Multivariate Statistics in Ecology and Quantitative Genetics

Analyzing gene expression data or other data with more parameters than observatios

Dirk Metzler & Noémie Becker

http://evol.bio.lmu.de/_statgen

July 2013

Contents

Contents

Affymetrix Microarrays

- Probe set for each gene
- Probe sets are spread over the chip
- For each probe (pm) a control probe (mm) where the

Contents

Robust Multi-Array Average (RMA)

Irizarry et al., 2003

- Assume that probe expression value consists of exponential signal + normally distributed random noise.
- Each array has its own mean background level of random noise.
- After background correction, probe expression values are normalized
- Then pm expression values are summarized to gene expression values

RMA

Model to summarize probe sets:

$$Y_{ijn} = \mu_{in} + \alpha_{jn} + \varepsilon_{ijn}$$

where n is the probe set, Y_{ijn} is are the log scaled probe expression values, α_{jn} is the probe affinity effect with $\sum_j \alpha_{jn} = 0$, and μ_{in} is the log expression measure for gene (probe set) n on array i.

Contents

Variance Stabilizing Normalization (VSN)

Huber et al., 2002

- Uses error model of Rocke and Durbin (2001)

$$Y_{ijn} = \alpha_i + \beta_{ij} \cdot e^{\eta_{ijn}} + \nu_{ijn}$$

where $\eta_{ijn} \sim \mathcal{N}(0, 1)$ and $\nu_{ijn} \sim \mathcal{N}(0, s_\nu^2)$

- This leads to the transformation

$$h_i(y_{ijn}) = \text{arsinh}(a_i + b_i y_{ijn})$$

where a_i and b_i have to be estimated from the data, and

$$\text{arsinh}(x) = \log\left(x + \sqrt{x^2 + 1}\right)$$

Contents

Stability of gene lists

For the selection of most promising genes from gene expression data different criteria may be applied, which may lead to different list.
One way to decide which criteria are useful is to assess their stability: Do they lead to similar gene lists if part of the data is changed?

Contents

Ridge Regression

If covariables are correlated, regression coefficients β_i can become very large and cancel each other. A way to avoid this:

$$\widehat{\beta} = \arg\min_{\beta} \left\{ \sum_i \left(y_i - \beta_0 - \sum_j x_{ij}\beta_j \right)^2 + \lambda \sum_j \beta_j^2 \right\}$$

or, in other words:

$$\widehat{\beta} = \arg \min_{\beta : \sum_i \beta_i^2 \leq s} \left\{ \sum_i \left(y_i - \beta_0 - \sum_j x_{ij}\beta_j \right)^2 \right\}$$

Caution: depends on scale! Normalize the data before applying this.

Solution for Ridge Regression:

$$\widehat{\beta} = \left(X^T X + \lambda I\right)^{-1} X^T y$$

(usual linear regression if $\lambda = 0$)

Alternative Motivation of Ridge Regression:
$\mathcal{N}(0, \tau^2)$-Prior on β_j

