CPXR – A Powerful New Method for both Regression Modeling and Regression Model Analysis

Guozhu Dong

PhD, Professor

Data Mining Research Lab

Kno.e.sis Center of Excellence

Wright State University

http://cecs.wright.edu/~gdong/

guozhu.dong@gmail.com

Outline

- Introduction
- Pattern aided regression models: PXR
- Diverse predictor-response relationships
- Contrast pattern aided regression algorithm: CPXR
- Experimental evaluation
- CPXR(Log): Logistic variant of CPXR
- Example applications of CPXR:
  - Water content prediction for soil
  - Traumatic brain injury (TBI) outcome prediction
  - Heart failure (HF) survival prediction
- Potentials of CPXR and insights

← New regression model type

←Reason why existing algorithms perform poorly?!

WRIGHT STATE UNIVERSITY

Prediction is important for insurance

- Success of insurance companies depends on
  - Robust risk profiling techniques
    - **Accurately estimate risk** on all policyholders
    - Premium level is aligned with risk level
    - Retain profitable policyholders
    - **Avoid losing money on risky policyholders**

Prediction is difficult

- Prediction is difficult, especially if it is about the future
  - Nils Bohr, Nobel laureate in Physics
  - Danish Proverb
- Those who have knowledge, don't predict. Those who predict, don't have knowledge.
  - Lao Tzu, **6th** Century BC Philosopher

Prediction is difficult

- Prediction is difficult, especially if it is about the future
  - Nils Bohr, Nobel laureate in Physics
  - Danish Proverb
- Those who have knowledge, don't predict. Those who predict, don't have knowledge.
  - Lao Tzu, **6th** Century BC Philosopher
- *Prediction can be difficult, especially when the modeler doesn't know how to handle diverse predictor-response relationships.*
  - Guozhu Dong ☺

WRIGHT STATE
UNIVERSITY

Prediction is difficult

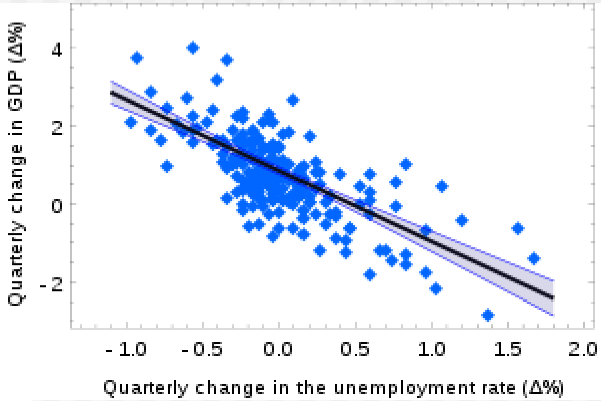
- Prediction is difficult, especially if it is about the future
  - Nils Bohr, Nobel laureate in Physics
  - Danish Proverb
- Those who have knowledge, don't predict. Those who predict, don't have knowledge.
  - Lao Tzu, **6th** Century BC Philosopher
- Prediction can be difficult. But prediction can be done fairly accurately, if one has the right tools & relevant knowledge.
  - Guozhu Dong ☺

CPXR can help

- Designing highly accurate risk models
  - Especially when the data is challenging:
    - has high dimension
    - has diverse predictor-response relationships
- Evaluating existing risk models
  - In **what way those models make big mistakes**
  - How to systematically correct those mistakes
- Identifying niche insurance opportunities
- Avoiding big losses on high risk customers

Preliminaries on prediction using regression

- Training dataset: {(x_i,y_i) | 1 <= i <=n}
  - x_i: vector of predictor variables
  - y_i: value of response variable
- Regression model evaluation

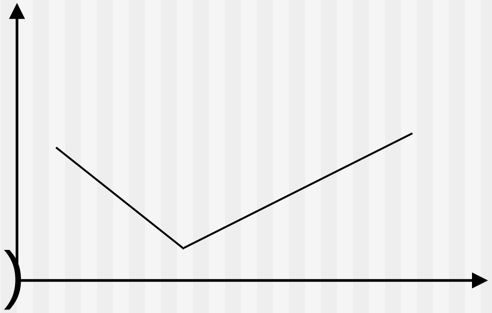


$$RMSE(f) = \sqrt{\frac{\sum_{i=1}^{n}(f(x_i) - y_i)^2}{n}}$$

$$R^2(f) = 1 - \frac{\sum_{i=1}^{n}(f(x_i) - y_i)^2}{\sum_{i=1}^{n}(y_i - avg_{j=1}^{n} y_j)^2}$$

Many methods/model types have been proposed

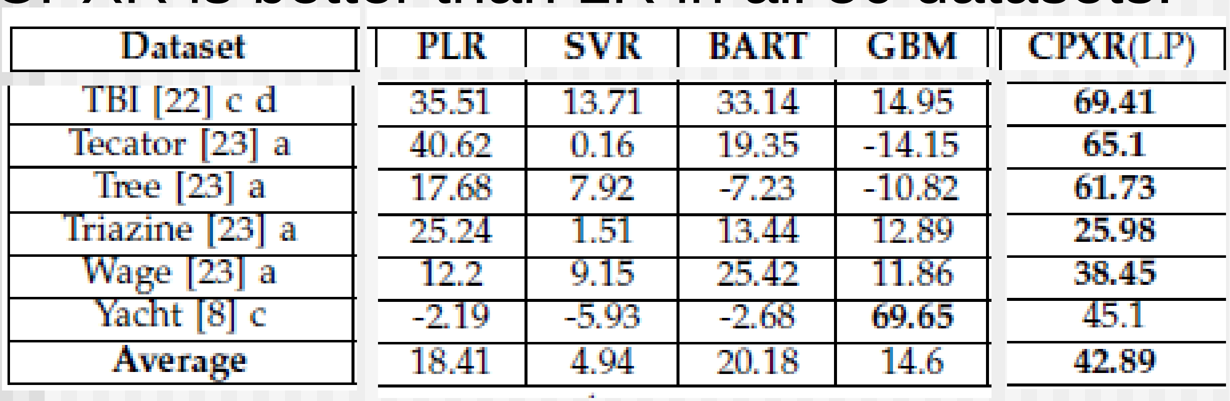
- Linear regression (LR): a linear function
- Piecewise linear regression (PLR)
  - several linear functions, each over an interval of (the same) one predictor variable



- Gradient boosting regression modeling (GBM)
  - multiple regression functions
  - generalization of AdaBoost, adapted for regression
- BART (Bayesian Additive Regression Trees)
  - a large number of regression trees
- Support vector regression (SVR): SVM like, but minimizing prediction error
- Neural networks and many others …

Peek at performance of CPXR (1)

- **CPXR: highest accuracy in 41 out of 50 datasets**
- Average RMSE reduction (relative to LR) of 42% in 50 datasets, much higher than the best competing method
- CPXR achieved 60+% RMSE reduction in 10 out the 50.
- CPXR is better than LR in all 50 datasets.

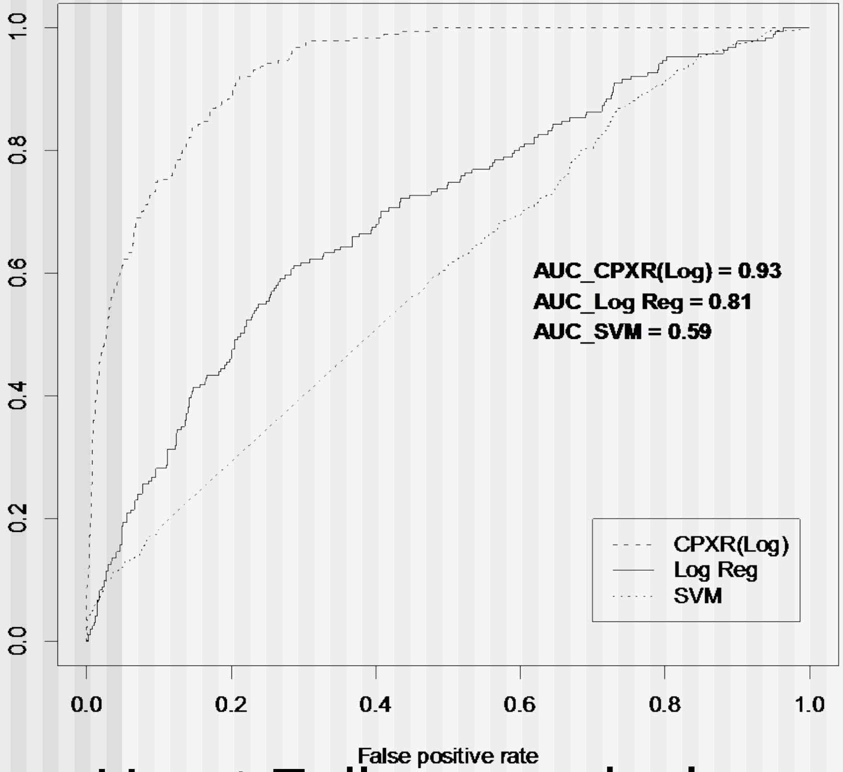


Dataset	PLR	SVR	BART	GBM	CPXR(LP)
TBI [22] c d	35.51	13.71	33.14	14.95	**69.41**
Tecator [23] a	40.62	0.16	19.35	-14.15	**65.1**
Tree [23] a	17.68	7.92	-7.23	-10.82	**61.73**
Triazine [23] a	25.24	1.51	13.44	12.89	**25.98**
Wage [23] a	12.2	9.15	25.42	11.86	**38.45**
Yacht [8] c	-2.19	-5.93	-2.68	**69.65**	45.1
Average	18.41	4.94	20.18	14.6	**42.89**

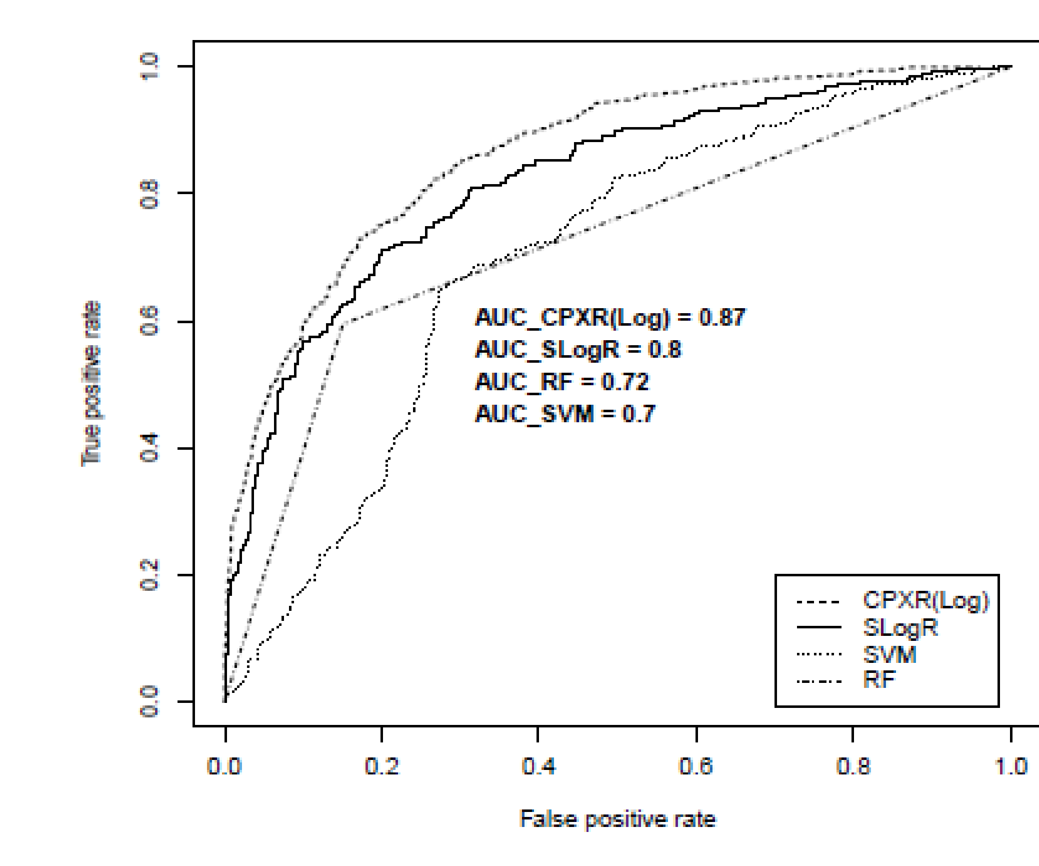
RMSE reduction of M:
[RMSE(LR)-RMSE(M)] / RMSE(LR)

