

Singular Value Decomposition for Feature Selection in cDNA Arrays

Dirk Klingbiel (AG Ruiz / University of Lübeck)
(klingb@molgen.mpg.de)

Supervisor: Lutz Dümbgen (University of Berne)

03.03.03



Contents

- Preliminaries (Data, Nomenclature, SVD)
- Methods (Visualization of V , A Test Statistic)
- Outlook

Preliminaries: The Data

- Data from cDNA nylon arrays
 - very noisy
 - highly variable
 - many genes, low replication
- $X \in \mathbb{R}^{p \times n}$: data matrix (of intensities)
 - p rows representing the clones
 - n columns representing the experiments
 - $n \ll p$: the well known *curse of dimensionality*
 - suitably normalized (e.g. using VSN)

Preliminaries: Nomenclature

- L groups
 - forming a partition of the set of experiments
 - numbered from $1, 2, \dots, L$
 - group i contains n_i experiments

$$\implies \sum_{i=1}^L n_i = n$$

- Let $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n)$ denote a matrix in $\mathbb{R}^{p \times n}$. Then $\mathbf{A}_k = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k) \in \mathbb{R}^{p \times k}$ stands for the matrix comprising the first k columns of matrix \mathbf{A}
- \mathbf{I}_d denotes the identity matrix of size $d \times d$

Preliminaries: SVD

Let A denote a matrix in $\mathbb{R}^{p \times n}$. Then it can be factorized as

$$A = \mathbf{U} \mathbf{D} \mathbf{V}^T = \sum_{i=1}^d \sigma_i \mathbf{u}_i \mathbf{v}_i^T, \quad (1)$$

with $d = \min(p, n)$ and

$$\begin{aligned} \mathbf{U} &= (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_d) \in \mathbb{R}^{p \times d}, & \mathbf{U}^T \mathbf{U} &= \mathbf{I}_d, \\ \mathbf{D} &= \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_d), & \sigma_1 &\geq \sigma_2 \geq \dots \geq \sigma_d \geq 0, \\ \mathbf{V} &= (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_d) \in \mathbb{R}^{n \times d}, & \mathbf{V}^T \mathbf{V} &= \mathbf{I}_d. \end{aligned}$$

The quantities σ_i are called the *singular values* of A , and the columns of \mathbf{U} and \mathbf{V} are called the *left and right singular vectors* of A .

Preliminaries: Properties of the SVD

- Total variance explained by factor σ_j is $\sigma_j^2 / \sum_{i=1}^d \sigma_i^2$
- $\|A\|_2 = \sigma_1$, $\|A\|_{\text{Frobenius}} = \left(\sum_{i=1}^d \sigma_i^2 \right)^{1/2}$
- Best rank ℓ approximation of A :

$$\min_{\text{rank}(B)=\ell} \|A - B\|_2 = \left\| A - \sum_{i=1}^{\ell} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \right\|_2 = \sigma_{\ell+1}$$
- First singular components explain variance caused by differences from zero \implies center matrix before SVD

Preliminaries: Applications of SVD

- Solutions to linear Equations
- Noisy signal filtering
- Compression
- Time series analysis
- Has, of course, been applied to gene expression data before

Preliminaries: Why use SVD?

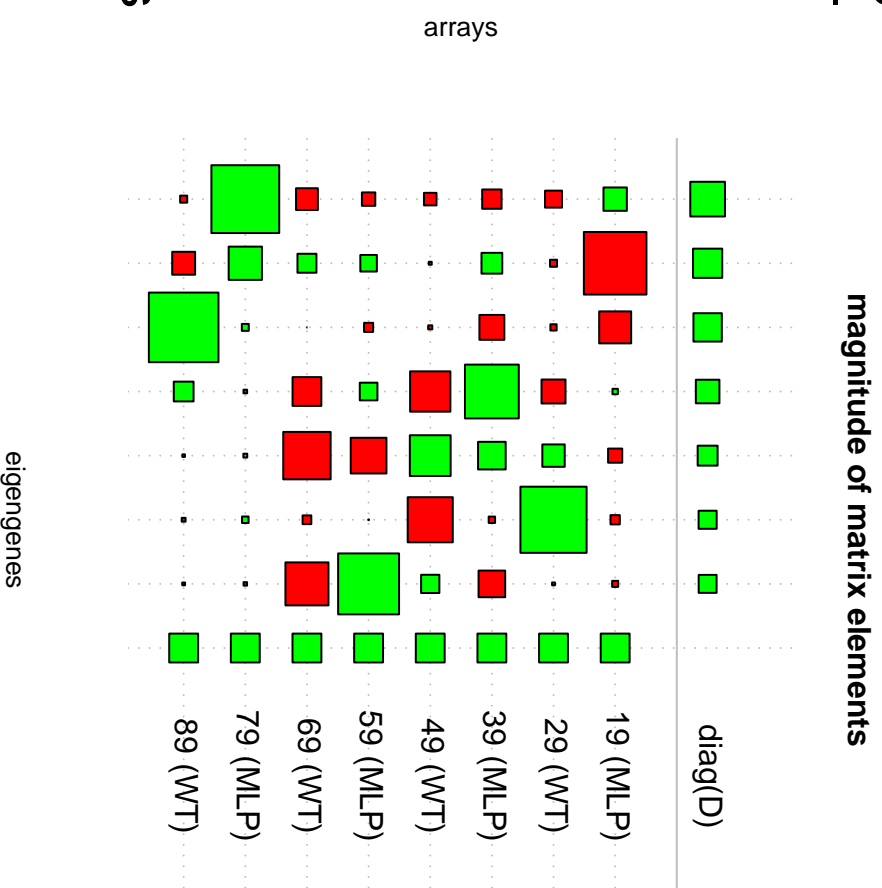
- Robust method
- Can be used for reduction of dimensionality
- Assumption: “genuine” biological variance bigger than (at least) some of the “experimental noise” (biological and technical variability)

