Principal Components Regression

Josh Fontes

PCR Outline

1. PCA on design matrix
2. Use subset of principal components as regressors
   - Regression assumptions (residuals normally distributed and constant variance, observations are independent of each other, etc…)

RECALL:

- PCA transforms original predictor variables to a set of linearly uncorrelated variables (principal components)
- Use a subset to preserve as much variance as possible while reducing dimensions of dataset

$\Rightarrow$ Multicollinearity reduces

Mathematical process

Design Matrix

$X_{k \times p}$

SVD on centered design matrix

$\bar{X} = X - \bar{x}$

$\bar{X} = U \Sigma V^T$

Principal Components

$Z_{k \times r} = \bar{X}_{k \times p} V_{p \times r}$

Regression

$y = Z\beta_z + \varepsilon$

$\beta_z = (Z'Z)^{-1} Z'y$

Derive β_x

$\hat{y} = Z\beta_Z = (\bar{X}V)\beta_Z = \bar{X}(V\beta_Z) = \bar{X}\beta_X$

$\Leftrightarrow \quad \beta_X = V\beta_Z$

(optional)

UCI Superconductivity dataset

- Elements dataset:
  - Degree of each element in conductors
  - Response:
    - Critical Temperature (Kelvin)
- Features dataset (design matrix):
  - Predictors:
    - Features of conductor materials

Response: $y_{21,263 \times 1}$

Design Matrix: $X_{21,263 \times 81}$

Some Features

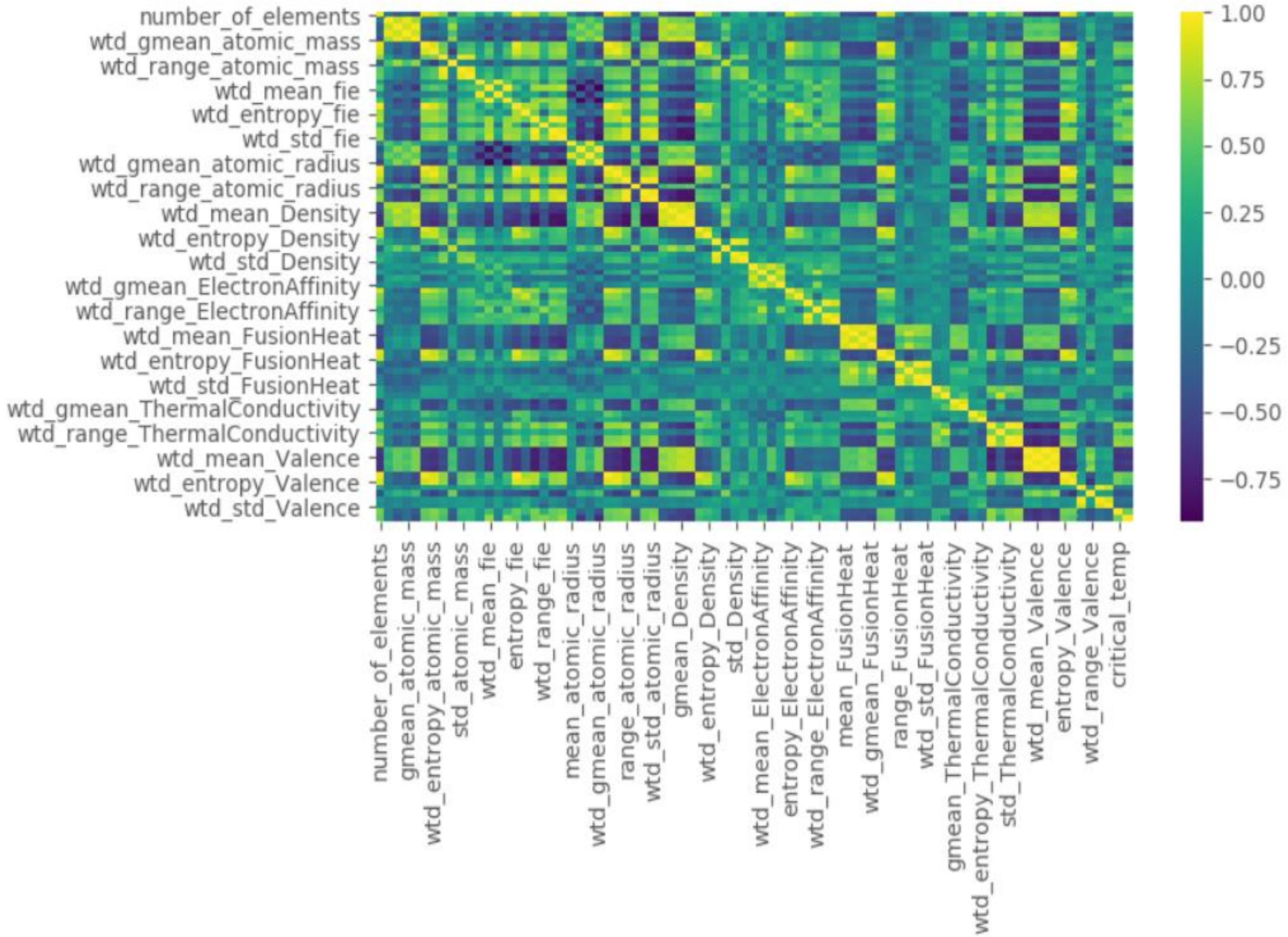
Atomic Mass
First Ionization Energy
Atomic Radius
Density
Electron Affinity
Fusion Heat
Thermal Conductivity
Valence