Peek 2: AUC of ROC curves for CPXR(Log) vs other methods on HF and TBI

this is about classification



Heart Failure survival

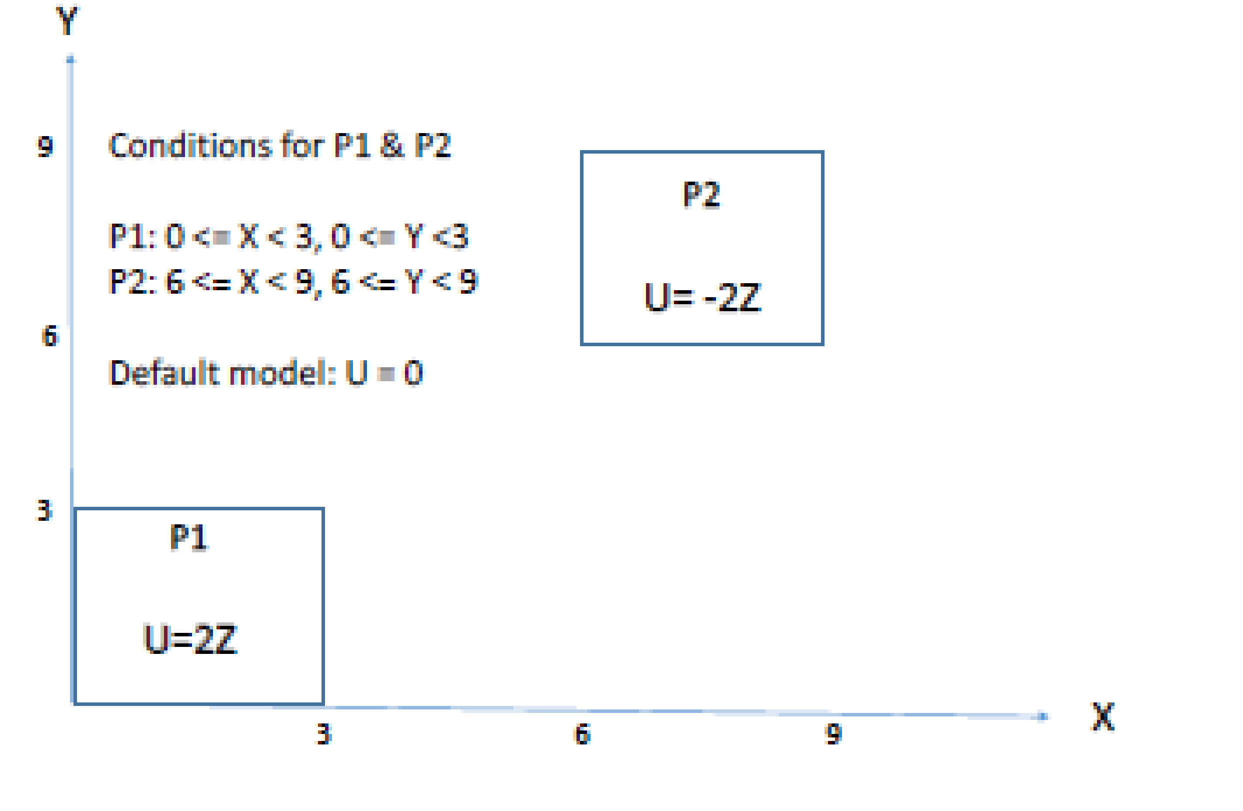


Traumatic Brain Injury outcome

Diverse predictor-response relationships

- Definition: The data (for an application) contains diverse predictor-response relationships if it contains substantially different subgroups with highly different best-fit group-specific local models [Dong+TaslimitehraniTKDE15]
- PXR is good because it is designed to capture diverse predictor-response relationships in a natural manner
  - A PXR model = several pattern & local regression model pairs
- We believe: Diverse predictor-response relationships are the main reason why best state-of-the-art regression methods often perform poorly

Illustration: Diverse Predictor Response Relationships and PXR



Conditions for P1 & P2

P1: 0 <= X < 3, 0 <= Y <3
P2: 6 <= X < 9, 6 <= Y < 9

Default model: U = 0

There are 3 predictor variables: X,Y,Z

U is response variable

- Data in P1-region, model is U = 2Z
- Data in P2-region: model is U = -2Z
- Data outside P1 and P2: model is U=0

Preliminaries: Patterns & contrast patterns

- Pattern: A (simple) condition (on individual objects)
  - **EG: age <= 35 & rank = full professor**
  - It describes all full professors whose age <=35.

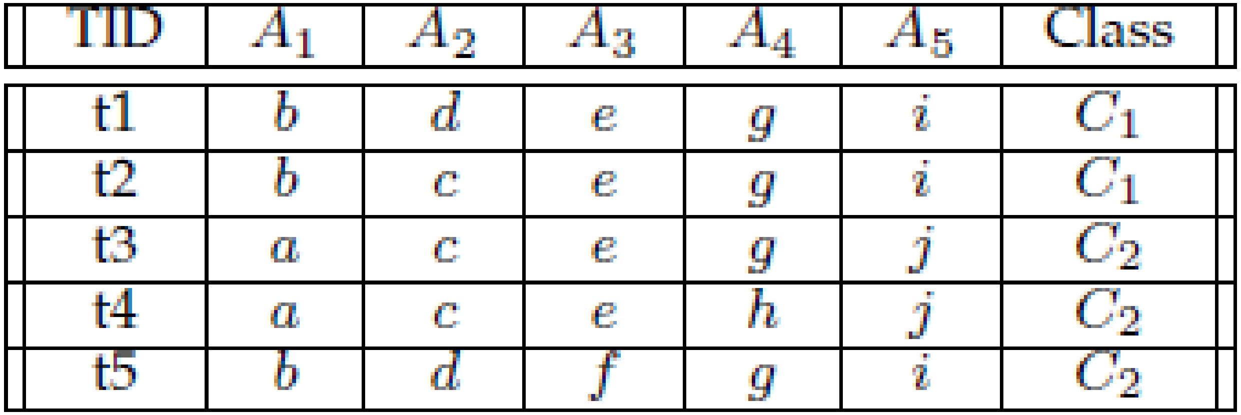
  - A pattern describes a subgroup of data using a low dimensional constraint [data is in higher dimensional space]
  - It is a logical unit in the conceptual thinking about objects

- Contrast patterns: patterns/conditions that distinguish objects in different classes
  - Class: the good, the bad

WRIGHT STATE UNIVERSITY

Contrast patterns – example

- CP: A1=b & A3=e
  - It matches → all C1 objects. →
  - It matches zero C2 objects
  - Its mds={t1,t2}



TID	A_1	A_2	A_3	A_4	A_5	Class
t1	b	d	e	g	i	C_1
t2	b	c	e	g	i	C_1
t3	a	c	e	g	j	C_2
t4	a	c	e	h	j	C_2
t5	b	d	f	g	i	C_2

- Generally: A contrast pattern matches many more objects in one class than in the other classes
  - aka emerging pattern [I first studied EPs in 1998]
- mds(P): the set of objects matching P.
  - An equivalence class of patterns: A set of patterns having the same matching dataset (mds) (hence having same behavior)

Intuitively

- A pattern is a simple condition characterizing a set of objects (customers)
- A contrast pattern characterizes a set of objects of some given class C0
  - If you see an object matching the CP, then it is likely the object belongs to C0
  - We will define the classes for use in building regression models

Why contrast pattern based approaches are successful & have big potentials?

- Contrasting is useful
  - Contrasting ➔ findings linked with risk indicator ➔ findings give advantage in survival/wellbeing
  - Contrasting: built into human (animal) instincts
- Focus on contrast patterns ➔ focus on important issues
  - Efficiency: Number of contrast patterns << number of frequent patterns
- **Opportunity**: Humans mostly used low dim CPs (old days) (?brain's computing power is low & humans had little data?)
  - Humans still depend on independence assumption for high dimension apps (!unfortunate!).
  - High (3—7) dimensional CPs capture novel important multi-variable interactions (insights), with impact on outcome

Pattern aided regression model (formal)

- A PXR model is represented by a tuple

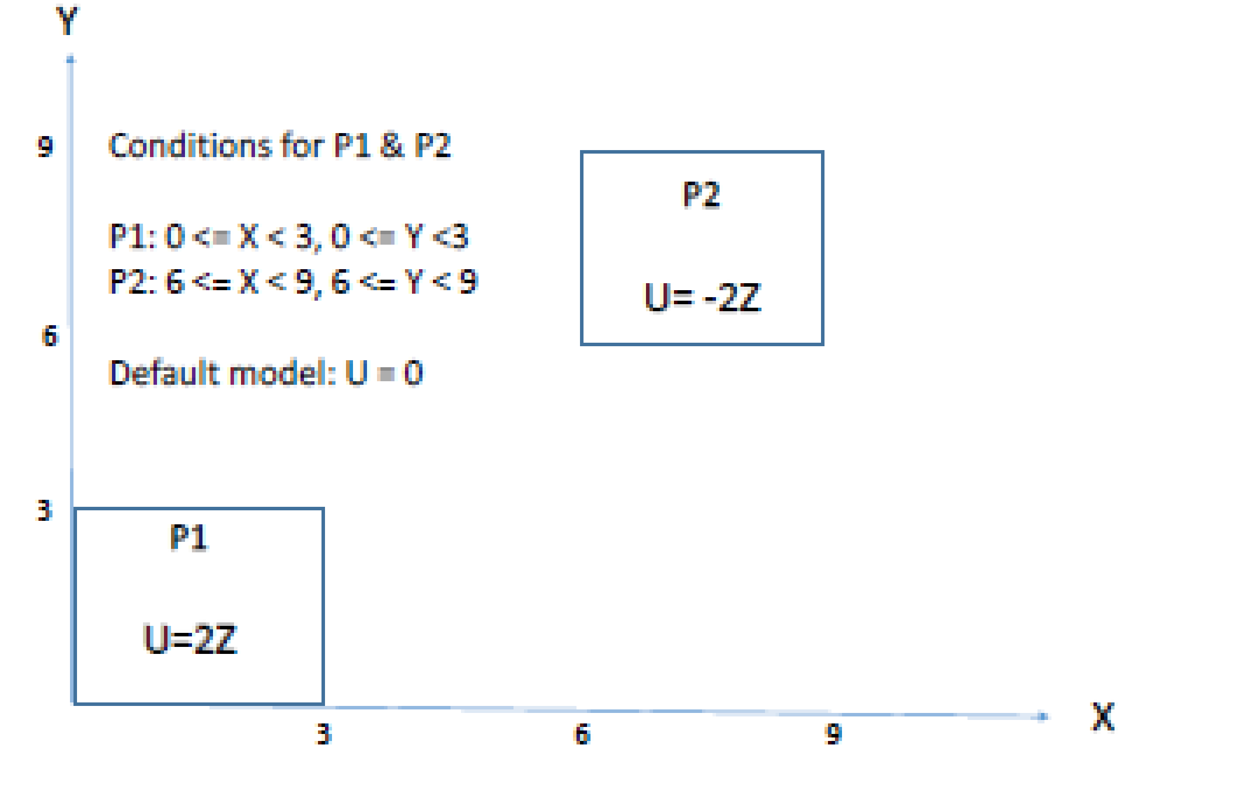
$$PM = ((P_1, f_1, w_1), \ldots, (P_k, f_k, w_k), f_d),$$

  - Each P_i is a pattern
  - Each f_i is a local regression model, learned from, and to be applied to, only data satisfying P_i,
  - f_d is a default local regression model
  - Each w_i is weight for f_i
- The regression function of PM is defined by

$$f_{PM}(x) = \begin{cases} \frac{\Sigma_{P_i \in \pi_x} w_i f_i(x)}{\Sigma_{P_i \in \pi_x} w_i} & \text{if } \pi_x \neq \emptyset \\ f_d(x) & \text{otherwise} \end{cases}$$

$$\pi_x = \{P_i \mid 1 \leq i \leq k,\ x \text{ satisfies } P_i\}$$

A simple PXR model


Conditions for P1 & P2

P1: 0 <= X < 3, 0 <= Y <3
P2: 6 <= X < 9, 6 <= Y < 9

Default model: U = 0

PM=((P1,f1,1), (P2,f2,1), fd)

There are 3 predictor variables: X,Y,Z

- For (X=2,Y=2,Z=5), PXR model returns 10 P1
- For (X=7,Y=8,Z=5), PXR model returns -10 P2
- For (X=2,Y=8,Z=5), PXR model returns 0 neither P1 nor P2
- weights not important here since there Is no overlap between mds(P1), mds(P2)
-

Another simple PXR model

- u,v z: predictor variables, y: response variable
- $((P_1, f_1, w_1), (P_2, f_2, w_2), f_d)$ is a PXR model

$$P_1 : 2 \le u < 5 \,\&\, 3 \le v < 7$$

$$P_2 : 6 \le z < 9$$

$$f_1 : y = 4 + 3z$$

$$f_2 : y = 5 + 2u$$

$$f_d : y = 2u + 4v$$

$$w_1 = 0.6,\ w_2 = 0.3,$$

1. mds(P1) and mds(P2) have overlap

- Weighted average is used to derive PXR's value on (u=3,v=4,z=7)

2. P1 and P2 use different sets of variables

WRIGHT STATE UNIVERSITY

Discussion (1)

