ETH zürich

Schweizer Paraplegiker Zentrum

Schweizer Paraplegiker Forschung

CAI LAB

Generalizability of ML Models: Bias-Variance Tradeoff, Model and Feature Selection

Yanke Li

2024.06

Outline for Topics Today

1. **Generalizability**
2. **Bias-Variance Trade-off**
   - Definition
   - Explanation
3. **Model Selection**
   - Overview of Techniques
   - Cross Validation
4. **Regularization**
   - Overview
5. **Feature Selection**
   - Overview

Generalizability of Machine Learning Models

Generalizability: ability of the model to perform well on unseen data

How can we empirically evaluate generalizability of trained machine learning (ML) model?

- Evaluation on a separated dataset (test data) which is different from the train data
- If it performs well on the separated dataset, the ML model is considered to be able to generalize.

What will influence the generalizability of ML models?

Bias-Variance Trade-off

- **Bias** measures the error due to erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).
- Mathematically, bias is defined for a model f estimating a target y as:

$$Bias(f) = E[(f(x) - E[y|x])]$$

- In simpler terms, it's the difference between the expected prediction of our model and the true output we try to predict.

- **Variance** measures how much the predictions for a given point vary between different realizations of the model. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting).
- Mathematically:

$$Variance(f) = E[(f(x) - E[f(x)])^2]$$

- This is the variability of model prediction for a given data point.

Explanation:

- The expected prediction error for any machine learning algorithm can be decomposed into three parts: Bias, Variance, and a noise term.
- For a regression problem, the expected mean squared error at a point x is given by:

$$E\left[\left(y - f(x)\right)^2\right] = Bias^2(f) + Variance(f) + \sigma^2$$

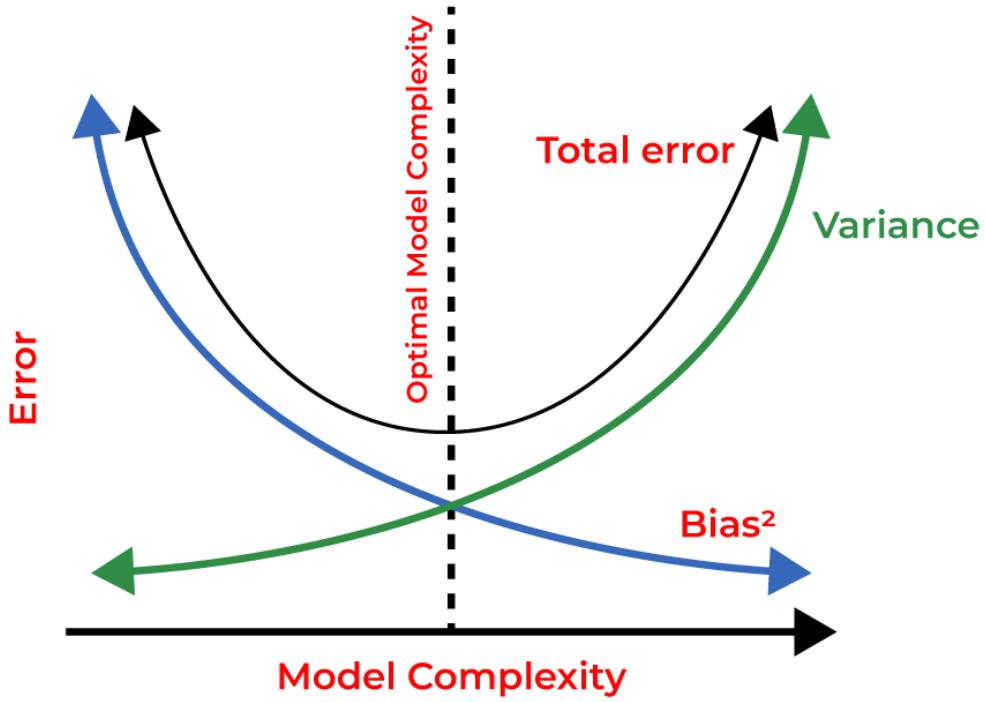
- where σ^2 represents the irreducible error inherent in the problem itself due to noise in the data.

Interpretation:

High Bias, Low Variance Models: pay little attention to the training data and oversimplify the model, leading to to underfitting. They have a high error on training data and test data.

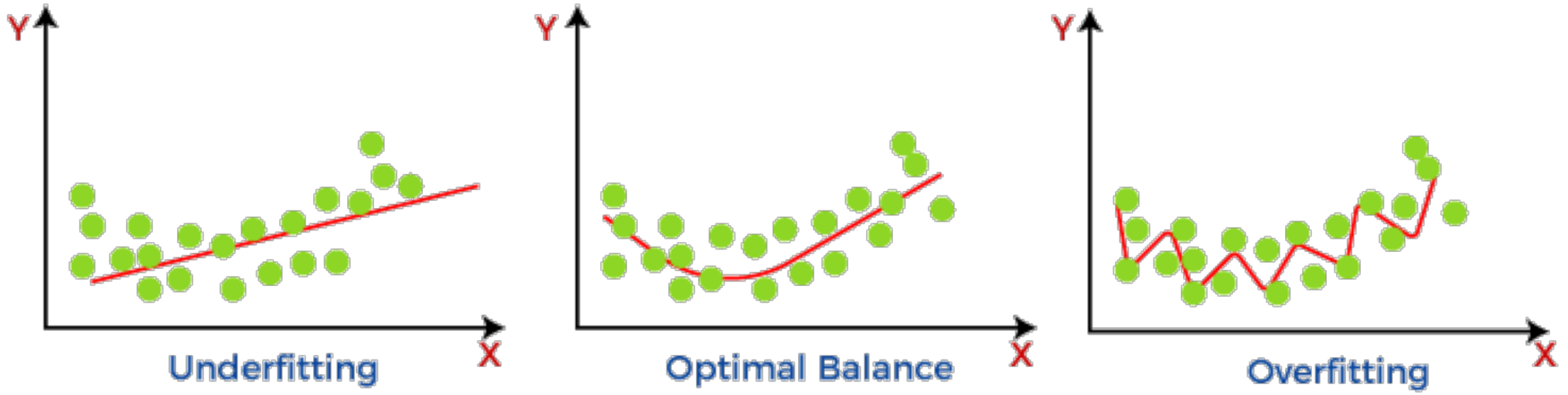
Low Bias, High Variance Models: pay too much attention to the training data and capture noise as well as the underlying data structure, leading to overfitting. They perform well on training data but poorly on any unseen test data.

Trade-off:
Aiming to find a balance between these two errors is the essence of the bias-variance tradeoff. We want a model that sufficiently captures the important trends (low bias) without fitting excessively to the training data noise (low variance).

