Predicting Quality-relevant Variables in Industrial Process via Causality Analysis

Reporter: Liang Cao

Supervisor: Bhushan Gopaluni

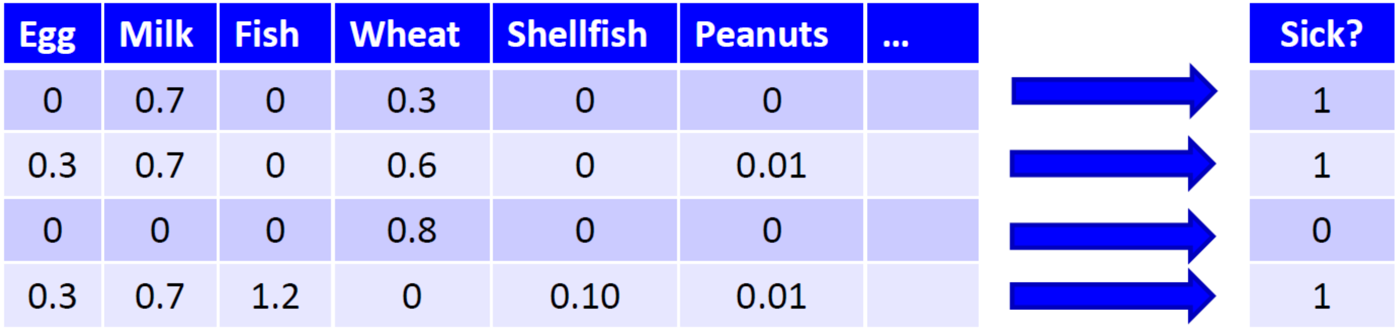
Department of Chemical and Biological Engineering,
University of British Columbia

2020.02.14

Situation

✓ Chemical process usually has thousands of process variables (X) available to predict quality-relevant variables (Y).

✓ How to find the features of X that are important for predicting Y (which features of X helps predict Y) is one of the most important problems in ML, but very messy.

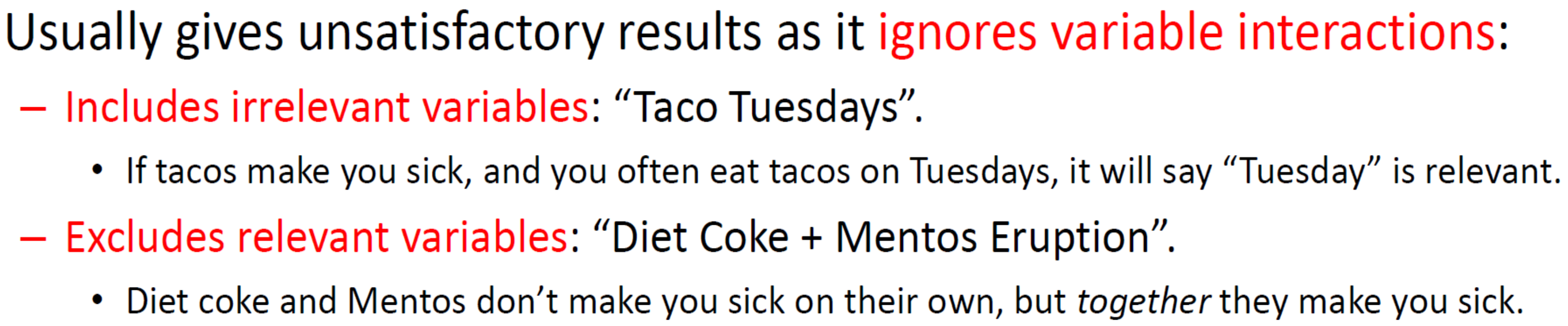


Egg	Milk	Fish	Wheat	Shellfish	Peanuts	…	Sick?
0	0.7	0	0.3	0	0		1
0.3	0.7	0	0.6	0	0.01		1
0	0	0	0.8	0	0		0
0.3	0.7	1.2	0	0.10	0.01		1

✓ We want to know which foods are important for predicting “sick”

Problem

- ✓ A common way to do feature selection is compute the correlation between feature values X_i and Y, if the correlation is above certain value, take these features.