Contents

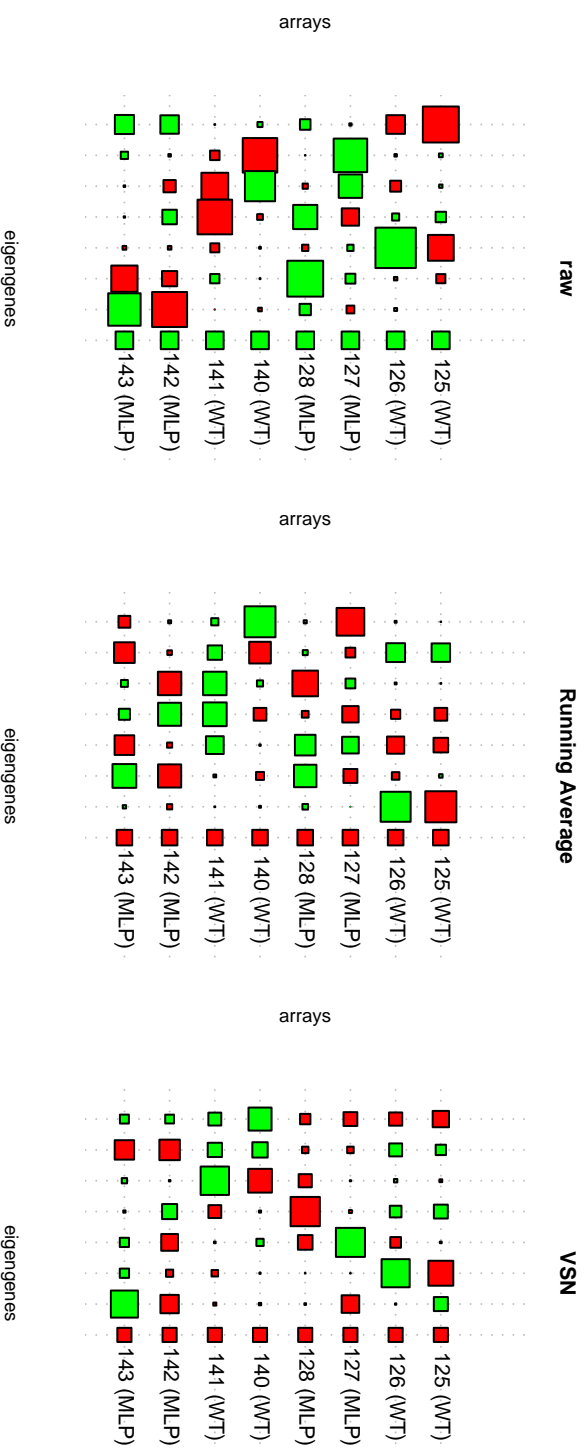
- Preliminaries (Data, Nomenclature, SVD)
- **Methods (Visualization of V , A Test Statistic)**
- Outlook

Methods: Visualization of V —Example 2

- Mouse experiment by Henning Witt (same setup, different individuals, different day)
- Things look different here
- In fact, things look a mess
- The right singular components reflect—uhm...



Methods: Visualization of V —Normalization



- The influence of different normalization methods applied to the data from the first experiment

Methods: Visualization of V —Conclusion

- SVD can—more or less clearly—reflect the biological variance
- Other experiments showed to be not as beautiful as experiment 1, but also not as messy as experiment 2
- If one singular component is not sufficient, take more!
- Normalization matters!
- SVD **CAN** reflect the biological variance!!!

Contents

- Preliminaries (Data, Nomenclature, SVD)
- **Methods (Visualization of V , A Test Statistic)**
- **Outlook**

Methods: The Test Statistic \tilde{u} —Algorithm (1)

(i) Transform the rows of the data matrix to have a mean of zero and perform SVD to the centered matrix:

- $\tilde{\mathbf{X}} := \mathbf{X} - \boldsymbol{\mu}_{\text{row}}(\mathbf{X})$

Otherwise we would obtain “boring” first singular components

- $\boldsymbol{\mu}_{\text{row}}(\mathbf{X})$ is the vector containing the means of each row from \mathbf{X}
- $\tilde{\mathbf{X}} = \mathbf{U}\mathbf{D}\mathbf{V}^T$

Methods: The Test Statistic \tilde{u} —Algorithm (2)

- (ii) For each of the L groups (or any set of groups) define a base vector of length n , using a centered indicator function:
- $\mathbf{b}_j := (1 \{\text{array } i \text{ from group } j\} - n(j))_{i=1,2,\dots,n}$ $j = 1, 2, \dots, L$
 - $n(j)$ is the relative amount of experiments from group j ,
i.e. $n(j) := n_j/n$.

Methods: The Test Statistic \tilde{u} —Algorithm (3)

(iii) For each base vector find a vector $\tilde{\mathbf{v}}_j \in \text{span}(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k)$ so that the angle between $\tilde{\mathbf{v}}_j$ and \mathbf{b}_j (or equivalently $\|\mathbf{V}_k \theta_j - \mathbf{b}_j\|^2$) becomes minimal:

$$\bullet \|\mathbf{V}_k \theta_j - \mathbf{b}_j\|^2 \stackrel{!}{=} \min, \quad \theta_j \in \mathbb{R}^k$$

- We obtain the orthogonal projection of \mathbf{b}_j onto $\text{span}(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k)$:

$$\theta_j = \mathbf{V}_k^T \mathbf{b}_j$$

- Hence, $\tilde{\mathbf{v}}_j = \mathbf{V}_k \theta_j = \mathbf{V}_k \mathbf{V}_k^T \mathbf{b}_j$

- Original idea: Find $\tilde{\mathbf{b}}_j \in \text{span}(\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_L)$

Methods: The Test Statistic \tilde{u} —Algorithm (4)

- (iv) Optionally check correlation between \tilde{v}_j and b_j
- If correlation too low, increase k and go back to step (iii):
 - Experiments 1 and 2 revisited:

k	1	2	3	4	5	6	7	8
correlation								
experiment 1	0.11	0.94	0.95	0.98	0.99	0.99	1.00	1.00
experiment 2	0.51	0.51	0.72	0.96	0.97	0.98	1.00	1.00

- So, taking ≥ 4 singular vectors experiment 2 doesn't look too bad

Methods: The Test Statistic \tilde{u} —Algorithm (5)

