Subjective Video Quality Test via Crowdsourcing

Semester Project

Yinan Shi, Fengze Han, Zhong Chen, Hao Xu
Supervisor: Dr.-Ing. Christian Keimel, Dipl.-Ing. Thomas Volk, Philipp Paukner

Institute for Data Processing, Technical University of Munich

July 24, 2018

Table of contents

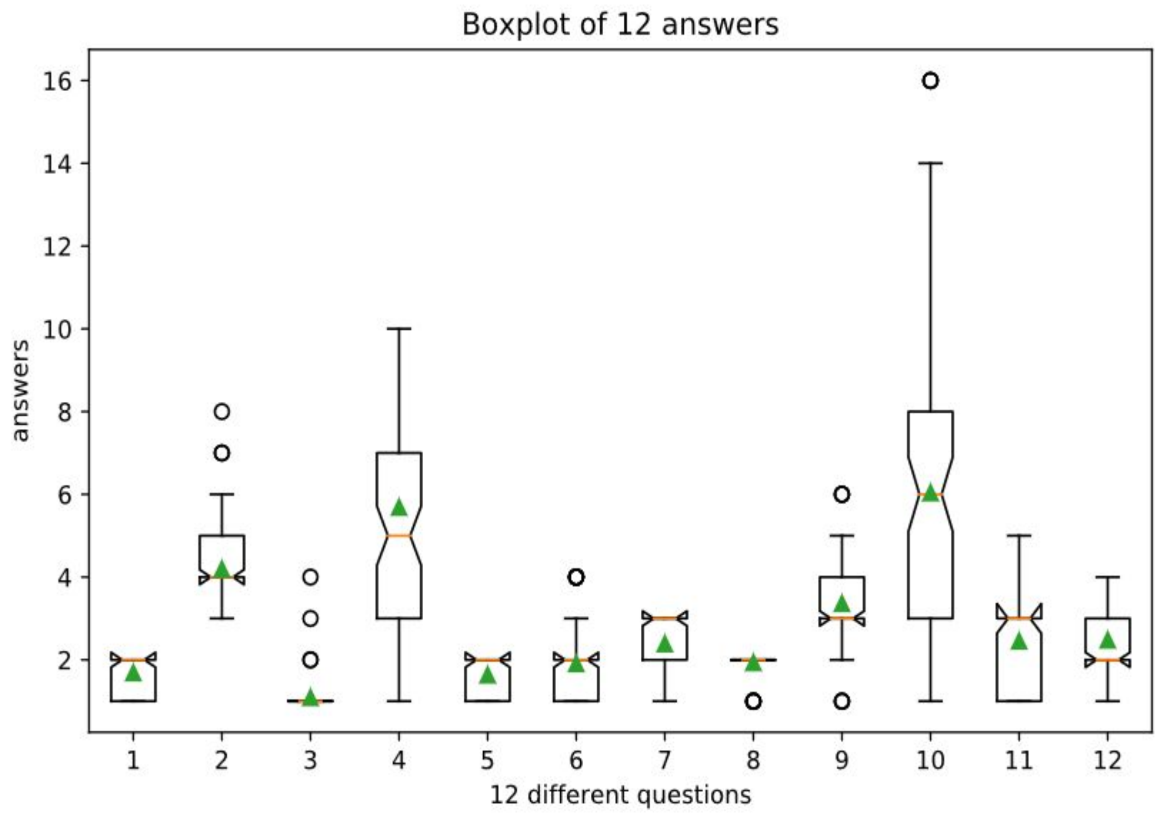
Part A: Crowdsourcing Results

- Demographics and Streaming Habits

Question 1: mostly Male

Question 3: mostly America

Question 8: almost 100% in evening



Part A: Crowdsourcing Results

- Demographics and Streaming Habits

Question 1: mostly Male

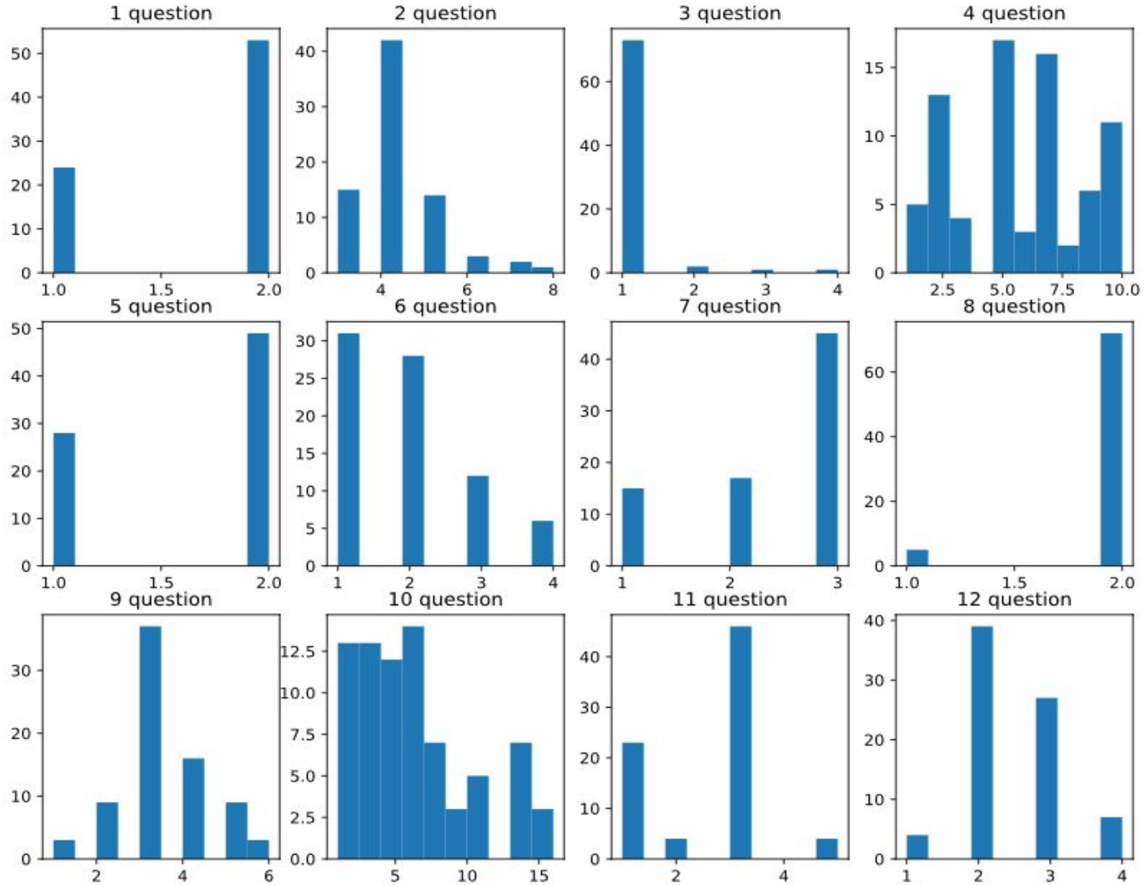
Question 3: mostly America

Question 8: almost 100% in evening

Question 2: 90% age 18-44
