Pre-validation for assessing the added predictive value of high-dimensional molecular data in binary classification

Eva Endres

Department of Statistics,
Ludwig Maximilian's University Munich

June 26, 2014

Outline

Background

Pre-validation

Assessment of the added predictive value

Practical application

Results

Summary

Outline

Background

Pre-validation

Assessment of the added predictive value

Practical application

Results

Summary

Biological foundations

- Human genome is estimated to consist of about 20,500 genes
- Genes are sections of the DNA which in turn forms the 46 human chromosomes
- Genes control the production of amino acids/proteins
- Gene expression determines the phenotype
  → Structurally/functionally heterogeneous cells
- Measurement of gene expression with the aid of microarray technology
  → Indication about presence or future development of diseases ($\hat{=}$ phenotype)

Statistical background

- Emphasis on **binary classification**, e.g. prognosis/diagnosis in cancer research
- Goal: Creation of a function that assigns a class to each new observation
- Logistic regression model: Estimation of the (conditional) probability

$$P(y_i = 1|\mathbf{z}_i) = \frac{\exp(\gamma_0 + \gamma_1 \cdot z_{i1} + \gamma_2 \cdot z_{i2} + \ldots + \gamma_q \cdot z_{iq})}{1 + \exp(\gamma_0 + \gamma_1 \cdot z_{i1} + \gamma_2 \cdot z_{i2} + \ldots + \gamma_q \cdot z_{iq})}$$

- **Linear predictor** may include **clinical and molecular information**
  - → Combination of predictors with different dimensionalities
  - → High-dimensionality of the molecular predictors

Statistical background

- **High-dimensionality** of the molecular predictors
  - Variable selection, dimension reduction, regularization techniques
  - Here: Least absolute shrinkage and selection operator and supervised principal component analysis
- **Combination** of clinical and molecular predictors:
  - Aggregation of the molecular predictors to one new component, the (linear) omics score

$$x_{score,i} = w_1 \cdot x_{i1} + w_2 \cdot x_{i2} + \ldots + w_p \cdot x_{ip}$$

  - Omics score is considered as new predictor

$$\eta_i = \underbrace{\gamma_0 + \gamma_1 \cdot z_{i1} + \gamma_2 \cdot z_{i2} + \ldots + \gamma_q \cdot z_{iq}}_{\text{clinical model}} + \beta_{score} \cdot x_{score,i}$$

⇒ Does the inclusion of the omics score in the prediction model improve its predictive ability?

Statistical background

- Question concerning the **added predictive value** of the omics score compared to well-established clinical predictors
- Validation of the added predictive value usually needs independent validation data
- What if there is **no validation data available**?
- Assessment of the added predictive value on the same data set that was used to derive the score
  - → Omics score overfits the data at hand
  - → Strongly biased results in favor of the omics score i.e., the score might seem more important than it actually is
  - ⇒ **Pre-validation**:
    Embedding score generation into a pre-validation loop ensures a fair comparison of the different predictors

Outline

Background

Pre-validation

Assessment of the added predictive value

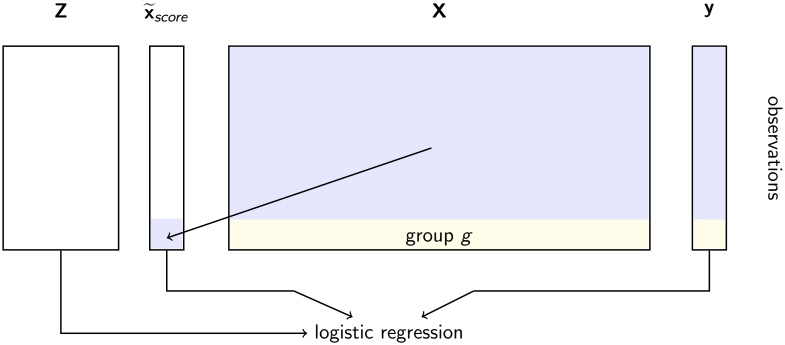
Practical application

Results

Summary

Pre-validation

Fundamental idea



$$\eta_i = \gamma_0 + \gamma_1 \cdot z_{i1} + \ldots + \gamma_q \cdot z_{iq} + \widetilde{\beta}_{score} \cdot \widetilde{x}_{score,i}$$

Pre-validation

Algorithm

1. Divide the present observations into G approximately equal-sized groups.
2. Set group g aside.
   Use the gene expression levels of the remaining observations to obtain a rule f for generating the molecular score.
3. Apply this rule on the left-out observations of group g which yields the pre-validated molecular score.

