NATIONAL RESEARCH UNIVERSITY

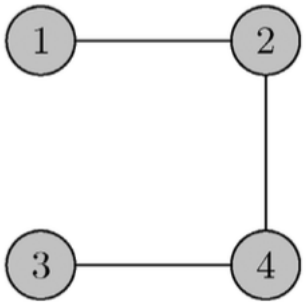
Statistical uncertainty in Gaussian graphical models selection

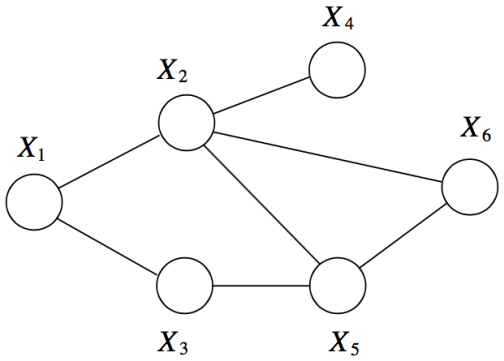
Гречихин Иван Сергеевич, аспирант 1 года

Руководитель: Колданов Александр Петрович

Graphical models

- Means for visualization





- Applications:
  - Informatics and Bioinformatics, Economics, etc.

Goals of research

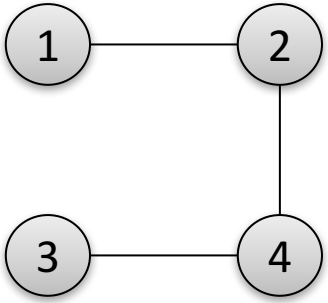
- Theoretical review of statistical uncertainty of identification procedures in Gaussian graphical models selection:
  - various errors and risk functions (FWER, FDR, FDP, …)
  - additive risk function и соответствующая ей функция риска
  - unbiasness of statistical identification procedures
  - robustness of statistical identification procedures properties to different distributions
  - optimality of known procedures regarding risk function for different types of statistical procedures
- Practice:
  - Mathematical modelling for uncovering statistical properties of different identification procedures. Theoretical and numerical comparison of the algorithms and their properties.

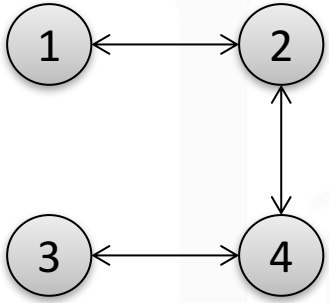
Literature

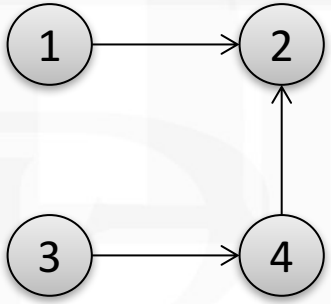
- Reviews
  - Mathias Drton, Michael D. Perlman. (2007) *Multiple Testing and Error Control in Gaussian Graphical Model Selection*. Statistical Science, Vol. 22, No. 3, 430–449.
  - Michael I. Jordan. *Graphical Models*. (2004) Statistical Science, Vol 19, No. 3, 140-155.
  - Mathias Drton, Marloes H. Maathuis. (2017) *Structure Learning in Graphical Modeling*. Annual Review of Statistics and Its Application, Vol. 4, 365-393.
- Statistical procedures
  - Mathias Drton, Michael D. Perlman. (2007) *Multiple Testing and Error Control in Gaussian Graphical Model Selection*. Statistical Science, Vol. 22, No. 3, 430–449.
  - Mathias Drton, Michael D. Perlman. (2008) *A SINful approach to Gaussian graphical model selection*. Journal of Statistical Planning and Inference, Vol. 138, 1179-1200.
  - Anna Gottard, Simona Pacillo. *Robust concentration graph model selection*. (2010) Computational Statistics and Data Analysis, Vol. 54, 3070-3079.
- Goodness-of-fit procedures
  - Juliane Schafer and Korbinian Strimmer. (2005) *An empirical Bayes approach to inferring large-scale gene association networks*. Bioinformatics, Vol. 21, No 6, 754-764.
  - Khondker Z.S et al. (2013) *The Bayesian Covariance Lasso.* Stat Interface, 6(2), 243–259.
  - Hastie T., Tibshirani R., Friedman J. (2009) *Undirected Graphical Models*. In: The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, NY

Types of graphical models

- Undirected: conditional independence, concentration/precision matrix, partial correlations
- Bidirected: marginal independence, covariance/correlation matrix
- Directed Acyclic graphs: conditional independence on a set of parent nodes







