LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

Munich Center for Machine Learning

A re-examination of inference methods for the Generalization Error

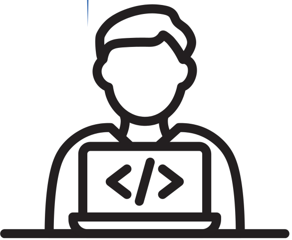
Hannah Schulz-Kümpel, Sebastian Fischer

September 13th, 2023

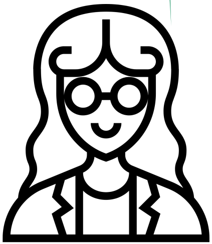
Statistische Woche 2023 | Methodology of Statistical Surveys (Session 8)

Hey, is there a quantity that quantifies how accurately a model that we fit will predict on new data that we haven't seen yet?

Sure Thing!



Practitioners



Theoreticians

Definition of Generalization error I

Given

- A *sequence of observations* $\mathcal{D} = \left\{ \left(\mathrm{x}^{(1)}, y^{(1)}\right), \dots, \left(\mathrm{x}^{(n)}, y^{(n)}\right) \right\}$ with $\left(\mathrm{x}^{(i)}, y^{(i)}\right) \in \mathcal{X} \times \mathcal{Y} \quad \forall i \in \{1, \dots, n\}$ and each $\left(\mathrm{x}^{(i)}, y^{(i)}\right)$ being an independent draw from a distribution $\mathbb{P}_{xy}$, i.e. the sequence of observations $\mathcal{D}$ is a realization of a random matrix $\boldsymbol{\mathcal{D}} \sim \bigotimes_{i=1}^{n} \mathbb{P}_{xy}$.

- A *point predictor* $\hat{f}_{\mathcal{I},\mathcal{D}} : \mathcal{X} \longrightarrow \mathbb{R}, \quad x \longmapsto \hat{f}_{\mathcal{I},\mathcal{D}}(x)$ where $\mathcal{I}$ denotes the "algorithm" (such as logistic regression) fit on $\mathcal{D}$, which we refer to as **inducer**.

- A *loss function* $L : \mathcal{Y} \times \mathbb{R} \longrightarrow \mathbb{R}$

Definition of Generalization error II

Generalization error may be used as an umbrella term for the following two quantities:

$$\textbf{prediction error (PE)} : \qquad \mathbb{E}[L(\boldsymbol{y}^*, \hat{f}_{\mathcal{I},\mathcal{D}}(\boldsymbol{x}^*))|\boldsymbol{\mathcal{D}} = \mathcal{D}]$$

$$\textbf{expected prediction error (ePE)} : \qquad \mathbb{E}\left[\mathbb{E}[L(\boldsymbol{y}^*, \hat{f}_{\mathcal{I},\boldsymbol{\mathcal{D}}}(\boldsymbol{x}^*))|\boldsymbol{\mathcal{D}}]\right],$$

with $(\boldsymbol{x}^*, \boldsymbol{y}^*) \sim \mathbb{P}_{xy}$ a *random variable* distributed according to the same distribution as every observation in $\mathcal{D}$.

Definition of Generalization error II

We are interested in point estimates and confidence intervals for one of the following quantities, for which **generalization error** may be used as an umbrella term.

The expected pointwise loss given a specific data set:

how well suited a specific model that has been fit on a specific data set will be on average for predictions on data stemming from the same data generating process as said data set.

prediction error (PE) : $\mathbb{E}[L(\boldsymbol{y}^*, \hat{f}_{\mathcal{I},\mathcal{D}}(\boldsymbol{x}^*))|\boldsymbol{\mathcal{D}} = \mathcal{D}]$

expected prediction error (ePE) : $\mathbb{E}\left[\mathbb{E}[L(\boldsymbol{y}^*, \hat{f}_{\mathcal{I},\mathcal{D}}(\boldsymbol{x}^*))|\boldsymbol{\mathcal{D}}]\right]$,

with $(\boldsymbol{x}^*, \boldsymbol{y}^*) \sim \mathbb{P}_{xy}$ a *random variable* distributed according to the same distribution as every observation in $\mathcal{D}$.

