A new approach for merging gene expression datasets

Marie-Christine Roubaud, Bruno Torrésani

A NEW APPROACH FOR MERGING GENE EXPRESSION DATASETS

Marie-Christine Roubaud and Bruno Torrésani

Université de Provence
Laboratoire d'Analyse, Topologie et Probabilités
CMI, 39 rue Joliot-Curie, 13453 Marseille Cedex 13, France

ABSTRACT

We propose a new approach for merging gene expression data originating from independent microarray experiments. The proposed approach is based upon a model assuming dataset-independent gene expression distribution, and dataset-dependent observation noise and nonlinear observation functions. The estimation algorithm combines smoothing spline estimation for the observation functions with an iterative method for gene expression estimation. The approach is illustrated by numerical results on simulation studies and real data originating from prostate cancer datasets.

Index Terms— Gene expression, Microarray data, Smoothing spline regression, Empirical Bayes estimation

1. INTRODUCTION

Microarray experiments provide indirect measurements (generally through radioactivity or fluorescence intensities) of the quantity of RNA produced by large sets of genes in controlled conditions. As such, they are expected to allow deeper understanding of gene regulation, as well as important prognostic tools in a number of pathologies.

Unfortunately, microarray expression measurements turn out to be highly sensitive to experimental conditions, and important reproducibility problems have been encountered. For example, it is very difficult to aggregate datasets originating from different experiments, even in situations where the biological objective and experimental setup are similar. For these reasons, it has been argued by several authors that better results are obtained by merging the results of several studies, rather than performing similar studies on aggregated datasets. The main shortcoming of such approaches is that they do not completely exploit the variability present in the aggregated dataset. For example in differential analysis, several approaches are based on combining adjusted p-values. Statistical tests to detect differentially expressed genes are performed on each dataset independently and the problem of small groups is not resolved. There is cooperation between the different experiments only at the final stage and relevant information is likely to be lost.

We consider here the dataset aggregation problem, and propose a pre-processing aiming at reducing the between-study variability, in the spirit of standard microarray normalization methods. The pre-processing is based on an explicit modeling of both the gene expression values and the transformations induced by different experiments. Such a modeling allows us to propose estimation methods for the former and the latter, which we illustrate on both simulated and real data. This work builds on prior work by the same authors, in which a similar, though simpler modeling was proposed for estimating so-called *rectification functions* (i.e. reciprocal functions of the observation functions).

This contribution is organized as follows. The model is presented in section 2 and an estimation algorithm is developed in 3. Numerical results are discussed in section 4.

2. MODEL DEFINITION AND ESTIMATION

2.1. The model

We assume we are given several datasets, corresponding to different studies $k = 1, \dots K$. Each dataset k consists in N_c^k arrays, hereafter termed *conditions*. After suitable pre-processing if necessary, we are led to a set of common genes $g = 1, \dots N_g$; for each experiment, we denote by $c = 1, \dots N_c^k$ the corresponding conditions.

The observations therefore take the form $\boldsymbol{y} = y_{g,c}^k$, denoting the measured expression level of gene g in condition c of experiment k. The main assumption of the model is that measured expression values from the various datasets are realizations of random variables (the "true" expression values"), which differ by
- experiment-dependent observation noise
- experiment-dependent observation function, assumed to be non-linear and smooth.

The observed values are then modeled as follows:

- **Observations:** $\boldsymbol{y} = \{y_{gc}^k\}$, of the form

$$y_{gc}^k = f_k(x_{gc}^k) + u_{gc}^k \,, \qquad u^k \sim \mathcal{N}(0, \tau_k^2) \,,$$

  where the observation noise variances τ_k^2 are unknown, and the underlying gene expressions $\boldsymbol{x}$ and observation functions $\boldsymbol{f}$ are described below.

