Controlling a confounding effect in predictive analysis

Darya Chyzhyk[1], Gaël Varoquaux[1], Bertrand Thirion[1] and Michael Milham[2]

[1]Parietal - Inria / CEA. Paris-Saclay University. France
[2]Center for Biomedical Imaging and Neuromodulation, Nathan S. Kline Institute for Psychiatric Research, USA

EuroSciPy 2019, Bilbao

September 2rd - 6th, 2019

Outline

Introduction

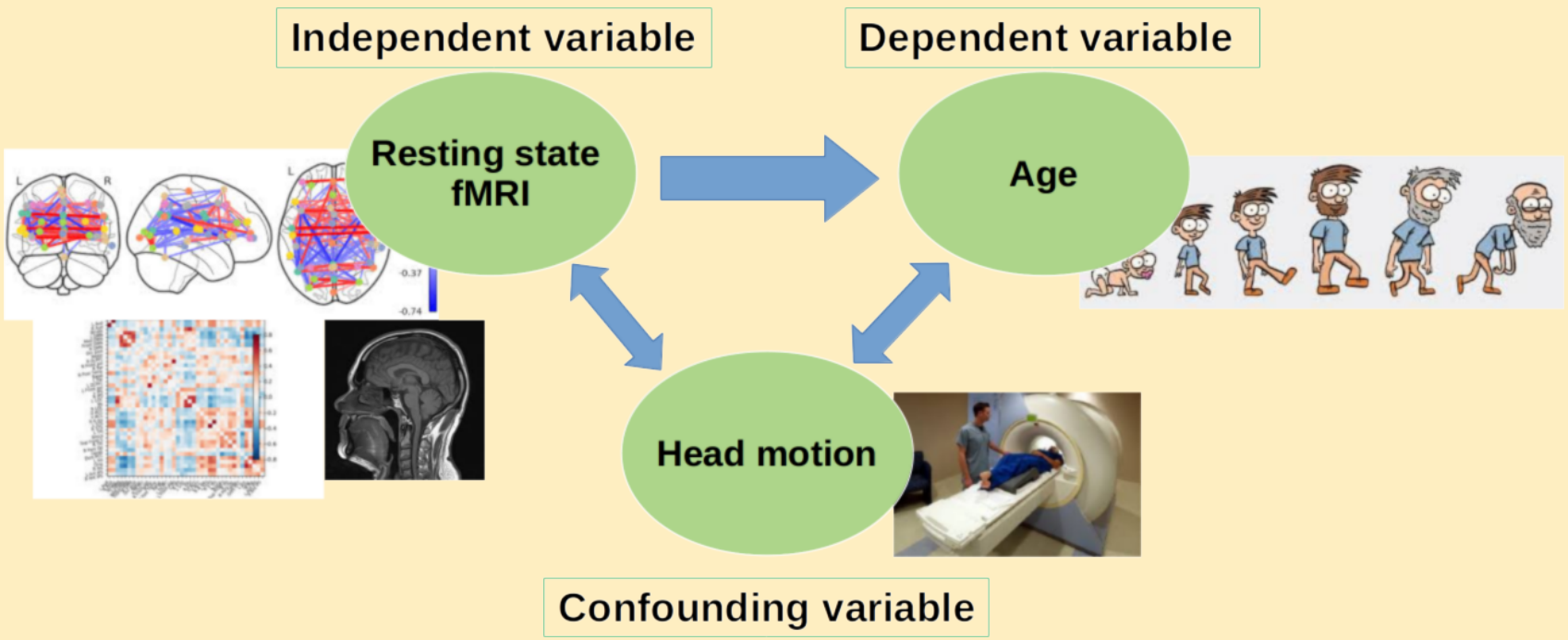
- Predictive models applied on **brain** images can extract imaging **biomarkers** of pathologies.
- Successful prediction may be driven by a **confounding** effect that is correlated with the effect of interest.

Example

Power JD et al showed that in-scanner **head motion** produces a significant **confound** for rest fMRI: motion creates systematic differences in brain signals, and in-scanner motion varies with subjects' **age**.