Definition of Generalization error II

We are interested in point estimates and confidence intervals for one of the following quantities, for which **generalization error** may be used as an umbrella term.

The expected pointwise loss given a specific data set:

how well suited a specific model that has been fit on a specific data set will be on average for predictions on data stemming from the same data generating process as said data set.

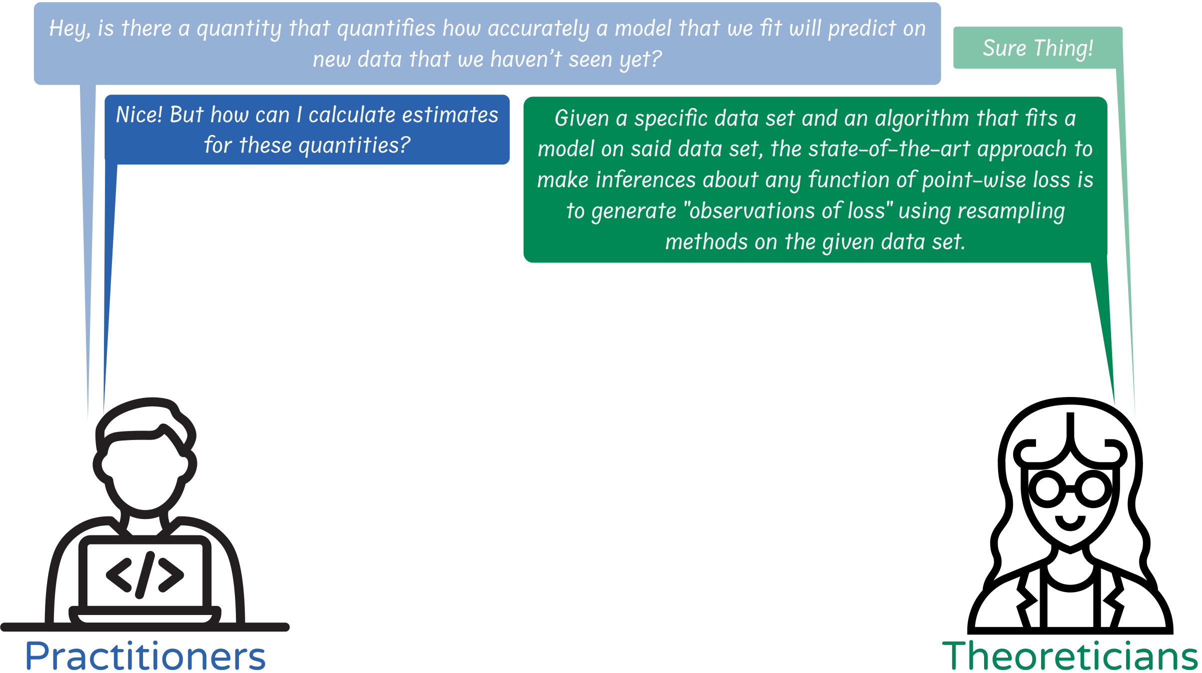
prediction error (PE) : $\mathbb{E}[L(\boldsymbol{y}^*, \hat{f}_{\mathcal{I},\mathcal{D}}(\boldsymbol{x}^*))|\boldsymbol{\mathcal{D}} = \mathcal{D}]$

expected prediction error (ePE) : $\mathbb{E}\left[\mathbb{E}[L(\boldsymbol{y}^*, \hat{f}_{\mathcal{I},\mathcal{D}}(\boldsymbol{x}^*))|\boldsymbol{\mathcal{D}}]\right]$,

The expectation of above conditional expected pointwise loss over all data sets of the same size:

how well suited models that have been fit using a certain algorithm on data sets of size n will be on average for predictions on data stemming from the same data generating process.

with $(\boldsymbol{x}^*, \boldsymbol{y}^*) \sim \mathbb{P}_{xy}$ a *random variable* distributed according to the same distribution as every observation in $\mathcal{D}$.