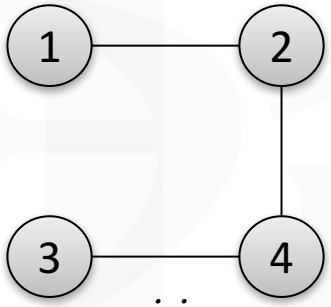
Undirected graphical models

$$Y = (Y_1, ..., Y_p) \sim N(\mu, \Sigma) \text{ - random vector}$$

σ_{ij} - elements of Σ

σ^{ij} - elements of $C = \Sigma^{-1}$

$$\rho^{ij} = \frac{\sigma^{ij}}{\sqrt{\sigma^{ii} * \sigma^{jj}}}$$



$$(i,j) \notin E \iff Y_i \perp\!\!\!\perp Y_j \mid Y_{V/\{i,j\}} \iff \rho^{ij} = 0$$

C	1	2	3	4
1	1	0.3	0	0
2	0.3	1	0	0.4
3	0	0	1	0.1
4	0	0.4	0.1	1

Statistical procedures

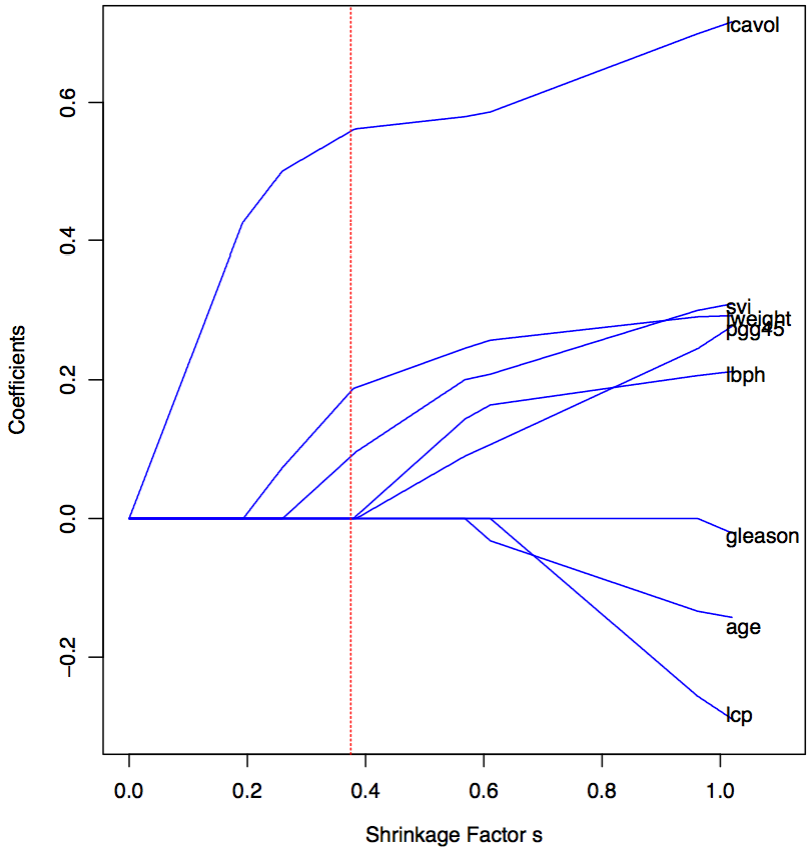
$$h_{ij} : \rho^{ij} = 0 \quad vs \quad k_{ij} : \rho^{ij} \neq 0$$

- Multiple testing procedure
  - Testing every hypothesis at predetermined significance level
- Family wise error rate (FWER) controlling procedures with p-value adjustment *
  - Possible adjustments are: Bonferroni adjustment, Sidak adjustment, Holm step-down procedures
- Standard measures of errors
  - Type I and II Errors (False Positives (FP) and False Negatives (FN))
  - Relations between FP, FN, True Positives (TP) and True Negatives (TN)
  - Risk function (Linear combination of errors)
- *Mathias Drton, Michael D. Perlman. (2007) *Multiple Testing and Error Control in Gaussian Graphical Model Selection*. Statistical Science, Vol. 22, No. 3, 430–449.

Optimisation procedures

$$\log\det C - trace(SC) - \lambda\|C\|_1$$

- L1 regularization for concentration matrix*
  - After optimising L1-regularized function, a number of parameters become zero.