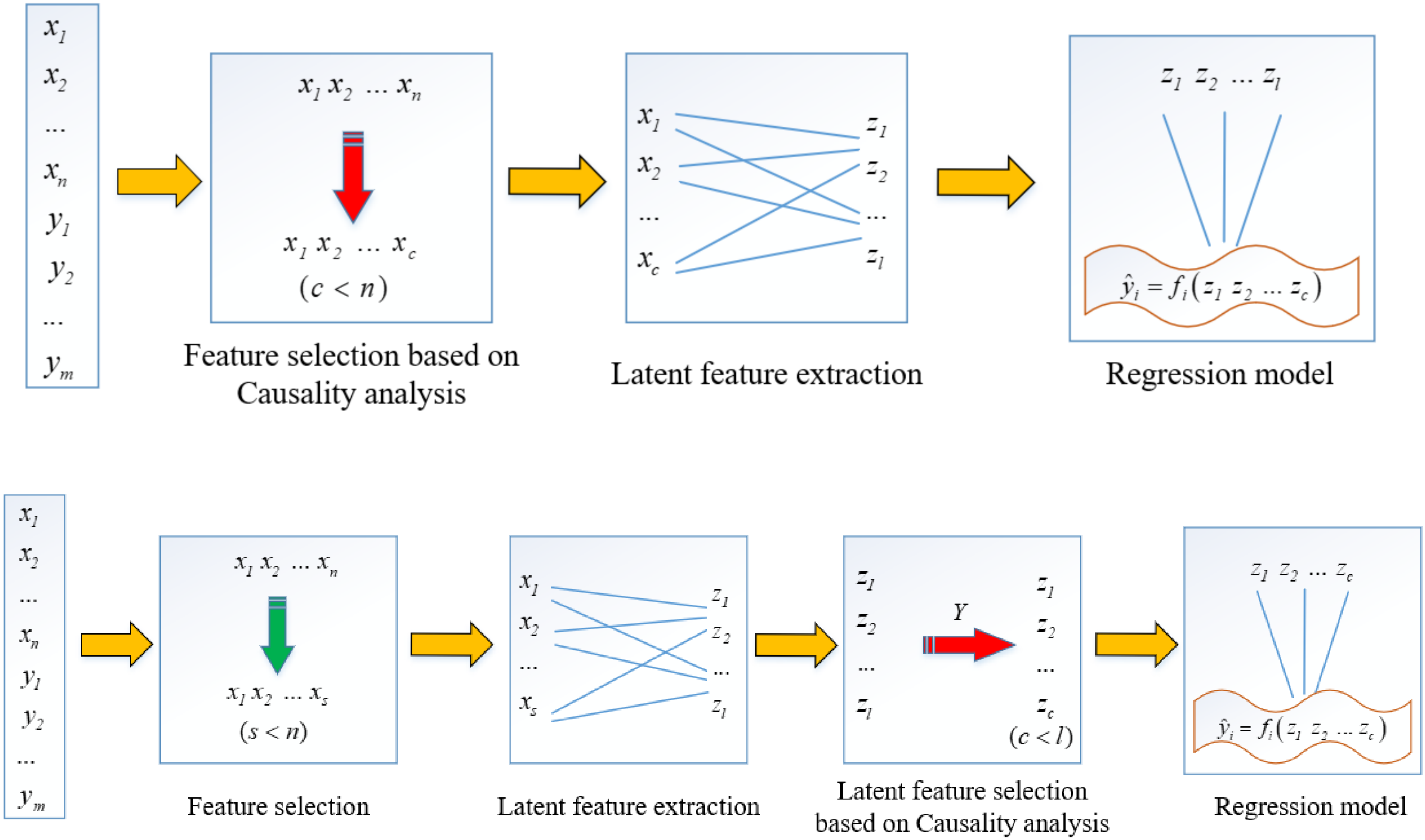
- ✓ 

Usually gives unsatisfactory results as it ignores variable interactions:

- Includes irrelevant variables: “Taco Tuesdays”.
  - If tacos make you sick, and you often eat tacos on Tuesdays, it will say “Tuesday” is relevant.
- Excludes relevant variables: “Diet Coke + Mentos Eruption”.
  - Diet coke and Mentos don’t make you sick on their own, but *together* they make you sick.

- ✓ To build simpler, more powerful, more interpretable model, we use causality analysis to find causal features.

How



Feature selection based on Causality analysis

Latent feature extraction

Regression model

Feature selection

Latent feature extraction

Latent feature selection based on Causality analysis

Regression model

Fig.1 Flowchart of two kinds of causality-based inferential sensors

How

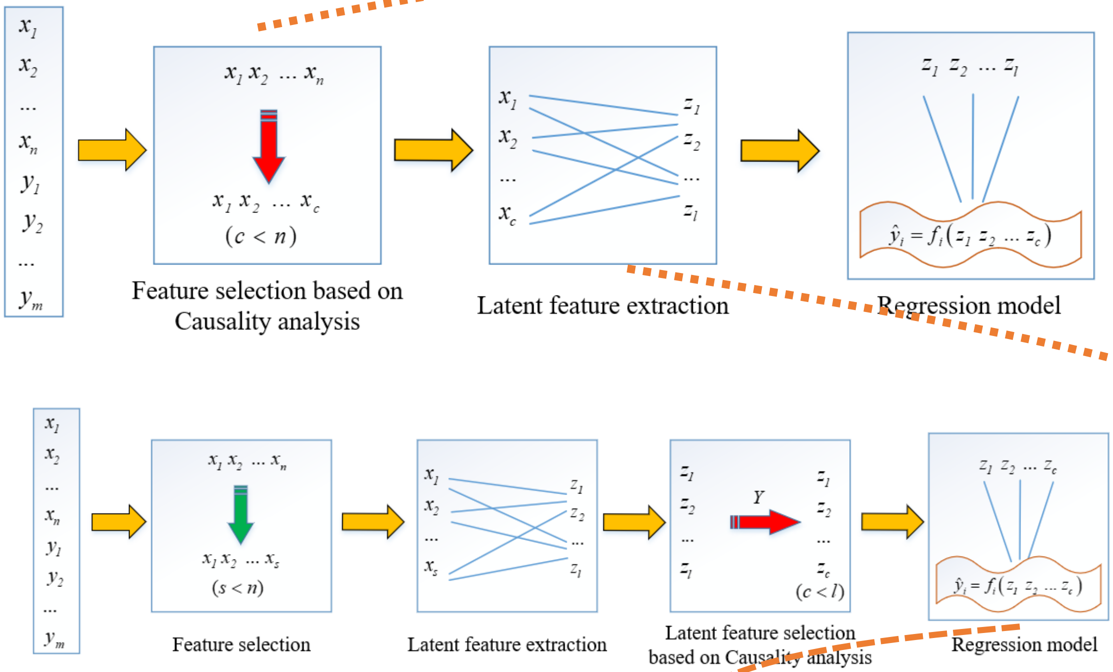


Fig.1 Flowchart of two kinds of causality-based inferential sensors

Causality analysis methods:

Granger Causality analysis
Transfer Entropy
…..

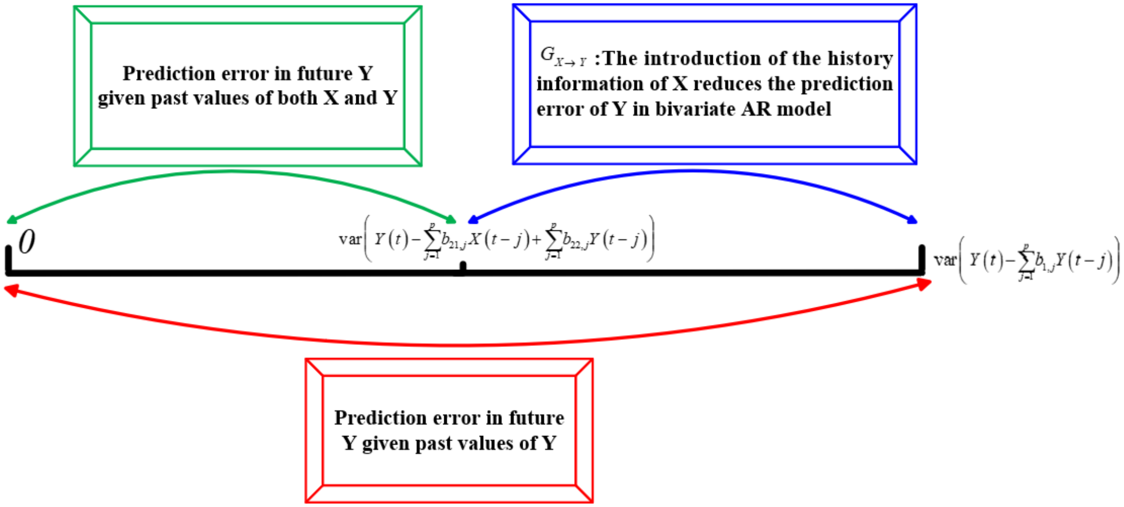
Latent feature extraction methods:

PCA ICA SFA CVA
…..

