Hybrid Unsupervised-Supervised Machine Learning Models for Materials Science

Rose K. Cersonsky

Laboratory of Computational Science and Modeling (COSMO)

École Polytechnique Fédérale de Lausanne (EPFL)

COSMO
COMPUTATIONAL SCIENCE AND MODELLING

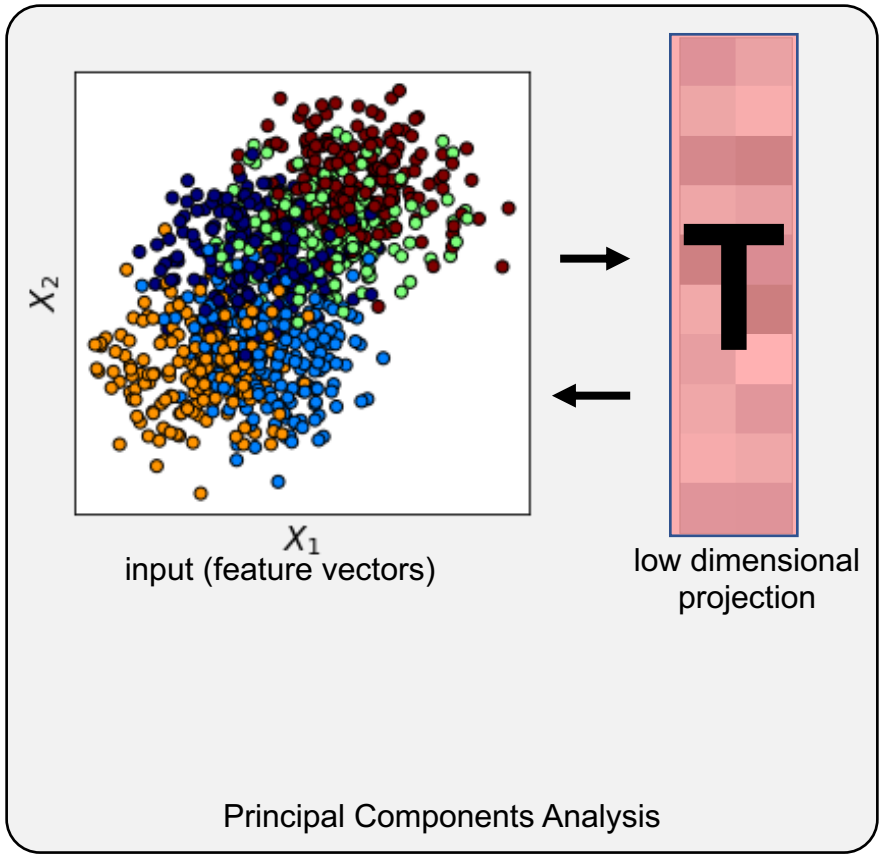
A couple words on notation…

$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \dots \end{bmatrix}$	A matrix containing as rows **the fingerprints of a set of structures**
$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \dots \end{bmatrix}$	A matrix containing as rows **the target properties for a set of structures**

$\mathbf{P}_{\mathrm{AB}}$	A matrix that projects from space $\mathbf{A}$ to space $\mathbf{B}$
$\mathbf{T} = \mathbf{X}\mathbf{P}_{\mathrm{XT}}$	A matrix containing as rows **the latent-space projection of a set of structures**

Principal Components Analysis (PCA)

PCA determines an information-rich set of features to represent a larger set of features.



$$\ell = \|\mathbf{X} - \mathbf{X}\,\mathbf{P}_{\mathrm{XT}}\,\mathbf{P}_{\mathrm{TX}}\|^2$$

This is solved by constructing the projectors from the eigendecomposition of either the Gram matrix K or the covariance C (analogous to the SVD of X)

$$\mathbf{K} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$$

gram matrix

$$\mathbf{C} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$$

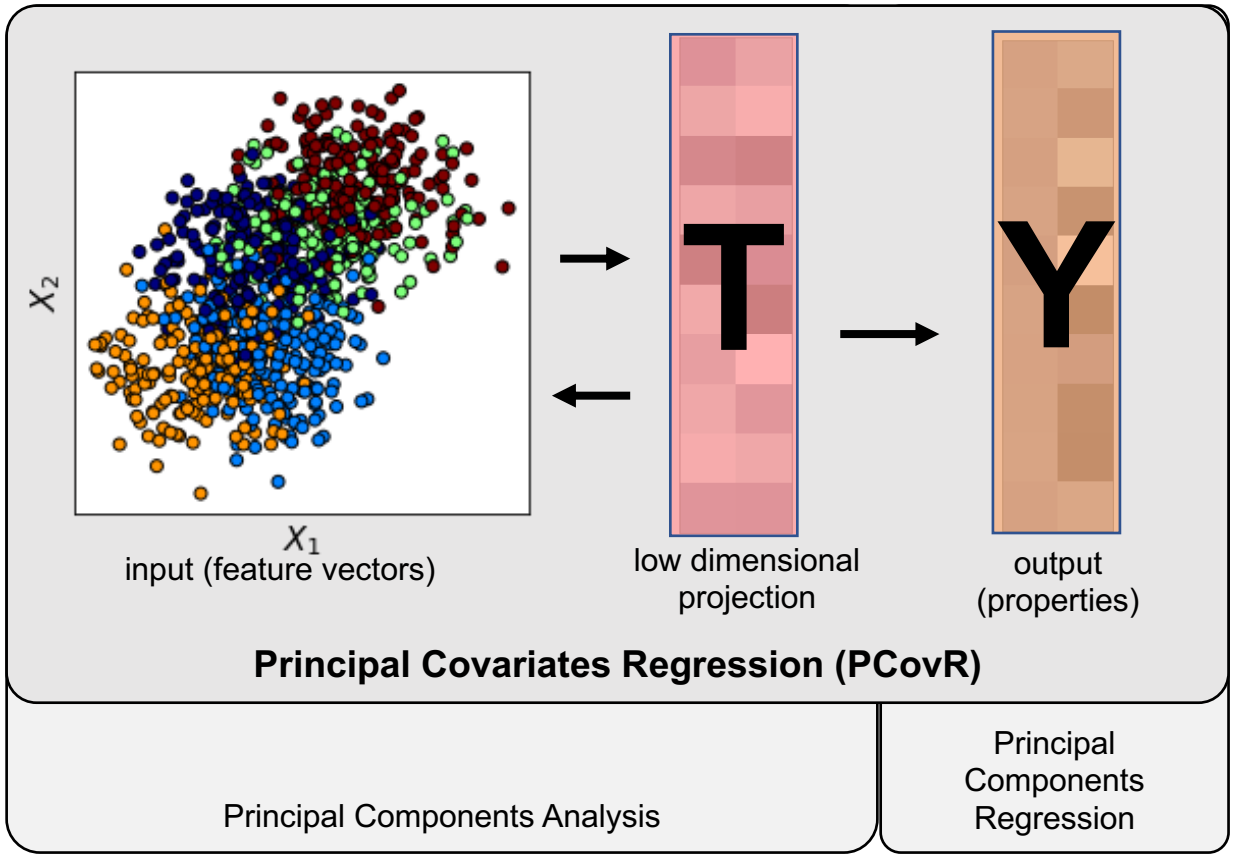
covariance matrix

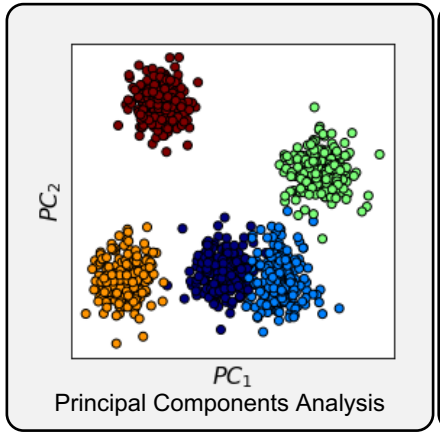
S. de Jong, H.A.L. Kiers, Chemom. intell. lab. syst. 14 (1992) 155-164.
scikit-cosmo.readthedocs.io

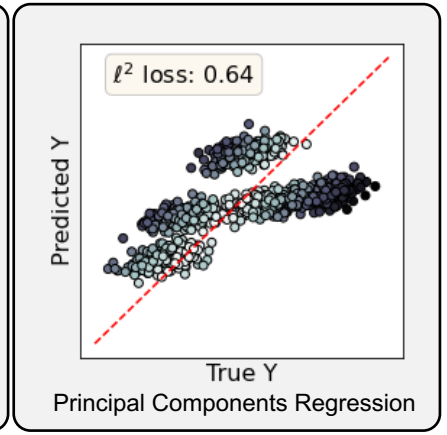
Inputs: sklearn.datasets.make_blobs
Regression Model: RidgeCV(cv=5)

Principal Covariates Regression (PCovR)

is a dimensionality reduction technique that determines a latent-space projection that incorporate saspects of supervised learning.