- *Hastie T., Tibshirani R., Friedman J. (2009) *Undirected Graphical Models*. In: The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, NY
-
- .

Experiments

- Repetition and extension of experiments from the article of Drton & Perlman (2007)
  - Concentration matrix 7x7 with 9 non-zero elements, generated from the interval [0.2, 0.55]
  - Comparison of FWER at level alpha = 0.1 for the number of observations from 25 to 500
  - Bonferroni Adjustment: $\pi_{ij}^{Bonf} = \min\{1, \pi_{ij} * C_p^2\}$
  - Sidak Adjustment: $\pi_{ij}^{Sidak} = 1 - (1 - \pi_{ij})^{C_p^2}$
- Comparison of ROC AUC for different statistical procedures
  - AUC is an area under curve
  - Curve is plotted from data, where X = FP/FP+TN, Y = TP/TP+FN
- Estimation of risk function for different statistical procedures

Results

q=0.2, Type I Error Number	100	200	300	400	500
Bonferroni	0.076	0.088	0.072	0.066	0.082
Bon. Holm	0.078	0.09	0.076	0.071	0.085
Sidak	0.079	0.096	0.085	0.077	0.091
Sidak Holm	0.08	0.1	0.092	0.084	0.092

q=0.2, Type II Error Number	100	200	300	400	500
Bonferroni	36.522	28.451	25.676	23.846	22.574
Bon. Holm	36.363	28.336	25.574	23.716	22.425
Sidak	36.414	28.401	25.645	23.805	22.515
Sidak Holm	36.284	28.284	25.529	23.671	22.361

Results

q=0.6, Type I Error Number	100	200	300	400	500
Bonferroni	0.032	0.044	0.031	0.045	0.044
Bon. Holm	0.033	0.052	0.044	0.068	0.061
Sidak	0.032	0.046	0.033	0.05	0.047
Sidak Holm	0.034	0.056	0.046	0.071	0.063

q=0.6, Type II Error Number	100	200	300	400	500
Bonferroni	157.201	129.713	114.537	105.175	98.14
Bon. Holm	156.8	128.633	113.328	103.769	96.631
Sidak	156.959	129.413	114.299	104.945	97.895
Sidak Holm	156.554	128.347	113.071	103.523	96.375