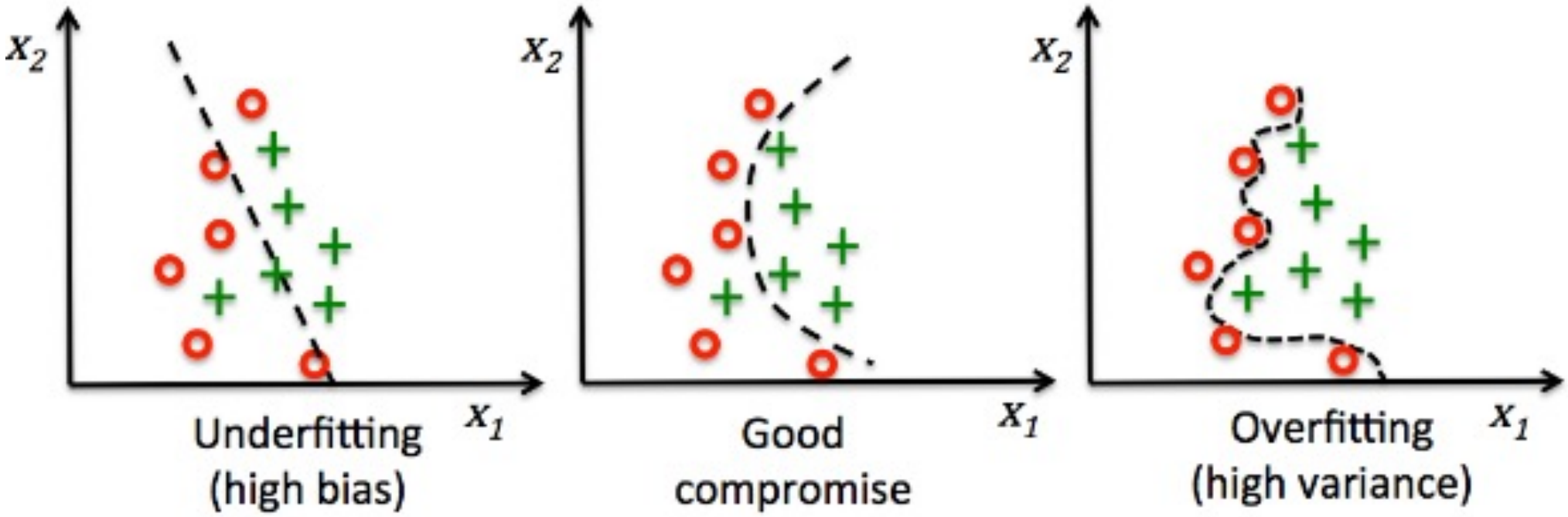


Bias-Variance Trade-off and Model Selection

- **Simple Models**: usually exhibit high bias and low variance. They might not capture complex patterns well, leading to underperformance (underfitting) even on the training data
- **Complex Models**: low bias because they are flexible enough to capture complex patterns in the data. However, they tend to have high variance, meaning their performance can degrade on unseen data due to overfitting.



Bias-Variance Trade-off and Model Selection



Techniques for Model Selection

Resampling Methods: seek to estimate the performance of a model (or more precisely, the model development process) on out-of-sample data.

- Random train/test split
- **Cross-Validation**
- Bootstrap

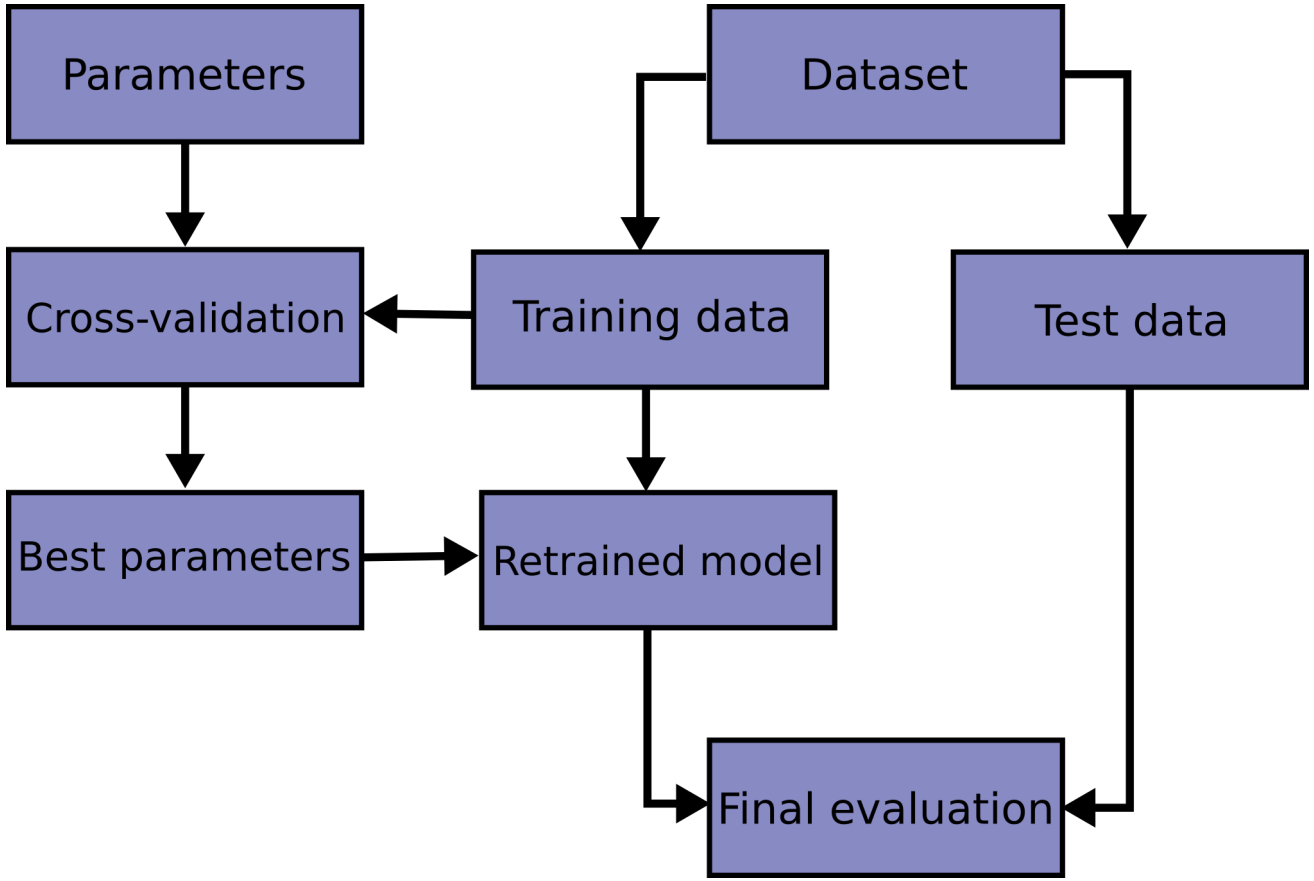
Probabilistic Methods: involve analytically scoring a candidate model using both its performance on the training dataset and the complexity of the model.

- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Minimum Description Length (MDL)
- Structural Risk Minimization (SRM)

Cross Validation

General Purpose - Finding the Best Hyperparameters

Often used to estimate a model's generalization performance. By training the model on several randomly split subsets of the dataset and validating it on other subsets, we can assess both bias and variance, helping to select a model that generalizes well to unseen data.



5-Fold Cross Validation

K-Fold Cross Validation: General Procedure

1. Shuffle the dataset randomly.
2. Split the dataset into k groups
3. For each unique group:
   - Take the group as a hold out or test data set
   - Take the remaining groups as a training data set
   - Fit a model on the training set and evaluate it on the test set
   - Retain the evaluation score and discard the model
4. Summarize the skill of the model using the sample of model evaluation scores

Importantly, each observation in the data sample is assigned to an individual group and stays in that group for the duration of the procedure. This means that each sample is given the opportunity to be used in the hold out set 1 time and used to train the model k-1 times.