(v) Apply the coefficients from step (iii) to \mathbf{U}_k :

- $\tilde{\mathbf{u}}_j := \mathbf{U}_k \theta_j = \mathbf{U}_k \mathbf{V}_k^T \mathbf{b}_j$

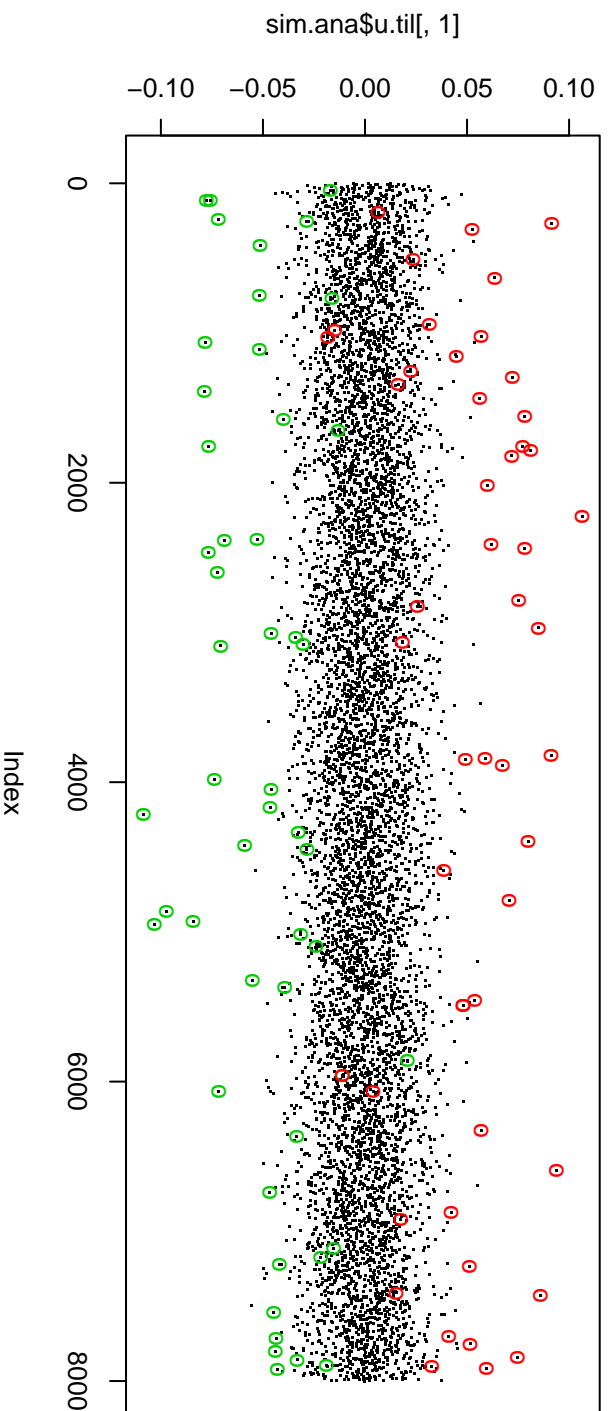
- That's it!

Methods: The Test Statistic \tilde{u} —Application (1)

- Now apply algorithm to data
- P-values are taken from Welch's t-test
- Adjustment of p-values by R's `p.adjust` function, method is `fdr`

Methods: The Test Statistic \tilde{u} —Application (2)

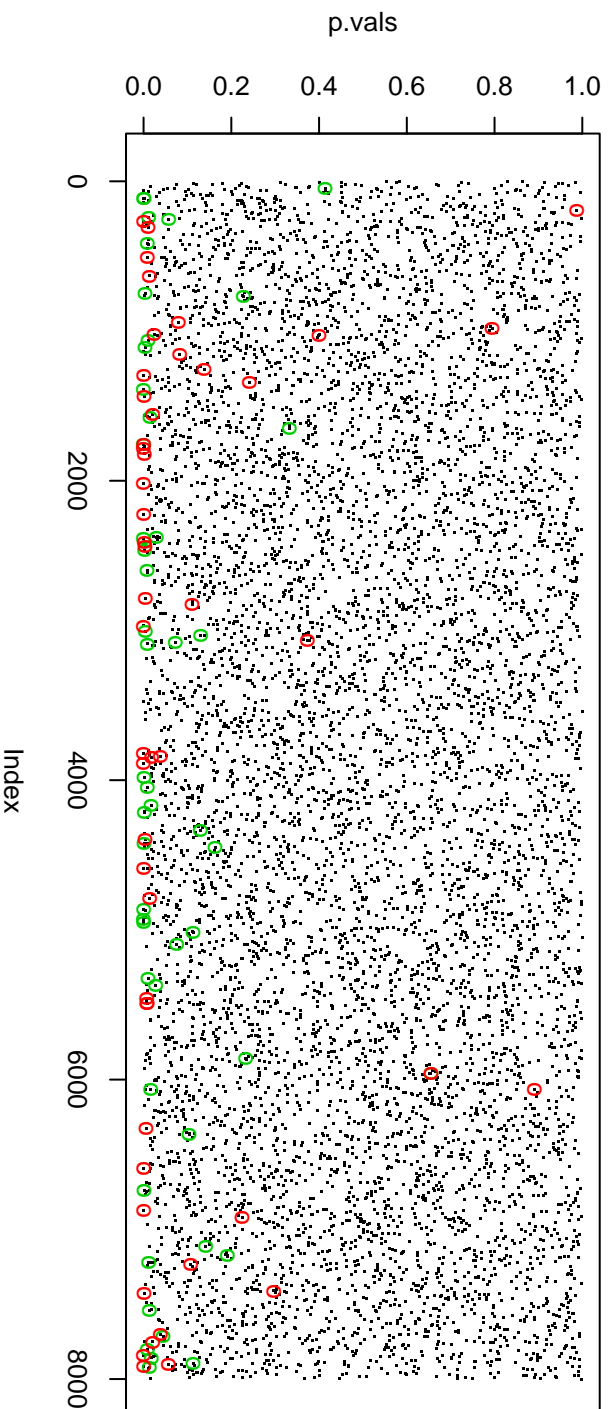
\tilde{u} -tilde for the simulation experiment



