Full versus incomplete cross-validation: measuring the impact of imperfect separation between training and test sets in prediction error estimation

Roman Hornung

Joint work with Christoph Bernau, Caroline Truntzer, Thomas Stadler and Anne-Laure Boulesteix

LMU Munich
Department of Medical Informatics, Biometry and Epidemiology

August, 28th, 2014

Introduction

Addon procedures

Full versus incomplete CV

New measure CVIIM

Illustration

Summary & Conclusion

1 Introduction

2 Addon procedures

3 Full versus incomplete CV

4 New measure CVIIM

5 Illustration

6 Summary & Conclusion

Introduction

- Modern technologies, most prominently microarrays, enable the measurement of the expression of **thousands or many thousands of genes for each unit of investigation**.
- Data sets are generated in which such measurements are agglomerated for units (patients, tissues,...) affected by a **disease of interest** and for unaffected controls.

Introduction

- This data contains **empirical information** on (previously) unknown systematic differences between the two groups.
- By the aid of classification methods we can thus empirically construct **prediction rules** for the purpose of predicting the status (diseased or healthy) of new units.
- Due to limited sample sizes and information contained in gene expression such prediction rules make errors.

Introduction

- Our general interest lies in a correct **estimation of the expected error frequency**.
- ↯ Procedures which tend to result in too small estimated error frequencies can result in overoptimistic ($\Rightarrow$ dangerous!) conclusions regarding the performance of a prediction rule.

Prediction rule in general

Full versus incomplete cross-validation

Roman Hornung et al.

Introduction

Addon procedures

Full versus incomplete CV

New measure CVIIM

Illustration

Summary & Conclusion

Sample: $\boldsymbol{S} = \{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\} \sim P^n$

$\boldsymbol{X}_i$: ($LONG$) vector of (raw!) gene expressions ("**covariates**")

Y_i : status of patient ("**class**") ($i = 1, \ldots, n$)

Prediction rule fitted on "**training data**" $\boldsymbol{S}$:

$$\hat{g}_{\boldsymbol{S}} : \mathcal{X} \mapsto \mathcal{Y} = \{1, 2\}$$

Predicted status of new patient with gene expression vector $\boldsymbol{x} \in \mathcal{X}$:

$$\hat{g}_{\boldsymbol{S}}(\boldsymbol{x}) = \hat{y}$$

Exemplary analysis for obtaining a prediction rule

- Data material: 100 DNA microarrays of breast tissues, 50 affected with early stage breast cancer and 50 unaffected
- Analysis:
  1. Normalization using the RMA method. $\Rightarrow$ 47,000 different expression variables
  2. t-test based selection of the 500 most informative variables
  3. Cross-validation based selection of the optimal cost parameter for the Support Vector Machine (SVM) classification method
  4. Fitting the SVM classification method using the optimized cost parameter

$\Rightarrow$ Three preliminary steps before fitting the actual classification method - these steps are part of prediction rule $\hat{g}_{\boldsymbol{s}}(\cdot)$.

"Addon procedures" for preliminary steps

Full versus incomplete cross-validation

Roman Hornung et al.

Introduction

Addon procedures

Full versus incomplete CV

New measure CVIIM

Illustration

Summary & Conclusion

- In general: for the purpose of prediction each step performed for obtaining a prediction rule has to be done on new units as well

- Naive approach: 1) Pool training data with new units, 2) re-perform all preliminary steps, 3) fit the classification method anew on the training data

  ↯ a) Impossible for steps the fitting of which requires the target variable; b) prediction rule is commonly kept fixed

  $\Rightarrow$ All steps have to be integrated into the constructed prediction rule $\hat{g}_{\boldsymbol{s}}(\cdot)$.

"Addon procedures" for preliminary steps

- "Addon procedures": New units made subject to exactly the same procedure as those in the training data, but new units not involved in the adaption of the procedure to the training data.

- Example - Addon procedure for variable selection:

  The same variables are chosen for new units than for those in the training data, but only the training data is used to determine, which variables are used.