Variations on Cross Validation

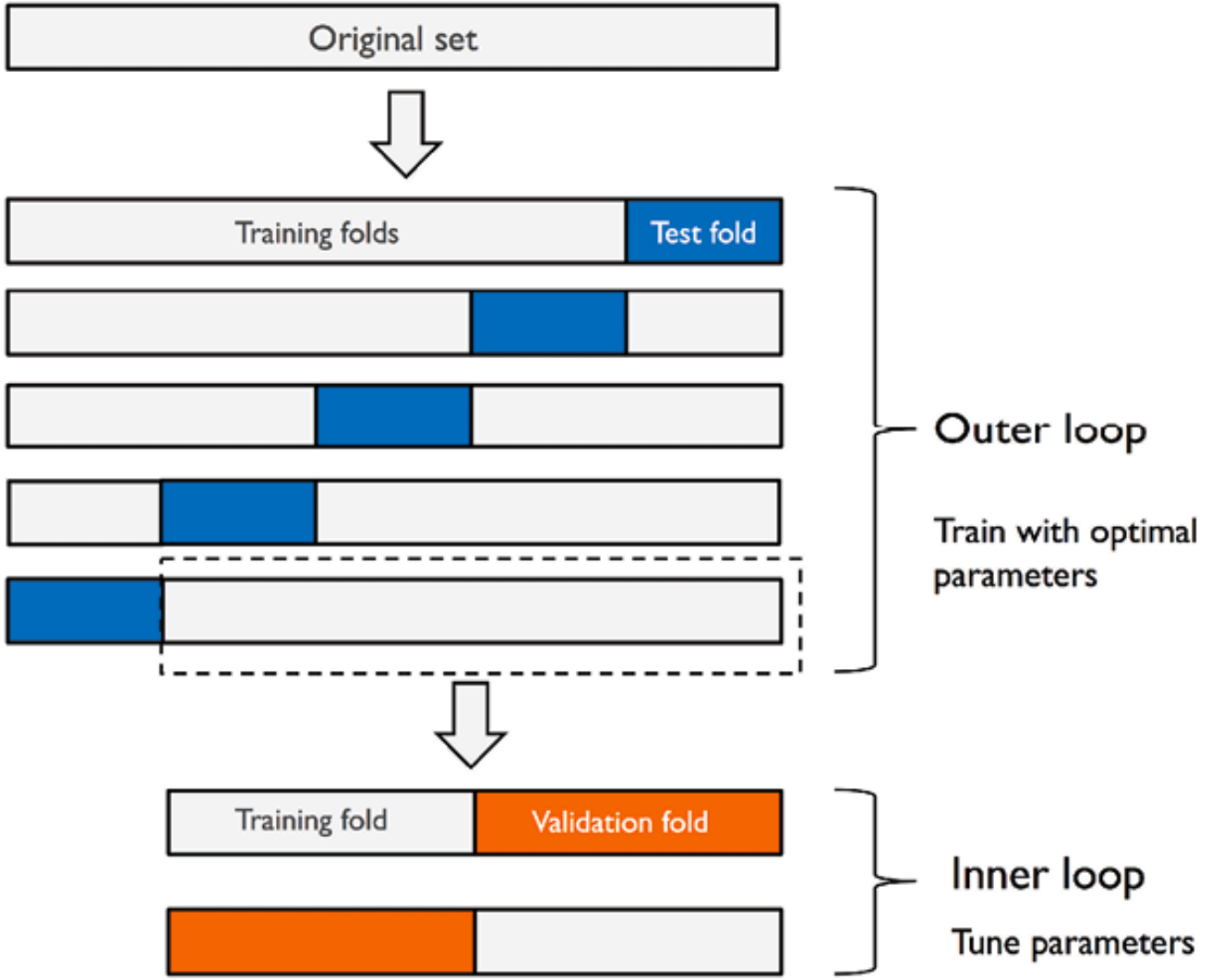
There are several variations on the k-fold cross validation procedure.

Train/Test Split: Taken to one extreme, k may be set to 2 such that a single train/test split is created to evaluate the model.
LOOCV: Taken to another extreme, k may be set to the total number of observations in the dataset such that each observation is given a chance to be the held out of the dataset. This is called leave-one-out cross-validation, or LOOCV for short.
Stratified: The splitting of data into folds may be governed by criteria such as ensuring that each fold has the same proportion of observations with a given categorical value, such as the class outcome value. This is called stratified cross-validation.
Repeated: This is where the k-fold cross-validation procedure is repeated n times, where importantly, the data sample is shuffled prior to each repetition, which results in a different split of the sample.
Nested: This is where k-fold cross-validation is performed within each fold of cross-validation, often to perform hyperparameter tuning during model evaluation. This is called nested cross-validation or double cross-validation.

Nested Cross Validation

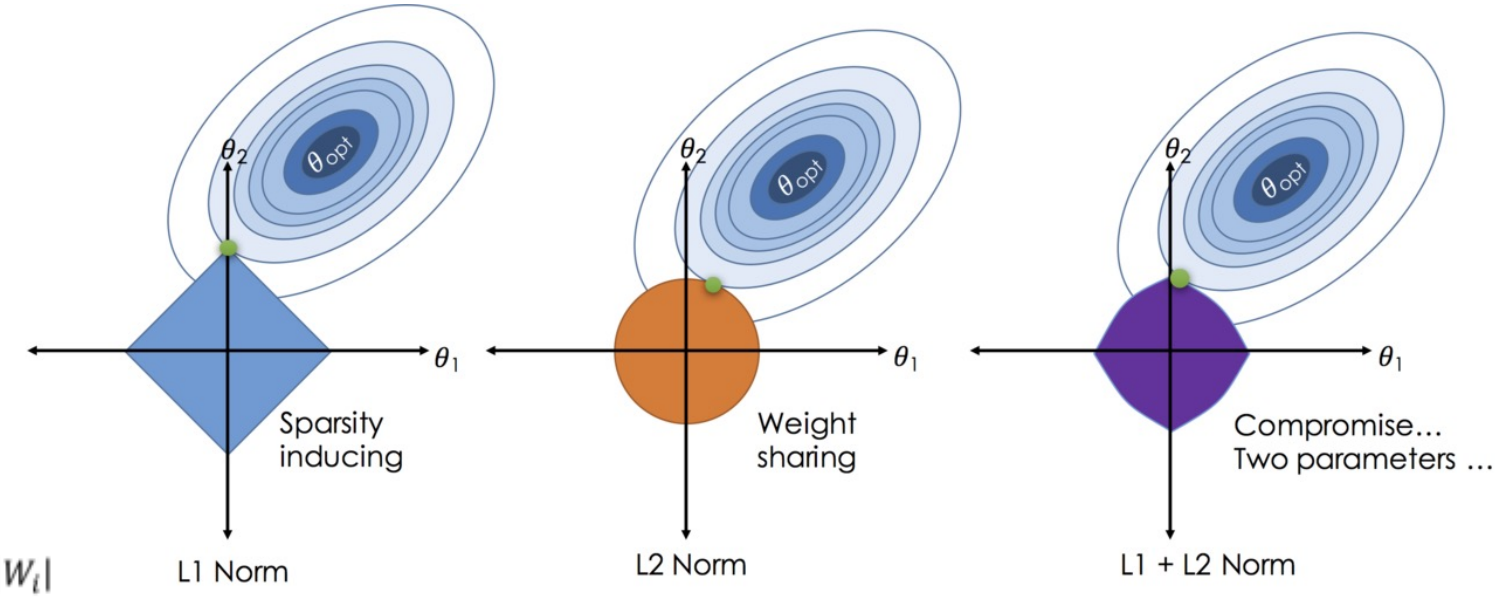
CV can be used both when optimizing the hyperparameters of a model on a dataset, and when comparing and selecting a model for the dataset. When the same cross-validation procedure and dataset are used to both tune and select a model, it is likely to lead to an optimistically biased evaluation of the model performance.

One approach to overcoming this bias is to nest the hyperparameter optimization procedure under the model selection procedure. This is called double cross-validation or nested cross-validation and is the preferred way to evaluate and compare tuned machine learning models.



Regularization

- L1-regularization: feature selection
- L2-regularization: weight decay



L1 Regularization

$$\text{Modified loss function} = \text{Loss function} + \lambda \sum_{i=1}^{n} |W_i|$$

L2 Regularization

$$\text{Modified loss function} = \text{Loss function} + \lambda \sum_{i=1}^{n} W_i^2$$

Techniques like L1 and L2 regularization are used to add a penalty on model complexity. Regularization can force a model with high variance to become simpler, thus reducing variance at the cost of increasing bias, which can be beneficial if the model is overfitting.

Feature Selection

In many high-dimensional problems, we may prefer not to work with all potentially available features.

Why?

- **Interpretability:** aim to understand the model or identify important features (risk factors)
- **Generalization:** simpler models may generalize better
- **Storage/computation/cost:** don't need to store or acquire data for unused features

How?

- Filter Methods
- Wrapper Methods
- Embedded Methods

Thank You!

