Statistical Methods - Model evaluation, comparison and selection

Lecture 11

Mattias Villani

Sveriges Riksbank and Stockholm University

June 1, 2010

Overview Lecture 11

- Automatic Bayesian knot selection in spline regression.
- Model evaluation

Automatic Bayesian knot selection in spline regression

- Selecting the knots in a spline regression is exactly variable/covariate selection in linear regression.
- Introduce variable selection indicators, I_j such that $I_j = 0 \iff \beta_j = 0$ and

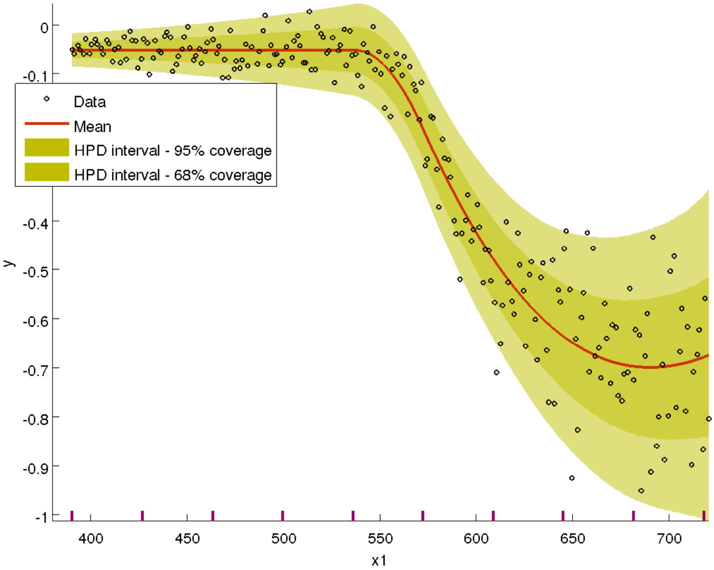
$$\beta_j \sim N(0, \lambda^{-2}) \text{ if } I_j = 1.$$

- Need a prior on $I_1, ..., I_K$. Simple choice: $I_1, ..., I_K \overset{iid}{\sim} Bernoulli(\theta)$.
- Simulate from the posterior distribution:

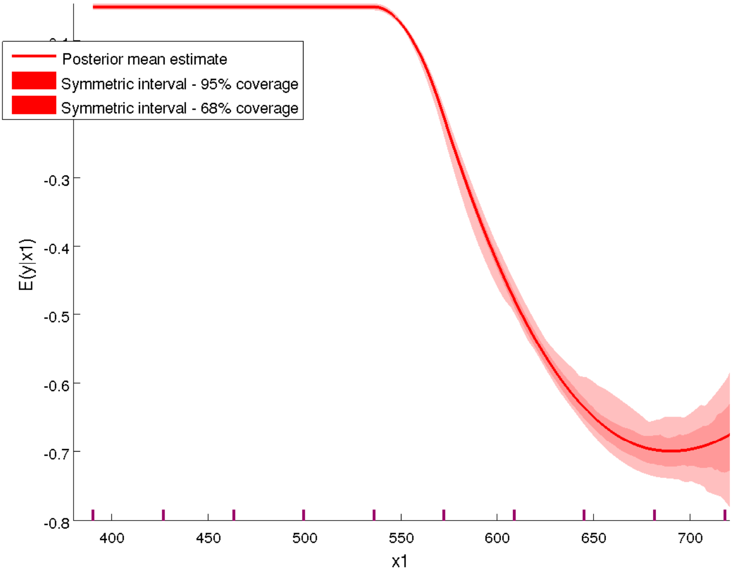
$$p(\beta, I | y, x) = p(\beta | I, y, x) p(I | y, x).$$

