Three approaches to control weighted FDR and FWER in survival analysis

Przemyslaw.Biecek@gmail.com, MIM University of Warsaw, Poland, Europe

Outline

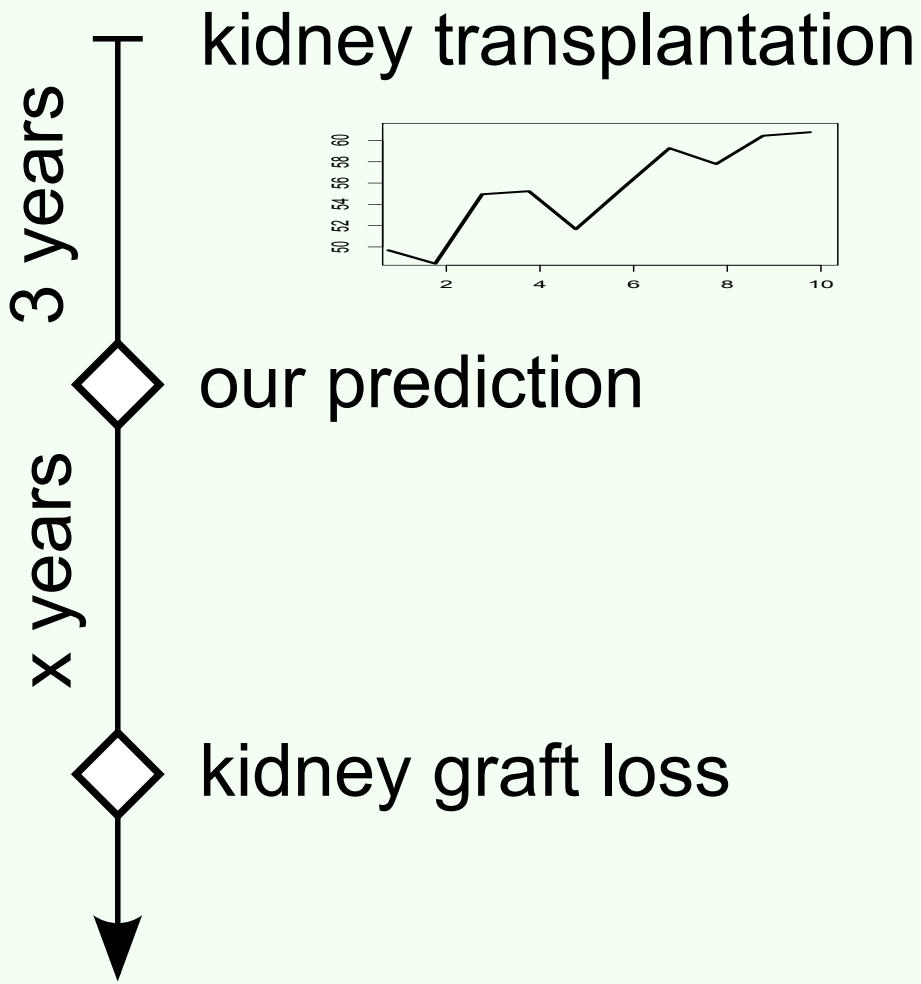
1. Dataset, problem and motivation
2. Our study
   - Step 1: censored vs. longitudinal variable,
   - model parameterization,
   - multiple hypothesis testing, part 1,
   - basis selection problem.
   - Step 2: binary vs. censored vs. longitudinal variable,
   - multiple hypothesis testing, part 2.
3. Results
4. Summary

Dataset

	A	B	C	D	E	F	G	H	I	J	K
1	sex	age	time followup	lost	MDRD7	MDRD6m	MDRD12m	MDRD18m	MDRD24m	MDRD30m	MDRD36m
2	K	56	5		40	49	40	48	44	59	49
3	M	43	5	2008	39	51	50	51	56	65	60
4	K	25	5,5		43	56	68	51	45	56	60
5	M	49	5,5		27	48	50	44	52	56	46
6	M	32	5,5	2005	48	48	52	52	48	45	48
7	...	...	...	...	...	...	...	...	...	...	...
8	K	48	8	2006	28	37	36	34	31	28	32
9	K	23	8		19	40	68	57	68	48	46
10	K	35	10,3		36	53	58	53	63	69	65
11	M	44	10,4	2003	44	66	66	60	72	78	77
12	M	53	10,5	2003	16	32	44	52	55	48	55
13	M	43	10,5	2006	25	48	47	48	50	52	49
14	M	32	14,2	2004	20	57	63	63	60	60	63
15	K	22	15	2003	40	50	58	62	53	68	74

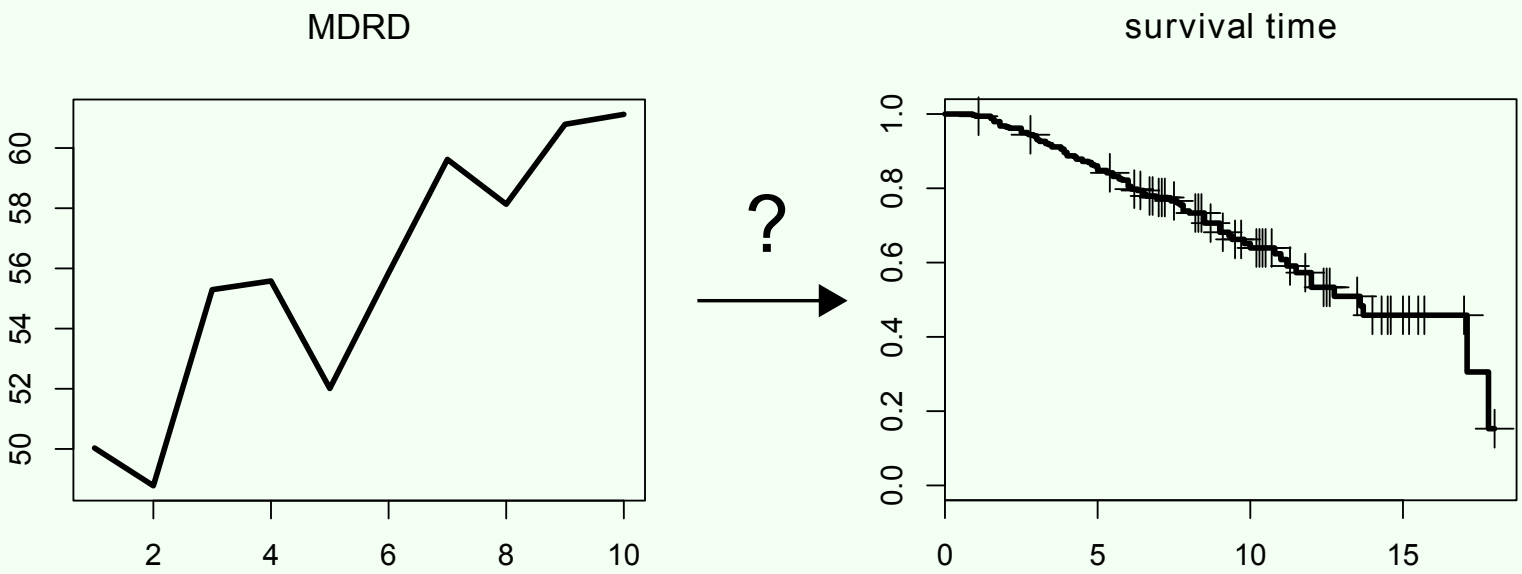
Dataset: 342 patients after kidney transplantation.
Every patient is examined every 6 months (many traits are measured, e.g. MDRD), for every patient also kidney lifetime is measured (censored variable $Y_i = min(T_i, Z_i)$).
All patients are observed at least for 3 years.

Dataset



During this three years patients are egzamined every six months, many (>300) traits are measured.
After third year we want to evaluate the graft function and predict the graft lifetime.

General model, step I



Nephrologists know that there is a significant relation between these two variables, but till now there is no golden model for this relation.

How we can find (and model) a relation between these two variables?