50% age 25-34

Question 6: ¾ extremely often and very often

Question 7: 60% no difference
23% on weekends



Part A: Crowdsourcing Results

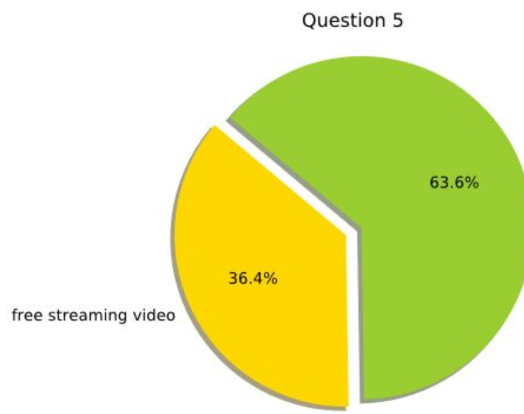
- Demographics and Streaming Habits

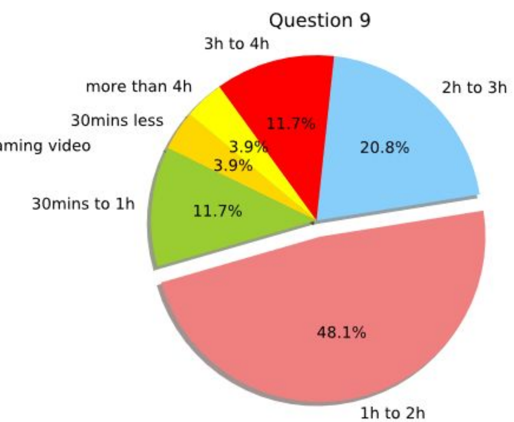
Question 5: 36.4% free
63.6% paid

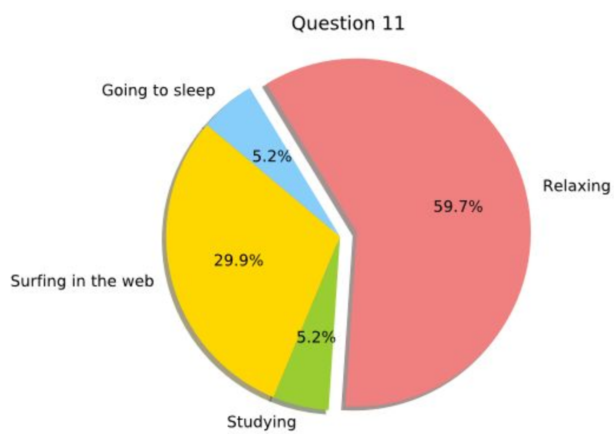
Question 9: half 1-2 h
85% > 1h

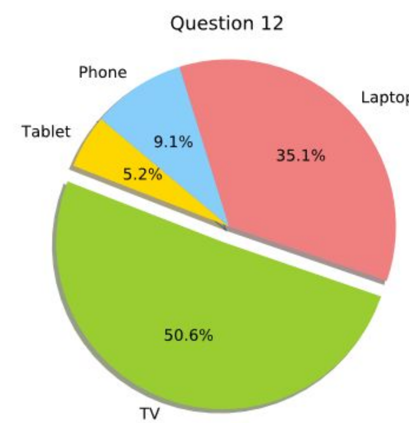
Question 11: 60% relaxing
30% surfing

Question 12: ½ TV
35% Laptop





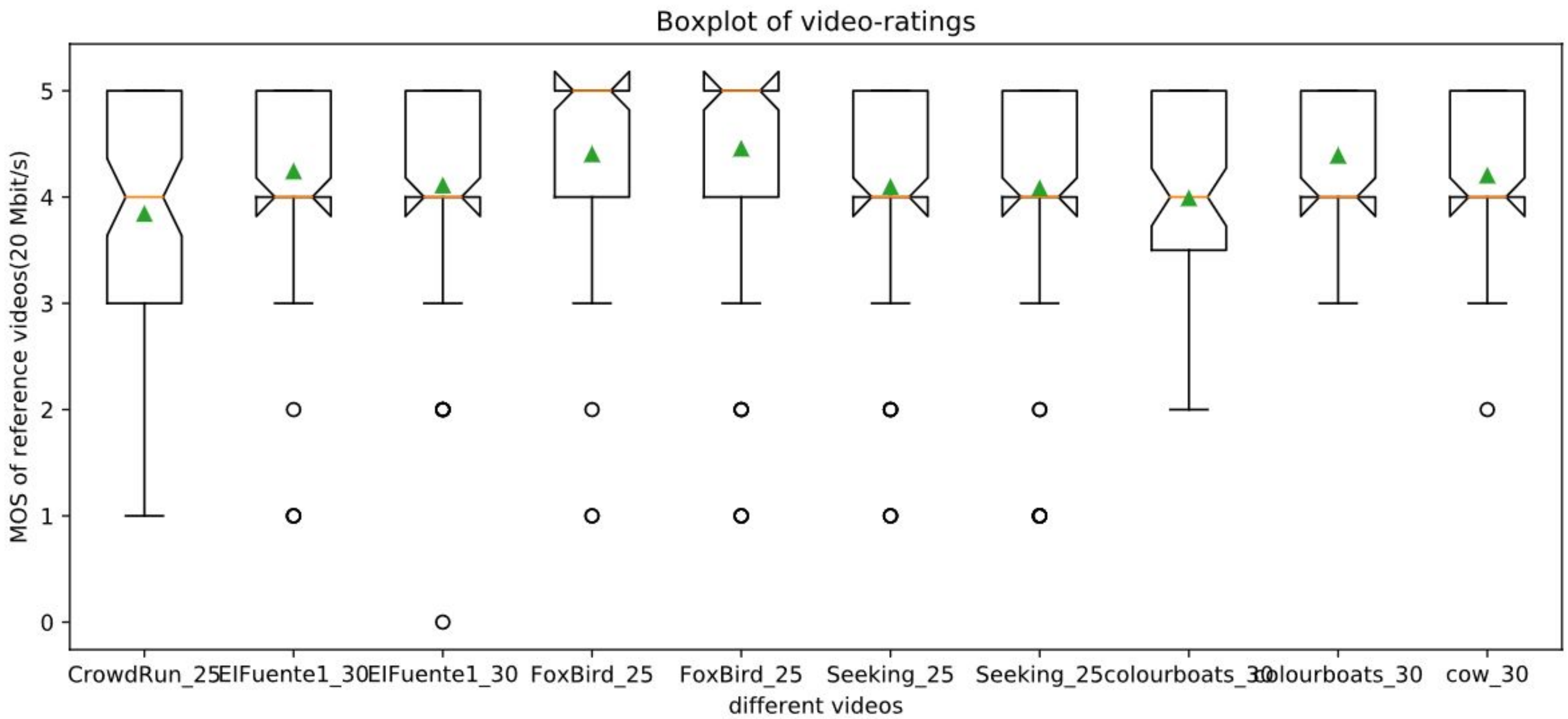




Part A: Crowdsourcing Results

- Quality Ratings

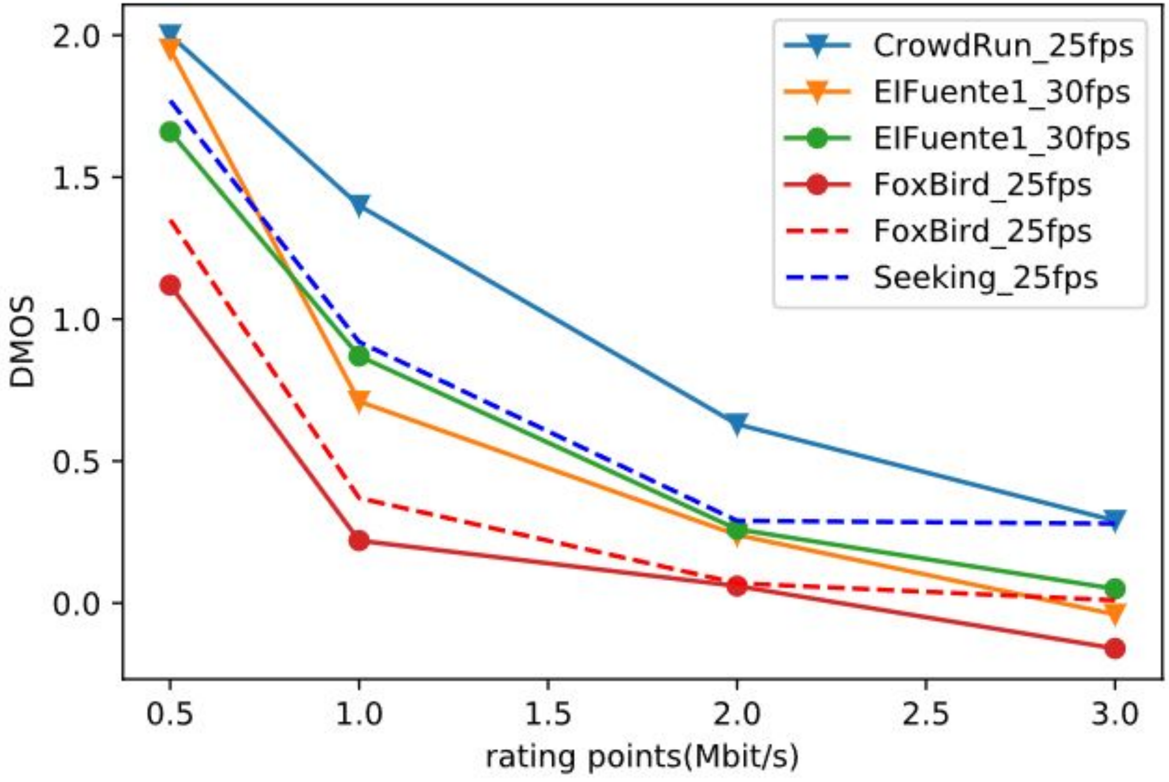
X: MOS of Reference(20 Mbit/s)   Y: different videos in different experiments



