IPF-LASSO: integrative L_1-penalized regression with penalty factors for prediction based on multi-omics data

Anne-Laure Boulesteix
joint with Riccardo De Bin, Xiaoyu Jiang, Mathias Fuchs,
Simon Klau, Tobias Herold, Vindi Jurinovic

Institute for Medical Informatics, Biometry and Epidemiology
Ludwig-Maximilians-University Munich

Toulouse, June 24th 2016

Prediction models with multi-omic data

Y	$X_1^{(1)}$	...	$X_{p_1}^{(1)}$	$X_1^{(2)}$	...	$X_{p_2}^{(2)}$	...	$X_1^{(M)}$	...	$X_{p_M}^{(M)}$
0	...	...	...	...	...	...	...	...	...	...
0	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...
1	...	...	...	...	...	...	...	...	...	...
1	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...

- Response variable Y: e.g., responder status, survival time
- $X_1^{(m)}, \ldots, X_{p_m}^{(m)}$ form the mth group of clinical or "omics" variables, termed "modality"

Goal: Constructing (and evaluating) a prediction model for Y based on $X_1^{(1)}, \ldots, X_{p_1}^{(1)}, \ldots, X_1^{(M)}, \ldots, X_{p_M}^{(M)}$

Examples

clinical	low-dim
transcriptomic	high-dim
miRNA	high-dim
methylation	high-dim
SNP	high-dim
copy number variation	high-dim
metabolomic	high-dim
proteomic	high-dim

The "naive" strategy

- Ignore the modality structure, i.e. treat all variables $X_1^{(1)}, \ldots, X_{p_1}^{(1)}, \ldots, X_1^{(M)}, \ldots, X_{p_M}^{(M)}$ equally.
- Apply a prediction method, for example fit a L_1-penalized regression model (lasso):

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} -\ell(\boldsymbol{\beta}) + \lambda \sum_{m=1}^{M} \sum_{j=1}^{p_m} |\beta_j^{(m)}|$$

where ℓ is the log-likelihood and λ a penalty parameter.

Separate models

- **Problem:** In most cases the modalities are not equally relevant to the prediction problem, and ideally one wants to take this information into account.
- A small relevant modality may "get lost" within the variables from a large irrelevant modality.
- **Separate models** for each modality which are ultimately combined into a single prediction rule are an answer to this problem (Zhao et al., Brief Bioinf 2014), but also sub-optimal.

Overview

- One low- and one high-dimensional modality

  De Bin et al. (Stat Med 2014)
- Several high-dimensional modalities

  Boulesteix et al. (TechRep 2015)
- Other topics
  - validation

    De Bin et al. (BMC Med Res Meth 2014)
  - stability

    De Bin et al. (Biometrics 2015)
  - benchmarking

    Boulesteix et al. (Am Stat 2015)

    Boulesteix et al. (PLOS Comp Biol 2015)

Special case: one low and one high-dimensional modality

clinical	**omics**
$X_1^{(1)}, \ldots, X_{p_1}^{(1)}$	$X_1^{(2)}, \ldots, X_{p_2}^{(2)}$
low-dim ($p_1 < n$)	high-dim ($p_2 \gg n$)
cheap	expensive
well-investigated	explorative
highly relevant	???

- Differences have to be taken into account.
- Naive strategy is inappropriate.

The "residual" strategy

- Fit a (linear, logistic, Cox) model of the form

$$Y \sim X_1^{(1)} + \cdots + X_{p_1}^{(1)}$$

- Fit an omics-based model to the residuals of this model using lasso regression (or boosting, etc), i.e. consider the linear predictor $\sum_{j=1}^{p_1} \hat{\beta}_j^{(1)} X_j^{(1)}$ as an offset when estimating $\beta_1^{(2)}, \ldots, \beta_{p_2}^{(2)}$.
- Interpretation: omics variables are only used as "complement" to the clinical variables.