PROCESS

- Very large, high multicollinearity dataset

	VIF	features
0	79.644423	number_of_elements
1	414.277383	mean_atomic_mass
2	818.370293	wtd_mean_atomic_mass
3	444.203673	gmean_atomic_mass
4	879.861538	wtd_gmean_atomic_mass
...	...	...
76	307.311308	wtd_entropy_Valence
77	56.759455	range_Valence
78	26.150907	wtd_range_Valence
79	96.823865	std_Valence
80	51.827597	wtd_std_Valence

81 rows × 2 columns



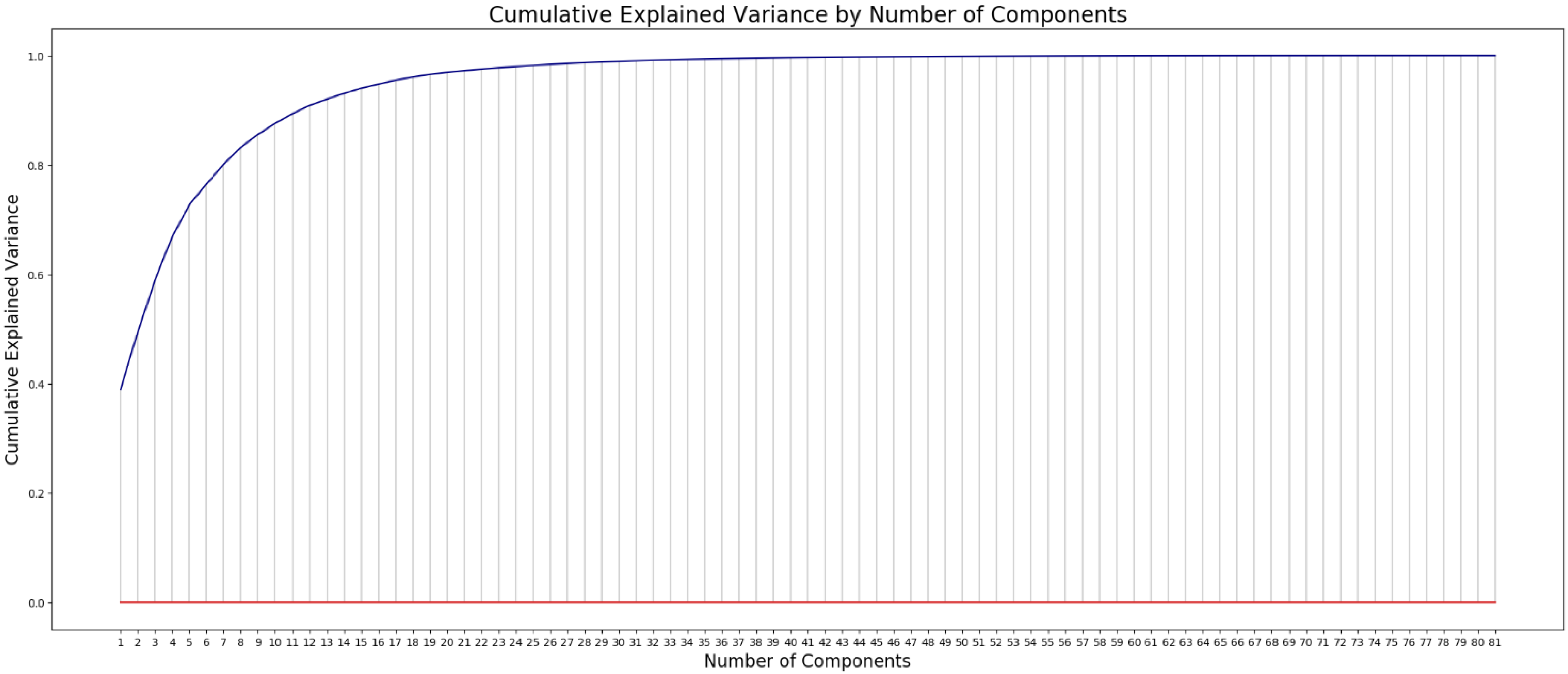
1. Standardize design matrix
2. PCA on design matrix
3. Subset principal components (30)