Part A: Crowdsourcing Results

- Quality Ratings

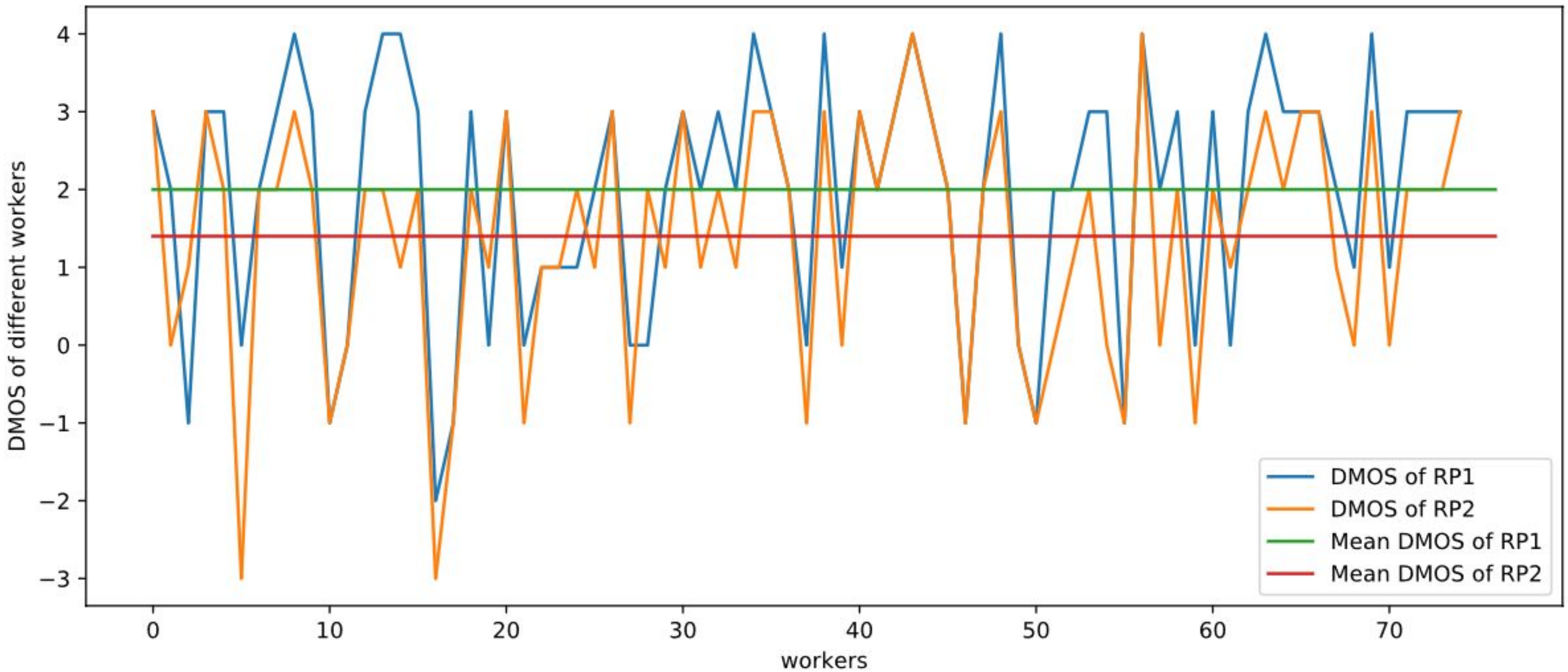
RP0 Reference: 20 Mbit/s *RP1: 500 kbit/s* *RP2: 1 Mbit/s* *RP3: 2 Mbit/s* *RP4: 3 Mbit/s*



Part A: Crowdsourcing Results

- Interactions between Streaming Habits/Quality Ratings

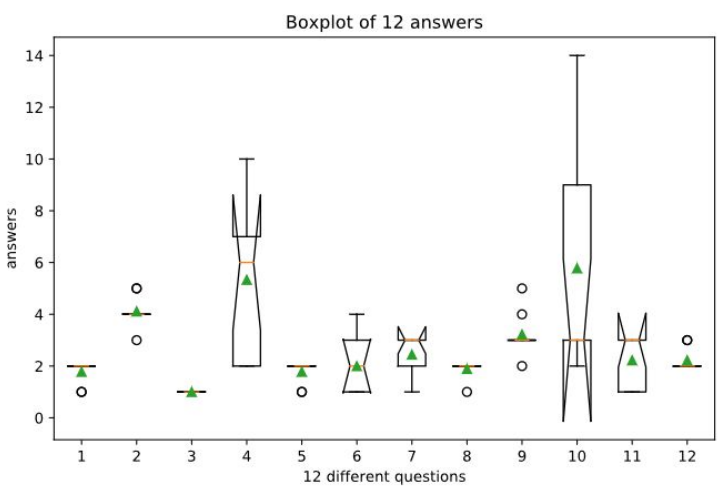
Worker 5, 10, 16, 21, 27, 37, 46, 50, 55	*relative small DMOS*
Worker 8, 34, 38, 48, 56, 66	*relative large DMOS*

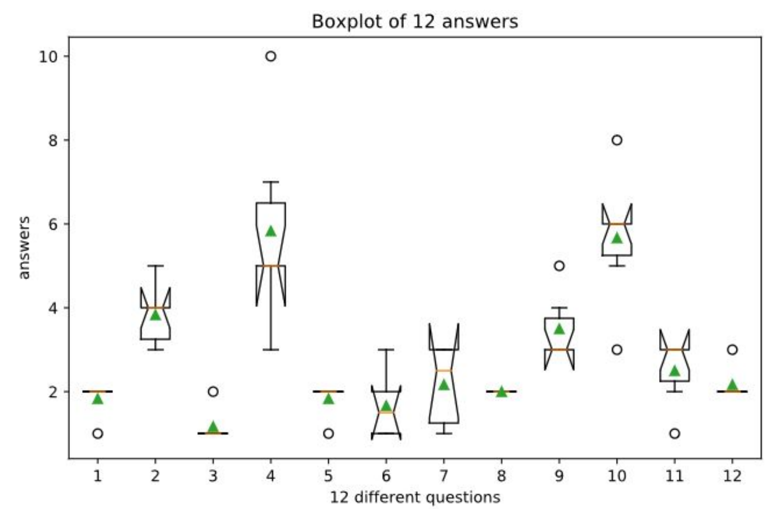


Part A: Crowdsourcing Results

- Interactions between Streaming Habits/Quality Ratings

Worker 5, 10, 16, 21, 27, 37, 46, 50, 55 *relative small DMOS* *right Figure*
Worker 8, 34, 38, 48, 56, 66 *relative large DMOS* *left Figure*