Introduction

Example



Introduction

- We introduce a non-parametric approach to **control for a confounding effect** in a predictive model
- It is based on forming a **test set** on which the effect of interest is **independent** from the confounding effect.
- We show that using a linear model to remove the effect of **age** on the **brain signals** leads to pessimistic scores on **fluid intelligence** prediction

Outline

Formalizing the problem

We consider on n subjects:

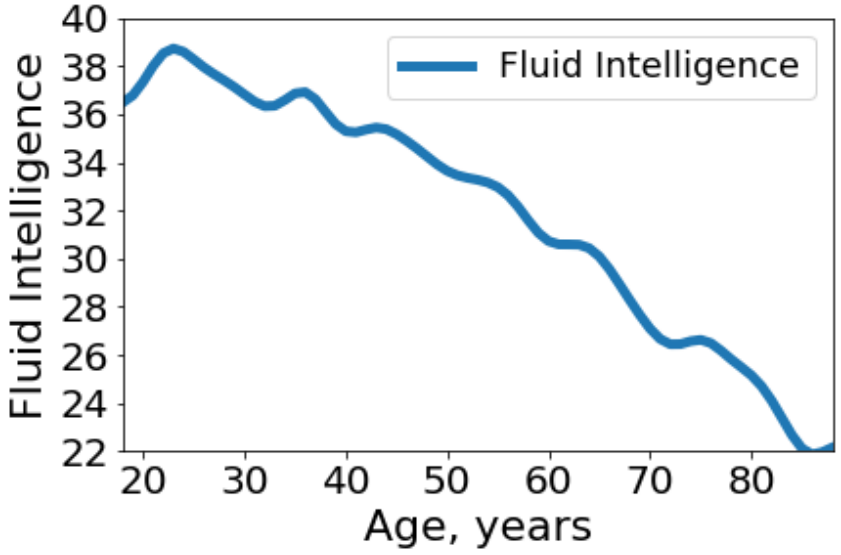
- $\mathbf{X} \in \mathbb{R}^{n \times p}$ – brain signals,
- $\mathbf{y} \in \mathbb{R}^{n}$ – the biomarker target,
- $\mathbf{z} \in \mathbb{R}^{n}$ – a confounding effect.

- An imaging biomarker predicts $\mathbf{y}$ from $\mathbf{X}$ independently of $\mathbf{z}$.
- If $\mathbf{y}$ and $\mathbf{z}$ are not independent, we have to account for this effect.
- Prediction of a target $\mathbf{y}$ mediated by a phenotypic information $\mathbf{z}$ may be misleading or useless.

Formalizing the problem

Example

Fluid intelligence declines with age $\Longrightarrow$ Link between brain structures **X** and fluid intelligence **y** is effected by confound age **z**



The results of the prediction model can be useless.

The **problem** that we focus on here it to test whether we can predict y from X rather than z.

Existing approaches

Deconfounding

The classical procedure – based on the **general linear model** variables that are correlated

- **Weakness**: designed to control **in-sample** properties, while predictive models are designed for **out-of-sample** prediction.
- Deconfounding **jointly**: breaks the statistical validity of cross-validation by coupling the train and the test set
- Powerful model **overfit**: the deconfounding procedure may remove signal of interest, unrelated to the confound.

