Prediction approaches for partly missing multi-omics covariate data: An overview and an empirical comparison study

Roman Hornung[1,2], Frederik Ludwigs[1], Jonas Hagenberg[1,3], Anne-Laure Boulesteix[1,2]

CM Statistics 2023, Berlin

December, 17th, 2023

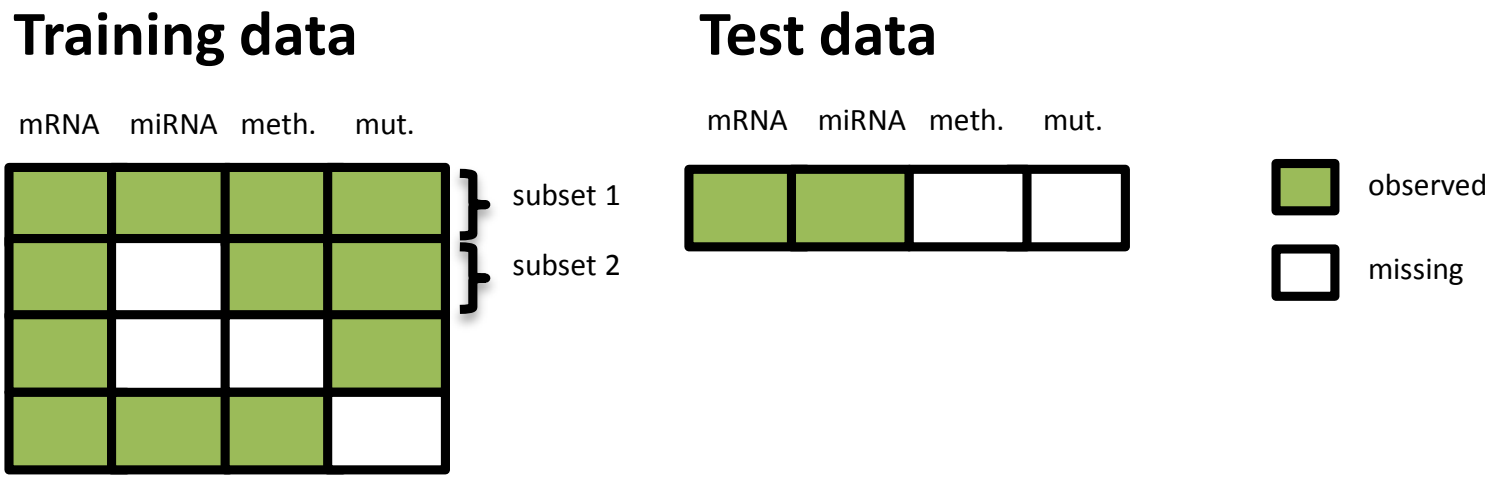
[1] **I**nstitute for Medical Information Processing, **B**iometry and **E**pidemiology, LMU Munich, Munich, Germany

[2] Munich Center for Machine Learning (MCML), Munich, Germany

[3] Department of Translational Research in Psychiatry, Max Planck Institute of Psychiatry, Munich, Germany

Introduction

- **Practical issue** with **multi-omics prediction** rules: **block-wise missing values** in training and test data



- A **number of methods exist** to handle block-wise missingness in prediction.
- This presentation will **overview** these methods and share a **benchmark study** comparing some of them.

Types of methods for handling block-wise missingness

Three categories of methods for handling block-wise missingness:

- Naive methods
- Imputation-based methods
- Methods that deal with the missingness pattern in the modeling

Naive methods

- **Complete case approach**: Use only observations without missing values; main **disadvantage**: potential **exclusion of many** observations, possibly all, depending on missingness patterns

- **Single block approach**: Use **only** the **data type** that has the **best** cross-validated **performance**.