$Z_{21,263 \times 30}$

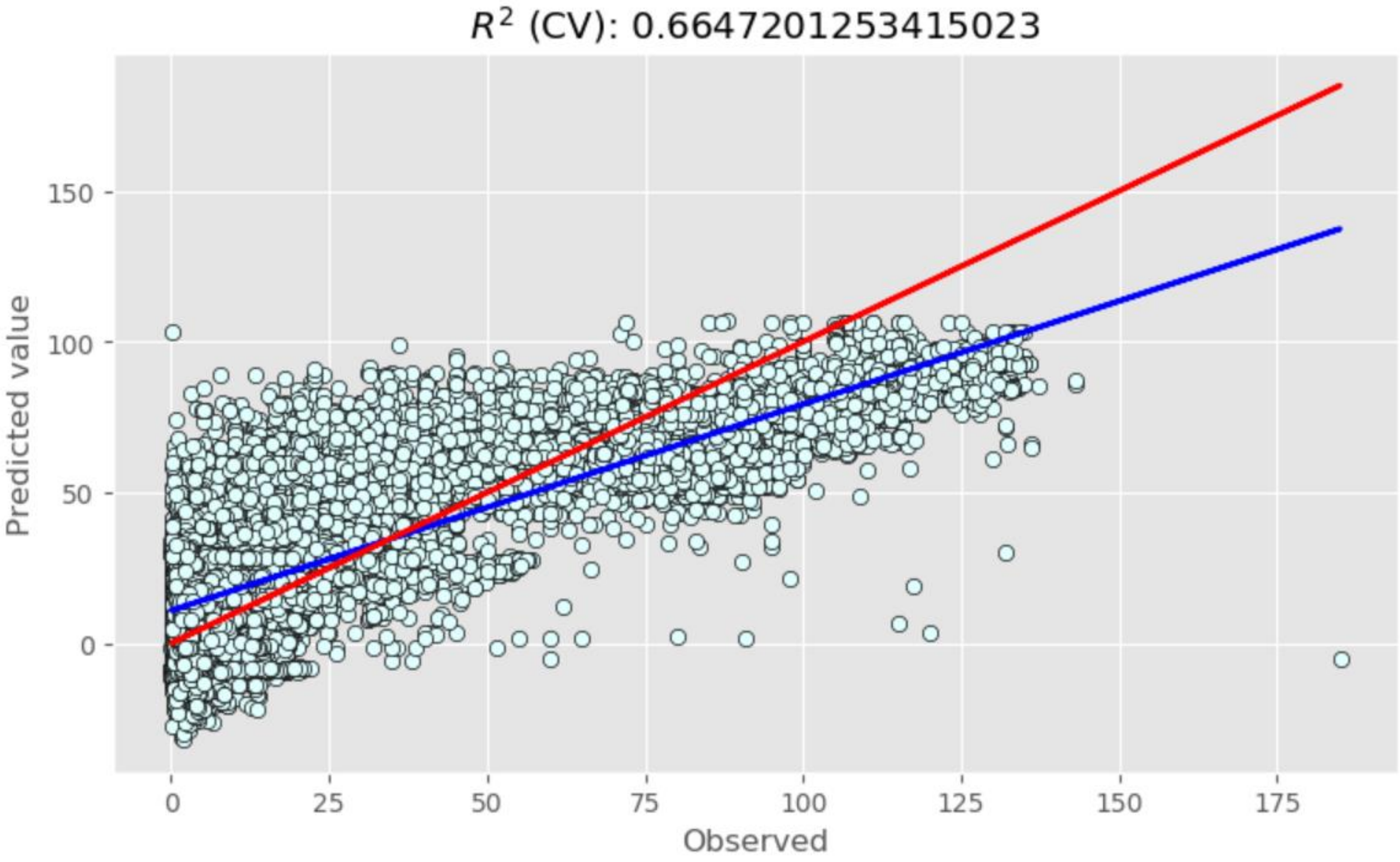
Proportion of total variance explained by the *k*th principal component:

