Model exploration via conditional visualisation

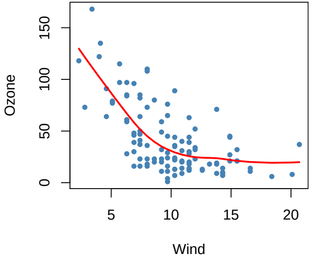
Catherine Hurley
Maynooth University Ireland
joint with Mark O' Connell, Katarina Domijan

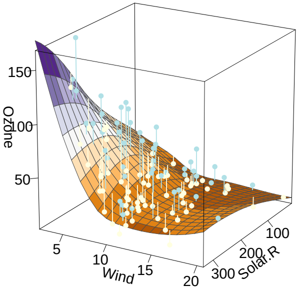
May 17 2018

Model exploration– why?

- See model in action, students and analysts
- Understanding black box behaviour
- Exploring lack of fit
- Compare fits
- Build better models

Conditional model visualisation, beyond 3d?





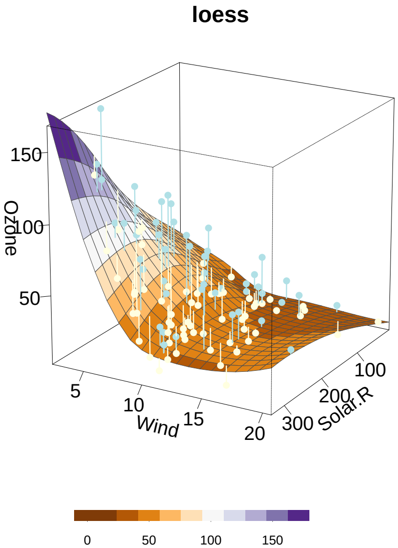
- Our approach: reduce dimensionality by conditioning

Outline

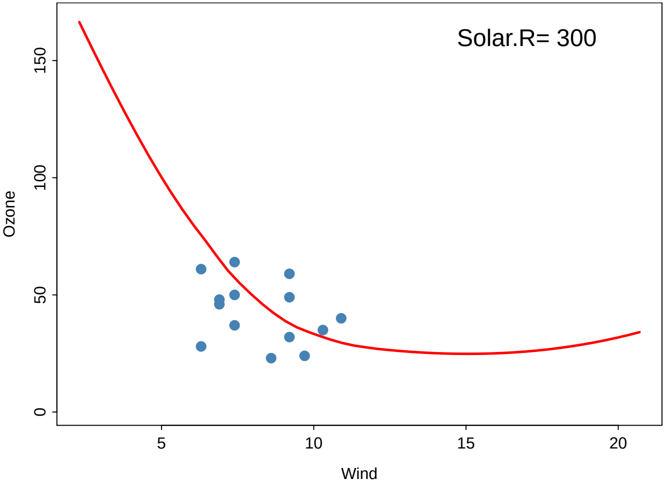
- Introductory example: Air quality data
- Condvis shiny app
- Example: Salary data
- Condtour: Animated tours of predictor space
- Case study: Glaucoma data

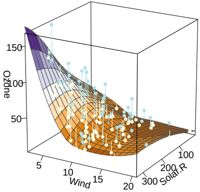
Introductory example: Air quality data

```
f2 <- loess(Ozone~Solar.R+Wind, data=airquality)
```