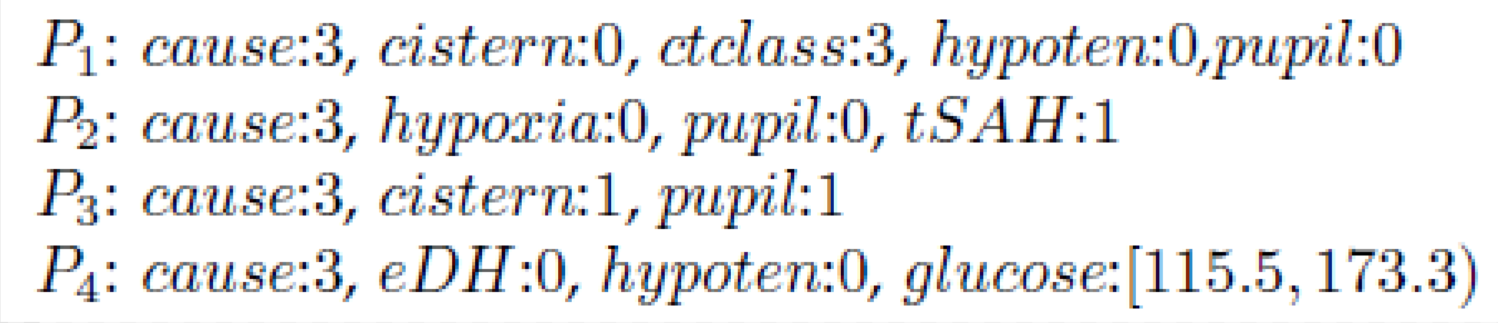
- PXR is a strict and flexible generalization of PLR
  - PLR can be viewed as trying to model diverse predictor-response relationships, but it is limited in modeling capabilities and computing algorithms
    - *Not aware of earlier researchers* mentioning DPR
- PXR models are easy to understand
  - Often a PXR model uses very few patterns (e.g. 7)
  - Often we only need to use simple local regression models such as LR or PLR
- PXR models can describe accurate models (expressive)

Discussion (2): PXR and diverse predictor-response relationships

- Different pattern-model pairs in a PXR can involve different sets of variables
- Different pattern-model pairs can be associated with highly different local regression functions
- Each pattern-model pair captures a highly distinct kind of behavior (predictor-response relationship)

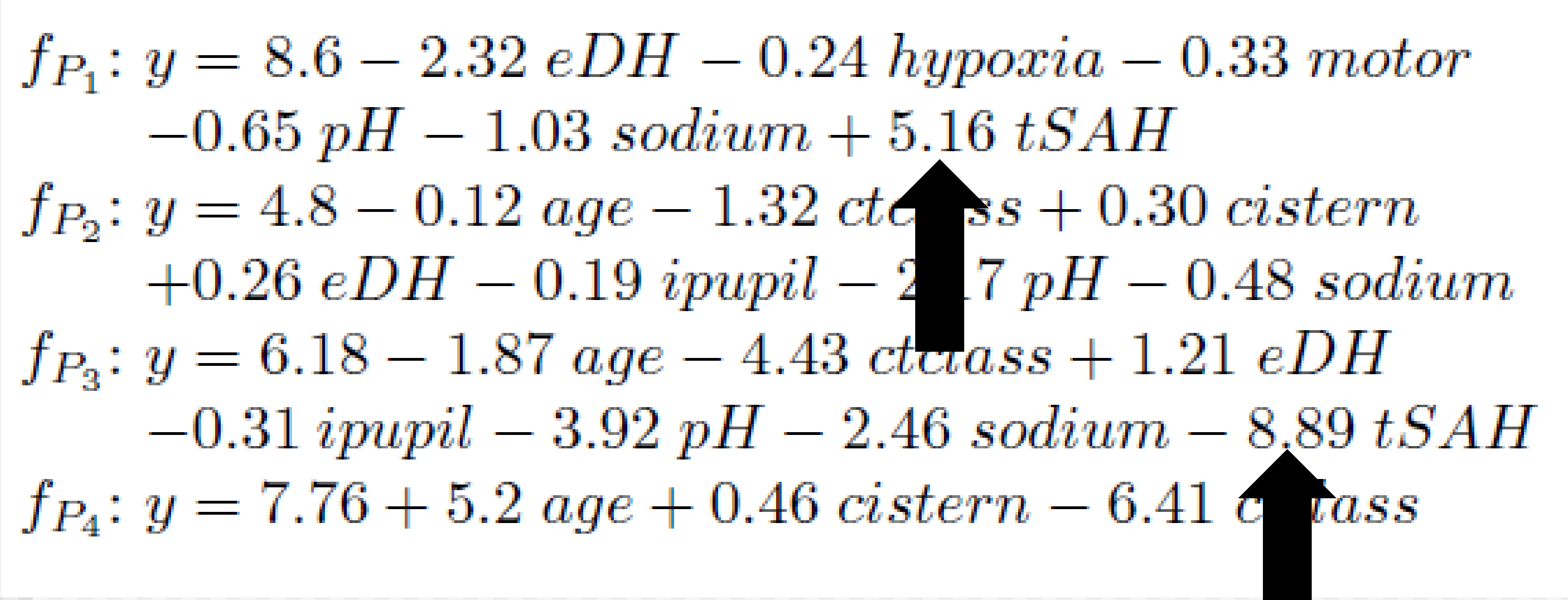
PXR can represent diverse predictor-response relationships, just by the way it is defined

DPR in TBI (traumatic brain injury)



P_1: *cause*:3, *cistern*:0, *ctclass*:3, *hypoten*:0,*pupil*:0
P_2: *cause*:3, *hypoxia*:0, *pupil*:0, *tSAH*:1
P_3: *cause*:3, *cistern*:1, *pupil*:1
P_4: *cause*:3, *eDH*:0, *hypoten*:0, *glucose*:[115.5, 173.3)

tSAH has very different coefficients



$$f_{P_1}: y = 8.6 - 2.32\,eDH - 0.24\,hypoxia - 0.33\,motor - 0.65\,pH - 1.03\,sodium + 5.16\,tSAH$$

$$f_{P_2}: y = 4.8 - 0.12\,age - 1.32\,ctclass + 0.30\,cistern + 0.26\,eDH - 0.19\,ipupil - 2.17\,pH - 0.48\,sodium$$

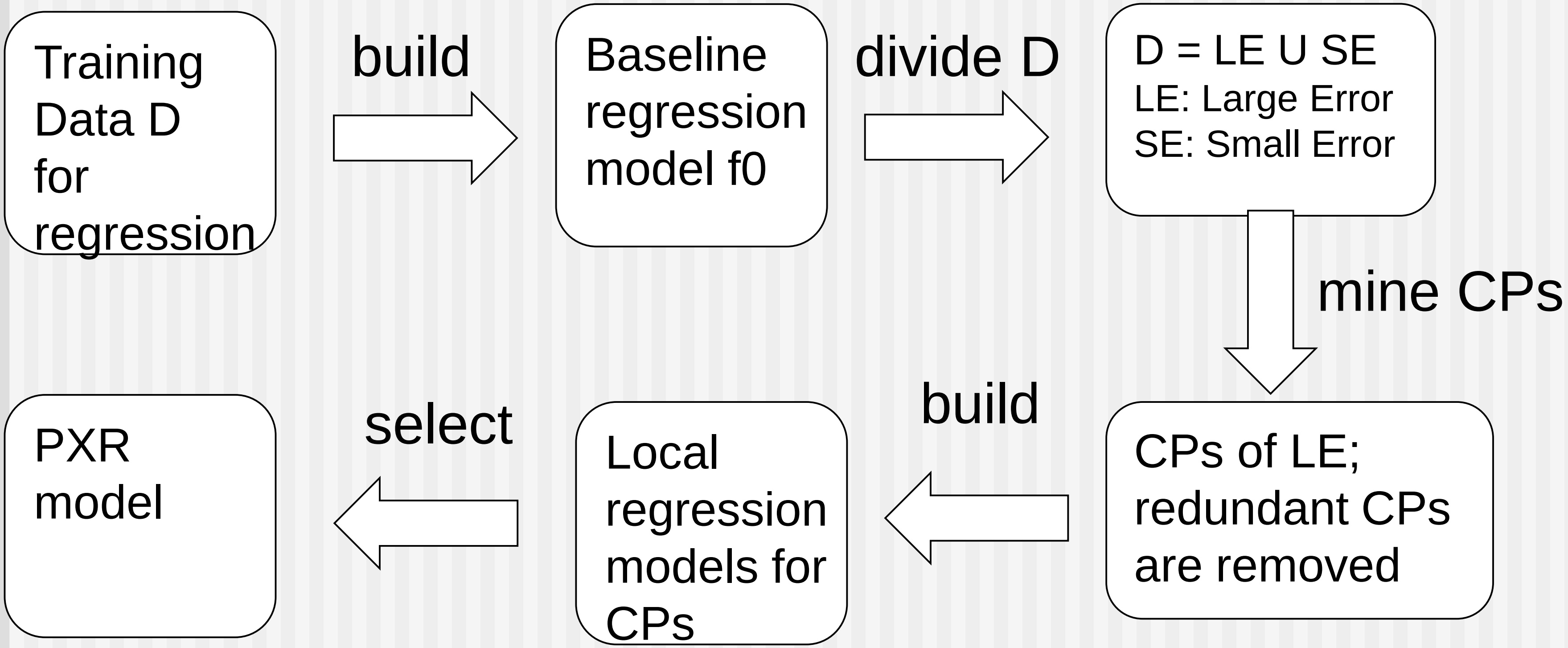
$$f_{P_3}: y = 6.18 - 1.87\,age - 4.43\,ctclass + 1.21\,eDH - 0.31\,ipupil - 3.92\,pH - 2.46\,sodium - 8.89\,tSAH$$

$$f_{P_4}: y = 7.76 + 5.2\,age + 0.46\,cistern - 6.41\,ctclass$$

Discussion (3): Diverse predictor-response relationships

- Diverse predictor-response relationships occur often in real life, for data with >=3 dimensions
- Diverse predictor-response relationships may be neutralized at the global level [e.g. the PXR in the pictorial example]
- Existing regression methods are weak in the presence of diverse predictor-response relationships

CPXR: It builds PXR models as follows



Training Data D for regression

build

Baseline regression model f0

divide D

D = LE U SE
LE: Large Error
SE: Small Error

mine CPs

CPs of LE; redundant CPs are removed

build

Local regression models for CPs

select

PXR model

Key ideas of CPXR

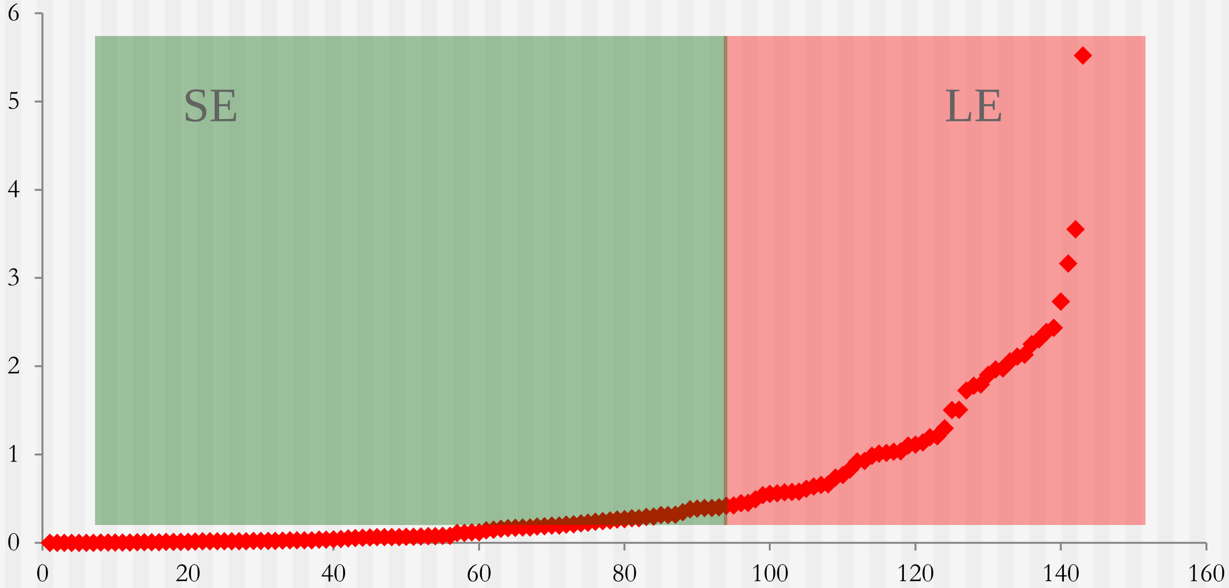
- Use contrast patterns P (that match more cases in LE than in SE)
- Identify an opportunity represented by a contrast patterns P, where
  - f0 makes lots of errors (on P’s matching data)
  - and a local model corrects those errors in a significant way

How CPXR builds PXR models: (D,f0) ➔ (LE,SE) ➔ Ps and fs ➔ PXR

- Starting with training dataset
- Build a baseline regression model f0
- Split data into LE (large error) and SE (small error), based on f0's prediction error [LE and SE are the classes for CPs]
- Mine CPs; Remove redundant CPs that are not very useful
- Build corresponding local regression models for remaining CPs, and select several CPs to construct PXR
- Many technical details, including variable binning, splitting data, search objectives, baseline model type, local model type …