Similar problems - longitudinal data and survival study

Similar analyses are performed also for genetic/genomic data.
An active area of genomic research is related with statistical methods for linking gene expressions to various clinical and phenotypic characteristics such as clinical events: death or recovery.
E.g. in the article **„Survival analysis of longitudinal microarrays”** N. Rajicic, D. Finkelstein, D. Schoenfeld et. al., Bioinformatics 2006.
Authors deal with the problem of determining the relation between longitudinal variable such as gene expressions measured in series of timepoints and censored variable as lifetime (objects with different lifetimes will have different number of measurements).

Popular approaches in such studys

- The most popular approach is to apply the random effects model for the longitudinal data. This model provides a way to incorporate subject-specific random intercept and slope:

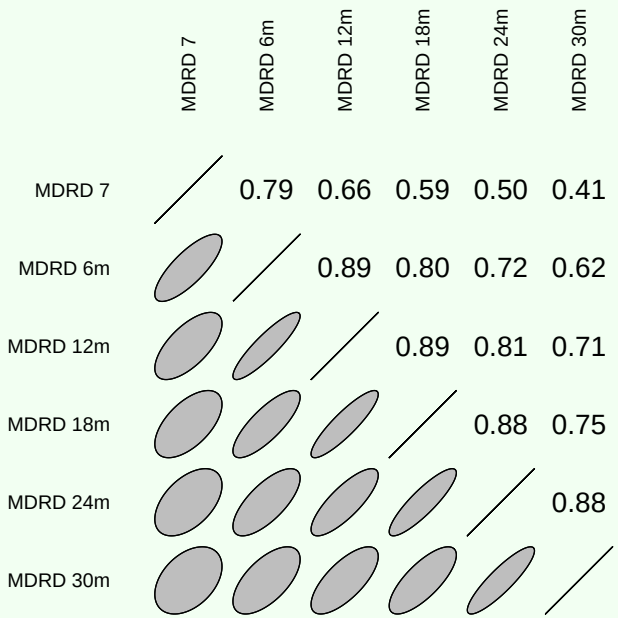
$$X_i(t) = \alpha_{0i} + \alpha_{1i}t, i = 1...n.$$

- The fixed effects model for the longitudinal data, in this case effect for each time point is estimated what not necessary „the question" (this will be discussed).

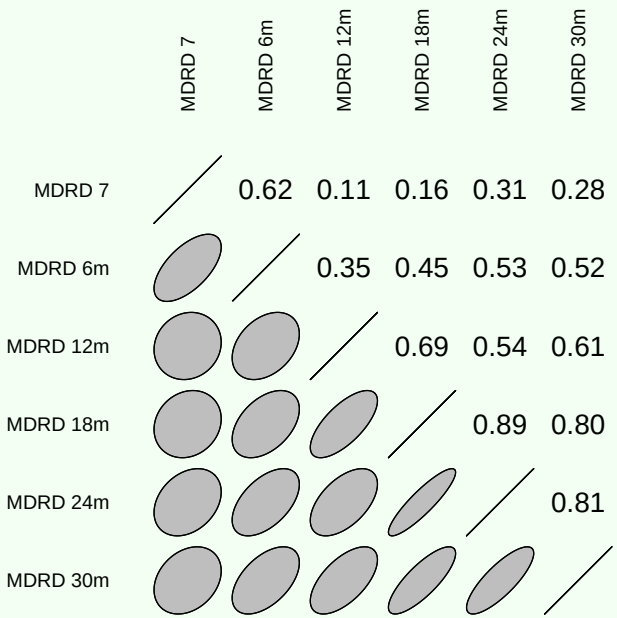
$$X_i(t) = \beta_0 + \beta_1 MDRD_1 + \beta_2 MDRD_2 + ...$$

- The fixed effects model with artificial variables calculated as mean, variance, linear trend etc of original variables (but how to construct this artificial variables?).
- Assume Cox proportional hazard model and analyse each time interval independently.

MDRD correlations



test statistics correlations



Eigen values for correlation matrices

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
MDRD	**5.44**	**0.89**	**0.31**	0.17	0.09	0.08	0.06
test statistics	**4.29**	**1.42**	**0.56**	0.37	0.25	0.08	0.06

Why we don't want to use original MDRD values

There is a lot reasons to avoid raw MDRD:

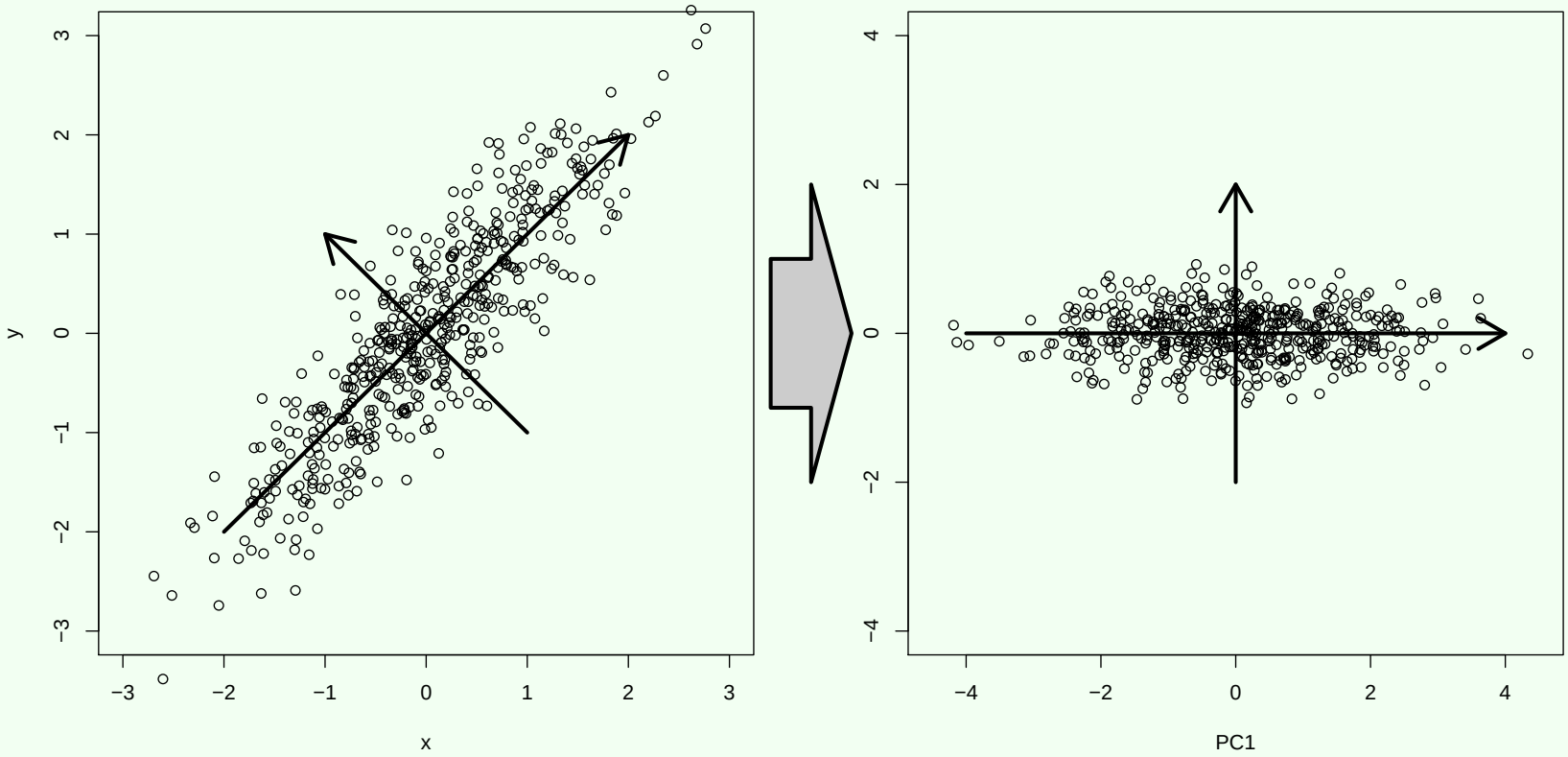
- high correlation $\rightarrow$ conservative control of error rate, unstable results, higher variance of estimated regression coefficients,
- high fraction of missing data $\rightarrow$ models for different number of variables have different number of observations, and are incomparable, only $\approx 66\%$ of patients have complete all measurements,
- this is not „the question"!