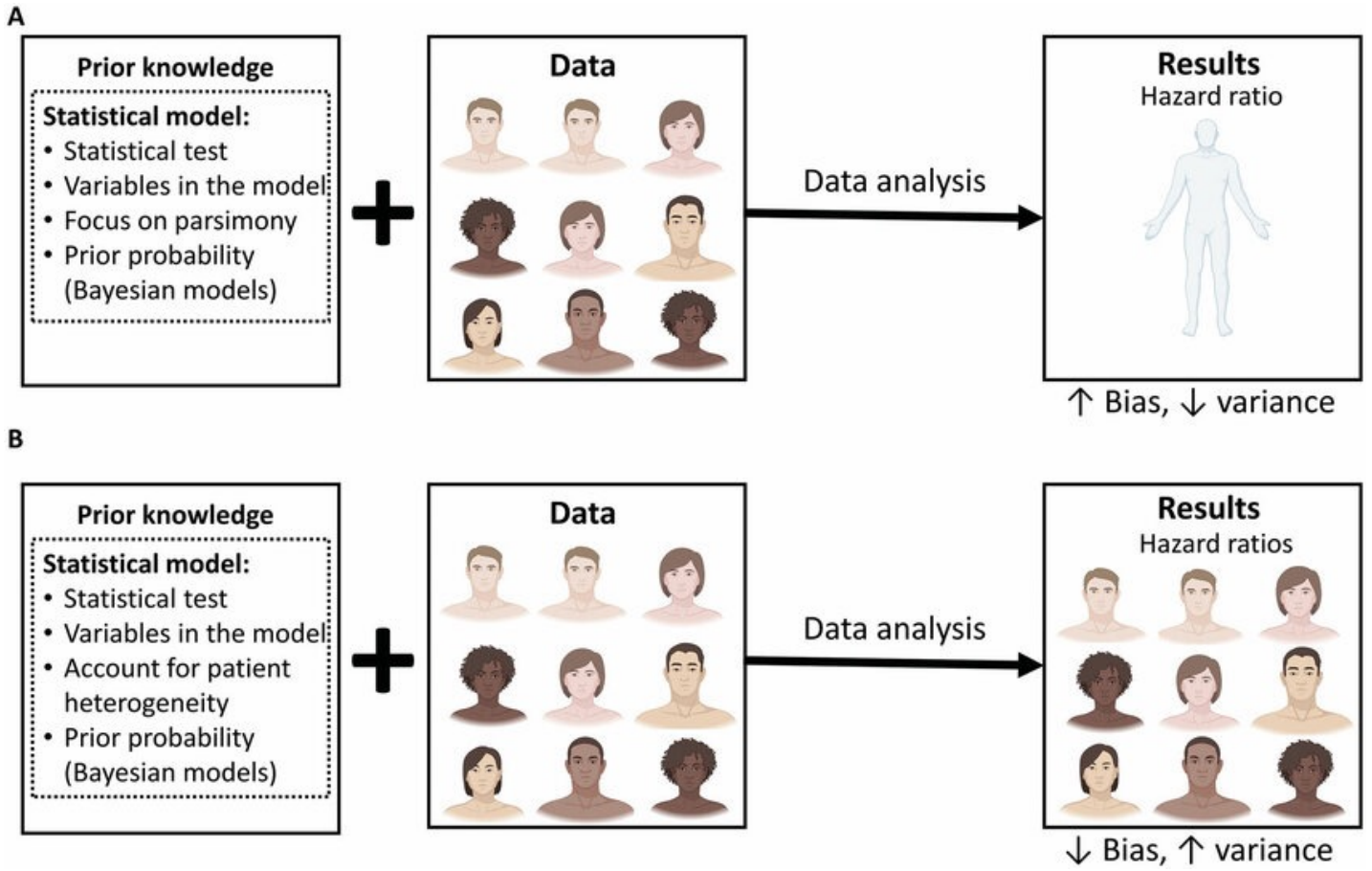
Yanke Li
SCAI PhD Researcher
yanke.li@hest.ethz.ch

SCAI Lab
Direktionssekretariat SPZ
Guido A. Zäch Strasse 1
6207 Nottwil
Switzerland

Sensory-Motor Systems Lab
Gloriastrasse 37/ 39
ETH GLC G19
8092 Zurich
Switzerland

Healthcare Domain-Specific Considerations (extend more pages)
(model and sampling, splitting)

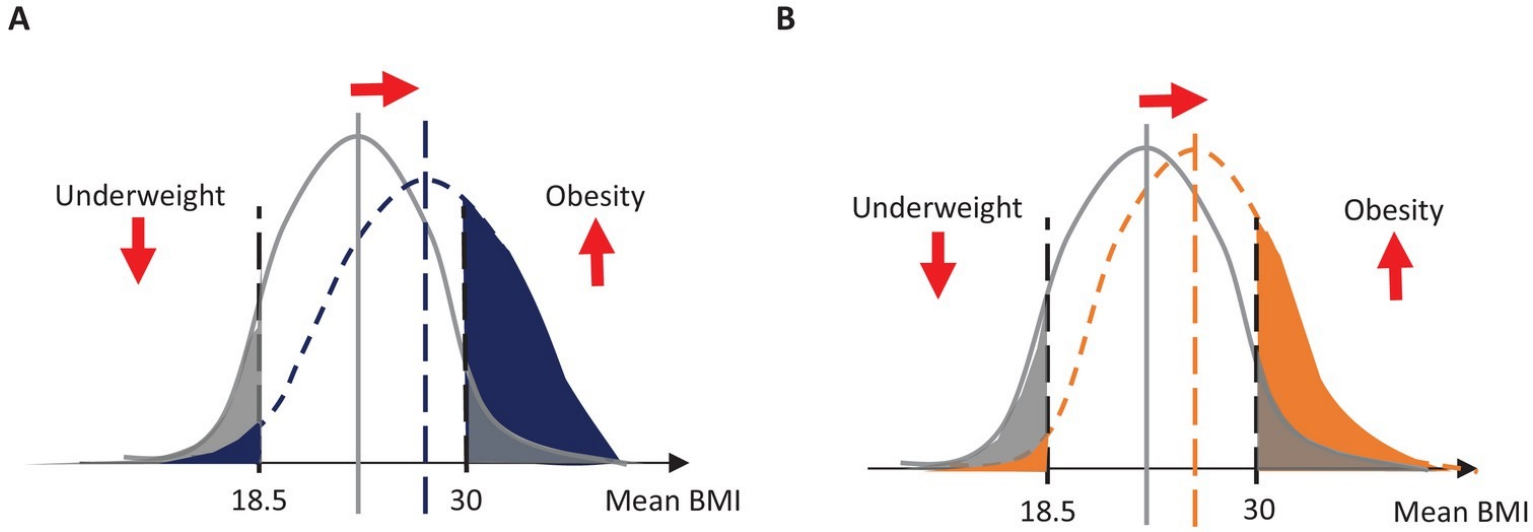
In fields like healthcare, where model reliability can critically impact outcomes, selecting a model with a careful consideration of the bias-variance trade-off is crucial. It's not only about achieving the highest possible accuracy on the training data but ensuring that predictions are robust and consistent under varying conditions and on new, unseen data.



Distribution Shift in Healthcare and Rehabilitation (not needed)

Distribution shift is a critical challenge in healthcare and rehabilitation where the model trained on one set of data (source domain) may not perform well when deployed in a different setting (target domain) due to changes in the underlying data distribution.

