SSPA - Pilot data based sensitivity analysis for high-dimensional data

Maarten van Iterson (Bioinformatics PhD Student)

Center for Human and Clinical Genetics
Leiden University Medical Center

November 16, 2010

What is the appropriate sample size for an experiment?

- sample variability
- effect size
- proportion of features of interest

Basically two ways:

- simulation study
- pilot data

- adaptive FDR frame work (Benjamini and Hochberg (1995) and Storey (2003))

- adaptive FDR frame work (Benjamini and Hochberg (1995) and Storey (2003))
- estimates of π_0 and distribution of effect sizes

- adaptive FDR frame work (Benjamini and Hochberg (1995) and Storey (2003))
- estimates of π_0 and distribution of effect sizes
  - Several good estimators of π_0 are known (Langaas *et al.* (2005), Storey (2002) and Ferreira and Zwinderman (2006))

- adaptive FDR frame work (Benjamini and Hochberg (1995) and Storey (2003))
- estimates of π_0 and distribution of effect sizes
  - Several good estimators of π_0 are known (Langaas *et al.* (2005), Storey (2002) and Ferreira and Zwinderman (2006))
  - density of effect sizes is estimated by a deconvolution estimator (Delaigle and Gijbels (2007))

- adaptive FDR frame work (Benjamini and Hochberg (1995) and Storey (2003))
- estimates of π_0 and distribution of effect sizes
  - Several good estimators of π_0 are known (Langaas *et al.* (2005), Storey (2002) and Ferreira and Zwinderman (2006))
  - density of effect sizes is estimated by a deconvolution estimator (Delaigle and Gijbels (2007))
- power and sample size analysis

- adaptive FDR frame work (Benjamini and Hochberg (1995) and Storey (2003))
- estimates of π_0 and distribution of effect sizes
  - Several good estimators of π_0 are known (Langaas *et al.* (2005), Storey (2002) and Ferreira and Zwinderman (2006))
  - density of effect sizes is estimated by a deconvolution estimator (Delaigle and Gijbels (2007))
- power and sample size analysis
- all estimators are based on p-values or test statistics, the assumption is either normal or Student's t

- adaptive FDR frame work (Benjamini and Hochberg (1995) and Storey (2003))
- estimates of π_0 and distribution of effect sizes
  - Several good estimators of π_0 are known (Langaas *et al.* (2005), Storey (2002) and Ferreira and Zwinderman (2006))
  - density of effect sizes is estimated by a deconvolution estimator (Delaigle and Gijbels (2007))
- power and sample size analysis
- all estimators are based on p-values or test statistics, the assumption is either normal or Student's t
- currently under development moderated t (*limma*) and nonparametric approach

implementation of the deconvolution

mixture representation for the test statistics:

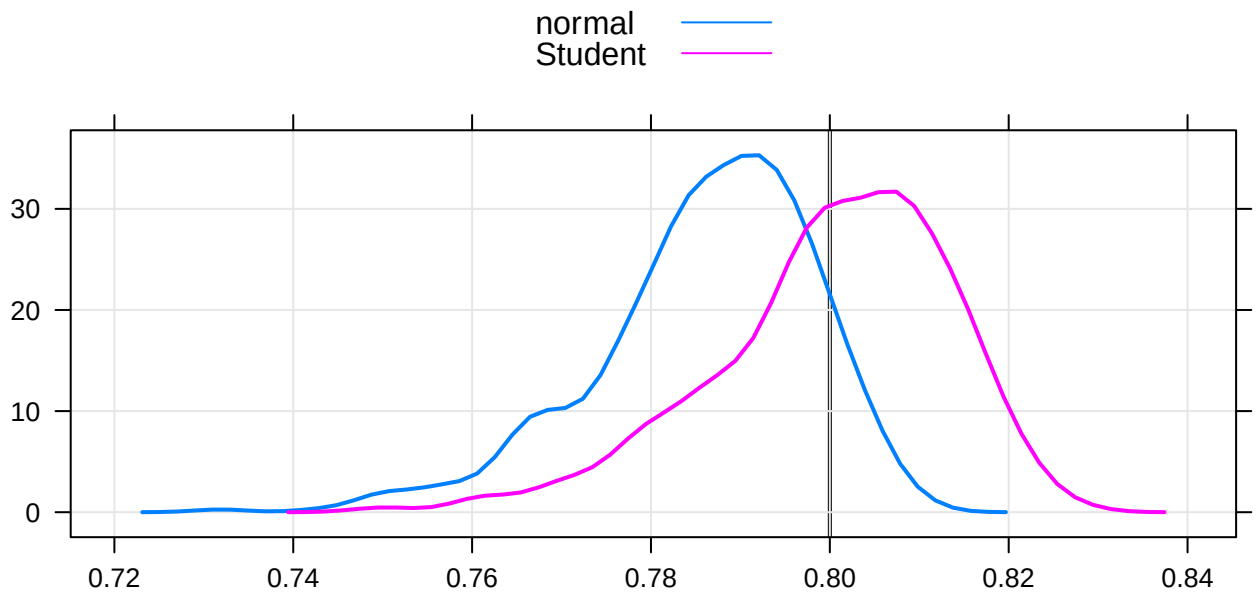
$$f(x) = \pi_0 f_0(x) + (1 - \pi_0) \int_{-\infty}^{+\infty} f_0(x - \theta\sqrt{N})\lambda(\theta)d\theta \qquad (1)$$

deconvolution estimator:

$$\hat{\lambda}(\theta) = \frac{\sqrt{N}}{2\pi} \int_{-\infty}^{+\infty} e^{-it\theta\sqrt{N}} \chi_K(at) \frac{\hat{\chi}_h(t)}{\chi_{f_0}}(t)dt, \qquad (2)$$