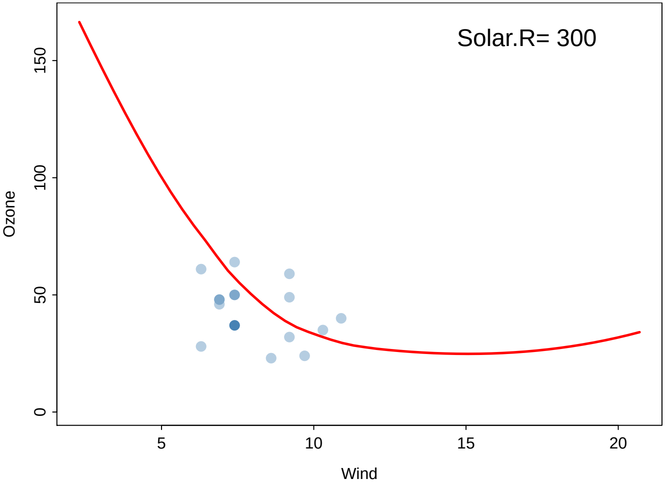
Ozone v Wind,condition on Solar.R = 300

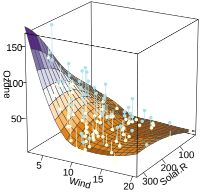




Show (Ozone, Wind) points with Solar.R $\approx$ 300

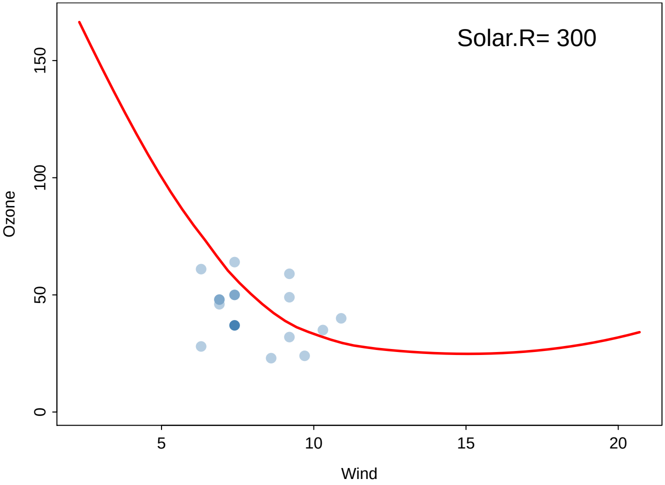
Ozone v Wind, condition on Solar.R = 300

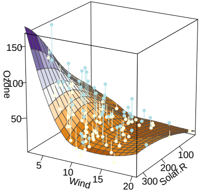




Fade (Ozone, Wind) points by distance from Solar.R $\approx$ 300

Ozone v Wind, condition on Solar.R animation





Fade (Ozone, Wind) points by distance from selected Solar.R value

Condvis setup

- response y
- fit f
- p predictors, say x_1, x_2, x_3, x_4
- one (or two) section predictors, say x_1
- remainder are conditioning predictors, here x_2, x_3, x_4

Condvis setup

- response y
- fit f
- p predictors, say x_1, x_2, x_3, x_4
- one (or two) section predictors, say x_1
- remainder are conditioning predictors, here x_2, x_3, x_4

- set $x_2 = u_2, x_3 = u_3, x_4 = u_4$
- let x_1^r be a sequence covering range of x_1
- plot $f(x_1^r, u_2, u_3, u_4)$ versus x_1^r
- superimpose points (y, x_1) whose (x_2, x_3, x_4) values are near (u_2, u_3, u_4)

Condvis setup

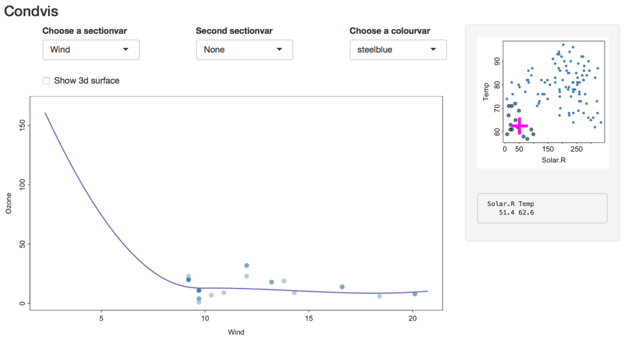
- response y
- fit f
- p predictors, say x_1, x_2, x_3, x_4
- one (or two) section predictors, say x_1
- remainder are conditioning predictors, here x_2, x_3, x_4

- set $x_2 = u_2, x_3 = u_3, x_4 = u_4$
- let x_1^r be a sequence covering range of x_1
- plot $f(x_1^r, u_2, u_3, u_4)$ versus x_1^r
- superimpose points (y, x_1) whose (x_2, x_3, x_4) values are near (u_2, u_3, u_4)

