Testing against a high-dimensional alternative

Aldo Solari

Statistical Learning

These slides have been borrowed from Prof. Jelle Goeman and Prof. Hans van Houwelingen. Thanks!

Testing against a high-dimensional alternative

Jelle Goeman Hans van Houwelingen

Department of Medical Statistics
Leiden University Medical Center

Lancaster, 2006-10-31

Outline

Outline

Testing in a high-dimensional model

Testing in linear regression

- Response vector y ($n \times 1$) and design matrix X ($n \times p$)
- Model: $y \sim \mathcal{N}(X\beta, \sigma^2 I)$
- Regression coefficients β ($p \times 1$) and variance σ^2 unknown
- How to test $H_0 : \beta = \mathbf{0}$ against $H_A : \beta \neq \mathbf{0}$?

Testing in a high-dimensional model

Testing in linear regression

- Response vector y ($n \times 1$) and design matrix X ($n \times p$)
- Model: $y \sim \mathcal{N}(X\beta, \sigma^2 I)$
- Regression coefficients β ($p \times 1$) and variance σ^2 unknown
- How to test $H_0 : \beta = \mathbf{0}$ against $H_A : \beta \neq \mathbf{0}$?

- Well-known solution: the F-test

Testing in a high-dimensional model

Testing in linear regression

- Response vector y ($n \times 1$) and design matrix X ($n \times p$)
- Model: $y \sim \mathcal{N}(X\beta, \sigma^2 I)$
- Regression coefficients β ($p \times 1$) and variance σ^2 unknown
- How to test $H_0 : \beta = \mathbf{0}$ against $H_A : \beta \neq \mathbf{0}$?

- Well-known solution: the F-test
- This works well for p (much) smaller than n
- What if p is close to n?

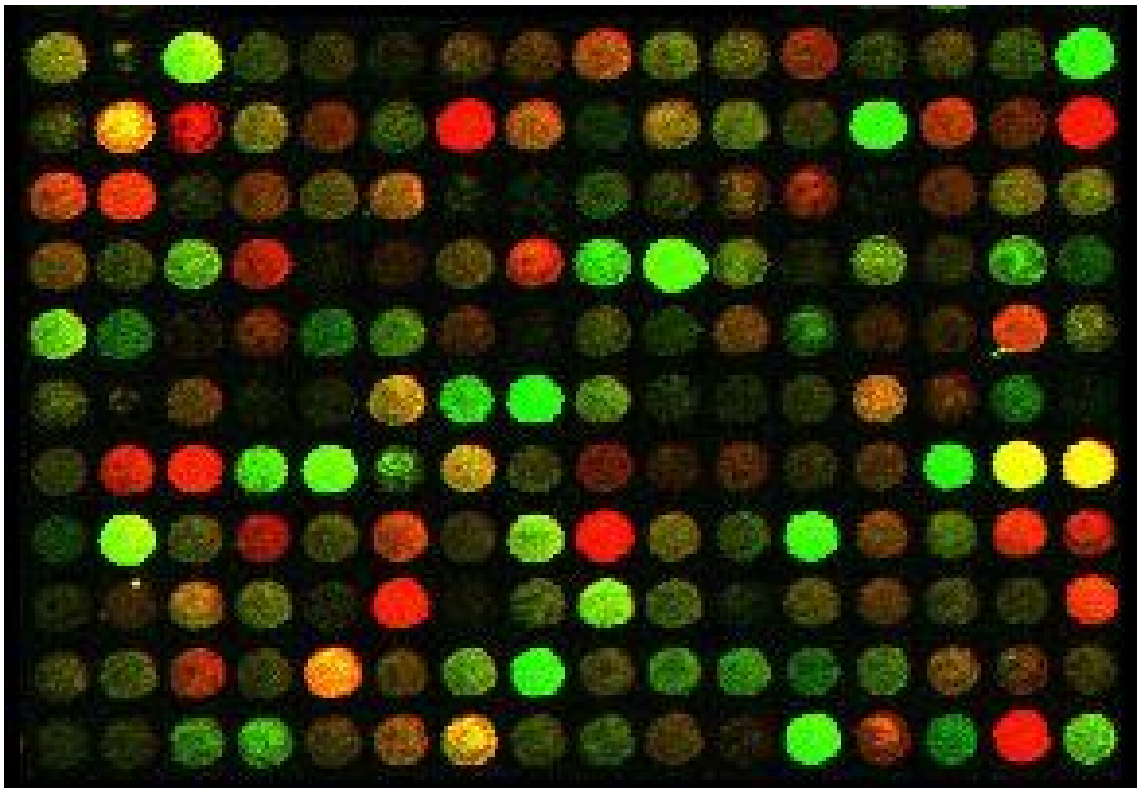
Testing in a high-dimensional model

Testing in linear regression

- Response vector y ($n \times 1$) and design matrix X ($n \times p$)
- Model: $y \sim \mathcal{N}(X\beta, \sigma^2 I)$
- Regression coefficients β ($p \times 1$) and variance σ^2 unknown
- How to test $H_0 : \beta = \mathbf{0}$ against $H_A : \beta \neq \mathbf{0}$?