- Some “differentially expressed genes ” are lost in the middle

Methods: The Test Statistic \tilde{u} —Application (3)

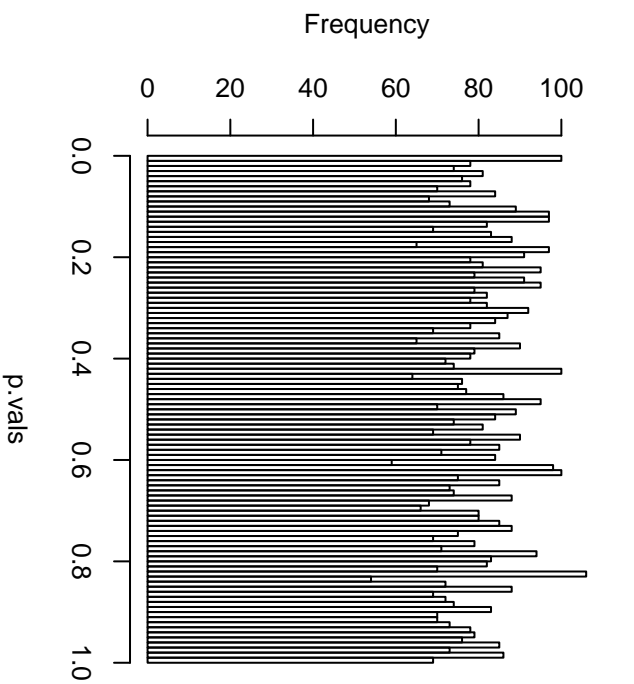
P-Values (simulation)



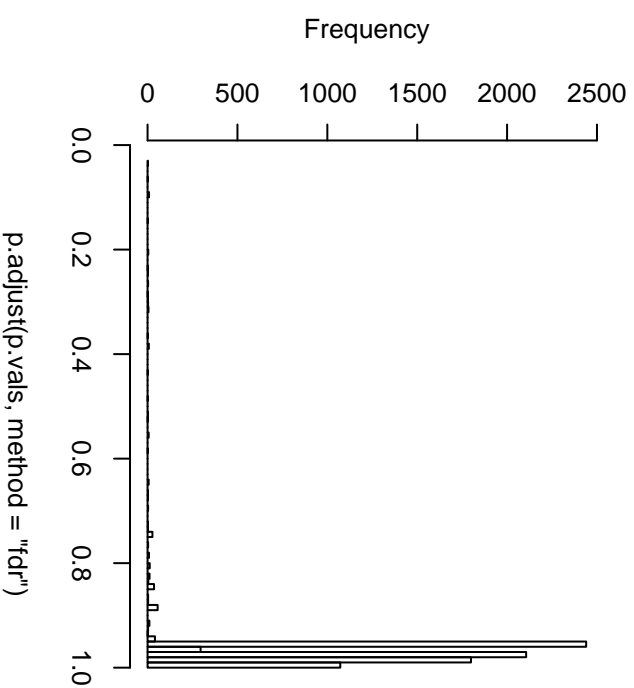
- Some “differentially expressed genes ” are lost in the middle, *and...*

Methods: The Test Statistic \tilde{u} —Application (4)

P-Values (simulation, non-adjusted)



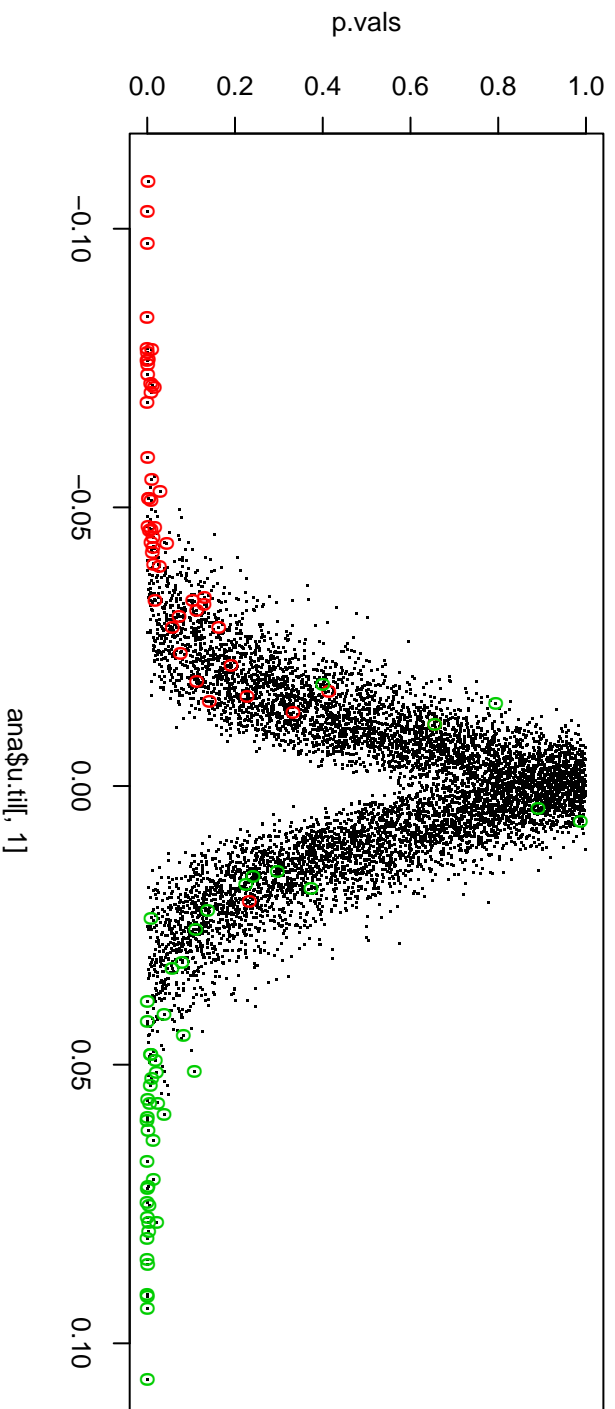
P-Values (simulation, adjusted)