$$\widetilde{\mathbf{x}}_{score}^{[o(g)]} = \hat{f}_{\mathbf{X}^{[-o(g)]},\mathbf{y}^{[-o(g)]}}(\mathbf{X}^{[o(g)]})$$

4. Repeat steps 2-3 for each group $g = 1, \ldots, G$.

Pre-validation

Least absolute shrinkage and selection operator

$$\hat{\boldsymbol{\beta}}_{Lasso} = \arg\min_{\boldsymbol{\beta}} \left\{ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \cdot ||\boldsymbol{\beta}||_1 \right\}$$

- Shrinks some coefficients, sets others to zero
- Good prediction accuracy and good interpretability of the regression results
- Handles the high-dimensionality of the molecular data
- Computational feasible
- *Lasso*-score:

$$x_{score,i} = \hat{\beta}_{Lasso,1} \cdot x_{i1} + \ldots + \hat{\beta}_{Lasso,p} \cdot x_{ip}$$

Pre-validation

Least absolute shrinkage and selection operator

1. Divide the available observations into G approximately equal-sized groups.
2. Leave group g out and perform a *Lasso*-regression on the remaining observations to derive the vector $\hat{\boldsymbol{\beta}}_{Lasso}^{[-o(g)]}$ including the regression coefficients of each molecular predictor.
3. Compute the pre-validated molecular score for person $i \in o(g)$ as weighted sum over all molecular predictors

$$\widetilde{x}_{score,i}^{[o(g)]} = \hat{\beta}_{Lasso,1}^{[-o(g)]} \cdot x_{i1}^{[o(g)]} + \ldots + \hat{\beta}_{Lasso,p}^{[-o(g)]} \cdot x_{ip}^{[o(g)]}.$$

4. Repeat steps 2-3 for every group $g = 1, \ldots, G$.

Pre-validation

Supervised principal component analysis

- Revelation of the latent structure of the data set ,i.e. groups of genes with similar expression profiles
- Uncorrelated linear combinations of the original predictors capture the largest proportion of variance

  → Dimension reduction with slightly loss of information
- Principal components are not necessarily related to the outcome
- **Supervised** principal component analysis
  - Use only molecular predictors which are related to the outcome for the principal component analysis
  - Perform an univariate variable selection (here: Wald test) and use only the first k gene expressions of the toplist → $\mathbf{X} \in \mathbb{R}^{n \times k}$

Pre-validation

Supervised principal component analysis

1. Divide the available observations into G approximately equal-sized groups.
2. Leave group g out and
   - 2.1 perform an univariate variable selection on the remaining observations to obtain a toplist of the molecular predictors;
   - 2.2 perform a principal component analysis on the basis of the first $k = 25$ predictors from the toplist;
   - 2.3 use the first m principal components as independent covariates in a multivariate logisitic regression model to estimate the vector $\hat{\boldsymbol{\beta}}_{superPC}^{[-o(g)]} \left(\in \mathbb{R}^{m \times 1}\right)$ of regression coefficients.
3. Compute the pre-validated molecular score for person $i \in o(g)$ as weighted sum over the first m principal components

$$\widetilde{x}_{score,i}^{[o(g)]} = \hat{\beta}_{superPC,1}^{[-o(g)]} \cdot \phi_{i1}^{[o(g)]} + \ldots + \hat{\beta}_{superPC,m}^{[-o(g)]} \cdot \phi_{im}^{[o(g)]}$$

4. Repeat steps 2-3 for every group $g = 1, \ldots, G$.

Outline

- Background
- Pre-validation
- Assessment of the added predictive value
- Practical application
- Results
- Summary

Testing in multivariate regression model

- Multivariate logistic regression model

$$P(y_i = 1|\mathbf{z}_i, \mathbf{x}_i) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}, \text{ where}$$

$$\eta_i = \gamma_0 + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \ldots + \gamma_q z_{iq} + \beta_{score} x_{score,i}$$

  → Test the hypotheses $H_0 : \beta_{score} = 0$ vs. $H_1 : \beta_{score} \neq 0$
  → p-value $< \alpha \Rightarrow$ Omics data provides added predictive value
- Comparison of the omics scores derived with and without pre-validation
  → **Expectation**: $\beta_{score} > \widetilde{\beta}_{score}$ and $p < \widetilde{p}$ if the test is performed on the same data set that was used to build the score
- Disadvantage: p-value gives no indication about the predictive ability of a model