The "favoring" strategy

- Penalize only $X_1^{(2)}, \ldots, X_{p_2}^{(2)}$ in lasso:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} -\ell(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p_2} |\beta_j^{(2)}|.$$

- Intermediate between naive and residual strategies

The "dimension reduction" strategy

- Summarize $X_1^{(2)}, \ldots, X_{p_2}^{(2)}$ in form of a score $Z^{(2)}$ (or several components).
- Fit a (linear, logistic, Cox) model of the form

$$Y \sim X_1^{(1)} + \cdots + X_{p_1}^{(1)} + Z^{(2)}$$

- **Problem:** $Z^{(2)}$ overfits the data:
  - split the training data into two training subsets
  - or use pre-validation (Tibshirani & Efron, SAGMB 2002; Matsui et al., Clin Canc Res 2012)

Statistics in Medicine

Special Issue Paper

Received 31 October 2013, Accepted 31 May 2014 Published online 9 July 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6246

Investigating the prediction ability of survival models based on both clinical and omics data: two case studies

Riccardo De Bin,[a*†] Willi Sauerbrei[b] and Anne-Laure Boulesteix[a]

In biomedical literature, numerous prediction models for clinical outcomes have been developed based either on clinical data or, more recently, on high-throughput molecular data (omics data). Prediction models based on both types of data, however, are less common, although some recent studies suggest that a suitable combination of clinical and molecular information may lead to models with better predictive abilities. This is probably due to the fact that it is not straightforward to combine data with different characteristics and dimensions (poorly characterized high-dimensional omics data, well-investigated low-dimensional clinical data). In this paper, we analyze two publicly available datasets related to breast cancer and neuroblastoma, respectively, in order to show some possible ways to combine clinical and omics data into a prediction model of time-to-event outcome. Different strategies and statistical methods are exploited. The results are compared and discussed according to different criteria, including the discriminative ability of the models, computed on a validation dataset. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: clinical information; combining clinical and omics data; high-dimensional data; prediction models; survival analysis

Lasso with different penalties (cooperation with Novartis Biomarkers)

Y	$X_1^{(1)}$	...	$X_{p_1}^{(1)}$	$X_1^{(2)}$	...	$X_{p_2}^{(2)}$	...	$X_1^{(M)}$	...	$X_{p_M}^{(M)}$
0	...	...	...	...	...	...	...	...	...	...
0	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...
1	...	...	...	...	...	...	...	...	...	...
1	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...

- Rationale: In practice different modalities often have different information content.
- New 'IPF-Lasso' method (integrative lasso with penalty factors) minimizes

$$-\ell(\boldsymbol{\beta}) + \sum_{m=1}^{M} \lambda_m \sum_{j=1}^{p_m} |\beta_j^{(m)}|$$

where λ_m is the modality-specific penalty parameter.

Lasso with different penalties (cooperation with Novartis)

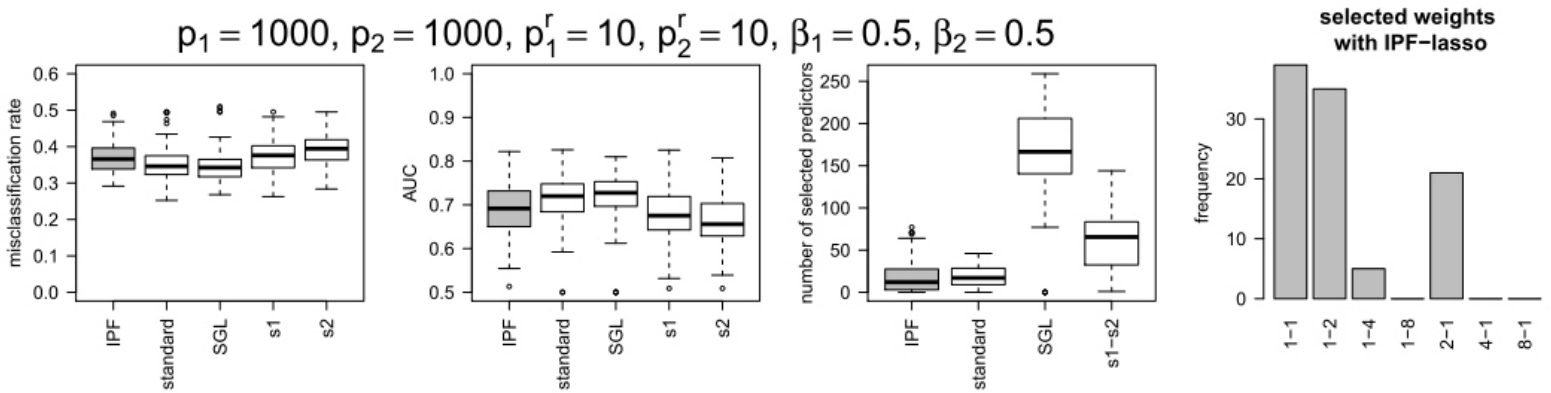
- **Estimation:** rescale the variables as