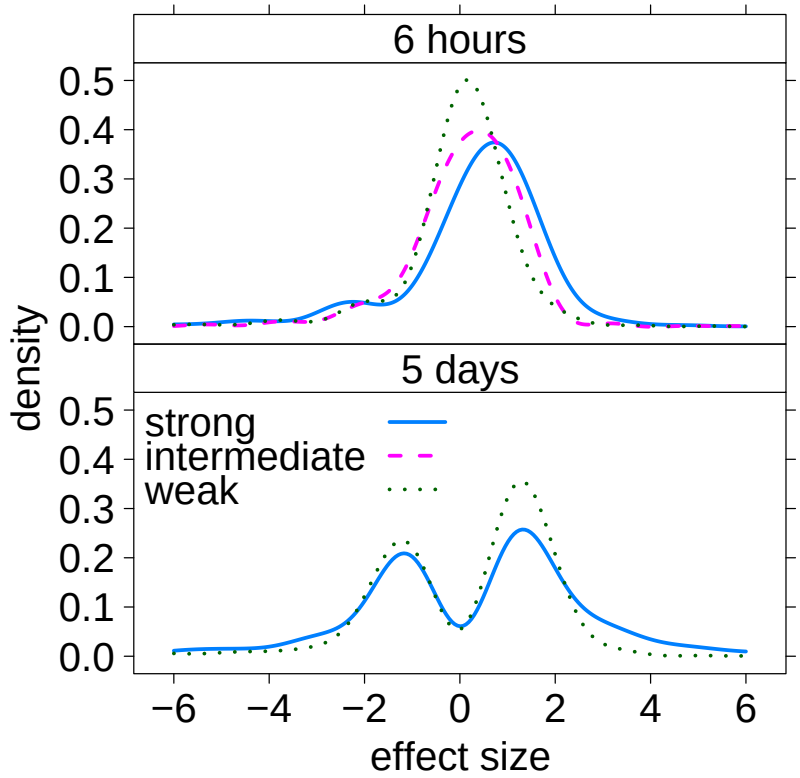
related to `density` using `fft` and `massdist` a C-function from the package *stats*.

Estimation of π_0



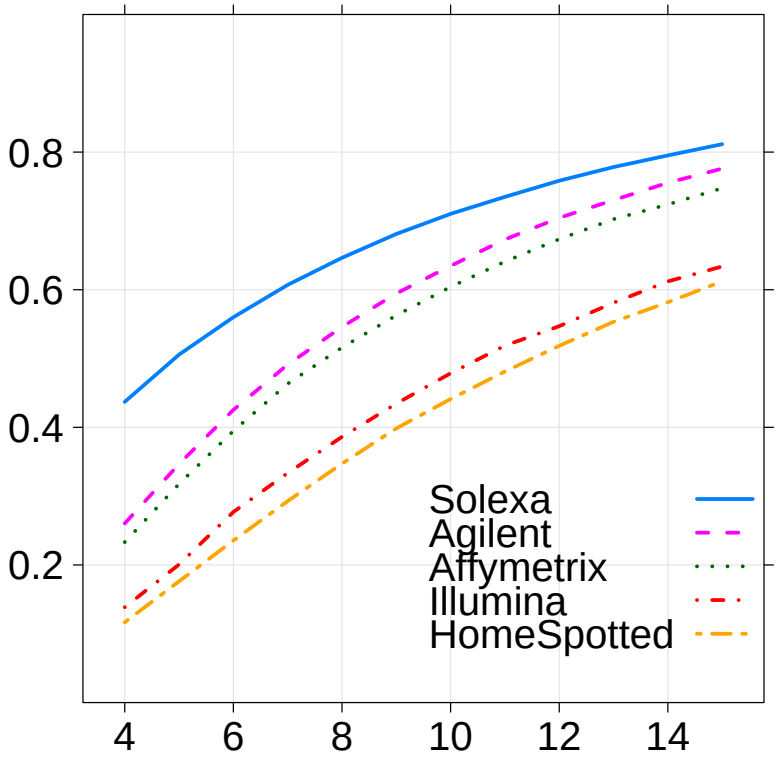
On simulated data using method by Langaas *et al.* (2005).

Estimation of density of effect sizes



Nutrigenomics experiment: using different compounds and exprosure times (van Iterson *et al.* (2009)).

Power curves



Different expression profiling platforms were compared (van Iterson *et al.* (2009)).

Conclusion:

- pilot data-based power and sample size analysis

Future plans:

- moderated *t* (*limma*) should be more suitable for small sample sizes
- nonparametric approach, main difficulty nonparametric null; bootstrap high-dimensional data with small sample sizes

M. van Iterson *et al.* Relative power and sample size analysis on gene expression profiling data. (2009), BMC Genomics, **10.**