Proposed method: confound-isolating cross-validation

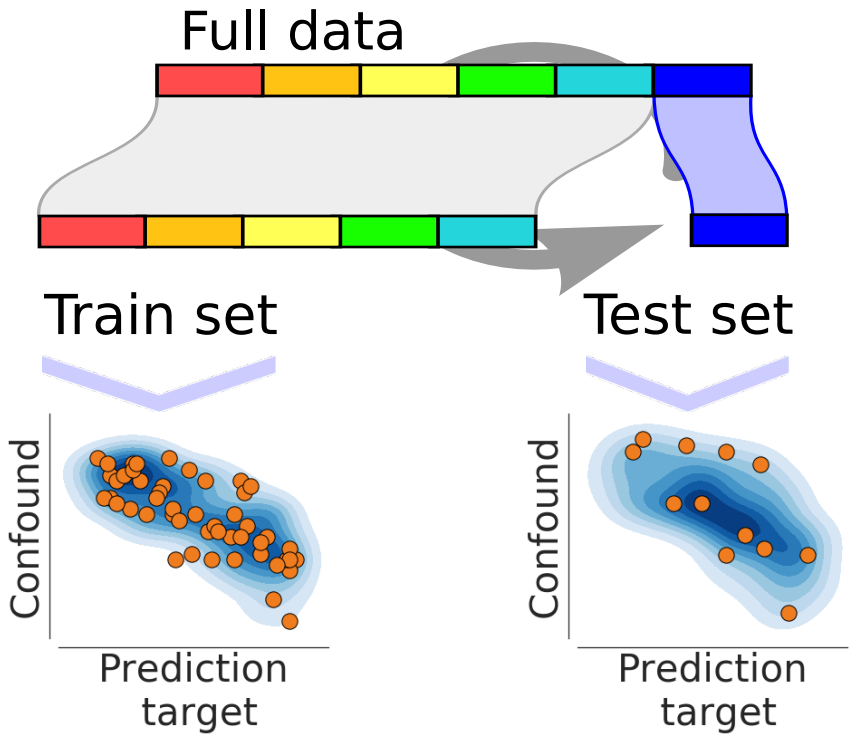
Strategy

- We use as a **test** set a subset $\mathcal{S}$ of the data such that $\mathbf{y}_{\mathcal{S}}$ and $\mathbf{z}_{\mathcal{S}}$ are close to **independent**
- The remainder – **training** set – we use to learn to predict $\mathbf{y}$ from $\mathbf{X}$

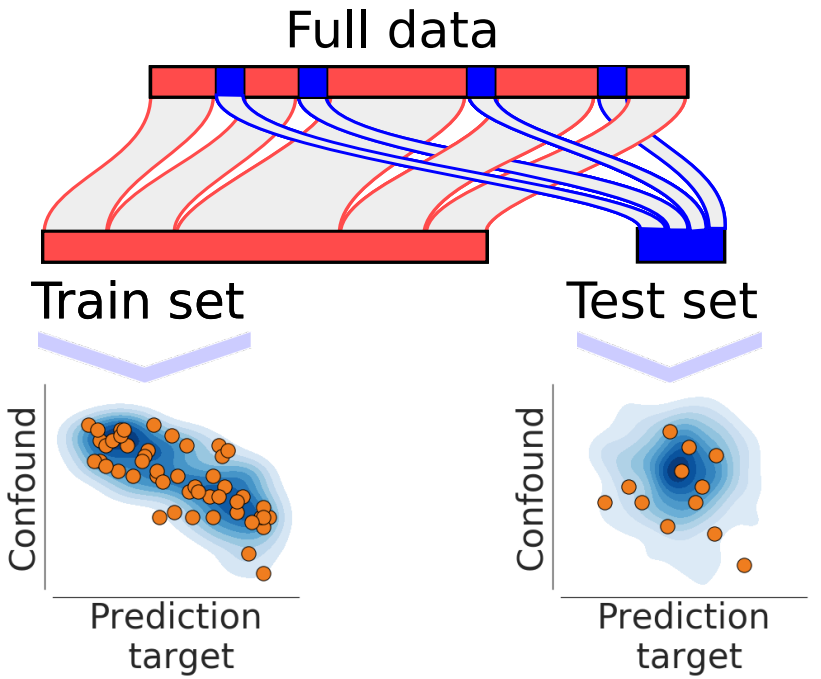
If the prediction generalizes to the test set $\mathcal{S}$, the learned relationship between $\mathbf{X}$ and $\mathbf{y}$ is not entirely mediated by $\mathbf{z}$.

Proposed method: confound-isolating cross-validation

K-fold cross-validation strategy



Confound-isolating cross-validation



Confound-isolating cross-validation

Goal – independence of $\mathbf{y}_{\mathcal{S}}$ and $\mathbf{z}_{\mathcal{S}}$

$$p(\mathbf{y}, \mathbf{z}) = p(\mathbf{y})\, p(\mathbf{z})$$

$p((\mathbf{y}, \mathbf{z}))$ – the joint probability function of $\mathbf{y}$ and $\mathbf{z}$,
$p(\mathbf{y})$ and $p(\mathbf{z})$ – the marginal probability distributions.

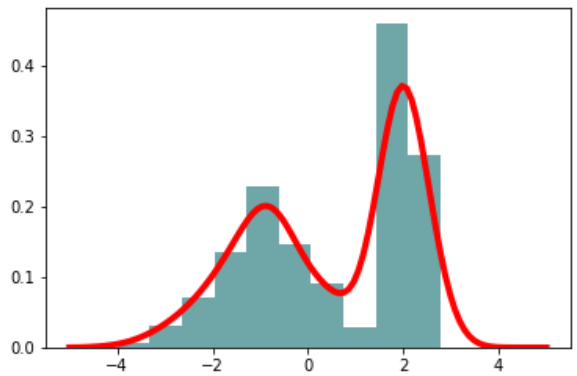
A related quantity is **mutual information** – level of dependency:

$$\mathbb{E}\left[\log\left(\frac{p((\mathbf{y},\mathbf{z}))}{p(\mathbf{y})p(\mathbf{z})}\right)\right]$$

Confound-isolating cross-validation

In practice

- estimate the probability density functions with a kernel-density estimator (KDE) using Gaussian kernels.