Imputation-based methods

Prediction approaches for partly missing multi-omics covariate data

Roman Hornung

Introduction

Overview of approaches

Benchmark study

Limitations and further points

- **Simplest** approach: Use **existing imputation methods** suitable for high-dimensional data.
- **Mdd-sPLS**: Performs a **PLS-type procedure** on each data type, **imputing missing covariates with** further **PLS** models; information sharing via block-unspecific parameters
- **Priority-LASSO-impute**: **sequential LASSO models** for each block with **imputed offsets** for unavailable data types
- **TOBMI**: *k*-nearest neighbor-based imputation
- Other methods: **matrix completion** (Thung et al. (2014), Cai et al. (2016), Linder and Zhang (2019)); **latent-structure based** methods (Zhang et al. (2020), Yang et al. (2020))

Methods that deal with the missingness pattern in the modeling

- Ingalhalikar et al. (2012) and **multi-source random forests**: **Train separate models** for **different block combinations** and use weighted average predictions.
- **IMSF**, **iSFS**, **MMPFS**: similar, incorporating **block combination-unspecific parameters**
- **Block-wise random forests**: similar, but **training by block**
- Multi-source random forests, iMSF, and MMPFS do **not require retraining** under **varying test data missingness** patterns.
- **Hypergraph-based methods**: Liu et al. (2017), Dong et al. (2021), Chen & Zhang (2020)

Design of the benchmark study

- **Compared methods**: complete case approach, single block approach, TOBMI, mdd-sPLS, block-wise random forests, multi-source random forests, priority-LASSO-impute
- **13 multi-omics datasets** from The Cancer Genome Atlas (TCGA) without (block-wise) missing values
- **Data types**: clinical covariates, mRNA, miRNA, copy-number variation
- **Outcome**: TP53 mutation (yes vs. no)
- **Repeated random splitting** into **training** and **test** data and induction of **different patterns of block-wise missingness**
- **Performance metrics**: accuracy, Brier score, AUC