Hey, is there a quantity that quantifies how accurately a model that we fit will predict on new data that we haven't seen yet?
Sure Thing!
Nice! But how can I calculate estimates for these quantities?
Given a specific data set and an algorithm that fits a model on said data set, the state-of-the-art approach to make inferences about any function of point-wise loss is to generate "observations of loss" using resampling methods on the given data set.
Practitioners
Theoreticians

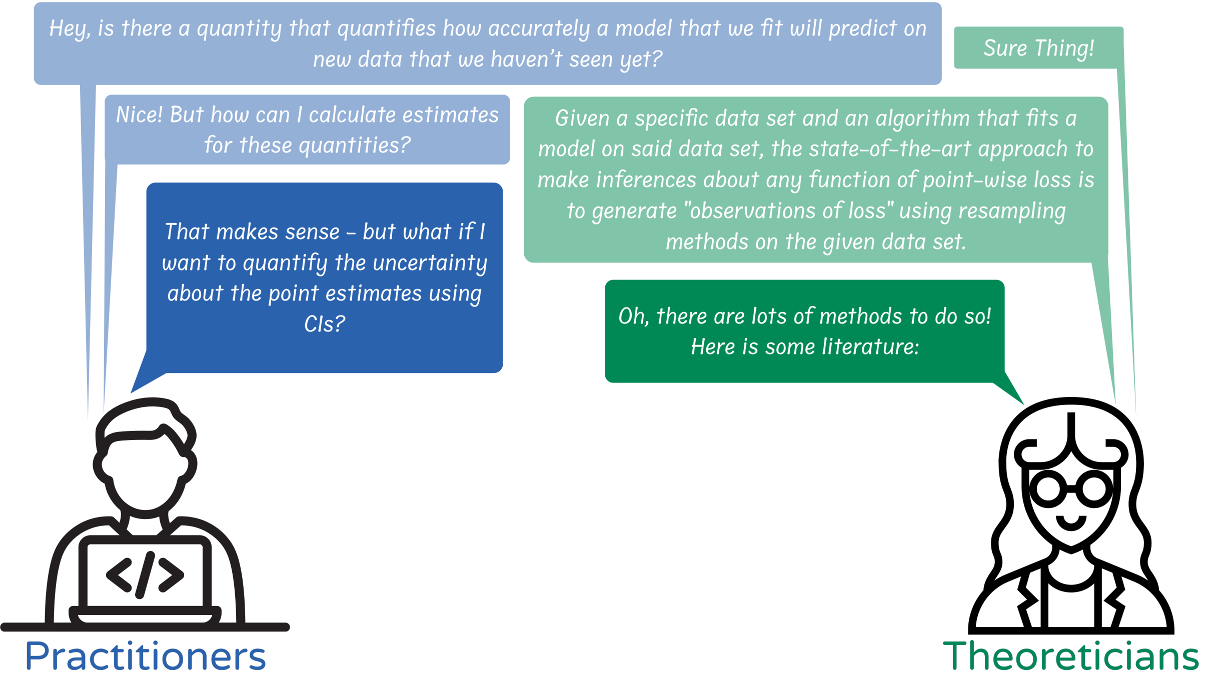
Inference based on resampling I

- With resampling, there are two issues in need of addressing:
    - Any standard point estimate for the generalization error based on refitting an algorithm on resampled data will be more appropriate for the ePE than for the PE (*weak correlation*)
    - Any resampling creates dependence structures in our inference data, with the exact structure depending on the resampling method.