- ... there are lots of false positives (or almost no significant genes)

Methods: The Test Statistic \tilde{u} —Application (5)

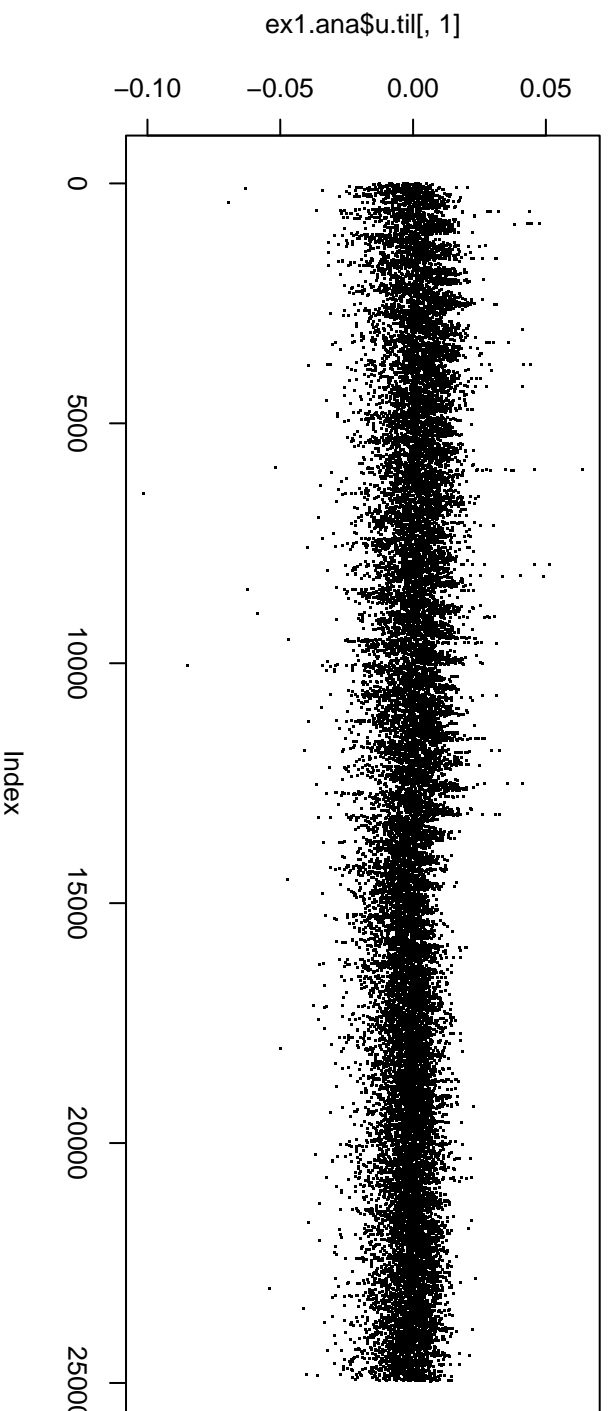
\tilde{u} -tilde vs P-Values (simulation)



- Comparison of P-Values and \tilde{u}

Methods: The Test Statistic \tilde{u} —Application (6)

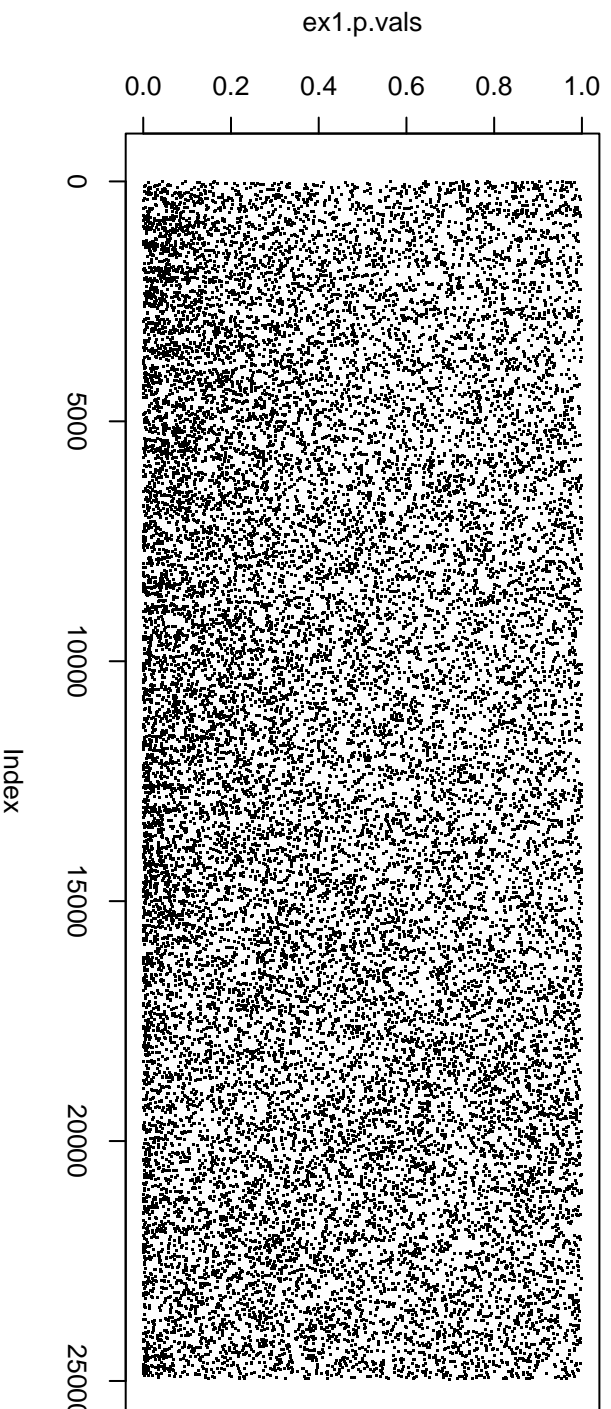
\tilde{u} —tilde for experiment 1



- Strange oscillations occur

Methods: The Test Statistic \tilde{u} —Application (7)