$$X_j^{*(m)} = X_j^{(m)} / \lambda_m$$

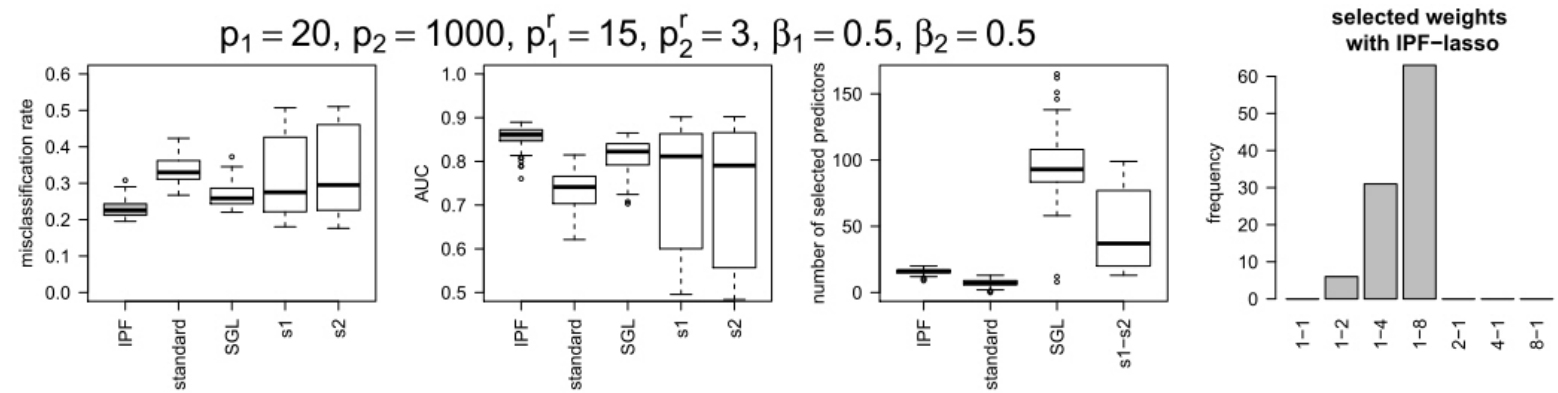
  and use standard estimation algorithm (e.g., 'glmnet').
- **Choice of λ_m, $m = 1, \ldots, M$:**
  - fully data-driven: cross-validation
  - taking other aspects into account (e.g., cost)
- **Implementation**: R package 'ipflasso' based on 'glmnet'

- New method performs worse than standard lasso if modalities are similar in terms of prediction accuracy and better otherwise:

setting A



setting F



IPF-LASSO's features

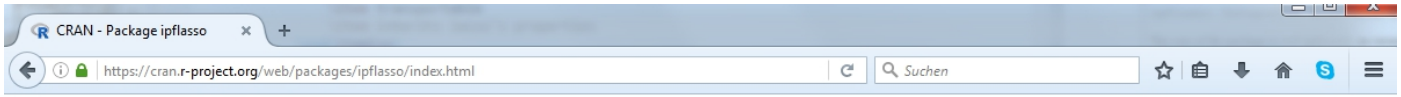
- sparse
- flexible
- fast
- transportable
- inherits lasso's properties

IPF-lasso: integrative L_1-penalized regression with penalty factors for prediction based on multi-omics data

Anne-Laure Boulesteix[1], Riccardo De Bin[1], Xiaoyu Jiang[2], Mathias Fuchs[1]

[1] Department of Medical Informatics, Biometry and Epidemiology, University of Munich (LMU), Marchioninistr. 15, D-81377 Munich, Germany. boulesteix@ibe.med.uni-muenchen.de

[2] Novartis Biomarker Development



`ipflasso: Integrative Lasso with Penalty Factors`

The core of the package is cvr2.ipflasso(), an extension of glmnet to be used when the (large) set of available predictors is partitioned into several modalities which potentially differ with respect to their information content in terms of prediction. For example, in biomedical applications patient outcome such as survival time or response to therapy may have to be predicted based on, say, mRNA data, miRNA data, methylation data, CNV data, clinical data, etc. The clinical predictors are on average often much more important for outcome prediction than the mRNA data. The ipflasso method takes this problem into account by using different penalty parameters for predictors from different modalities. The ratio between the different penalty parameters can be chosen by cross-validation.