Artificial variables as linear combinations of original data

Since original MDRD measurements are correlated we may want to construct new uncorelated variables. There is many ways to construct such variables, let us consider following approaches:

- New variables are constructed in PCA fashion, in the next step we use them in the survival regression (PCR).
  Disadvantages: transformation is data dependent and do not incorporate information about time relationship,
  Instead of PCA one may consider other transformations like:
  - Non-negative matrix factorization (NMF)
  - Independent component analysis (ICA)
- New variables are constructed as projections on to k dimensional orthonormal basis. There is many orthonormal bases in space of all functions on interval [0,1].
  We may use different orthonormal bases, following results are presented for Legendre polynomials and cosine functions ($f_j(x) = \pi jx$, $j > 0$ and $f_0(x) \equiv 1$).

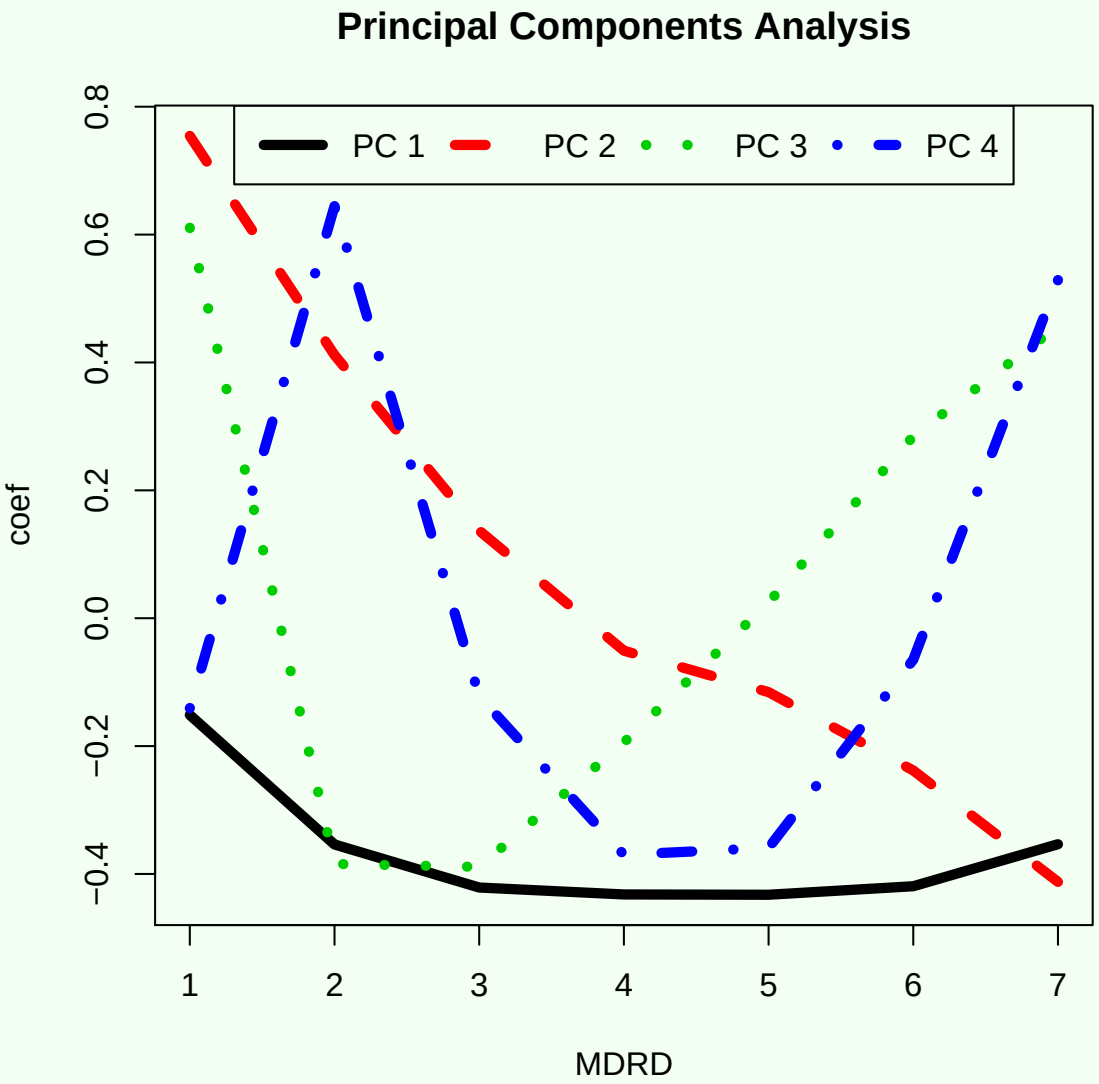
Principal Components Analysis

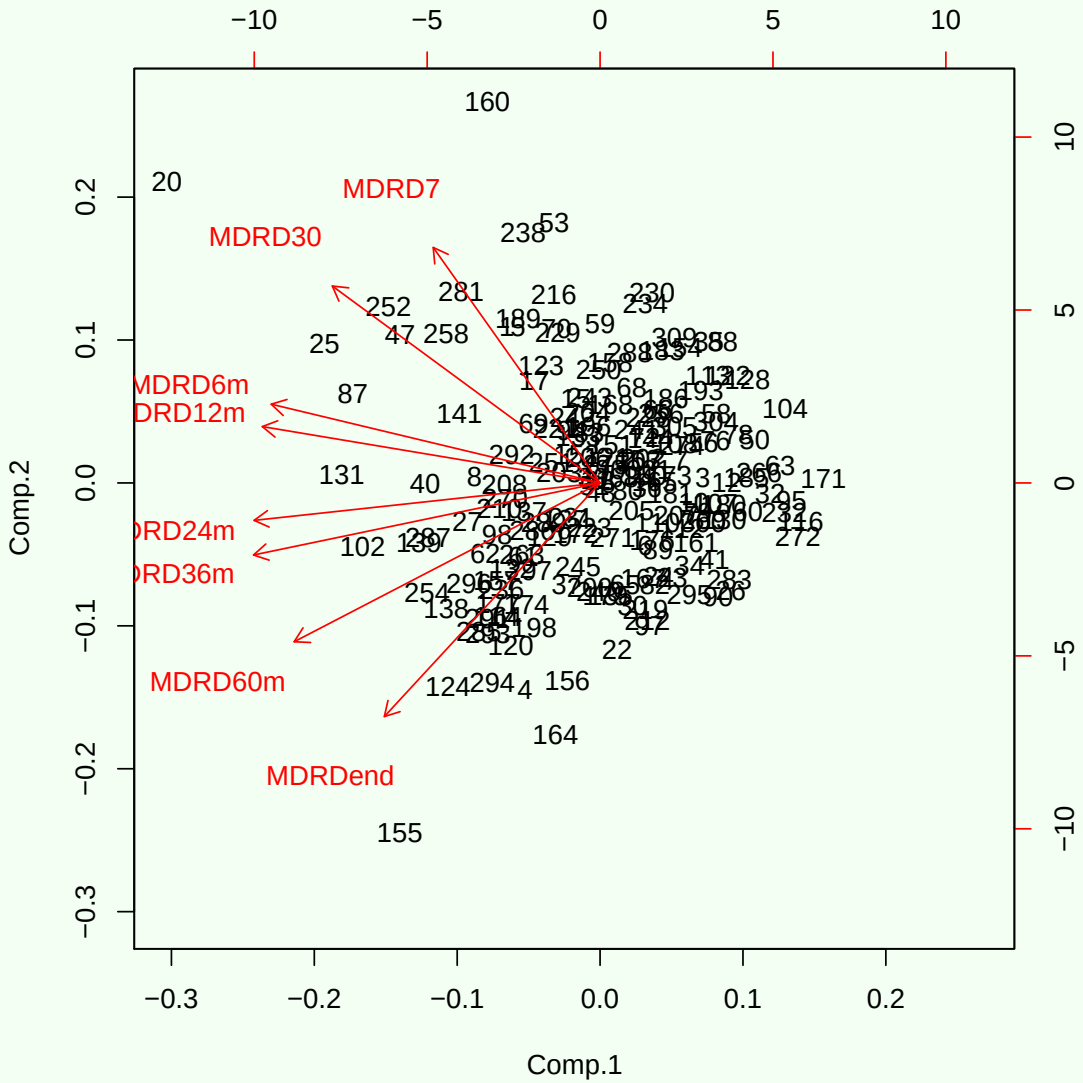


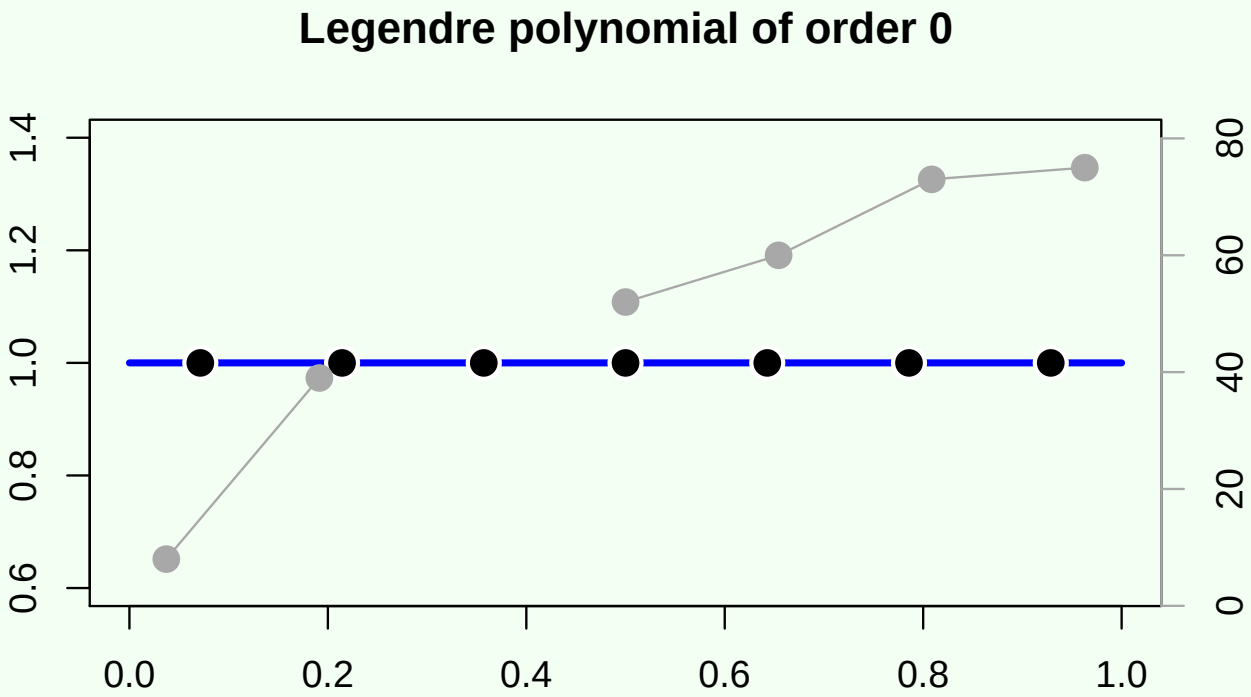
Principal Components Analysis

```
> loadings(princomp(daneMDRD))

Loadings:
        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
MDRD7   -0.265  0.828 -0.492
MDRD6m  -0.427 -0.123         0.510  0.503  0.537
MDRD12m -0.452 -0.195 -0.147  0.539 -0.402 -0.532
MDRD18m -0.452 -0.195 -0.147  0.539 -0.402 -0.532
MDRD24m -0.390 -0.272 -0.224 -0.384 -0.552  0.521
MDRD30m -0.452 -0.195 -0.147  0.539 -0.402 -0.532
MDRD36m -0.384 -0.313 -0.289 -0.497  0.527 -0.382
```

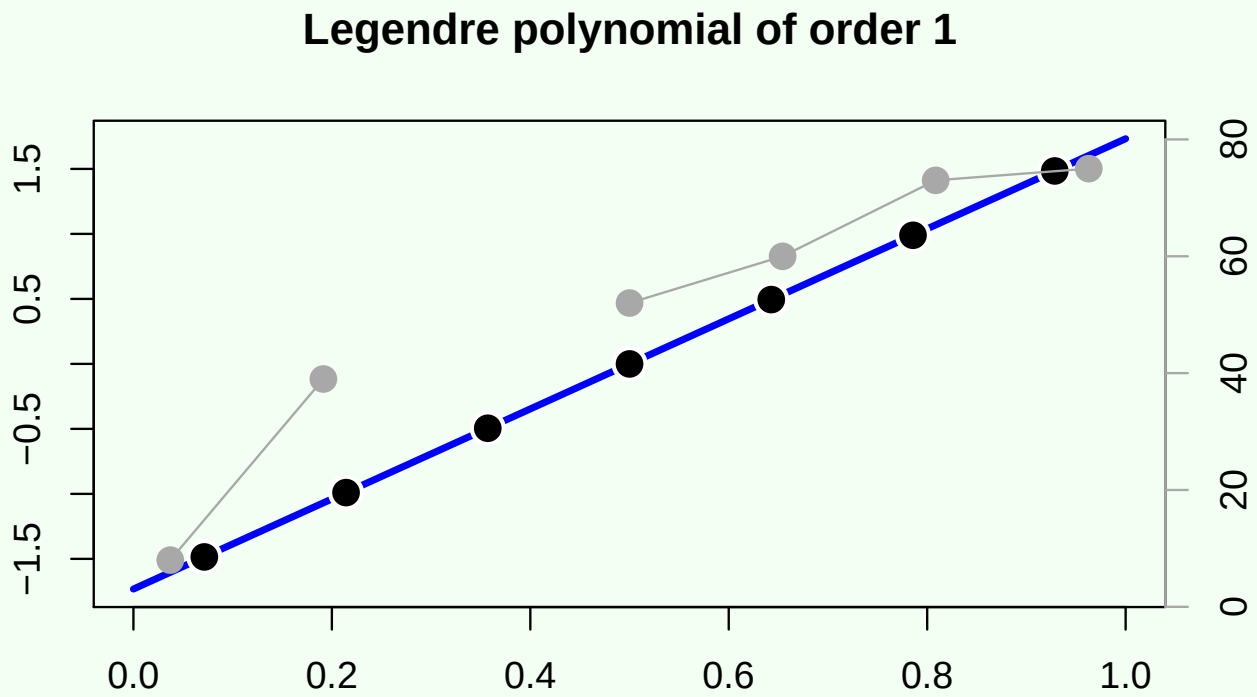

Principal Components Analysis
PC 1
PC 2
PC 3
PC 4
coef
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
1
2
3
4
5
6
7
MDRD