Regression methods:

LS, Bayesian, Decision tree, SVM, Neural networks
…..

How



Prediction error in future Y given past values of both X and Y

$G_{X\to Y}$:The introduction of the history information of X reduces the prediction error of Y in bivariate AR model

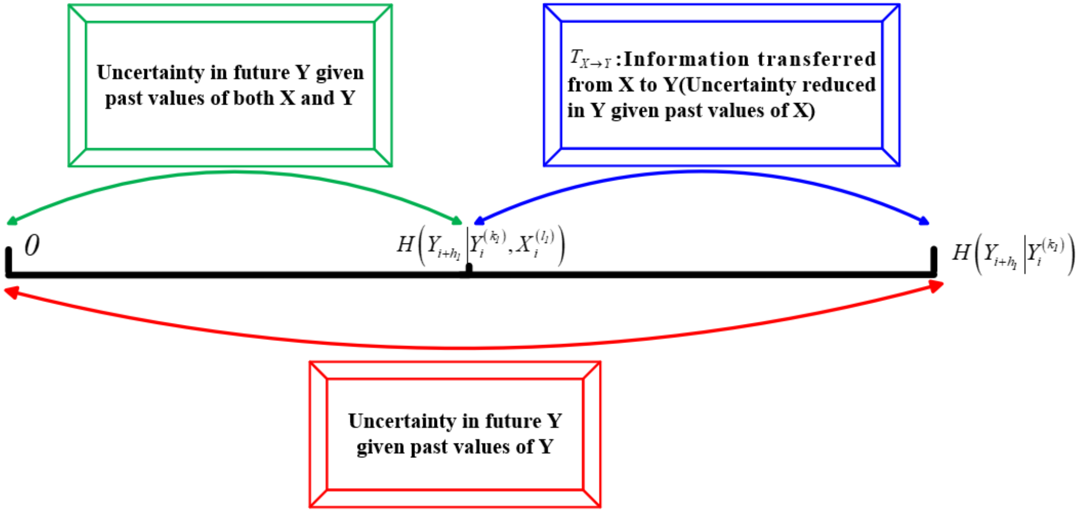
0

$\mathrm{var}\left(Y(t)-\sum_{j=1}^{p} b_{21,j} X(t-j)+\sum_{j=1}^{p} b_{22,j} Y(t-j) \right)$

$\mathrm{var}\left(Y(t)-\sum_{j=1}^{p} b_{1,j} Y(t-j) \right)$

Prediction error in future Y given past values of Y

GCA



Uncertainty in future Y given past values of both X and Y

$T_{X\to Y}$:Information transferred from X to Y(Uncertainty reduced in Y given past values of X)

0

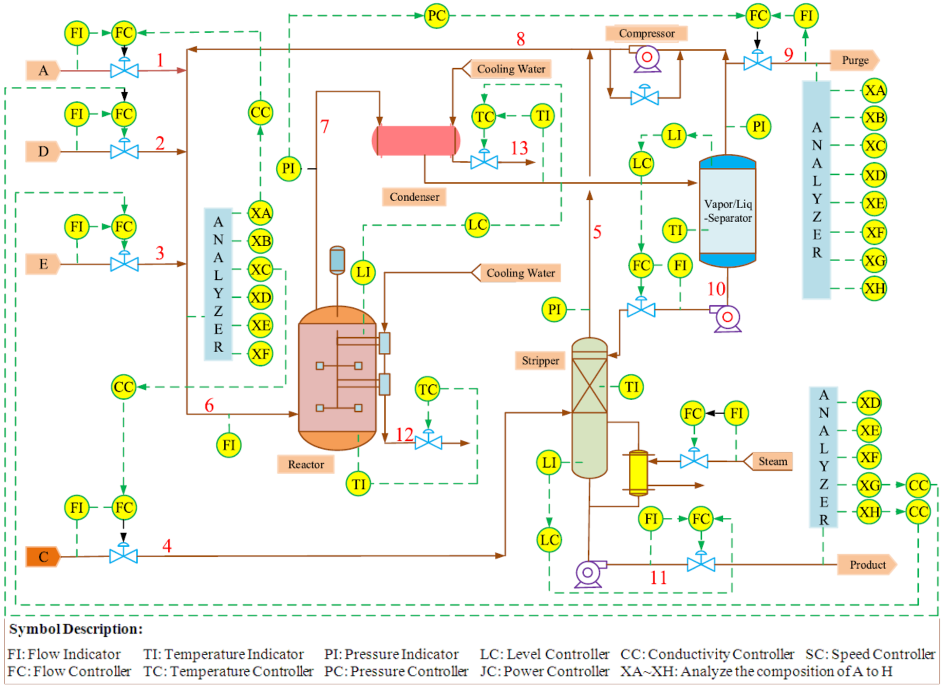
$H\left(Y_{i+h_1} \middle| Y_i^{(k_1)}, X_i^{(l_1)} \right)$