- Well-known solution: the F-test
- This works well for p (much) smaller than n
- What if p is close to n?
- What if p is larger than n?

Motivation: microarray data



What is a microarray?

- Microarray measures 'gene expression'
- Central dogma of molecular biology

$$\text{DNA} \rightarrow \text{RNA} \rightarrow \text{protein}$$

- DNA is the same in every cell at every time
  But need for proteins is different
  $\rightarrow$ amount of RNA is different
- Gene expression $\approx$ concentration of RNA of a specific gene
- Microarray: simultaneous measurement of gene expression of around 20,000 genes in one tissue

A typical microarray experiment

- Make microarrays of tissue samples of n patients
- Sample size n usually between 10 and 300.
- Typical research question:
  Which genes are associated with the response?

Typical procedure

1. Do 20,000 univariate tests
2. Correct for multiple testing
3. Report a list of associated genes

Alternative: focus on pathways

- Pathway = set of genes with a similar function in the cell
  - Apoptosis: programmed cell death
  - Cell Cycle: the process of cell division
  - Angiogenesis: generation of blood vessels
- Our research question:
  which "pathways" are associated with y?
- More general: take any predefined set of genes
  - Genes with similar annotation in Gene Ontology / KEGG
  - Genes with similar chromosomal location
  - All genes on the microarray
- p = #genes: anything from 1 to 20,000 or more

Testing association of a pathway with a response

- Gene set = a set of covariates $(x_1, \ldots, x_p)$
  $\rightarrow$ design matrix X $(n \times p)$
- How to test for association $(x_1, \ldots, x_p)$ and y?
- $(x_1, \ldots, x_p)$ and y are associated $\iff$
  part of the variance of y can be predicted using $(x_1, \ldots, x_p)$
- In a linear model: $y \sim \mathcal{N}(X\beta, \sigma^2 I)$
- No association:

$$H_0 : \beta = \mathbf{0}$$

- Some association:

$$H_A : \beta \neq \mathbf{0}$$

Outline

The alternative hypothesis

Alternative: "at least one $\beta_i \neq 0$": very general if p large

Statistical problems

- Classical tests break down if $p > n$
- There are $\boldsymbol{\beta} \neq \mathbf{0}$ which have $\mathbf{r} = X\boldsymbol{\beta} = \mathbf{0}$
  $\rightarrow$ it is impossible to have power against all alternatives

Focus the power

- Focus the power using a distribution on β:
- Larger density of β
  $=$ alternative β is more interesting to detect

General Unprejudicedness Assumptions:

1. Alternative β as interesting as $-\beta$
2. Alternative β as interesting as permutation $\pi(\beta)$

Focus the power

- Focus the power using a distribution on β:
- Larger density of β
  $=$ alternative β is more interesting to detect

General Unprejudicedness Assumptions:

1. Alternative β as interesting as $-\beta$
2. Alternative β as interesting as permutation $\pi(\beta)$

- Consequence: take β with $\mathrm{E}(\beta) = \mathbf{0}$ and $\mathrm{E}(\beta\beta') = \tau^2 I$
- The distribution of β may have any shape
  $\rightarrow$ We don't want to assume any shape

A model with random regression coefficients

Inference via the marginal model: integrate β out

$\rightarrow$ p parameters in β become one parameter τ^2

New alternative

Test $H_0 : \tau^2 = 0$ against $H_A : \tau^2 > 0$

A model with random regression coefficients

Inference via the marginal model: integrate β out
$\rightarrow$ p parameters in β become one parameter τ^2

New alternative

Test $H_0 : \tau^2 = 0$ against $H_A : \tau^2 > 0$

Model interpretation 1: Empirical Bayes

Look at the distribution of β as a prior
to Inference on β goes through hyperparameter τ^2

A model with random regression coefficients

Inference via the marginal model: integrate β out
$\rightarrow$ p parameters in β become one parameter τ^2