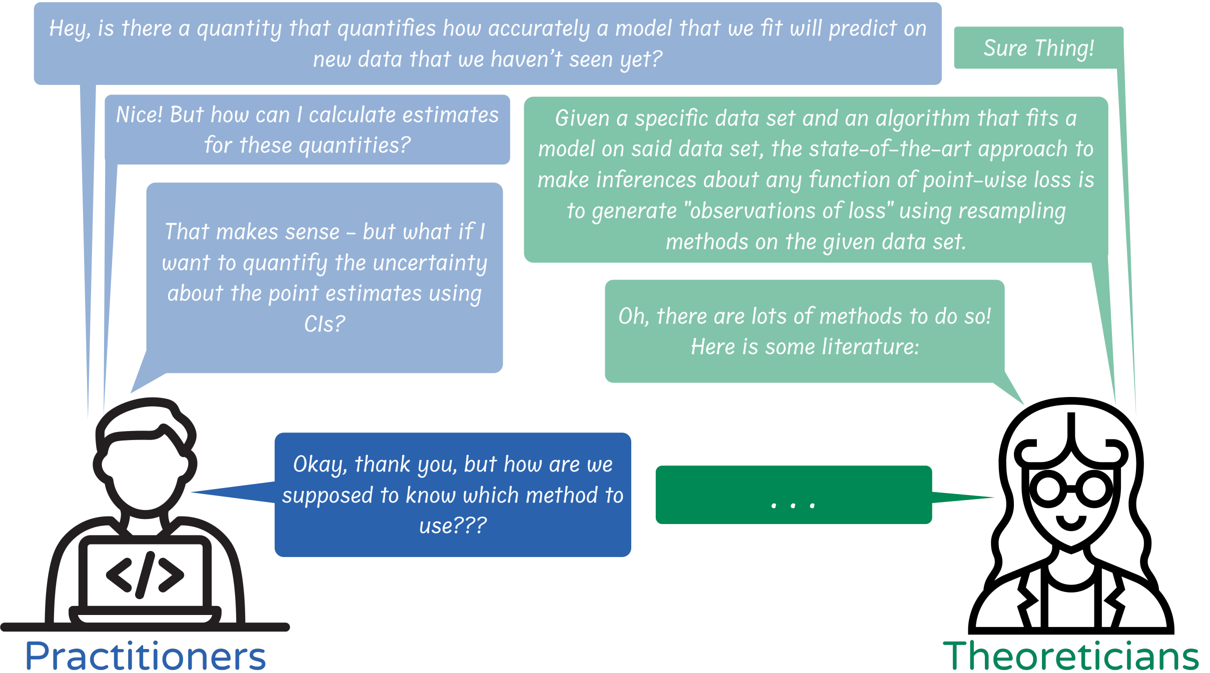
Inference based on resampling II

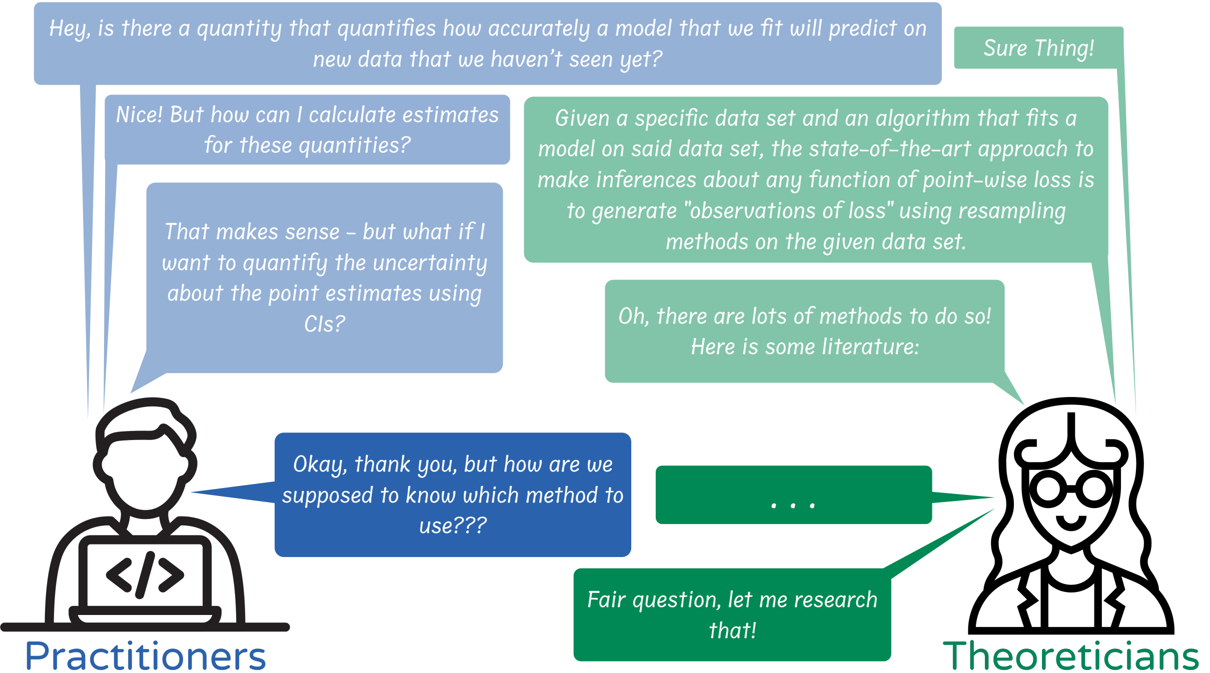
- **When one is only interested in point estimates**, these issues are negligible because
  - As ePE is in some sense the expectation of PE, one could argue that any point estimate of ePE may serve as a valid, if less accurate, point estimate of PE.
  - Due to the properties of the expected value, the dependence structures of the "loss-observations" do not problematically affect the common point estimates for the ePE