$H\left(Y_{i+h_1} \middle| Y_i^{(k_1)} \right)$

Uncertainty in future Y given past values of Y

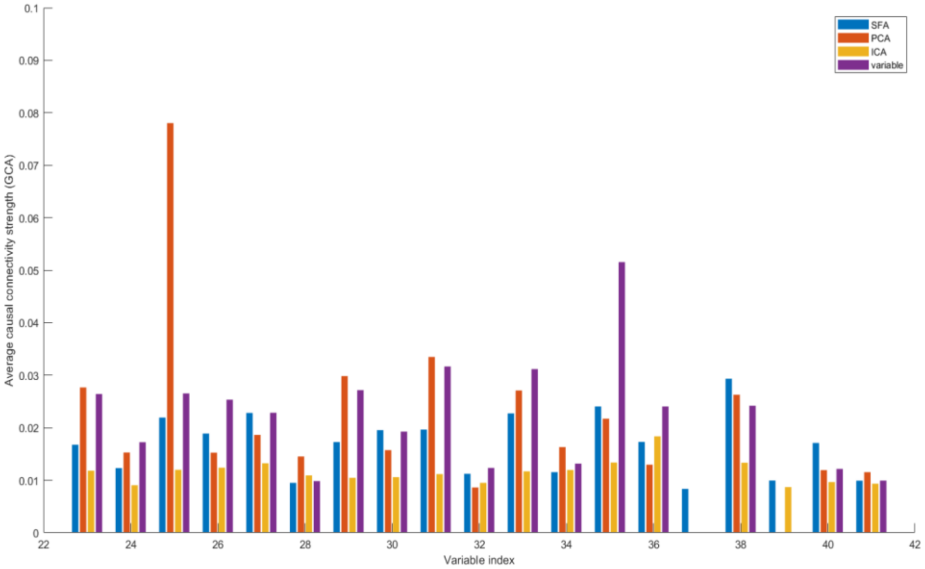
TE

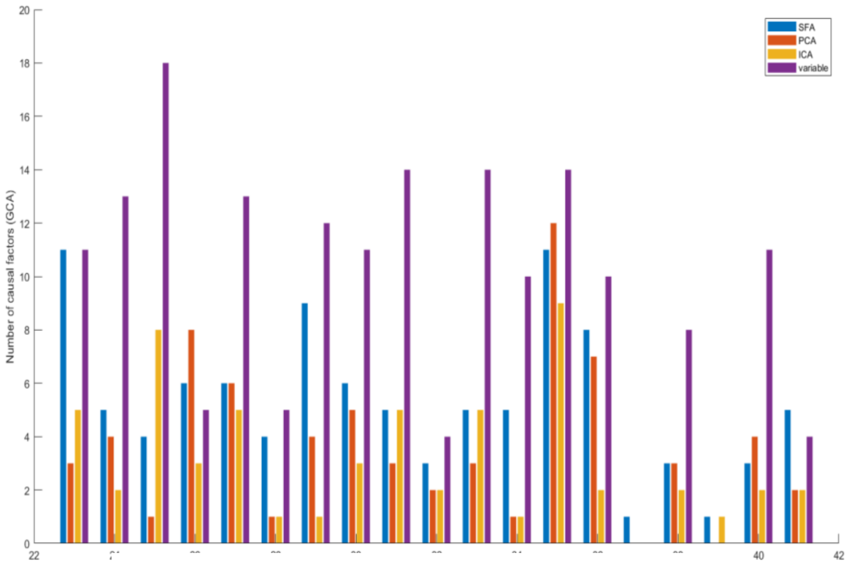
Results

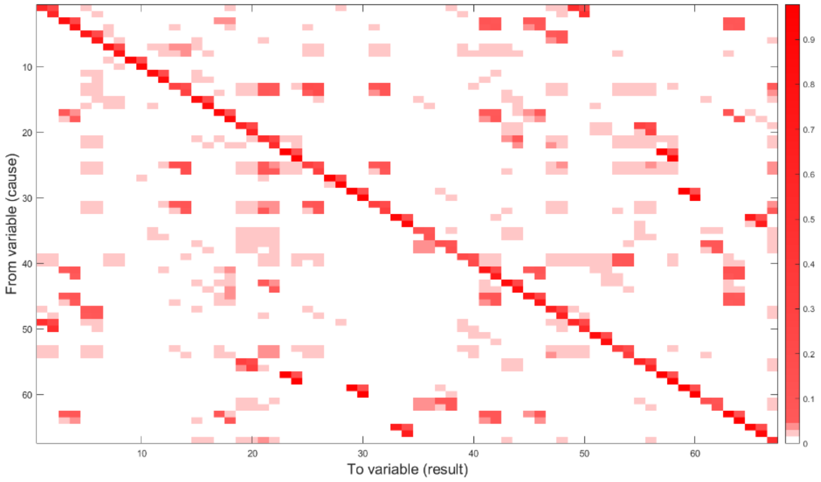


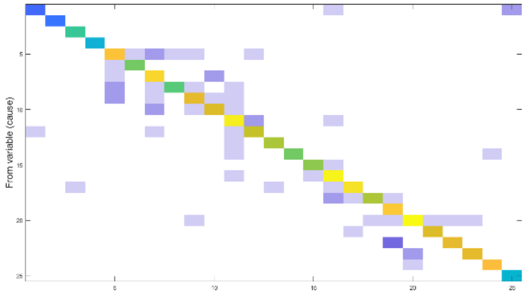
There are 52 different variables in this process, among which 33 variables can be measured in real time while another 19 variables need to be analyzed respectively. Hence, 33 variables are chosen as the process data and 19 variables are seen as the quality-relevant variables to be predicted (only use normal data, no fault data). We choose and 33 process variables as X and the variable 31 as Y.

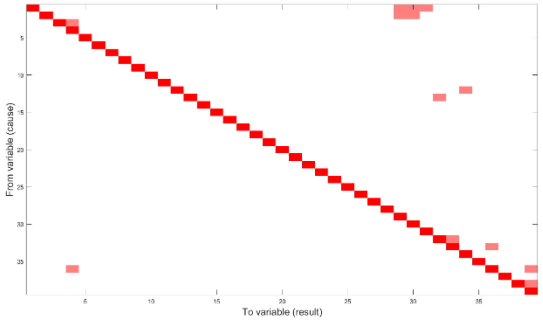
Results


SFA
PCA
ICA
variable
Average causal connectivity strength (GCA)
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
22
24
26
28
30
32
34
36
38
40
42
Variable index