Estimating the error of prediction rules by cross-validation (CV)

Full versus incomplete cross-validation
Roman Hornung et al.
Introduction
Addon procedures
Full versus incomplete CV
New measure CVIIM
Illustration
Summary & Conclusion

- Procedure (In practice: repeat and take the average of the results):
  1. Split the data set $\boldsymbol{s}$ into K (approximately) equally sized folds $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K$
  2. For $k = 1, \ldots, K$: Use the units in $\boldsymbol{s} \backslash \boldsymbol{s}_k$ for constructing the prediction rule and the units in $\boldsymbol{s}_k$ as test data.
  3. Average the misclassification rates out of the K splittings in (2).

- **Incomplete CV**: One or more preliminary steps performed before CV

- Correct CV is termed as **Full CV**.

Incomplete cross-validation

Full versus incomplete cross-validation
Roman Hornung et al.
Introduction
Addon procedures
Full versus incomplete CV
New measure CVIIM
Illustration
Summary & Conclusion

- Implication: Part of the prediction rule already constructed on the whole data set
  $\Rightarrow$ Violation of the training and test set principle of CV
- Can lead to severe underestimation of the true error on independent data
- Incomplete CV known to be severely downwardly biased for the case of variable selection

Incomplete cross-validation

Full versus incomplete cross-validation
Roman Hornung et al.
Introduction
Addon procedures
Full versus incomplete CV
New measure CVIIM
Illustration
Summary & Conclusion

- Issue previously unexamined for other preliminary steps in the literature (to our knowledge)
- Preliminary steps almost always conducted before CV

  Examples: normalization of gene expression data, imputation of missing values, variable filtering by variance, dichotomization of continuous variables, data-driven determination of powers of fractional polynomials, sophisticated preprocessing steps for imaging data, ...
- No certainty on the extent of downward bias through incomplete CV with respect to such steps

A new measure of the impact of CV incompleteness (CVIIM)

Full versus incomplete cross-validation

Roman Hornung et al.

Introduction

Addon procedures

Full versus incomplete CV

New measure CVIIM

Illustration

Summary & Conclusion

- ↯ Full CV can be computationally intensive and procedures to integrate certain steps into CV are often not implemented.
- For some steps the extent to which the true error is underestimated by incomplete CV is marginal.
- Desirable: Spot cases, where full CV can be avoided generally and cases where incomplete CV is especially dangerous.

  $\Rightarrow$ Development of simple measure for the degree of bias induced by incomplete CV with respect to specific steps.

A new measure of the impact of CV incompleteness (CVIIM)

Full versus incomplete cross-validation

Roman Hornung et al.

Introduction

Addon procedures

Full versus incomplete CV

New measure CVIIM

Illustration

Summary & Conclusion

Our new measure CVIIM (standing for "Cross-Validation Incompleteness Impact Measure") is estimated by

$$\text{CVIIM}_{\boldsymbol{s},n,K} = 1 - \frac{\text{Incomplete CV error estimate}}{\text{Full CV error estimate}}$$

Set to zero if Incomplete CV error $>$ Full CV error or Full CV error $= 0$

$\text{CVIIM}_{\boldsymbol{s},n,K} \in [0,1]$. Larger values of $\text{CVIIM}_{\boldsymbol{s},n,K}$ are associated with a stronger underestimation of the true error.

Interpretation: Relative reduction of estimated error when performing incomplete CV in comparison to full CV

A new measure of the impact of CV incompleteness (CVIIM)

Full versus incomplete cross-validation

Roman Hornung et al.

Introduction

Addon procedures

Full versus incomplete CV

New measure CVIIM

Illustration

Summary & Conclusion

- Rules of thumb: $\text{CVIIM}_{\boldsymbol{s},n,K} \in$
  $[0, 0.02] \Rightarrow \sim$ no bias, $]0.02, 0.1] \Rightarrow$ weak bias,
  $]0.1, 0.2] \Rightarrow$ medium bias, $]0.2, 0.4] \Rightarrow$ strong bias,
  $]0.4, 1] \Rightarrow$ very strong bias.

- $\text{CVIIM}_{\boldsymbol{s},n,K}$ dependent on data distribution P
  $\Rightarrow$ Calculate $\text{CVIIM}_{\boldsymbol{s},n,K}$ for several data sets before drawing general conclusions regarding the investigated step.

