Modern approaches for component-wise boosting:

Automation, efficiency, and distributed computing with application to the medical domain

---

Daniel Schalk

March 24, 2023

Supervisor: Prof. Dr. Bernd Bischl
Referees: Prof. Dr. Matthias Schmid, PD Dr. Fabian Scheipl
Chair of the examination panel: Prof. Dr. Christian Heumann

Overview

Publications

List with all Publications

Structure of the talk

- Background
    - S1
- Efficiency
- Automation
- Distributed computing

Background

Terminology

- p-dimensional covariate or feature vector $\mathbf{x} = (x_1, \dots, x_p) \in \mathcal{X} = \mathcal{X}_1 \times \dots \times$ and target variable $y \in \mathcal{Y}$.
- Data set $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)}) \mid i = 1, \dots, n\}$ with $(\mathbf{x}^{(i)}, y^{(i)})$ sampled from an unknown probability distribution $\mathbb{P}_{xy}$.
- True underlying relationship $f : \mathcal{X}^p \to \mathbb{R}$, $\mathbf{x} \mapsto f(\mathbf{x})$.
- Goal of Machine Learning (ML) is to estimate a model $\hat{f} = \text{arg min}_f \mathcal{R}_{\text{emp}}(f|\mathcal{D})$ with
  - Empirical risk $\mathcal{R}_{\text{emp}}(f|\mathcal{D}) = n^{-1} \sum_{(\mathbf{x}, y) \in \mathcal{D}} L(y, \hat{f}(\mathbf{x}))$ and
  - Loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, $(y, \hat{y}) \mapsto L(y, \hat{y})$.
- The inducer $\mathcal{I} : \mathbb{D} \times \Lambda \to \mathcal{F}$, $(\mathcal{D}, \boldsymbol{\lambda}) \mapsto \hat{f} = \mathcal{I}_{\boldsymbol{\lambda}}(\mathcal{D})$ gets a data set $\mathcal{D} \in \mathbb{D}$ with hyperparameters (HPs) $\boldsymbol{\lambda} \in \Lambda$.

Gradient boosting

- Aim is to fit a model $\hat{f}$ by conducting functional gradient descent $\hat{f}^{[m-1]} = \hat{f}^{[m]} + \nu \hat{b}^{[m]}$.
- $\hat{b}$ is a base learner fitted to pseudo residuals $r^{[m](i)} = - \left. \frac{\partial L(y^{(i)}, f(\mathrm{x}^{(i)}))}{\partial f(\mathrm{x}^{(i)})} \right|_{f=\hat{f}^{[m-1]}}$, $i \in \{1, \dots, n\}$
- The pseudo residuals contain information in which direction to move $\hat{f}^{[m]}$ for a better fit to the training data $\mathcal{D}$.
- The fitting is initialized with $\hat{f}^{[0]}(\mathrm{x}) = \arg\min_{c \in \mathcal{Y}} \mathcal{R}_{\text{emp}}(c|\mathcal{D})$ and repeated M times or until an early stopping criterion is met.

Component-wise gradient boosting – Basics

- Compared to GB, component-wise gradient boosting (CWB) can choose from a set of K base learners $b \in \{b_1, \dots, b_K\}$.
- Often, $b_1, \dots, b_K$ are chosen to be (interpretable) statistical models and hence f corresponds to a generalized additive model $f(\mathbf{x}) = f_0 + \sum_{k=1}^{K} b_k(\mathbf{x})$ with intercept $_0$.
- Advantages of CWB:
  - Feasible to get fit in high-dimensional feature spaces ($p \gg n$).
  - An inherent (unbiased) feature selection.
  - Interpretable/explainable partial feature effects (depending on the choice of base learners).

Component-wise gradient boosting – Algorithm

Algorithm 1 Vanilla CWB algorithm

Input Train data $\mathcal{D}$, learning rate ν, number of boosting iterations M, loss function L, base learners $b_1, \dots, b_K$

Output Model $\hat{f} = \hat{f}^{[M]}$

```
procedure CWB($\mathcal{D}, \nu, L, b_1, \dots, b_K$)
    Initialize: $f_0 = \hat{f}^{[0]}(\mathsf{x}) = \text{arg min}_{c \in \mathcal{Y}} \mathcal{R}_{\text{emp}}(c|\mathcal{D})$
    while $m \leq M$ do
        $r^{[m](i)} = -\left.\frac{\partial L(y^{(i)}, f(\mathsf{x}^{(i)}))}{\partial f(\mathsf{x}^{(i)})}\right|_{f=\hat{f}^{[m-1]}}, \quad \forall i \in \{1, \dots, n\}$
        for $k \in \{1, \dots, K\}$ do
            $\hat{\boldsymbol{\theta}}_k^{[m]} = \left(Z_k^{\mathsf{T}} Z_k + K_k\right)^{-1} Z_k^{\mathsf{T}} \boldsymbol{r}^{[m]}$
            $\text{SSE}_k = \sum_{i=1}^{n} (r^{[m](i)} - b_k(\mathsf{x}^{(i)}|\hat{\boldsymbol{\theta}}_k^{[m]}))^2$
        $k^{[m]} = \text{arg min}_{k \in \{1, \dots, K\}} \text{SSE}_k$
        $\hat{f}^{[m]}(\mathsf{x}) = \hat{f}^{[m-1]}(\mathsf{x}) + \nu b_{k^{[m]}}(\mathsf{x}|\hat{\boldsymbol{\theta}}_{k^{[m]}}^{[m]})$
    return $\hat{f} = \hat{f}^{[M]}$
```