    - at most, we are conditioning on a slightly smaller data set, e.g. in CV.


Hey, is there a quantity that quantifies how accurately a model that we fit will predict on new data that we haven't seen yet?
Sure Thing!
Nice! But how can I calculate estimates for these quantities?
Given a specific data set and an algorithm that fits a model on said data set, the state-of-the-art approach to make inferences about any function of point-wise loss is to generate "observations of loss" using resampling methods on the given data set.
That makes sense – but what if I want to quantify the uncertainty about the point estimates using CIs?
Oh, there are lots of methods to do so! Here is some literature:
Practitioners
Theoreticians


Bootstrap methods
at the very least Jiang, Varma, and Simon (2008) as well as the .632+ Bootstrap Method from Efron and Tibshirani (1997)
Practitioners
Bayle (2020): two different variance estimators
Bates, Hastie, and Tibshirani (2021): Nested CV
The standard holdout estimator
CV based methods
Naive Estimator as implemented in glmnet (Breiman et al. (1984))
Repeated Subsampling
Austern and Zhou (2020)
Dietterich (1998): 2 × 5 Cross-Validation
Nadeau and Bengio (2003): Corrected t-test and conservative z-test


Hey, is there a quantity that quantifies how accurately a model that we fit will predict on new data that we haven't seen yet?
Sure Thing!
Nice! But how can I calculate estimates for these quantities?
Given a specific data set and an algorithm that fits a model on said data set, the state-of-the-art approach to make inferences about any function of point-wise loss is to generate "observations of loss" using resampling methods on the given data set.
That makes sense – but what if I want to quantify the uncertainty about the point estimates using CIs?
Oh, there are lots of methods to do so! Here is some literature:
Okay, thank you, but how are we supposed to know which method to use???
. . .
Practitioners
Theoreticians


Hey, is there a quantity that quantifies how accurately a model that we fit will predict on new data that we haven't seen yet?
Sure Thing!
Nice! But how can I calculate estimates for these quantities?
Given a specific data set and an algorithm that fits a model on said data set, the state-of-the-art approach to make inferences about any function of point-wise loss is to generate "observations of loss" using resampling methods on the given data set.
That makes sense - but what if I want to quantify the uncertainty about the point estimates using CIs?
Oh, there are lots of methods to do so! Here is some literature:
Okay, thank you, but how are we supposed to know which method to use???
. . .
Fair question, let me research that!
Practitioners
Theoreticians

Only the following 2 CIs are asymptotically exact!

Bootstrap methods

at the very least Jiang, Varma, and Simon (2008) as well as the .632+ Bootstrap Method from Efron and

Practitioners

Bayle (2020): two different variance estimators

Bates, Hastie, and Tibshirani (2021): Nested CV

CV based methods

The standard holdout estimator

Naive Estimator as implemented in glmnet (Breiman et al.

Repeated Subsampling

Austern and Zhou (2020)

Nadeau and Bengio (2003): Corrected t-test and conservative z-

Dietterich (1998): 2 × 5 Cross-Validation

The considered methods fall into two categories

Category A Those where asymptotical exactness of the CI w.r.t. a clearly defined proxy quantity was either proven by the original authors or could be verified by us.