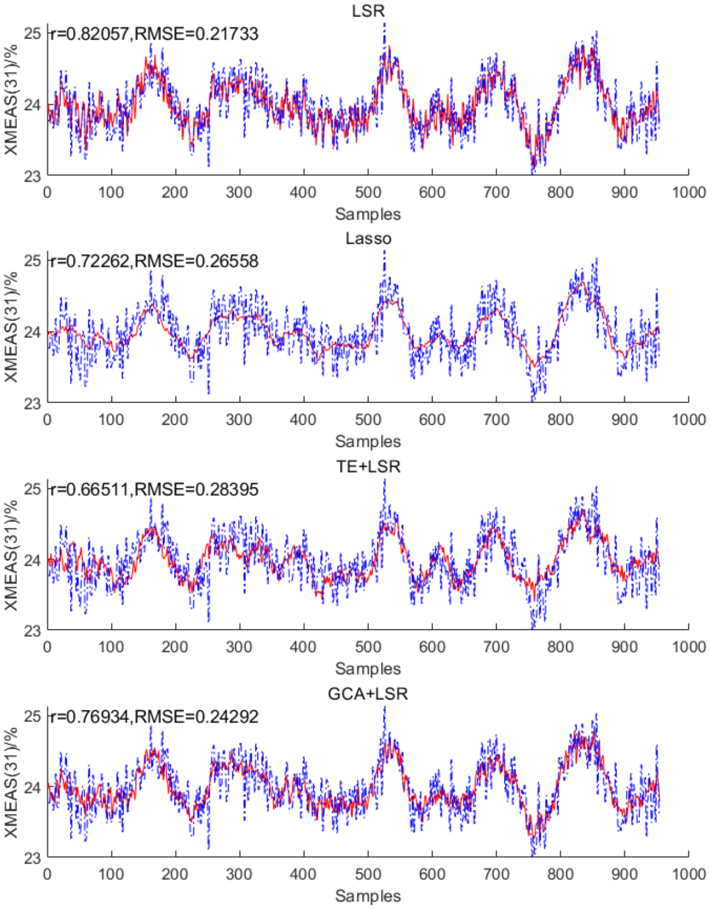

SFA
PCA
ICA
variable
Number of causal factors (GCA)
20
18
16
14
12
10
8
6
4
2
0
22
40
42

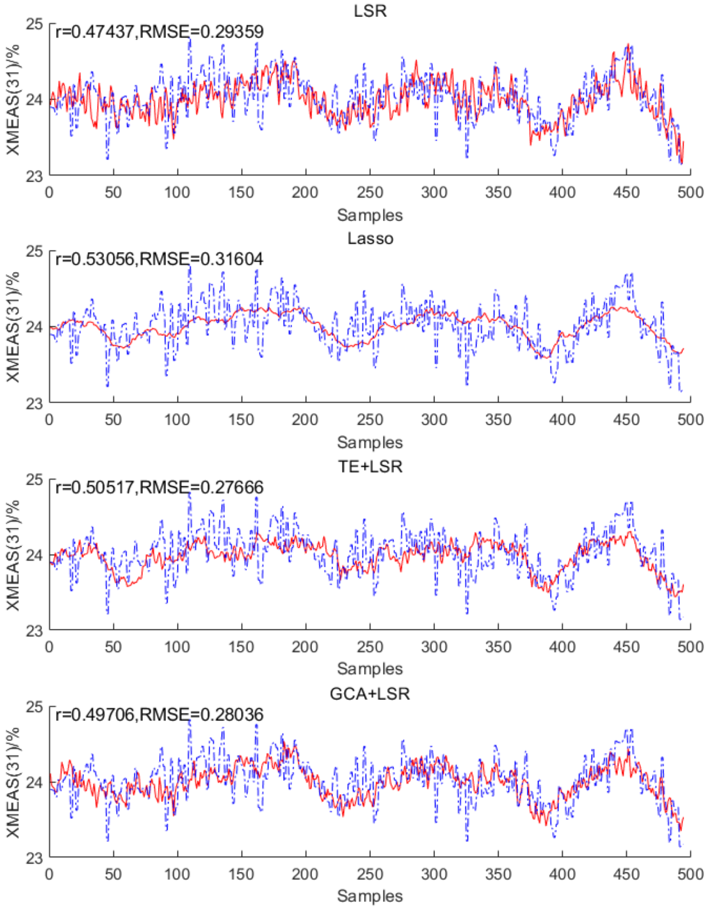

From variable (cause)
10
20
30
40
50
60
To variable (result)
10
20
30
40
50
60
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0


