$	$\delta_g \times \phi_j$

We investigated the fit for each structure by means of the DIC.

Strategy of model checking

First level of the model

The structure of the bias attributable to each method can be multiplicative or additive

Additive bias	Multiplicative bias
$\delta_g + \xi_j$	$\delta_g \times \phi_j$

We investigated the fit for each structure by means of the DIC.

Second level of the model

Incorporating the right structure on the variance in the model is a key point. We considered two different structures for σ_{gjk}^2 :

- a multiplicative function of the gene specific variance
- a polynomial function of the gene specific variance and of the global gene expression

Multiplicative structure	Polynomial structure
$\exp(\lambda_{jk}) \times \sigma_{gk}^2$	$\exp(\lambda_{0jk} + \sum_v \lambda_{vjk} \times \bar{y}_g^v) \times \sigma_{gk}^2$

We investigated the fit of each structure using the Mixed Posterior Predictive check

Finding differentially expressed genes from the output of a Bayesian Hierarchical model

$$H_0^g : \delta_g = 0 \quad \text{vs} \quad H_1^g : \delta_g \neq 0$$

We calculated a standardised distribution of differential expression:

$$z_g = \frac{\delta_g}{\sqrt{w_g}}$$

- δ_g is the measure of differential expression obtained from the model
- $w_g = \frac{2}{RJ^2} \sum_{j=1}^J (\sigma_{gj1}^2 + \sigma_{gj2}^2)$ is the variance of δ_g which is a function of σ_{gjk}^2

The Tail Posterior Probability is defined as:

$$p(z_g; z_\alpha) = \Pr(|z_g| > z_\alpha \mid \mathbf{y}_g)$$

where α is the chosen quantile of the standard normal distribution (usually $\alpha = 0.05$ and consequently $z_\alpha = 1.96$)

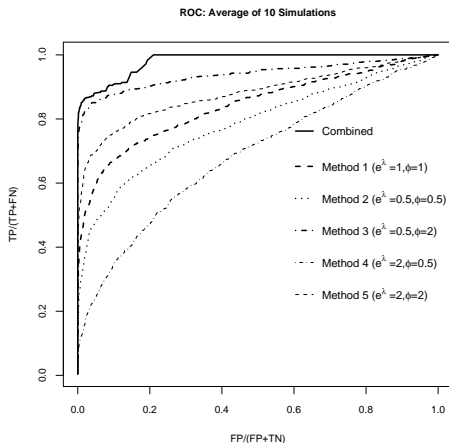
The histogram of $p(z_g; z_\alpha)$ is used to define a cut off as the differentially expressed genes are identified by a local peak on its right tail.

Bochkina N, Richardson S: Tail Posterior Probability for Inference in Pairwise and Multiclass Gene Expression Data. Biometrics 2007, 63(4): 1117 - 1125

Performance of the model: simulated data

Simulated log expression of 1000 genes (20% differentially expressed) and 5 pre-processing methods.

The simulation is repeated 10 times.



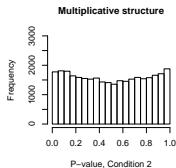
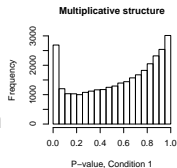
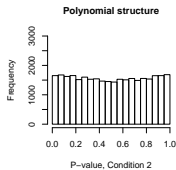
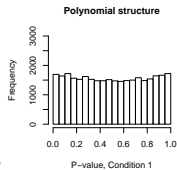
The **combined** method shows a greater sensitivity and specificity than each **single** pre-processing method.

Biological experiment

- Experiment to study the effect of high fat diet (HFD) versus normal fat diet (NFD) on mice adipose tissue
- 12488 genes, 2 conditions, 3 replicates for each condition
- Three pre-processing methods commonly used in the biological literature: MAS5, RMA, dChip
- Cut off of 0.98 on the tail posterior probability for the combined method leads to 292 differentially expressed genes

Biological experiment: Mixed Posterior Predictive check

- The Mixed Posterior Predictive check generates an empirical p-value for each gene
- Under the null hypothesis of the model being true, the distribution of the p-values should be approximately uniform
- A poor model fit is indicated by the presence of a notable pattern in the plot, suggesting a systematic difference between the observed values and the predicted ones.



The more complex model where the variance is a polynomial function of the global expression shows a better fit to the data in hand

Biological experiment: differential expression

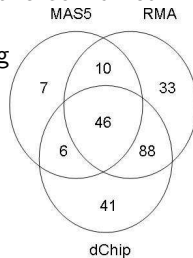
292 genes are found differentially expressed by the combined method.