Component-wise gradient boosting – Example

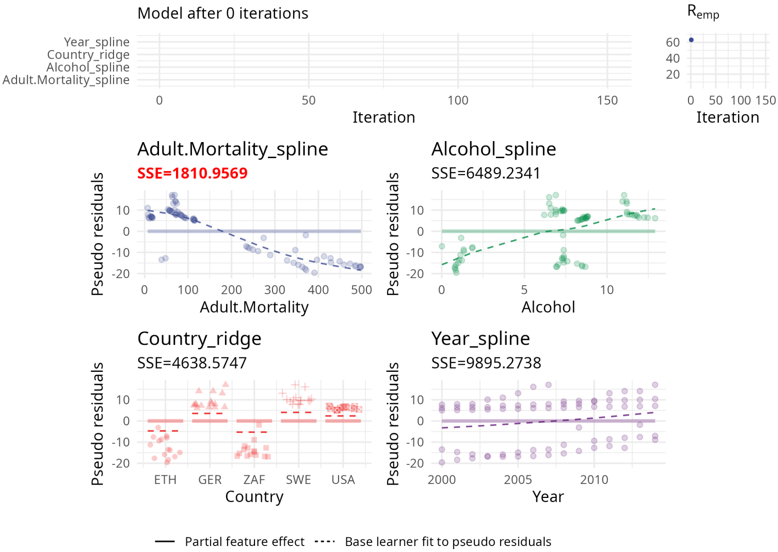
Example throughout this presentation is a subset of a WHO data set[1] about life expectation in years per country:

- Target variable is `Life.expectancy` in years.
- Features are `Country`, `Year`, `Alcohol` recorded per capital (15+) consumption (in liters of pure alcohol), and `Adult.Mortality` rates of both sexes of dying between 15 and 60 years per 1000 population.
- Numerical features `Year`, `Alcohol` and `Adult.Mortality` are modeled as P-splines (Eilers and Marx, 1996) and `Country` as one-hot-encoded linear model with ridge penalty.

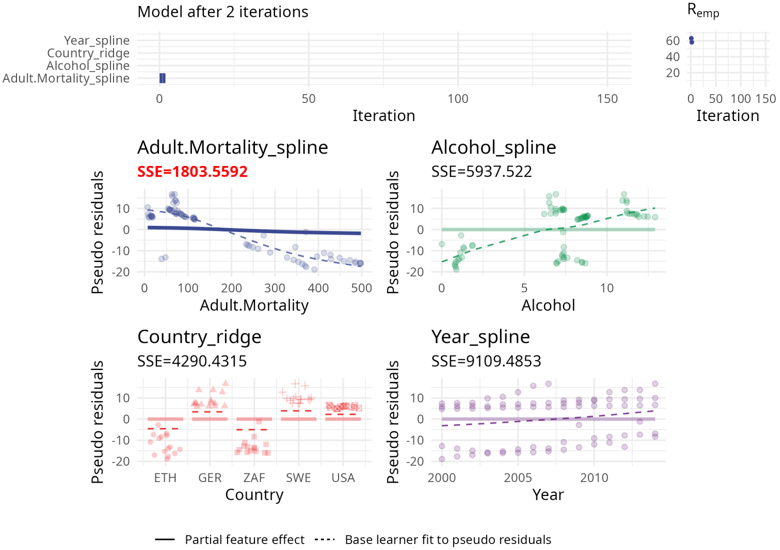
---

[1] Full description and data is available at `kaggle.com/datasets/kumarajarshi/life-expectancy-who`

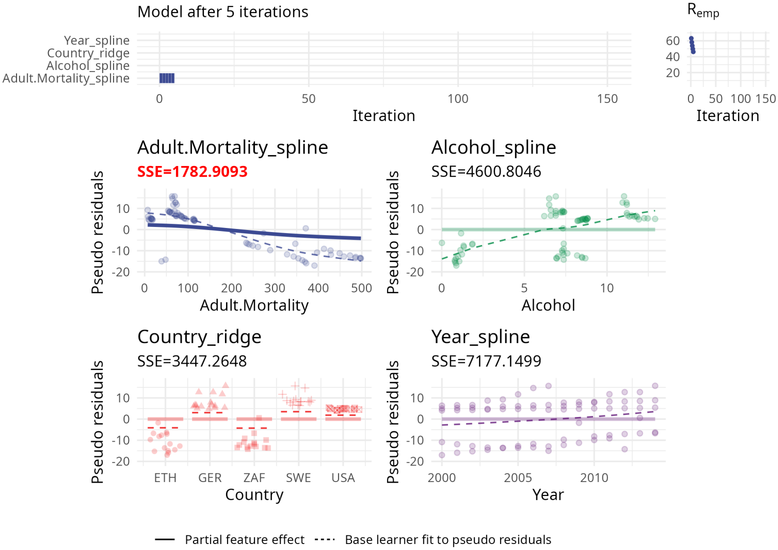
Component-wise gradient boosting – Example



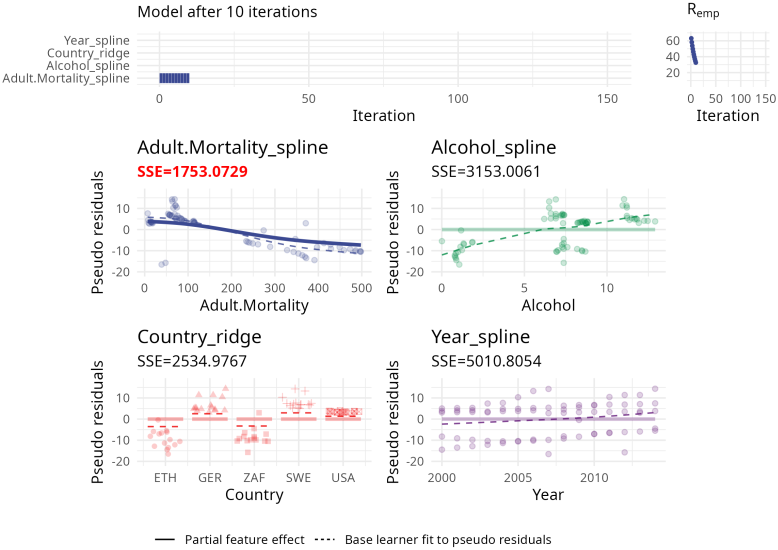
Component-wise gradient boosting – Example