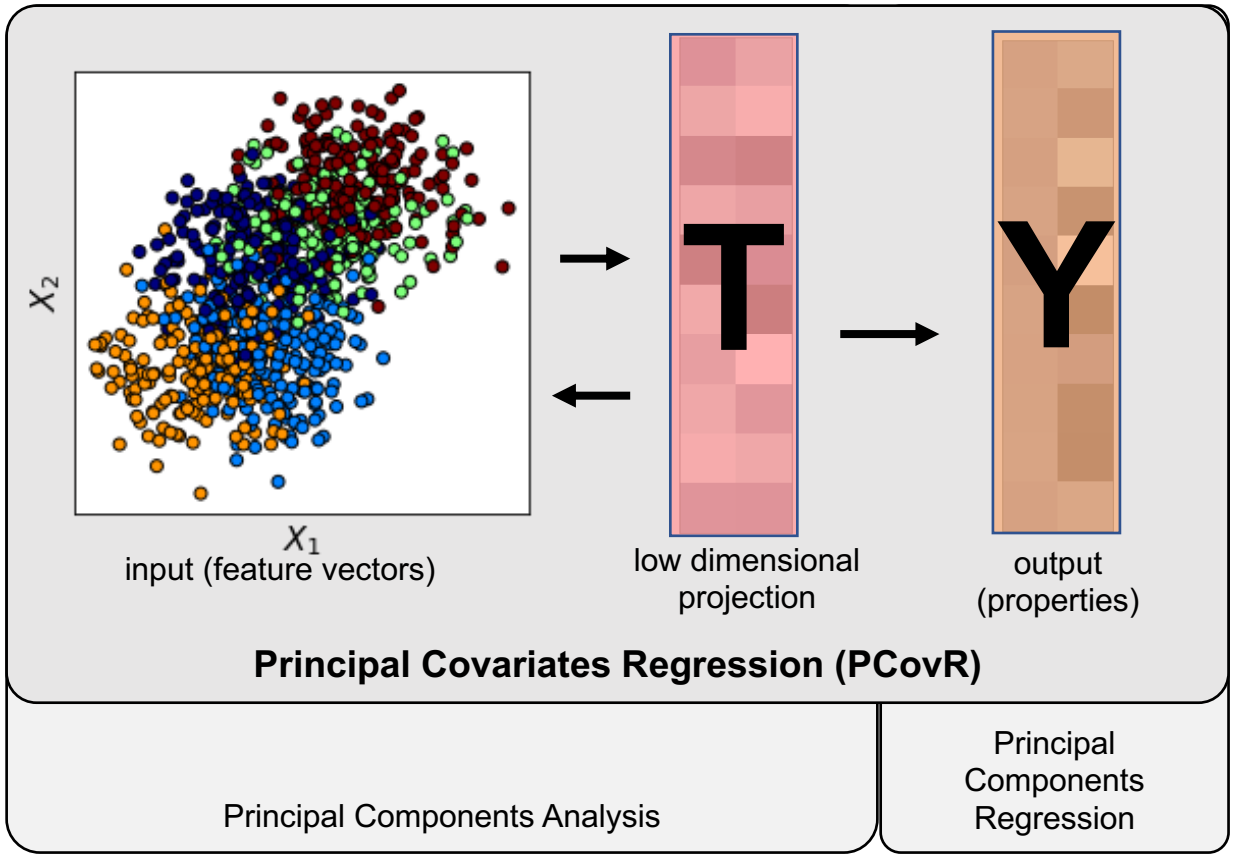


S. de Jong, H.A.L. Kiers, Chemom. intell. lab. syst. 14 (1992) 155-164.
scikit-cosmo.readthedocs.io

Inputs: sklearn.datasets.make_blobs
Regression Model: RidgeCV(cv=5)

Principal Covariates Regression (PCovR)

is a dimensionality reduction technique that determines a latent-space projection that incorporate saspects of supervised learning.



loss in reconstructing X

$$\ell = \alpha\|\mathbf{X} - \mathbf{X}\,\mathbf{P}_{\mathrm{XT}}\,\mathbf{P}_{\mathrm{TX}}\|^2 + (1-\alpha)\|\mathbf{Y} - \mathbf{X}\,\mathbf{P}_{\mathrm{XT}}\,\mathbf{P}_{\mathrm{TY}}\|^2$$

loss in reconstructing Y

This is solved by constructing the projectors from the eigendecomposition of either a **modified** Gram matrix or a **modified** covariance

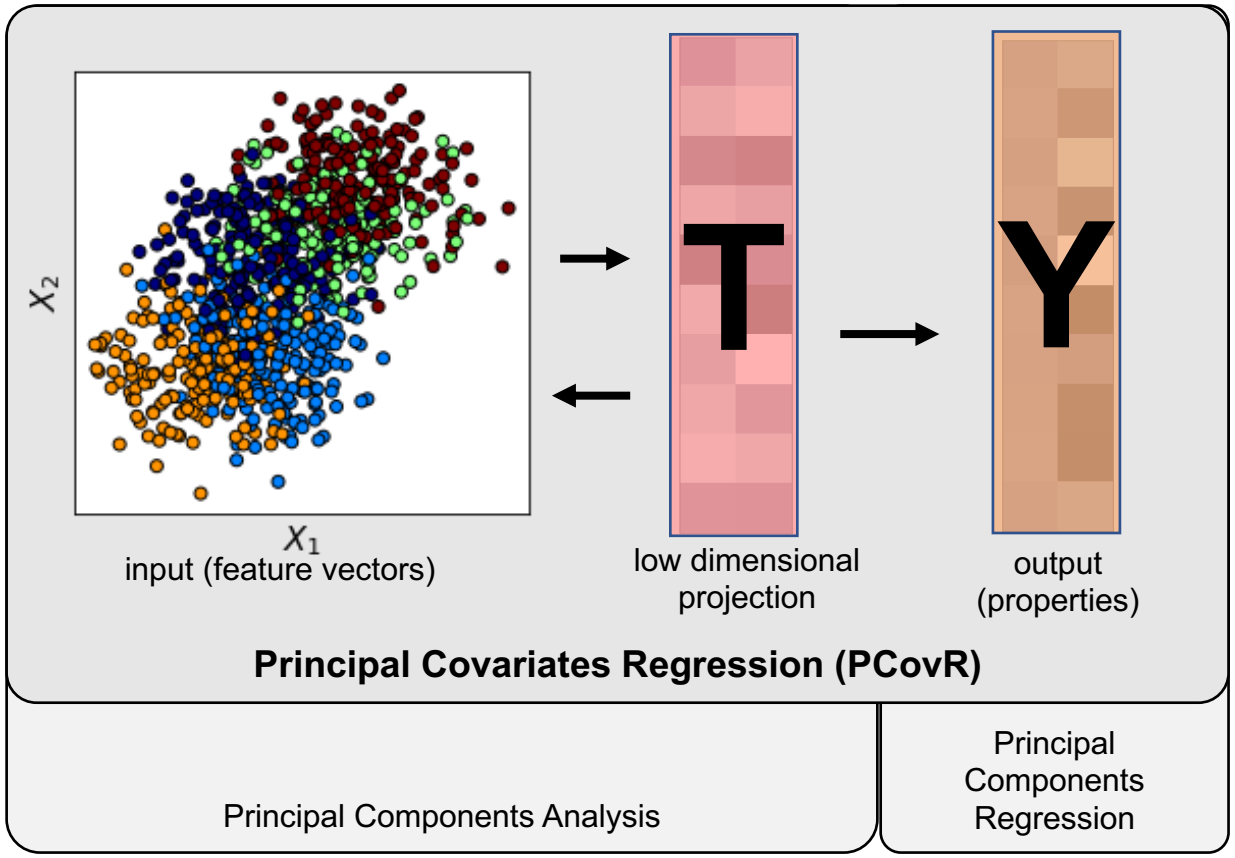
$$\mathbf{K} \rightarrow \tilde{\mathbf{K}} \qquad \mathbf{C} \rightarrow \tilde{\mathbf{C}}$$