- **"True" gene expressions:** $\boldsymbol{x} = \{x_{gc}^k\}$, of the form

$$x_{gc}^k = \mu_g + \delta_{gc}^k \,, \qquad \delta_g \sim \mathcal{N}(0, \sigma_g^2)$$

  where the gene average expressions μ_g and variances σ_g^2 are unknown.

- **Observation functions:** the observation functions are supposed to be smooth functions $\boldsymbol{f} = \{f_k,\ k = 1, \dots K\}$, modeled as spline functions f_k, with smoothness enforcing prior probability $\ln p(f_k) \sim -\lambda_k \|f_k''\|_2^2$, controlled by some parameter λ_k.

Given these assumptions, the log posterior probability reads

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{f}|\boldsymbol{y}) = \mathcal{L}^{(1)} + \mathcal{L}^{(2)} + \mathcal{L}^{(3)} \,, \qquad \text{with}$$

$$\left\{ \begin{array}{rcccl} \mathcal{L}^{(1)} & = & \displaystyle\sum_{k,g} \mathcal{L}_g^{(1);k} & = & \displaystyle -\sum_{k,g} \frac{1}{2\tau_k^2} \sum_c [y_{gc}^k - f_k(x_{gc}^k)]^2 \\ \mathcal{L}^{(2)} & = & \displaystyle\sum_{k,g} \mathcal{L}_g^{(2);k} & = & \displaystyle -\sum_{g} \frac{1}{2\sigma_g^2} \sum_{k,c} [x_{gc}^k - \mu_g]^2 \\ \mathcal{L}^{(3)} & = & \displaystyle\sum_{k} \mathcal{L}^{(3);k} & = & \displaystyle -\sum_{k} \lambda_k \int_{-\infty}^{\infty} |f_k''(x)|^2 \, dx \end{array} \right.$$

2.2. Observation functions estimation

Assume the "true expression values" are known, the problem of estimating the observation function reduces to the minimization with respect to $\boldsymbol{f} = \{f_k \in H^2(\mathbb{R}), k = 1, \dots K\}$ of the quantity $\Gamma[\boldsymbol{f}] = \sum_k \left[\left(\sum_g \mathcal{L}_g^{(1);k} \right) + \mathcal{L}^{(3);k} \right]$ and decouples as K optimisation problems: for $k = 1, \dots K$,

$$\min_{f_k} \left\{ \frac{1}{\tau_k^2} \sum_g \left[y_{gc}^k - f_k(x_{gc}^k) \right]^2 + \lambda_k \int |f_k''(x)|^2 \, dx \right\} \,.$$

The latter are actually smoothing spline estimation problems, for which efficient algorithms are available. Notice that once the spline has been estimated, its derivative is readily available. These estimations are performed on a set of genes with small variance across samples in each experiment, termed below *invariant gene set* .

2.3. Means and variances estimation.

The average gene expressions μ_g are re-estimated at each step of the algorithm as sample averages of the estimated gene expressions.

The estimation of variance components is a difficult task, as many gene variances σ_g^2 are to be estimated. An iterated MINQUE [4] (i.e. REML) approach restricted to the *invariant gene set* (see Remark 1 below) is used, that turns out to yield sensible estimates for the observation noise variances τ_k^2. Unfortunately, the corresponding estimates for gene variances σ_g^2 we resort to the MINQUE approach,

The gene variances σ_g^2 are estimated using sample estimates from the initialization (see Remark 1 below).

2.4. Adjustment: intrinsic gene expression values estimation.

Given the observation functions f_k, the genes are decoupled, and the estimation reduces to minimizing for each g

$$\Phi_g(\boldsymbol{x}) = \sum_{k,c} \left\{ \frac{1}{2\tau_k^2} \left[y_{gc}^k - f_k(x_{gc}^k) \right]^2 + \frac{1}{2\sigma_g^2} \left[x_{gc}^k - \mu_g \right]^2 \right\} .$$