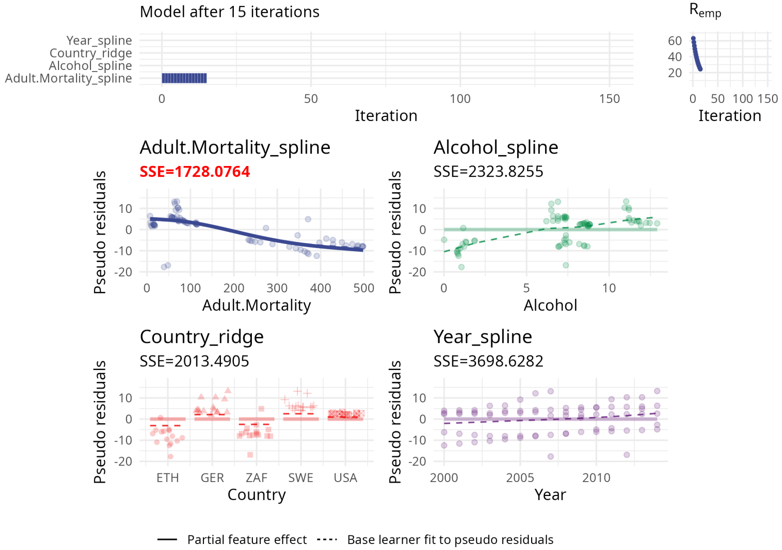
Component-wise gradient boosting – Example



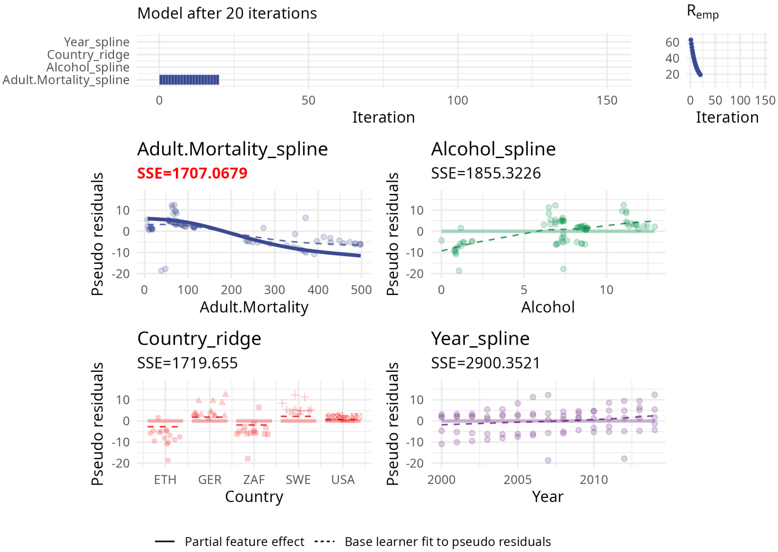
Component-wise gradient boosting – Example



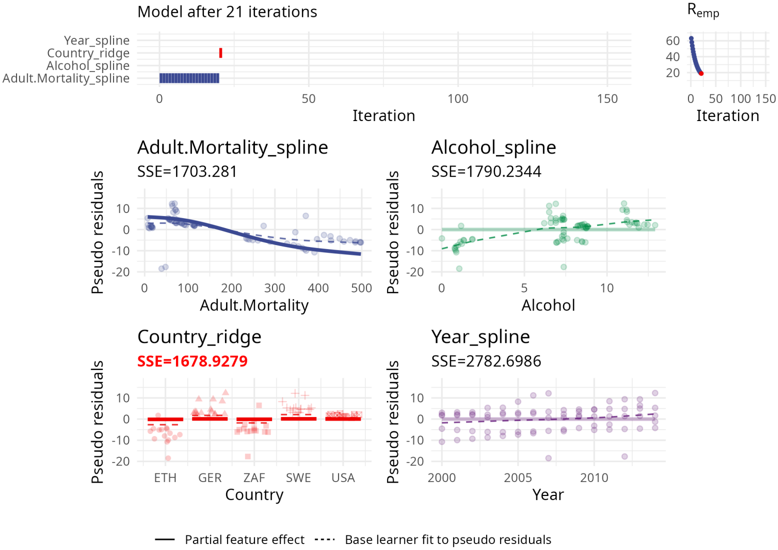
Component-wise gradient boosting – Example



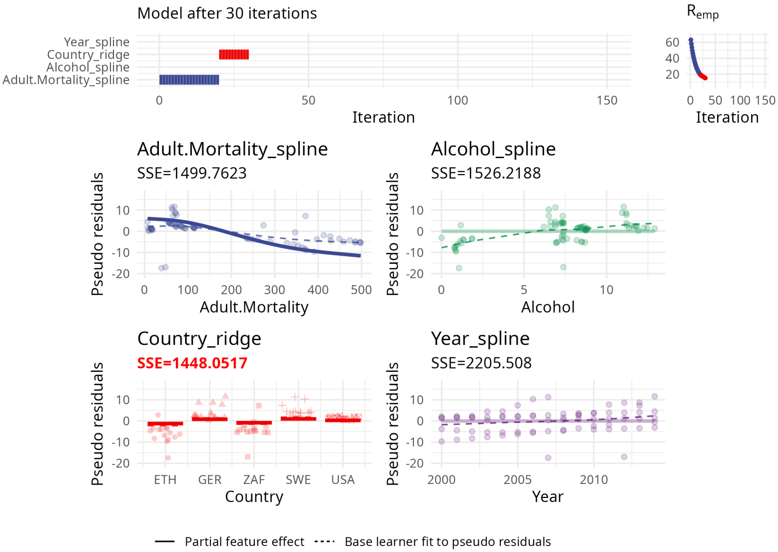
Component-wise gradient boosting – Example