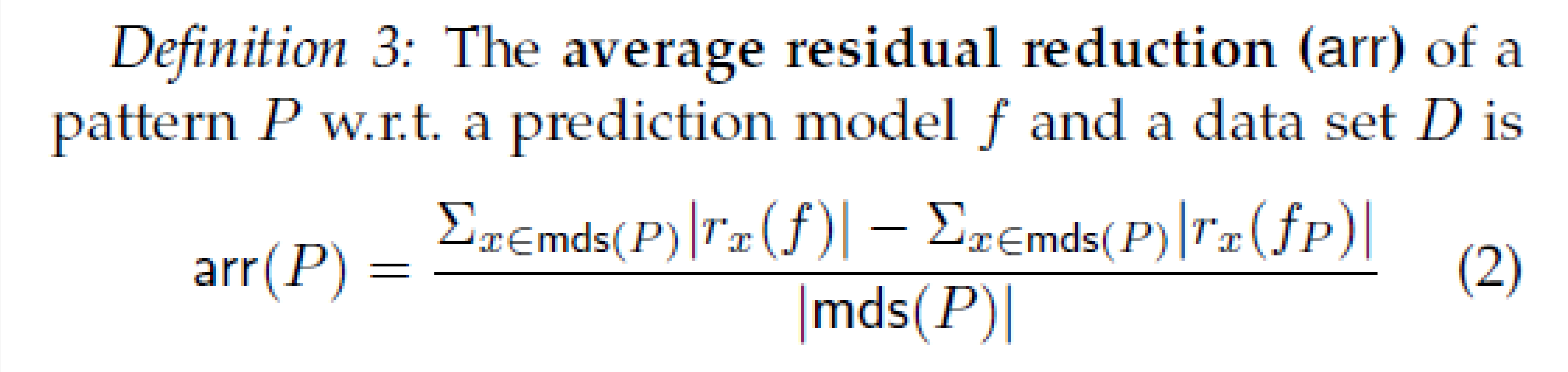
CPXR Technical Details (1)

- Divided D into LE and SE
  - using residual ratio ρ: total residuals of LE cases over total residuals of all cases



CPXR Technical Details (2)

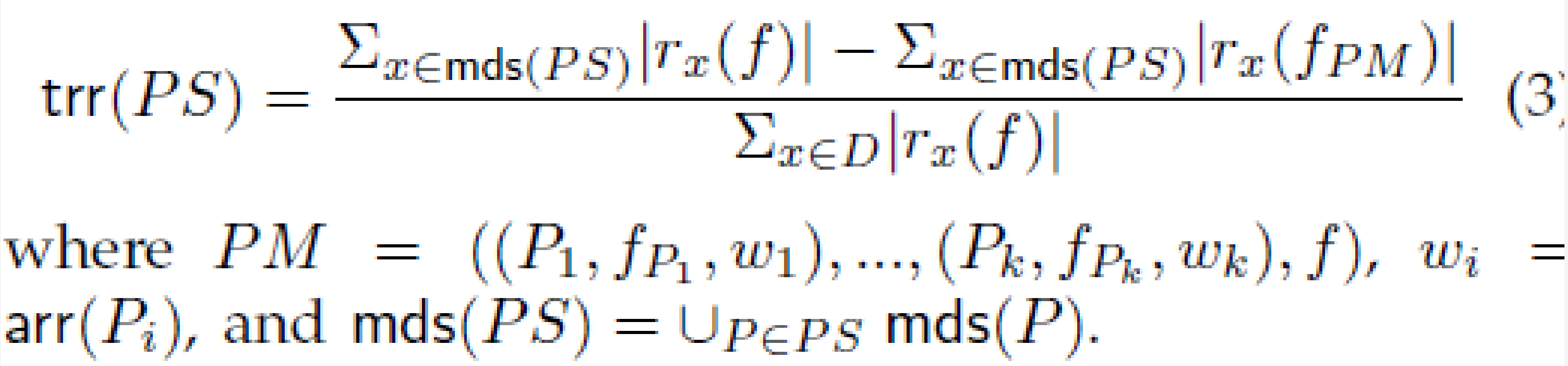
- Selecting good patterns



Definition 3: The **average residual reduction** (arr) of a pattern P w.r.t. a prediction model f and a data set D is

$$\mathsf{arr}(P) = \frac{\Sigma_{x \in \mathsf{mds}(P)} |r_x(f)| - \Sigma_{x \in \mathsf{mds}(P)} |r_x(f_P)|}{|\mathsf{mds}(P)|} \qquad (2)$$

- Selecting desirable sets of patterns



$$\mathsf{trr}(PS) = \frac{\Sigma_{x \in \mathsf{mds}(PS)} |r_x(f)| - \Sigma_{x \in \mathsf{mds}(PS)} |r_x(f_{PM})|}{\Sigma_{x \in D} |r_x(f)|} \qquad (3)$$

where $PM = ((P_1, f_{P_1}, w_1), ..., (P_k, f_{P_k}, w_k), f)$, $w_i = \mathsf{arr}(P_i)$, and $\mathsf{mds}(PS) = \cup_{P \in PS}\ \mathsf{mds}(P)$.

CPXR Technical Details (3)

- CPXR iteratively adds/replaces one CP in PS

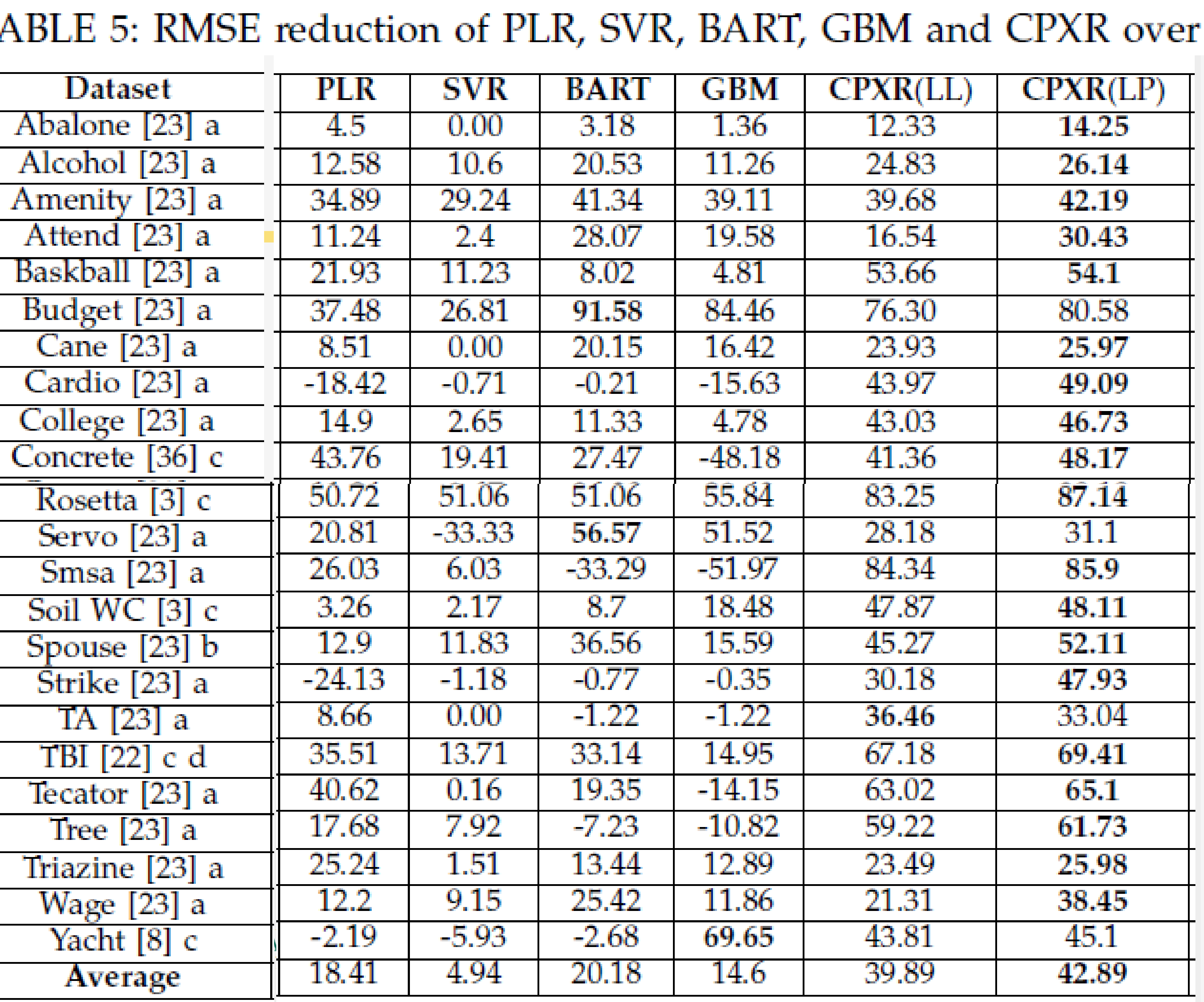
$$imp(PS, P_-, P_+) = \text{trr}(f_{PM'}, f) - \text{trr}(f_{PM(PS,f)}, f)$$

where $PM' = PM(PS - \{P_-\} \cup \{P_+\}, f)$.

  - Adding: |PS|=1, |PS|=2, …, |PS|=k
- To help avoid overfitting, CPXR does not select a CP P if |mds(P)| <= # variables
- CPXR uses equivalence class and Jaccard similarity to reduce the number of candidate CPs

Summary of empirical results

- CPXR is highly accurate for building regression models
  - Outperforms other methods, often by big margins
    - On accuracy
    - On overfitting
    - On. sensitivity to noise
- We used 50 real datasets used in previous studies by others, and 20+ synthetic ones
- Exp says: Diverse predictor-response relationships occur often in the real world, for data with >=3 dimensions (when RMSE reduction > 25%)



ABLE 5: RMSE reduction of PLR, SVR, BART, GBM and CPXR over RMSE of

Dataset	PLR	SVR	BART	GBM	CPXR(LL)	CPXR(LP)
Abalone [23] a	4.5	0.00	3.18	1.36	12.33	**14.25**
Alcohol [23] a	12.58	10.6	20.53	11.26	24.83	**26.14**
Amenity [23] a	34.89	29.24	41.34	39.11	39.68	**42.19**
Attend [23] a	11.24	2.4	28.07	19.58	16.54	**30.43**
Baskball [23] a	21.93	11.23	8.02	4.81	53.66	**54.1**
Budget [23] a	37.48	26.81	**91.58**	84.46	76.30	80.58
Cane [23] a	8.51	0.00	20.15	16.42	23.93	**25.97**
Cardio [23] a	-18.42	-0.71	-0.21	-15.63	43.97	**49.09**
College [23] a	14.9	2.65	11.33	4.78	43.03	**46.73**
Concrete [36] c	43.76	19.41	27.47	-48.18	41.36	**48.17**
Rosetta [3] c	50.72	51.06	51.06	55.84	83.25	**87.14**
Servo [23] a	20.81	-33.33	**56.57**	51.52	28.18	31.1
Smsa [23] a	26.03	6.03	-33.29	-51.97	84.34	**85.9**
Soil WC [3] c	3.26	2.17	8.7	18.48	47.87	**48.11**
Spouse [23] b	12.9	11.83	36.56	15.59	45.27	**52.11**
Strike [23] a	-24.13	-1.18	-0.77	-0.35	30.18	**47.93**
TA [23] a	8.66	0.00	-1.22	-1.22	**36.46**	33.04
TBI [22] c d	35.51	13.71	33.14	14.95	67.18	**69.41**
Tecator [23] a	40.62	0.16	19.35	-14.15	63.02	**65.1**
Tree [23] a	17.68	7.92	-7.23	-10.82	59.22	**61.73**
Triazine [23] a	25.24	1.51	13.44	12.89	23.49	**25.98**
Wage [23] a	12.2	9.15	25.42	11.86	21.31	**38.45**
Yacht [8] c	-2.19	-5.93	-2.68	**69.65**	43.81	45.1
Average	18.41	4.94	20.18	14.6	39.89	**42.89**