Version:	0.1
Depends:	glmnet, survival
Published:	2015-11-24
Author:	Anne-Laure Boulesteix, Mathias Fuchs
Maintainer:	Anne-Laure Boulesteix <boulesteix at ibe.med.uni-muenchen.de>
BugReports:	NA
License:	GPL-2 | GPL-3 [expanded from: GPL]
URL:	NA
NeedsCompilation:	no
CRAN checks:	ipflasso results

`Downloads:`

Reference manual:	ipflasso.pdf
Package source:	ipflasso_0.1.tar.gz
Windows binaries:	r-devel: ipflasso_0.1.zip, r-release: ipflasso_0.1.zip, r-oldrel: ipflasso_0.1.zip
OS X Mavericks binaries:	r-release: ipflasso_0.1.tgz, r-oldrel: ipflasso_0.1.tgz

---

`cvr2.ipflasso` *Cross-validated integrative lasso with cross-validated penalty factors*

Description

Runs cvr.glmnet giving different penalty factors to predictors from different blocks and chooses the penalty factors by cross-validation from the list `pflist` of candidates.

Usage

```
cvr2.ipflasso(X, Y, family, type.measure, standardize=TRUE,
              alpha=1, blocks, pflist, nfolds, ncv,
              nzeromax = +Inf, plot=FALSE)
```

Limitations of IPF-LASSO

- CV is computationally expensive when M is large!
- tends to select variables from many/all modalities

Alternative strategy: adopting the offset strategy in multi-modality settings (master's thesis Simon Klau, co-supervised by Tobias Herold and Vindi Jurinovic)

Offset strategy in multi-modality settings

Recall: offset strategy

- Fit a (linear, logistic, Cox) model of the form

$$Y \sim X_1^{(1)} + \cdots + X_{p_1}^{(1)}$$

- Fit an omics-based model to the residuals of this model using lasso regression (or boosting, etc), i.e. consider the linear predictor $\sum_{j=1}^{p_1} \hat{\beta}_j^{(1)} X_j^{(1)}$ as an offset when estimating $\beta_1^{(2)}, \ldots, \beta_{p_2}^{(2)}$.

Extension:

- consider the linear predictor $\sum_{\ell=1}^{m-1} \sum_{j=1}^{p_\ell} \hat{\beta}_j^{(\ell)} X_j^{(\ell)}$ as an offset when estimating $\beta_1^{(m)}, \ldots, \beta_{p_m}^{(m)}$.

What is validation of added predictive value?

- accuracy of combined model

  $\rightarrow$ apply model to independent test data and compute accuracy
- accuracy improvement of combined vs. clinical model

  $\rightarrow$ apply both models to independent test data and compute/compare accuracies
- effect of an omics score

  $\rightarrow$ fit model to clinical variables and omics score using independent data and test coefficient of omics score
- importance of an omics score for prediction

  $\rightarrow$ estimate accuracy of two models using cross-validation within independent test dataset: (i) clinical variables, (ii) clinical variables + score.

De Bin *et al. BMC Medical Research Methodology* 2014, **14**:117
http://www.biomedcentral.com/1471-2288/14/117

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Added predictive value of omics data: specific issues related to validation illustrated by two case studies

Riccardo De Bin[1*], Tobias Herold[2,3] and Anne-Laure Boulesteix[1]

Abstract

Background: In the last years, the importance of independent validation of the prediction ability of a new gene signature has been largely recognized. Recently, with the development of gene signatures which integrate rather than replace the clinical predictors in the prediction rule, the focus has been moved to the validation of the added predictive value of a gene signature, i.e. to the verification that the inclusion of the new gene signature in a prediction model is able to improve its prediction ability.

Methods: The high-dimensional nature of the data from which a new signature is derived raises challenging issues and necessitates the modification of classical methods to adapt them to this framework. Here we show how to validate the added predictive value of a signature derived from high-dimensional data and critically discuss the impact of the choice of methods on the results.

