Using FARMS for summarization
Using I/NI-calls for gene filtering

Djork-Arné Clevert

May 13, 2010

Contents

1 Introduction

The *farms* package provides a new summarization algorithm called FARMS - Factor Analysis for Robust Microarray Summarization and a novel unsupervised feature selection criterion called I/NI-calls.

2 FARMS

The summarization method is based on a factor analysis model for which a Bayesian Maximum a Posteriori method optimizes the model parameters under the assumption of Gaussian measurement noise Hochreiter et al. (2006). Thereafter, the RNA concentration is estimated from the model. *farms* does not use background correction and uses either quantile normalization Bolstad et al. (2003) or cyclic loess Yang et al. (2002); Dudoit et al. (2002). Neverthess any other *affy* preprocessing method can be applied as well. *farms* uses quantile normalization as default normalization procedure because it is computational efficient. It does not apply PM corrections and uses PMs only. We set the hyperparameters of the prior distribution by default to **weigth** $=$ **0.5**, **mu** $=$ **0**. We further set the default values for the maximal EM-iterations to **cyc** $=$ **100** and the termination criteriea to **tol** $=$ **0.00001**, which express that the iteration will stop if the change of $\mathbf{var}(\mathbf{z}|\mathbf{x})$ after the update step is smaller than that tolerance value. If probes of a probe set are governed by a common latent variable, then we associate this variable with the mRNA concentration and its variation with the mRNA variation, i.e. with the signal. Intuitively speaking, if probes of a probe set change synchronously across the arrays then this effect is very unlikely produced by noise and one should assume they are driven by a signal. But if no common variable exists, the covariance structure can solely explain by the noise variance and implies that factor loadings are zero. In this case the expression values will be constant. Some post-processing methods e.g. t-tests face problems with constant results, therefore we introduced a boolean parameter called **robust**, which prevent results with zero variance. This parameter is by default set to TRUE. **Nevertheless, we highly recommend to filter out nonrelevant probe sets by applying**

I/NI-calls, as described in section 3. For the sake of convenience *farms* package provides three wrapper function for *affy-* `expresso`:

- `q.farms` is a wrapper function to `expresso` and uses no background correction and quantile normalization as default normalization procedure.
- `l.farms` performs like `q.farms`, but uses loess normalization as default normalization procedure.
- The function `exp.farms` is a transparent wrapper to `expresso` and permits further preprocessing options.

Note: If you use this package please cite Hochreiter et al. (2006) and Talloen et al. (2007). This package is only free for non-commercial users. Non-academic users **MUST** have a valid license.

2.1 Getting Started

As usual, it is necessary to load the package.

```
> library(farms)

Changes in FARMS:
Version 1.3.0: Added I/NI-calls for filtering
Version 1.3.1: Adjusted Hyperparameters for alternative CDFs,
                           probes set standardized, weighted mean
Version 1.4.0: Works now with R >= 2.8 and Bioconductor 2.3,
                           Changed termination criterion, initialization values,
                           factors and loadings scaled, added argument robust
For all changes previous to 1.3.0, see the farms vignette.
For updates please check http://www.bioinf.jku.at/software/farms/farms.html

> library(affydata)
```

In the following, we use the `affybatch.example` data set as it is provided by the *affy* package to illustrate how to compute expression measures with *farms*.

```
> data(Dilution)
> eset <- q.farms(Dilution)

background correction: none
normalization: quantiles
PM/MM correction : pmonly
expression values: farms
background correcting...done.
normalizing...
The downloaded packages are in
        /var/folders/aw/awsycpQ+En8YZO809RqAS++++TI/-Tmp-//RtmpUPpwtg/downloaded_packages
done.
12625 ids to be processed
|                    |
|####################|
```

This will store expression values, in the object `eset`, as an object of class `exprSet` (see the *Biobase* package).

```
> data(Dilution)
> eset <- exp.farms(Dilution, bgcorrect.method = "rma", pmcorrect.method = "pmonly",
+     normalize.method = "constant")

background correction: rma
normalization: constant
PM/MM correction : pmonly
expression values: farms
background correcting...done.
normalizing...done.
12625 ids to be processed
|                    |
|####################|
```

The available preprocessing options can be queried by using `normalize.AffyBatch.methods`, `pmcorrect.methods` or `bgcorrect.methods`.

3 I/NI calls

In this section, we show how to apply the I/NI-calls to a data set. Informative/ non-informative (I/NI) calls is an objective feature filtering technique for Affymetrix GeneChips. It uses the multiple probes measuring the same target mRNA as repeated measures to quantify the signal-to-noise ratio of that specific probe set. By incorporating probe level information to assess the noisy nature of probe sets, I/NI calls provide a highly powerful and objective tool for gene filtering. I/NI calls consequently offers a key solution to the main problem in the analysis of high-dimensional microarray data, being multiple testing and overfitting. I/NI calls can be used in combination with summarization techniques like FARMS, but also with any other summarization technique like MAS5 or (GC)RMA.