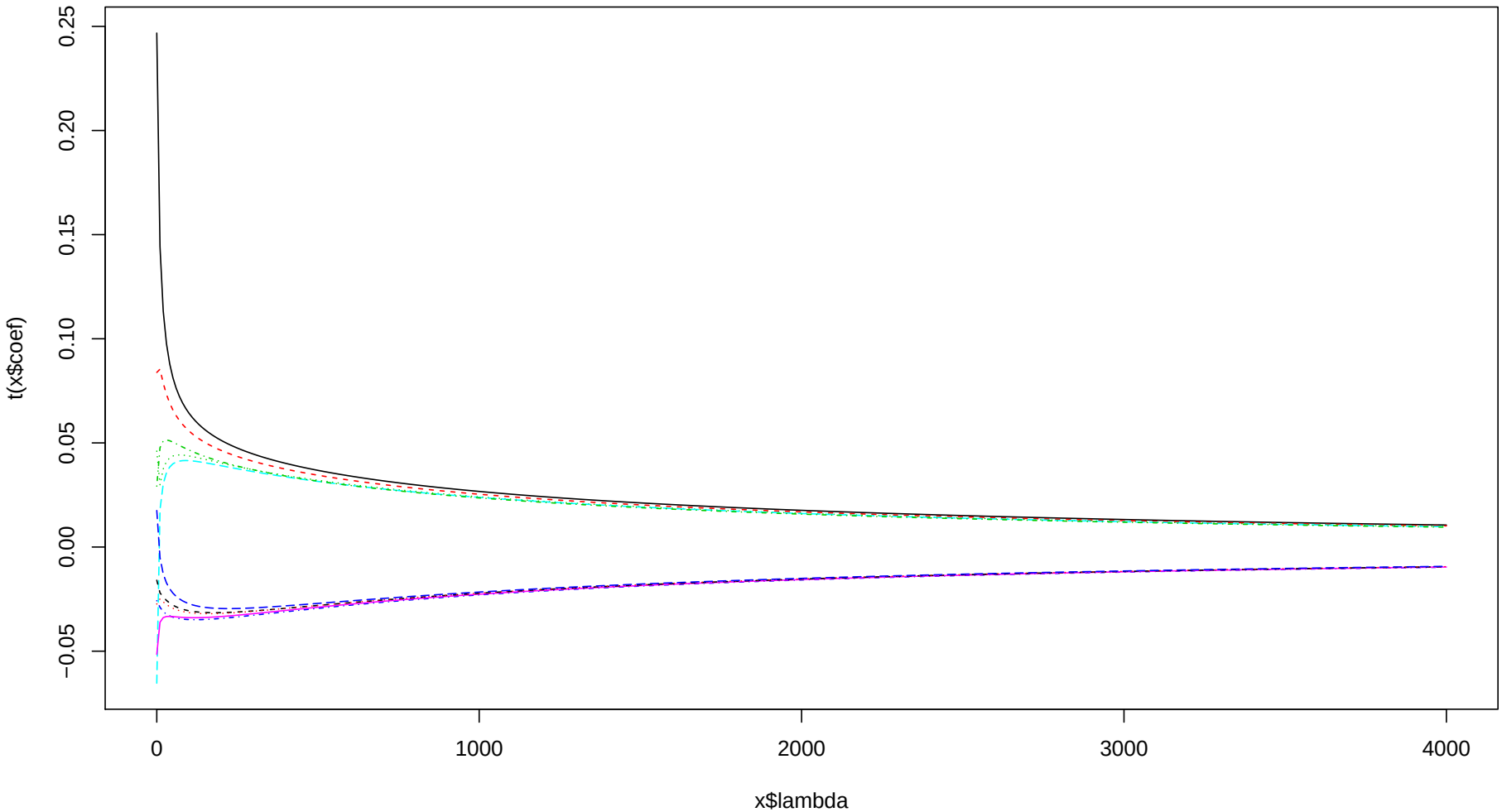
$$y_i \sim \mathcal{N}(\beta_0 + \sum_j x_{ij}\beta_j \ , \ \sigma^2)$$

Then the $\widehat{\beta}$ are means of the posterior distribution, where $\lambda = \sigma^2/\tau^2$.

