

The background is a dark blue gradient with a subtle pattern of small white dots. On the left side, there are several concentric circles and a curved scale with numerical markings from 160 to 260. The scale is white and curves from the top left towards the bottom left. There are also some faint, larger concentric circles and arrows scattered across the background.

BASIC STATS IV

CROSS VALIDATION

CHECKING IN WITH THE PROS

Ivezic et al

- use leave-one-out cross validation to find optimal bandwidth for KDE (“hyperparameter”)
- basic ~~cross~~ validation:
 - Training = 50-70%
 - CV = half rest
 - Test = other half rest
- our activity:
 - definition of error not obvious (no “rms”)
 - lots of data not used in final calibration
 - swapping sample defns

SAMSI astrostatisticians

- better ways to find optimal bandwidth for KDE (in R, translatable to python)
- optimal bandwidth depends on question (produce KDE density plot, 1 vs. 2 Gaussian model, etc.)
- default cross validation = **k-fold** (leave-one-out is $k=N_{pts}$)
 - divide into k subsamples
 - train k models, each time leaving out one subsample for CV – compute median model
 - for robust errors, train $k*(k-1)$ models saving out a test subsample each time too

PROJECT REQUIREMENT

- Use k-fold (multi-sample) cross-validation
- Compute errors