Que1: Mostly male
Que4: Business /Industrial and manufacturing/
Law Enforcement and Armed Forces

Que6: more often
Que7: also during the week
Que9: spend more time(more hours)
Que10: talk shows/comedy

Part B: Video Quality Metric

- Features collection using VMAF metric
- Model selection
- Feature preprocessing
- Performance
- Discussion and Conclusion

Features collection using VMAF

- Selected scores:
  - VIF scores: scale 0 ~ 3
  - Adm scores: DLM and AIM
  - Motion scores
  - SSIM and MS-SSIM
  - PSNR
  - Other scores: Bagging score

Features collection using VMAF

- Selected scores:
  - VIF scores: scale 0 ~ 3
    - a image quality metric: measurement of information fidelity loss. Combine the loss of fidelity in each one of 4 scales.
  - Adm2 scores: *DLM* and *AIM*
    - *Detail loss Metric* and Additive impairment measure; image quality metric
  - Motion
    - measure of temporal difference between adjacent frames: average absolute pixel difference for luminance component

Features collection using VMAF

- Selected scores:
    - SSIM and MS-SSIM
        - Luminance Comparison
        - Contrast Comparison
        - Structure Comparison

          $$SSIM = I(S,\hat{S})c(S,\hat{S})s(S,\hat{S})$$
        - *MS*: multiscale subsample

Features collection using VMAF

- Selected scores:
  - PSNR: *peak signal to noise ratio*

    $$20log_{10}(MAX_1) - 10log_{10}(MSE)$$
  - Other scores: Bagging scores
    - Bootstrapping aggregation for feature extraction: std, mean, vmaf, etc.

Model Selection

- Principal Component Regression
  - Extract latent component from features
- Partial Least Squares Regression
  - Extract latent component both from the features and also from the groundtruth

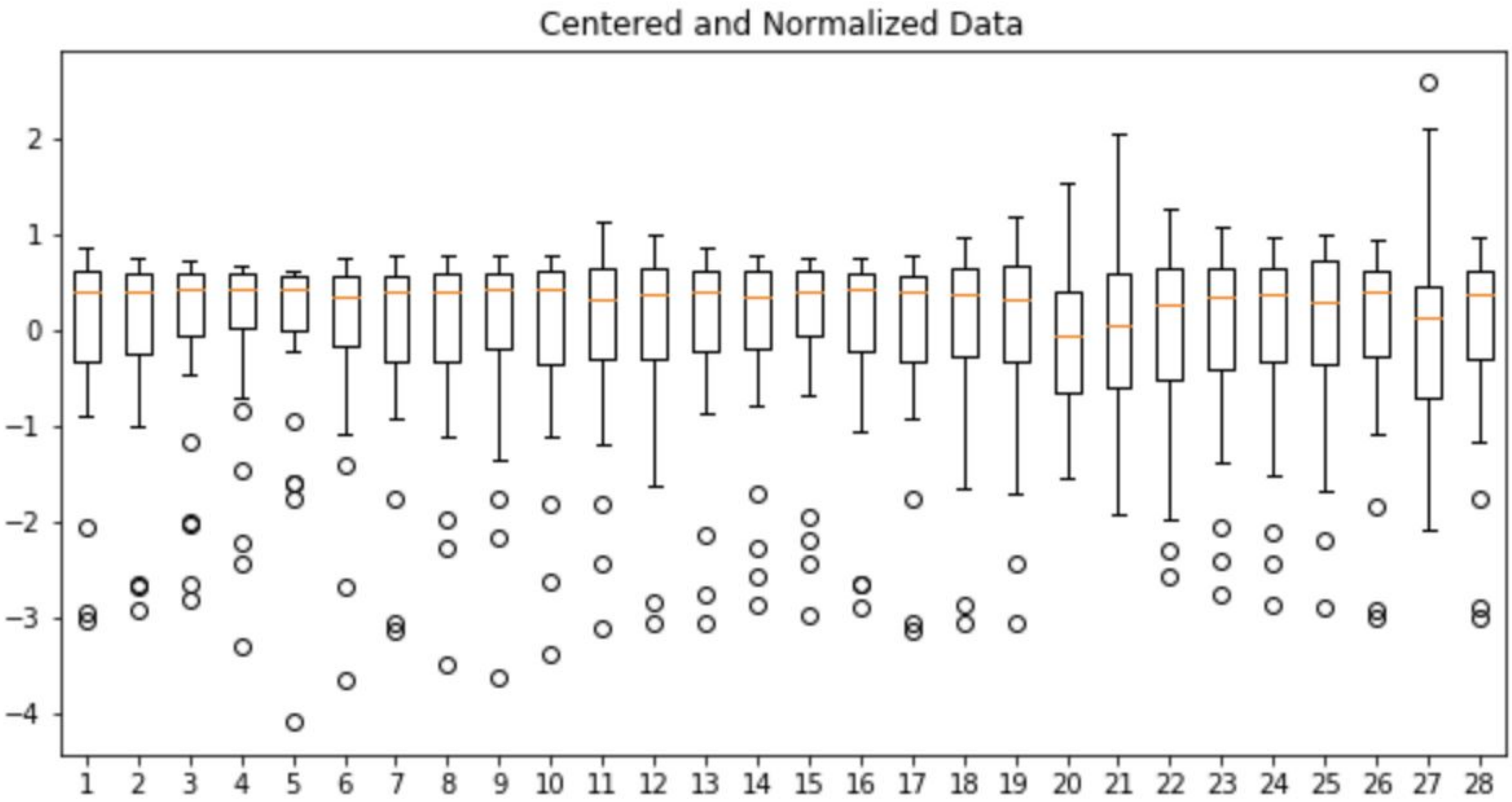
Feature Preprocessing

- Standardizing
  - This method attribute data assumes a Gaussian distribution of input features and "standardizes" to a mean of 0 and a standard deviation of 1

    + : Ensures insensitivity of the model to the original scale of variance for the data

$$z = \frac{x - \mu}{\sigma}$$

Feature Preprocessing



Feature Preprocessing

- Visualising (1st - 7th feature for example)