MDRD7
MDRD30
MDRD6m
DRD12m
DRD24m
DRD36m
MDRD60m
MDRDend
Comp.1
Comp.2



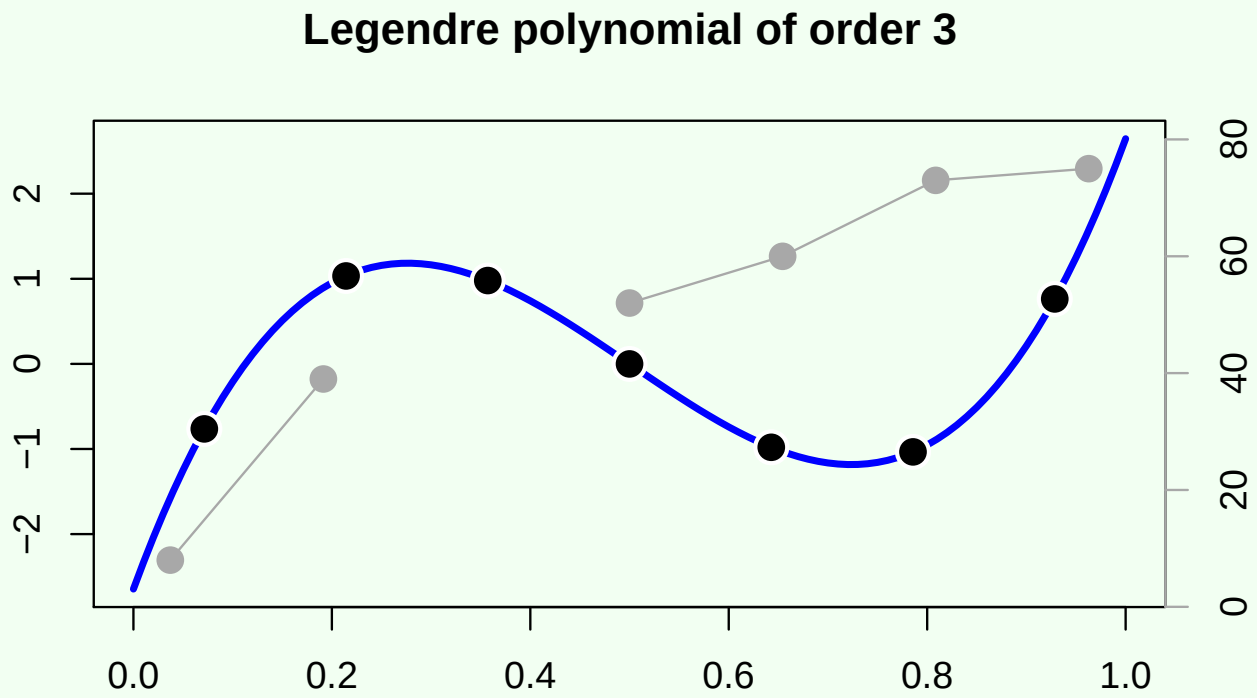
$f_0(x)$	1	1	1	1	1	1	1
MDRD	8	39	NA	52	60	73	75

$$f_0(x) \circ MDRD = 51.166$$



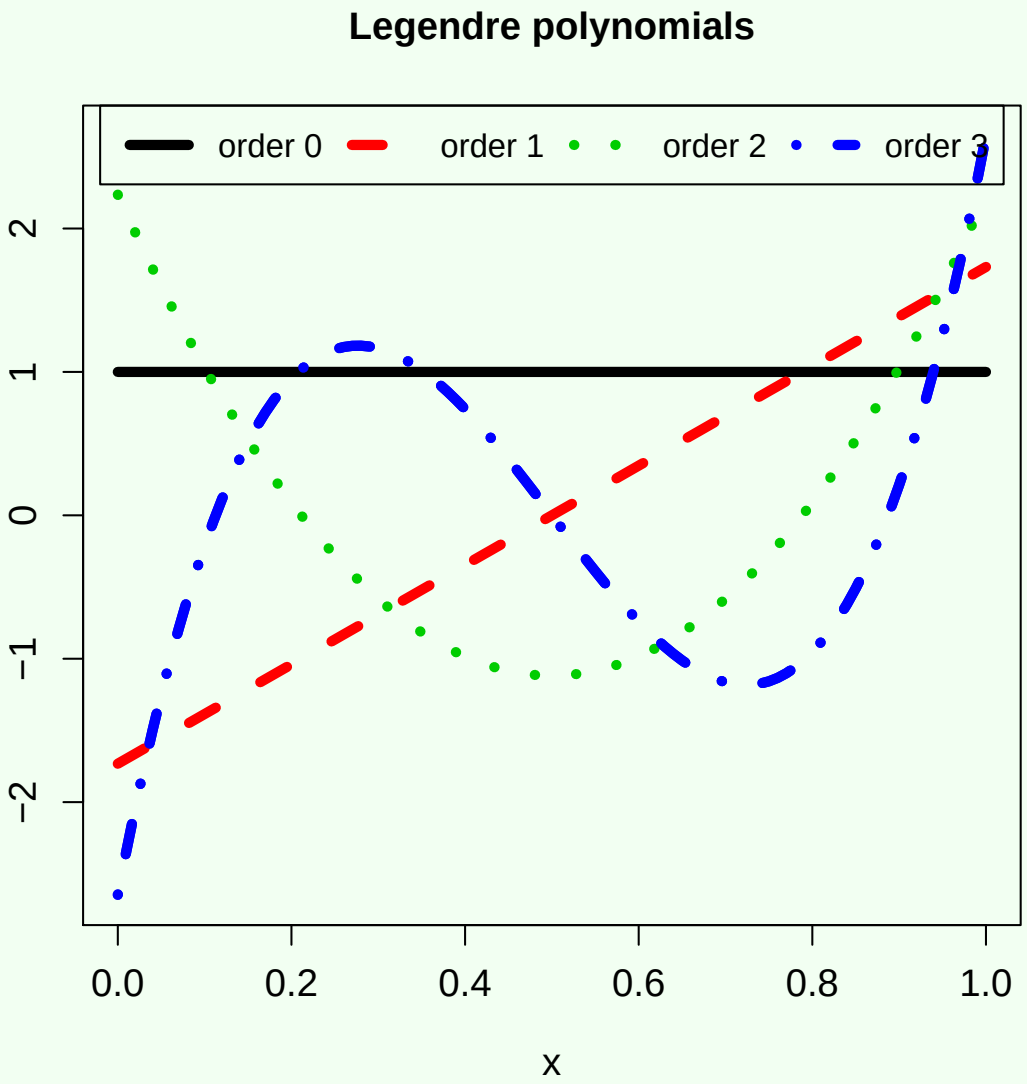
$f_1(x)$	-1.484	-0.989	-0.494	0	0.494	0.989	1.484
MDRD	8	39	NA	52	60	73	75