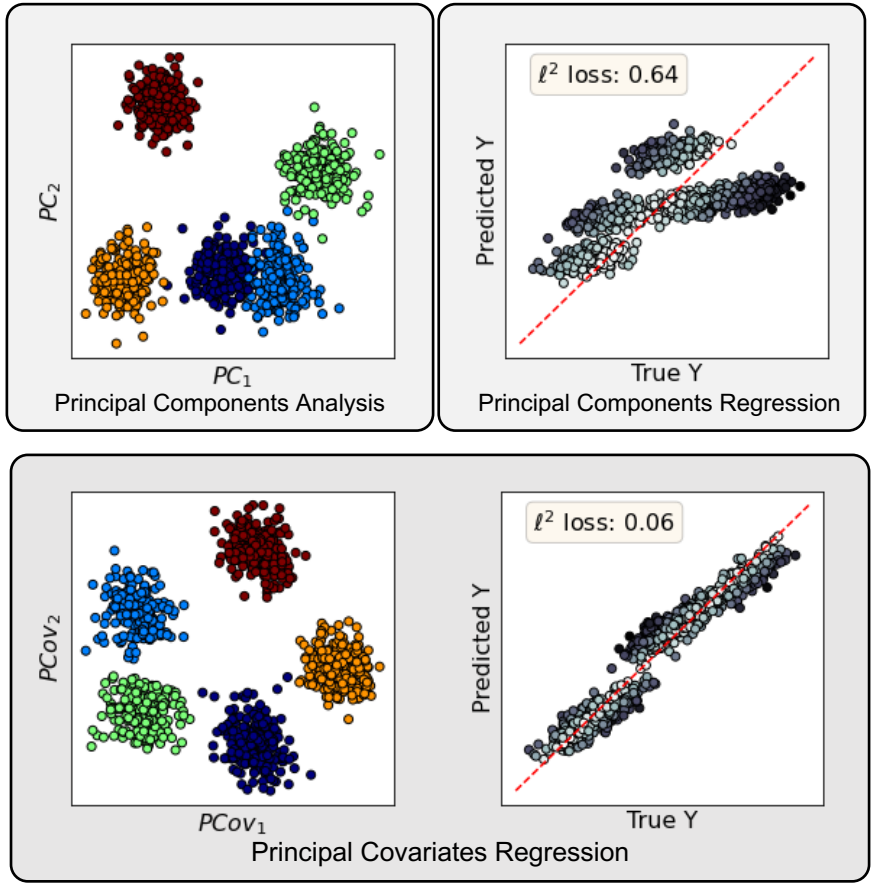
$$\tilde{\mathbf{K}} = \alpha\mathbf{X}\mathbf{X}^{\mathrm{T}} + (1-\alpha)\hat{\mathbf{Y}}\hat{\mathbf{Y}}^{\mathrm{T}} \qquad \tilde{\mathbf{C}} = (\mathbf{C}^{-1/2}\mathbf{X}^{\mathrm{T}})\tilde{\mathbf{K}}(\mathbf{X}\mathbf{C}^{-1/2})$$

S. de Jong, H.A.L. Kiers, Chemom. intell. lab. syst. 14 (1992) 155-164.
scikit-cosmo.readthedocs.io

Principal Covariates Regression (PCovR)

is a dimensionality reduction technique that determines a latent-space projection that incorporate saspects of supervised learning.

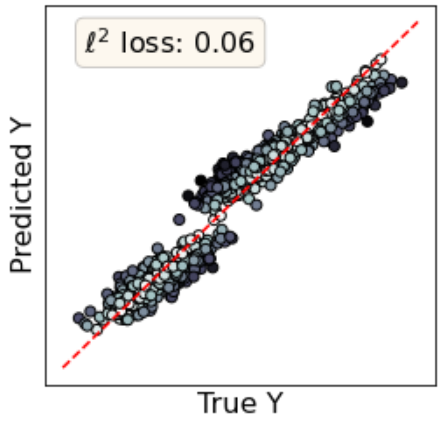


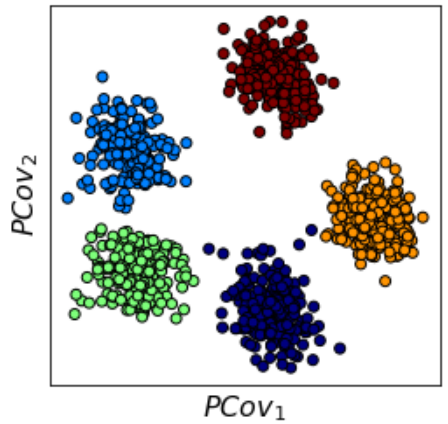


S. de Jong, H.A.L. Kiers, Chemom. intell. lab. syst. 14 (1992) 155-164.
scikit-cosmo.readthedocs.io

Inputs: sklearn.datasets.make_blobs
Regression Model: RidgeCV(cv=5)

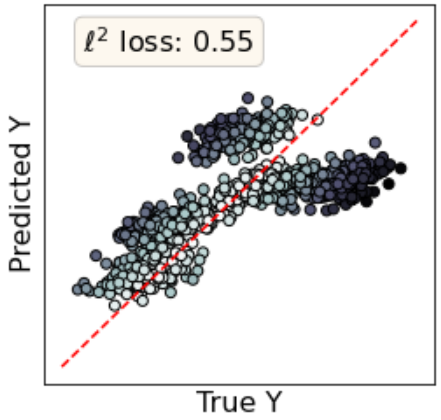
PCovR is controlled by a mixing parameter α that weights the regression and decomposition tasks.