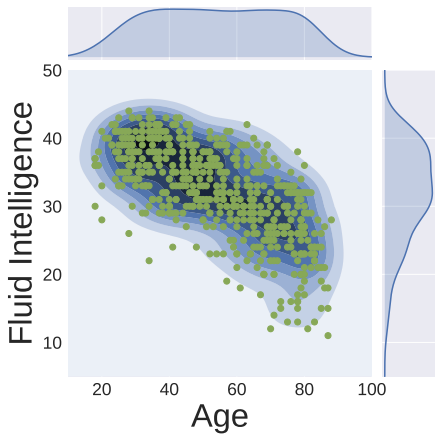
  - iteratively create the test $\mathcal{S}$ set by removing subjects
    - at each iteration we have matching problem $p(\mathbf{y}_{\mathcal{S}}, \mathbf{z}_{\mathcal{S}})$ and $p(y\mathcal{S})\, p(z\mathcal{S})$
    - resolve using **importance sampling**: we draw randomly 4 subjects to discard with a probability $\frac{p(\mathbf{y}_{\mathcal{S}}, \mathbf{z}_{\mathcal{S}})}{p(\mathbf{y}\mathcal{S})\, p(\mathbf{z}\mathcal{S})}$

Outline

A rest-fMRI dataset

We consider the Cambridge Centre for Ageing and Neuroscience (**CamCan**) data.

- CamCan data displays a strong relation between fluid intelligence and age
- When extracting biomarkers of fluid intelligence, the danger is to simply predict age.



Prediction from rest-fMRI functional connectivity

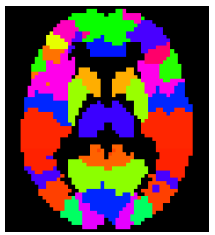
- each row of $\mathbf{X}$ is a vectorized form of the functional connectivity matrix for each subject
- the target vector $\mathbf{y}$ is the fluid intelligence score
- the confound $\mathbf{z}$ is the age in years.

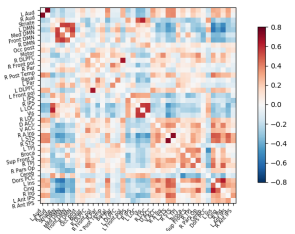
Prediction pipeline

We use functional-connectivity matrices as brain imaging signals to build our biomarkers

These matrices were generated by:

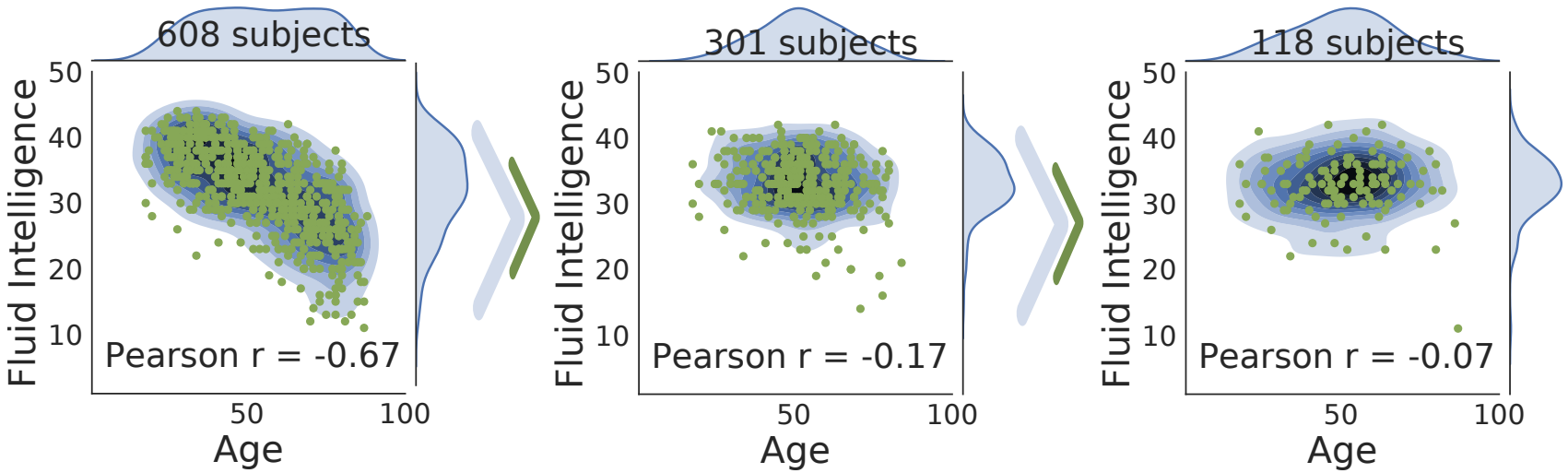
- extracting the time series from predefined **atlases BASC** with 64 regions
- calculating connectivity matrices using **tangent connectivity measure**