- modify (u_2, u_3, u_4) and watch plot change

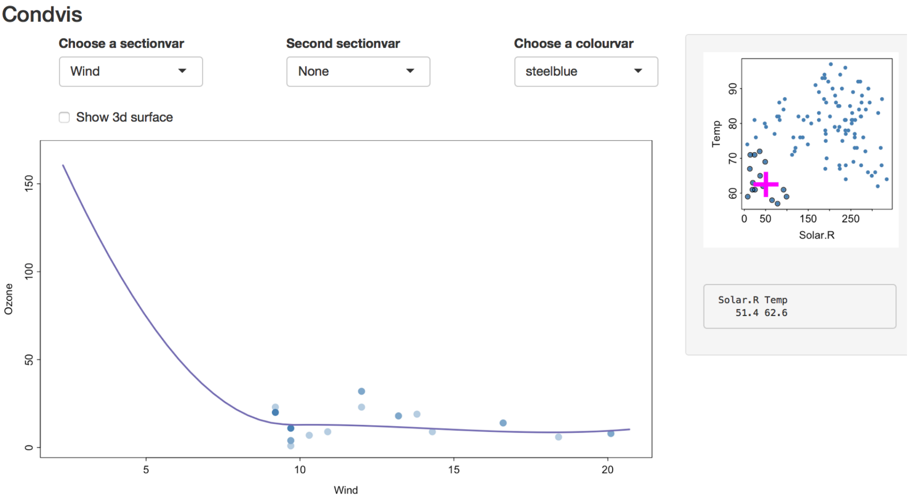
Condvis shiny app

```
f3 <- loess(Ozone~Solar.R+Wind+ Temp, data=airquality)
condvis(ozone, f3, sectionvar="Wind")
```



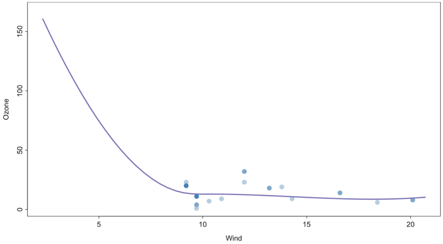
- Main panel shows a plot of fit on the section, with superimposed points
- Right hand panel shows plots of conditioning predictors
- User interacts with conditioning predictors to change their values
- Plot of fit on the section changes to reflect the new condition

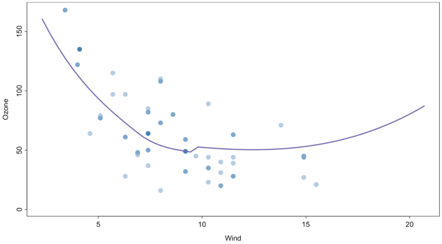
Condvis: Air quality data



Condvis: Air quality data

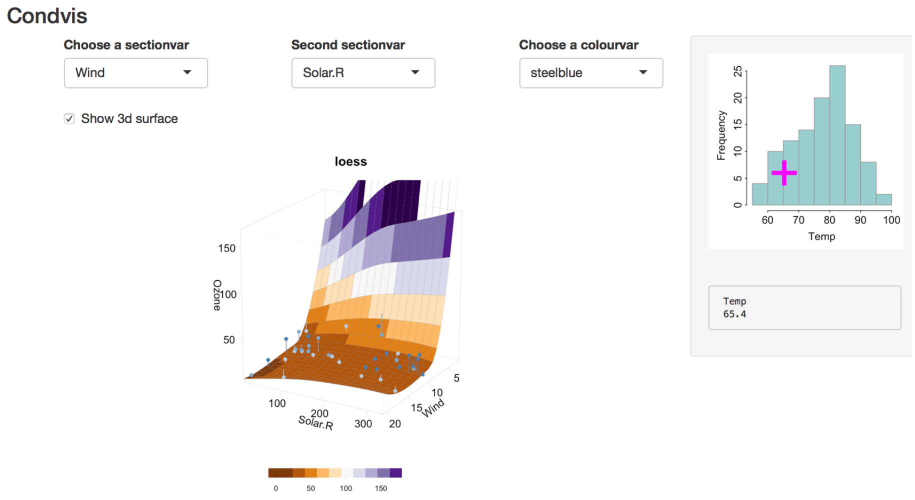
- Fitted relationship between Ozone and Wind depends on values of Solar.R and Temp
- Some regions have little or no data
- Extrapolation