Results: The analysis of the added predictive value of two gene signatures developed in two recent studies on the survival of leukemia patients allows us to illustrate and empirically compare different validation techniques in the high-dimensional framework.

Conclusions: The issues related to the high-dimensional nature of the omics predictors space affect the validation process. An analysis procedure based on repeated cross-validation is suggested.

Keywords: Added predictive value, Omics score, Prediction model, Time-to-event data, Validation

Stability investigations using bootstrap samples

- Variable selection is unstable. The selected model may change a lot when the data change a little.
- **Common approach:** Repeat variable selection on bootstrap samples:

→ variable inclusion frequencies

→ model frequencies

Table 1: Glioma data: selection frequencies of the 10 top ranked models for bootstrap(n), bootstrap(m) and subsample(m), based on 10,000 pseudo-samples for $\alpha = 0.05$ and presented in decreasing sum of the three selection frequencies.

	bootstrap(n)		bootstrap(m)		subsample(m)	
model	rank	freq.	rank	freq.	rank	freq.
basic+kard1	2	124	1	326	1	1615
basic+kard1+epi	8	93	7	128	2	417
basic+kard1+surgd2	6	103	3	163	4	352
basic+kard1+sex	3	108	2	187	6	290
basic	140	15	8	123	3	398
basic+kard1+cort	5	106	4	148	5	298
basic+kard1+sex+epi	1	156	6	140	9	225
basic+cort+ops	22	62	4	148	7	264
basic+epi	54	33	12	104	8	242
basic+ops	101	20	9	121	12	189
basic*	717	2	10	117	10	205
basic+kard1+cort+ops	7	97	15	93	15	134
basic+gradd2+kard1+cort	8	93	43	40	23	84
basic+gradd2+kard1+cort+ops	3	108	55	33	55	35
basic+kard1+surgd2+sex+epi	10	89	52	35	67	27

basic=intercept+gradd1+age+surgd1; basic*=intercept+gradd2+age+surgd1

Motivation of our project: Bootstrap has problems, subsamples may be more appropriate.

Problem: inflated type-1 error for tests performed on bootstrap samples

- Z-test: $Z = \sqrt{n}(\overline{x} - \mu_0)/\sigma \sim \mathcal{N}(0,1)$ under $H_0 : \ \mu = \mu_0$.
- For $Z^* = \sqrt{n}(\overline{x}^* - \mu_0)/\sigma$ computed **from a bootstrap sample**, we have under H_0

$$E(Z^*) \ = \ E(E(Z^*|\hat{F})) \ = \ E(Z) \ = \ 0$$

and

$$\begin{aligned} V(Z^*) &= V(E(Z^*|\hat{F})) + E(V(Z^*|\hat{F})) \\ &= V(Z|\hat{F}) + E(V(Z)) \\ &= 2 \end{aligned}$$

Janitza et al., Biometrical Journal 2016.

Impact on bootstrap-based variable selection

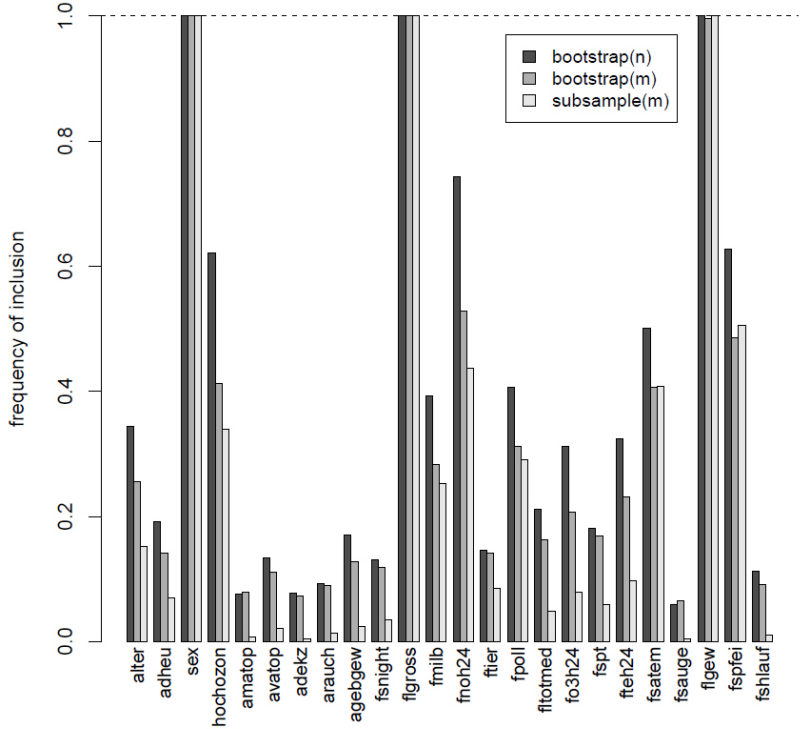


Figure 2: Ozone data: inclusion frequencies, based on 10,000 pseudo-samples, for all the 24 available variables. The results refer to the case $\alpha = 0.05$.