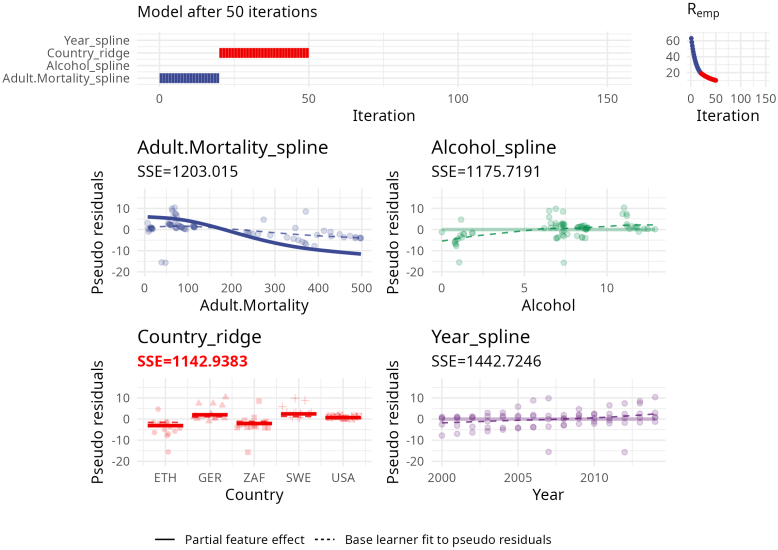
Component-wise gradient boosting – Example



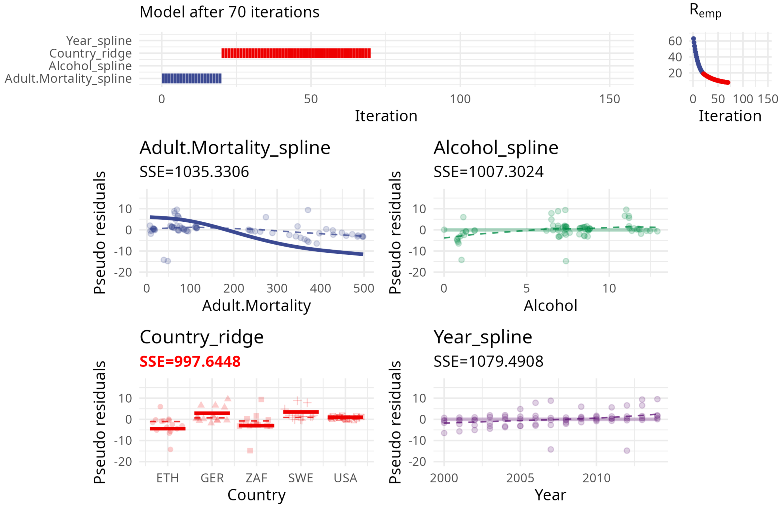
Component-wise gradient boosting – Example