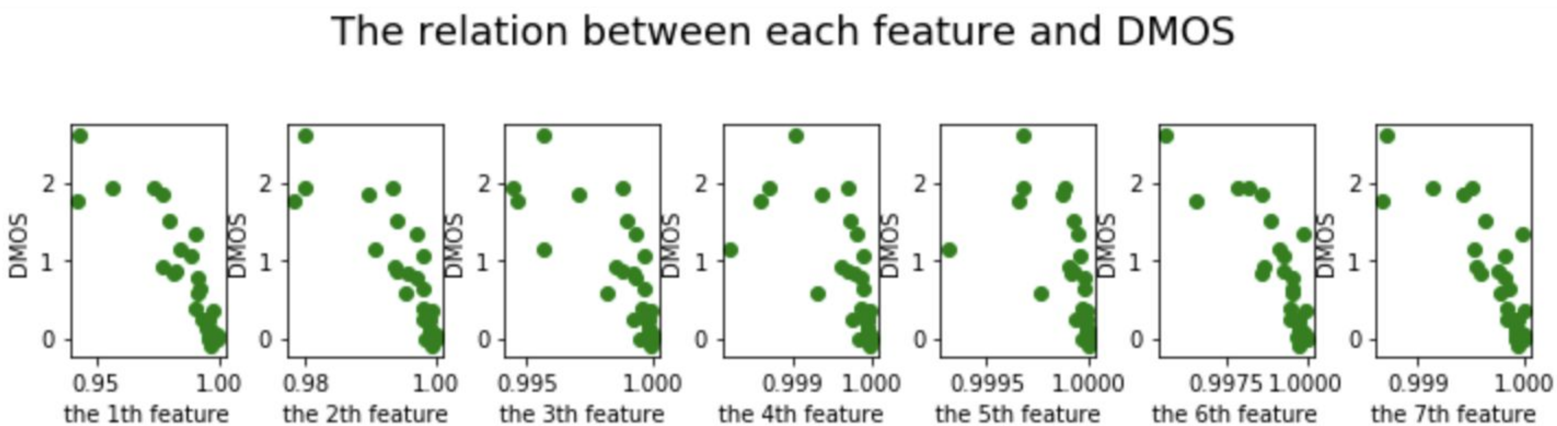
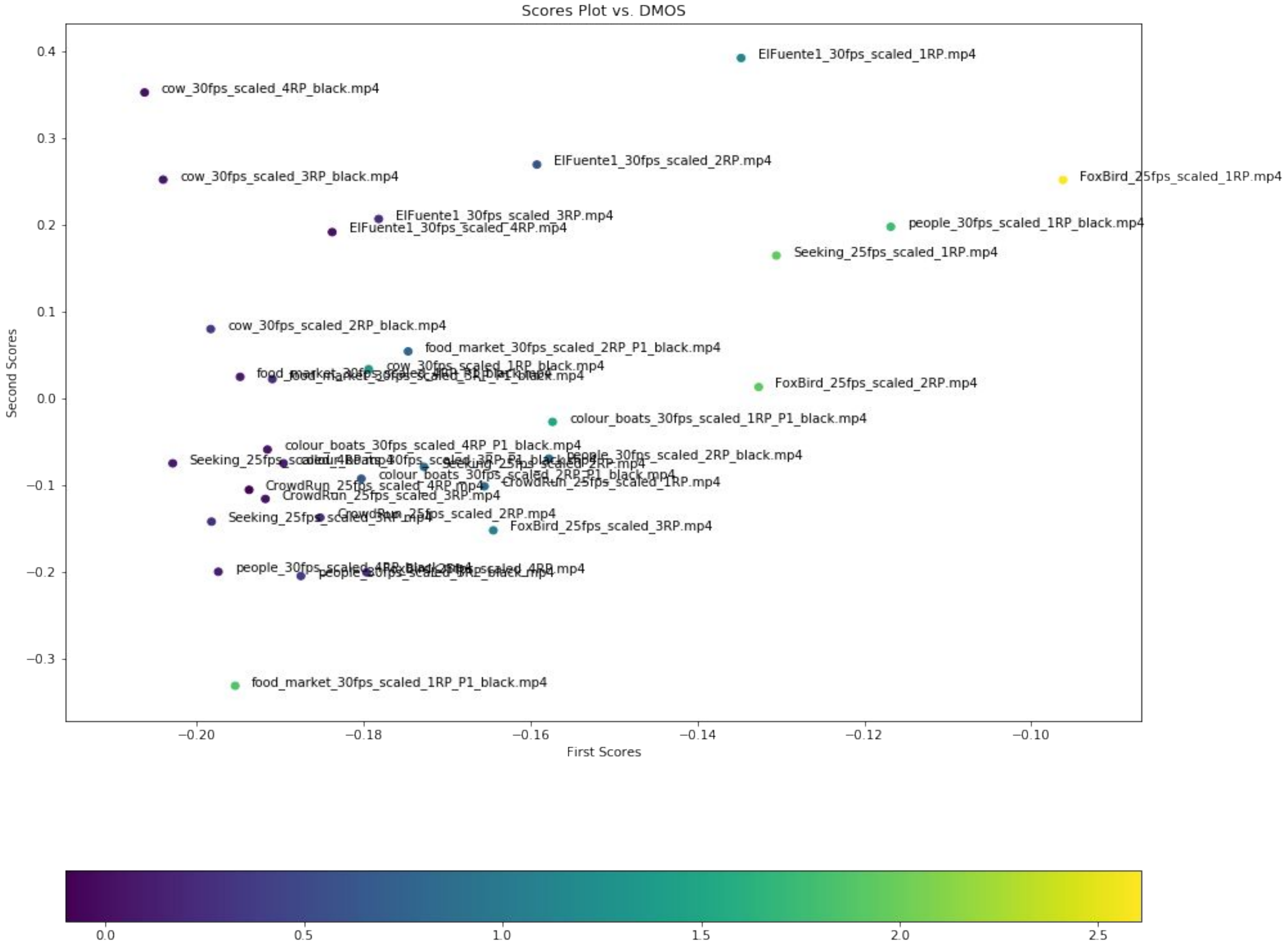


Table 1: Top 3 features with high influence

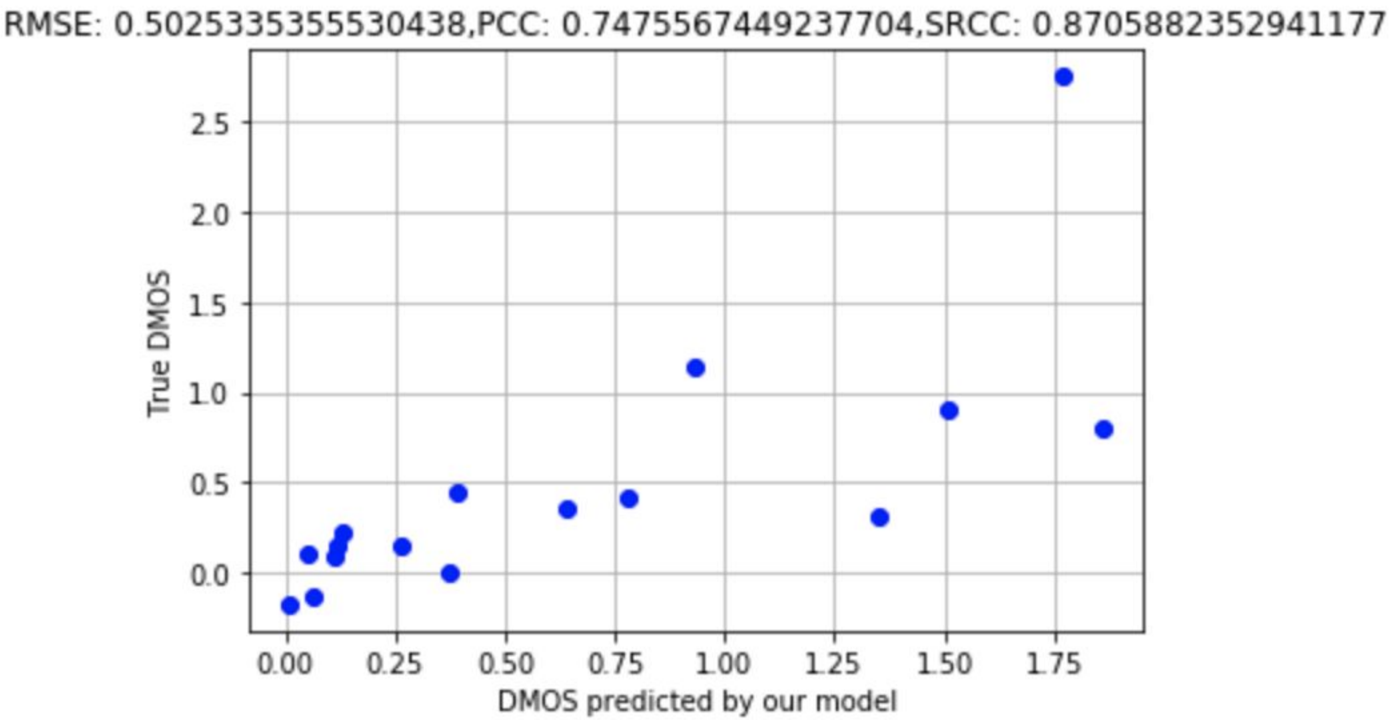
Feature name	Covariance value (with DMOS)
MS_SSIM_feature_ms_ssim_l_scale4_score	0.03218822469883268
MS_SSIM_feature_ms_ssim_s_scale0_score	0.026682828773759194
SSIM_feature_ssim_l_score	0.02605933998903838