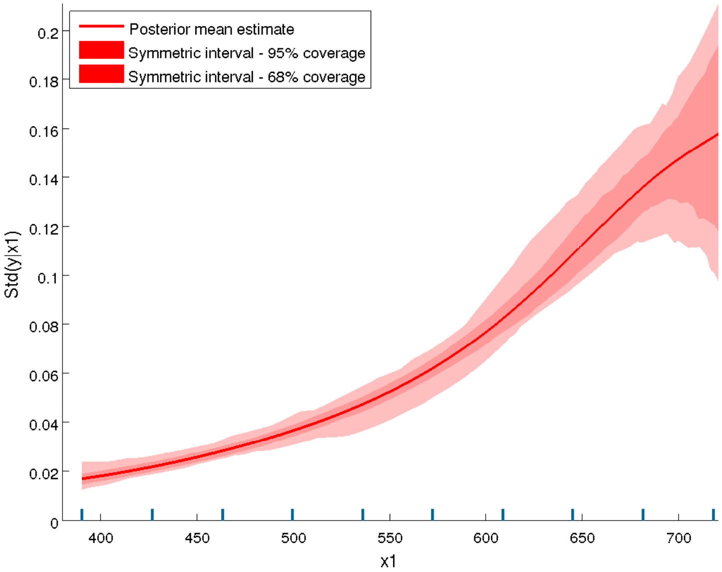
- Automatic model averaging, all in one simulation run.
- Can be generalized to non-linear models, GLMs and more.

Bayesian knot selection - Lidar example



Bayesian knot selection - Lidar example





Posterior mean estimate
Symmetric interval - 95% coverage
Symmetric interval - 68% coverage
Std(y|x1)
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
0.18
0.2
400
450
500
550
600
650
700
x1

Models - why?

- We now know how to **compare** models.
- But how do we know if any given model is 'any good'?
- George Box: 'All models are false, but some are useful'.
- What is the purpose of the model:
  - Prediction. Interpretation may be of lesser concern. Black-box approach may be sufficient. Extrapolation?
  - Abstraction to aid in thinking about a phenomena. Prediction accuracy may be of lesser concern (?).
  - Compact description of complex phenomena. Computational cost of model evaluation may be a concern.

How to evaluate a model?

- Out-of-sample predictions.
  - Evaluation to point forecasts (RMSE, MAE etc)
  - Evaluation of interval forecasts (Interval coverage probability. 'QQ-plot'.)
  - Evaluation of density forecasts. (Log Predictive Density Score, normalized forecast errors etc)
- Simulate data from an estimated model.
  - General idea: if $p(y|\theta)$ is a 'good' model, then the data actually observed should not differ 'too much' from simulated data from $p(y|\hat{\theta})$.
  - Weak test, but often useful: Many, many models cannot even fit the data when their parameters are estimated on the same data set.
  - But simulated data from the model with an $\bar{\theta} \neq \hat{\theta}$ may be very different from those obtained with $\theta = \hat{\theta}$, even when $\bar{\theta}$ is a 'likely' value.

Posterior predictive analysis

- Bayesian solution: simulate data from many different from posterior predictive distribution:

$$p(y^{rep}|y) = \int p(y^{rep}|\theta)p(\theta|y)d\theta.$$

- *Posterior predictive density* is a weighted likelihood function, using the posterior density as a weighing function.
- Difficult to compare y and y^{rep} because of dimensionality. Solution: compare low-dimensional 'test' statistic $T(y, \theta)$ to $T(y^{rep}, \theta)$. Not a problem for a Bayesian that the statistic depends on θ.
- A posterior predictive analysis evaluates the full probability model consisting of both the likelihood *and* prior distribution.

- Algorithm for simulating from the posterior predictive density $p[T(y^{rep})|y]$:

1 Draw a $\theta^{(1)}$ from the posterior $p(\theta|y)$.

  2 Simulate a data-replicate $y^{(1)}$ from $p(y^{rep}|\theta^{(1)})$.
  3 Compute $T(y^{(1)})$.
  4 Repeat steps 1-3 a large number of times to obtain a sample from $T(y^{rep})$.

- We may now compare the observed statistic $T(y)$ with the distribution of $T(y^{rep})$. Ex. $\Pr[T(y^{rep}) \geq T(y)]$ (*Posterior predictive p-value*) or informal graphical analysis.

- Ex. 1. Model: $y_1, ..., y_n \overset{iid}{\sim} N(\mu, \sigma^2)$. $T(y) = \max_i |y_i|$.
- Ex. 2. Assumption of no reciprocity in social networks/statistical graphs. Model: $y_{ij} = 1|\theta \overset{iid}{\sim} Bernoulli(\theta)$. $T(y) =$proportion of reciprocated node pairs.
- Ex. 3. ARIMA-process. $T(y)$ may be the autocorrelation function.
- Ex. 4. Poisson regression. $T(y)$ frequency distribution of the response counts. Proportions of zero counts.

Posterior predictive analysis - Normal model, max statistic