Looking at the data in low dimensions

1. Univariate continuous data

For each variable we can the four panel EDA univariate distribution views

Label panel should contain the variable name and units of measure and may include additional summary statistics such as the mean, median, standard deviation and so on.

Density plot panel can show modes, skewness (lack of symmetry), tail thickness and outliers.

The Normal Q_Q plots panel may show thin and/or thick tails and other departures from a robust fit line.

The boxplot panel shows the median, 1st and 3rd quartile interval, adjacent values interval and outliers

2. Bivariate scatterplot matrices

Can address overplotting to some extend with binning

Task 1: Look for: functional relationships

Enhance s scatterplots with smoothes

Task 2: look for density patterns

Views:

2.1 Kernel density estimates

Views contours and surfaces

Look for: Outliers, clusters, modes, tails

2.2 Binning plot

Views: gray scale and scagnostics

Use scagnostics to prioritize scatterplots for review.

With p variables number of variables pairs is p*(p-1)/2. This grow white the square of p. Pragmatically we can only decode and interpret a modest number of scatterplots.

Sensors and processed results have extended our vision. Algorithms can scan virtual plots and guides us to view those of greatest potential interest to us. While this methodology has not made it in the main stream it may nonetheless be useful.

Task 3: Consider variable reduction

We have limited ability to integrate information across plots. The judicious computation of indices and principal components as supervise principal component may help us see more.

Task 4: Assess the function domain for models

Look for: prediction regions not well supported by data and high leverage point

Augmented views include: Convex hull and alpha shapes

Task 5. Stratification and/or conditioned views

Scales of measurement
Creating summary indicator variables
Creating 0 or 1 variables
Creating sometimes 0 variables for regression.
Regression versus classification
Prediction

#

Scatterplots di