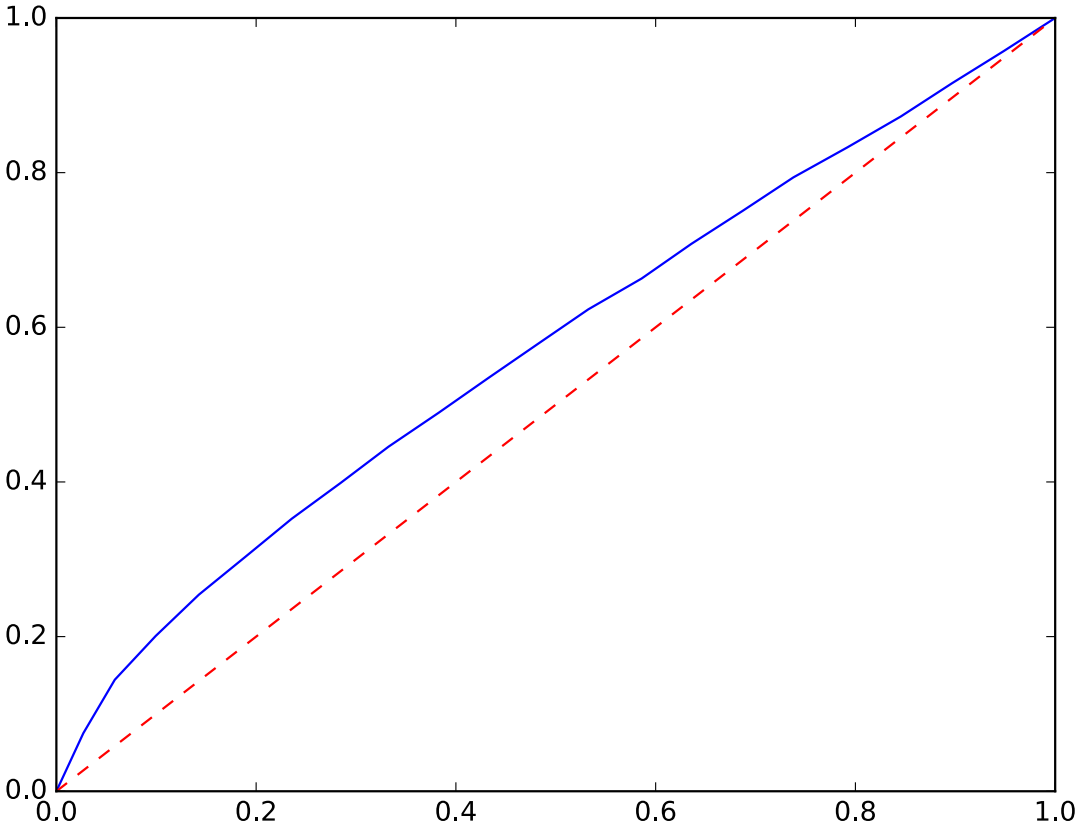
Results

q=0.95,Type I Error Number	100	200	300	400	500
Bonferroni	0.004	0.002	0.002	0.006	0.009
Bon. Holm	0.004	0.004	0.01	0.008	0.01
Sidak	0.004	0.002	0.002	0.006	0.009
Sidak Holm	0.004	0.005	0.01	0.009	0.01

q=0.95,Type II Error Number	100	200	300	400	500
Bonferroni	262.147	223.957	198.573	180.204	166.712
Bon. Holm	261.634	221.822	195.288	175.954	162.004
Sidak	261.815	223.51	198.107	179.762	166.256
Sidak Holm	261.278	221.341	194.786	175.395	161.472