- *"proxy quantity"* $\hat{=}$ A quantity that is neither PE nor ePE, but seems close to it.
- *For example, in Bayle (2020):*
  $n^{-1}\sum_{i=1}^{k}\sum_{j=1}^{n_{\text{test},i}}\mathbb{E}\left[L(y^{(J_{\text{test},i}[j])},\hat{f}_{\mathcal{I},\mathcal{D}_{\textbf{train},i}}(x^{(J_{\text{test},i}[j])}))|\mathcal{D}_{\text{train},i}\right]$

The considered methods fall into two categories

Category A Those where asymptotical exactness of the CI w.r.t. a clearly defined proxy quantity was either proven by the original authors or could be verified by us.

- *"proxy quantity"* $\hat{=}$ A quantity that is neither PE nor ePE, but seems close to it.
- *For example, in Bayle (2020):*
  $n^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_{\mathrm{test},i}} \mathbb{E}\Big[L(y^{(J_{\mathrm{test},i}[j])}, \hat{f}_{\mathcal{I},\mathcal{D}_{\mathbf{train},i}}(\mathrm{x}^{(J_{\mathrm{test},i}[j])}))|\mathcal{D}_{\mathrm{train},i}\Big]$

Category B Those where a "confidence interval" is constructed using, at least partially, contextual reasoning instead of asymptotics - often we are still able to identify a proxy quantity or at least verify whether the CI was intended to cover PE or ePE.

Given this discovery and the lack of neutral empirical comparison...

... we decided on a research plan consisting of the following elements:

1. Conceptual and Theoretical:
   - A comprehensive explanation of the topic "Confidence intervals for the generalization error" directed towards the broadest possible audience.
   - Formally examining the dependence structures in different resampling methods.
   - Working out the underlying assumptions for existing methods.
   - Identifying proxy quantities and verifying asymptotic exactness via formal proofs.

Given this discovery and the lack of neutral empirical comparison...

... we decided on a research plan consisting of the following elements:

2. Empirical:
   - Provide point estimates and CIs for the coverage frequency as well as quantifications of several other performance measures for a variety of "*CI for the GE*" methods, based on their application on a wide variety of settings **(main experiment)**. *Importantly, we conduct this study from a neutral standpoint*.
   - Perform an **exploratory analysis** of
     a. The distance between established proxy quantities and (e)PE, where applicable.
     b. The performance variability caused by e.g. the choice of inducer or data size.
   - Throughout, provide insights, guidelines, and interim results intended for the use in follow-up analyses, or generally further studies in the field.

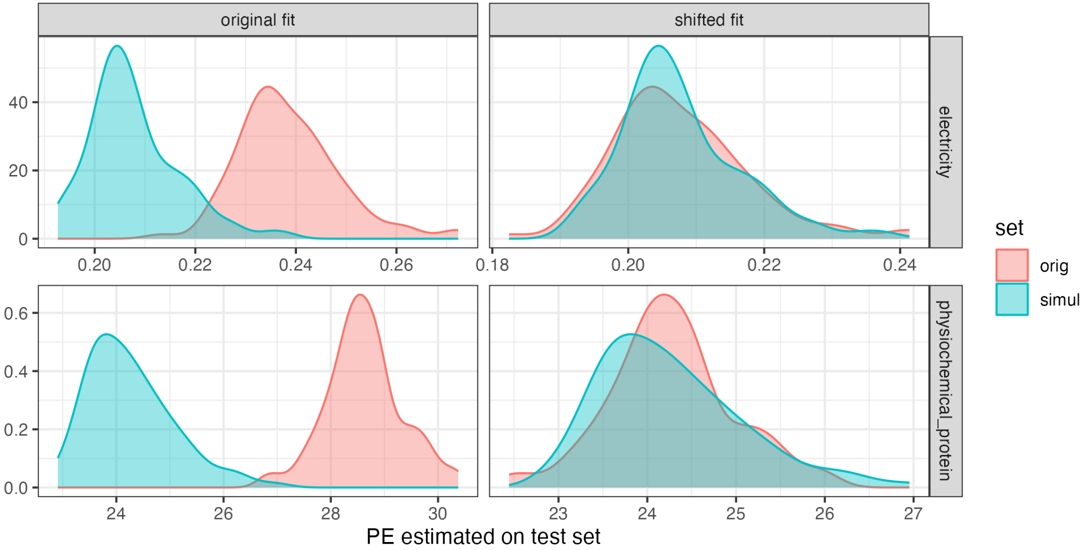
Some visualizations for two specific data sets