Illustration

- Various real-life data sets, mostly gene-expression data
- Investigated preliminary steps:
  1. Variable selection based on t-tests
  2. Variable filtering by variance
  3. Choice of tuning parameters for various classification methods
  4. Imputation using a variant of k-Nearest-Neighbors
  5. Normalization with the RMA method
  6. Principal Component Analysis

Illustration

- Considered classification methods: Nearest Shrunken Centroids, (Diagonal) Linear Discriminant Analysis, Random Forests
- In each case (incomplete) CVs repeated 300 times and the results averaged.
- Splitting ratios between the sizes of the training and test sets: 2:1 (3-fold CV), 4:1 (5-fold CV) and 9:1 (10-fold CV)

Overview of used data sets

Name	number of samples	number of variables	% diseased	type of variables	disease
`ProstatecTranscr`	102	12,625	51%	transcriptomic	prostate cancer
`HeadNeckcTranscr`	50	22,011	50%	transcriptomic	head and neck squamous
`LungcTranscr`	100	22,277	49%	transcriptomic	lung Adenocarcinoma
`SLETranscr`	36	47,231	56%	transcriptomic	systemic lupus erythematosus
`GenitInfCoww0`	51	21	71%	various	genital infection in cows
`GenitInfCoww1`	51	24	71%	various	genital infection in cows
`GenitInfCoww2`	51	27	71%	various	genital infection in cows
`GenitInfCoww3`	51	26	71%	various	genital infection in cows
`GenitInfCoww4`	51	27	71%	various	genital infection in cows
`ProstatecMethyl`	70	222	41%	methylation	prostate cancer
`ColoncTranscr`	47	22,283	53%	transcriptomic	colon cancer
`WilmsTumorTranscr`	100	22,283	42%	transcriptomic	Wilms' tumor

Results: Variable selection and variable filtering by variance

Full versus incomplete cross-validation

Roman Hornung et al.