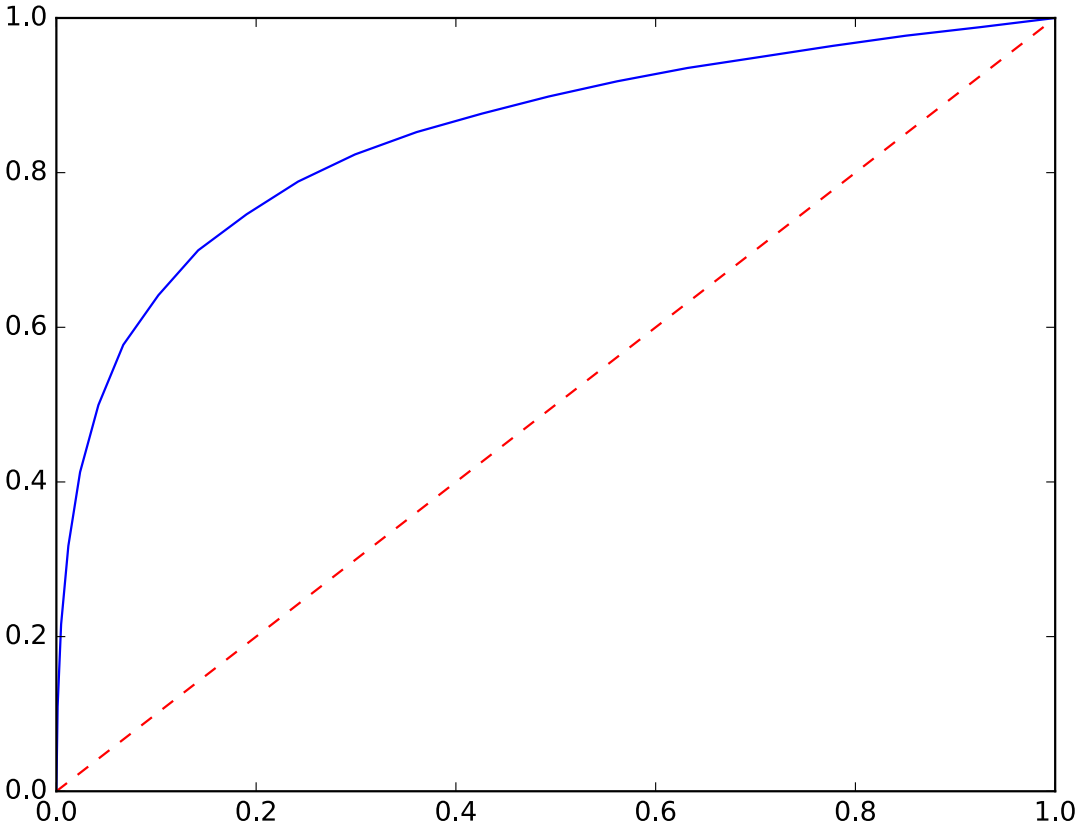
Results ROC AUC

ROC AUC, n = 10, multiple testing



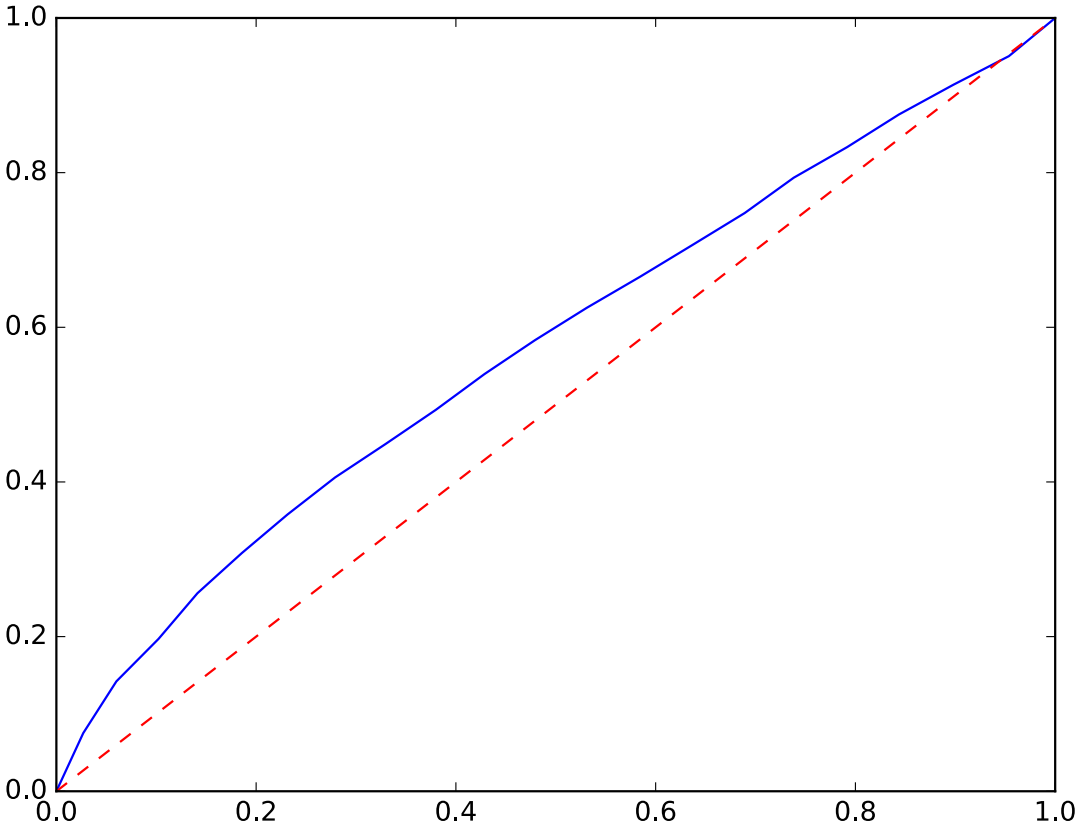
Results ROC AUC

ROC AUC, n = 40, multiple testing



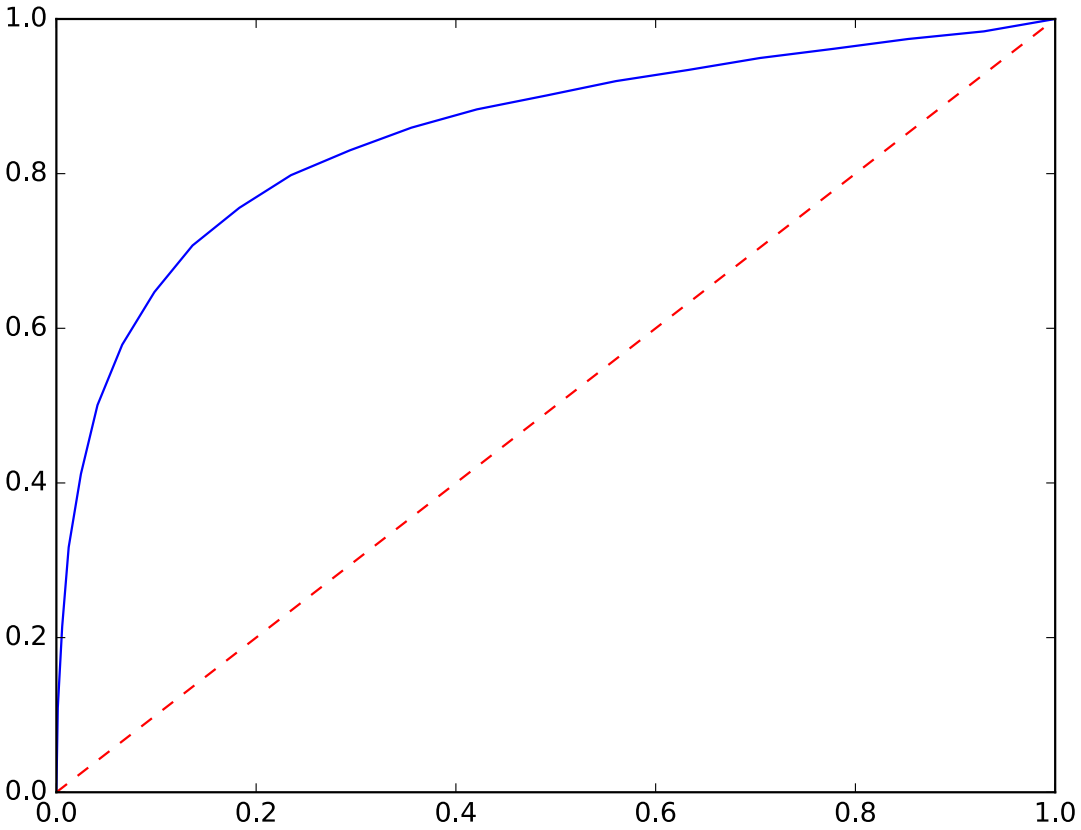
Results ROC AUC

ROC AUC, n = 10, Sidak adjustment