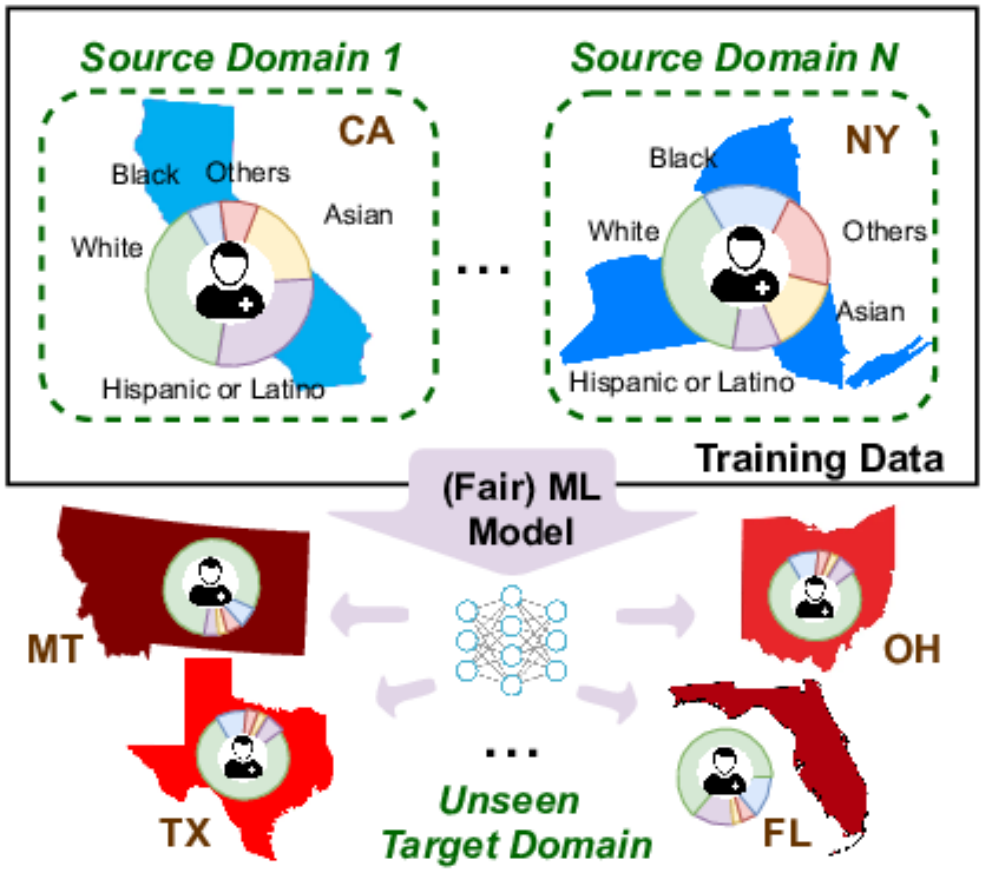
(more specific example in healthcare)



(A) Change in the prevalence of underweight and obesity if the distribution shifts, represented by a change in its mean and its shape. In this example, the change (shown as the difference between blue and gray) results in a small decrease of underweight and a large increase in obesity. (B) Change in the prevalence of underweight and obesity when only mean BMI changes (shown as the difference between orange and gray), without a change in the shape of the distribution.

NCD Risk Factor Collaboration, 2021

Domain Generalization in Healthcare (not needed)



ETH zürich

Schweizer Paraplegiker Zentrum

Schweizer Paraplegiker Forschung

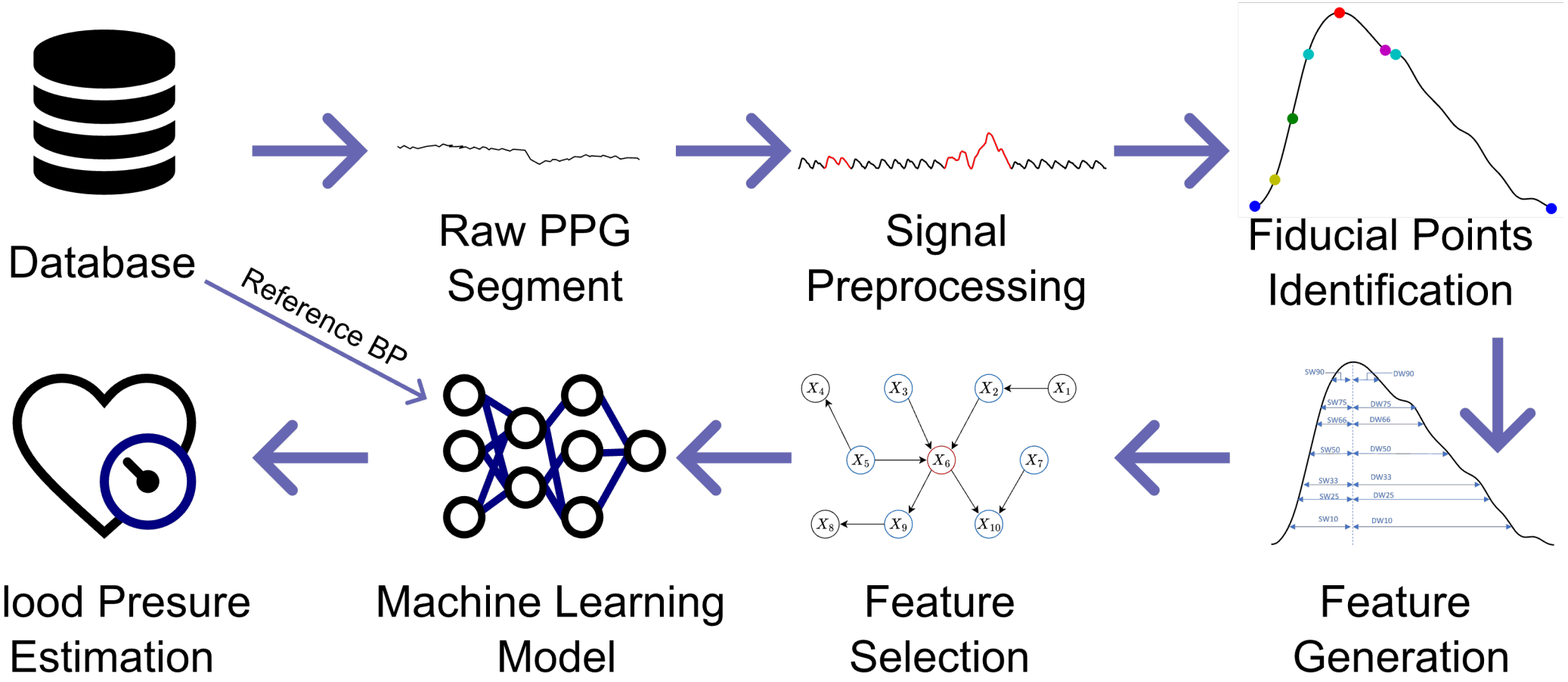
CAI LAB

Feature Selection Using Markov Blanket in Graphical Models

Yanke Li

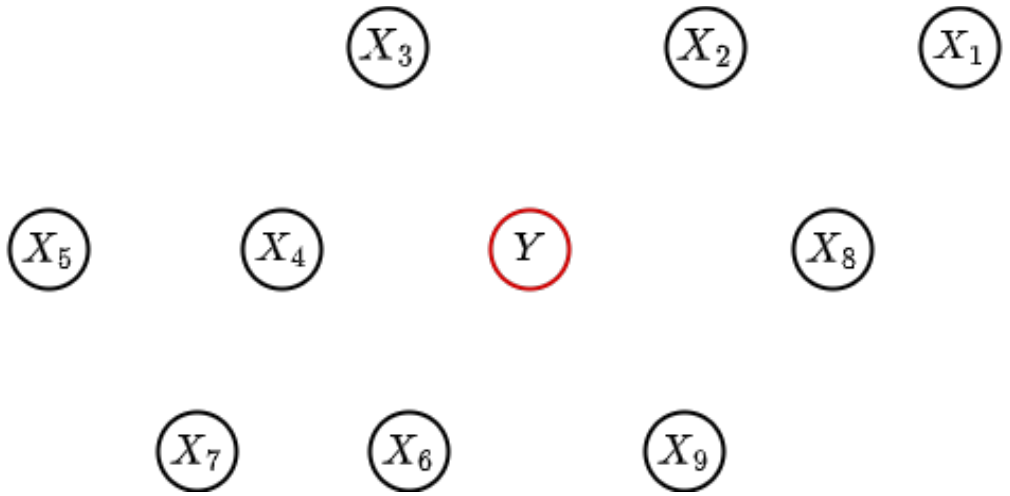
2024.06

Blood Pressure Estimation using Markov Blanket Feature Selection (include another paper about using causal graph for BP estimation: CiGNN)