>25%

No

Yes

Yes

Yes

Yes

Yes

Yes

Yes

Yes

Yes

Yes

Yes

Yes

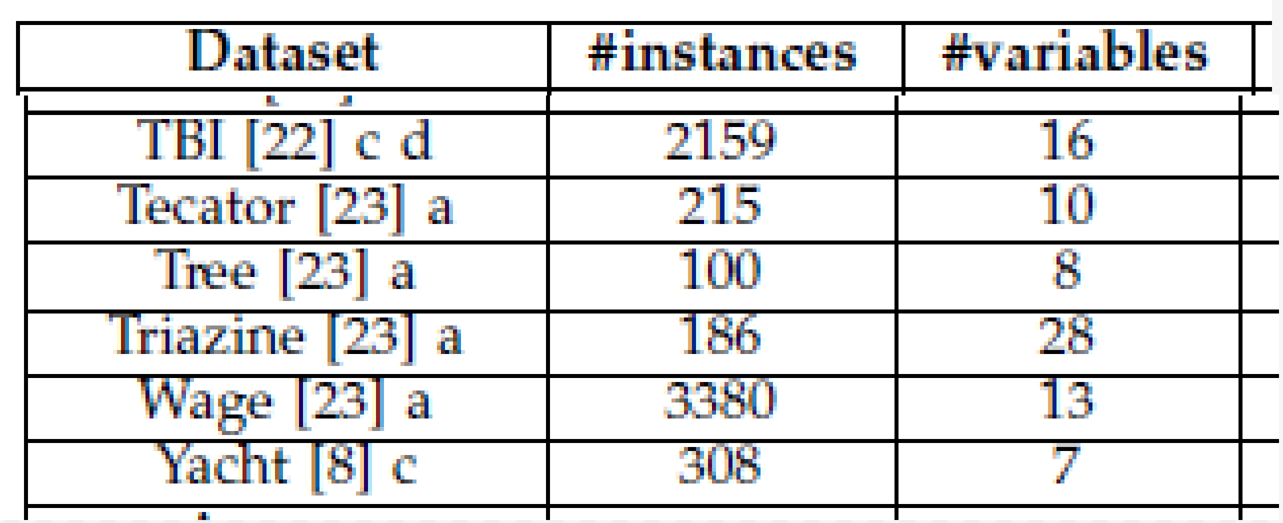
Yes

Yes

Yes

Experiments used 50 real datasets used in previous regression studies

Example dataset characteristics



Dataset	#instances	#variables
TBI [22] c d	2159	16
Tecator [23] a	215	10
Tree [23] a	100	8
Triazine [23] a	186	28
Wage [23] a	3380	13
Yacht [8] c	308	7

- We have worked with data with 80+ variables

CPXR achieved large RMSE reduction (accuracy improvement) consistently

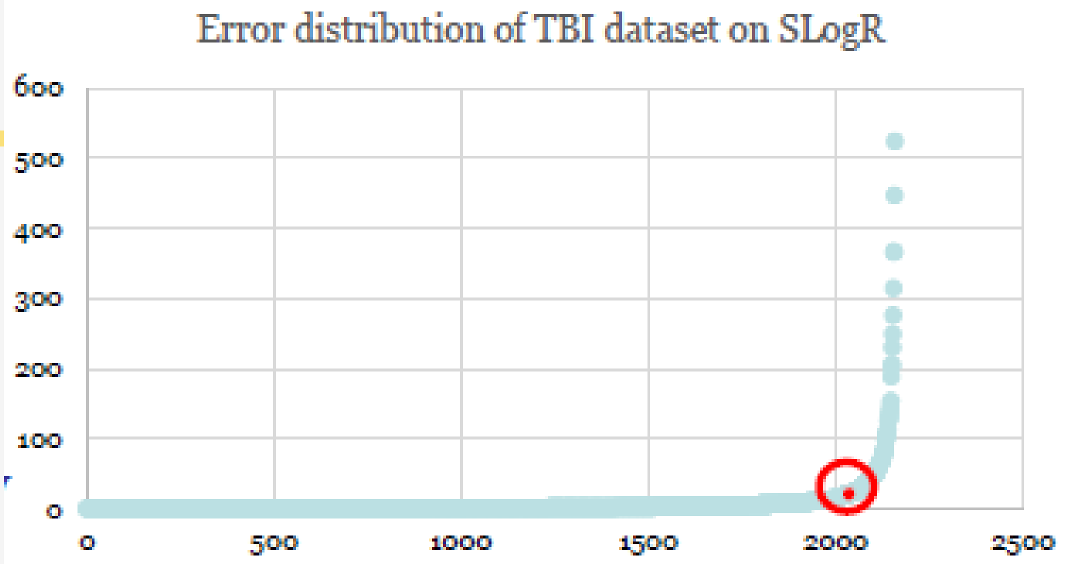
- **CPXR: highest accuracy in 41 out of 50 datasets** **(4 competitors)**
- Average RMSE reduction (relative to LR) of 42% in 50 datasets, much higher than that of best competing method
- CPXR achieved 60+% RMSE reduction in 10 out the 50.
- CPXR is better than LR in all 50 datasets.

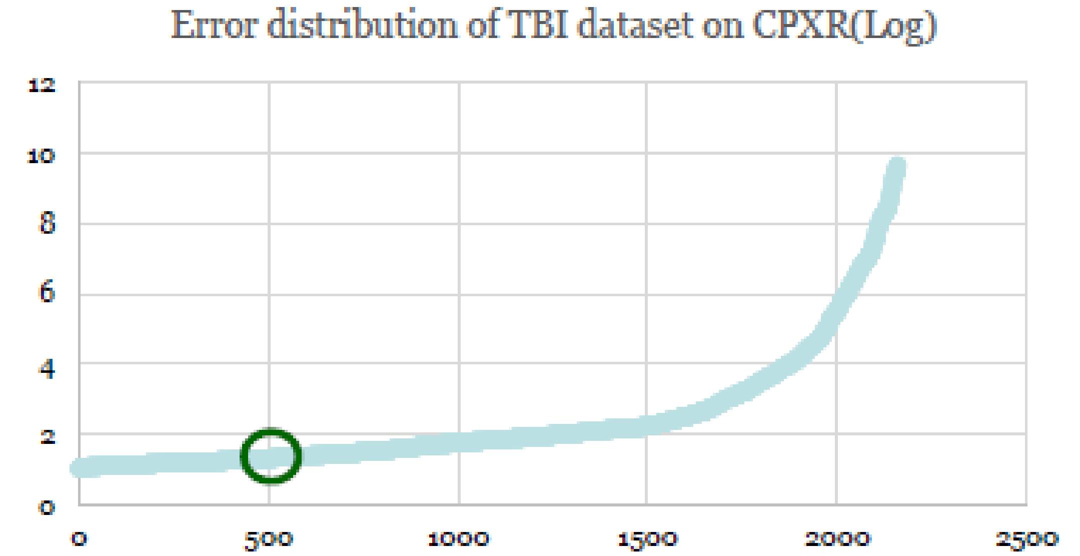
$$\text{RMSE Reduction by M} = \frac{\text{RMSE(LR)-RMSE(M)}}{\text{RMSE(LR)}}$$

We also tried other competitors but they are not competitive.
CPXR is not better than other methods on random data

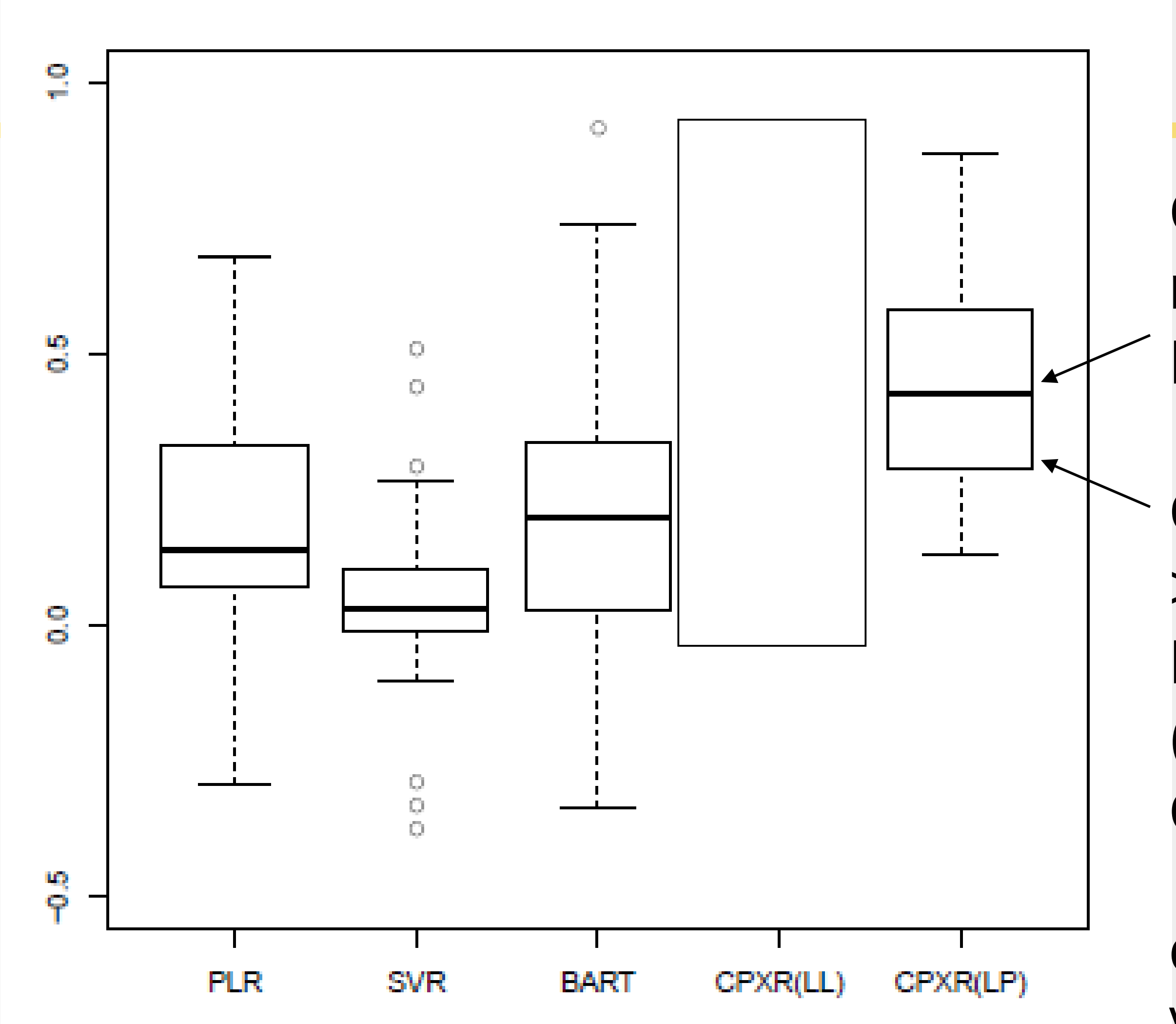
Residual reduction by CPXR

Y-axis: Pearson residuals ➔


Error distribution of TBI dataset on SLogR


Error distribution of TBI dataset on CPXR(Log)

Boxplot of RMSE reduction



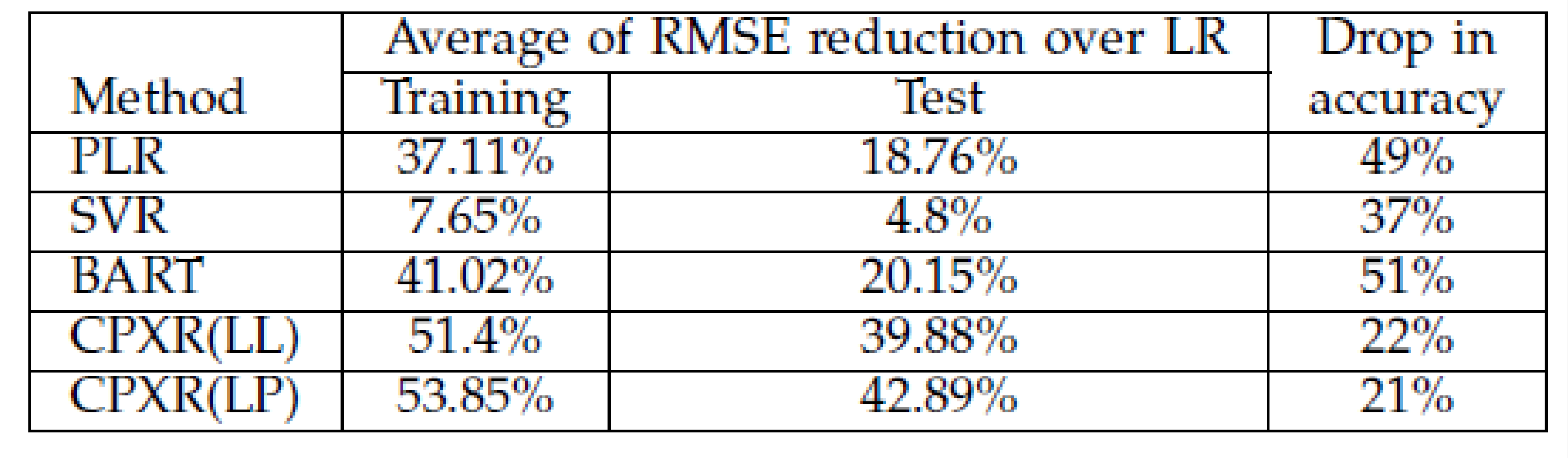
CPXR(LP)’s median > Q3 of PLR,LR,BART

CPXR(LP)’s Q1 > median of PLR,LR,BART (approx. their Q3)