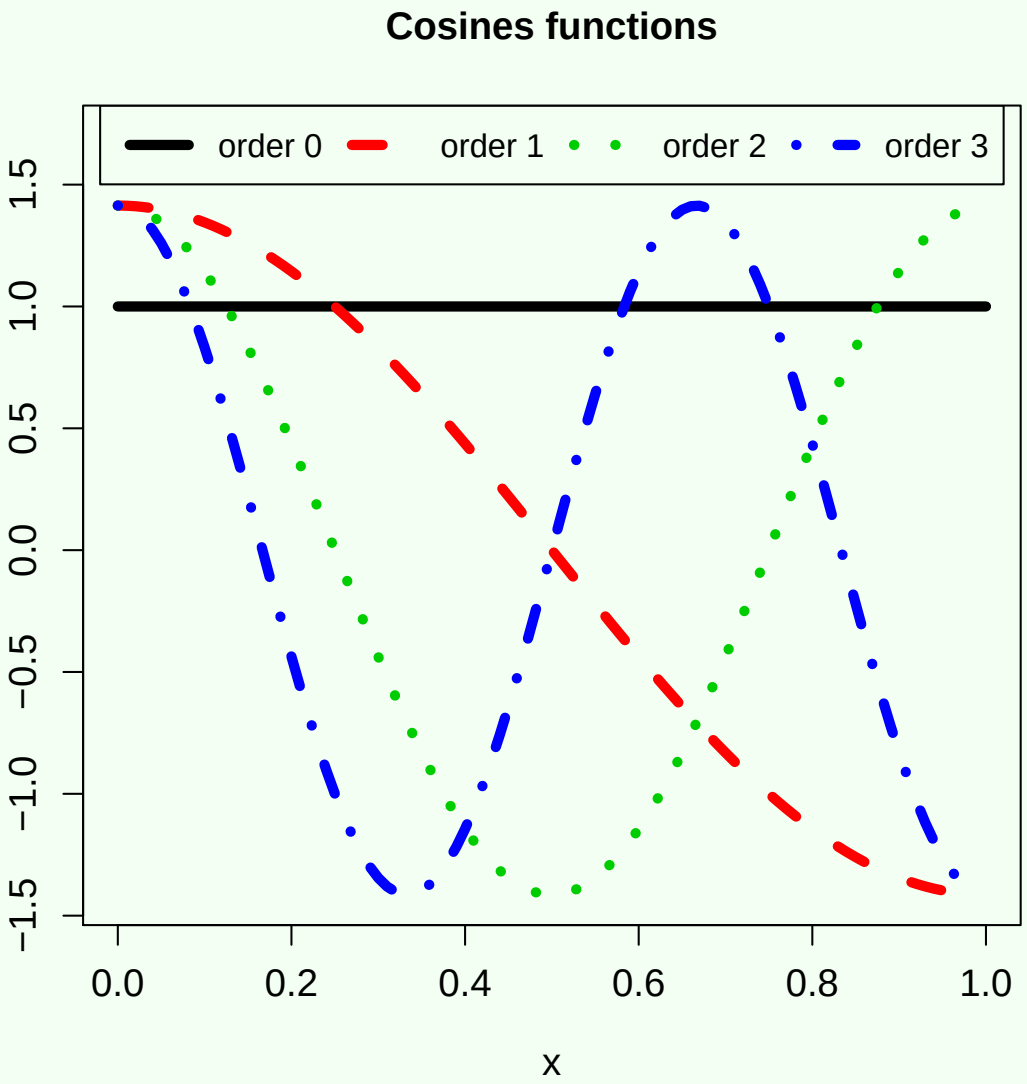
$$f_1(x) \circ MDRD = 27.13546$$



$f_3(x)$	-0.763	1.033	0.979	0	-0.979	-1.033	0.763
MDRD	8	39	NA	52	60	73	75

$$f_3(x) \circ MDRD = -7.126$$


Legendre polynomials
order 0
order 1
order 2
order 3
2
1
0
−1
−2
0.0
0.2
0.4
0.6
0.8
1.0
x

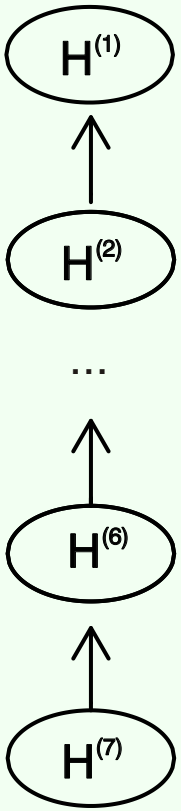

Cosines functions
order 0
order 1
order 2
order 3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0.0
0.2
0.4
0.6
0.8
1.0
x

Different importance

In all presented cases (PCA, Legendre polynomials and cosines) different coordinates have different meanings and different importance. In all cases first coordinates have most straightforward interpretation or explain higher proportion of variance (in case of PCA).
Thus, it is natural to treat them with different weights.
We propose two approaches to do this:

- Choose the subset of predictors of the form $\{f_0, ..., f_j\}$, so all coordinates from 0 to j, that explain all linear relations between MDRD and survival time.
- Assign different a priori probabilities for following coordinates and use them as weights in testing scheme, to construct a set of variables that explain all linear relations between MDRD and survival time.

Subsets of consecutive predictors



For the first strategy we can make use of following remarks:

Remark 1:

One may use follow up testing strategy with significance levels α for consecutive tests, and this strategy hold the FWER on level α.

Remark 2:

One may use follow up testing strategy with significance levels $\alpha/(1+\alpha)$ for consecutive tests, and this strategy hold the FDR on level α.

Weighted testing with a priori distribution for coordinates

The a priori probabilities a_i may be transformed into weights using transformation (Kustra, Biecek, Zagdański 2009)

$$w_i = \left(\frac{1 - a_i}{a_i}\right)^{1/c}.$$

where c is a constant and should be chosen that $\sum_i w_i = 1$.
If we can assign weights for following coordinates we may use following theorem to control FDR:

Theorem 3: (Genovese 2006)
If weights w_i fulfill $\sum_i w_i = 1$ then using modified p-values $q_i = p_i/w_i$ with standard step-up Benjamini-Hochberg procedure give control of FDR on level α.

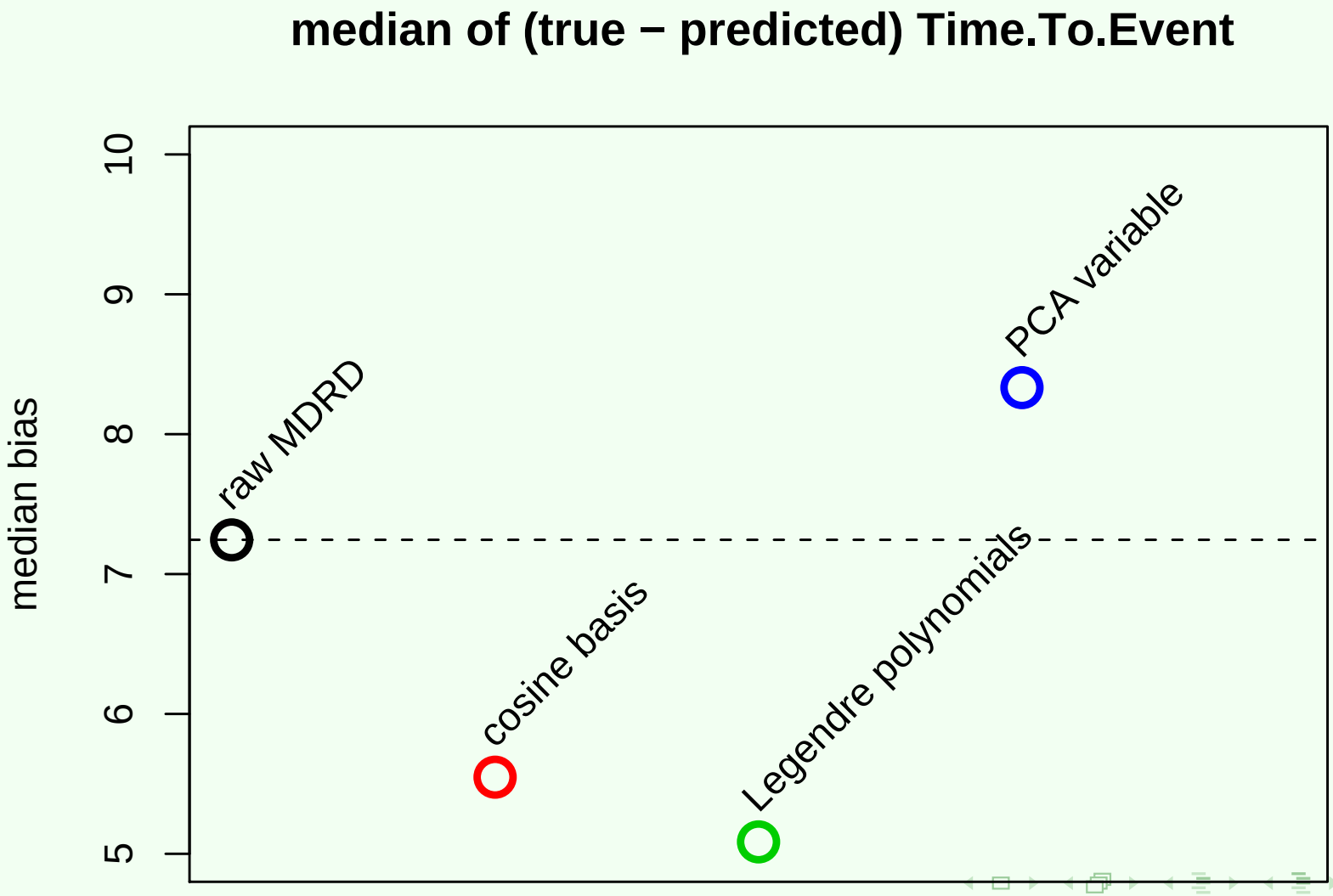
Results

Below we presents p-values for LRT for consecutive coordinates (model with k coordinates is compared with model with $k+1$).