As a prediction model we choose the standard **ridge regression** with nested cross-validation

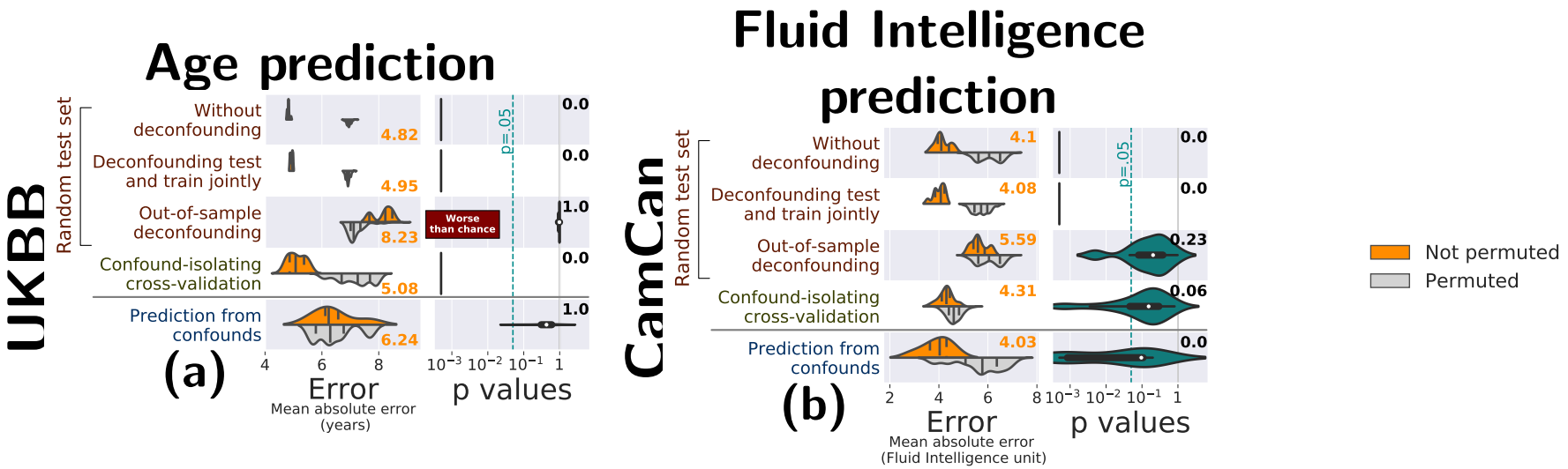
Results



Evolution of the test set created by Confound-isolating cross-validation: joint distribution of the target (Fluid intelligence) and the confound (Age)

We show the process of selecting proper samples for the test set: entire dataset, half of the iterations, the final test set.