Eigenvalues = λ

$$\frac{\lambda_k}{\sum \lambda_i}$$

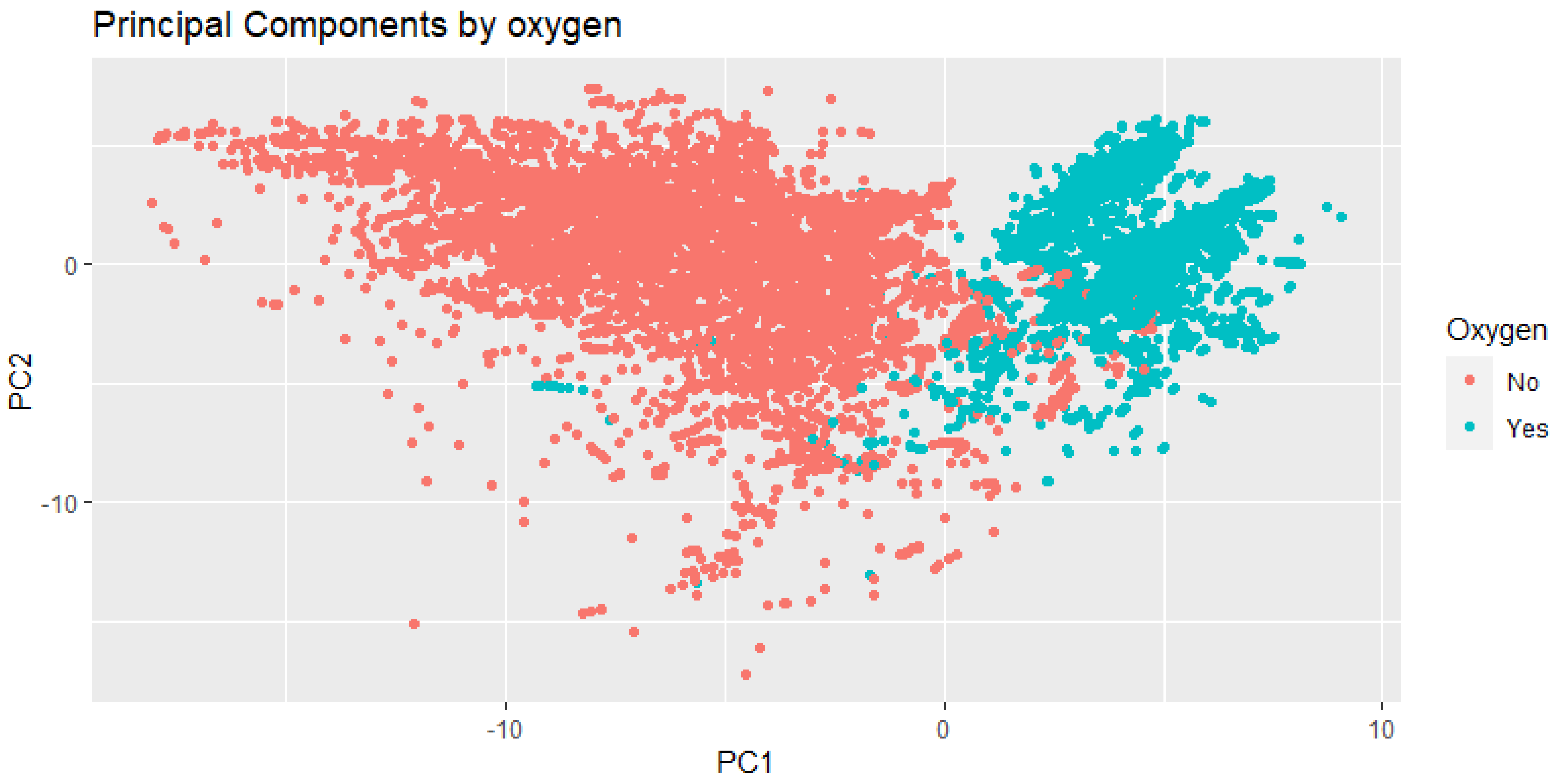


What about subsetting certain elements?