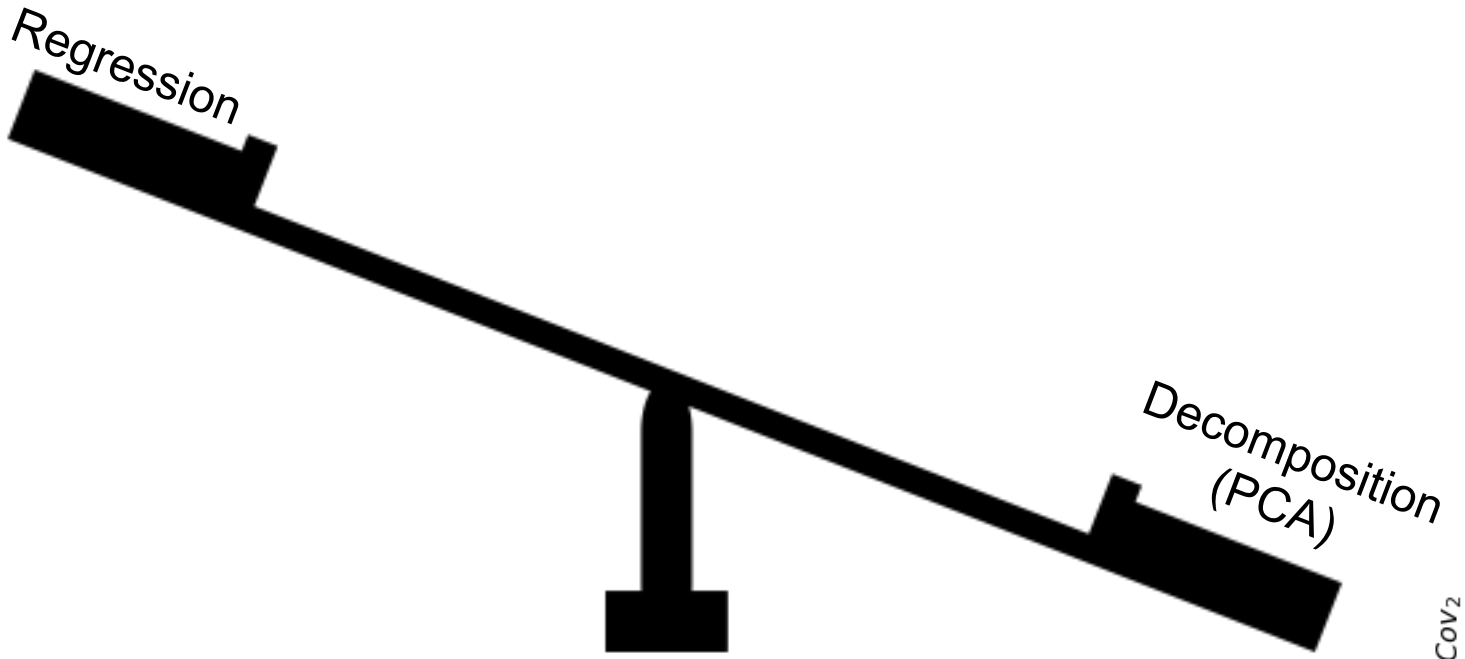




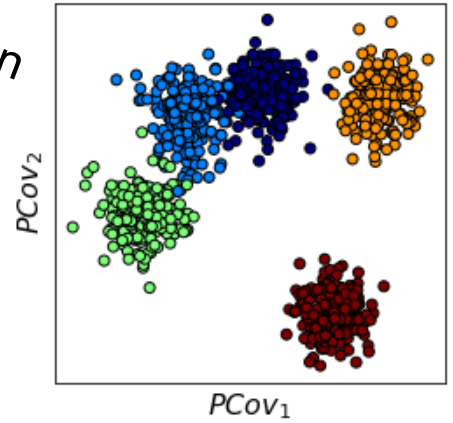
$$\alpha = 0.5$$

PCovR is controlled by a mixing parameter α that weights the regression and decomposition tasks.

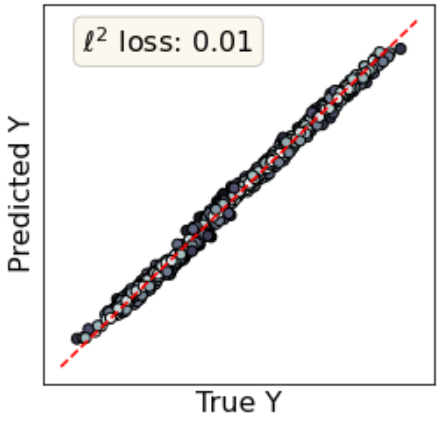


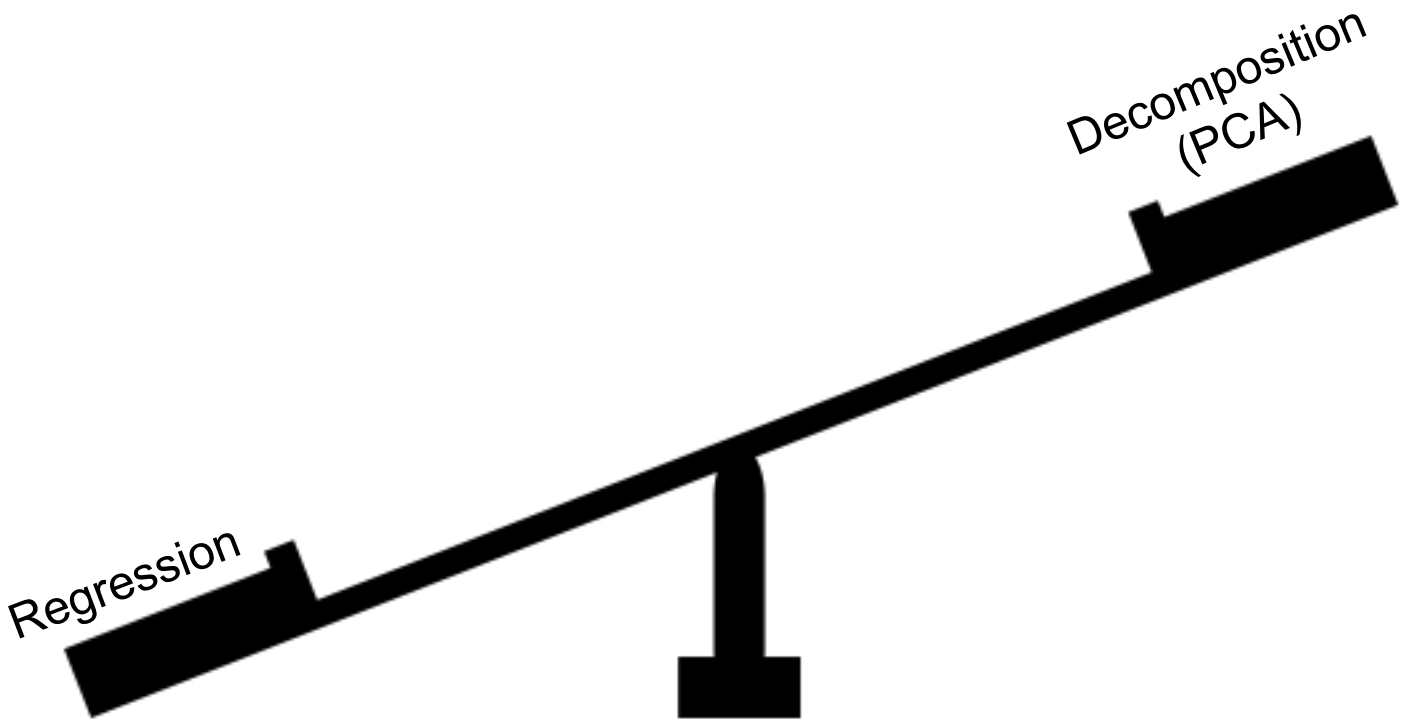


$\alpha = 0.1$

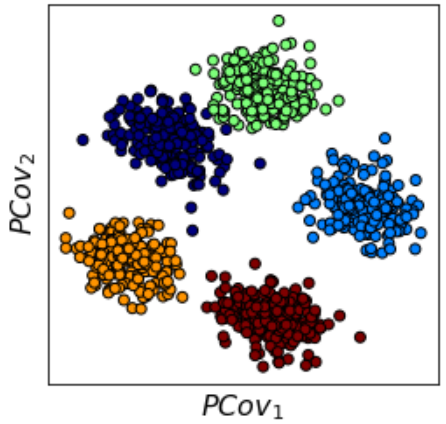


PCovR is controlled by a mixing parameter α that weights the regression and decomposition tasks.





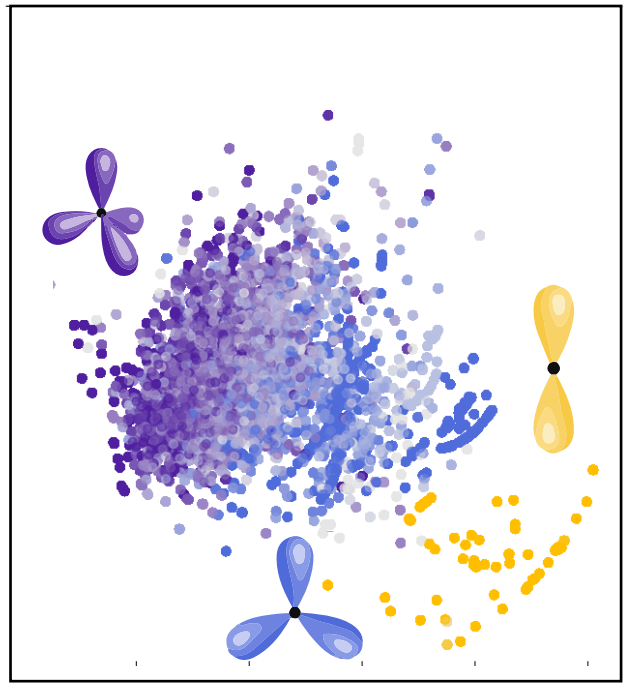
$\alpha = 0.9$



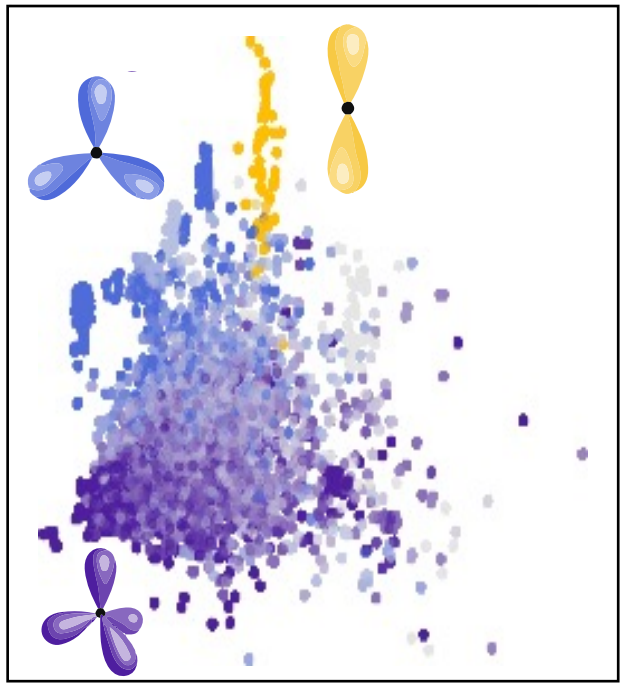
Kernel Principal Covariates Regression

Determines a low-dimension projection from a similarity kernel, considering target data when constructing the projection.