New alternative

Test $H_0 : \tau^2 = 0$ against $H_A : \tau^2 > 0$

Model interpretation 2: Penalized likelihood

- The distribution of β gives a penalty to the likelihood
- Do maximum likelihood on penalized likelihood
- Normal β: Ridge regression
- Double exponential β: LASSO

A model with random regression coefficients

Inference via the marginal model: integrate β out
$\rightarrow$ p parameters in β become one parameter τ^2

New alternative

Test $H_0 : \tau^2 = 0$ against $H_A : \tau^2 > 0$

Model interpretation 3: Random effects model

- β random $\rightarrow$ linear predictor $\mathbf{r} = X\beta$ random
- Moments: $\mathrm{E}(\mathbf{r}) = \mathbf{0}$; $\mathrm{E}(\mathbf{r}\mathbf{r}') = XX'$
- Look at $r_1, \ldots, r_n$ as a random subject effect
- Similar gene expressions $\rightarrow$ correlated random effect
- Similar gene expressions $\rightarrow$ correlated response

The Neyman Pearson Lemma

- Why integrate out the β parameters?

The Neyman-Pearson lemma:

The likelihood ratio test of $H_0 : \beta = \mathbf{0}$ against $H_A : \beta = \beta_A$ has optimal power to detect H_A among all tests of at most the same size

The Neyman Pearson Lemma

- Why integrate out the β parameters?

The Neyman-Pearson lemma:

The likelihood ratio test of $H_0 : \beta = \mathbf{0}$ against $H_A : \beta = \beta_A$ has optimal power to detect H_A among all tests of at most the same size

Corollary to Neyman-Pearson:

The likelihood ratio test of $H_0 : \tau^2 = 0$ against $H_A : \tau^2 = \tau_A^2$ has optimal expected power to detect $H_A : \beta = \beta_A$ among all tests of at most the same size α

- Expectation is taken over the distribution of β_A under H_A

The Neyman Pearson Lemma

- Why integrate out the β parameters?

The Neyman-Pearson lemma:

The likelihood ratio test of $H_0 : \beta = \mathbf{0}$ against $H_A : \beta = \beta_A$ has optimal power to detect H_A among all tests of at most the same size

Corollary to Neyman-Pearson:

The likelihood ratio test of $H_0 : \tau^2 = 0$ against $H_A : \tau^2 = \tau_A^2$ has optimal expected power to detect $H_A : \beta = \beta_A$ among all tests of at most the same size α

- Expectation is taken over the distribution of β_A under H_A
- Drawback: distribution of β must be fully specified

Score testing

- Score test always one-sided: $H_0 : \theta = 0$ against $H_A : \theta > 0$
- Score test $=$ limit for $\theta_1 \downarrow 0$ of the LR test of H_0 against $H_A : \theta = \theta_1$

Score testing

- Score test always one-sided: $H_0 : \theta = 0$ against $H_A : \theta > 0$
- Score test $=$ limit for $\theta_1 \downarrow 0$ of the LR test of H_0 against $H_A : \theta = \theta_1$
- Test statistic:

$$\lim_{\theta_1 \downarrow 0} \frac{\log L(\theta_1) - \log L(0)}{\theta_1} = \frac{d}{d\theta} \log L(0)$$

Score testing

- Score test always one-sided: $H_0 : \theta = 0$ against $H_A : \theta > 0$
- Score test $=$ limit for $\theta_1 \downarrow 0$ of the LR test of H_0 against $H_A : \theta = \theta_1$
- Test statistic:

$$\lim_{\theta_1 \downarrow 0} \frac{\log L(\theta_1) - \log L(0)}{\theta_1} = \frac{d}{d\theta} \log L(0)$$

- Score test also "locally most powerful test"

Corollary 2 to Neyman-Pearson:

The score test of $H_0 : \theta = 0$ has optimal slope of the power function in $\theta = 0$ among all tests of at most the same size

Score testing

- Score test always one-sided: $H_0 : \theta = 0$ against $H_A : \theta > 0$
- Score test $=$ limit for $\theta_1 \downarrow 0$ of the LR test of H_0 against $H_A : \theta = \theta_1$
- Test statistic:

$$\lim_{\theta_1 \downarrow 0} \frac{\log L(\theta_1) - \log L(0)}{\theta_1} = \frac{d}{d\theta} \log L(0)$$

- Score test also "locally most powerful test"

Corollary 2 to Neyman-Pearson:

The score test of $H_0 : \theta = 0$ has optimal slope of the power function in $\theta = 0$ among all tests of at most the same size