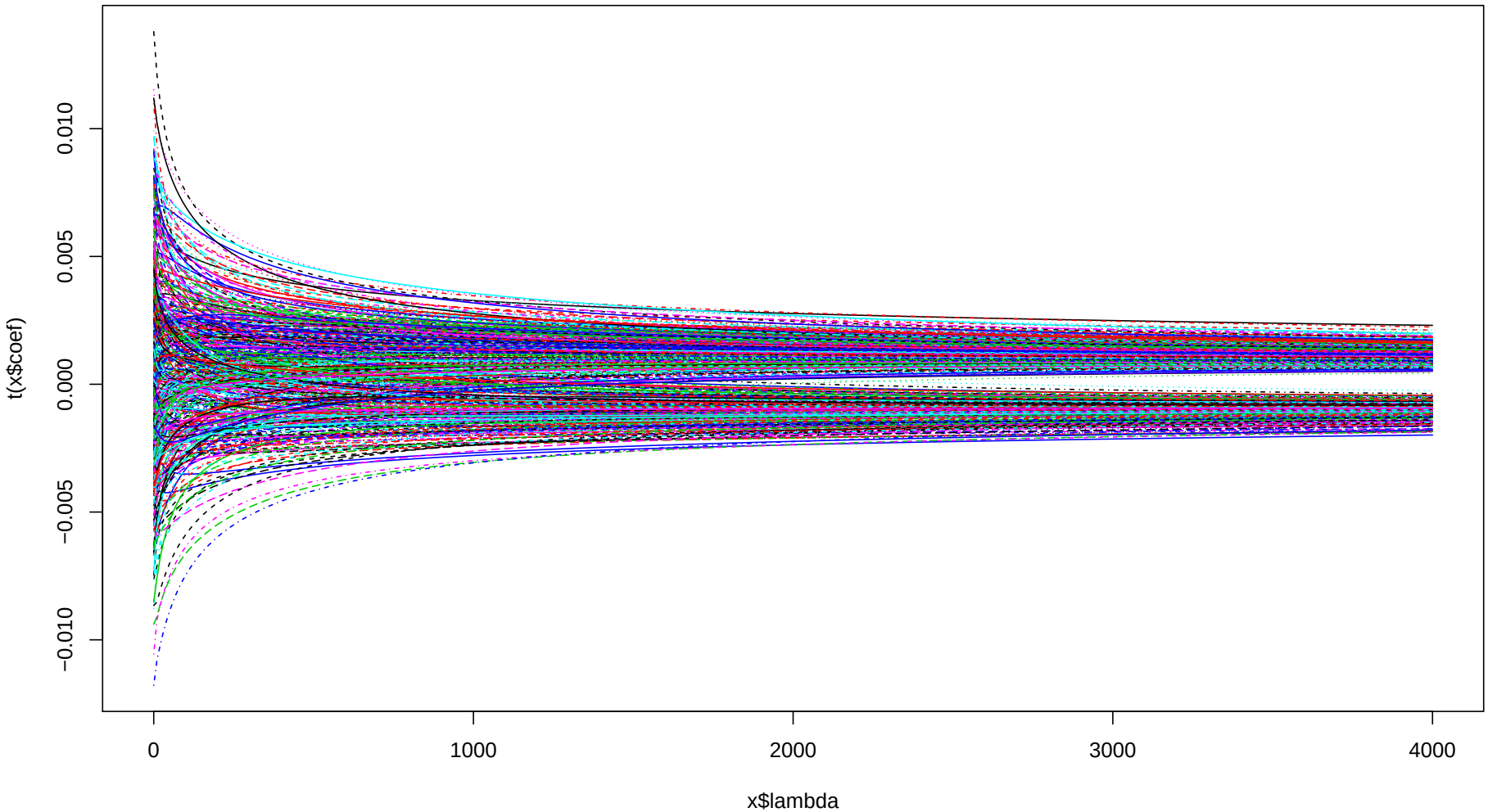
Geometric Interpretation of Ridge Regression:
All pricipal components are shrinked, the shorter they are, the more.
If d_j is the eigenvalue of the j-th principal component , shrink it by factor