Introduction

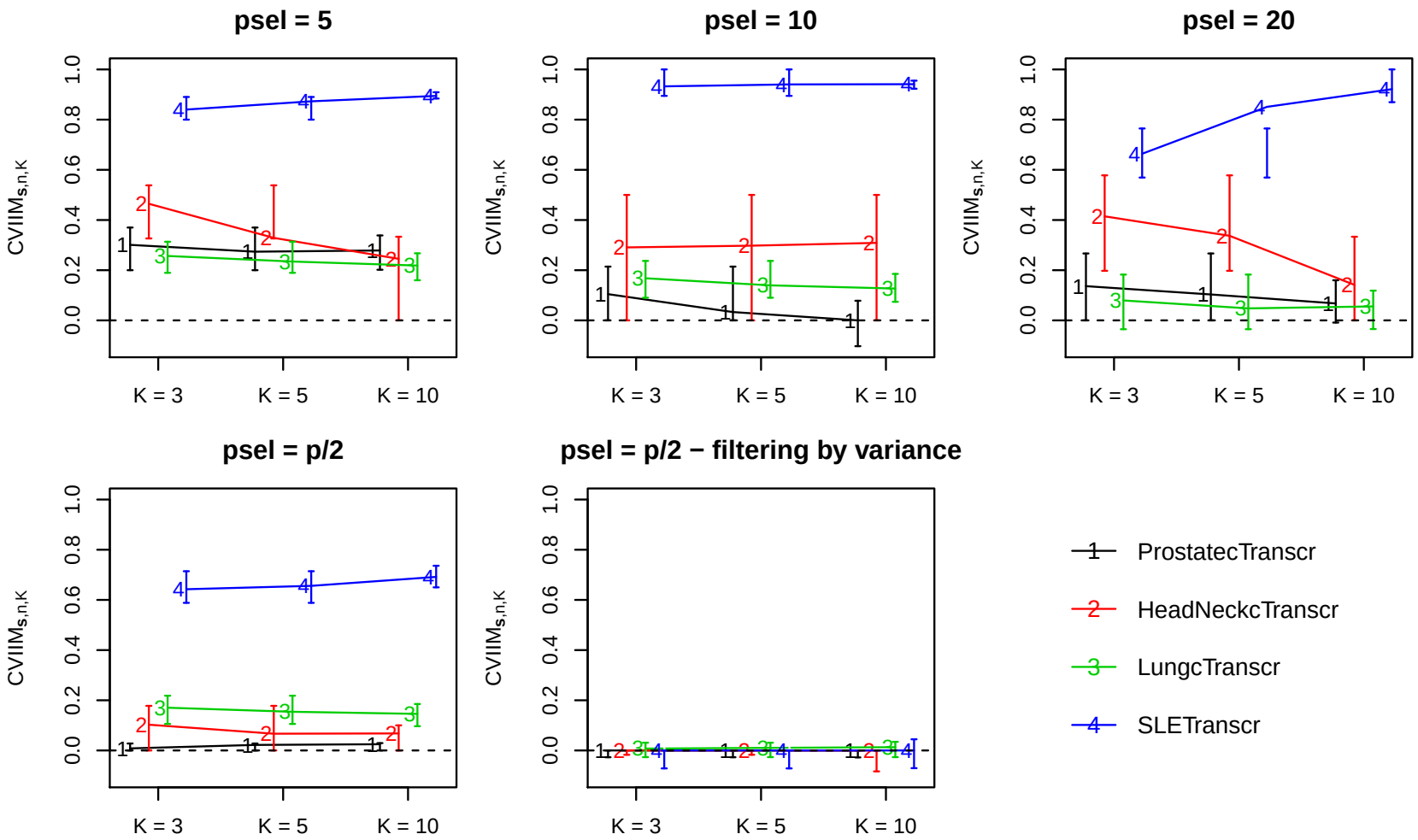
Addon procedures

Full versus incomplete CV

New measure CVIIM

Illustration

Summary & Conclusion



Results: Optimization of tuning parameters

Full versus incomplete cross-validation

Roman Hornung et al.