KPCA
(~KPCovR, $\alpha = 1.0$)



KPCovR, $\alpha = 0.5$



$$\tilde{\mathbf{K}} = \alpha \mathbf{X}\mathbf{X}^{\mathrm{T}} + (1-\alpha)\hat{\mathbf{Y}}\hat{\mathbf{Y}}^{\mathrm{T}}$$

gram matrix,
a.k.a. “linear kernel”

non-linear kernel

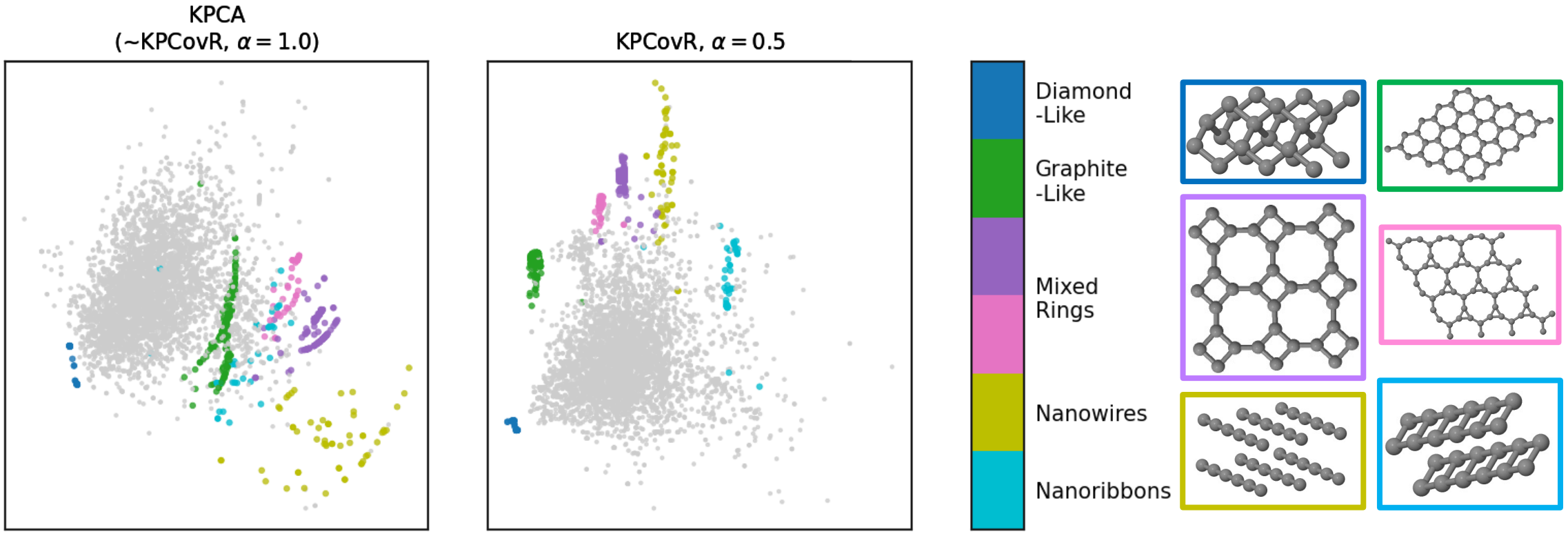
$$\mathrm{K}_{ij} = \mathrm{k}(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2}$$

Inputs: SOAP features of 10,000 AIRSS carbon crystals
Target: energies in [eV / atom]
Kernel Parameters: RBF kernel, γ=$10^{-3.8}$
(1/1) train / test split

B. A. Helfrecht, **RKC**, G. Fraux, and M. Ceriotti. 2020 Mach. Learn.: Sci. Technol. 1 045021
C. J. Pickard. AIRSS Data for Carbon at 10GPa and the C+N+H+O System at 1GPa (2020).
scikit-cosmo.readthedocs.io

Kernel Principal Covariates Regression

Determines a low-dimension projection from a similarity kernel, considering target data when constructing the projection.