Block-wise missingness patterns in training and test data

Prediction approaches for partly missing multi-omics covariate data

Roman Hornung

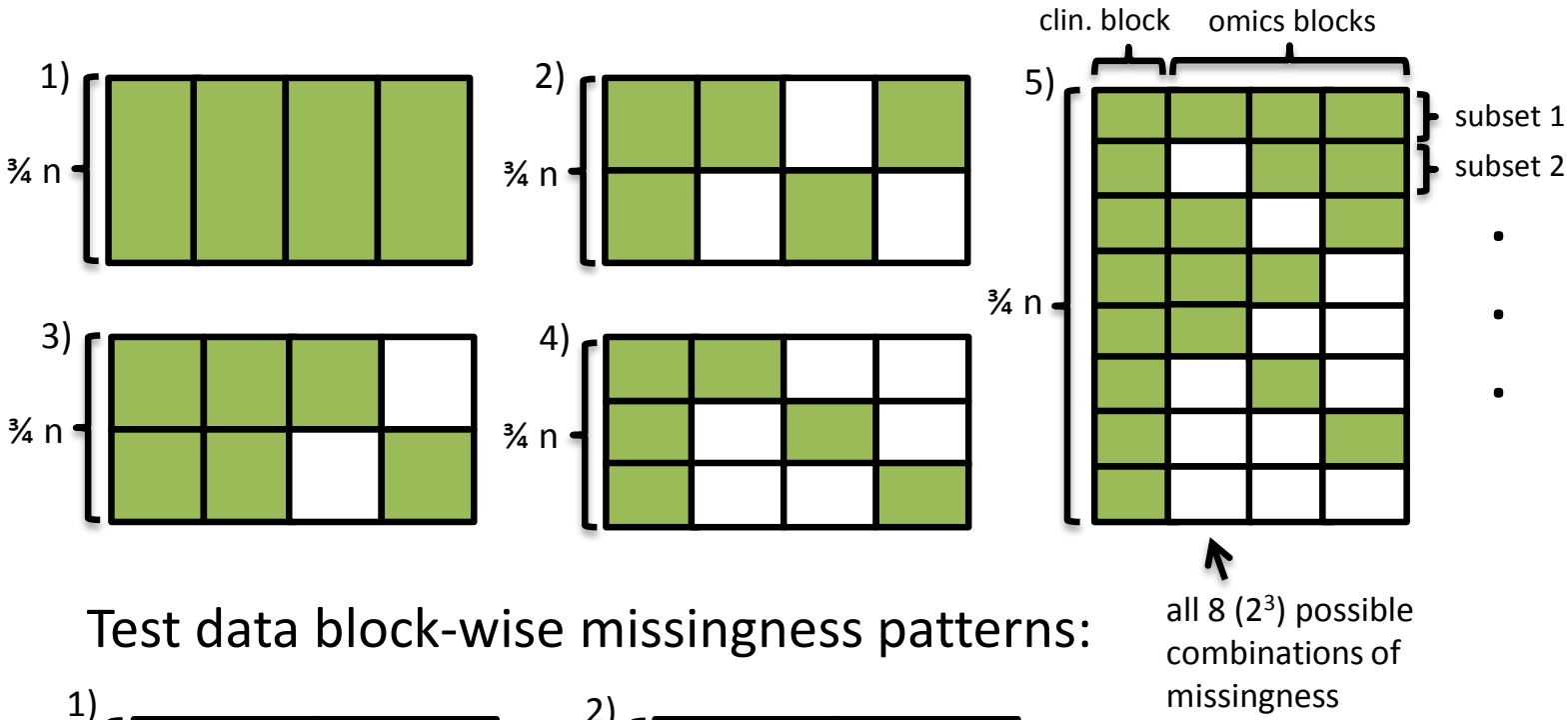
Introduction

Overview of approaches

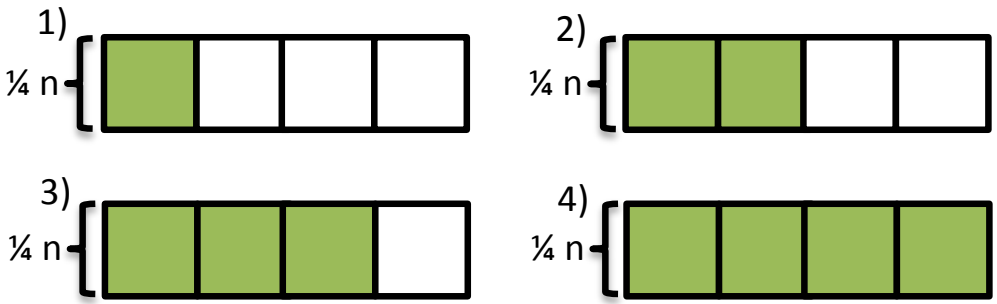
Benchmark study

Limitations and further points

Training data block-wise missingness patterns:

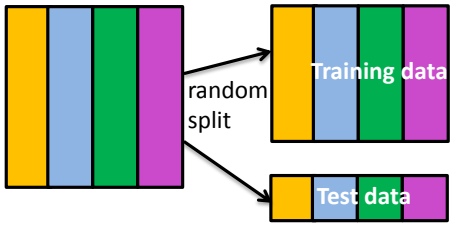


Test data block-wise missingness patterns:

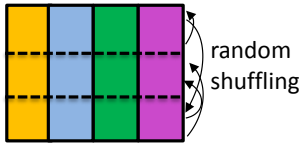


Procedure used for inducing block-wise missingness in the training and test data

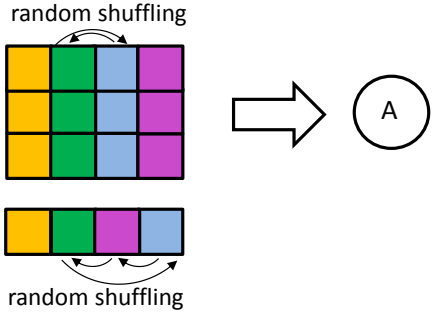
1) Split data set into training and test data (3:1 ratio).



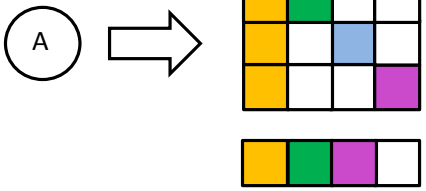
2) Assign subset memberships to the training observations randomly.



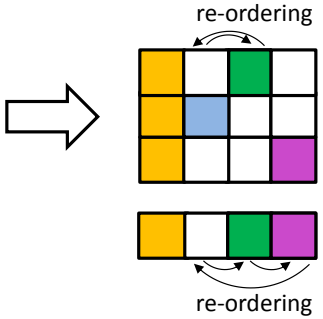
3) Permute omics blocks randomly (clinical block stays in the first position).



4) Delete values in the data according to the respective trmbp and tebmp.



5) Re-order the blocks to their original ordering.