Component-wise gradient boosting – Example

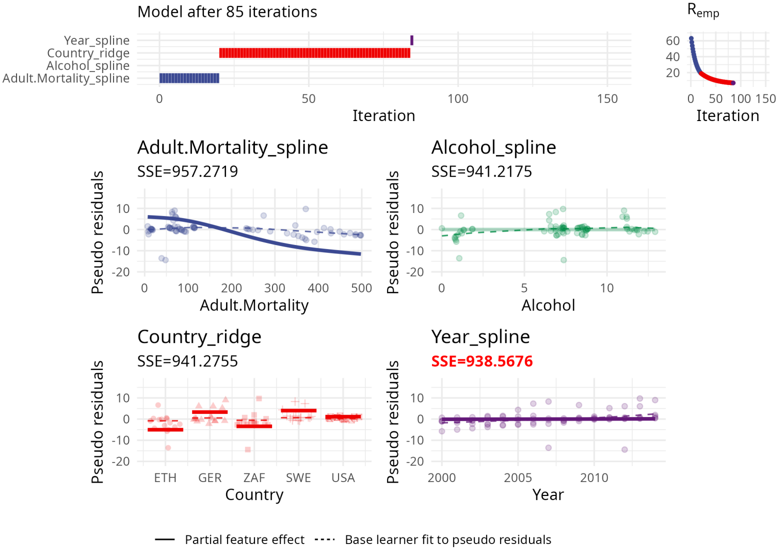


Component-wise gradient boosting – Example

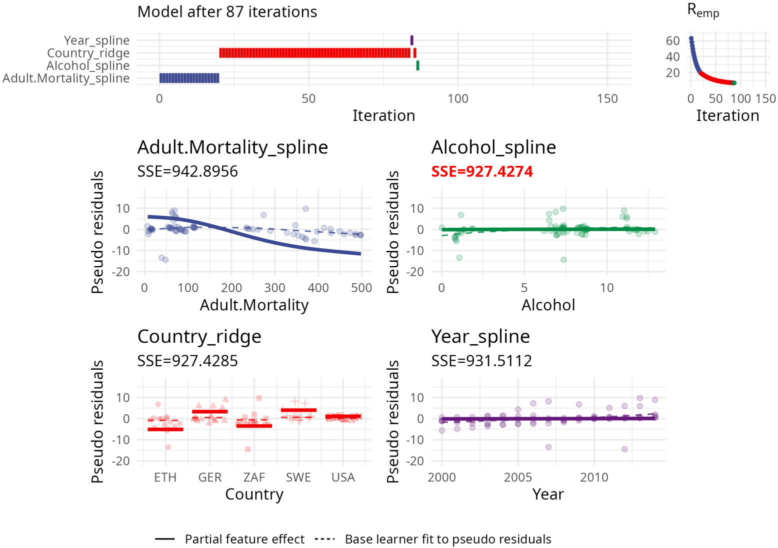


— Partial feature effect ---- Base learner fit to pseudo residuals

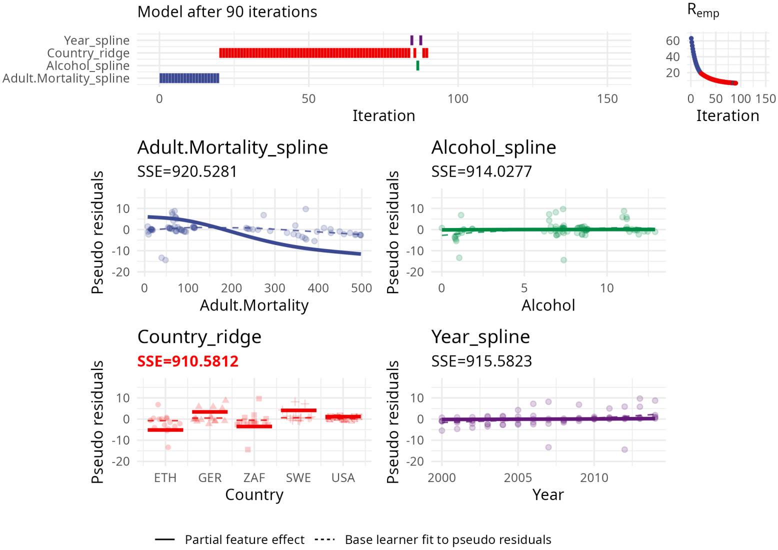
Component-wise gradient boosting – Example



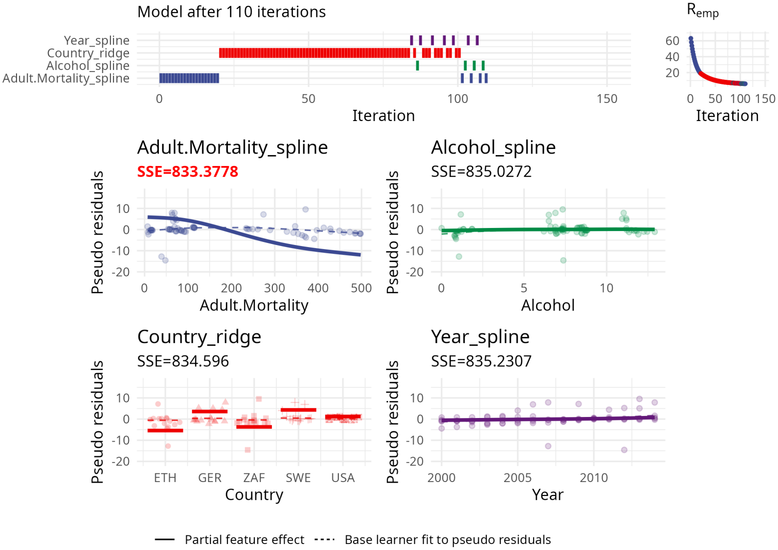
Component-wise gradient boosting – Example



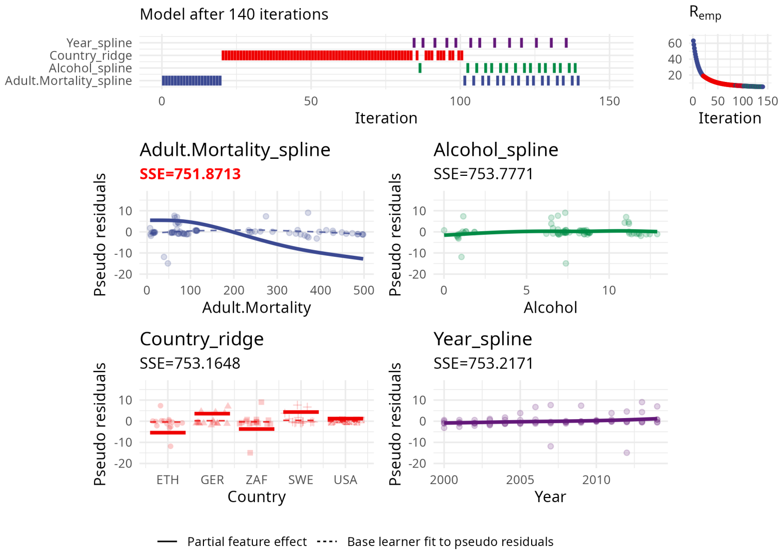
Component-wise gradient boosting – Example