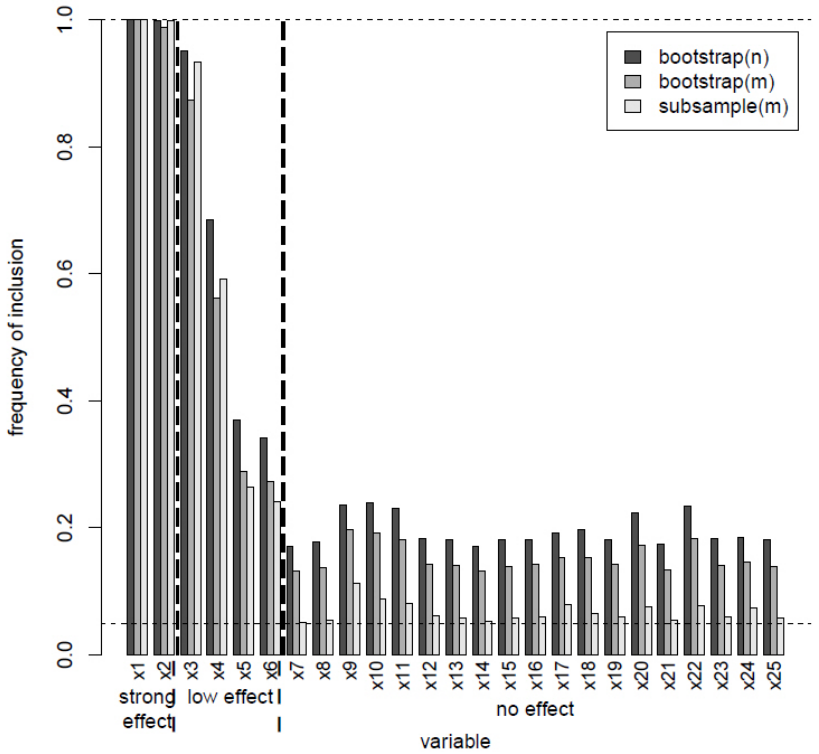
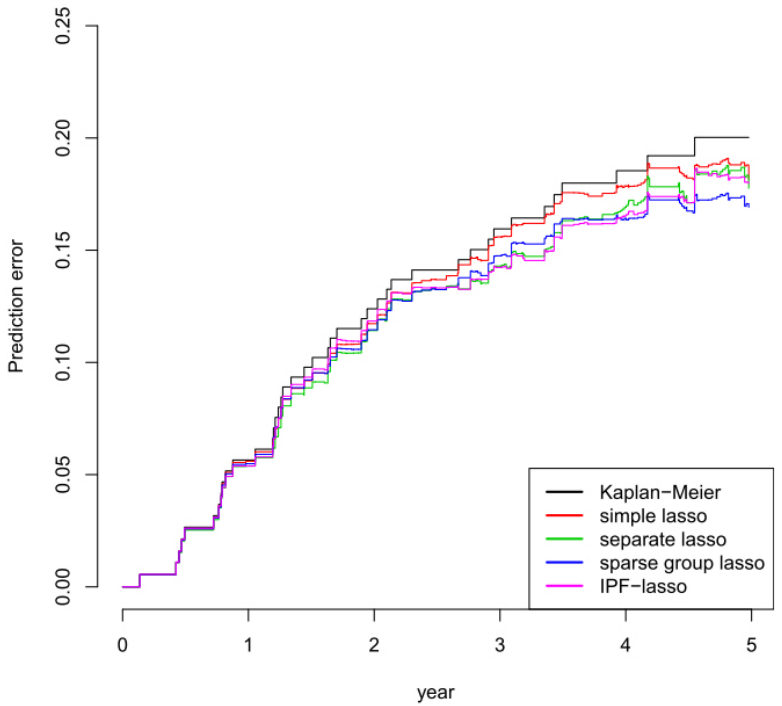


Figure 3: Simulated data: inclusion frequencies of the variables based on 1,000,000 pseudo-samples, 1,000 for each dataset.