CPXR(LL) is a variant of CPXR

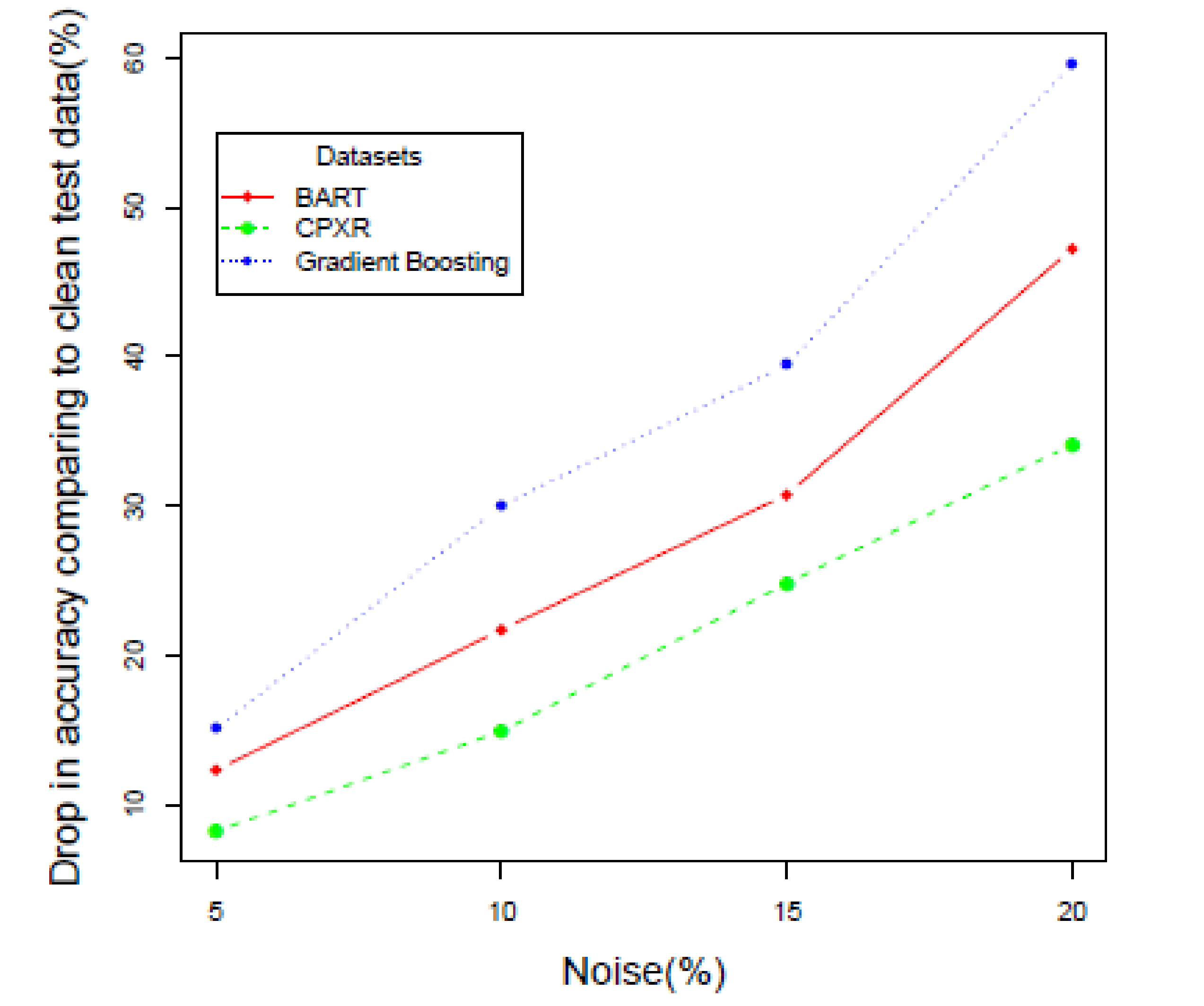
Evaluation on overfitting

- CPXR is more accurate than PLR, LR and BART on testing data; its model complexity is fairly low.
- CPXR has smaller relative accuracy drop (from training to test)



Method	Average of RMSE reduction over LR		Drop in accuracy
	Training	Test	
PLR	37.11%	18.76%	49%
SVR	7.65%	4.8%	37%
BART	41.02%	20.15%	51%
CPXR(LL)	51.4%	39.88%	22%
CPXR(LP)	53.85%	42.89%	21%

Evaluation on sensitivity to noise



Build PXR on clean training data

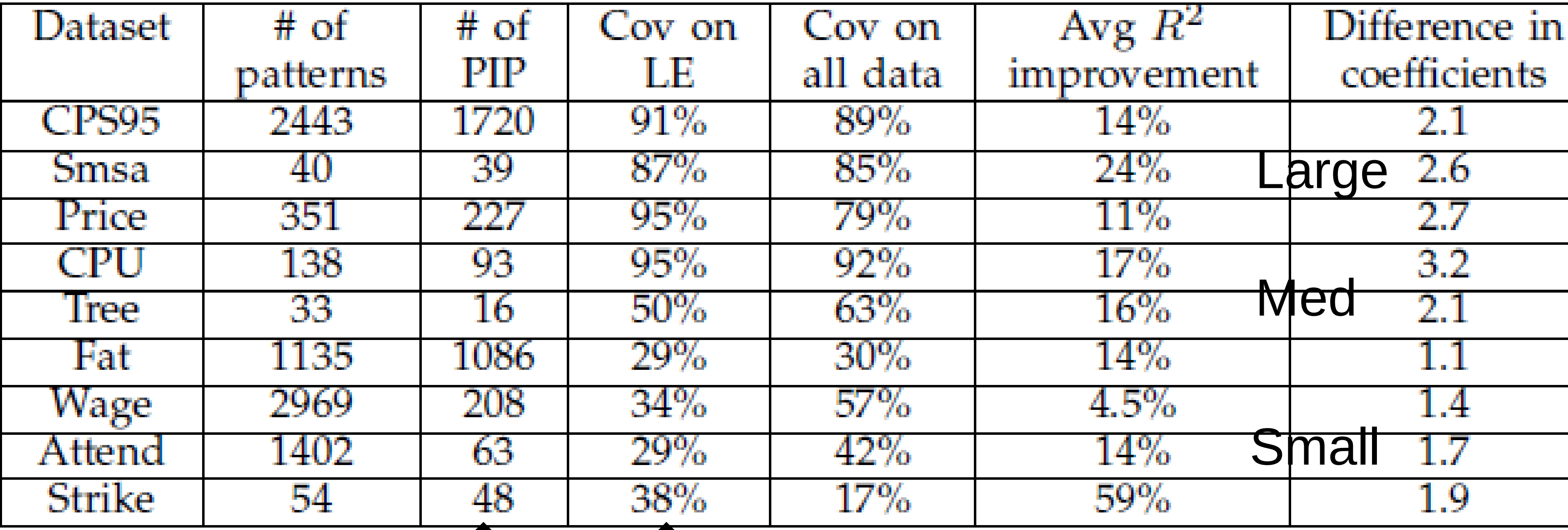
Compare accuracies on

- training data
- noise-added test data

- Smaller drop

➔ Less sensitive

CPXR's outperformance vs degree of diversity of PR-relationships



Dataset	# of patterns	# of PIP	Cov on LE	Cov on all data	Avg R^2 improvement	Difference in coefficients
CPS95	2443	1720	91%	89%	14%	2.1
Smsa	40	39	87%	85%	24%	2.6
Price	351	227	95%	79%	11%	2.7
CPU	138	93	95%	92%	17%	3.2
Tree	33	16	50%	63%	16%	2.1
Fat	1135	1086	29%	30%	14%	1.1
Wage	2969	208	34%	57%	4.5%	1.4
Attend	1402	63	29%	42%	14%	1.7
Strike	54	48	38%	17%	59%	1.9

Large

Med

Small

PIP: Positive impact pattern

local model reduces RMSE by 10+%

Diff = ratio of largest coefficien

of pairs of local models

Large: CPXR has large RMSE reduction;
Small: CPXR has low RMSE reduction

Diverse Predictor-Response Relationships May Neutralize Each Other at High Level

- Example: We considered a set S1={V1,V2,V3,V4} of 4 vars, then S2=S1 U {V5,V6}, on **soil water content data**
- LR model on S2 is not more accurate than LR model on S1
  - Ditto for PLR, SVR, BART
- CPXR’s model on S2 gives 20% RMSE improvement over CPXR’s model on S1
  - The new variables are involved in most of the diverse PR relationships (the patterns in the PXR model)
  - These relationships somehow cancelled each other’s effect, at the whole data set level. ➔ missed by LR etc.

Other usages of CPXR, besides building accurate prediction models

- Analysis on a given prediction model (next slide)
  - On what kinds of data it make large prediction errors
  - How to correct those prediction errors
- Do important models in science and medicine
  - **have systematic mistakes?**
- Analysis on comparing two given prediction models, w.r.t. their differences
- Discovering policy errors, niche opportunities, …
- Discovering true importance of variables
- Discovering intricate multi-variable interactions
- ……

WRIGHT STATE UNIVERSITY

How CPXR can help analyze given models

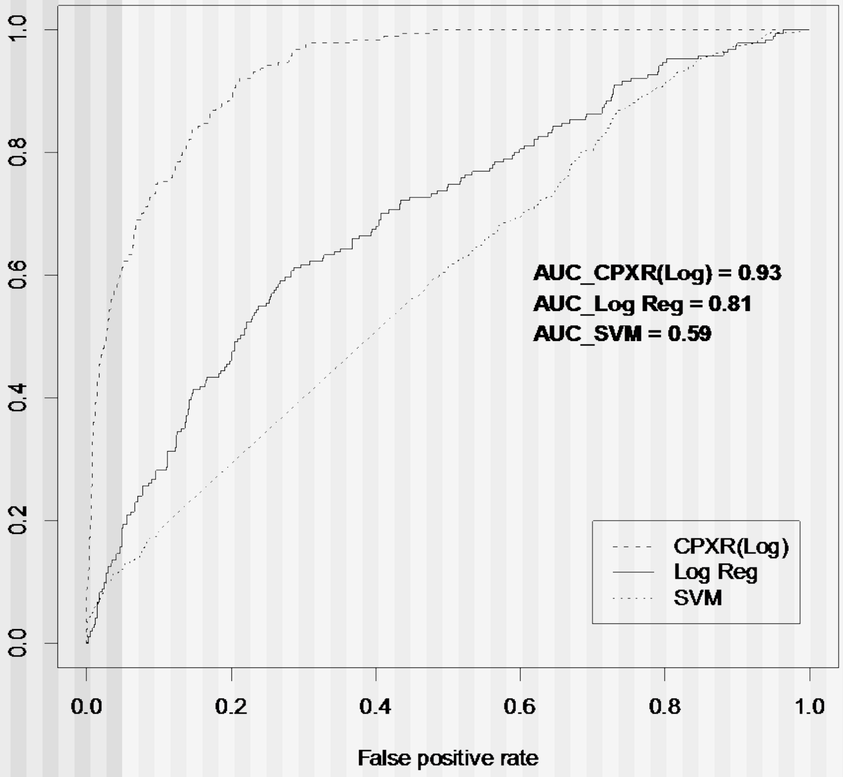
- Starting with training dataset
- Use a given model f0 as a baseline regression model
- Split data into LE (large error) and SE (small error), based on f0’s prediction error [these are the classes for CPs]
- Mine CPs; Remove redundant CPs that are not very useful
- Build corresponding local regression models for remaining CPs, and select CPs to construct PXR

- The patterns in the computed PXR model characterize how/where f0 makes significant prediction errors; the local models show how to correct those errors

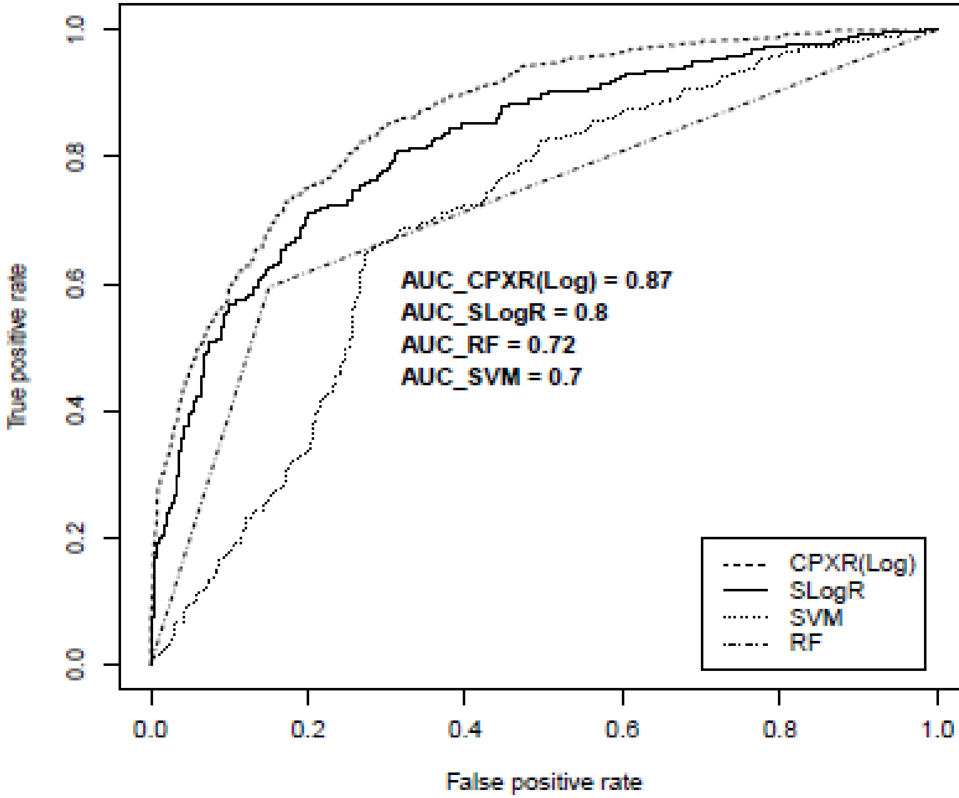
CPXR for logistic regression modeling and results on outcome prediction for HF/TBI