Global performance across considered block-wise missingness patterns

Prediction approaches for partly missing multi-omics covariate data

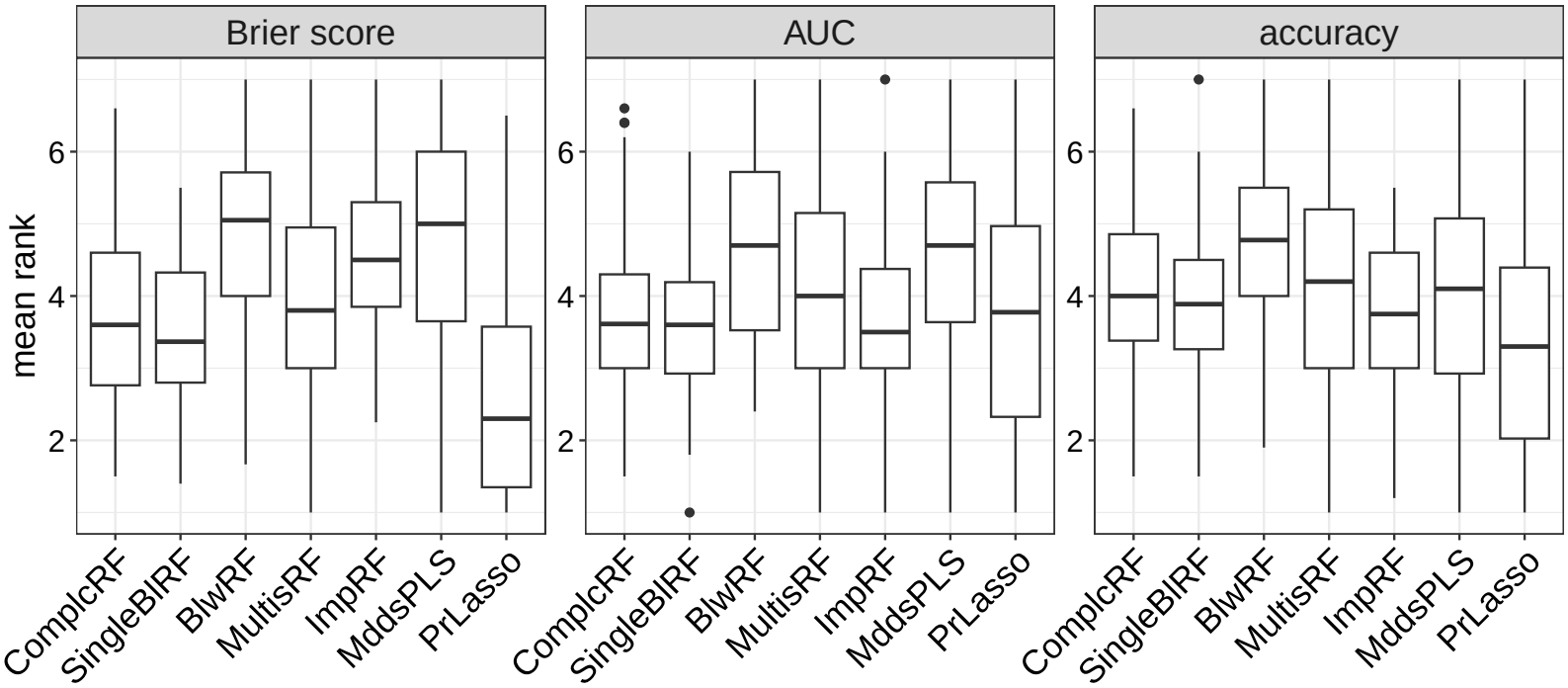
Roman Hornung

Introduction

Overview of approaches

Benchmark study

Limitations and further points



Results summary

- **Priority-LASSO-impute generally outperformed** other methods in most settings.
- Its effectiveness **likely** comes from **better use of** the **clinical** covariates.
- Block-wise random forests and mdd-sPLS, **treating blocks separately**, **underperformed**, indicating a need for collective block modeling.
- **Rankings** were **quite consistent across** missingness **patterns**, with minor variations by pattern and metric.
- **Performance with full and missing** data was **comparable**, showing effective use of available predictive information.

Limitations and further points

- **High variability in results** across datasets and metrics; **limited** training/test **data divisions**
- **Training and test data** from the same datasets. $\Rightarrow$ **same distributions**; however: **in practice**, omics data subject to **batch effects**
- Blocks were missing completely at random (**MCAR**) — **realistic** and **most methods can handle** missing at random (**MAR**) blocks.
- **Some compared methods** were developed **by us**. $\Rightarrow$ great care taken to **maintain study neutrality**
- **Preliminary results** suggest **better predictions with fewer omics types**, possibly affecting method rankings and the impact of block-wise missingness in the field.