Condvis: Air quality data

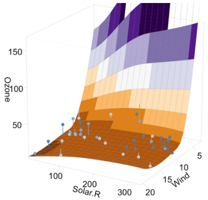
Make Solar.R a sectionvar

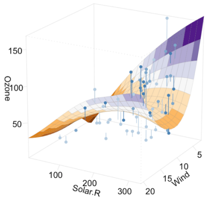


Condvis: Air quality data

Make Solar.R a sectionvar

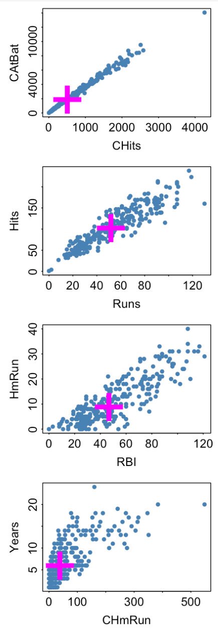
- The fitted relationship between Ozone and (Wind, Solar.R) depends on Temp
- Some regions have little or no data
- Extrapolation





Choosing conditions

- Shows 1d or 2d displays of many condition vars
- Pair predictors with dependence: goal is to avoid selecting empty sections
- Or, use PCP of condition vars, to condition on observations



Distances

condition values: u, observation: x_i

$$\boxed{d_i = d(u, x_i) = d_n(u, x_i) + \lambda M_f(u, x_i)}$$

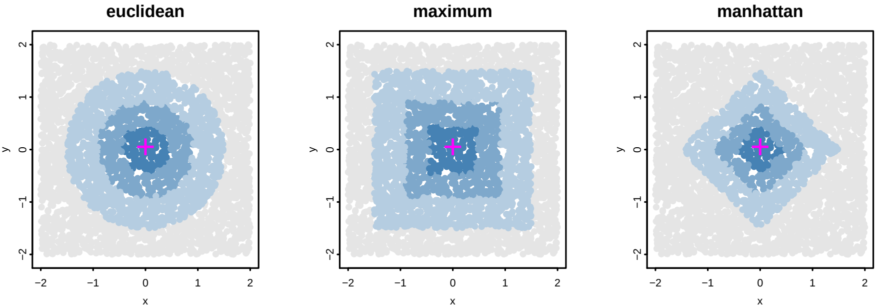
- d_n is the distance between numerical predictors
- M_f counts the number of mismatches between factors
- $\lambda \geq 0$

- Plot points with $d_i \leq \sigma$ (threshold)

$$\boxed{w_i = \max(0, 1 - d_i/\sigma)}$$

- Fade point colour proportional to w_i

Distances



- Choices for d_n
- Threshold $= 1.5$

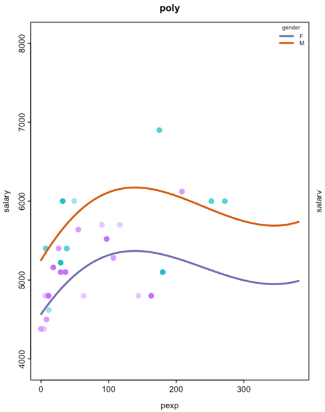
Showing sections

- Section plot types, nn, nf, nnn, nnf etc
- Confidence intervals
- Surfaces
- Multiple fits

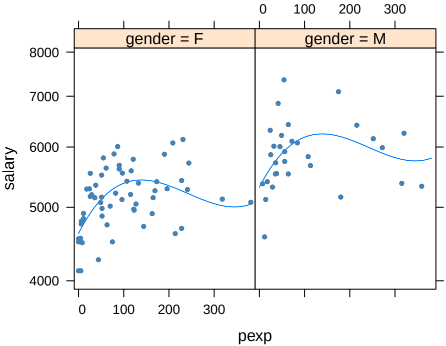
Example: Salary data

```
fit1 <- lm(log(salary) ~ gender+poly(pexp,3)+
           time+educ , data=sal)
condvis(sal, fit1, sectionvars=c("pexp","gender"))
```