Due to the non-linearity of the observation functions f_k, no closed-form expression exist for the solution, and we resort to an iterative algorithm. We assume that the mean μ_g and variances σ_g^2 and τ_k^2 are known, as well as the observation functions f_k. Suppose that we already have a first estimate, say $x_{gc}^k(t-1)$ of the gene expression values. A linearization of the observation functions f_k in the neighborhood yields the first order approximations $x_{gc}^k(t) = x_{gc}^k(t-1) + \epsilon_{gc}^k$, and

$$f_k(x_{gc}^k(t)) \approx f_k(x_{gc}^k(t-1)) + \epsilon_{gc}^k f_k'(x_{gc}^k(t-1)) ,$$

from which we deduce

$$\Phi_g \approx \sum_{k,c} \left\{ \frac{1}{2\tau_k^2} \left[\epsilon_{gc}^k f_k'(x_{gc}^k(t-1)) - (y_{gc}^k - f_k(x_{gc}^k(t-1))) \right]^2 + \frac{1}{2\sigma_g^2} \left[\epsilon_{gc}^k - (\mu_g - x_{gc}^k(t-1)) \right]^2 \right\}$$

The update of x_{gc}^k can therefore be obtained by optimizing the above expression, which yields

$$a_{gc}^k \epsilon_{gc}^k = \frac{f_k'(x_{gc}^k(t-1))}{\tau_k^2} \left(y_{gc}^k - f_k(x_{gc}^k(t-1)) \right) + \frac{1}{\sigma_g^2} \left(\mu_g - x_{gc}^k(t-1) \right) ,$$

where we have set $a_{gc}^k = \frac{\sigma_g^2 f_k'(x_{gc}^k(t-1))^2 + \tau_k^2}{\sigma_g^2 \tau_k^2}$. Set now

$$\alpha_{gc}^k = 1/a_{gc}^k \sigma_g^2 = 1/(1 + f_k'(x_{gc}^k(t-1))^2 \sigma_g^2/\tau_k^2) ,$$

Then $0 \leq \alpha_{gc}^k \leq 1$, the limits being attained in the extreme cases (no noise, or constant f_k). This yields the update rule $x_{gc}^k(t) = x_{gc}^k(t-1) + \epsilon_{gc}^k$, i.e.

$$x_{gc}^k(t) = \alpha_{gc}^k \mu_g + (1 - \alpha_{gc}^k) \left[x_{gc}^k(t-1) + \frac{1}{f_k'(x_{gc}^k(t-1))} \left(y_{gc}^k - f_k(x_{gc}^k(t-1)) \right) \right] .$$

i.e. a weighted average of the mean μ_g and the contribution of observations. This is similar to empirical Bayes type update rules, the difference being that the weights depend upon the observations, due to the nonlinearity of observation functions.

3. ALGORITHM AND IMPLEMENTATION

The proposed approach can be summarized as follows:

- **Initialization:** Start from a first estimate for the "true" expression values $x_{gc}^k(0)$. Estimate the gene means μ_g and variances σ_g^2 as in 2.3, and the observation functions as in 2.2.

- **Iteration t:** estimates $x_{gc}^k(t-1)$ are available.
  - Re-estimate the gene expressions $x_{gc}^k(t)$ as in 2.4.
  - Update the mean gene expressions μ_g as in 2.3.

The output of the algorithm consists in estimates $\hat{\boldsymbol{x}} = \{\hat{x}_{gc}^k\}$ for the expression datasets $\boldsymbol{x} = \{x_{gc}^k\}$, to be exploited for further analyses, together with estimates for the means μ_g, variances σ_g^2 and τ_k^2, and the observation functions f_k.