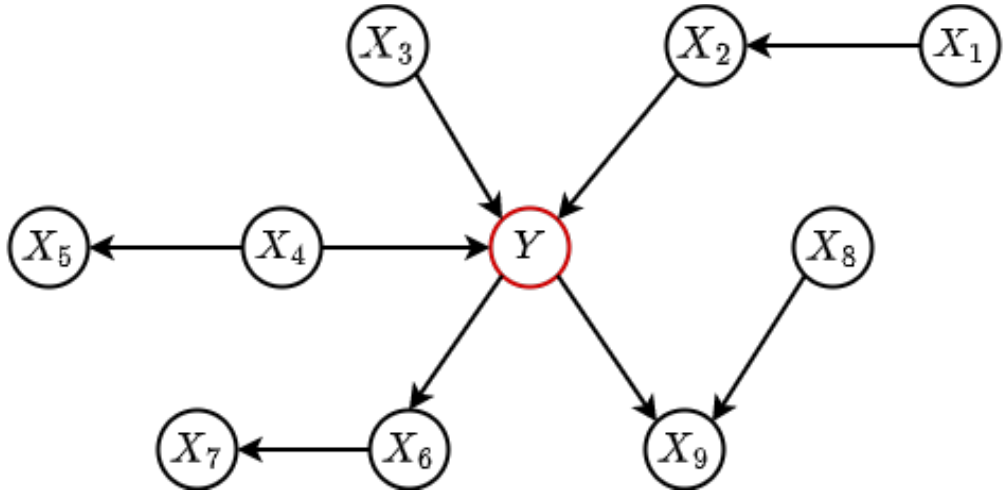


What is Markov Blanket in DAG?

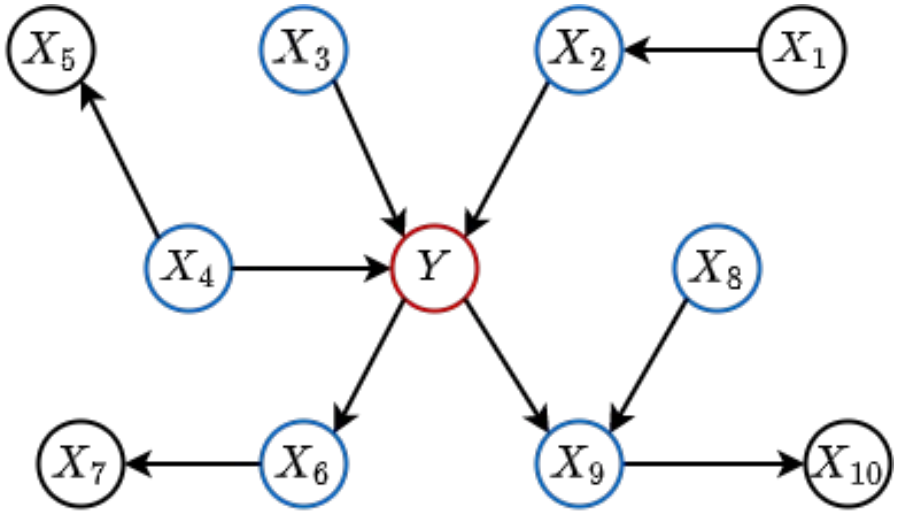
What is Markov Blanket in DAG?



What is Markov Blanket in DAG?



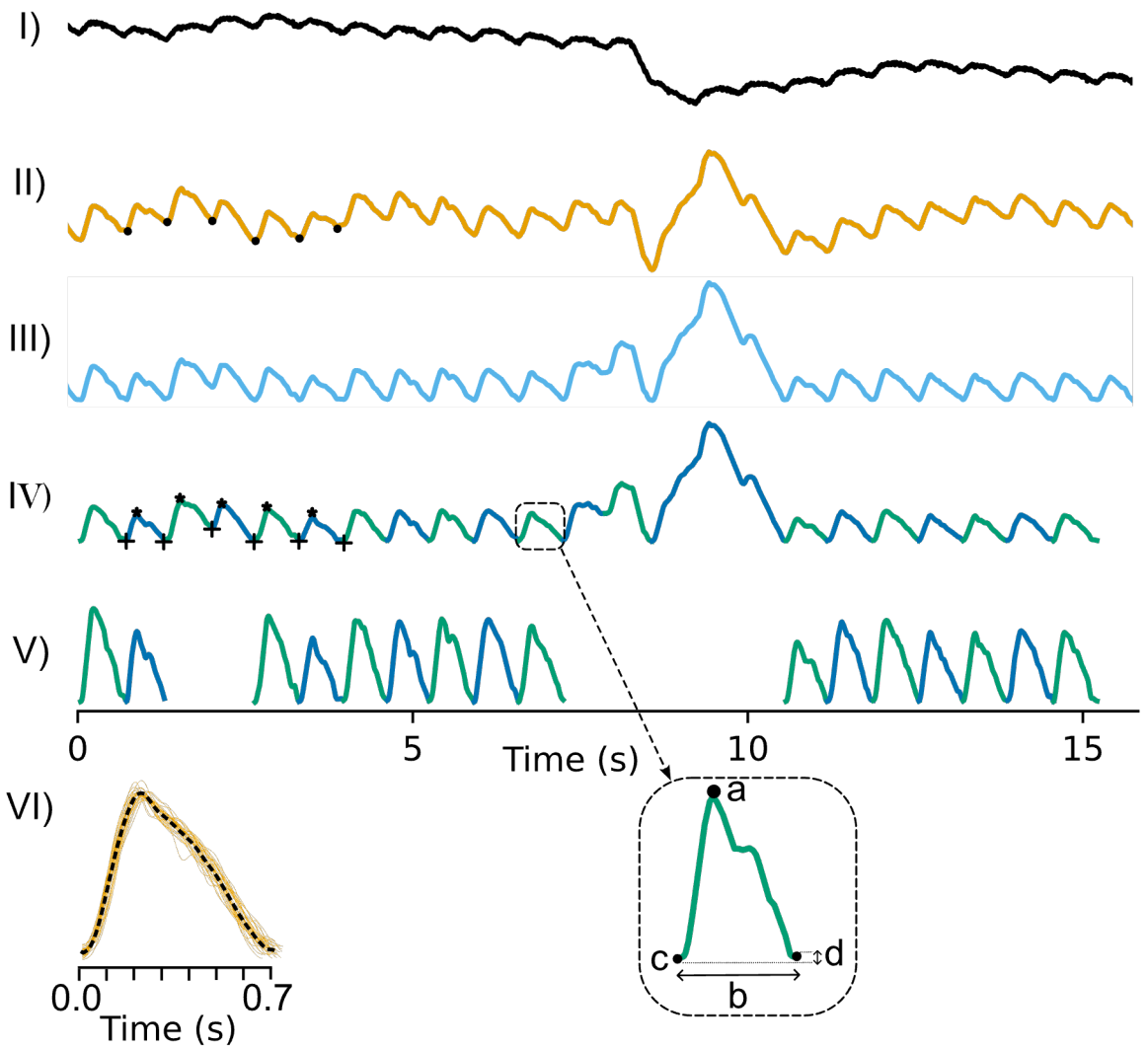
What is Markov Blanket in DAG?



Markov Blanket (MB) of Y is the set of nodes which are parents, children and spouses of the target node in the graph. Given the MB, all other features outside the MB would be redundant for prediction of Y (conditional independent of Y given MB).

- It was first proposed by [2] to discover only the MB from data without learning the full graph. Over the past decades, variant methods by extending and improving [2] have been developed.