Results

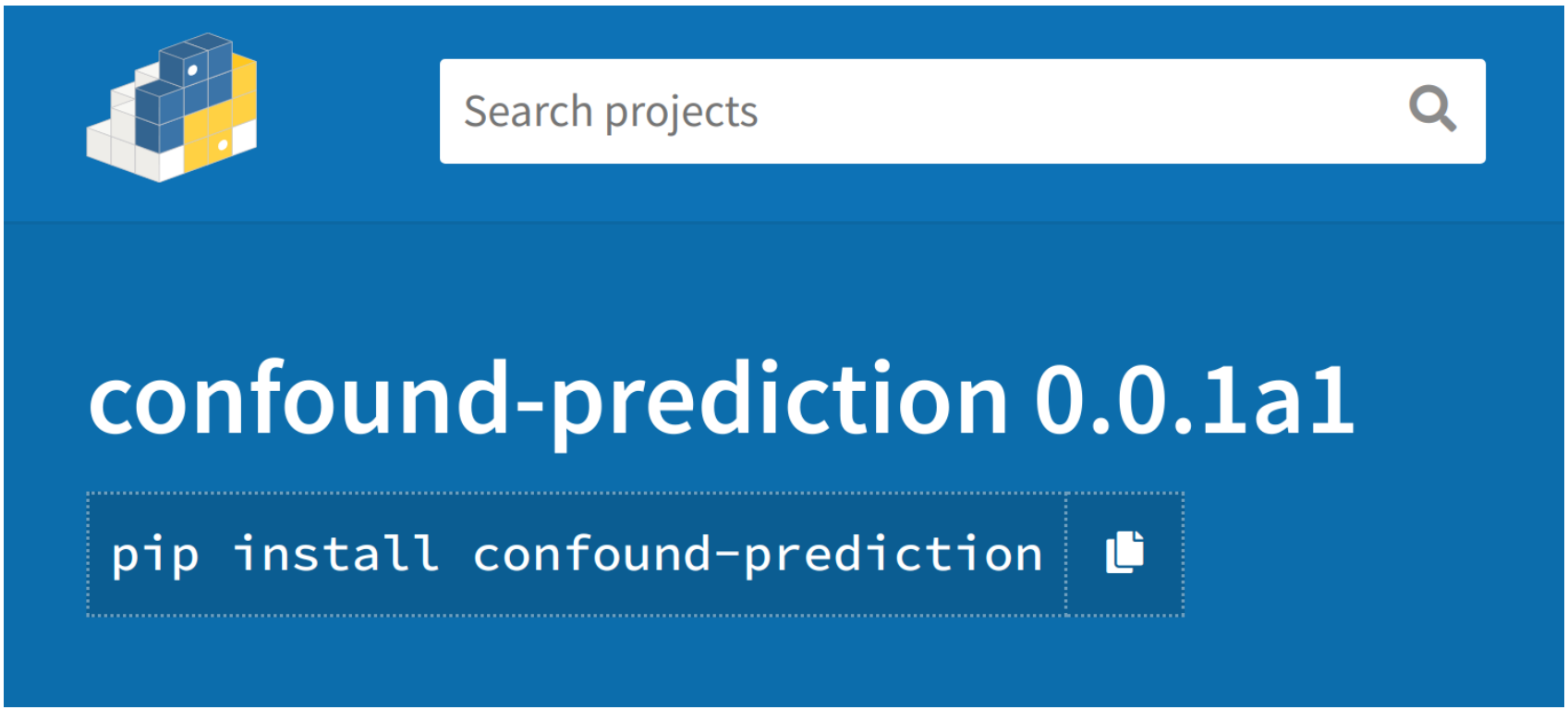


The results show

- random sampling capture the age information
- classical deconfounding gives significant but negative results
- proposed non-parametric Anti Mutual Information method is shows more favorable and less significant prediction.

Outline

pip install


Search projects
confound-prediction 0.0.1a1
pip install confound-prediction

Using *confound_prediction*

Create the test and training sets with *Confound Isolation sampling*

```
from confound_prediction.deconfounding import confound_isolating_cv

x_test, x_train, y_test, y_train, ds_test, ids_train = \
    confound_isolating_cv(X, y, z, random_seed=None, min_sample_size=None,
                          cv_folds=10, n_remove=None)
```

Using *confound_prediction*

Generate the data

```
from confound_prediction.data_simulation import simulate_confounded_data

X, y, z, = simulate_confounded_data(link_type='direct_link', n_samples=1000,
                                    n_features=100)
```