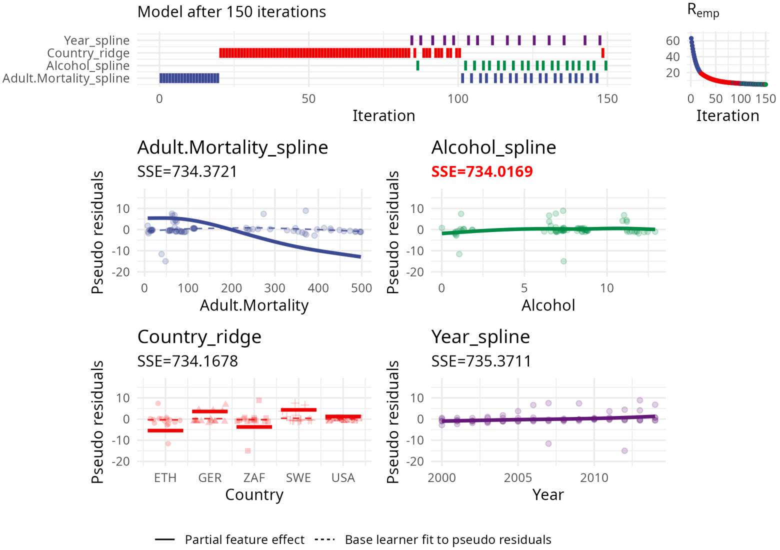
Component-wise gradient boosting – Example



Component-wise gradient boosting – Example



Component-wise gradient boosting – Example



Part I - Efficiency

Name paper here and what it solves

- Memory consumption

Part II - Interpretable ML Framework

Name paper here and what it solves

- Easy access etc.

Part III - Distributed Component-wise Boosting

Name paper here and what it solves

- Distributed etc.

Misc: Distributed Model evaluation, A few slides in the end that describes the issue (sometimes not feasible to just aggregate), how we solved it for the AUC, and what can be done in the future

Efficiency

Efficiency

About

Adaption

About

Results

About

Automation

Automation

About

Autocompboost

About

Results

About

Outlook

About

Distributed computing

Distributed computing

About

Adaption

About

Results

About

Outlook

About

Bla i

Bla (see, e.g., Pepe, 2003), or DeLong et al. (1988)

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

Audet, C. and Hare, W. (2017). *Derivative-free and blackbox optimization*, volume 2. Springer.

Barrett, R., Berry, M., Chan, T. F., Demmel, J., Donato, J., Dongarra, J., Eijkhout, V., Pozo, R., Romine, C., and Van der Vorst, H. (1994). *Templates for the solution of linear systems: building blocks for iterative methods*. SIAM.

Bates, D., Maechler, M., and Jagan, M. (2022). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.5-1.

Bekkerman, R., Bilenko, M., and Langford, J. (2011). *Scaling up machine learning: Parallel and distributed approaches*. Cambridge University Press.

Biau, G., Cadre, B., and Rouvìère, L. (2019). Accelerated gradient boosting. *Machine Learning*, 108(6):971–992.

References ii

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L., et al. (2021). Hyperparameter optimization: Foundations, algorithms, best practices and open challenges. *arXiv preprint arXiv:2107.05847*.

Bischl, B., Mersmann, O., Trautmann, H., and Weihs, C. (2012). Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evolutionary computation*, 20:249–75.

Bost, R., Popa, R. A., Tu, S., and Goldwasser, S. (2014). Machine learning classification over encrypted data. Cryptology ePrint Archive, Paper 2014/331. `https://eprint.iacr.org/2014/331`.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brier, G. W. et al. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Brockhaus, S., Rügamer, D., and Greven, S. (2020). Boosting functional regression models with fdboost. *Journal of Statistical Software*, 94(10):1–50.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, 44(1):108–132.

Bühlmann, P., Hothorn, T., et al. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical science*, 22(4):477–505.

Bühlmann, P. and Yu, B. (2003). Boosting with the L2 loss: regression and classification. *Journal of the American Statistical Association*, 98(462):324–339.

Buluc, A. and Gilbert, J. R. (2008). Challenges and advances in parallel sparse matrix-matrix multiplication. In *2008 37th International Conference on Parallel Processing*, pages 503–510.

References iii

Casalicchio, G. (2019). *On benchmark experiments and visualization methods for the evaluation and interpretation of machine learning models*. PhD dissertation, LMU Munich.

Chen, Y.-R., Rezapour, A., and Tzeng, W.-G. (2018). Privacy-preserving ridge regression on distributed data. *Information Sciences*, 451:34–49.

Choi, J., Walker, D. W., and Dongarra, J. J. (1994). Pumma: Parallel universal matrix multiplication algorithms on distributed memory concurrent computers. *Concurrency: Practice and Experience*, 6(7):543–570.

Coors, S., Schalk, D., Bischl, B., and Rügamer, D. (2021). Automatic componentwise boosting: An interpretable automl system. *ECML-PKDD Workshop on Automating Data Science*.

Cunha, M., Mendes, R., and Vilela, J. P. (2021). A survey of privacy-preserving mechanisms for heterogeneous data types. *Computer Science Review*, 41:100403.

Dagum, L. and Menon, R. (1998). Openmp: an industry standard api for shared-memory programming. *Computational Science & Engineering, IEEE*, 5(1):46–55.

Davis, T. A. (2006). *Direct methods for sparse linear systems*. SIAM.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845.

Drozdal, J., Weisz, J., Wang, D., Dass, G., Yao, B., Zhao, C., Muller, M., Ju, L., and Su, H. (2020). Trust in automl: Exploring information needs for establishing trust in automated machine learning systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI ’20, page 297–307, New York, NY, USA. Association for Computing Machinery.

References iv

Duff, I. S., Grimes, R. G., and Lewis, J. G. (1989). Sparse matrix test problems. *ACM Transactions on Mathematical Software (TOMS)*, 15(1):1–14.

Dwork, C. (2006). Differential privacy. In *International Colloquium on Automata, Languages, and Programming*, pages 1–12. Springer.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. (2006a). Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503. Springer.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006b). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer.

Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical science*, pages 89–102.

Fang, H. and Qian, Q. (2021). Privacy preserving machine learning with homomorphic encryption and federated learning. *Future Internet*, 13(4).