From variable (cause)
5
10
15
20
25
5
10
15
20
25

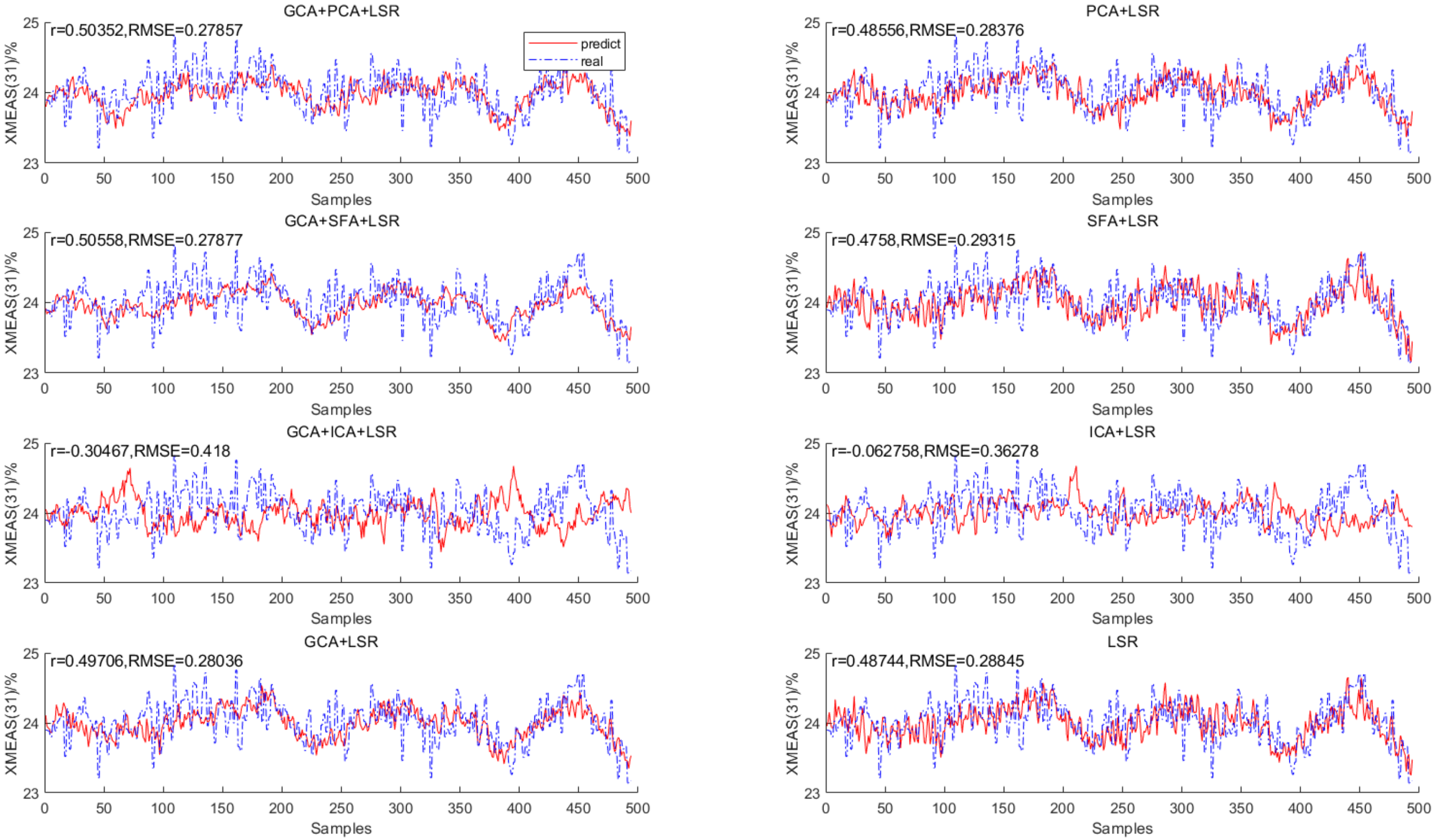

From variable (cause)
5
10
15
20
25
30
35
5
10
15
20
25
30
35
To variable (result)

Results

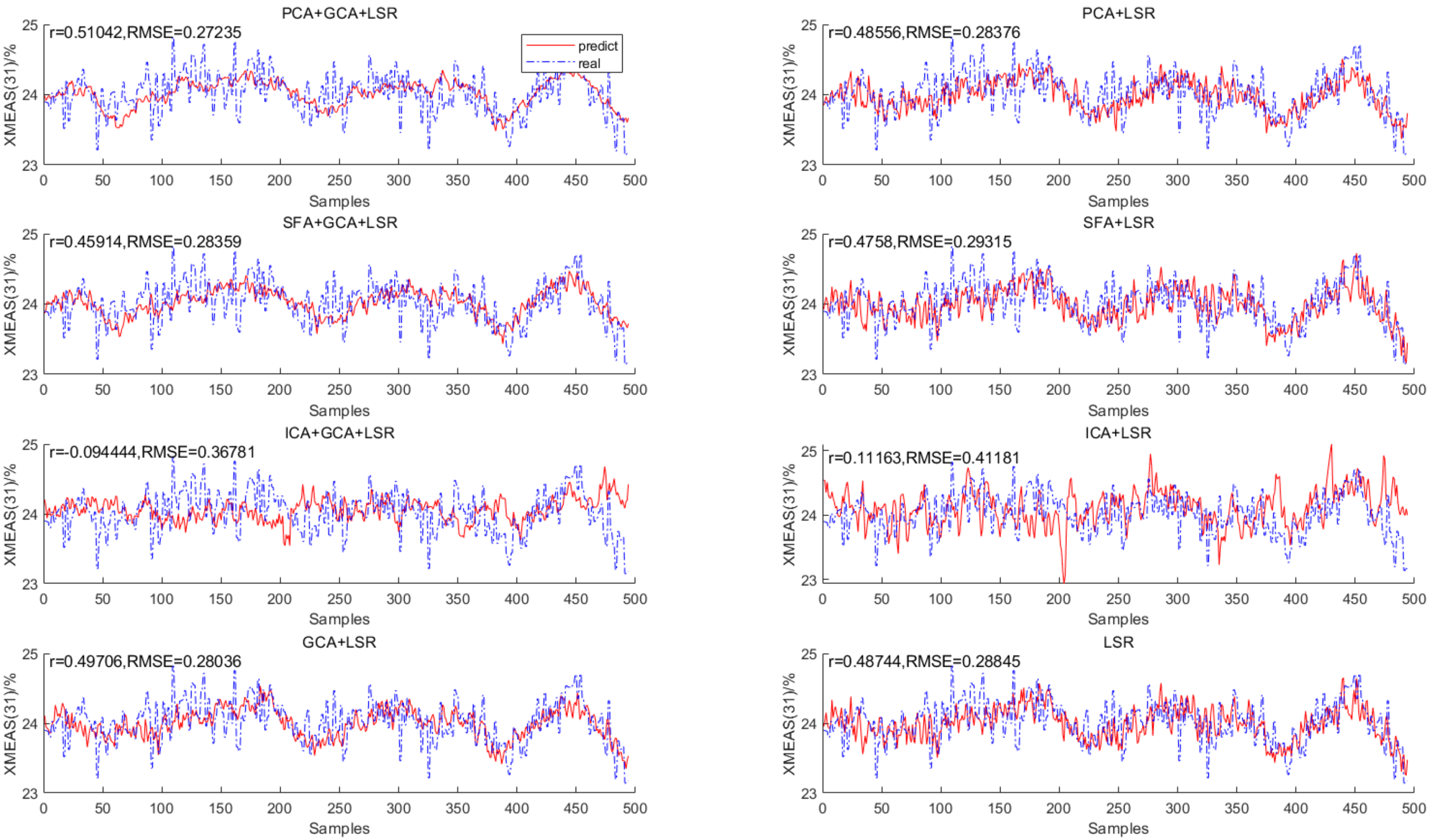

LSR
r=0.82057,RMSE=0.21733
Lasso
r=0.72262,RMSE=0.26558
TE+LSR
r=0.66511,RMSE=0.28395
GCA+LSR
r=0.76934,RMSE=0.24292
XMEAS(31)/%
23
24
25
Samples
0
100
200
300
400
500
600
700
800
900
1000


LSR
r=0.47437,RMSE=0.29359
Lasso
r=0.53056,RMSE=0.31604
TE+LSR
r=0.50517,RMSE=0.27666
GCA+LSR
r=0.49706,RMSE=0.28036
XMEAS(31)/%
23
24
25
Samples
0
50
100
150
200
250
300
350
400
450
500