Score value analysis



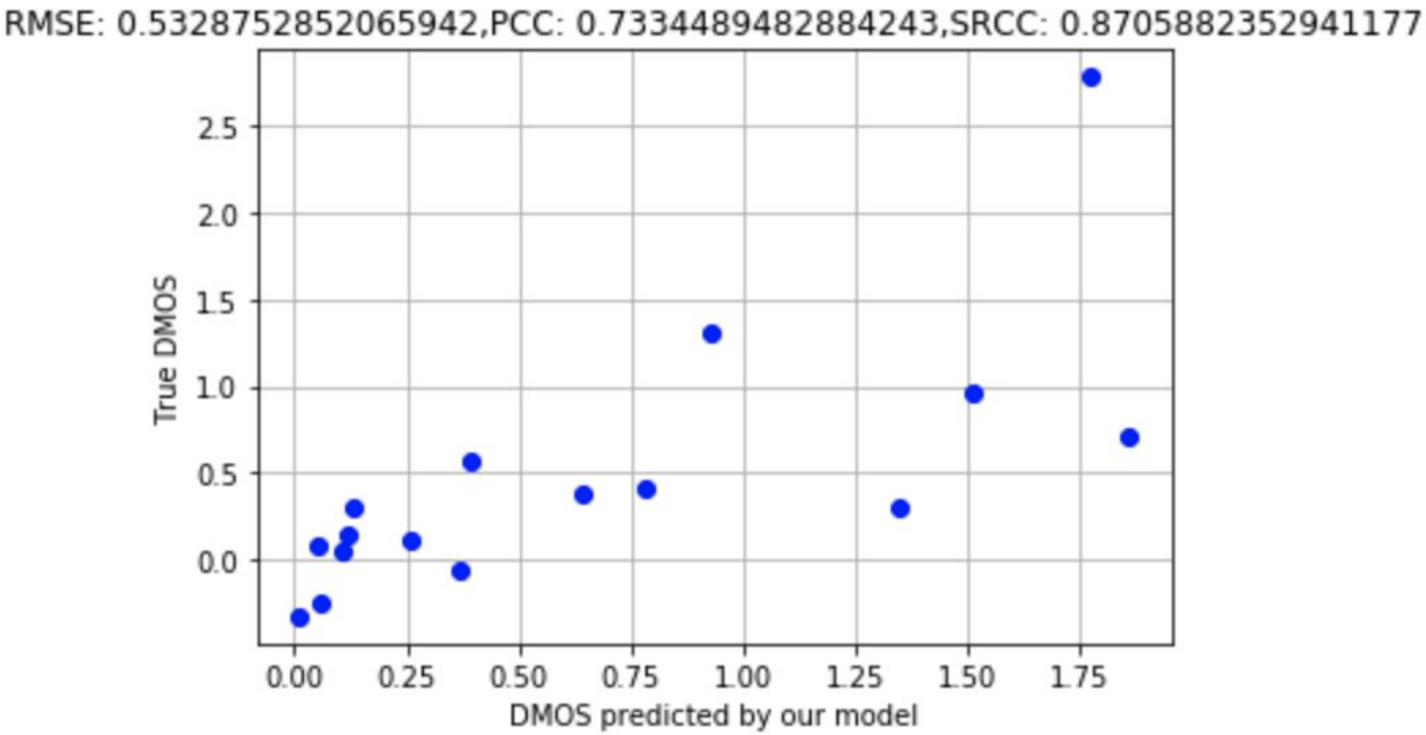
Performance

PCR model, when component is 1



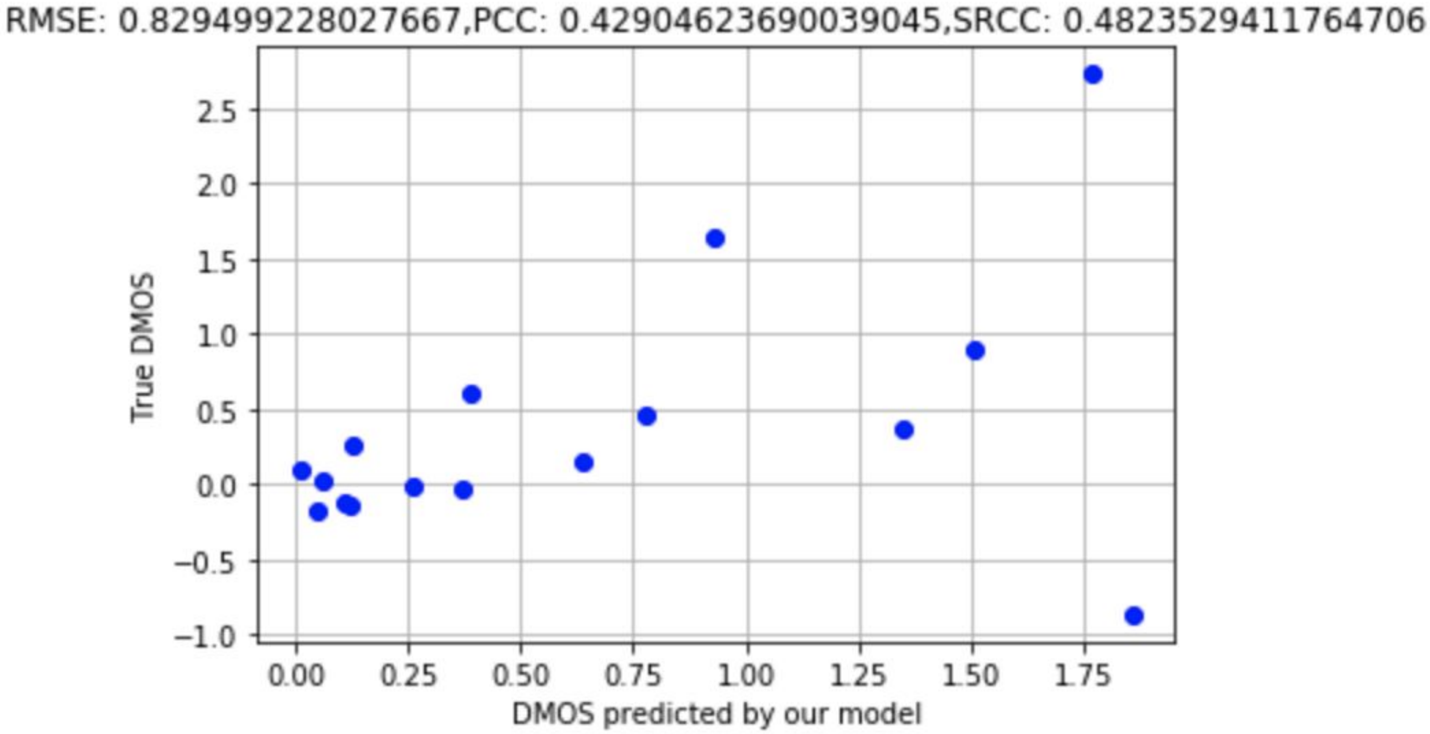
Performance

PCR model, when component is 2



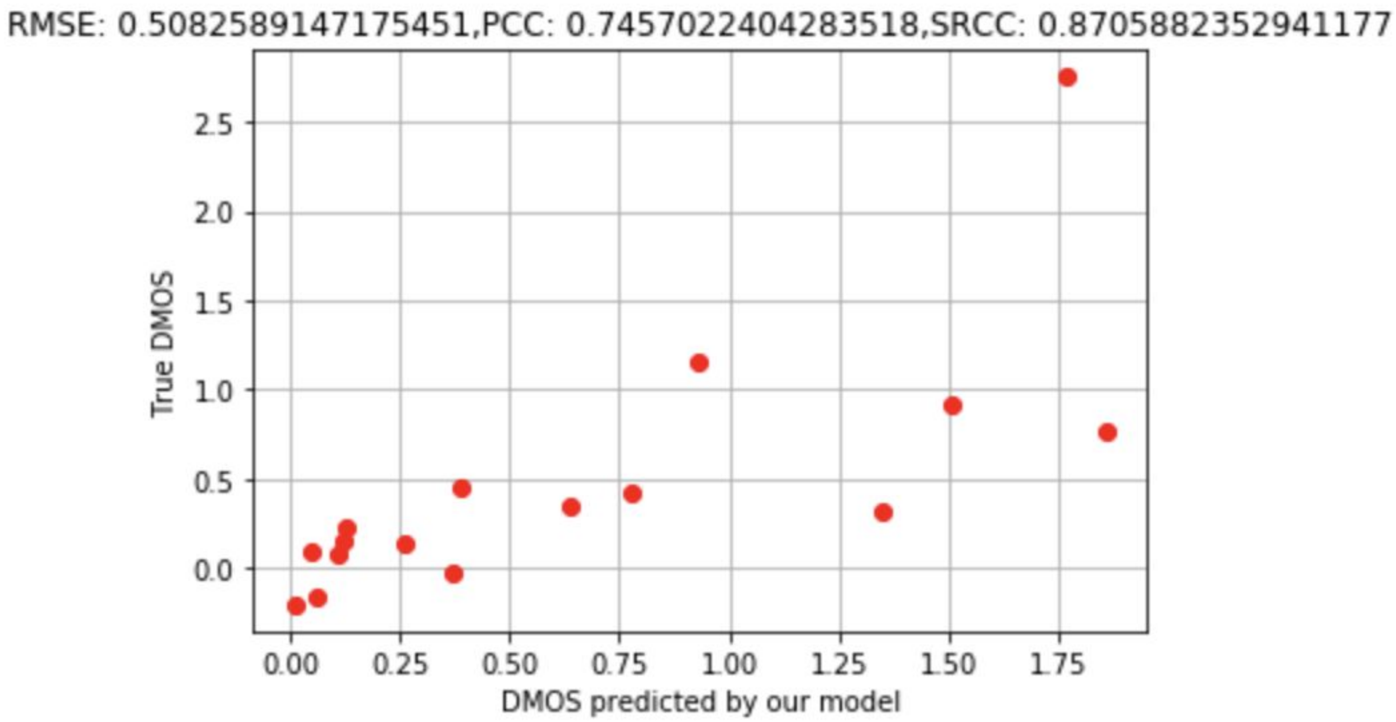
Performance

PCR model, when component is 10



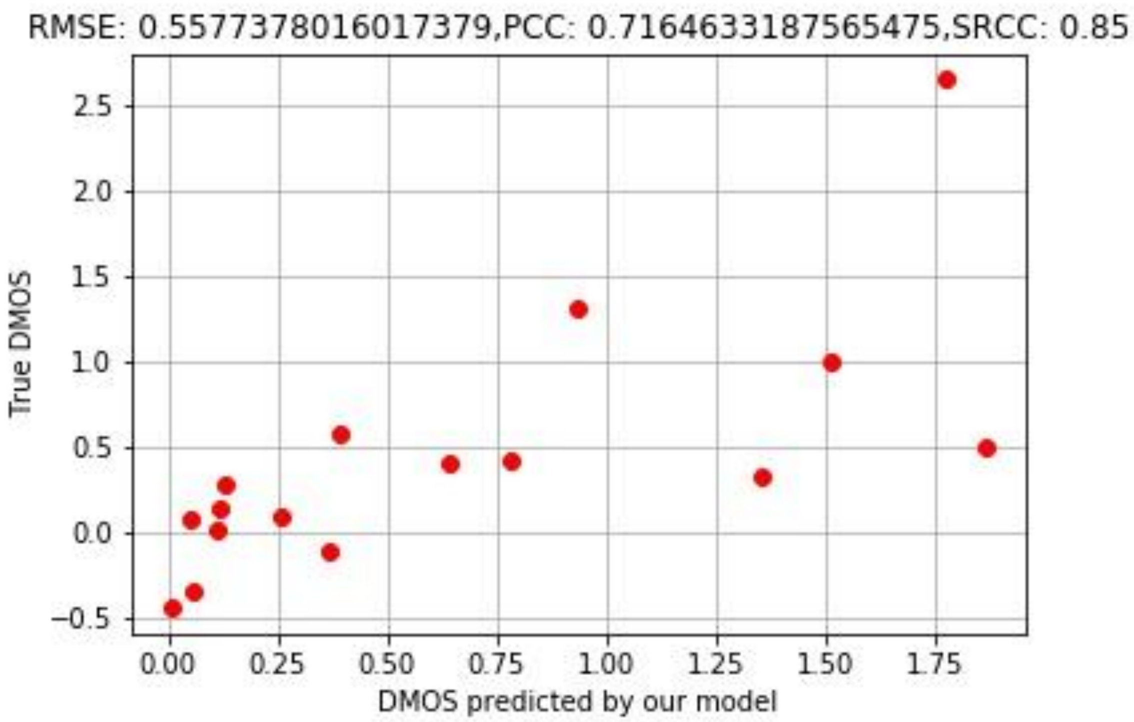
Performance

PLS model, when component is 1



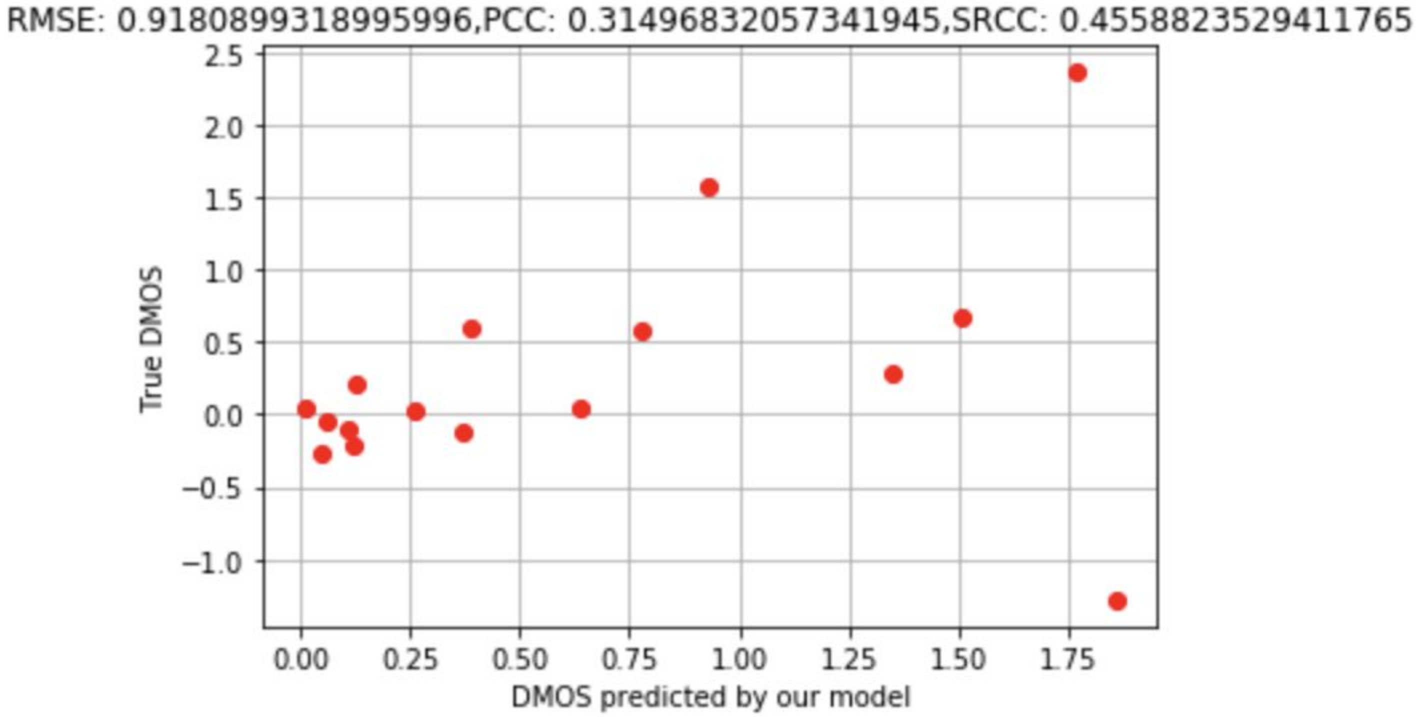
Performance

PLS model, when component is 2



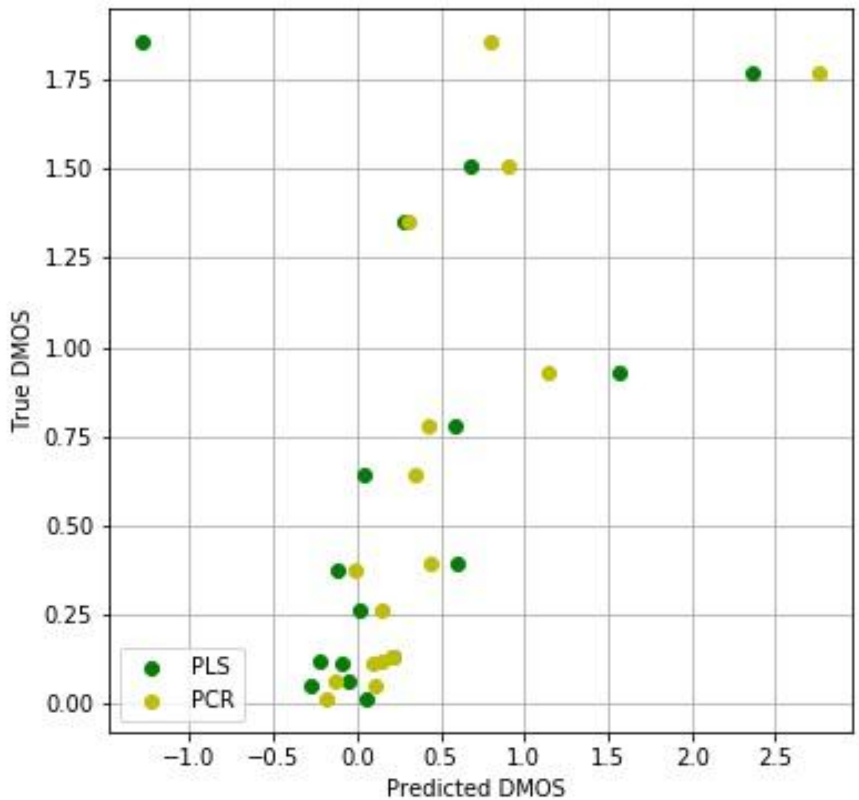
Performance

PLS model, when component is 10