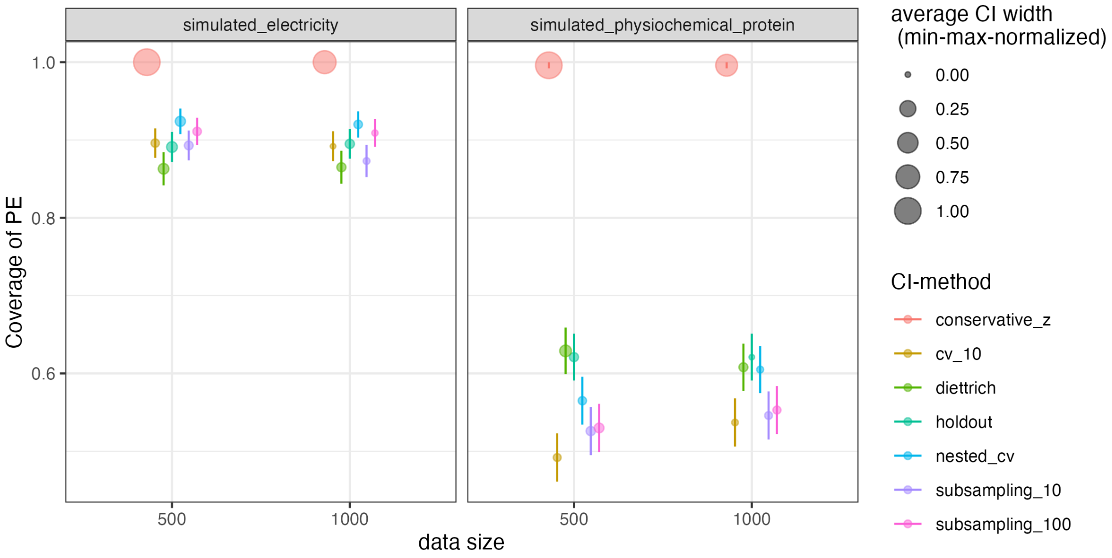
For the data sets

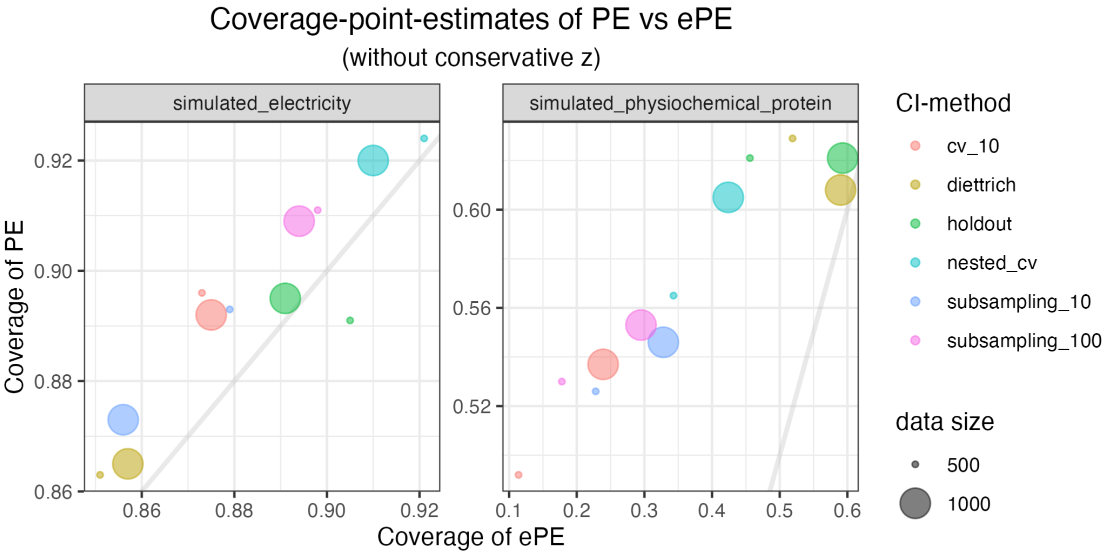
`Electricity` (suitable for logistic regression classification) and