Chicago bank discrimination data, 1979



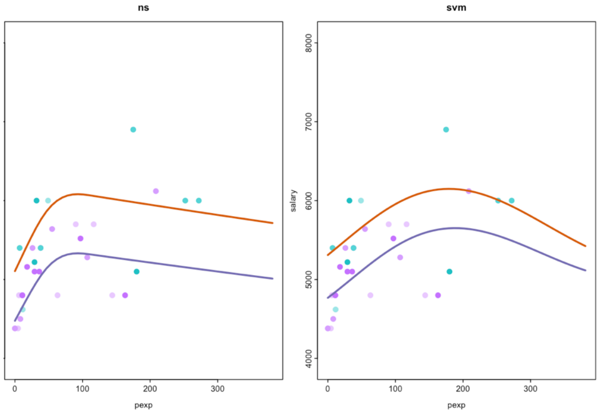
Not to be confused with: partial residual plot



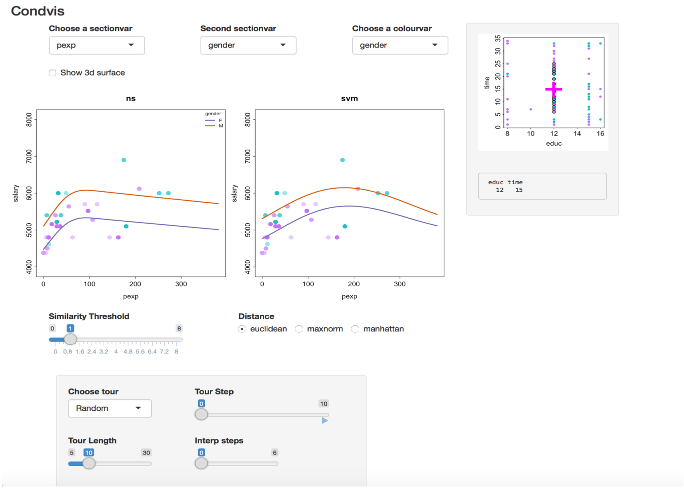
- Produced by effects package (Fox, Weisberg)
- Averages over educ and time
- Requires linear predictors

Example: Salary data, multiple fits

- ns uses a spline term for pexp
- svm is a support vector machine



Comparing fits, ns and svm



- M/F difference less with svm
- svm poor for low pexp

Challenges

Large p

- Use variable importance measures to select predictor subset for visualisation
- About 10 conditioning predictors is max for app
- Other predictors are fixed
- Pair predictors that are dependent
- Empty sections
- Use pre-calculated tours to visit conditioning space
  - Choose random sections
  - Use kmeans or similar and visit cluster centroids
  - Visit sections that have large residuals
  - Or, big disparities amoung fits