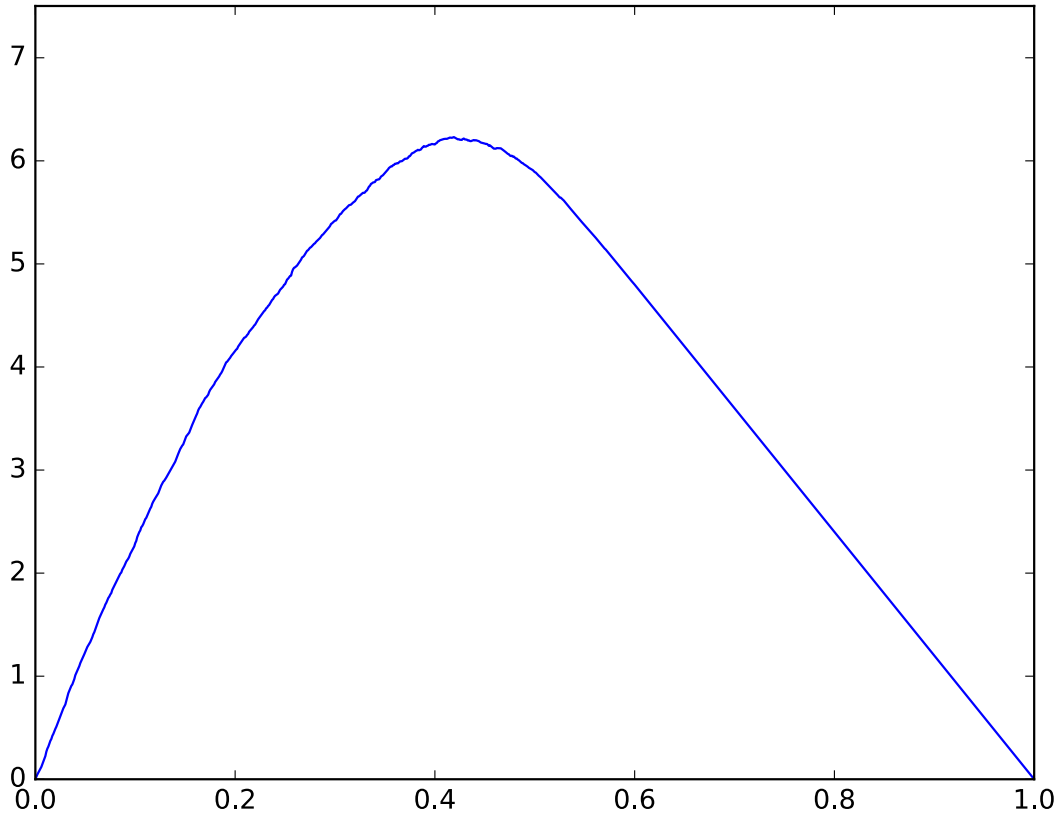
Results ROC AUC

ROC AUC, n = 40, Sidak adjustment



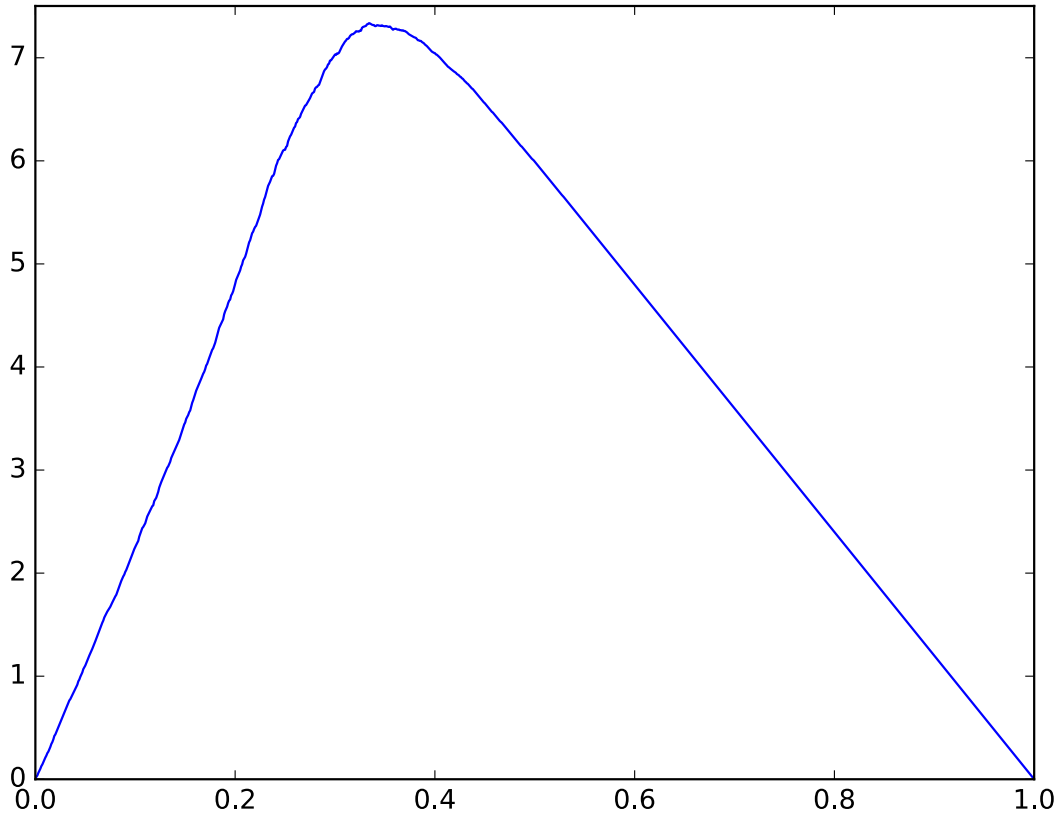
Results: Risk Function

Risk function, n = 10, Multiple Testing



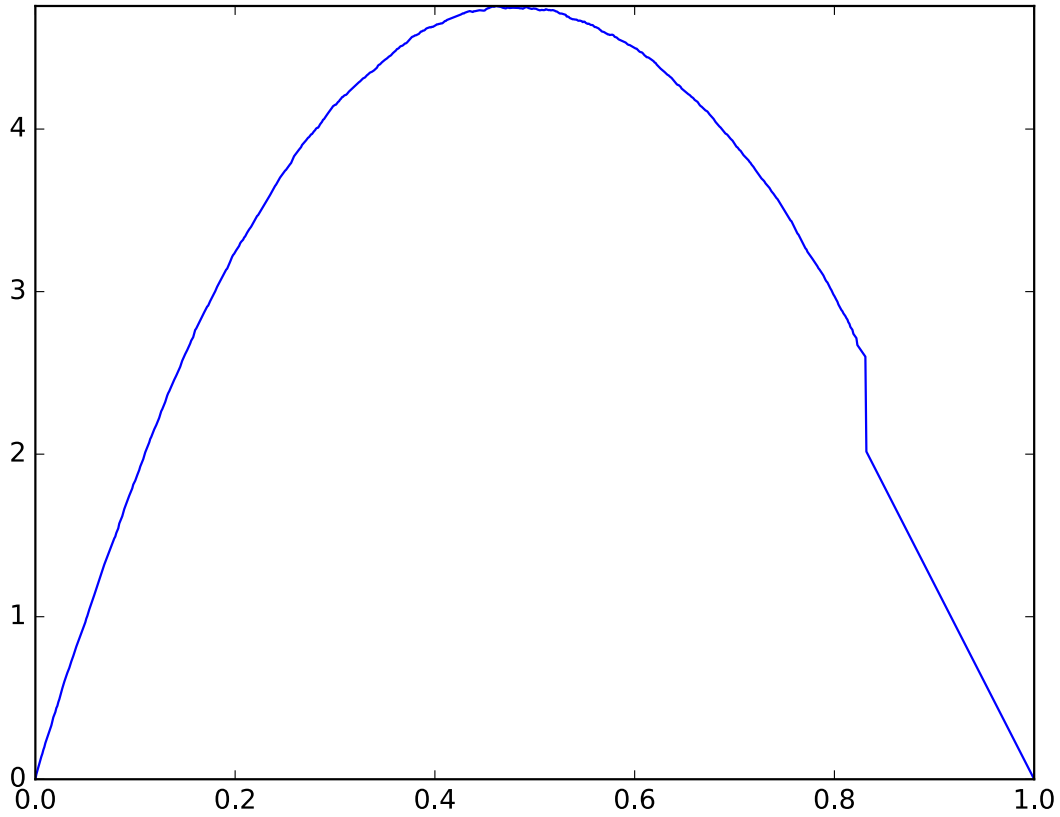
Results: Risk Function