Remark 1: For the initialization, since only $\boldsymbol{y}$ is available, we need a first estimate of the reciprocal of the observation functions. We use the estimate provided by the approach described in [5], which estimates *rectification functions* $\boldsymbol{\varphi} = \{\varphi_k\}$ (i.e. reciprocal functions of the observation functions, $\varphi_k = {f_k}^{-1}$). The estimation leads to another smoothing spline problem : optimize, with respect to the mean gene expressions μ_g and the rectification functions φ_k the quantity

$$\sum_{k=1}^{K}\sum_{g=1}^{N_g}\frac{1}{N_c^k}\sum_{c=1}^{N_c^k}\Big[\varphi_k(x_{gc}^k)-\mu_g\Big]^2+\sum_{k=1}^{K}\lambda_k\int|\varphi_k''(x)|^2\,dx\ .$$

The problem is solved by an iterative algorithm, in the same spirit as the approach described here.

The algorithm was implemented using the `R` statistical environment, from which we used the smoothing spline function `smooth.spline`. Bioinformatics related functions from the `Bioconductor` package [8] were also used, as well as the `multtest` package for multiple comparisons used below.

4. NUMERICAL RESULTS

We limit the above discussion to the case of $K = 2$ datasets to be merged. The approach was first validated using a simulated dataset, according to the model, using explicitly defined observation functions. Several choices for the variances and observation functions were tested. The corresponding numerical results (which we won't discuss further here due to the lack of space) allowed us to validate the approach.

4.1. Real data with artificial distortions

A test was performed using artificial observation functions, applied to real data. Namely, we chose a dataset with well understood biological outcome, splitted it into two well balanced subgroups (see below for details) and applied to the two subsets two different observation functions, before adding Gaussian observation noise. The goal was to study the impact of the deformation induced by the observation functions and the noise (which were chosen so as to hide the biological effects), and the ability of the algorithm to perform a sensible correction.

E. Coli expression data from the Covert *et al* study [1] were used. The data include expressions of 7295 genes under two different situations (20 aerobic and 22 anaerobic), and are particularly interesting in that they exhibit a clear variability between the two biological situations. Two subsets were created with both 10 aerobic and 11 anaerobic conditions, randomly chosen. Different non-linear transformations $f_1(x) = x^{0.7}$ and $f_2(x) = x^{1.4}$ were applied to the two so-created subsets (after standardization), and observation noise with variances $\tau_1^2 = \mathrm{var}(f_1(\boldsymbol{x}_1))/100$ and $\tau_2^2 = 9\,\mathrm{var}(f_2(\boldsymbol{x}_2))/100$ was added. A standard PCA (see Fig. 1) shows that in the new artificial dataset the biological variability is far dominated by the introduced distortions.