[2] Tsamardinos, I., Aliferis, C. F., Statnikov, A. R., & Statnikov, E. (2003, May). Algorithms for large scale Markov blanket discovery. In *FLAIRS conference* (Vol. 2, pp. 376-380).


I)
II)
III)
IV)
V)
0
5
Time (s)
10
15
VI)
0.0
0.7
Time (s)
a
b
c
d

Feature Selection using Markov Blanket for Blood Pressure Regression

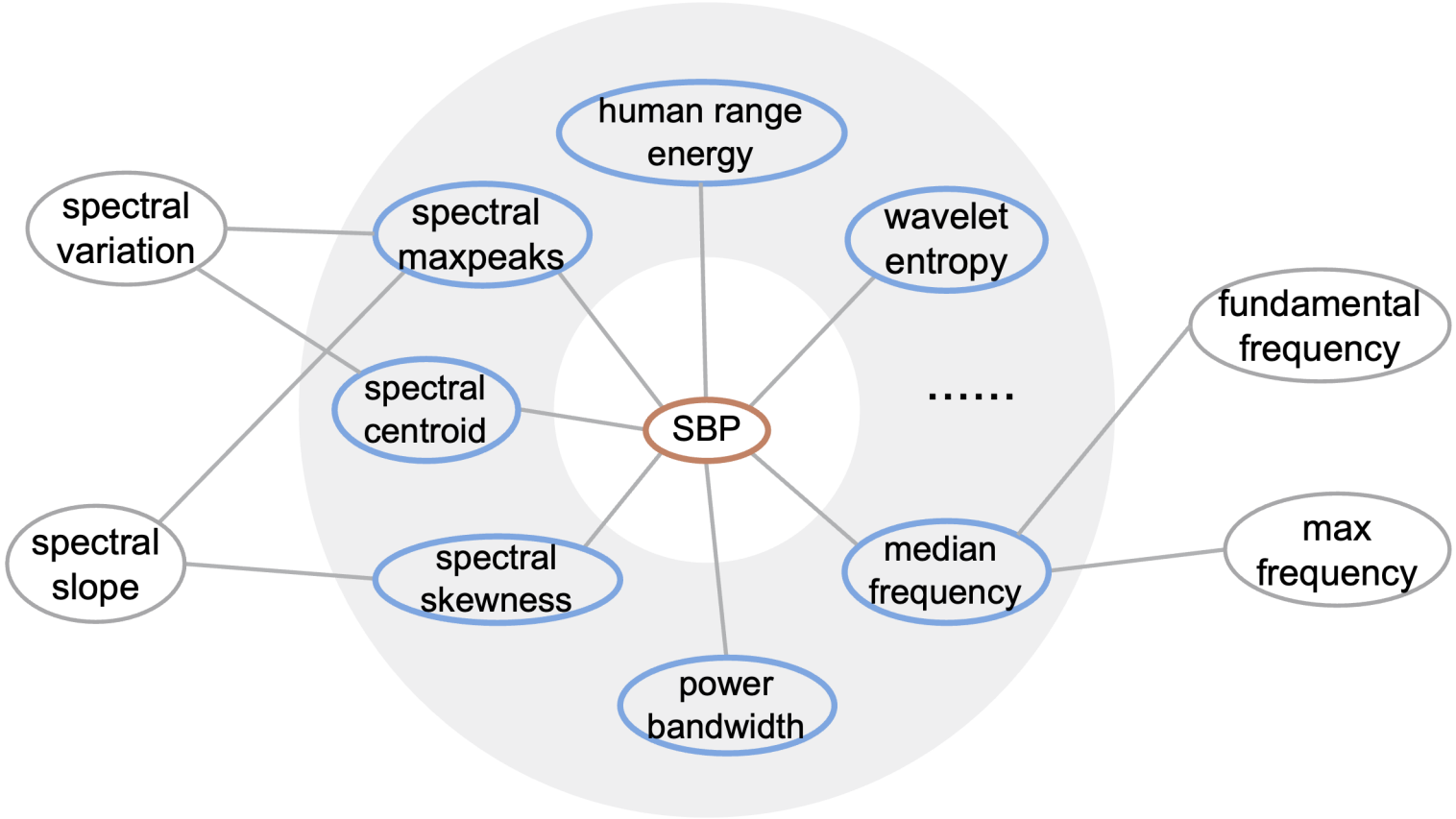
Objectives:

- Select stable features to estimate the blood pressure for different populations.
- Increase both robustness and accuracy of estimation against domain shift.

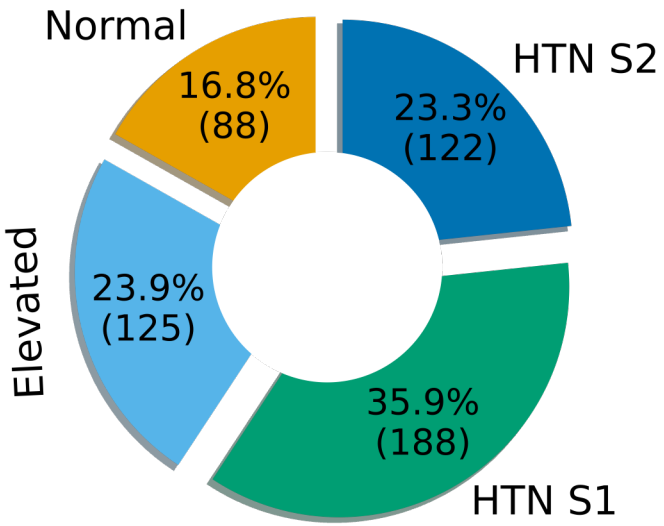
Assumptions:

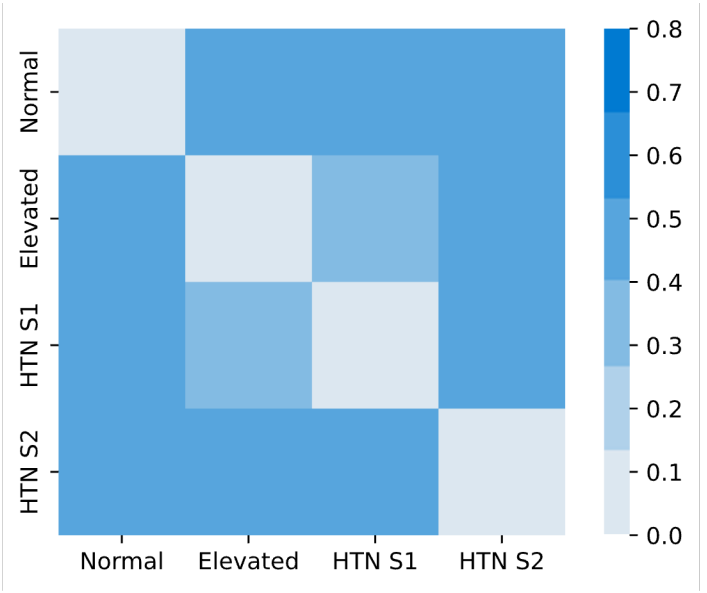
- Markovness and Faithfulness.
- Constant relationships of conditional independence across domains.
- Causal sufficiency (no latent confounders).