- Practical advantage: no need to estimate θ

The locally most powerful test

- Back to our model with random parameters
- Instead of a Likelihood Ratio test we do a score test

The locally most powerful test

- Back to our model with random parameters
- Instead of a Likelihood Ratio test we do a score test

Corollary 3 to Neyman-Pearson:

The score test of $H_0 : \tau^2 = 0$ against $H_A : \tau^2 > 0$ has optimal expected slope of the power function in $\beta = \mathbf{0}$ among all tests of at most the same size

- Expectation is w.r.t. uniformly choosing a random direction in p-space
- Practical advantage: no need to estimate τ^2

The high dimensional score test statistic

- Test statistic is $S = S(0)$ with $S(\tau^2) = \frac{\mathrm{d}}{\mathrm{d}\tau^2} \log \mathrm{E}_{\beta|\tau^2} L(\beta; y)$
- Is equal to:

$$S = \frac{1}{2} s' s - \frac{1}{2} \mathrm{trace}(\mathbf{I})$$

- $s = \frac{\partial}{\partial \beta} \log L(0; y)$, the score of β

  $\mathbf{I} = -\frac{\partial^2}{\partial \beta \partial \beta'} \log L(0; y)$ is the observed Fisher inform. of β
- Nice properties:

  —*Easy: no evaluation of p-dimensional integrals*

  —*Test statistic only depends on the first two moments of* β

The high dimensional score test statistic

- Test statistic is $S = S(0)$ with $S(\tau^2) = \frac{\mathrm{d}}{\mathrm{d}\tau^2} \log \mathrm{E}_{\beta|\tau^2} L(\beta; y)$
- Is equal to:

$$S = \frac{1}{2} s' s - \frac{1}{2} \mathrm{trace}(\mathbf{I})$$

- $s = \frac{\partial}{\partial \beta} \log L(0; y)$, the score of β

  $\mathbf{I} = -\frac{\partial^2}{\partial \beta \partial \beta'} \log L(0; y)$ is the observed Fisher inform. of β
- Nice properties:

  —*Easy: no evaluation of p-dimensional integrals*

  —*Test statistic only depends on the first two moments of* β
- Difficult part: What is the distribution of S?

  —Solved for linear (exact), GLM, Cox PH (asymptotic)

  —Alternative: permutations

Outline

Back to the linear model

- What is the locally most powerful test in the linear model?
- Get rid of the nuisance parameter $\sigma^2 \rightarrow$ profile likelihood
- For β with $\mathrm{E}\beta = \mathbf{0}$ and $\mathrm{E}(\beta\beta') = \tau^2 I$:

$$S = \frac{y'XX'y}{y'y}$$

- For β with $\mathrm{E}\beta = \mathbf{0}$ and $\mathrm{E}(\beta\beta') = \tau^2\Sigma$:
  Score test statistic:

$$S_\Sigma = \frac{y'X\Sigma X'y}{y'y}$$

The F-test revisited

- The F-test is only defined when $p < n$
- Equivalent test statistic $F = \frac{y'X(X'X)^{-1}X'y}{y'y}$ (beta distributed)

The F-test revisited

- The F-test is only defined when $p < n$
- Equivalent test statistic $F = \frac{y'X(X'X)^{-1}X'y}{y'y}$ (beta distributed)

The F-test as a score test

F-test is equivalent to a score test with $\Sigma = (X'X)^{-1}$

- $S = F$ in an orthogonal design (i.e. $X'X \propto I$)
- F-test optimizes power over a prior for β with $\mathrm{E}(\beta\beta') = \tau^2(X'X)^{-1}$ (g-prior) for τ^2 small

Outline

Discussion: gene expression data

Analyzing micorarray data in terms of pathways

- Enables use of prior biological knowledge
- Alleviates the multiple testing problem
- Gives better reproducible results across platforms
- Global Test opens door to real inference:
  $\rightarrow$ testing hypotheses on biological mechanisms based on theory or past research

Discussion: high dimensional testing

A locally most powerful test in high dimensions

- Useable whatever the dimensionality
- Good power against interesting alternatives
- Useful applications in and outside microarray data analysis
- Applicable in survival analysis and generalized linear models

Read more?


Goeman, Van de Geer, De Kort, Van Houwelingen (2004).
A global test for groups of genes.
Bioinformatics, **20** (1) 93–99.


Goeman, Van de Geer and Van Houwelingen (2006).
Testing against a high-dimensional alternative.
JRSSB, **68** (3) 477–493.


Goeman and Oosting.
R package: globaltest
www.bioconductor.org