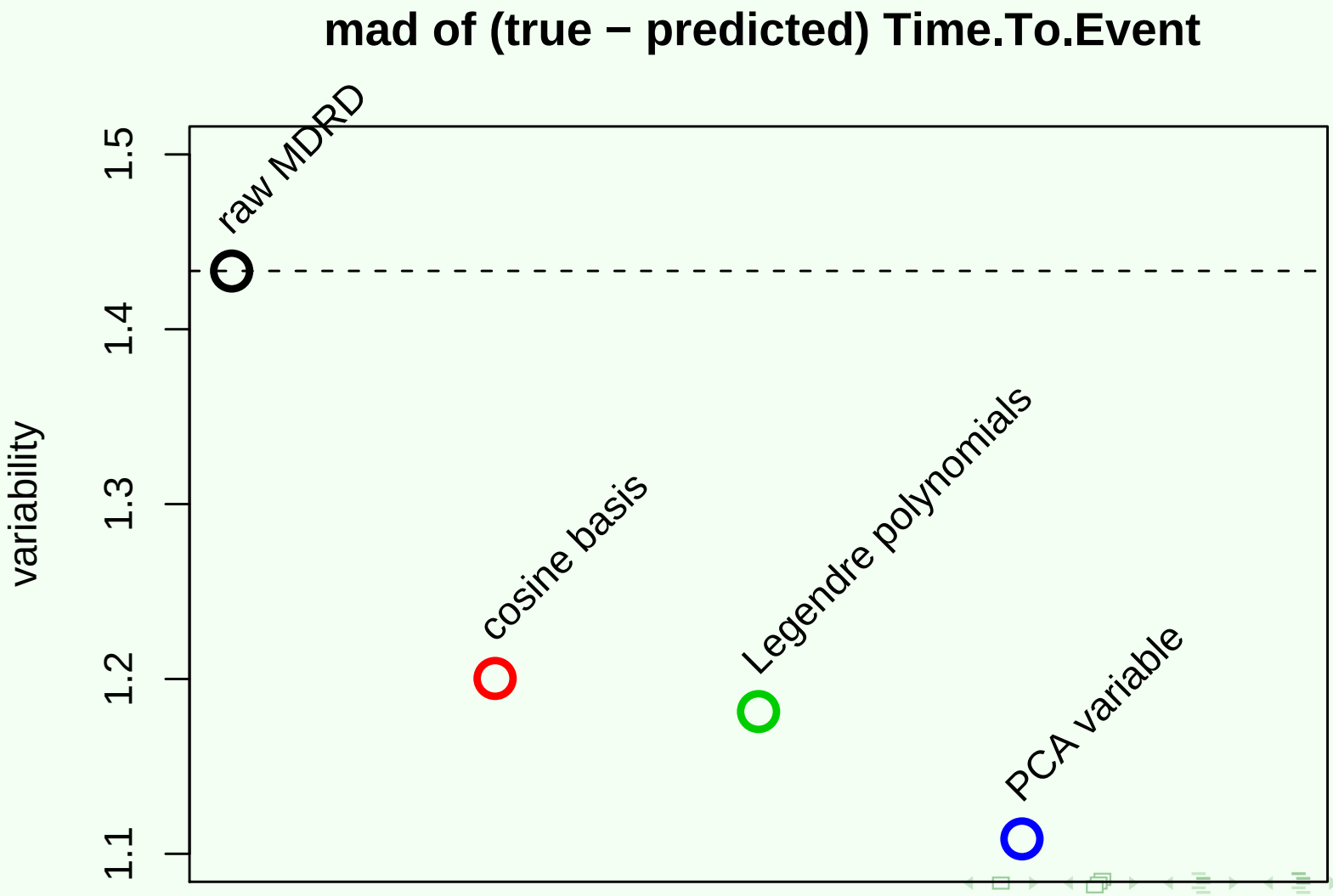
model size	cosines	Legendre polynomials	PCA
1	$\mathbf{1.1 * 10^{-7}}$	$\mathbf{1.1 * 10^{-7}}$	**0.0007**
2	$\mathbf{5.1 * 10^{-10}}$	$\mathbf{7.0 * 10^{-11}}$	**0.0027**
3	$\mathbf{7.5 * 10^{-10}}$	$\mathbf{1.4 * 10^{-8}}$	**0.037**
4	$\mathbf{2.4 * 10^{-10}}$	$\mathbf{1.2 * 10^{-10}}$	0.260
5	0.62	0.64	0.012
6	0.53	0.31	0.76
LRT	145.98	146.68	44.52
df	4	4	3

In our case both Legendre and cosines bases give similar results, we will choose Legendre basis in further step.

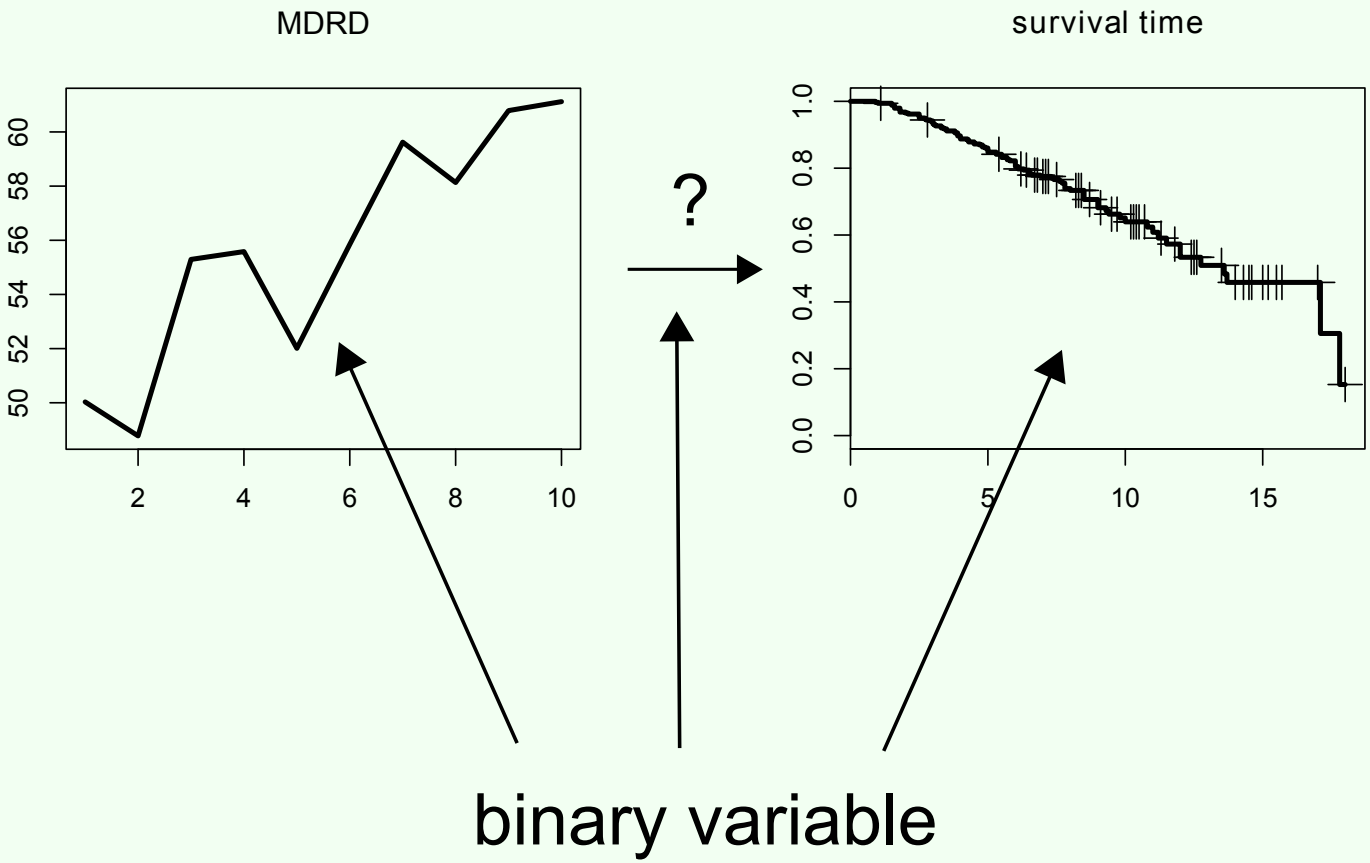
Median bias - the lower the better



Median absolute deviation - the lower the better



Second iteration



Nephrologists know that there are many additional factors which influence the relationship between MDRD and kidney lifetime.