`Physiochemical protein` (suitable for linear regression)

1. Estimate density of $\mathbb{P}_{xy}$ using LLM [method from Borisov et al. (2022)] and use it as data generating process (DGP)
2. Compare point estimates and CIs of the (e)PE coverage between the different methods
3. Compare the different Methods regarding coverage of PE **and** ePE


original fit
shifted fit
electricity
physiochemical_protein
set
orig
simul
PE estimated on test set
40
20
0
0.20
0.22
0.24
0.26
0.18
0.6
0.4
0.2
0.0
24
26
28
30
23
25
27


simulated_electricity
simulated_physiochemical_protein
1.0
0.8
0.6
Coverage of PE
500
1000
500
1000
data size
average CI width
(min-max-normalized)
0.00
0.25
0.50
0.75
1.00
CI-method
conservative_z
cv_10
diettrich
holdout
nested_cv
subsampling_10
subsampling_100


Coverage-point-estimates of PE vs ePE
(without conservative z)
simulated_electricity
simulated_physiochemical_protein
Coverage of PE
Coverage of ePE
0.92
0.90
0.88
0.86
0.86
0.88
0.90
0.92
0.60
0.56
0.52
0.1
0.2
0.3
0.4
0.5
0.6
CI-method
cv_10
diettrich
holdout
nested_cv
subsampling_10
subsampling_100
data size
500
1000

Outlook

- We are still in the process of completing our ambitious research project.
- The code-framework and computational resources for the empirical study are fully set up.
- We have already formally verified the first methods' proxy-quantities and generated several interesting hypotheses.
- Hopefully, at least some concrete guidelines for practitioners regarding which method to use will become known.

References

Austern, Morgane and Wenda Zhou (2020). "Asymptotics of cross-validation". In: *arXiv preprint arXiv:2001.11111*.

Bates, Stephen, Trevor Hastie, and Robert Tibshirani (Apr. 1, 2021). "Cross-Validation: What Does It Estimate and How Well Does It Do It?" In: arXiv: 2104.00673 [math, stat]. URL: http://arxiv.org/abs/2104.00673 (visited on 04/08/2021).

Bayle (2020). "Cross-validation Confidence Intervals for Test Error". In: DOI: 10.48550/ARXIV.2007.12671. URL: https://arxiv.org/abs/2007.12671.

Borisov, Vadim et al. (2022). "Language models are realistic tabular data generators". In: *arXiv preprint arXiv:2210.06280*.

Dietterich, Thomas G. (Oct. 1, 1998). "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms". In: *Neural Computation* 10.7, pp. 1895–1923. ISSN: 0899-7667. DOI: 10.1162/089976698300017197. URL: https://www.mitpressjournals.org/doi/10.1162/089976698300017197 (visited on 02/09/2021).

Efron, Bradley and Robert Tibshirani (1997). "Improvements on Cross-Validation: The .632+ Bootstrap Method". In: *Journal of the American Statistical Association* 92.438, pp. 548–560. ISSN: 01621459. URL: http://www.jstor.org/stable/2965703 (visited on 06/17/2023).

Friedman, Jerome, Robert Tibshirani, and Trevor Hastie (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1, pp. 1–22. DOI: 10.18637/jss.v033.i01.

Jiang, Wenyu, Sudhir Varma, and Richard Simon (2008). "Calculating confidence intervals for prediction error in microarray classification using resampling.". In: *Stat Appl Genet Mol Biol* 7.1, Article8.

Nadeau, Claude and Yoshua Bengio (Sept. 1, 2003). "Inference for the Generalization Error". In: *Machine Learning* 52.3, pp. 239–281. ISSN: 1573-0565. DOI: 10.1023/A:1024068626366. URL: https://doi.org/10.1023/A:1024068626366 (visited on 02/09/2021).