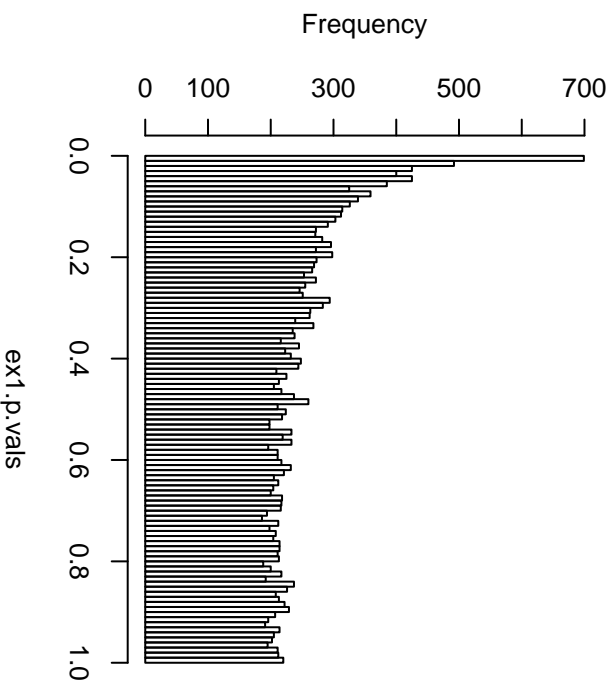
P-Values (experiment 1)



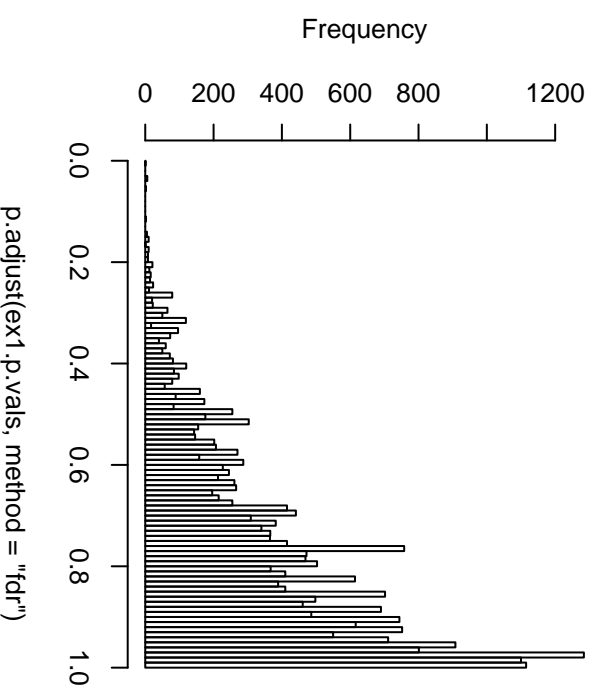
- No visible pattern

Methods: The Test Statistic \tilde{u} —Application (8)

P-Values (experiment 1, non-adjusted)



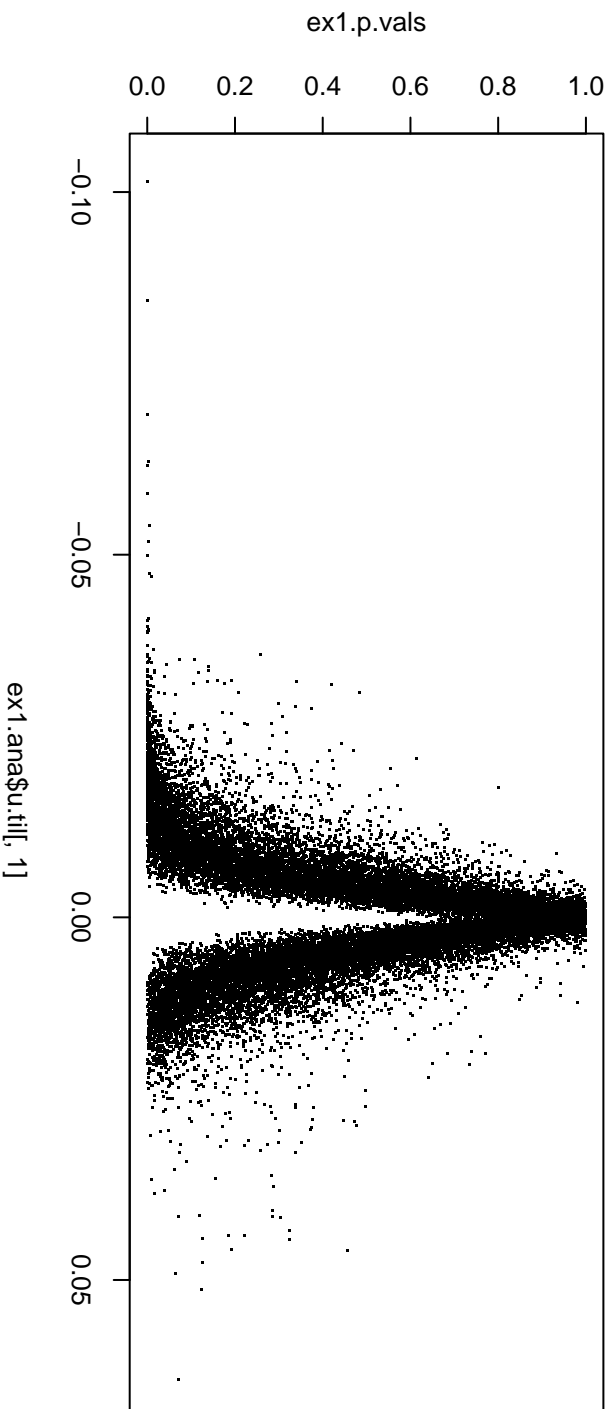
P-Values (experiment 1, adjusted)



- Many—or (nearly) no hits

Methods: The Test Statistic \tilde{u} —Application (9)