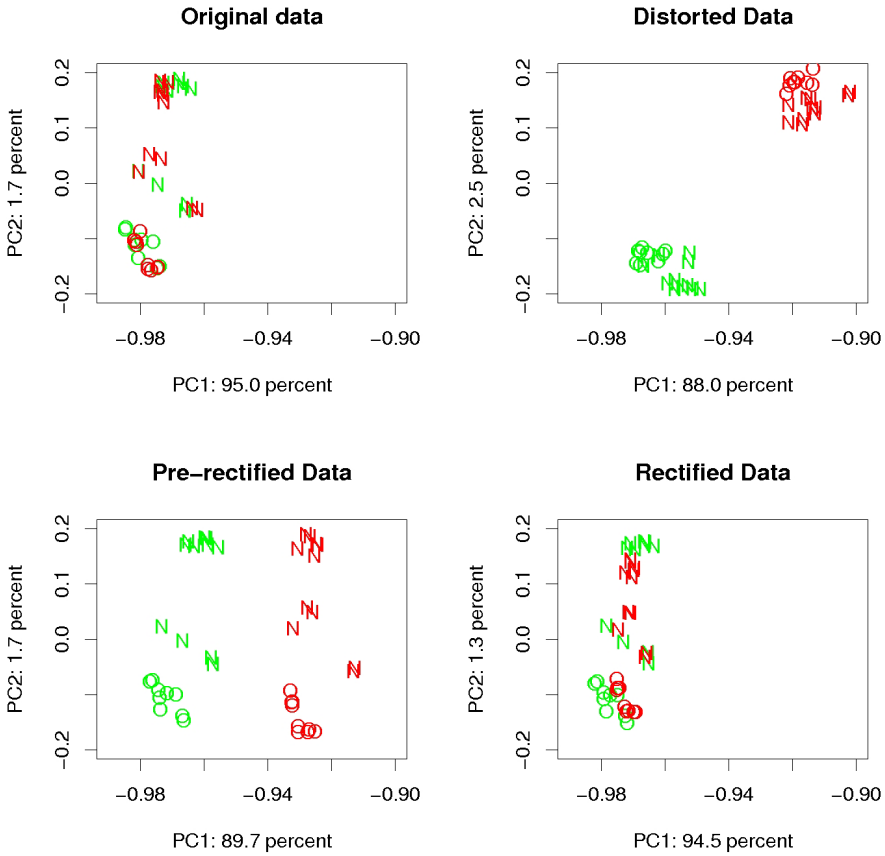

Fig. 1. Projections on the first principal plane. Top: original data (left) and distorted data (right). Bottom: rectified data: initialization (left) and processed data (right). O: aerobic; N: anaerobic. Green: dataset 1; red: dataset 2

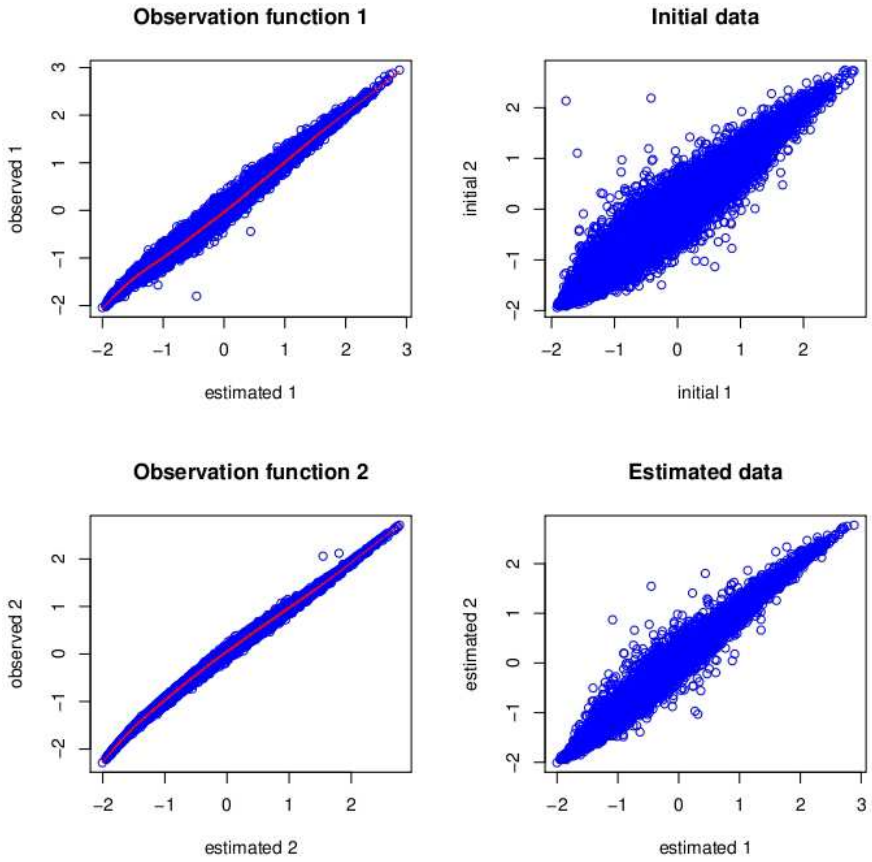

Fig. 2. Prostate datasets: observation functions (left), initial data (top right) and processed data (bottom right).