References – I

Cai, T., Cai, T. T., Zhang, A., 2016.
Structured matrix completion with applications to genomic data integration.
J. Am. Stat. Assoc. 111, 621–633.

Chen, J., Zhang, A., 2020.
Hgmf: Heterogeneous graph-based fusion for multimodal data with incompleteness.
In Proc. 26th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining, 1295–1305.

Dong, A., Li, Z., Wang, M., Shen, D., Liu, M., 2021.
High-order Laplacian regularized low-rank representation for multimodal dementia diagnosis.
Front. Neurosci. 15, 634124.

Dong, X., Lin, L., Zhang, R., Zhao, Y., Christiani, D. C., Wei, Y., Chen, F., 2019.
Tobmi: Trans-omics block missing data imputation using a k-nearest neighbor weighted approach.
Bioinformatics 35(8), 1278–1283.

Hagenberg, J., 2020.
Penalized regression approaches for prognostic modelling using multi-omics data with block-wise missing values.
Master's Thesis, University of Munich.

Ingalhalikar, M., Parker, W. A., Bloy, L., Roberts, T. P., Verma, R., 2012.
Using multiparametric data with missing features for learning patterns of pathology.
In Int. Conf. Medical Image Computing and Computer-Assisted Intervention, 468–475.

Klau, S., Jurinovic, V., Hornung, R., Herold, T., Boulesteix, A.-L., 2018.
Priority-lasso: A simple hierarchical approach to the prediction of clinical outcome using multi-omics data.
BMC Bioinformatics 19, 322.

References – II

Krautenbacher, N., Flach, N., Böck, A., Laubhahn, K., Laimighofer, M., Theis, F. J., Ankerst, D. P., Fuchs, C., Schaub, B., 2019.
A strategy for high-dimensional multivariable analysis classifies childhood asthma phenotypes from genetic, immunological, and environmental factors.
Allergy 74(7), 1364–1373.

Lan, Q., Jiang, S., 2021.
A method of credit evaluation modeling based on block-wise missing data.
Appl. Intell. 51(10), 6859–6880.

Linder, H., Zhang, Y., 2019.
Iterative integrated imputation for missing data and pathway models with applications to breast cancer subtypes.
Commun. Stat. Appl. Methods 26, 411–430.

Liu, M., Gao, Y., Yap, P.-T., Shen, D., 2017.
Multi-hypergraph learning for incomplete multimodality data.
IEEE J. Biomed. Health Inform. 22, 1197–1208.

Lorenzo, H., Saracco, J., Thiébaut, R., 2019.
Supervised learning for multi-block incomplete data.
arXiv:1901.04380.

Ludwigs, F., 2020.
A comparison study of prediction approaches for multiple training data sets and test data with block-wise missing values.
Master's Thesis, University of Munich.

References – III

Thung, K.-H., Wee, C.-Y., Yap, P.-T., Shen, D., Initiative, A. D. N., 2014.
Neurodegenerative disease diagnosis using incomplete multi-modality data via matrix shrinkage and completion.
NeuroImage 91, 386–400.

Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P. M., Ye, J., Alzheimer's Disease Neuroimaging Initiative, 2014.
Bi-level multi-source learning for heterogeneous block-wise missing data.
NeuroImage 102(Pt 1), 192–206.

Yang, X., Kim, Y.-J., Taub, M., Azevedo, R., Chi, M., 2020.
Prime: Block-wise missingness handling for multimodalities in intelligent tutoring systems.
In Int. Conf. Multimedia Modeling, 63–75.

Yuan, L., Wang, Y., Thompson, P. M., Narayan, V. A., Ye, J., Initiative, A. D. N., 2012.
Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data.
NeuroImage 61, 622–632.

Zhang, Y., Tang, N., Qu, A., 2020.
Imputed factor regression for high-dimensional block-wise missing data.
Statistica Sinica 30(2), 631–651.

Thank you for your attention!

Recommended reading

Hornung, R., Ludwigs, F., Hagenberg, J., Boulesteix, A.-L. 2023.
Prediction approaches for partly missing multi-omics covariate data: A literature review and an empirical comparison study.
Wiley Interdiscip. Rev. Comput. Stat. e1626.