De Bin et al., Biometrics 2016.

Benchmarking example: the real data study on IPF-lasso



- We applied IPF-lasso to three datasets (leukemia, breast cancer).
- For two of them IPF-lasso performed better, for one of them worse than competitor SGL.
- What to conclude?
- What to do? Report only the two good results? Or risk to get rejected?

Performed better on real data?

Typical sentence in abstracts of computational science articles:

"Our method performed better than existing methods on real data"

- Compute CV error of K methods for J data sets ($J \approx 2 - 10$)
- In machine learning: test difference in error rates using paired t-test or Wilcoxon signed rank test

Which null hypothesis is being tested?

$$H_0: \quad \boldsymbol{E}_{P^n}(\varepsilon(\hat{f}_{k_1}^S)) \;=\; \boldsymbol{E}_{P^n}(\varepsilon(\hat{f}_{k_2}^S)) \;?$$

No, since data sets are drawn from different P's $P_1, \ldots, P_J$!

In R...

- interface to openML (www.openml.org)
- packages 'mlr', package 'CMA'
- ...

What is being tested?

- Distribution P is now considered as the outcome of a random variable Φ, and size of data set n as the outcome of a random variable N.
- Then the hypothesis that is implicitly being tested when comparing methods k_1 and k_2 can be written as

$$\boldsymbol{E}(\varepsilon(k_1, \Phi, N)) = \boldsymbol{E}(\varepsilon(k_2, \Phi, N)),$$

where $\boldsymbol{E}$ denotes the expectation over the random variables Φ and N.

Boulesteix et al., The American Statistician 2015.

Test statistic and power considerations

- Test statistic (paired t-test):

$$T = \frac{\overline{\Delta e}}{\sqrt{\frac{1}{N}\frac{1}{N-1}\sum(\Delta e(D_j) - \overline{\Delta e})^2}},$$

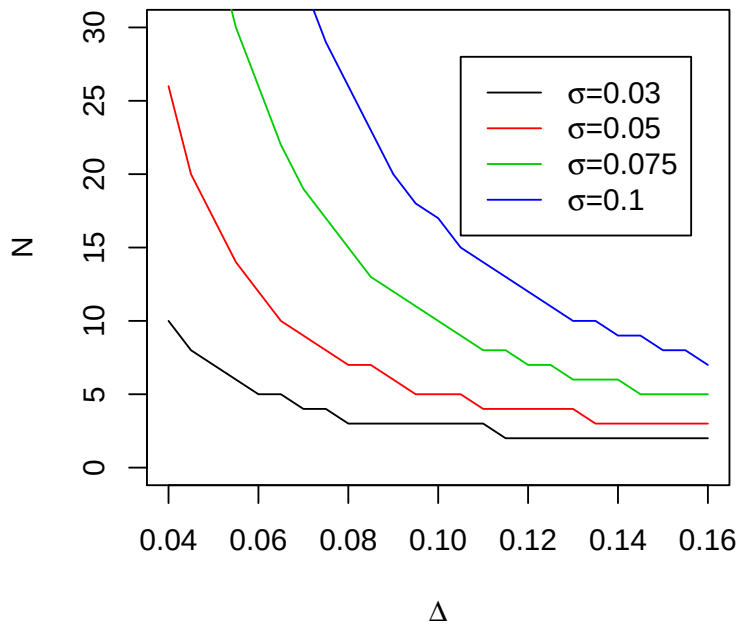
where $\Delta e(D_j)$ is the difference between estimated errors of methods k_2 and k_1 in data set D_j and $\overline{\Delta e}$ is the mean over data sets.

