Common Issues

Controlled vs naturalistic data

Lack of trials

Nonstationarity and long timescales

Trials/Sessions/Animals/Conditions

Binning

Missing data

Too much data

Multicollinearity

Gotcha: Normalization and correlations

Gotcha: Correlated training/test data

Gotcha: What is N?

Interpreting analysis results

Common Issues - Controlled vs naturalistic data

Controlled

Many trials/repetitions

Simplified task/stimuli

Usually uncorrelated stimuli

Experiment designed as simplest way to address specific scientific question

Can use “traditional” statistics

Unclear whether/how results generalize to natural settings

Naturalistic

Closer to natural setting

Often no trials or exact repetitions

Complex stimuli

Stimuli often correlated (see “multicollinearity”)

Experiment designed to evoke naturalistic behavior with little interference

Traditional statistics often insufficient

Common Issues - Trials/Sessions/Animals/Conditions

Common Issues - Nonstationarity and long timescales

Common Issues - Binning

Common Issues - Missing data

Common Issues - Too much data

Common Issues - Multicollinearity

Multicollinearity: highly correlated inputs

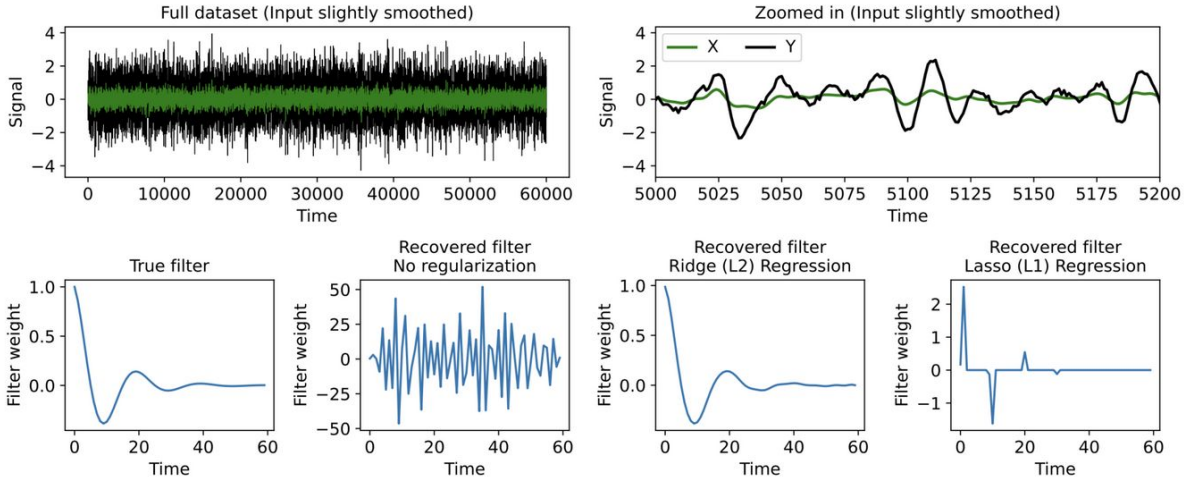
- Example: any time-series in which input timepoints are not i.i.d.
- *Extremely* when “input” is measured rather than controlled

Problem: ambiguity in which inputs to weight as predictors

- Example: x(t) and x(t+1) highly correlated → 2x(t) + 2x(t+1) ≈ x(t) + 3x(t+1) ≈ 4x(t) ≈ 4x(t+1)
- Many different solutions for how to weight inputs to predict output y(t)
- Unregularized regression techniques → weights highly sensitive to noise… often get e.g. 1000x(t) - 996x(t+1)

Solution: need “rule” for choosing which inputs to weight

- Ridge regression (L2 penalty): similar inputs → similar weights
- Lasso (L1 penalty): sparse set of nonzero weights on input
- Represent D-dim weight vector (filter) as combo of smaller number of weight vectors



Common Issues - Gotchas: normalization

Common Issues - Gotchas: correlated training/test data

Common Issues - Gotchas: what is N?

Common Issues - Interpreting data-driven analysis

Is my result a product of my data or my analysis?

Solution: run same analyses on null datasets