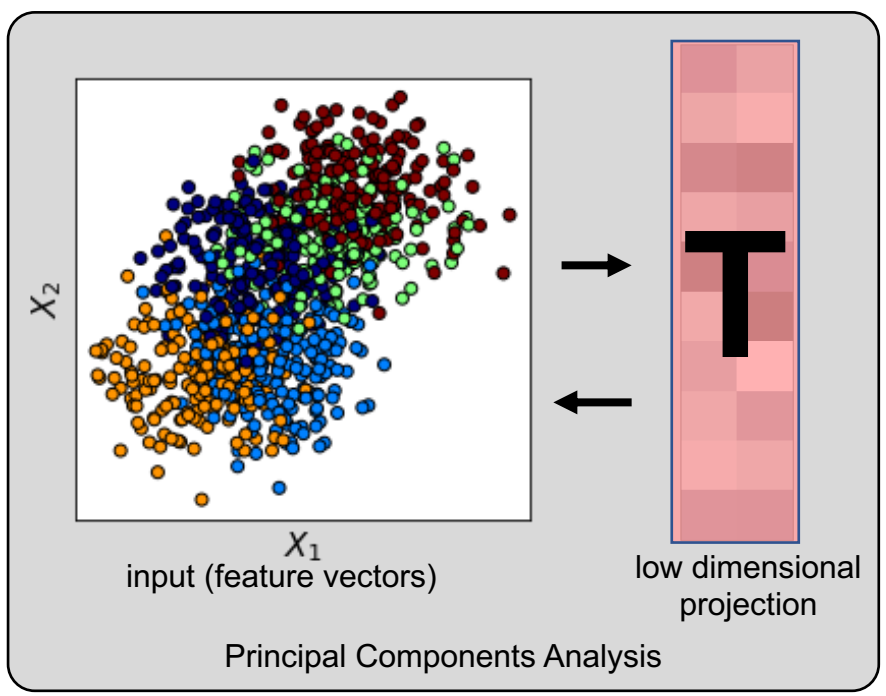


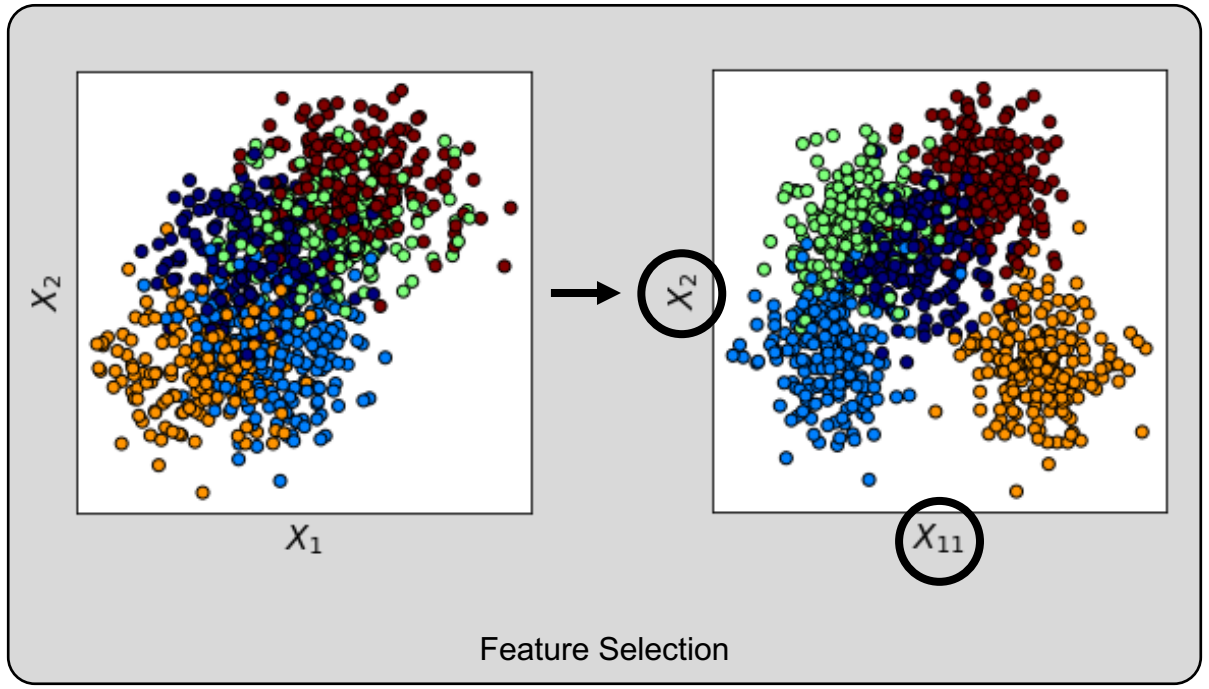
Inputs: SOAP features of 10,000 AIRSS carbon crystals
Target: energies in [eV / atom]
Kernel Parameters: RBF kernel, γ=$10^{-3.8}$
(1/1) train / test split

B. A. Helfrecht, **RKC**, G. Fraux, and M. Ceriotti. 2020 Mach. Learn.: Sci. Technol. 1 045021
C. J. Pickard. AIRSS Data for Carbon at 10GPa and the C+N+H+O System at 1GPa (2020).
scikit-cosmo.readthedocs.io

What if the features carry inherent meaning?

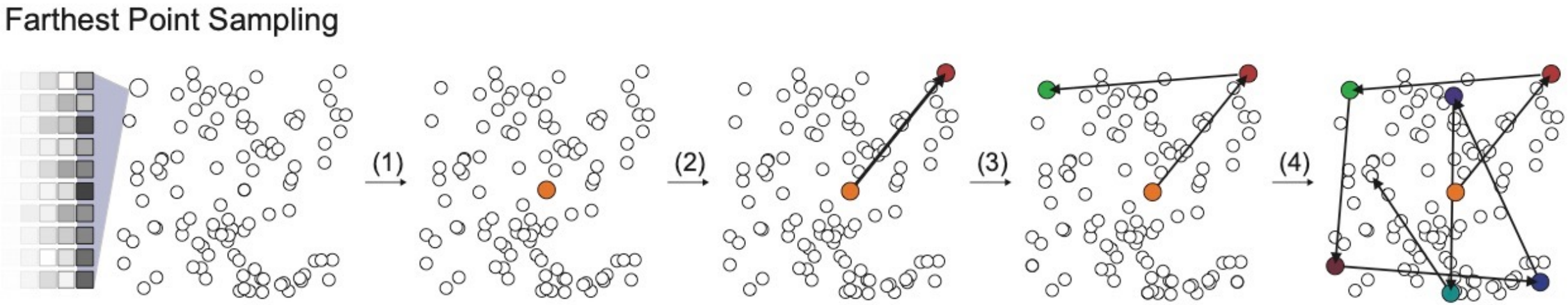
Many dimensionality reduction techniques construct a *new* set of features, but what if you want to just work with a subset of the old set?





Farthest Point Sampling (FPS)

FPS aims to select a diverse subset of features or samples that cover the greatest portion of sample or feature space.

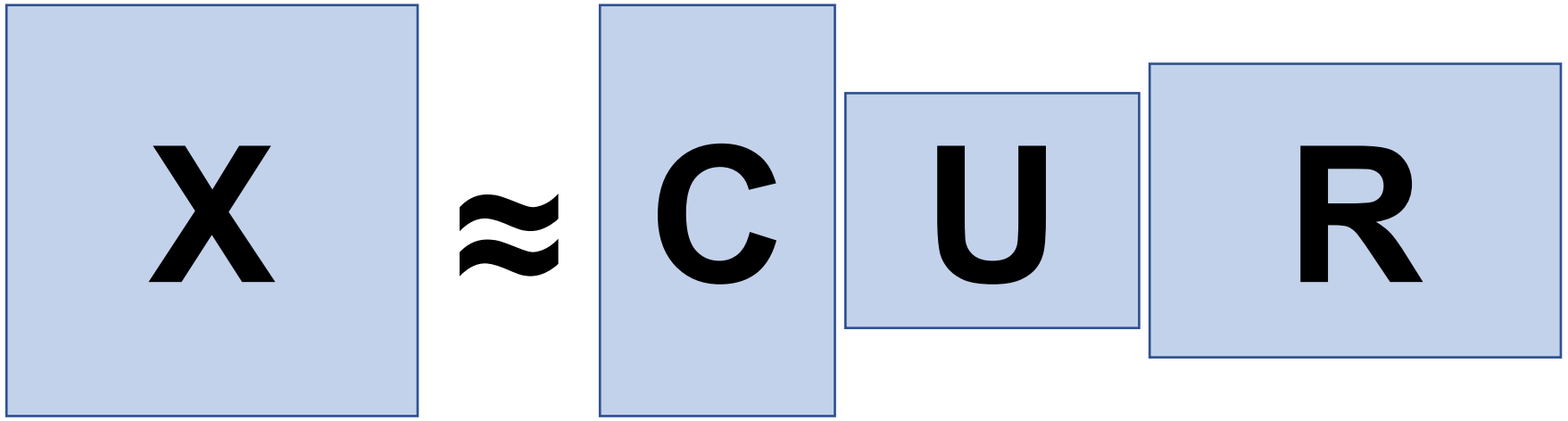


1. Choose a first point
2. Compute distance d
3. Choose point with highest $\min(d)$ to the selected points
4. Repeat 1-3 until you have enough features!

CUR Decomposition

Traditional CUR decomposition selection aims to select “important” features or samples from the overall distribution.

Head Canon:
“Columns und Rows” decomposition



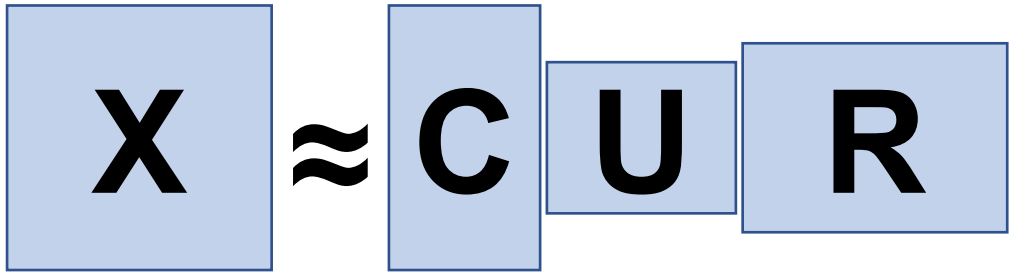
Original Matrix

Subset of Columns

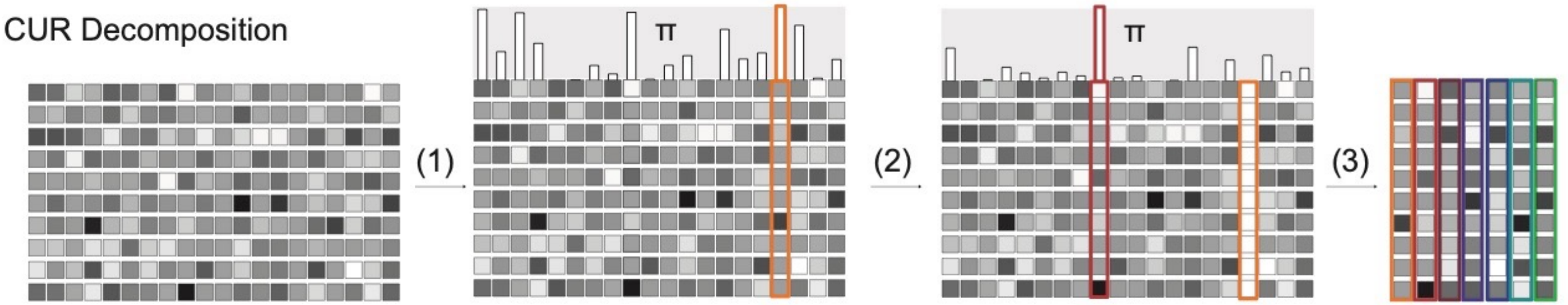
Subset of Rows

CUR Decomposition

Traditional CUR decomposition selection aims to select "important" features or samples from the overall distribution.