$$\frac{d_j^2}{d_j^2 + \lambda}.$$

Ridge regression with 10 genes



Ridge regression with 500 genes

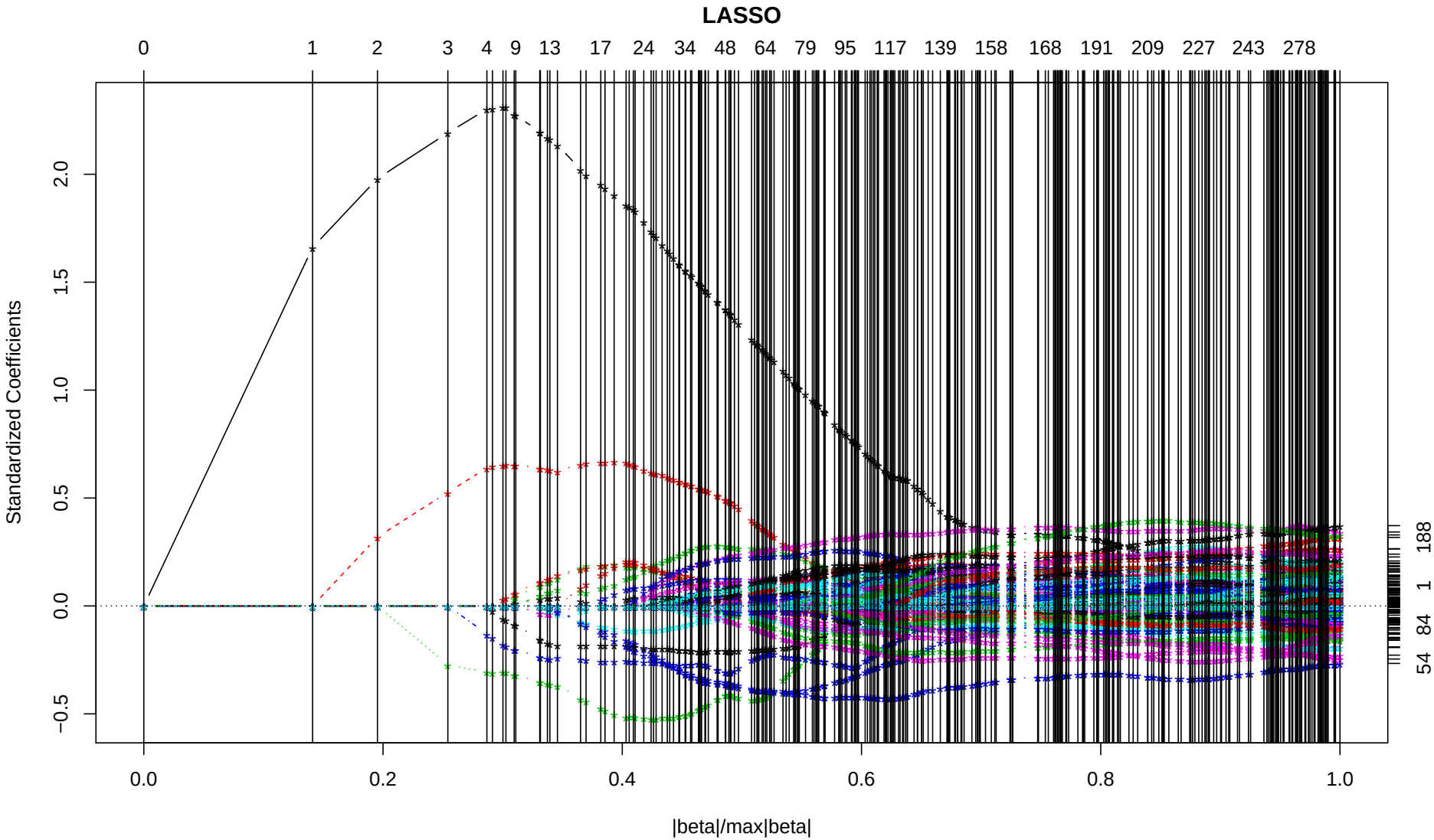


Similar method: **LASSO** (least absolute shrinkage and selector operator)

$$\widehat{\beta} = \arg \min_{\beta : \sum_i |\beta_i| \leq s} \left\{ \sum_i \left(y_i - \beta_0 - \sum_j x_{ij} \beta_j \right)^2 \right\}$$

the coefficients are set to 0.
"kind of continous subset selection"
(Hastie, Tibshirani, Friedman, 2001, *The Elements of Statistical Learning*)

Lasso with 500 genes



Contents

see R file