Introduction

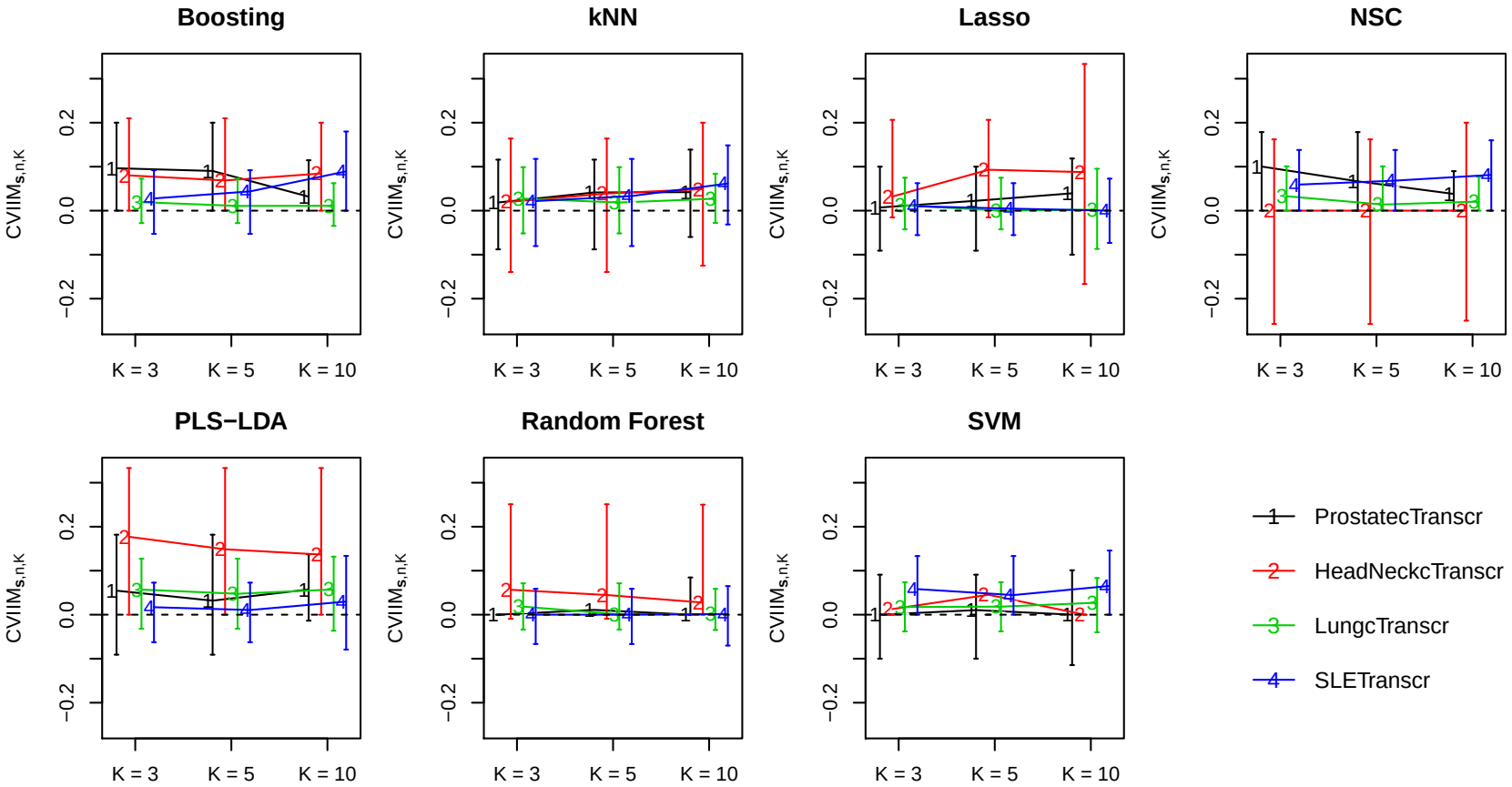
Addon procedures

Full versus incomplete CV

New measure CVIIM

Illustration

Summary & Conclusion



Results: Imputation of missing values, RMA normalization and principal component analysis

Full versus incomplete cross-validation

Roman Hornung et al.