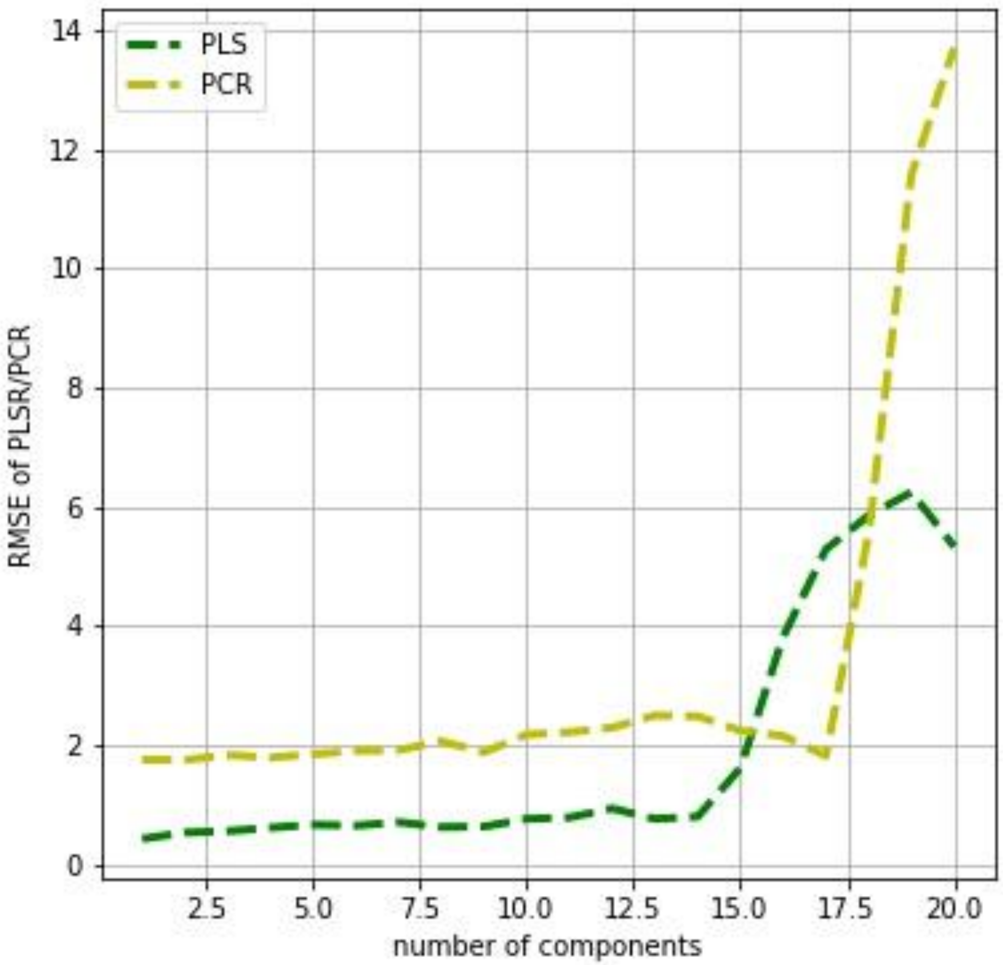
Performance

model comparison, when best component number is 1



Performance

Cross-validation: PLSR and PCR model test with different number of components



Performance

Model comparison between PLSR and PCR

	RMSE	PCC	SRCC
PLSR(pc=1)	0.50825	0.745702	0.870588
PLSR(pc=2)	0.55773	0.71646	0.850583
PLSR(pc=10)	0.91808	0.31496	0.45588
PCR(pc=1)	0.50253	0.74755	0.870588
PCR(pc=2)	0.53287	0.73344	0.870589
PCR(pc=10)	0.82949	0.42904	0.48235

Discussion and Conclusion

Model differences			
	Size of the database	**Quality of the database**	**Feature space**
PCR	• Performances well with small dataset • Much quicker	• Performances well when latent variables mainly correlated to feature dataset(X)	• More features reduce the influence of outlier • Smaller feature space may need more time to find the optimal component number
PLSR	• Performances good at bigger dataset • Need more computations	• Performances well when latent variables correlated not only to feature dataset(X) but also to predicted data(Y)	• Feature space scale has less influence when compared to PCR.

Thank you for attention!