Picking regressors

Christopher A. Sims
Princeton University
sims@princeton.edu

February 3, 2020

A class of problems

- A large class of potential models, possibly infinite-dimensional.
- We want to choose a model that is "simple", or "smooth", yet "performs well".
- Machine learning: Separate "training data" and "test data". Invoke a computer program that generates a model from training data, tests it on test data, improves the model based on those results.
- Frequentist non-parametric inference: Specify an infinite-dimensional parameter space and an estimator that maps data into models determined by the infinite-dimensional parameter. Derive stochastic properties of the estimator under some assumptions, under repeated samples.

- Bayesian non-parametric inference: Specify an infinite-dimensional parameter space and a model that maps parameters into probability distributions for the data. Look at data and apply Bayes rule.

Competition among approaches

- When the data set is large and complex, explicit probability modeling is hard.
- We might hope that the training vs test sample approach will let us assess the performance of algorithms without our having derived the algorithm from a probability model of the data.
- But it is in fact hard to do this reliably. If the test data are systematically used to improve models generated from the training data, the distinction between test and training data is eroded.
- Machine learning on large data sets often aims at performing some prediction problem "well", rather than claiming to have a model that

is in some sense optimal. For some purposes this is a plus, because it allows results for really large data sets, but for some purposes this is also a weakness, especially if we want to understand how the model works or interpret its structure.

Competition among approaches

- The most straightforward Bayesian approach is more demanding than most frequentist approaches, since it aims at the optimal estimator for the model, rather than starting with a convenient or appealing estimator and deriving its properties.

- But in many applications the fact that Bayesian inference aims at characterizing post-sample uncertainty, rather than at characterizing repeated-sample properties of estimators is central.

- Bayesian approaches can be useful without insisting on optimality.

- The "data science" literature has a growing branch called "ABC", for "Approximate Bayesian Computation".

A particular problem, posed by Xavier Sala-I-Martin

- We have data on output growth for N countries and also on M variables that might be predictive of income growth.
- N and M are the same order of magnitude.
- We proceed as if there were a true linear regression model based on a subset of the M predictive variables, and use a prior reflecting this.
- However, we assume that there will be many of the 2^M possible models that perform nearly as well as the best.

A fully Bayesian approach

- Fernández, Ley, and Steel (2001)
- Model i: $y = X\beta_i + \varepsilon$.
- Conjugate prior on non-zero elements of β_i, and in particular a "g-prior".
- $\beta \sim N(0, \sigma^2(gX'X)^{-1})$.
- The "explained sum of squares" in the regression model is $\beta' X' X \beta$. The g-prior implies an a priori expectation that the explained sum of squares will be k/g, where k is the number of non-zero elements in β.

Prior across models

- FLS put equal prior probability on every model — i.e. on every possible subset of the M variable indexes.
- This seems not to be favoring small models?
- In fact, the g-prior does favor smaller models, because it "expects" a higher R^2 for bigger models.
- If g is held fixed as sample size increases, this prior asymptotically favors smaller models by more than the BIC, and the prior fails to be dominated by the likelihood.

- It might make sense to favor smaller models further, by making the prior probability of a model proportional to p^k, where $p \in (0, 1)$ and k is the number of variables in the model.

Marginal data density for one model

FLS give the formula

$$\left(\frac{g}{1+g}\right)^{k/2}\left(\frac{\hat{u}'\hat{u}}{1+g}+\frac{gy'y}{1+g}\right)^{-(n-1)/2},$$

where $\hat{u}$ is the vector of OLS residuals and y is the dependent variable vector, k is the number of variables in the regression, and n is sample size.

FLS say that σ^2 and the constant term are "common" to all the models. This does not mean, though, that there is a single σ^2 and a single constant that applies to all models. There is a separate σ^2 and a separate constant for each model, but the same (flat) prior is applied to these parameters in each model.

Prior on constant

FLS say they used a flat prior on the constant. They implemented this, probably, by using all data as deviations from sample means. So in the formula for individual model mdd on the previous slide, the "y" vector is actually the deviation of the growth rate variable from its sample mean.

σ^2 and constant

- Flat prior on constant, $\log \sigma^2$.
- When comparing models, usually priors are important and results with "flat" priors are nonsense.
- However here σ^2 and the constant are common to all the models, so in model comparison the prior on these parameters cancels out, in a sense.
- The shape of the prior on these parameters does affect estimates, but the usual problem with flat priors — that the level of the pdf is arbitrary, and affects model comparisons — is not present.

Implementing MCMC

- Since each model is a standard normal linear model (SNLM), for which we can calculate the Bayes factor (integrated posterior density, or marginal data density (mdd)) analytically, we can make draws directly from the marginal posterior on "model number".

- To implement a random walk Metropolis MCMC, we need a rule for jumping between models that is symmetric — so the jump distribution $p(j \mid i)$ satisfies $p(j \mid i) = p(i \mid j)$.

- The jump rule FLS use is: Pick a variable number from a uniform distribution over variable numbers. If that variable is in the current model, remove it, otherwise add it, to generate the new model.

- The number of possible models is 2^M, where M is the number of variables. The MCMC chain will not visit all of them. It nonetheless is likely to visit all of those with high posterior probability.

*

References

Fernández, C., E. Ley, and M. F. J. Steel (2001): "Model Uncertainty In Cross-Country Growth Regressions," *Journal of Applied Econometrics*, 16(5), 563–576.