Evaluation of the prediction accuracy

- Discriminative ability determined via the area under the receiver operating characteristic curve
- Comparison of the prediction accuracy of the clinical and the combined prediction model
  - → $AUC_{clinical} < AUC_{combined} \Rightarrow$ Omics data provides added predictive value
- Comparison of the omics scores derived with and without pre-validation
  - → **Expectation**: $AUC_{combined} > \widetilde{AUC}_{combined}$ if the AUC is computed on the same data set that was used to build the score

Outline

Background

Pre-validation

Assessment of the added predictive value

Practical application

Results

Summary

Data simulation

- Simulation of $n = 200$ observations of $q = 10$ clinical and $p = 1000$ molecular predictors, where $(\mathbf{Z}, \mathbf{X}) \sim \text{MVN}(\mathbf{0}, \mathbf{R})$
- $\gamma = (-2, -1.5, -1, 1, 1.5, 2, 0, 0, 0, 0)^\top$ and $\boldsymbol{\beta} = (\underbrace{0.75, \ldots, 0.75}_{\text{1-20}}, \underbrace{0, \ldots, 0}_{\text{21-1000}})^\top$
- Response is a Bernoulli random variable, where

$$\mathrm{P}(\mathrm{y}_i = 1 | \mathbf{x}_i, \mathbf{z}_i) = \frac{\exp(\gamma_0 + \gamma_1 \cdot z_{i1} + \ldots + \gamma_q \cdot z_{iq} + \beta_1 \cdot x_{i1} + \ldots + \beta_p \cdot x_{ip})}{1 + \exp(\gamma_0 + \gamma_1 \cdot z_{i1} + \ldots + \gamma_q \cdot z_{iq} + \beta_1 \cdot x_{i1} + \ldots + \beta_p \cdot x_{ip})}$$

- Four settings:

		predictive ability of clinical data	
		high	low
predictive ability of molecular data	high	setting 1	setting 3
	low	setting 2	setting 4
	no	setting 5	setting 6

Hatzis' breast cancer data

- Prospective multicenter study conducted from June 2000 to March 2010 at the M. D. Anderson Cancer Center in Houston, Texas
- 310 patients with newly diagnosed ERBB2 - negative breast cancer treated with chemotherapy
- $$y = \begin{cases} 0 & \text{chemosensitivity} \qquad \text{(no or minimal residual disease)} \\ 1 & \text{chemoresistance} \quad \text{(moderate or extensive residual disease)} \end{cases}$$

  after neoadjuvant chemotherapy
- Clinical predictors: Age, progesterone receptor status, estrogen receptor status, tumor stage, nodal status and tumor grade
- 22,383 molecular predictors measured with the aid of gene expression microarrays from Affymetrix

Outline

Background

Pre-validation

Assessment of the added predictive value

Practical application

Results

Summary

Results

Simulation setting 1

			Without pre-validation	**5-fold pre-validation**
Lasso		β_{score}	2.2627	1.5183
		p_{score}	0.0001	0.0005
		AUC	0.9407	0.9059
superPC	without adjustment	β_{score}	1.2262	0.6225
		p_{score}	$1.73{\cdot}10^{-8}$	0.0007
		AUC	0.9712	0.9113
	with adjustment	β_{score}	1.6216	0.6703
		p_{score}	$7.87{\cdot}10^{-7}$	0.0057
		AUC	0.9817	0.9041

$AUC_{clinical} = 0.8548$

Results

Simulation setting 2

			Without pre-validation	5-fold pre-validation
Lasso		β_{score}	5.2785	0.4013
		p_{score}	0.2489	0.4522
		AUC	0.9923	0.9915
superPC	without adjustment	β_{score}	1.6925	-0.0747
		p_{score}	0.0246	0.4803
		AUC	0.9958	0.9914
	with adjustment	β_{score}	3.1455	-0.1375
		p_{score}	0.0102	0.4479
		AUC	0.9997	0.9914

$AUC_{clinical} = 0.9909$

Results

Simulation setting 3

			Without	**5-fold**
			pre-validation	**pre-validation**
Lasso		β_{score}	2.3631	1.4250
		p_{score}	$4.12 \cdot 10^{-5}$	0.0026
		AUC	0.9410	0.9018
superPC	without adjustment	β_{score}	1.2415	0.6232
		p_{score}	$2.27 \cdot 10^{-8}$	0.0001
		AUC	0.9705	0.9084
	with adjustment	β_{score}	1.688	0.7708
		p_{score}	$1.43 \cdot 10^{-6}$	0.0042
		AUC	0.9809	0.9097

