Data-Splitting Models for O3 Data

Q. Yu, S. N. MacEachern and M. Peruggia

Abstract

Daily measurements of ozone concentration and eight covariates were recorded in 1976 in the Los Angeles basin (Breiman and Friedman, 1985). We are interested in predicting ozone concentration. These data have been analyzed with sophisticated black box routines such as CART and Smoothing Splines, which are said to provide superior predictive performance. We explore the benefits of subjective (human) modeling by a methodology that avoids double use of the data. The data are split into three portions. Each of three analysts independently models one portion of the data, reporting their posterior distribution. Each posterior consists of several models, with distributions assigned both across models and over parameters within models. Each posterior is then updated with the portions of the data that were not used to construct the posterior. The analysts' posteriors are then synthesized via Bayes' Theorem to produce a final posterior distribution. The predictive abilities of the various modeling strategies are evaluated by their performance on test data. Subjective modeling is superior to automatic modeling. Combining predictions across analysts is superior to relying on a single analyst.

Data Splitting

- Motivation
  - Bayesian Model Averaging
  - Different Analysts Produce Vastly Different Models
  - Avoids Multiple Use of the Data
  - Reduces the Impact of Outliers

- Methods:
  - Split Data into Pieces
  - One Piece, One Analyst
  - Combine Analyses Using Bayes Theorem

Application to the Ozone Data

- Data
  - Daily measurements of ozone concentration and eight other variables in the Los Angeles basin for 330 days in 1976.
  - Data split into three sets at random, each with 110 observations.
  - Three analysts, each got one part of the data.
  - All three analysts used log(ozone) as the response.
- Goal: To provide a predictive model for ozone concentration.

Analyst 1

- EDA used to select and create predictors
- Non-linear predictors created
  - ibht (inversion base height)
    $(1200 - \mathrm{ibht})^2 \times I(\mathrm{ibht} < 1200)$
  - hmdt (humidity)
    $(45 - \mathrm{hmdt}) \times I(\mathrm{hmdt} < 45)$
- Variables coarsened to create indicators
  - vsty (visibility) 7 categories
  - $I(\mathrm{ibht} = 5000)$

Analyst 1 (Cont'd)

- Structured prior for indicators
  - Forces similar effects for similar values of vsty
- Periodic with 1 year period
  - $\sin(\mathrm{day.rad} + 8\pi/12)$
- Residual variance depends on predictors
- Posterior includes 8 models
- Equal weight to each model

Analyst 2

- Detrend response and predictors using LOESS
- Models built on residuals of detrended data
- Response is normal with mean $\mu_t = \beta X_t + \theta_t$
- Two distinct models for βX_t
  - Detrending creates new predictors
  - 1st model built subjectively with graphical methods has 4 main effects
  - 2nd model built using stepwise selection adds many predictors, including interactions

Analyst 2 (Cont'd)

- Conditional Autoregressive (CAR) structure on θ_t
- First and second order neighborhoods for θ_t
- Vague priors on model parameters
- Posterior includes 4 models
  - 2 regressions $\times$ 2 neighborhood structures
- Weights assigned to models subjectively
  - Simpler regression models and simpler CAR structures get more weight

Analyst 3

- Modification of Least Angle Regression (LARS) to select variables
  - LARS selects variables according to their correlation coefficients with current residuals
  - Modification 1: can choose only hierarchical models
  - Modification 2: force some variables to enter the model first
  - With different forced-in variables, different models were created

Analyst 3 (Cont'd)

- Only one main effect was forced into the model each time
- AIC, BIC, Cp were jointly considered to select models
- Bayesian linear regression models on selected variables
- Vague priors used for parameters
- Posterior over four models
- BIC used to choose weights for models

Model Comparison

- Sum of squared errors for log(ozone)
- Sum of absolute errors for log(ozone)
- Human models
  - Developed on one portion of the data
  - Updated with the second portion
  - Predictions for the third portion
  - Combined using Bayes theorem
- Automatic Methods
  - Fit on two thirds of the data
  - Predictions for the remaining portion

Comparison of Automatically Fitted Models vs. Human Models by Sum of Squared Errors for Log Ozone

Testing Data	data 1		data 2		data 3	
Updating Method	Once	10 by 10	Once	10 by 10	Once	10 by 10
ANALYST 1	-	-	12.76	12.72	14.96	14.25
ANALYST 2	18.00	17.48	-	-	12.13	12.03
ANALYST 3	15.96	16.07	14.21	14.32	-	-
MEAN HUMAN	16.98	16.78	13.49	13.52	13.55	13.14
COMBINED HUMAN	16.00	16.31	12.50	13.11	12.10	11.89
CART	27.51	28.43	17.87	17.01	19.37	19.51
BAGGING	19.63	19.16	14.94	14.19	16.28	15.51
SMOOTHING SPLINE	26.85	26.39	17.60	16.73	18.01	15.96

Best models in red, second best in blue

Comparison of Automatically Fitted Models with Human Models by Sum of Absolute Errors for Log Ozone

Testing Data	data 1		data 2		data 3	
Updating Method	Once	10 by 10	Once	10 by 10	Once	10 by 10
ANALYST 1	-	-	27.98	28.87	32.46	31.19
ANALYST 2	35.71	35.23	-	-	28.73	28.77
ANALYST 3	33.17	32.96	29.34	29.30	-	-
MEAN HUMAN	34.44	34.10	28.66	29.09	30.60	29.98
COMBINED HUMAN	33.62	33.29	27.16	29.23	28.87	27.95
CART	42.56	43.11	31.10	34.31	36.20	36.15
BAGGING	36.11	34.83	31.59	30.93	34.00	32.59
SMOOTHING SPLINE	43.25	42.42	33.25	32.41	34.89	32.17

Best models in red, second best in blue