Considered hypotheses

For each binary variable X_i we perform three tests:

- $H_0^{(1,i)}$: MDRD is not related with binary variable X_i.

  For this hypothesis we can use standard MANOVA test. Due to missing data we will perform MANOVA on constructed artificial variables. The reduction for number of variables improve power of this test.

Considered hypotheses

For each binary variable X_i we perform three tests:

- $H_0^{(1,i)}$: MDRD is not related with binary variable X_i.
- $H_0^{(2,i)}$: survival time is not related with binary variable X_i.

We use the log-rank test for survival data to compare two survival curves.

Considered hypotheses

For each binary variable X_i we perform three tests:

- $H_0^{(1,i)}$: MDRD is not related with binary variable X_i.
- $H_0^{(2,i)}$: survival time is not related with binary variable X_i.
- $H_0^{(3,i)}$: the relation between MDRD and survival time is the same for both levels of of binary variable X_i.

We use survival regression to test the hypothesis of interaction in the considered model.

Considered hypotheses

For each binary variable X_i we perform three tests:

- $H_0^{(1,i)}$: MDRD is not related with binary variable X_i.
- $H_0^{(2,i)}$: survival time is not related with binary variable X_i.
- $H_0^{(3,i)}$: the relation between MDRD and survival time is the same for both levels of of binary variable X_i.

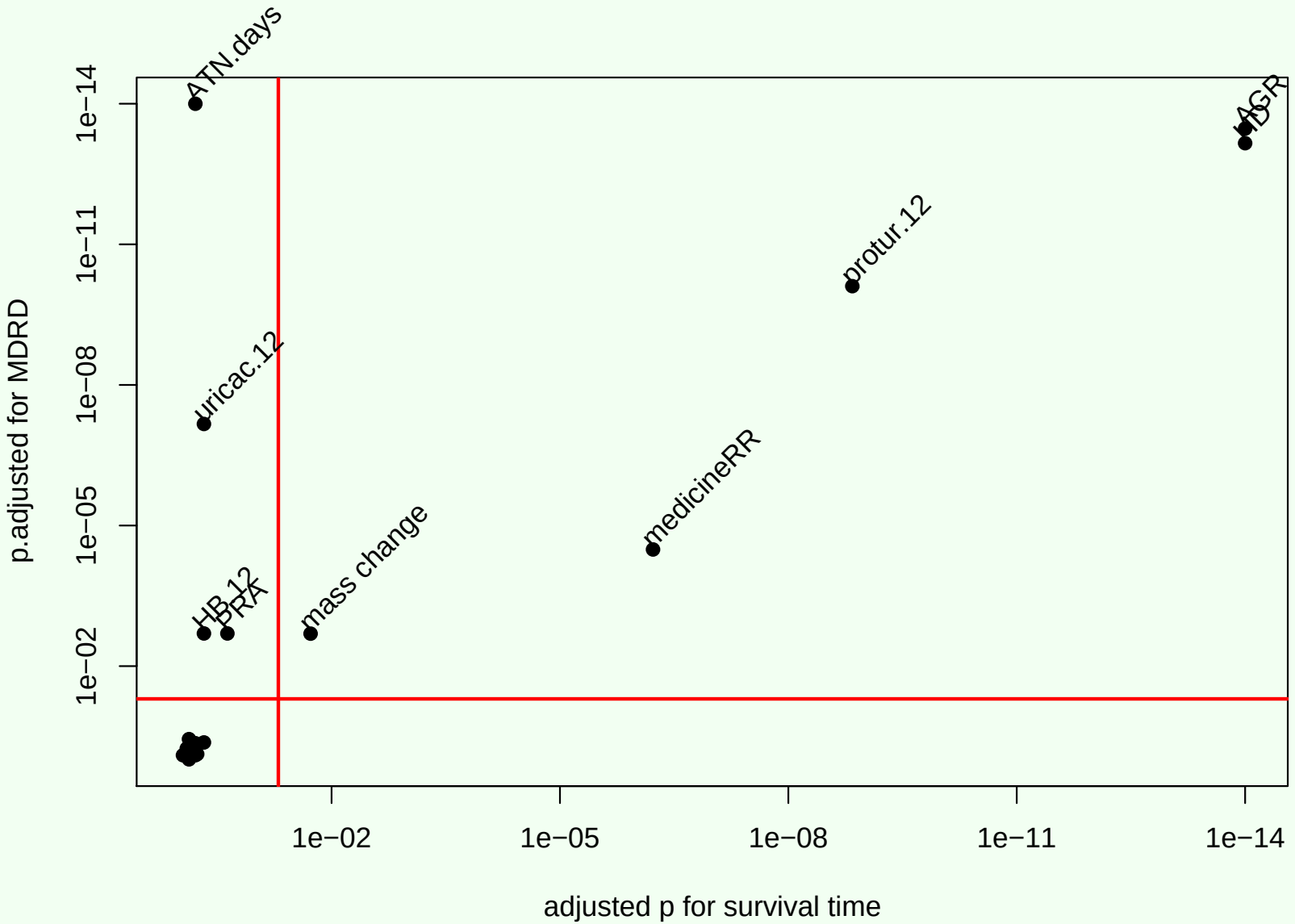
Thus for m binary variables we have $3 \times m$ test to perform.

Stratification

According to Lei Sun et al 2006 if we consider some number of groups of hypotheses, and we expect different proportion of true signals in different groups then we may perform testing procedure that controll FDR independently in each group of hypotheses. This is called stratification and gives better results (on average) than standard approach i.e. performing multiple testing procedure for all hypotheses together.
In our dataset we expected much more variables that somehow differentiate MDRD than variables that differentiate survival curves.

$$E(\sum H^{1,\cdot}) > E(\sum H^{2,\cdot})$$

We apply BH procedure independently to sets $H_0^{1,\cdot}$ and $H_0^{2,\cdot}$ and in our case stratification gives higher power.


ATN.days
uricac.12
HB.12
PRA
mass change
medicineRR
protur.12
AGR
HB
p.adjusted for MDRD
1e−02
1e−05
1e−08
1e−11
1e−14
adjusted p for survival time

Summary 1/2

In first part of this talk we have presented the model of relationship between longitudinal data and some (censored) variable

- we construct set of k new variables and then reduct this set by removing unimportant variables, smaller set of important variables is easier to deal with in next steps.
- The new variables are without missing data, we remove missing data in a very natural way.
- Models constructed for new variables are more accurate and are more stable.
- We may use weighted testing scheme, since in this case new variables have different importance.

Summary 2/2

In second part of this talk we have presented the model of relationship among binary variable, longitudinal data and some (censored) response variable

- We use results from the first part, new representation of data improve power of the MANOVA tests,
- Stratification of hypotheses gives higher power.

References

1. Ch.R. Genovese, K. Roeder, and L. Wasserman. False discovery control with p-value weighting. Biometrika, 93(3):509-524, 2006.
2. P. Biecek. Multiple Testing procedures for Hierarchically Related Hypotheses. PhD thesis 2007.
3. R. Kustra, P. Biecek, A. Zagdański. Optimal p-value ranking in multiple testing problems with applications to functional genomics. to appear, 2009
4. L. Sun, R.V. Craiu, A.D. Paterson, and S.B. Bull. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. Genetic Epidemiology, 30:519-530, 2006.
5. N. Rajicic, D. Finkelstein, D. Schoenfeld et. al. Survival analysis of longitudinal microarrays. Bioinformatics 2006.