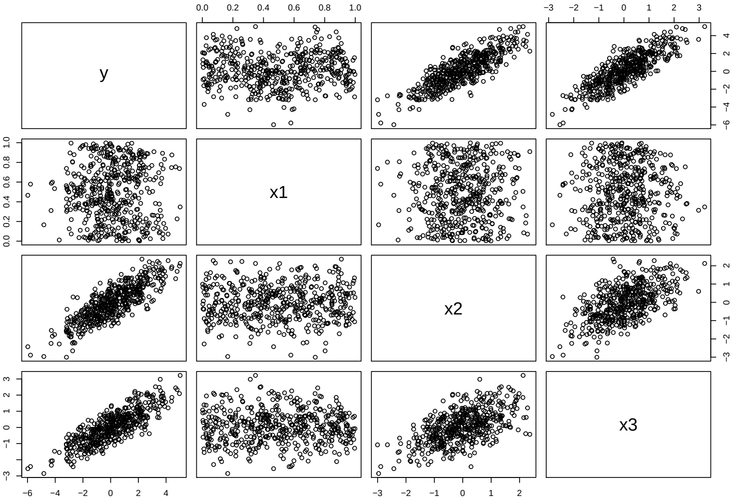
Challenges

Large n

- Use subset or binning in condition plots, for speed
- Examples up to $n = 50K$

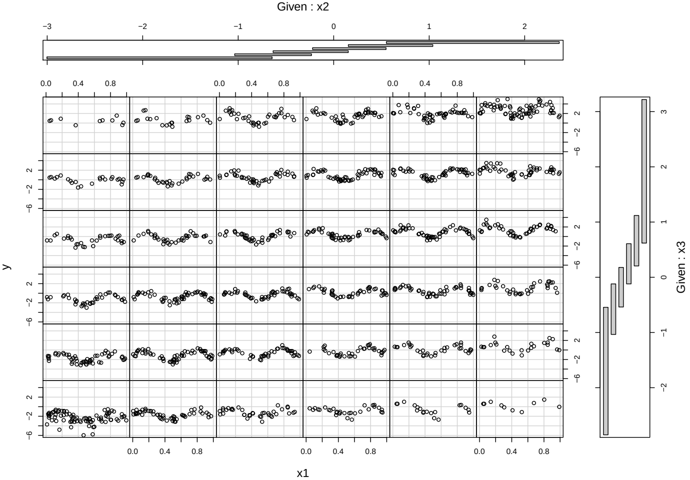
Tour of kmeans centroids

- Artifical data: $y = \sin(x_1) + x_2 + x_3$



Trellis

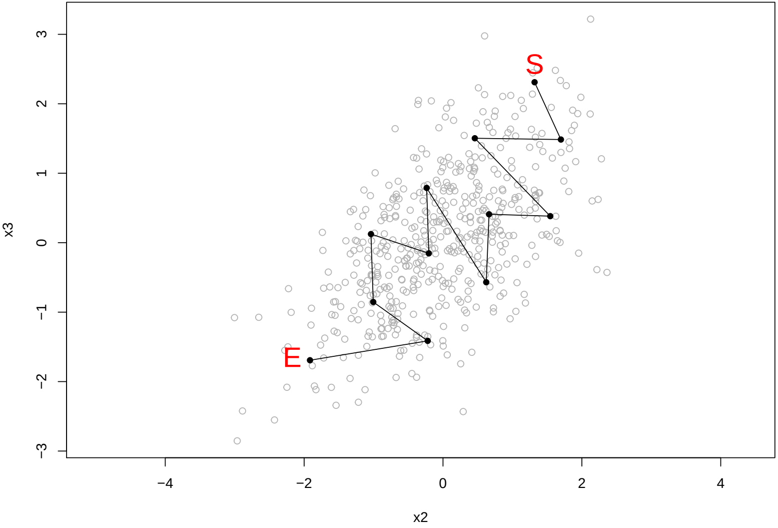
- Visualise dependence of y on x1, conditional on x2 and x3