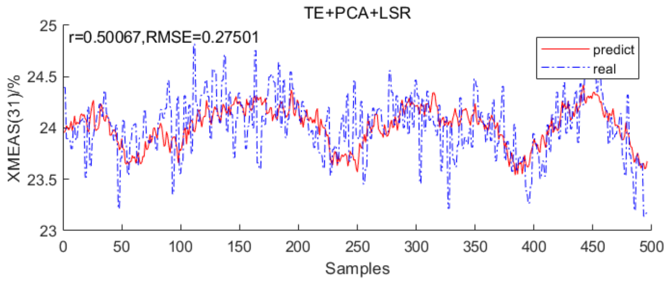
Results

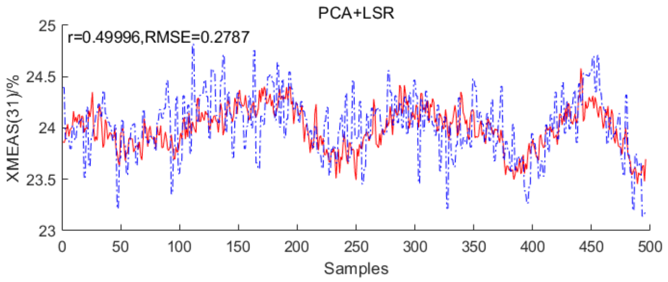

GCA+PCA+LSR
r=0.50352,RMSE=0.27857
predict
real
PCA+LSR
r=0.48556,RMSE=0.28376
GCA+SFA+LSR
r=0.50558,RMSE=0.27877
SFA+LSR
r=0.4758,RMSE=0.29315
GCA+ICA+LSR
r=-0.30467,RMSE=0.418
ICA+LSR
r=-0.062758,RMSE=0.36278
GCA+LSR
r=0.49706,RMSE=0.28036
LSR
r=0.48744,RMSE=0.28845
XMEAS(31)/%
23
24
25
Samples
0
50
100
150
200
250
300
350
400
450
500

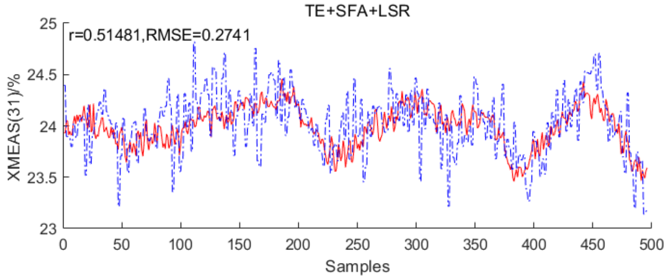
Results

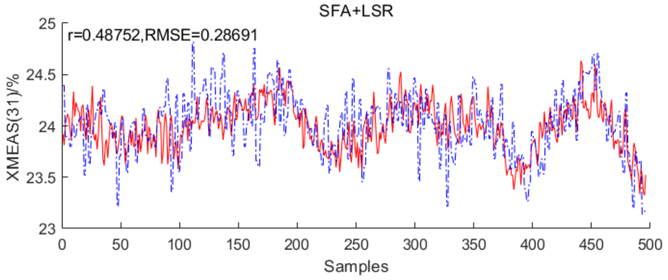

PCA+GCA+LSR
r=0.51042,RMSE=0.27235
predict
real
PCA+LSR
r=0.48556,RMSE=0.28376
SFA+GCA+LSR
r=0.45914,RMSE=0.28359
SFA+LSR
r=0.4758,RMSE=0.29315
ICA+GCA+LSR
r=-0.094444,RMSE=0.36781
ICA+LSR
r=0.11163,RMSE=0.41181
GCA+LSR
r=0.49706,RMSE=0.28036
LSR
r=0.48744,RMSE=0.28845
XMEAS(31)/%
25
24
23
Samples
0
50
100
150
200
250
300
350
400
450
500

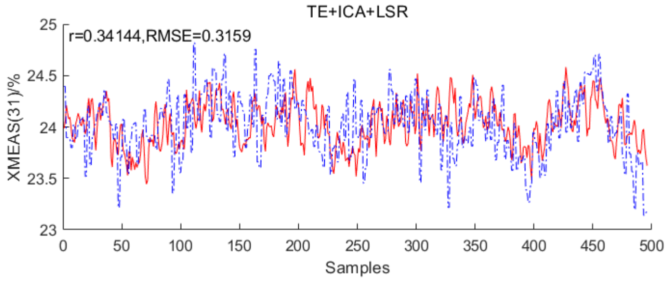
Results

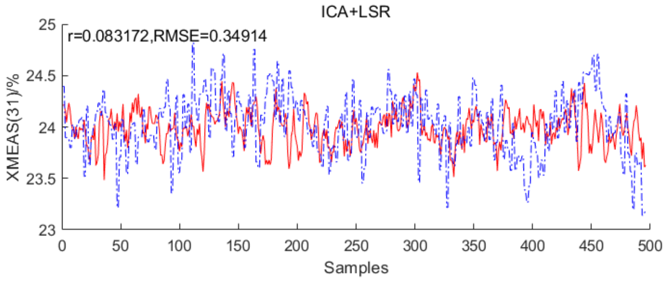

TE+PCA+LSR
r=0.50067,RMSE=0.27501
predict
real
XMEAS(31)/%
25
24.5
24
23.5
23
0
50
100
150
200
250
300
350
400
450
500
Samples

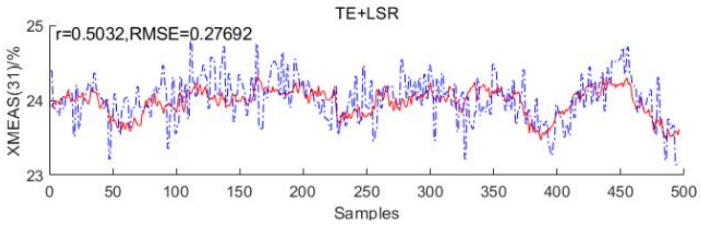

PCA+LSR
r=0.49996,RMSE=0.2787
XMEAS(31)/%
25
24.5
24
23.5
23
0
50
100
150
200
250
300
350
400
450
500
Samples