Deconfounding

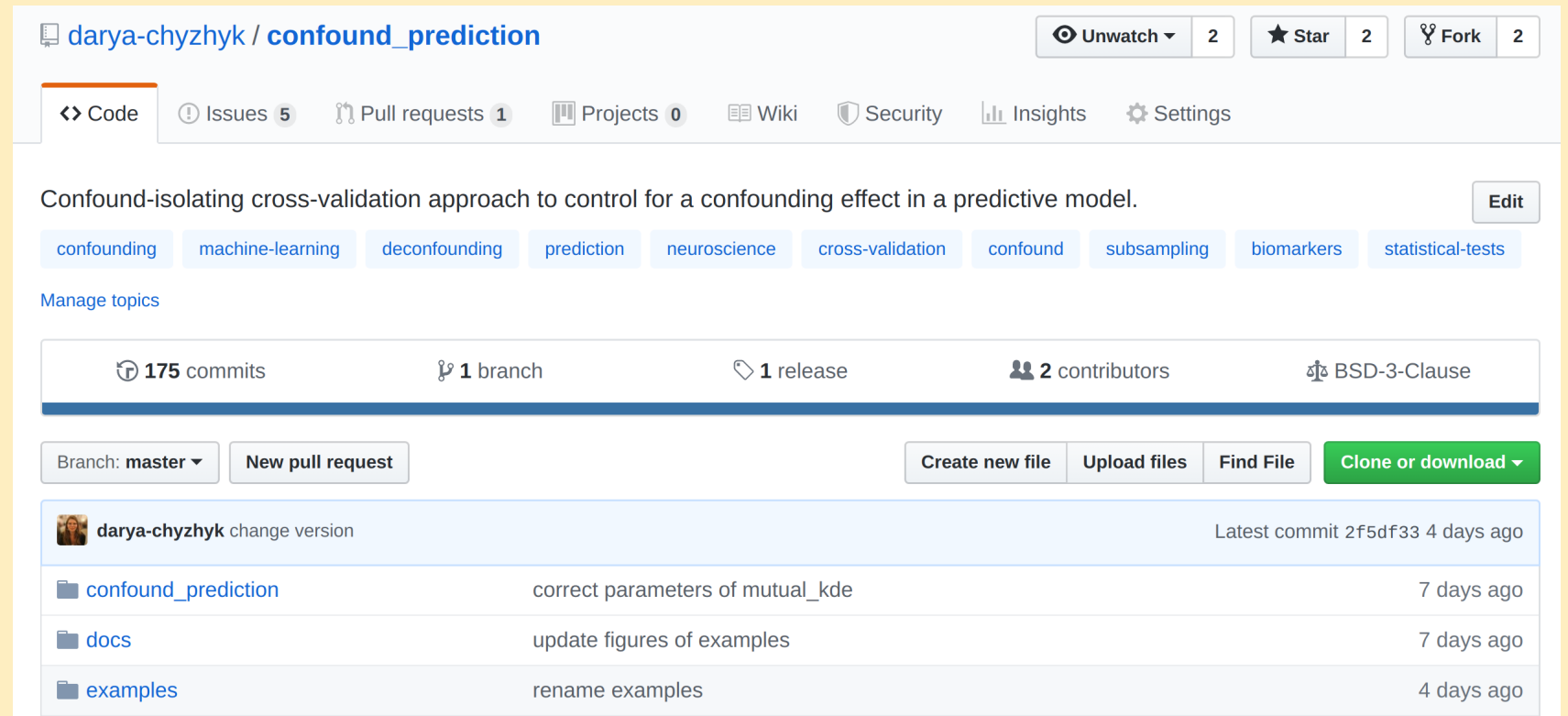
```
from confound_prediction.deconfounding import confound_regressout

x_test, x_train, y_test, y_train, _, _ = \
    confound_regressout(X, y, z, type_deconfound='out_of_sample',
                        min_sample_size=None, cv_folds=10, n_remove=None)

x_test, x_train, y_test, y_train, _, _ = \
    confound_regressout(X, y, z, type_deconfound='False',
                        min_sample_size=None, cv_folds=10, n_remove=None)
```

Github

Create the test and training sets with *Confound Isolation sampling*



Outline

Conclusions

- We consider the **problem** of building biomarkers in the presence of confounding effects that contribute to prediction.
- Deconfounding approaches used in **standard GLM-based analysis** can easily lead to pessimistic evaluations
- Our approach uses **anti mutual information** sampling to craft a test set on which the effect of interest is independent from the confound.
- It enables a correct test of the predictive power from brain imaging without killing potentially useful shared signal.
- It is **non parametric** and does not rely of a linear confounding model.
- We demonstrate the use of the method on large sample **resting-state fMRI**, predicting fluid intelligence from functional connectivity with effect of age.

Thank you!

Controlling a confounding effect in predictive analysis

Darya Chyzhyk

Twitter: @DaryaChyzhyk
GitHub: darya-chyzhyk