- The PXR-CPXR approach is not limited to using LR and PLR.
- We adapted it for logistic regression to get CPXR(Log)
- We used CPXR(Log) for outcome prediction for traumatic brain injury patients & for survival prediction for heart failure patients.
- CPXR(Log) is much more accurate than standard logistic regression and SVM
- CPXR(Log) also identifies important variables that are considered unimportant by standard logistic regression

AUC of ROC curves for CPXR(Log) and other methods (on HF and TBI)



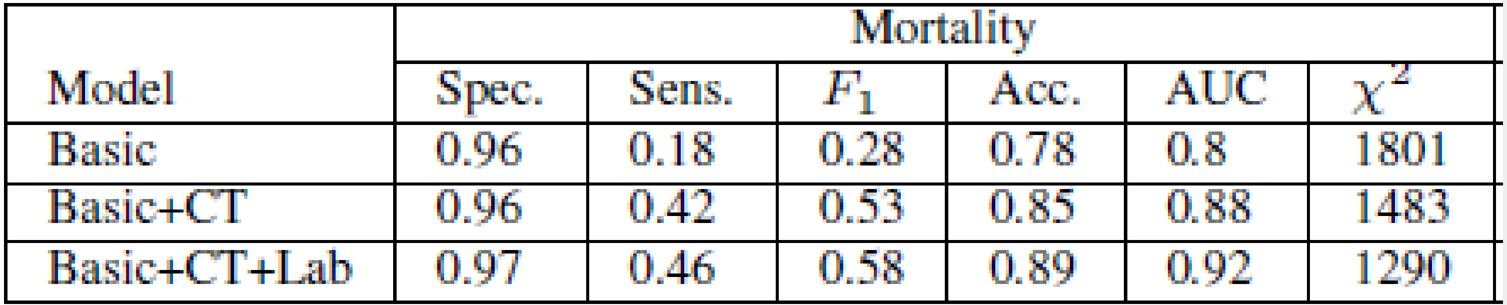
HF



TBI

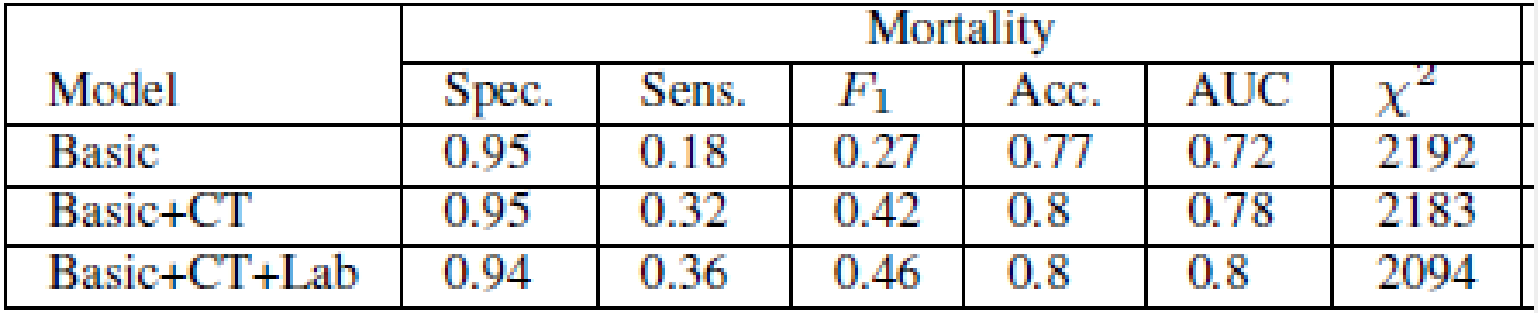
CPXR typically gets larger improvement when using more variables

TABLE II: CPXR(Log) performance on accuracy



Model	Mortality					
	Spec.	Sens.	F_1	Acc.	AUC	χ^2
Basic	0.96	0.18	0.28	0.78	0.8	1801
Basic+CT	0.96	0.42	0.53	0.85	0.88	1483
Basic+CT+Lab	0.97	0.46	0.58	0.89	0.92	1290

TABLE I: SLogR performance on accuracy



Model	Mortality					
	Spec.	Sens.	F_1	Acc.	AUC	χ^2
Basic	0.95	0.18	0.27	0.77	0.72	2192
Basic+CT	0.95	0.32	0.42	0.8	0.78	2183
Basic+CT+Lab	0.94	0.36	0.46	0.8	0.8	2094

Basic,CT, Lab:
3 sets of variables

Basic+CT:
Union of CT and Basic

CPXR identifies useful variables ignored by SLogR

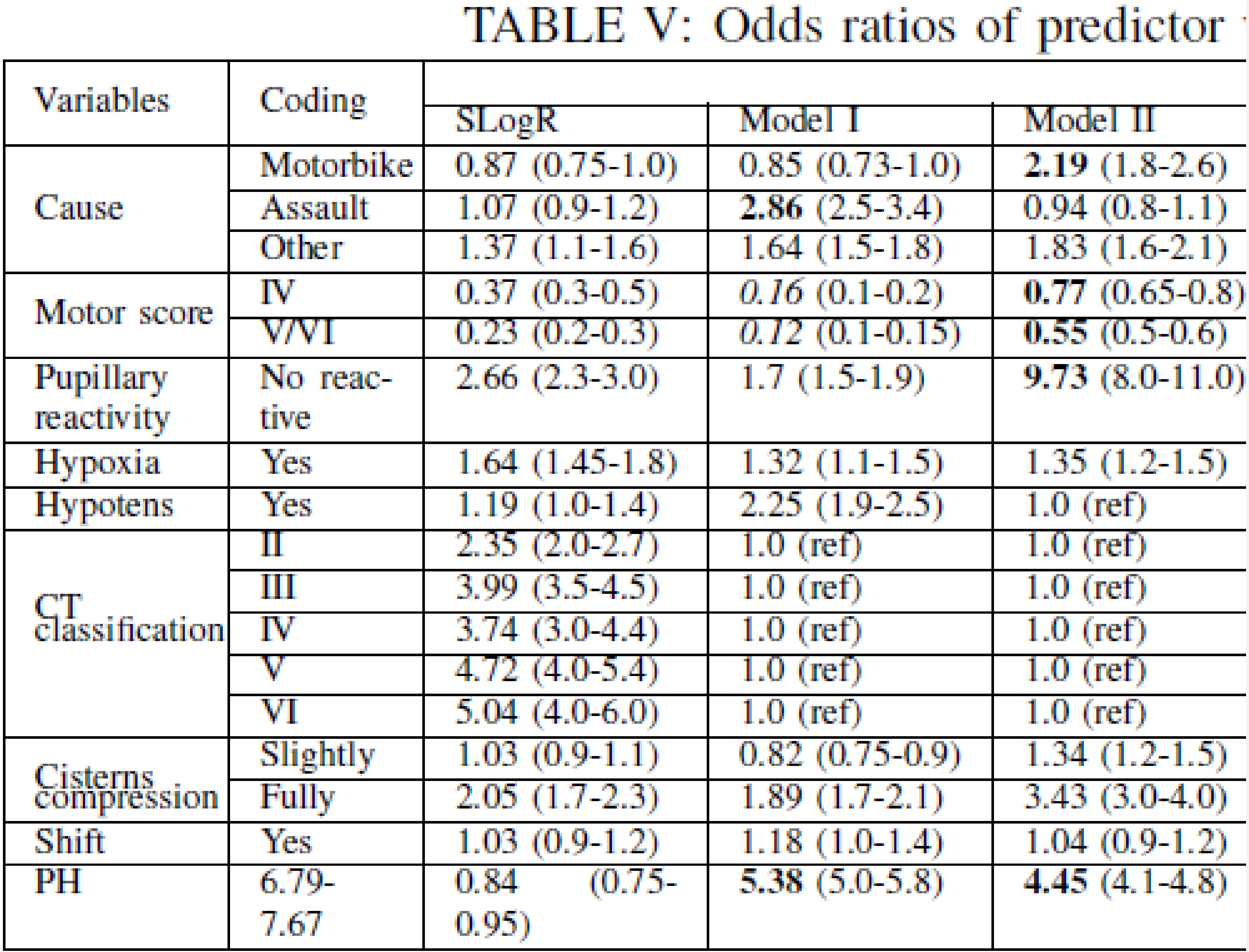


TABLE V: Odds ratios of predictor

Variables	Coding	SLogR	Model I	Model II
Cause	Motorbike	0.87 (0.75-1.0)	0.85 (0.73-1.0)	**2.19** (1.8-2.6)
	Assault	1.07 (0.9-1.2)	**2.86** (2.5-3.4)	0.94 (0.8-1.1)
	Other	1.37 (1.1-1.6)	1.64 (1.5-1.8)	1.83 (1.6-2.1)
Motor score	IV	0.37 (0.3-0.5)	*0.16* (0.1-0.2)	**0.77** (0.65-0.8)
	V/VI	0.23 (0.2-0.3)	*0.12* (0.1-0.15)	**0.55** (0.5-0.6)
Pupillary reactivity	No reactive	2.66 (2.3-3.0)	1.7 (1.5-1.9)	**9.73** (8.0-11.0)
Hypoxia	Yes	1.64 (1.45-1.8)	1.32 (1.1-1.5)	1.35 (1.2-1.5)
Hypotens	Yes	1.19 (1.0-1.4)	2.25 (1.9-2.5)	1.0 (ref)
CT classification	II	2.35 (2.0-2.7)	1.0 (ref)	1.0 (ref)
	III	3.99 (3.5-4.5)	1.0 (ref)	1.0 (ref)
	IV	3.74 (3.0-4.4)	1.0 (ref)	1.0 (ref)
	V	4.72 (4.0-5.4)	1.0 (ref)	1.0 (ref)
	VI	5.04 (4.0-6.0)	1.0 (ref)	1.0 (ref)
Cisterns compression	Slightly	1.03 (0.9-1.1)	0.82 (0.75-0.9)	1.34 (1.2-1.5)
	Fully	2.05 (1.7-2.3)	1.89 (1.7-2.1)	3.43 (3.0-4.0)
Shift	Yes	1.03 (0.9-1.2)	1.18 (1.0-1.4)	1.04 (0.9-1.2)
PH	6.79-7.67	0.84 (0.75-0.95)	**5.38** (5.0-5.8)	**4.45** (4.1-4.8)