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Feurer, M. and Hutter, F. (2019). Hyperparameter optimization. In *Automated machine learning*, pages 3–33. Springer, Cham.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., and Hutter, F. (2015). Efficient and robust automated machine learning. *Advances in neural information processing systems*, 28.

References v

Flach, P. (2012). *Machine learning: the art and science of algorithms that make sense of data*. Cambridge university press.

Freitas, A. A. (2019). Automated machine learning for studying the trade-off between predictive accuracy and interpretability. In Holzinger, A., Kieseberg, P., Tjoa, A. M., and Weippl, E., editors, *Machine Learning and Knowledge Extraction*, pages 48–66, Cham. Springer International Publishing.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Gambs, S., Kégl, B., and Aïmeur, E. (2007). Privacy-preserving boosting. *Data Mining and Knowledge Discovery*, 14(1):131–170.

Gaye, A., Marcon, Y., Isaeva, J., LaFlamme, P., Turner, A., Jones, E. M., Minion, J., Boyd, A. W., Newby, C. J., Nuotio, M.-L., et al. (2014). Datashield: taking the analysis to the data, not the data to the analysis. *International journal of epidemiology*, 43(6):1929–1944.

Gong, M., Xie, Y., Pan, K., Feng, K., and Qin, A. (2020). A survey on differentially private machine learning [review article]. *IEEE Computational Intelligence Magazine*, 15(2):49–64.

Gordon, D. F. and Desjardins, M. (1995). Evaluation and selection of biases in machine learning. *Machine learning*, 20(1):5–22.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge.

References vi

Hofner, B., Hothorn, T., Kneib, T., and Schmid, M. (2011). A framework for unbiased model selection based on boosting. *Journal of Computational and Graphical Statistics*, 20(4):956–971.

Hofner, B., Mayr, A., and Schmid, M. (2016). gamboostLSS: An R package for model building and variable selection in the GAMLSS framework. *Journal of Statistical Software*, 74(1).

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2010). Model-based boosting 2.0. *The Journal of Machine Learning Research*, 11:2109–2113.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2020). *mboost: Model-based boosting*. R package version 2.9-7.

Hutter, F., Kotthoff, L., and Vanschoren, J. (2019). *Automated machine learning: methods, systems, challenges*. Springer Nature.

Jayaraman, B. and Evans, D. (2019). Evaluating differentially private machine learning in practice. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 1895–1912.

John, G. H. (1995). Robust decision trees: Removing outliers from databases. In *KDD*, volume 95, pages 174–179.

Karr, A. F., Lin, X., Sanil, A. P., and Reiter, J. P. (2005). Secure regression on distributed databases. *Journal of Computational and Graphical Statistics*, 14(2):263–279.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kotthoff, L., Thornton, C., Hoos, H. H., Hutter, F., and Leyton-Brown, K. (2017). Auto-weka 2.0: Automatic model selection and hyperparameter optimization in weka. *Journal of Machine Learning Research*, 18(25):1–5.

References vii

Kotthoff, L., Thornton, C., Hoos, H. H., Hutter, F., and Leyton-Brown, K. (2019). Auto-weka: Automatic model selection and hyperparameter optimization in weka. In *Automated machine learning*, pages 81–95. Springer, Cham.

Lang, S., Umlauf, N., Wechselberger, P., Harttgen, K., and Kneib, T. (2014). Multilevel structured additive regression. *Statistics and Computing*, 24(2):223–238.

Lazarevic, A. and Obradovic, Z. (2001). The distributed boosting algorithm. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 311–316.

Li, J., Kuang, X., Lin, S., Ma, X., and Tang, Y. (2020). Privacy preservation for machine learning training and classification based on homomorphic encryption schemes. *Information Sciences*, 526:166–179.

Li, Y., Jiang, X., Wang, S., Xiong, H., and Ohno-Machado, L. (2016). Vertical grid logistic regression (vertigo). *Journal of the American Medical Informatics Association*, 23(3):570–579.

Li, Z. and Wood, S. N. (2020). Faster model matrix crossproducts for large generalized linear models with discretized covariates. *Statistics and Computing*, 30(1):19–25.

Liew, B. X., Rügamer, D., Abichandani, D., and De Nunzio, A. M. (2020a). Classifying individuals with and without patellofemoral pain syndrome using ground force profiles – Development of a method using functional data boosting. *Gait & Posture*, 80:90–95.

Liew, B. X., Rügamer, D., Stocker, A., and De Nunzio, A. M. (2020b). Classifying neck pain status using scalar and functional biomechanical variables – Development of a method using functional data boosting. *Gait & posture*, 76:146–150.

References viii

Liu, W. and Vinter, B. (2014). An efficient gpu general sparse matrix-matrix multiplication for irregular data. In *2014 IEEE 28th International Parallel and Distributed Processing Symposium*, pages 370–381.

Lu, H., Karimireddy, S. P., Ponomareva, N., and Mirrokni, V. (2020). Accelerating gradient boosting machines. In *International Conference on Artificial Intelligence and Statistics*, pages 516–526. PMLR.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Luo, C., Islam, M., Sheils, N. E., Buresh, J., Reps, J., Schuemie, M. J., Ryan, P. B., Edmondson, M., Duan, R., Tong, J., et al. (2022). Dlmm as a lossless one-shot algorithm for collaborative multi-site distributed linear mixed models. *Nature Communications*, 13(1):1–10.

Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. In Singh, A. and Zhu, J., editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR.

Mohassel, P. and Zhang, Y. (2017). Secureml: A system for scalable privacy-preserving machine learning. In *2017 IEEE symposium on security and privacy (SP)*, pages 19–38. IEEE.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

References ix

Nesterov, Y. (1983). A method for solving the convex programming problem with convergence rate $O(1/k^2)$.