Biological experiment: differential expression

292 genes are found differentially expressed by the combined method.

Considering the ranked list of 292 genes for each method and their intersection we have the following remarks:

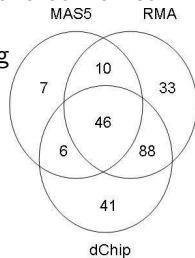
- All genes (46) in the intersection are found also by the combined method
- 185 genes are found by the combined method and one or two of the single methods.



Biological experiment: differential expression

292 genes are found differentially expressed by the combined method.

Considering the ranked list of 292 genes for each method and their intersection we have the following remarks:



- All genes (46) in the intersection are found also by the combined method
- 185 genes are found by the combined method and one or two of the single methods.

Additionally, 61 genes are found by the combined method

These are characterised:

- by a probability greater than 0.5 for at least two methods
- by a substantial different variability between the two conditions and a relatively small fold change

Considering each pre-processing method separately, these fold changes do not reach the top of the list, but the combined strategy increases their significance, by synthesising the evidence from the three pre-processing methods.

Annotation

First 292 genes with the largest posterior probability				
	Combined model	MAS5	RMA	dChip
Biological Processes	181	146	171	168
Molecular Functions	193	154	180	173
Cellular Components	179	145	173	169
KEGG pathways	223	182	215	205

Annotation

First 292 genes with the largest posterior probability				
	Combined model	MAS5	RMA	dChip
Biological Processes	181	146	171	168
Molecular Functions	193	154	180	173
Cellular Components	179	145	173	169
KEGG pathways	223	182	215	205

- For the combined method the most represented biological processes are the metabolic ones, functions associated with the response of the body to a change in the diet (*cellular metabolism, primary metabolism, macromolecular metabolism*). Using the Fisher Exact test 7 processes are enriched compared to MAS5, 5 compared to RMA and 3 compared to dChip.

Annotation

First 292 genes with the largest posterior probability				
	Combined model	MAS5	RMA	dChip
Biological Processes	181	146	171	168
Molecular Functions	193	154	180	173
Cellular Components	179	145	173	169
KEGG pathways	223	182	215	205

- For the combined method the most represented biological processes are the metabolic ones, functions associated with the response of the body to a change in the diet (*cellular metabolism, primary metabolism, macromolecular metabolism*). Using the Fisher Exact test 7 processes are enriched compared to MAS5, 5 compared to RMA and 3 compared to dChip.
- The most represented KEGG pathways are related to immune response and oxidation (*Antigen processing and presentation, MAPK signalling pathway, PPAR signaling pathway*), biological regulators of physiological functions as energy metabolism, insulin action, immunity and inflammation and known from the literature to be associated with high fat diet.

General Remarks

- Combining pre-processing methods is an example of how a measurement error approach can be used to identify the signal to noise ratio in gene expression studies

General Remarks

- Combining pre-processing methods is an example of how a measurement error approach can be used to identify the signal to noise ratio in gene expression studies
- Using a Hierarchical Bayesian allows the borrowing of information between genes and pre-processing methods and the inclusion of uncertainty at each level of the model.

General Remarks

- Combining pre-processing methods is an example of how a measurement error approach can be used to identify the signal to noise ratio in gene expression studies
- Using a Hierarchical Bayesian allows the borrowing of information between genes and pre-processing methods and the inclusion of uncertainty at each level of the model.
- Understanding the context and the variability sources is an important step in the modelling and model checking is essential to choose the best structure. We believe that the Mixed Posterior Predictive check is effective for model criticism as it provides a simple measure of discrepancy which is easy to visualise in an histogram.

General Remarks

- Combining pre-processing methods is an example of how a measurement error approach can be used to identify the signal to noise ratio in gene expression studies
- Using a Hierarchical Bayesian allows the borrowing of information between genes and pre-processing methods and the inclusion of uncertainty at each level of the model.
- Understanding the context and the variability sources is an important step in the modelling and model checking is essential to choose the best structure. We believe that the Mixed Posterior Predictive check is effective for model criticism as it provides a simple measure of discrepancy which is easy to visualise in an histogram.
- The method presented is general and applicable to a wide range of problems:
 - ↪ Multi-class studies
 - ↪ Main interest on the signal
 - ↪ Different types of chips (e.g. Agilent, Illumina)
 - ↪ Different pre-processing methods

For more information:

Blangiardo, M., Richardson, S.: A Bayesian calibration model for combining different pre-processing methods in Affymetrix chips.
BMC Bioinformatics 2008, 9:512
<http://www.biomedcentral.com/1471-2105/9/512/abstract>

THANKS!