Abstract:

"Circular inference" is a pejorative coined for methods in which a hypothesis is selected after looking at the data, but the inferential procedures treat it as if it was determined in advance. Unfortunately, many throughput screening experiments in genomics or neuroimaging seek to do exactly this: identify regions (bumps) of high signal in the data and evaluate these found regions using the same data. Simple estimators that ignore the selection will be biased; when the data is non-stationary, this bias can vary dramatically between different regions. Nevertheless, methods for evaluating and comparing selected regions are crucial, because typically only a handful of regions can be further explored in tailored follow up studies.

In this talk I describe a new conditional inference approach for characterizing these found regions by estimating their population parameters. Our method explicitly models the selection procedure, and simulates from the conditional distribution to estimate the underlying parameters. Efficient strategies for providing p-value, estimators and intervals will be discussed, as well as power versus accuracy tradeoffs. I will demonstrate the new method for estimating bumps in a comparison of DNA-methylation patterns across tissue type.

This is joint work with Jonathan Taylor and Rafael Irizarry.