- Sinusoidal pattern for y vs x1
- some intervals are large, with few points

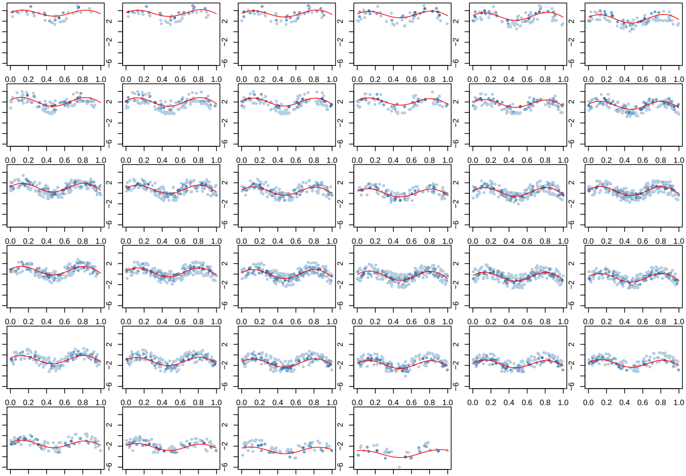
Kmeans of x2, x3

12 centres, ordered to form a path

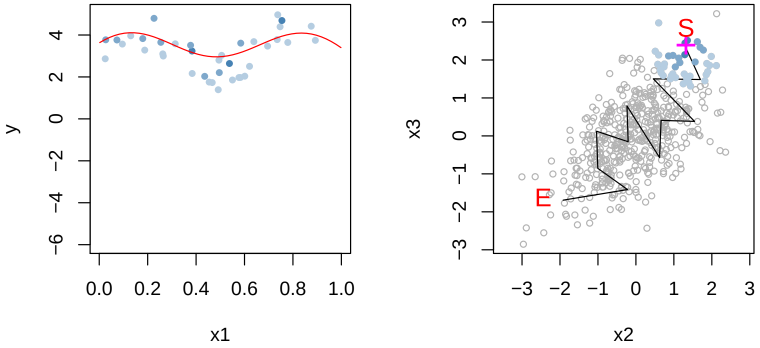


Condition on kmeans path

- pathlength is 34, 12 centres, plus two interpolated points
- fit is svm

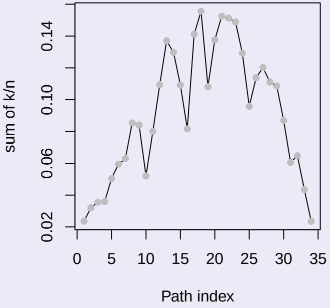


Condition on kmeans path

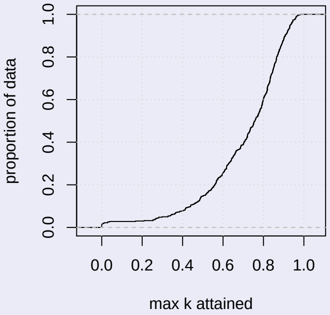


Tour diagnostics

- Shows the similarity weights at each plot on the tour
- Max is about 15% of the data



- cdf of similarity weights
- Very few cases have sim below .2



Glaucoma data

PLOS study: machine learning models for glaucoma diagnosis

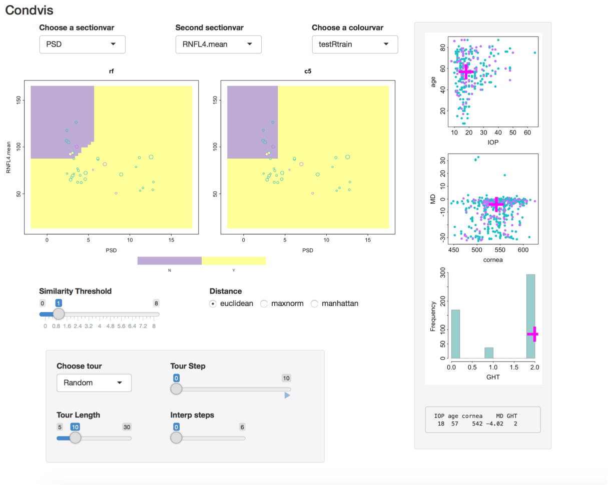
- 399 training and 100 test cases
- 60% of both have glaucoma
- response is glaucoma Y/N (based on optic disc and vis. field)
- predictors
    - age
    - IOP: ocularpressure
    - MD: vis. field measure
    - PSD: vis. field measure
    - GHT: vis. field measure
    - cornea: cornea thickness
    - RNFL4.mean: retinal nerve fiber layer thickness

Models: random forest, tree (C5.0),
svm, knn

Citation: Kim SJ, Cho KJ, Oh S (2017) Development of machine learning models for diagnosis of glaucoma. PLoS ONE 12(5): e0177726. https://doi.org/10.1371/journal.pone.0177726

Glaucoma data,compare rf and c5

- section vars have highest varImp: PSD and RNFL4.mean
- Show both training (green) and test data (pink)
- Point size represents distance (instead of fade)



Glaucoma data

- Where are wrong predictions?
- Reduce threshold to zero: points on section only

Glaucoma data

- Where are wrong predictions?
- Precalculate tour for condvis
- We see
  - Mostly false positive, for c5
  - Mostly at MD ≈ 0
  - low IOP

Concluding remarks

- Condvis is for
  - interactively exploring and comparing model fits
  - assessing if data supports the model
- Condvis works for any fit for which predict method exists or can be provided
- Augment fit with CI for those fits that provide it
- Bayesian fits:
  - plot median of posterior distribution of $\mathrm{E}(\mathrm{y}|\mathrm{x})$
  - or, with MCMC, plot median of sample from the posterior
- Conditional visualisation for any display
  - Related to brushing

Concluding remarks

- condvis is on CRAN
- uses base R interactive graphics/shiny
- new shiny only front-end in progress
- Extensions: nested predictors
- Reference: O'Connell et al. "Conditional Visualization for Statistical Models: An Introduction to the condvis Package in R". JSS 2017

catherine.hurley@mu.ie