Risk function, n = 40, Multiple Testing



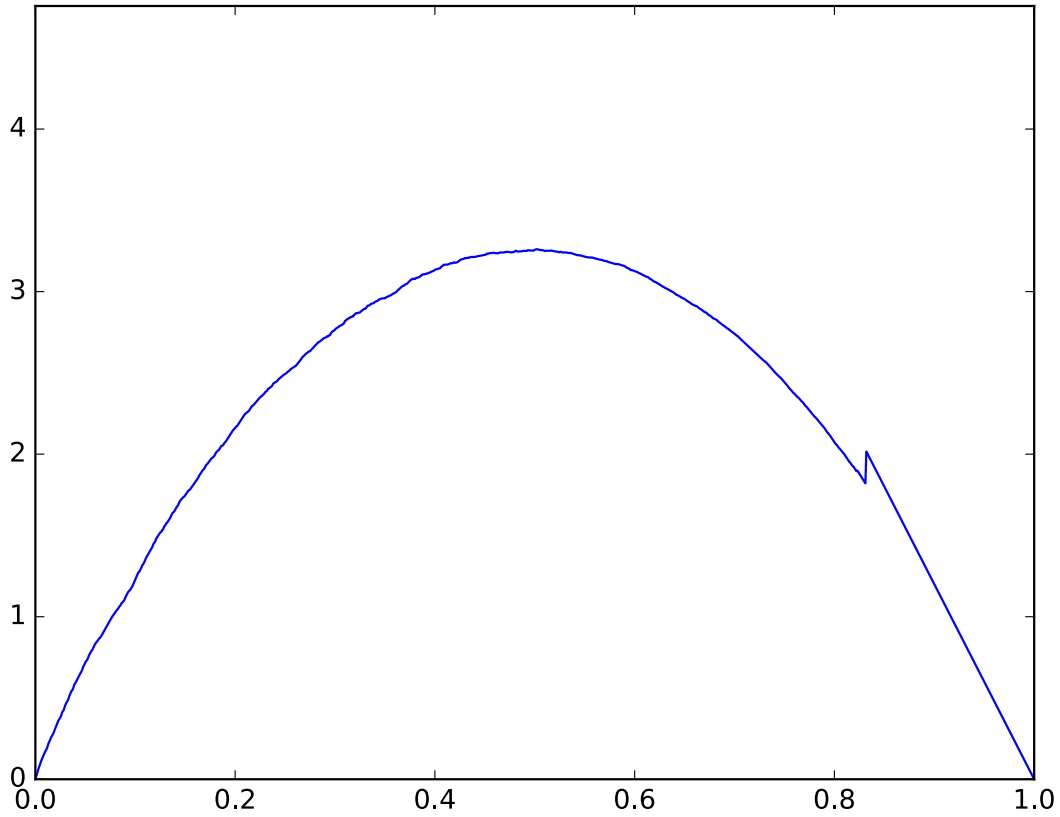
Results: Risk Function

Risk function, n = 10, Sidak



Results: Risk Function

Risk function, n = 40, Sidak



Future plans

- Mathematical modelling of optimisation procedures, especially with L1 regularization
- Experiments with higher dimensionality and different error measures
- Development of distribution free procedures, which expected error rates do not depend on the distribution of the data

NATIONAL RESEARCH
UNIVERSITY

Thank you
for your attention!