Original Matrix | Subset of Columns | Subset of Rows

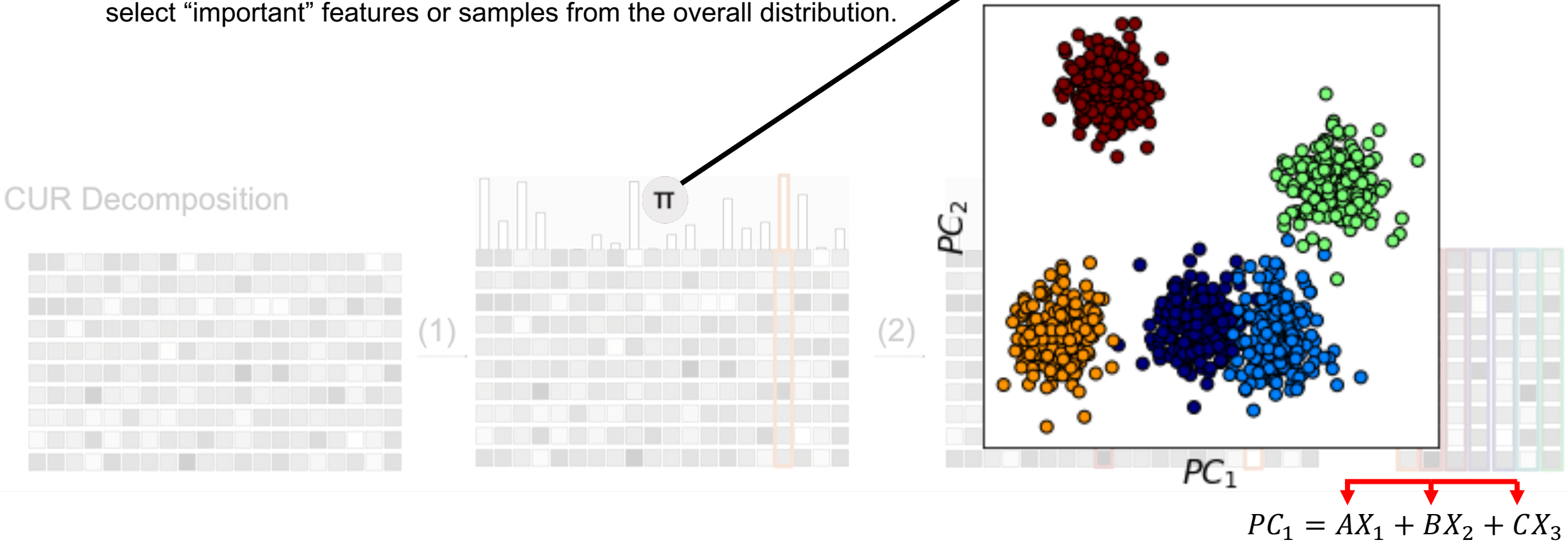


1. Compute importance score π
2. Choose column with highest π
3. Orthogonalize with respect to last chosen column.
4. Repeat 1-3 until you have enough features!

CUR Decomposition

Traditional CUR decomposition selection aims to select “important” features or samples from the overall distribution.

How do we calculate π?



PCov-FPS and Pcov-CUR

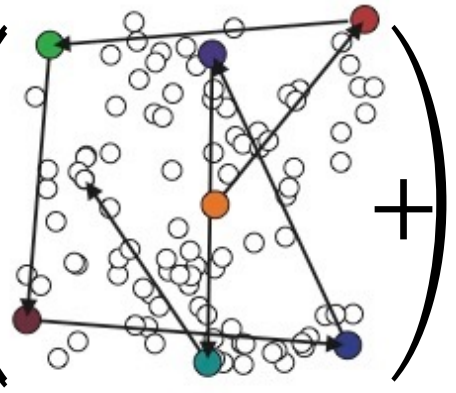
Both FPS and CUR can be translated to PCovR space for both feature (and sample) selection.

$$\tilde{\mathbf{C}} = (\mathbf{C}^{-1/2}\mathbf{X}^{\mathbf{T}})\tilde{\mathbf{K}}(\mathbf{X}\mathbf{C}^{-1/2})$$

feature selection

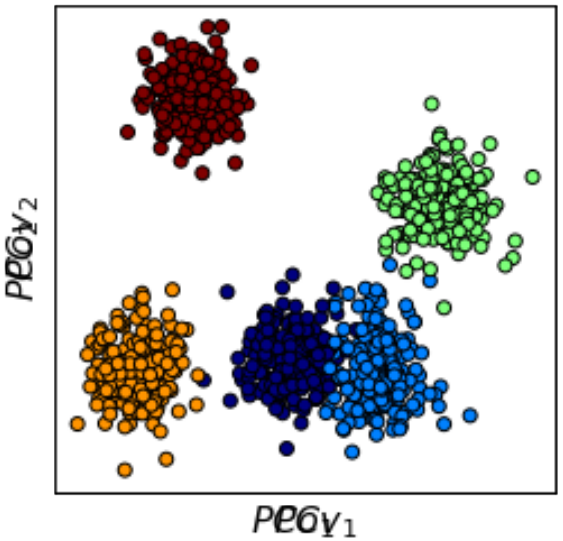
Farthest Point Sampling (FPS)

$$\mathbf{d} = f\left(\quad + \quad \text{Contribution to Regression Weights} \right)$$