Distribution Distance

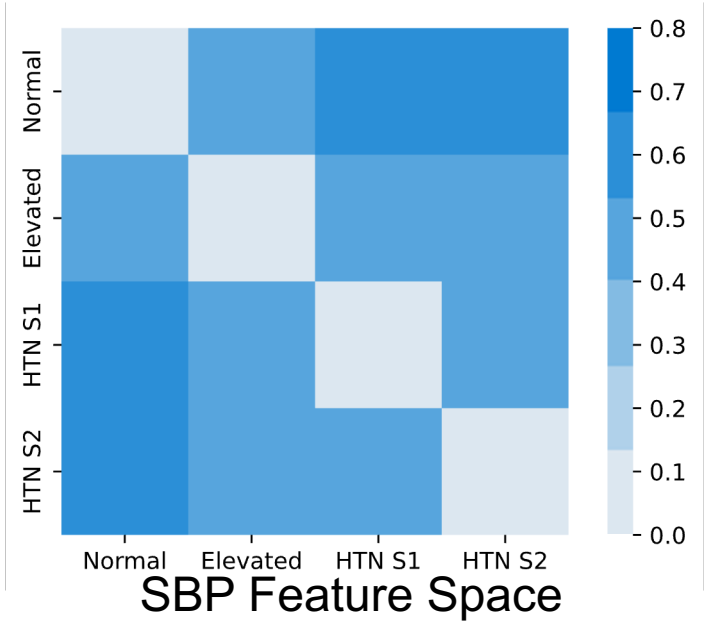


Data Distribution Difference

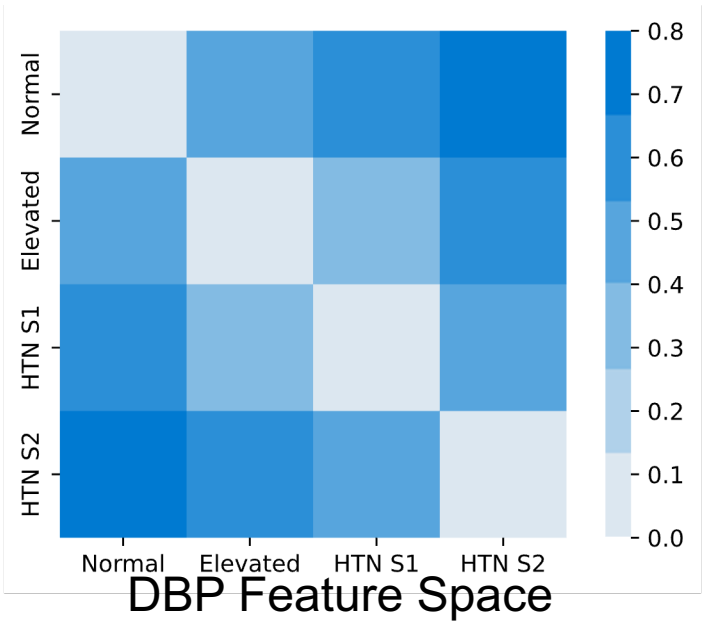




Full Feature Space



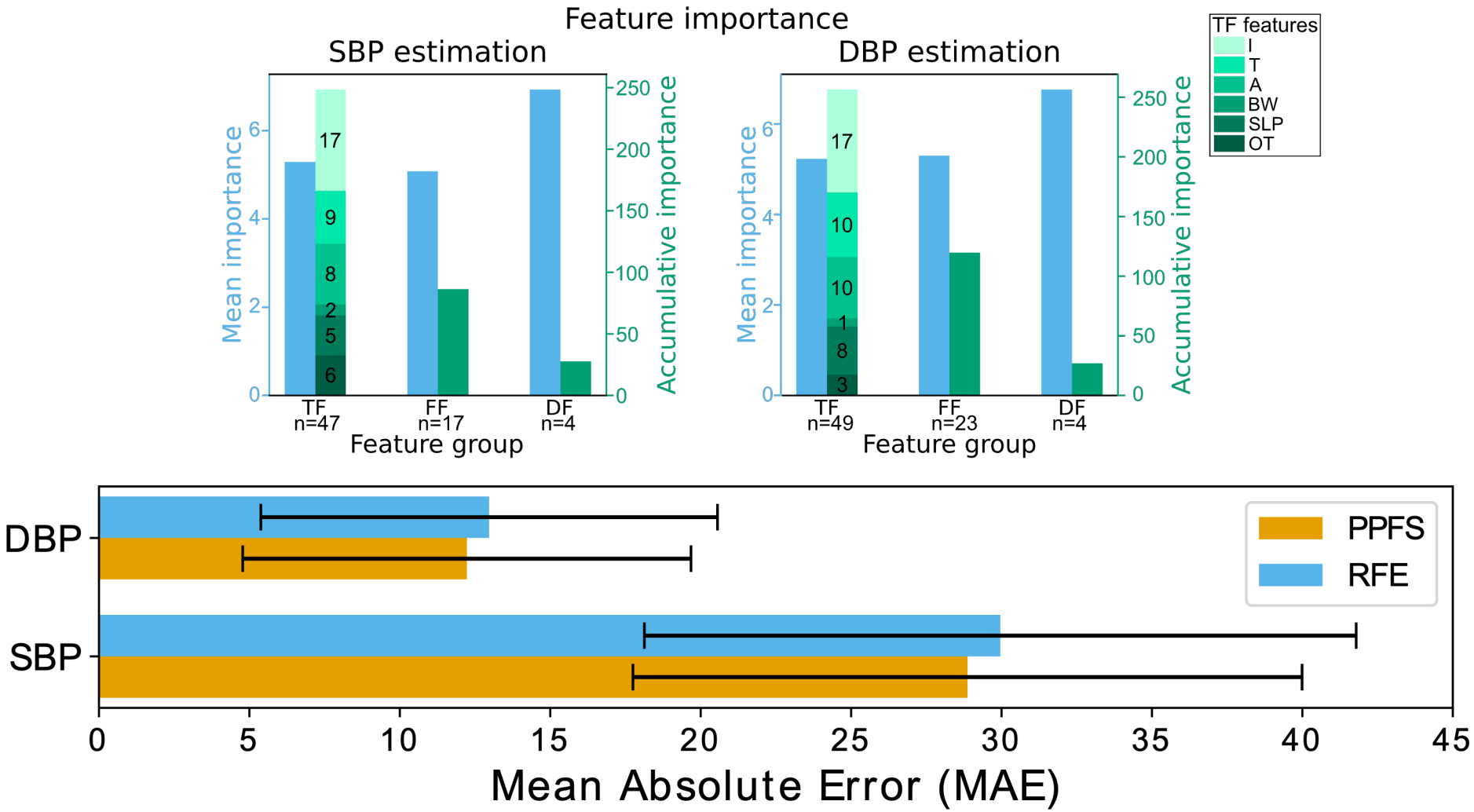
SBP Feature Space



DBP Feature Space

A. Cisnal, Y. Li, B. Fuchs, M. Ejtehadi, R. Riener, and D. Paez-Granados. "Robust Feature Selection for Continuous BP Estimation in Multiple Populations: Towards Cuffless Ambulatory BP Monitoring". In: Under Review - IEEE Journal of Biomedical and Health Informatics (2023). DOi: 10.36227/techrxiv.24112650

Aurora BP Estimation



A. Cisnal, Y. Li, B. Fuchs, M. Ejtehadi, R. Riener, and D. Paez-Granados. “Robust Feature Selection for Continuous BP Estimation in Multiple Populations: Towards Cuffless Ambulatory BP Monitoring”. In: Under Review - IEEE Journal of Biomedical and Health Informatics (2023). DOI: 10.36227/techrxiv. 24112650

Thank You!

References

[1] Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in genetics*, *10*, 524.

[2] Tsamardinos, I., Aliferis, C. F., Statnikov, A. R., & Statnikov, E. (2003, May). Algorithms for large scale Markov blanket discovery. In *FLAIRS conference* (Vol. 2, pp. 376-380).

[3] Magliacane, S., Van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., & Mooij, J. M. (2018). Domain adaptation by using causal inference to predict invariant conditional distributions. *Advances in neural information processing systems*, *31*.

[4] Peters, J., Bühlmann, P., & Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *78*(5), 947-1012.

Yanke Li
SCAI PhD Researcher
yanke.li@hest.ethz.ch

SCAI Lab
Direktionssekretariat SPZ
Guido A. Zäch Strasse 1
6207 Nottwil
Switzerland

Sensory-Motor Systems Lab
Gloriastrasse 37/ 39
ETH GLC G19
8092 Zurich
Switzerland