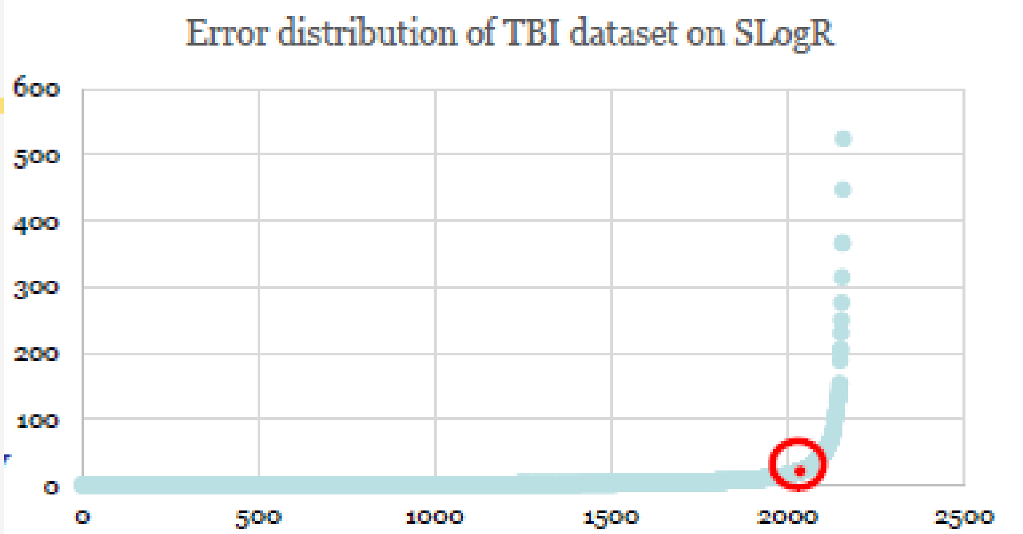
Summary of strength of PXR/CPXR

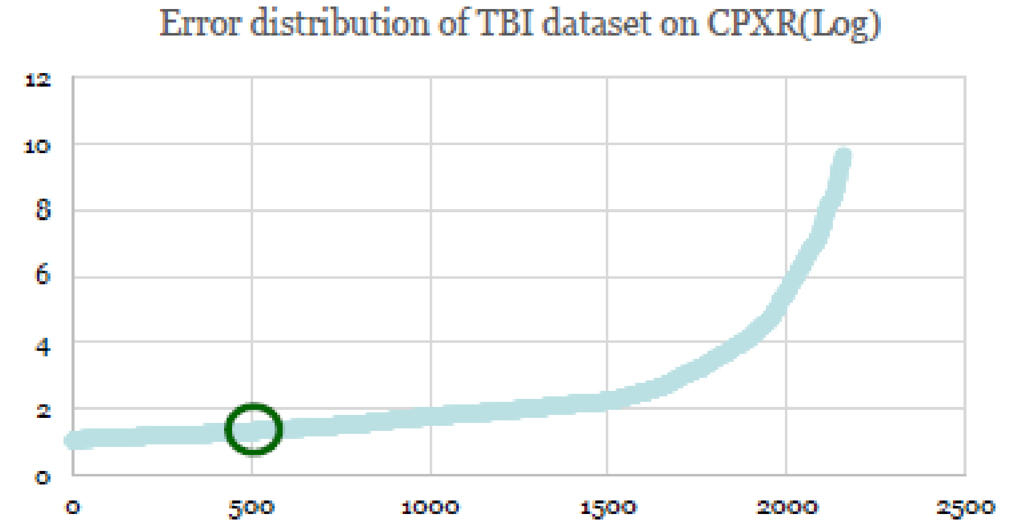
- Building accurate models (much better than existing methods)
  - Also easier to interpret
- Giving “conditions where given model makes large prediction errors” & “ways to correct those errors”
- Using pattern-model pairs to model/represent diverse predictor-response variable relationships
- The approach is highly suited to high dimensional data
- Offering insights to mistakes in business strategies
- Main idea: Using patterns to identify data groups where given model makes large prediction errors that can be corrected systematically

Observation: For many prediction models, the residuals contain exploitable structures!

- How often you hear “the residuals are noise”
- It is incorrect
- There is structure “in the residuals”!

Y-axis: Pearson residuals ➔


Error distribution of TBI dataset on SLogR


Error distribution of TBI dataset on CPXR(Log)

Comparison with boosting and ensemble methods

- CPXR can be viewed as **opportunity driven boosting**
  - It considers one pattern’s matching dataset at a time and looks to use patterns that represent “good opportunities”
  - Standard boosting considers all “incorrectly predicted data” at once (less accurate)
- CPXR builds **small, easy to explain committees**, using TRR optimization as objective
  - Other ensemble methods typically rely on chance to produce committees; produce large committees (less accurate)

If you cannot build accurate model for your data/problem

- You may think “the data does not contain enough information for prediction”
- You should not give up:
  - The data may contain useful information
  - But existing methods did not find them
- CPXR may be able to help!

Outline

- Introduction
- Pattern aided regression models: PXR

← New regression model type

- Diverse predictor-response relationships

←Reason why existing algorithms perform poorly?!

- Contrast pattern aided regression algorithm: CPXR
- Experimental evaluation
- CPXR(Log): Logistic variant of CPXR
- Example applications of CPXR:
  - Water content prediction for soil
  - Traumatic brain injury (TBI) outcome prediction
  - Heart failure (HF) survival prediction
- Potentials of CPXR and insights

WRIGHT STATE UNIVERSITY

http://cecs.wright.edu/~gdong/

Questions

Next step: help companies to

- build new accurate risk models
- analyze/improve existing risk models

Publications related to CPXR

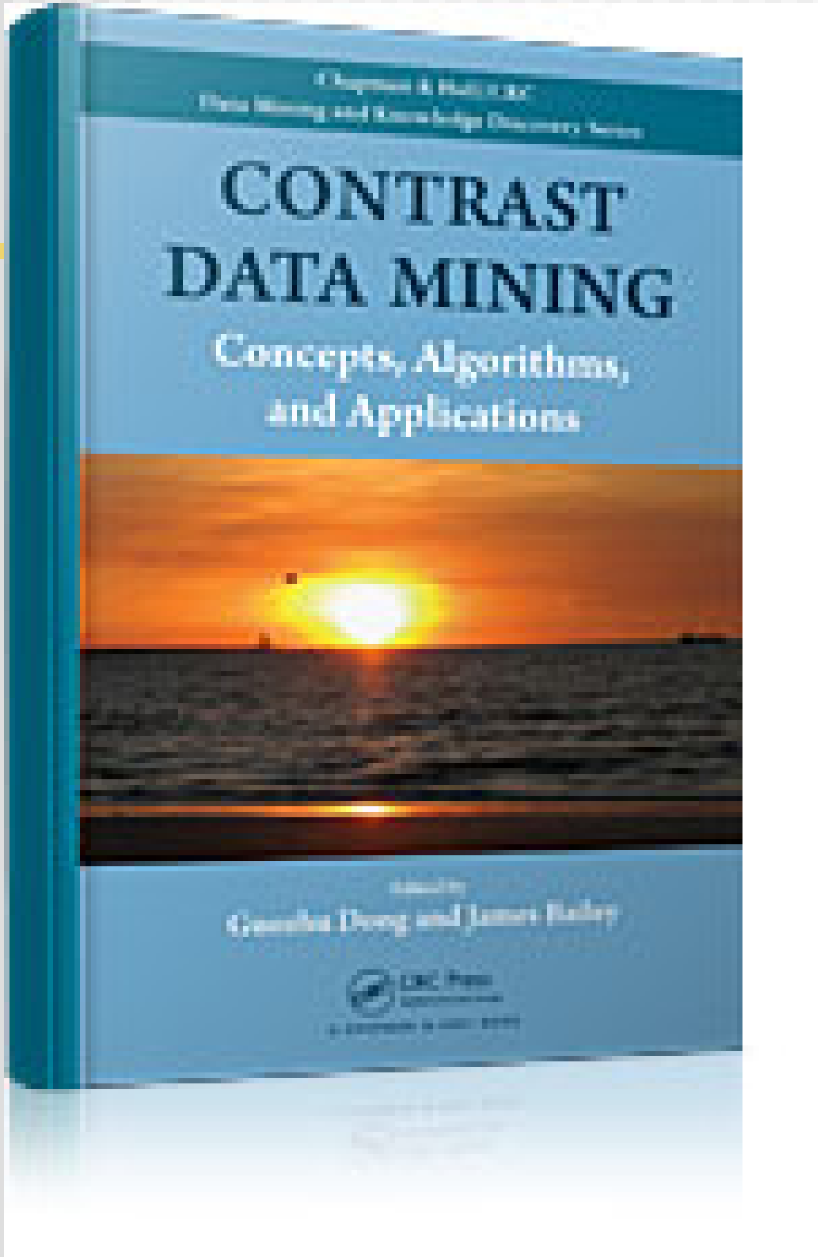
Guozhu Dong and Vahid Taslimitehrani. Pattern-Aided Regression Modeling and Prediction Model Analysis. IEEE Transactions on Knowledge and Data Engineering. Vol 27, Issue 9, pp 2452–2465, 2015.

Guozhu Dong and Vahid Taslimitehrani. Pattern Aided Classification. To appear in SIAM International Conference on Data Mining (SDM), 2016.

Behzad Ghanbarian, Vahid Taslimitehrani, Guozhu Dong, Yakov A. Pachepsky. Sample dimensions effect on prediction of soil water retention curve and saturated hydraulic conductivity. Journal of Hydrology. 528 (2015) 127–137.

Vahid Taslimitehrani and Guozhu Dong. A New CPXR Based Logistic Regression Method and Clinical Prognostic Modeling Results Using the Method on Traumatic Brain Injury. IEEE International Conference on BioInformatics and BioEngineering (BIBE). Nov. 2014. **Winner of Best Student Paper Award**.

Vahid Taslimitehrani, Guozhu Dong, Naveen L. Pereira, Maryam Panahiazar, Jyotishman Pathak. Developing EHR-driven Heart Failure Risk Prediction Models using CPXR(Log) with the Probabilistic Loss Function. To appear in Journal of Biomedical Informatics.


Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series
CONTRAST DATA MINING
Concepts, Algorithms, and Applications
Edited by
Guozhu Dong and James Bailey
CRC Press

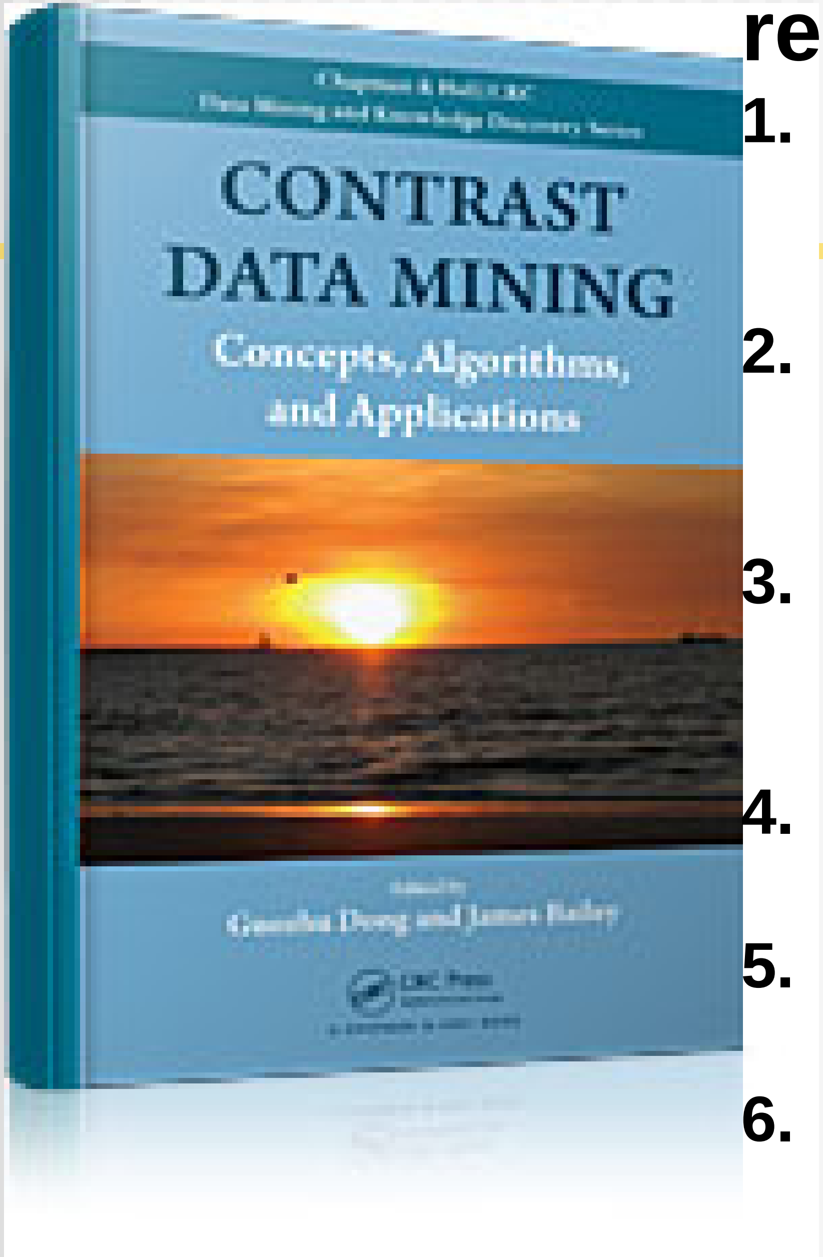
Related to CPXR: We published this book in 2012; 3 out of 6 parts on applications

1. Preliminaries and Measures on Contrasts
2. Contrast Mining Algorithms
3. Mining Generalized Contrasts
4. Contrast Mining for Classification & Clustering
5. Contrast Mining for Bioinformatics & Chemoinformatics
6. Contrast Mining for Special Application Domains

44 contributing authors, from ~dozen countries; not comprehensive

Methods used by many scientists.

Guozhu Dong’s recent results in this area


Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series
CONTRAST DATA MINING
Concepts, Algorithms, and Applications
Edited by
Guozhu Dong and James Bailey
CRC Press

1. **CAEP-style classification**: discriminative power aggregation of emerging patterns
2. **Outlier detection / intrusion detection**: almost model free; using discriminative pattern length
3. **Clustering quality evaluation** using patterns (quality, abundance, diversity): no distance function
4. **CP based clustering** and cluster description: no distance func needed
5. **Interaction based gene/SNP ranking** for complex diseases
6. **Contrast pattern aided regression:** Effectively handling diverse predictor-response relationships