Running the proposed procedure on the distorted data, the output data $\hat{x}_{gc}^k$ turn out to reproduce fairly accurately the original data x_{gc}^k (before distortion). The PCA performed on distorted data y_{gc}^k, data after initialization and processed data $\hat{x}_{gc}^k$ shows that the processing has permitted to recover the biological features as the first source of variability (see Fig. 1).

4.2. Real data

The algorithm was tested on two datasets of prostate cancer expression data, namely Singh et al [6] and Stuart et al [7]. After pre-processing (reduction of the Singh dataset to a subset of arrays whose correlations to the median array exceed 90%, and reduction to common genes), the two-datasets consist in respectively 61 (32 tumor and 29 normal) and 86 (37 tumor and 49 normal) conditions, with 12625 genes. The proposed algorithm was run on these two datasets. The result of the processing is shown in Fig. 2, where we display the estimated observation functions, together with the initial and processed data.

Differential analysis was performed on the real dataset, and the processed dataset. After filtering out the 30% least variable genes, we used t-test, with FDR correction for multiple testing ($\alpha = 5\%$, 2000 bootstrap samples). Differentially expressed genes were seeked for the real dataset and the processed dataset, as well as the individual subsets $\boldsymbol{y}_1$ (Singh) and $\boldsymbol{y}_2$ (Stuart), and the processed subsets.

The individual datasets yield poorly compatible results, as the number of common differential genes is quite small (see Table 1 for details). The processing barely improves the results in this respect, the two processed subsets

	Singh		Stuart		Inter.		Merged	
	+	-	+	-	+	-	+	-
Real data	17	80	57	123	12	11	108	311
Processed	20	47	87	123	14	13	134	327

Table 1. Differential analysis on Prostate datasets. Numbers of differential genes for the individual datasets (Singh and Stuart), numbers of common differential genes (Inter) and Numbers of differential genes for the merged datasets.

yielding 4 more differentially expressed genes than the real data. On the other hand, the number of differentially expressed genes found on the merged processed datasets is significantly higher than the number of differential genes in the merged real data. A closer analysis shows that dataset 1 (Singh) has experienced bigger corrections than dataset 2 (Stuart). This is not surprising, since the Singh data are far less correlated than the Stuart data, and the algorithm has to correct for it.

5. CONCLUSIONS

We have described in this paper a new approach for merging gene expression datasets originating from different studies. Our results show that the proposed approach is able to correct, to some extent, for study-dependent non-linear distortions and observation noise.

A key question in this procedure is actually the estimation of gene expression variances σ_g^2. This is known to be a difficult problem, given the generally low number of conditions, and the strategy used for this estimation turns out to strongly influence the final results. Several approaches have been proposed in the literature, and inclusion of these into our model will be the goal of further developments.

6. REFERENCES

[1] M.W. Covert et al, Integrating high-throughput and computational data elucidates bacterial networks, *Nature*, **429** (2004) p. 92.

[2] F. Hong, and R. Breitling, R. (2008) A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments, *Bioinformatics*, **24**:3 (2008), p. 374.

[3] R.A. Irizarry, Multiple-laboratory comparison of microarray platforms, *Nature Methods* **2**:6 (2005), p. 477.

[4] C.R. Rao, Estimation of variance and covariance components MINQUE theory". *J. Multivar. Anal.* **1** (1971), p. 257.

[5] M.-C. Roubaud and B. Torrésani, Approche variationnelle pour la fusion de jeux de données d'expression génique, proceedings of the 41-th french *Journées de Statistique*, SFdS, Bordeaux (2009). `http://hal.archives-ouvertes.fr/inria-00386695/`

[6] D. Singh et al, Gene expression correlates of clinical prostate cancer behavior *Cancer Cell* **1** (2002), p. 203.

[7] R. O. Stuart et al, In silico dissection of cell-type-associated patterns of gene expression in prostate cancer, *Proceedings of the National Academy of Sciences of the USA* **101** (2004), p. 615.

[8] Bioconductor, opensource software for bioinformatics, `http://www.bioconductor.org/`