The following example shows how this summarization method can be used as a filtering tool, based on informative / non-informative calls.

```
> data(Dilution)
> eset <- q.farms(Dilution)

background correction: none
normalization: quantiles
PM/MM correction : pmonly
expression values: farms
background correcting...done.
normalizing...done.
12625 ids to be processed
|                    |
|####################|

> INIs <- INIcalls(eset)
> summary(INIs)

Summary
Informative probe sets      : 8.42%
Non-Informative probe sets  : 91.58%
```

```
> I_data <- getI_Eset(INIs)
> I_data

ExpressionSet (storageMode: lockedEnvironment)
assayData: 1063 features, 4 samples
  element names: exprs, se.exprs
protocolData: none
phenoData
  sampleNames: 20A, 20B, 10A, 10B
  varLabels and varMetadata description:
    liver: amount of liver RNA hybridized to array in micrograms
    sn19: amount of central nervous system RNA hybridized to array in micrograms

    scanner: ID number of scanner used
featureData: none
experimentData: use 'experimentData(object)'
Annotation: hgu95av2

> plotINIs(INIs)
```

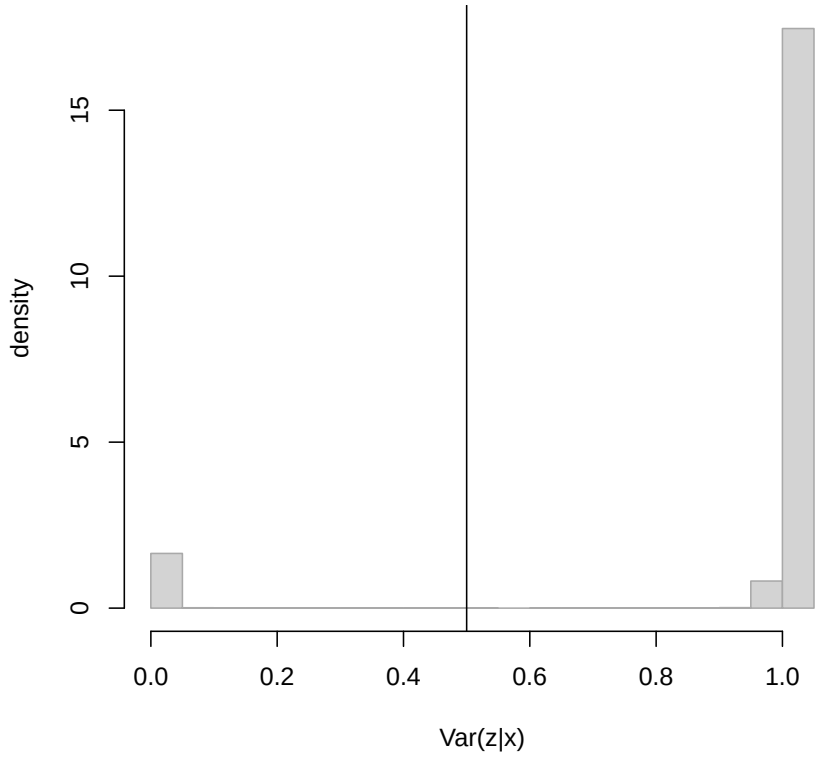

Figure 1: Histogram of $\mathbf{var}(\mathbf{z}|\mathbf{x})$ for the dilution data set, that is provided in affy.

Enjoy!

References

B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed. Statistical methods for identifying genes with differential expression in replicate cDNA microarray experiments. *Stat. Sin*, 12(1):111–139, 2002.

Sepp Hochreiter, Djork-Arne Clevert, and Klaus Obermayer. A new summarization method for affymetrix probe level data. *Bioinformatics*, page btl033, 2006. doi: 10.1093/bioinformatics/btl033. URL `http://bioinformatics.oxfordjournals.org/cgi/content/abstract/btl033v1`.

Willem Talloen, Djork-Arne Clevert, Sepp Hochreiter, Dhammika Amaratunga, Luc Bijnens, Stefan Kass, and Hinrich W.H. Goehlmann. I/ni-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data. *Bioinformatics*, page btm478, 2007. doi: doi:10.1093/bioinformatics/btm478. URL `http://bioinformatics.oxfordjournals.org/cgi/content/short/btm478v1`.

Y. H. Yang, S. Dudoit, P. Luu, D. Lin, V. Peng, J. Ngai, and T. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30(4):e15, 2002.