Nori, H., Jenkins, S., Koch, P., and Caruana, R. (2019). Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*.

Pepe, M. S. (2000). An interpretation for the roc curve and inference using glm procedures. *Biometrics*, 56(2):352–359.

Pepe, M. S. (2003). The statistical evaluation of medical tests for classification and prediction. *Journal of the American Statistical Association*.

Pfisterer, F. (2022). *Democratizing Machine Learning – Contributions in AutoML and Fairness*. PhD thesis, LMU Munich.

Pfisterer, F., Thomas, J., and Bischl, B. (2019). Towards human centered automl. *arXiv preprint arXiv:1911.02391*.

Prasser, F., Kohlbacher, O., Mansmann, U., Bauer, B., and Kuhn, K. A. (2018). Data integration for future medicine (difuture). *Methods Inf Med*, 57(S01):e57–e65.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). ”why should i trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

References x

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Rügamer, D., Brockhaus, S., Gentsch, K., Scherer, K., and Greven, S. (2018). Boosting factor-specific functional historical models for the detection of synchronization in bioelectrical signals. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(3):621–642.

Saintigny, P., Zhang, L., Fan, Y.-H., El-Naggar, A. K., Papadimitrakopoulou, V. A., Feng, L., Lee, J. J., Kim, E. S., Hong, W. K., and Mao, L. (2011). Gene expression profiling predicts the development of oral cancer. *Cancer Prevention Research*, 4(2):218–229.

Samarati, P. and Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.

Sanderson, C. and Curtin, R. (2016). Armadillo: a template-based c++ library for linear algebra. *Journal of Open Source Software*, 1(2):26.

Sanderson, C. and Curtin, R. (2018). A user-friendly hybrid sparse matrix class in c++. In *International Congress on Mathematical Software*, pages 422–430. Springer.

Schalk, D., Bischl, B., and Rügamer, D. (2022a). Accelerated componentwise gradient boosting using efficient data representation and momentum-based optimization. *Journal of Computational and Graphical Statistics*.

Schalk, D., Bischl, B., and Rügamer, D. (2022b). Privacy-preserving and lossless distributed estimation of high-dimensional generalized additive mixed models. *arXiv preprint arXiv:2210.07723. Currently under review in the Journal of Computational and Graphical Statistics*.

References xi

Schalk, D., Hoffmann, V. S., Bischl, B., and Mansmann, U. (2022c). Distributed non-disclosive validation of predictive models by a modified roc-glm. *arXiv preprint arXiv:2203.10828*.

Schalk, D., Hoffmann, V. S., Bischl, B., and Mansmann, U. (2022d). dsBinVal: Conducting distributed ROC analysis using DataSHIELD. *Currently under review in the Journal of Open Source Software,* *`github.com/openjournals/joss-reviews/issues/4545`*.

Schalk, D., Thomas, J., and Bischl, B. (2018). compboost: Modular framework for component-wise boosting. *Journal of Open Source Software*, 3(30):967.

Schmid, M. and Hothorn, T. (2008). Boosting additive models using component-wise p-splines. *Computational Statistics & Data Analysis*, 53(2):298–311.

Shahnaz, R., Usman, A., and Chughtai, I. R. (2005). Review of storage techniques for sparse matrices. In *2005 Pakistan Section Multitopic Conference*, pages 1–7.

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment*, 62(1):77–89.

Sun, X., Zhang, P., Liu, J. K., Yu, J., and Xie, W. (2020). Private machine learning classification based on fully homomorphic encryption. *IEEE Transactions on Emerging Topics in Computing*, 8(2):352–364.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570.

Thomas, J., Coors, S., and Bischl, B. (2018). Automatic gradient boosting. *ICML AutoML Workshop*.

Thomas, J., Hepp, T., Mayr, A., and Bischl, B. (2017). Probing for sparse and fast variable selection with model-based boosting. *Computational and mathematical methods in medicine*, 2017.

References xii

Thornton, C., Hutter, F., Hoos, H. H., and Leyton-Brown, K. (2013). Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 847–855.

Tutz, G. and Gertheiss, J. (2016). Regularized regression for categorical data. *Statistical Modelling*, 16(3):161–200.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

Verbraeken, J., Wolting, M., Katzy, J., Kloppenburg, J., Verbelen, T., and Rellermeyer, J. S. (2020). A survey on distributed machine learning. *Acm computing surveys (csur)*, 53(2):1–33.

Vuk, M. and Curk, T. (2006). Roc curve, lift chart and calibration plot. *Metodoloski zvezki*, 3(1):89.

Wang, Q. and Kurz, D. (2022). Reconstructing training data from diverse ml models by ensemble inversion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2909–2917.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

Wood, S. N., Li, Z., Shaddick, G., and Augustin, N. H. (2017). Generalized additive models for gigadata: Modeling the u.k. black smoke network daily data. *Journal of the American Statistical Association*, 112(519):1199–1210.

Xanthopoulos, I., Tsamardinos, I., Christophides, V., Simon, E., and Salinger, A. (2020). Putting the human back in the automl loop. In *EDBT/ICDT Workshops*.

Yan, Z., Zachrison, K. S., Schwamm, L. H., Estrada, J. J., and Duan, R. (2022). Fed-glmm: A privacy-preserving and computation-efficient federated algorithm for generalized linear mixed models to analyze correlated electronic health records data. *medRxiv*.

References xiii

Zhu, R., Jiang, C., Wang, X., Wang, S., Zheng, H., and Tang, H. (2020). Privacy-preserving construction of generalized linear mixed model for biomedical computation. *Bioinformatics*, 36(Supplement_1):i128–i135.

Backup slides

Backup

bla