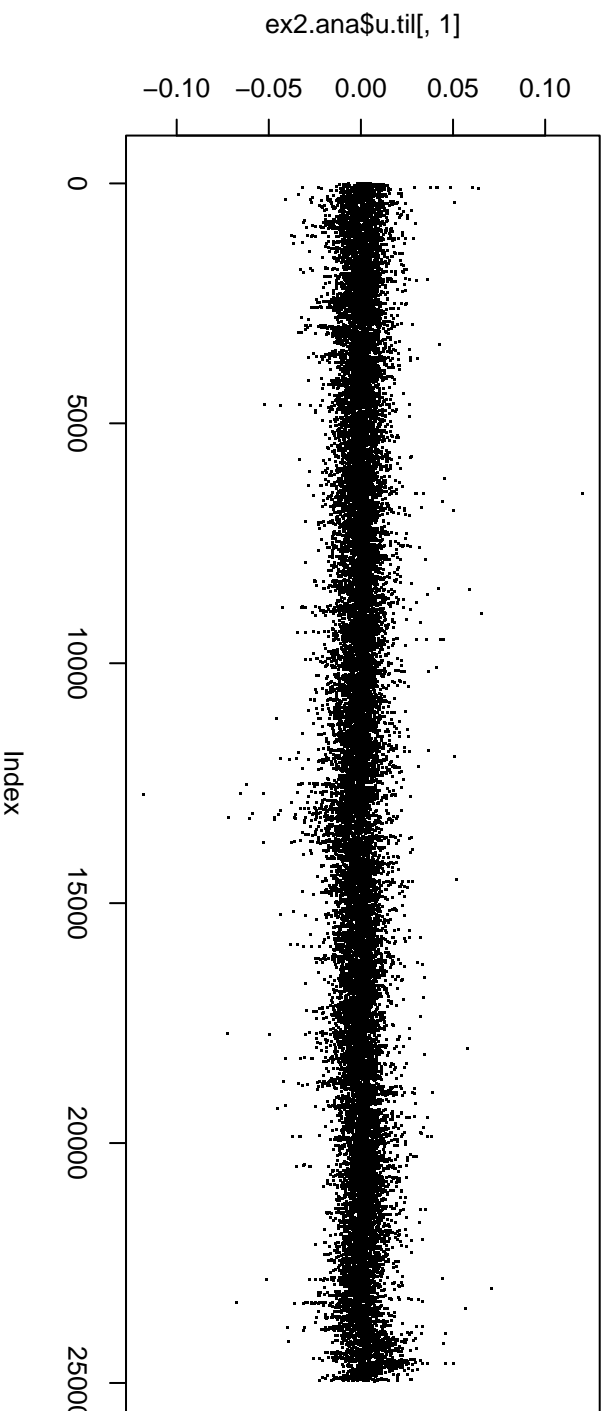
\tilde{u} -tilde vs P-Values (experiment 1)



- Comparison of P-Values and \tilde{u}

Methods: The Test Statistic \tilde{u} —Application (10)

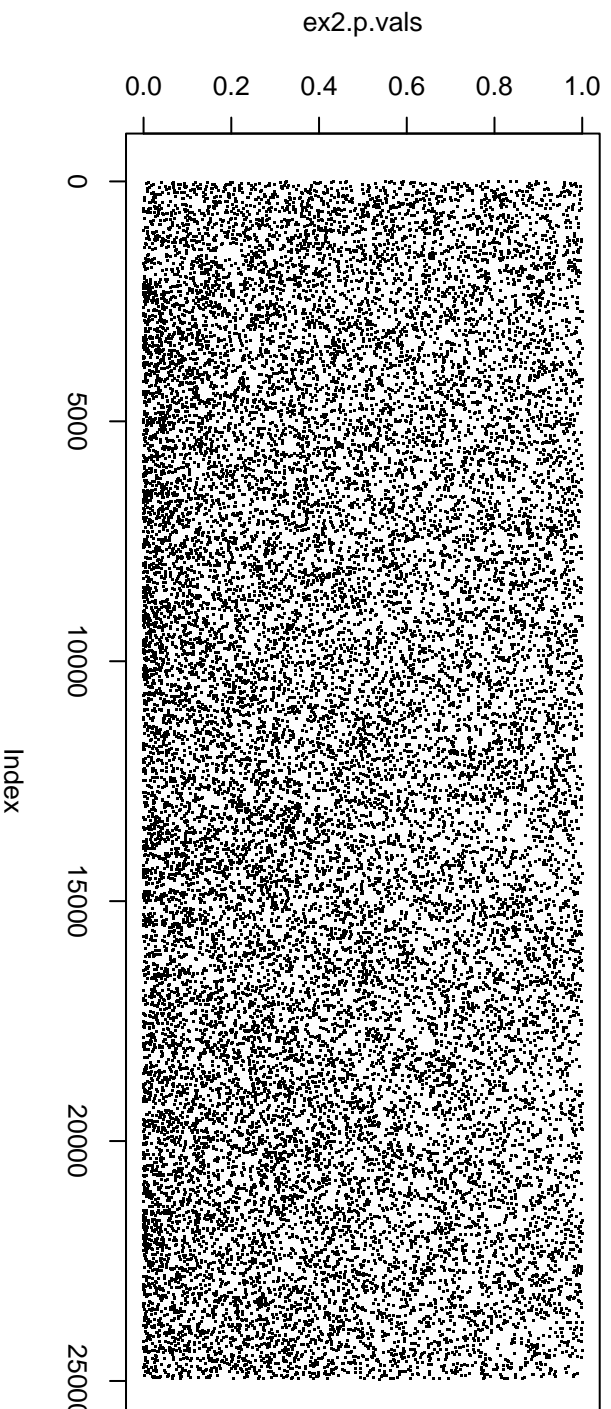
\tilde{u} —tilde for experiment 2



- Again, strange oscillations occur (not as strong)

Methods: The Test Statistic \tilde{u} —Application (11)