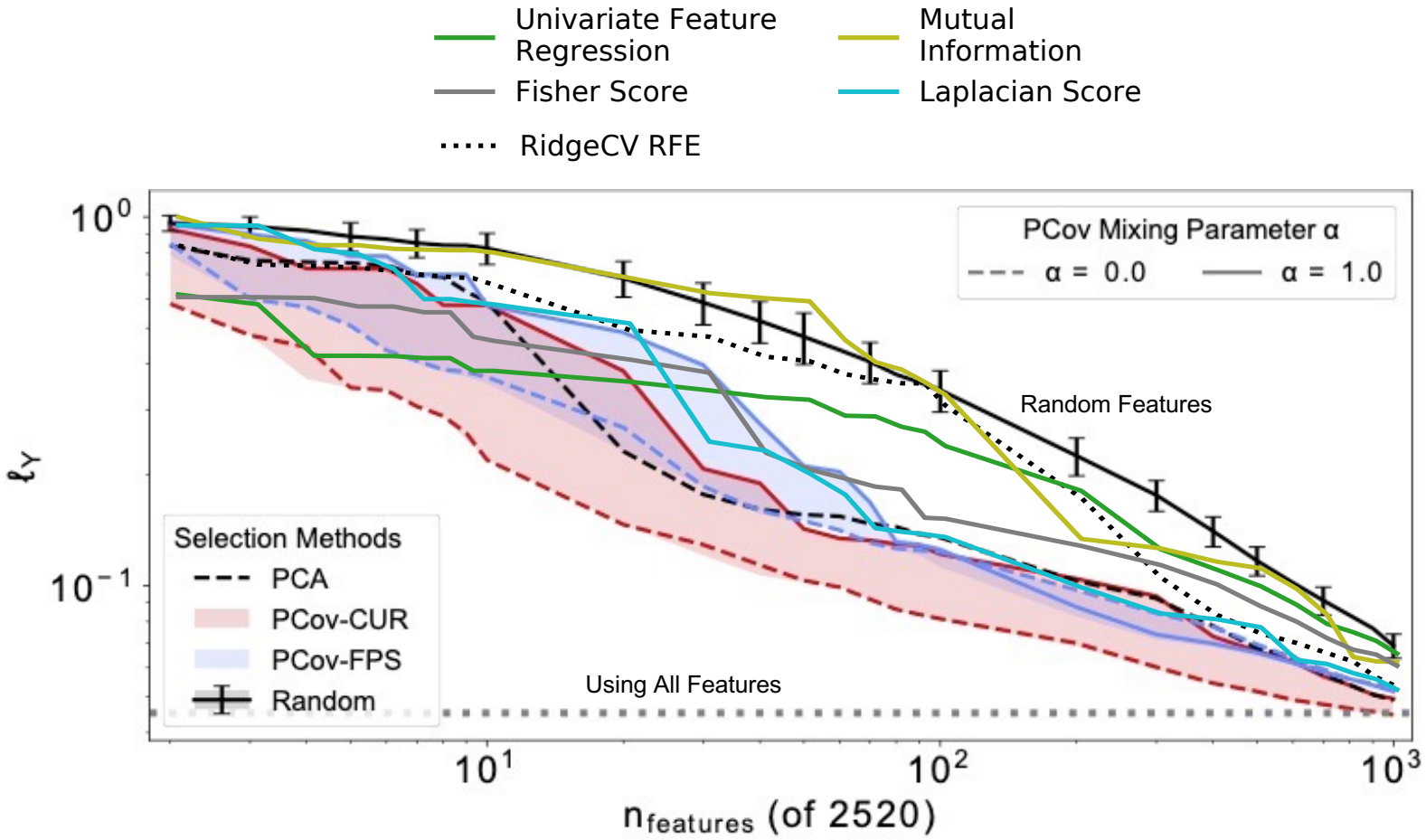
CUR Decomposition

$$\pi = f\left(\quad \right)$$



Linear Regression

Using PCov-style feature selection will universally out-perform common feature selection metrics available via popular packages.

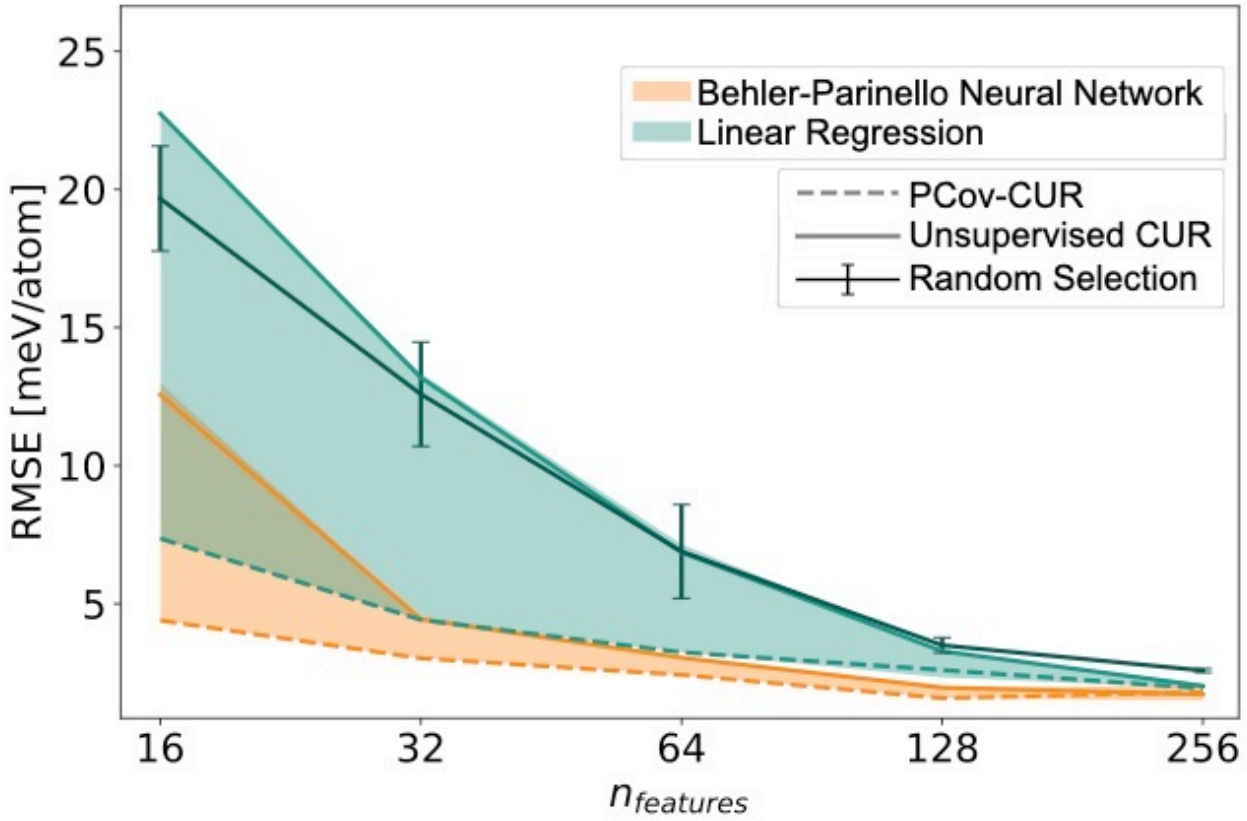


Inputs: SOAP vectors for small molecules containing C + H + N + O, (9 / 1) train / test split
Target: NMR chemical shieldings in ppm
Model used: 5-fold cross-validated linear ridge regression

RKC, et al 2021 Mach. Learn.: Sci. Technol. 2 035038
scikit-cosmo.readthedocs.io

Behler-Parinello Neural Networks

Introducing supervised aspects to feature selection invariably improves regression performance – even in non-linear models -- such as determining energies and forces using a neural network.



Inputs: symmetry functions of benzene rings from a simulation trajectory, (7/2/1) train / validation / test split

Target: energies in [meV / atom]

Models used: 5-fold cross-validated linear ridge regression, Behler-Parinello Neural Network

RKC, et al 2021 Mach. Learn.: Sci. Technol. 2 035038
scikit-cosmo.readthedocs.io

scikit-COSMO

scikit-COSMO is a collection of scikit-learn compatible utilities that implement methods developed at COSMO.

scikit-cosmo.readthedocs.io
https://www.github.com/cosmo-epfl/scikit-cosmo/

kernel-tutorials

A set of utilities and pedagogic notebooks for the use of linear and kernel methods in atomistic modeling
https://www.github.com/cosmo-epfl/kernel-tutorials/

librascal

A scalable and versatile library to generate representations for atomic-scale learning
https://www.github.com/cosmo-epfl/librascal/

chemiscope

chemiscope is an interactive structure/property explorer for materials and molecules. The goal of chemiscope is to provide interactive exploration of large databases of materials and molecules and help researchers to find structure-properties correlations inside such databases.
chemiscope.org

Hybrid Unsupervised-Supervised Machine Learning Models for Materials Science

Rose K. Cersonsky

RoseCersonsky.com

B. A. Helfrecht, **RKC**, G. Fraux, and M. Ceriotti
"Structure-property maps with Kernel principal covariates regression."
2020 Mach. Learn.: Sci. Technol. 1045021.
https://iopscience.iop.org/article/10.1088/2632-2153/aba9ef

RKC, B. A Helfrecht, E. A. Engel, and M. Ceriotti .
"Improving Sample and Feature Selection with Principal Covariates Regression"
2021 Mach. Learn.: Sci. Technol. 2 035038
https://doi.org/10.1088/2632-2153/abfe7c.

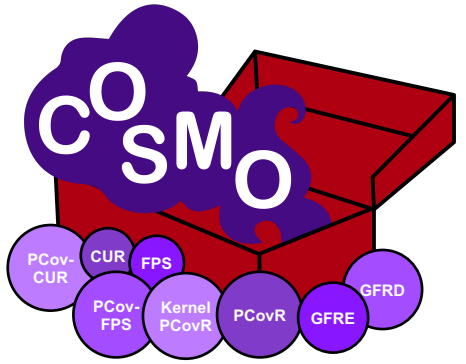
G. Fraux, **RKC**, M. Ceriotti. "Chemiscope"
2020 Journal of Open Source Software, 5(51), 2117.
https://doi.org/10.21105/joss.02117

S. de Jong, H.A.L. Kiers
"Principal Covariates Regression: Part 1."
Chemom. intell. lab. syst. 14 (1992) 155-164.
https://doi.org/10.1016/0169-7439(92)80100-I

scikit-COSMO

scikit-COSMO is a collection of scikit-learn compatible utilities that implement methods developed at COSMO.

scikit-cosmo.readthedocs.io
https://www.github.com/cosmo-epfl/scikit-cosmo/



Come see me at MRS!
BI02 & CH04