Introduction

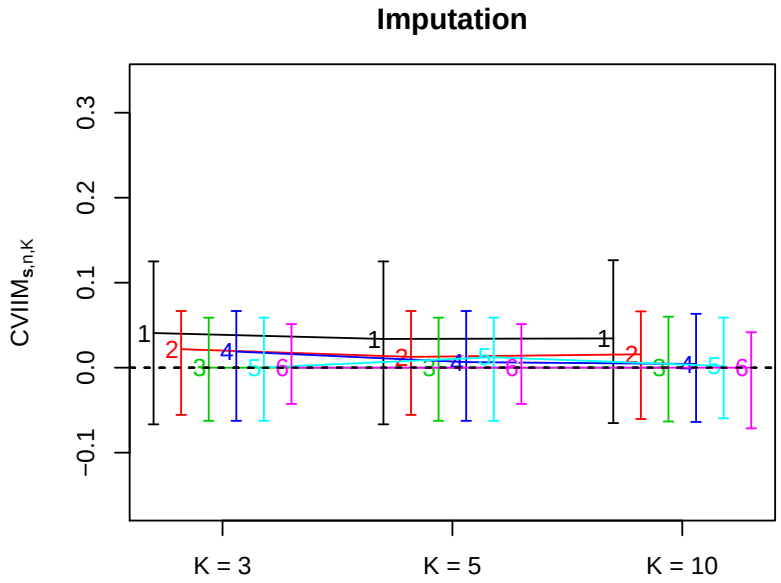
Addon procedures

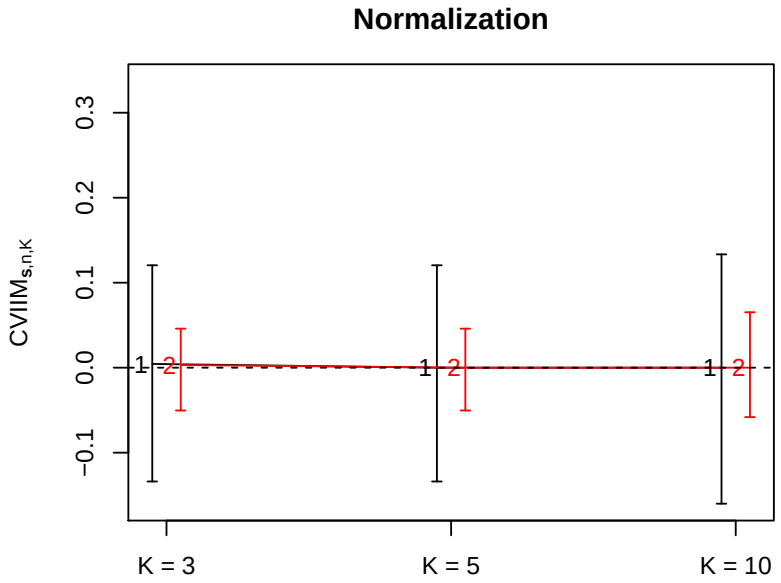
Full versus incomplete CV

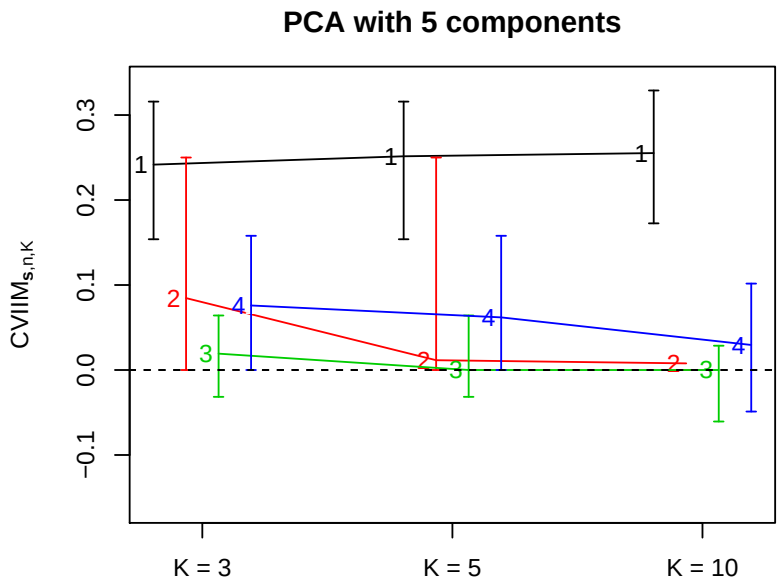
New measure CVIIM

Illustration

Summary & Conclusion







Imputation

1 GenitInfCoww0
2 GenitInfCoww1
3 GenitInfCoww2
4 GenitInfCoww3
5 GenitInfCoww4
6 ProstatecMethyl

Normalization

1 ColoncTranscr
2 WilmsTumorTranscr

PCA

1 ProstatecTranscr
2 HeadNeckcTranscr
3 LungcTranscr
4 SLETranscr

Summary & Conclusion

Full versus incomplete cross-validation

Roman Hornung et al.

Introduction

Addon procedures

Full versus incomplete CV

New measure CVIIM

Illustration

Summary & Conclusion

- Data preparation steps very often done before CV
  ↯ violation of the separation of training and test data.
  $\Rightarrow$ Over-optimistic conclusions possible

- Impact very different for different steps - expected to be smaller for steps disregarding the target variable, but not necessarily the case - relatively high for PCA in our illustration

- New measure CVIIM to assess this impact

- Constantly arising new types of molecular data will require specialized data preparation steps, for which the impact of CV incompleteness will have to be assessed.

Thank you for your attention!

References

Ambroise, C. and McLachlan, G. J. (2002).

Selection bias in gene extraction on the basis of microarray gene-expression data.

Proc. Nat. Acad. Sci. **99,** 6562–6566.

Kostka, D. and Spang, R. (2008).

Microarray based diagnosis profits from better documentation of gene expression signatures.

PLoS Computational Biology **4,** e22.

Simon, R., Radmacher, M. D., Dobbin, K., and McShane, L. M. (2003).

Pitfalls in the use of dna microarray data for diagnostic and prognostic classification.

Journal of the National Cancer Institute **95,** 14–18.

References

Varma, S. and Simon, R. (2006).

Bias in error estimation when using cross-validation for model selection.

BMC Bioinformatics **7,** 91.

Technical Report available online:

Roman Hornung, Christoph Bernau, Caroline Truntzer, Thomas Stadler, and Anne-Laure Boulesteix (2014).

Full versus incomplete cross-validation: measuring the impact of imperfect separation between training and test sets in prediction error estimation.

Department of Statistics, LMU, Technical Report **159**.