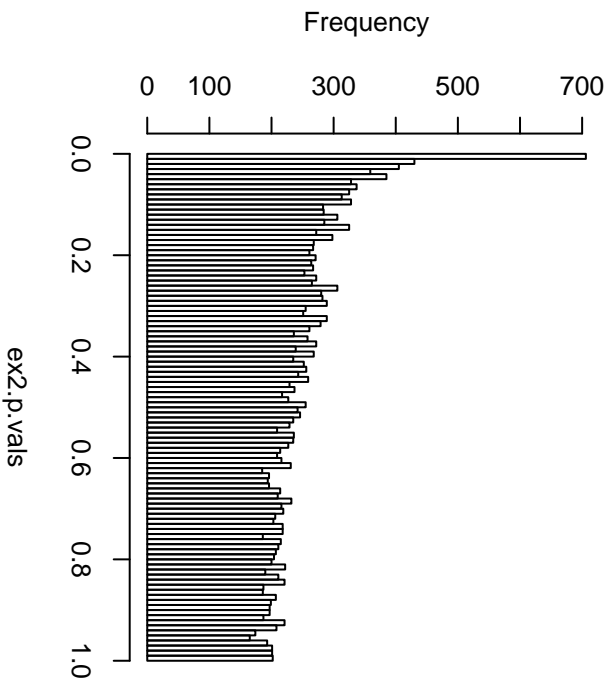
P-Values (experiment 2)



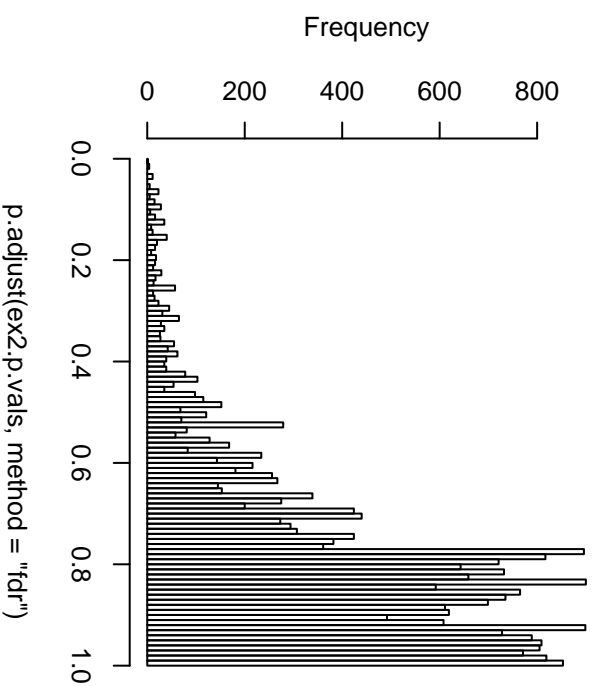
- Again, no visible pattern

Methods: The Test Statistic \tilde{u} —Application (12)

P-Values (experiment 2, non-adjusted)



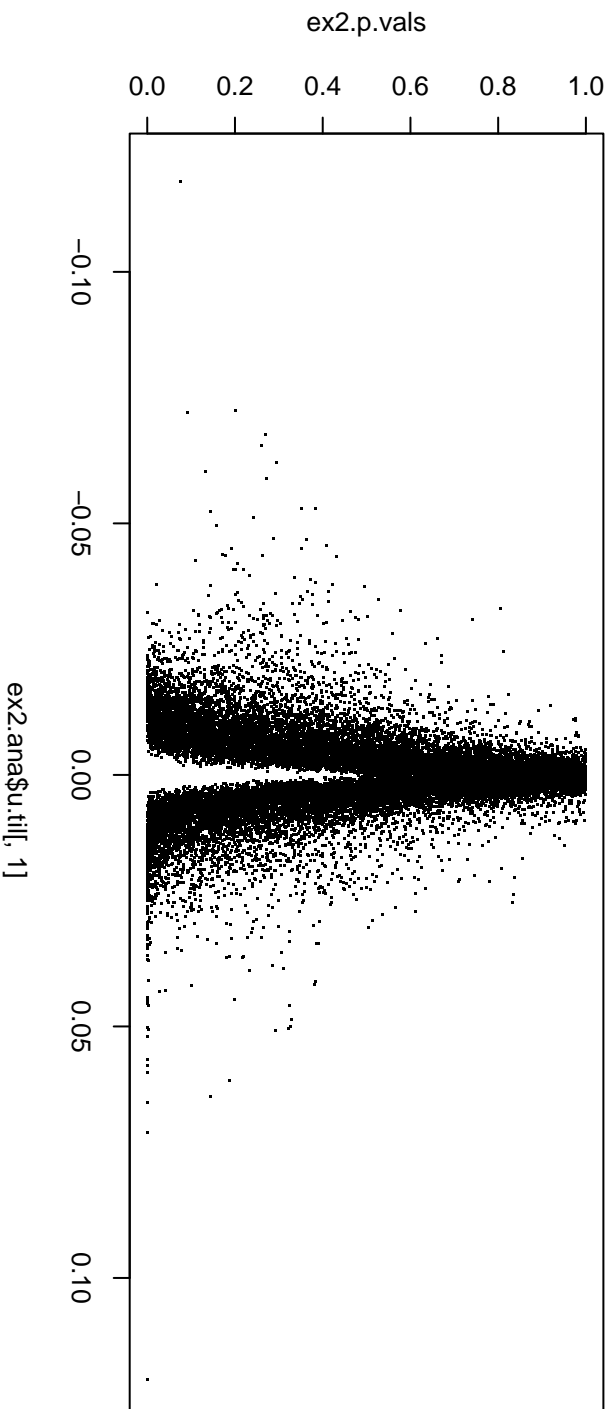
P-Values (experiment 2, adjusted)



- Many—or very few hits

Methods: The Test Statistic \tilde{u} —Application (13)

\tilde{u} -tilde vs P-Values (experiment 2)



- Comparison of P-Values and \tilde{u}

Methods: The Test Statistic \tilde{u} —Conclusion

- Simulation shows that the method works
- It seems to be rather specific
- It discovers spatial effects not visible with p-values
- The sets of selected genes will be different from those selected by p-values

Outlook—or better: TODO

- Find out about oscillations
- Application as cross validation/classification method
- Automatic threshold determination for \tilde{u}
- Permutation test (FDR)
- Compare results to “conventionally analyzed” data
- Write diploma thesis