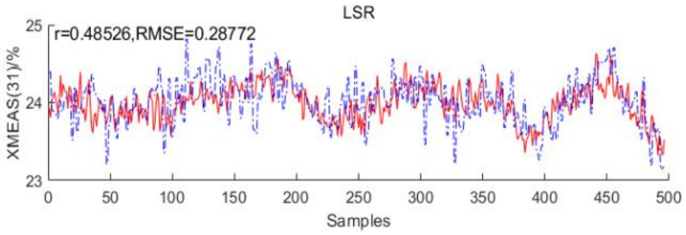

TE+SFA+LSR
r=0.51481,RMSE=0.2741
XMEAS(31)/%
25
24.5
24
23.5
23
0
50
100
150
200
250
300
350
400
450
500
Samples


SFA+LSR
r=0.48752,RMSE=0.28691
XMEAS(31)/%
25
24.5
24
23.5
23
0
50
100
150
200
250
300
350
400
450
500
Samples

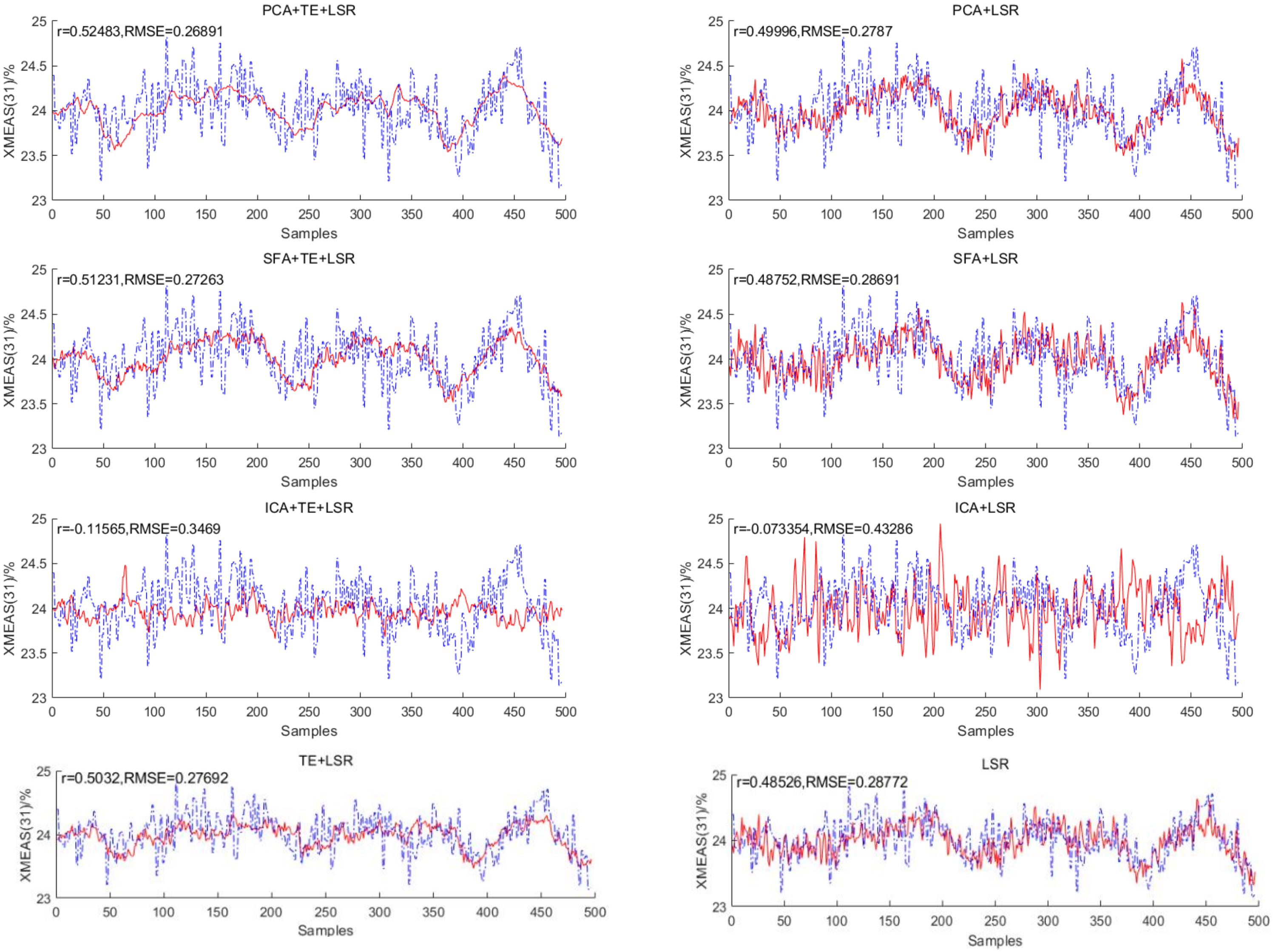

TE+ICA+LSR
r=0.34144,RMSE=0.3159
XMEAS(31)/%
25
24.5
24
23.5
23
0
50
100
150
200
250
300
350
400
450
500
Samples


ICA+LSR
r=0.083172,RMSE=0.34914
XMEAS(31)/%
25
24.5
24
23.5
23
0
50
100
150
200
250
300
350
400
450
500
Samples


TE+LSR
r=0.5032,RMSE=0.27692
XMEAS(31)/%
25
24
23
0
50
100
150
200
250
300
350
400
450
500
Samples


LSR
r=0.48526,RMSE=0.28772
XMEAS(31)/%
25
24
23
0
50
100
150
200
250
300
350
400
450
500
Samples

Results


PCA+TE+LSR
r=0.52483,RMSE=0.26891
PCA+LSR
r=0.49996,RMSE=0.2787
SFA+TE+LSR
r=0.51231,RMSE=0.27263
SFA+LSR
r=0.48752,RMSE=0.28691
ICA+TE+LSR
r=-0.11565,RMSE=0.3469
ICA+LSR
r=-0.073354,RMSE=0.43286
TE+LSR
r=0.5032,RMSE=0.27692
LSR
r=0.48526,RMSE=0.28772
XMEAS(31)/%
Samples
23
23.5
24
24.5
25
0
50
100
150
200
250
300
350
400
450
500