$AUC_{clinical} = 0.84378$

Results

Simulation setting 4

			Without pre-validation	5-fold pre-validation
Lasso		β_{score}	7.9535	4.3370
		p_{score}	0.0023	0.0379
		AUC	0.9836	0.9782
superPC	without adjustment	β_{score}	1.0901	0.5738
		p_{score}	0.0003	0.0108
		AUC	0.9929	0.9819
	with adjustment	β_{score}	1.4043	0.3803
		p_{score}	0.0004	0.1668
		AUC	0.9980	0.9765

$AUC_{clinical} = 0.9704$

Results

Simulation setting 5

			Without pre-validation	**5-fold pre-validation**
Lasso		β_{score}	0.0069	-0.0013
		p_{score}	0.3545	0.4489
		AUC	0.9547	0.9541
superPC	without adjustment	β_{score}	1.0068	-0.0198
		p_{score}	$4.23{\cdot}10^{-5}$	0.4648
		AUC	0.9795	0.9538
	with adjustment	β_{score}	3.7685	-0.0428
		p_{score}	$1.38{\cdot}10^{-4}$	0.4399
		AUC	0.9949	0.9539

$AUC_{clinical} = 0.9526$

Results

Simulation setting 6

			Without	**5-fold**
			pre-validation	**pre-validation**
Lasso		β_{score}	-0.1055	-0.0158
		p_{score}	0.3240	0.4140
		AUC	0.9621	0.9606
superPC	without adjustment	β_{score}	1.0069	-0.0616
		p_{score}	0.0004	0.4304
		AUC	0.9825	0.9605
	with adjustment	β_{score}	4.2160	-0.1443
		p_{score}	0.0002	0.3958
		AUC	0.9962	0.9608

$AUC_{clinical} = 0.9591$

Results

Hatzis' breast cancer data

			Without pre-validation	5-fold pre-validation
Lasso		β_{score}	0.3572	0.0403
		p_{score}	0.0988	0.3482
		AUC	0.7803	0.7749
superPC	without adjustment	β_{score}	1.1229	0.4468
		p_{score}	$2.43 \cdot 10^{-7}$	0.0120
		AUC	0.8408	0.7858
	with adjustment	β_{score}	1.0223	0.0956
		p_{score}	$4.68 \cdot 10^{-11}$	0.3487
		AUC	0.8887	0.7739

$AUC_{clinical} = 0.7718$

Outline

Background

Pre-validation

Assessment of the added predictive value

Practical application

Results

Summary

Summary

- Main tasks:
  - **Investigation** and **comparison** of the added predictive value of omics scores derived with and without pre-validation
    - → Pre-validation generally seems to reduce overfitting
    - → Strengthening of the clinical predictors cannot be confirmed
    - → None of the pre-validated scores shows significance if molecular data has no predictive ability
  - **Simulation** studies and analysis of real breast cancer data
  - **Implementation** of all applied methods in R
- Perspective:
  - → Modifications of the simulation
  - → Methods for binary classification
  - → Methods for score generation
  - → Implementation of the permutation test for pre-validation

Bair, E. and Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene expression data, *Public Library of Science Biology* **2**: 511–522.

Boulesteix, A.-L. and Sauerbrei, W. (2011). Added predictive value of high-throughput molecular data to clinical data and its validation, *Briefings in Informatics* **12**: 215–229.

De Bin, R., Herold, T. and Boulesteix, A.-L. (2014). Added predictive value of omics data: Specific issues related to validation illustrated by two case studies, *Technical Report 154*, Department of Statistics, University of Munich.

De Bin, R., Sauerbrei, W. and Boulesteix, A.-L. (2014). Investingating the prediction ability of survival models based on both clinical and omics data: Two case studies, *Technical Report 153*, Department of Statistics, University of Munich.

Hatzis, C. and et al. (2011). A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer, *The Journal of the American Medical Association* **305**: 1873–1881.

Tibshirani, R. (1996). Regression shrinkage and selection vias the lasso, *Journal of the Royal Statistical Society, Series B* **58**: 267–288.

Tibshirani, R. and Efron, B. (2002). Pre-validation and inference in microarrays, *Statistical Applications in Genetics and Molecular Biology* **1**: 1–18.