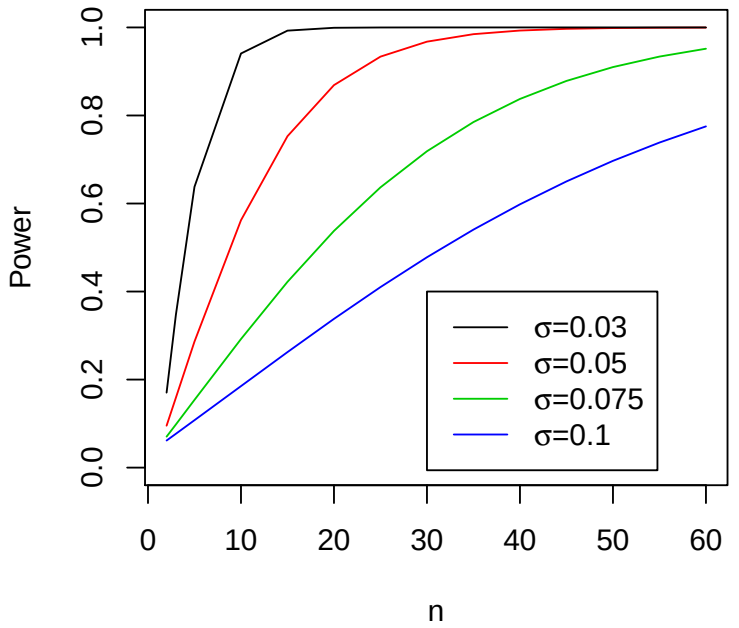
- Power calculation for "sample size" N (number of data sets)

Boulesteix et al., The American Statistician 2015.

Number of data sets and power

$$N \approx \frac{(z_{1-\beta} + z_{1-\alpha})^2}{\Delta^2/\sigma^2}$$





PLOS COMPUTATIONAL BIOLOGY

EDITORIAL

Ten Simple Rules for Reducing Overoptimistic Reporting in Methodological Computational Research

Anne-Laure Boulesteix*

Institute for Medical Informatics, Biometry and Epidemiology, Ludwig Maximilians University, Munich, Germany

- Rule 1: Assess the New Method
- Rule 2: Compare the New Method to the Best
- Rule 3: **Consider Enough Datasets**
- Rule 4: **Do Not "Fish" for Datasets**
- Rule 5: **Think of the No-Free-Lunch Theorem and Report Limitations**
- Rule 6: **Consider Several Criteria**
- Rule 7: Validate Using Independent Data
- Rule 8: Design Simulations Appropriately
- Rule 9: Provide All Information
- Rule 10: Read the Other Ten Simple Rules Articles

Parallel to computational/clinical research
Plea for benchmarking

- **Making the world better**
  - **Clin:** new interventions that improve health outcomes
  - **Comp:** new methods that make results of statistical analyses closer to the truth
- **Comparison studies**
  - **Clin:** validation studies, phase III, phase IV, meta-analyses
  - **Comp:** well-conducted benchmark studies

Would we take medicines evaluated in underpowered phase I studies conducted by a single team?

Boulesteix et al., 2013. PLOS ONE.

Boulesteix, 2013. Bioinformatics.

Publication Bias in Methodological Computational Research

Anne-Laure Boulesteix[1], Veronika Stierle[1] and Alexander Hapfelmeier[2]

[1]Department of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilian University, Munich, Germany. [2]Department of Medical Statistics and Epidemiology, Klinikum rechts der Isar Technical University of Munich, Munich, Germany.

Supplementary Issue: Statistical Systems Theory in Cancer Modeling, Diagnosis, and Therapy

ABSTRACT: The problem of publication bias has long been discussed in research fields such as medicine. There is a consensus that publication bias is a reality and that solutions should be found to reduce it. In methodological computational research, including cancer informatics, publication bias may also be at work. The publication of negative research findings is certainly also a relevant issue, but has attracted very little attention to date. The present paper aims at providing a new formal framework to describe the notion of publication bias in the context of methodological computational research, facilitate and stimulate discussions on this topic, and increase awareness in the scientific community. We report an exemplary pilot study that aims at gaining experiences with the collection and analysis of information on unpublished research efforts with respect to publication bias, and we outline the encountered problems. Based on these experiences, we try to formalize the notion of publication bias.

KEYWORDS: epistemology, publication practice, false research findings, overoptimism

Thank you for your attention!

Thanks to:

- Colleagues: R. De Bin, M. Eugster, M. Fuchs, T. Herold, S. Janitza, X. Jiang, V. Jurinovic, S. Klau, S. Lauer, W. Sauerbrei,
- German Research Foundation (DFG), Novartis Biomarkers