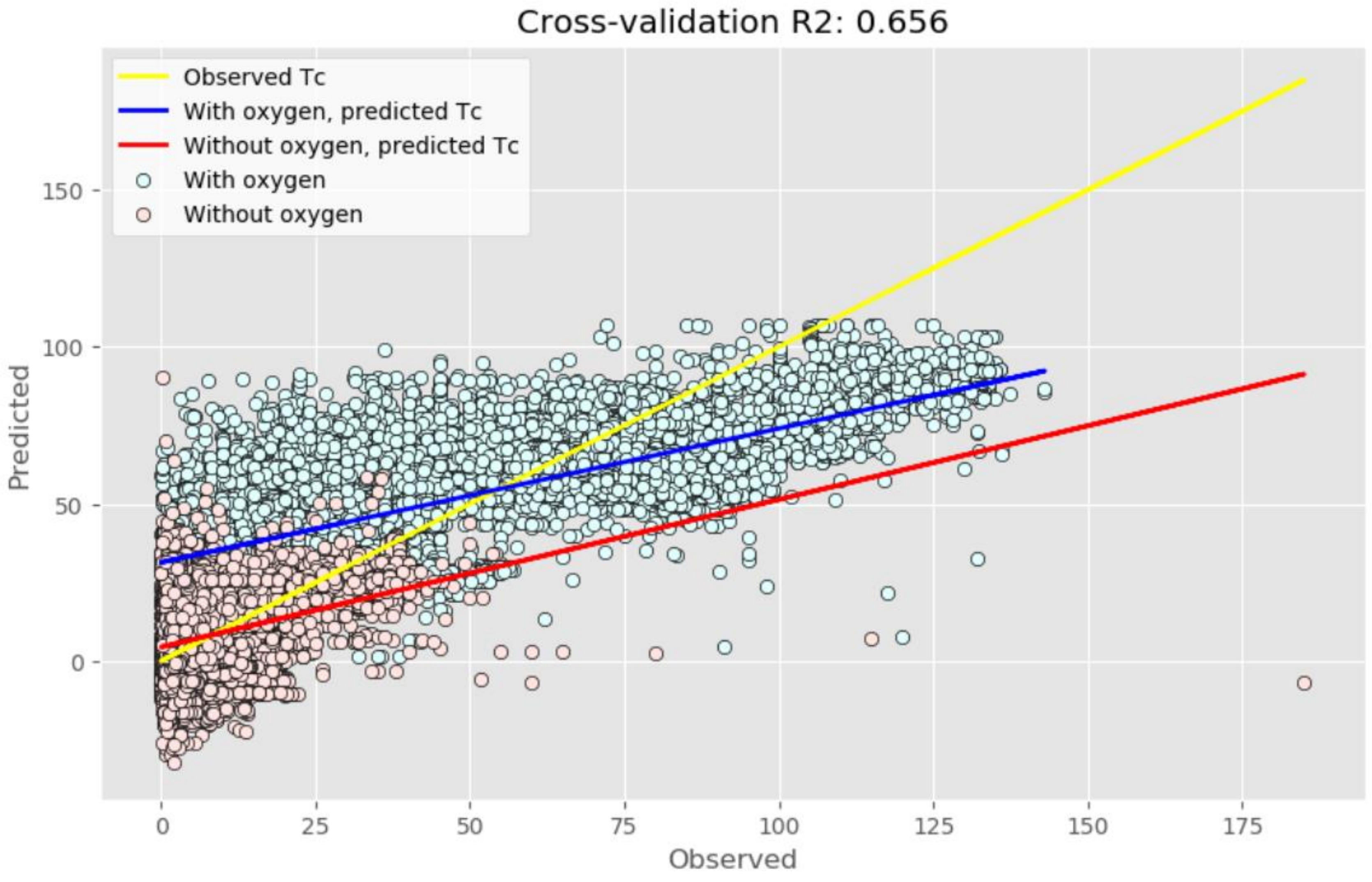

```
R2: 0.684
R2 CV: 0.665
MSE: 370.615
MSE CV: 393.386
Intercept: 34.42121913535249
Coefficients: [-4.0043751  -1.98810954  2.40200866 -1.25055204  2.56626636  3.42724176
 -0.76400443 -0.57528784 -1.9102399   0.09084165 -3.07758419 -1.93249279
  1.08631023  0.57218435  4.00960152 -0.30450643 -0.22193757 -4.80502217
  2.75139291  8.10565135  2.9911842   2.35121154 13.20875786 -1.22877303
 11.53167589  4.61912283 -0.10150983  1.47960276 -0.32030094 -3.50477045]
```

Oxygen


Principal Components by oxygen
PC2
PC1
0
-10
-10
0
10
Oxygen
No
Yes

With oxygen content used as categorical variable

```
R2: 0.686
R2 CV: 0.656
MSE: 368.548
MSE CV: 403.463
```



Discussion

- Cross-validation measures were not great
  - Residual analysis needed, transformations needed, influential points, etc…

- PCR advantages
  - Reduces multicollinearity
  - Excellent for data visualization
- PCR disadvantages
  - Mathematically complex to build models
  - Difficult to interpret
  - Categorical variables

Questions

- Multicollinearity reduced by subsetting?

OR

- by the fact that principal components are linearly uncorrelated?

Found multiple conflicting sources on this

Thank you for your attention!

URL References:

Relevant Paper:

https://www.sciencedirect.com/science/article/pii/S0927025618304877?via%3Dihub

Other sources used:

https://plotly.com/r/3d-scatter-plots/

https://www.rpubs.com/bpiccolo/pcaplots

https://nirpyresearch.com/pcr-vs-ridge-regression-nir-data-python/

https://www.nature.com/articles/31656

https://www.globalspec.com/learnmore/materials_chemicals_adhesives/electrical_optical_specialty_materials/superconductors_superconducting_materials

https://rpubs.com/esobolewska/pcr-step-by-step

https://en.wikipedia.org/wiki/Principal_component_regression

https://digitalcommons.wayne.edu/cgi/viewcontent.cgi?article=1166&context=jmasm

http://siret.ms.mff.cuni.cz/sites/default/files/doc/david.hoksza/lectures/vis/04-pca.pdf