Complex Model for Complex Data via the Bayesian Approach

Feng Li

Central University of Finance and Economics, Beijing, China

A brief review of statistical learning

Statistical methods have been developed rapidly in the past twenty years. One driving factor of this development is that more and more complicated high-dimensional data require sophisticated data analysis methods. A noticeably successful case is the machine learning field which is now wildly used in industry. Another reason are the dramatic advancements in the statistical computational environment. Computationally intensive methods that in the past could only be run on expensive super computers are now possible to run on a standard PC. This has created an enormous momentum for Bayesian analysis where complex models are typically analyzed with modern computer-intensive simulation methods.

Traditional linear models with Gaussian assumptions are challenged by the new large complicated datasets, which have in turn generated interests in new approaches with flexible model with less restrictive assumptions. Moreover, research has shifted the attention from merely modeling the mean and variance of the data to sophisticated modeling of skewness, tail-dependence and outliers. However such work demands efficient inference tools. The development of highly efficient Markov chain Monte Carlo (MCMC) methods has reduced the barrier. Moreover, the Bayesian approach provides a natural way for prediction, model comparison and evaluation of complicated models, and has the additional advantage of being intimately connected with decision making.

The Bayesian density estimation

In statistics, density estimation is the procedure of estimating an unknown density p(y) from observed data. The very early stage of density estimation techniques traces back to the usage of histograms, later followed by kernel density estimation in which the shape of the data is approximated through a kernel function with a smoothing parameter. However due to the difficulty in specifying the bandwidth in kernel density estimation, mixture models have become a popular alternative approach. A mixture density is a combination of different densities with different weights. Mixture densities can be used to capture data characteristics such as multi-modality, fat tails, and skewness.

[Fig-1 about here]

Using mixture of normal densities (thin lines) to mimic a flexible density (bold line)

The conditional density estimation concentrates on modeling the relationship between a response *y* and set of covariates *x* through a conditional density function p(y|x). In the simplest case, the Gaussian linear regression $y = x' \beta + \varepsilon$ is a trivially equivalent to modeling p(y|x) by a Gaussian density with mean function $\mu = x' \beta$ and constant variance.

In Bayesian statistics, inference of an unknown quantity θ, say p(θ|y), combines data information y, p(y|θ) with prior beliefs about θ, p(θ). In many simple statistical models with vague priors that play a minimal role in the posterior distribution, Bayesian inference draws similar conclusions to those obtained from a traditional frequentist approach. The Bayesian approach is however more easily extended to more complicated models using MCMC simulation techniques. In principle, MCMC can be applied to many hard to estimate models. However the efficiency heavily depends on how efficient the MCMC algorithm is. This is especially true in nonlineaer models with many correlated parameters.

A key factor for evaluating a method's performance is to check how it balances the trade-off of goodness-of-fit and overfitting. It is common that a model wins in goodness-of-fit but fails in prediction. Variable selection is a technique that is commonly used in such context. Historically the purposes for using variable selection are to select meaningful covariates that contributes to the model, inhibit ill-behaved design matrices, and to prevent model overfitting. Methods like backward and forward selections are standard routines in most statistical software packages. However the drawbacks are obvious in those techniques, e.g. the selection depends heavily on the starting points, which becomes more problematic with high dimensional data with many covariates. Most current methods rely on Bayesian variable selection via MCMC. A standard Bayesian variable selection approach is to augment the regression model with a variable selection indicator for each covariate. For the purpose of overcoming problems with overfitting, shrinkage estimation can also be used as an alternative, or even complementary, approach to variable selection. A shrinkage estimator shrinks the regression coefficients towards zero rather than eliminating the covariate completely. One way to select a proper value of the shrinkage is by cross-validation.

Bayesian models for complex data

Modeling the volatility and variability in financial data has been a highly active research area since the seminal paper by Engle introduced the ARCH model, and there are large financial markets for volatility-based instruments. Financial data, such as stock market returns, are typically heavy tailed and subject to volatility clustering, i.e. time-varying variance, skewness and kurtosis that evolve in a very persistent fashion or financial crisis with an unprecedented volatility. Bayesian modeling such data requires sophisticated MCMC treatment, but in return, we obtain more insights of the problem through the model where other methods can hardly tackle.

[Fig-2 about here]

Time series plot of the posterior median and 95% probability intervals for kurtosis in terms of degrees of freedom of the return distribution for S&P 500 stock returns.

In physics, a type of real dataset comes from a technique that uses laser-emitted light to detect chemical compounds in the atmosphere (LIDAR, Light Detection And Ranging). The response variable (logratio) consists of 221 observations on the log ratio of received light from two laser sources: one at the resonance frequency of the target compound, and the other from a frequency off this target frequency. The predictor is the distance traveled before the light is reflected back to its source (range).

Our aim is to model the predictive density p(logratio | range). A smooth mixture of asymmetric densities is used to model such predictive density which involves in a large number of parameters. It is therefore likely to over-fit the data unless model complexity is controlled effectively. Bayesian variable selection in all parameters can lead to important simplifications of the mixture components. Not only does this control complexity for a given number of components, but it also simplifies the existing components if an additional component is added to the model.

[Fig-3 about here]

Smooth mixture models for the LIDAR data. The figure displays the actual data overlayed on predictive regions and the predictive mean.

In finance applications, a firm's leverage (fraction of external financing) is usually modeled as a function of the proportion of fixed assets, the firm's market value in relation to its book value, firm sales and profits. The relationships between leverage and the covariates are highly nonlinear. There are also outliers. Strong nonlinearities seem to be a quite general feature of balance sheet data, but only a handful articles have suggested using nonlinear/nonparametric models. One attempt is to extend the regression model by introducing a lot of auxiliary variables, aka *splines.* A nonlinear curve/surface can then be constructed by choosing the correct number of splines and placing them in the right covarite space. Nonetheless, correctly allocating the splines in covarite space is not trivial. Bayesian methods treat the locations as unknown parameters that efficiently allocate the splines and therefore keep the number of splines to be a minimum. Compared with traditional deterministic spline approach, the Bayesian approach allows the splines to move freely in the covariate space and provides a dynamic surface with the measurement of surface uncertainty.

[Fig-4.1, Fig-4.2 about here]

The posterior mean (left) and standard deviation (right) of the posterior surface for the model for firm leverage data. The subplot to the right also shows an overlay of the covariate observations.

A model bigger than an elephant?

In the 1950s, linear regression model that was considered as very advanced is now the standard course content for university students. The data are much more complicated nowadays not only because the volume increases but also the structure is much more complicated. Very high-dimensional data that are mixing with numeric, character strings, images or videos are not rare anymore. Sophisticated models are essential for such situation. In principle, the complicated model should be able to capture more complicated data features but estimating and interpreting such model is not obvious. Personally speaking, there is huge space to explore computationally and statistically. Statistical models that can adapt to modern computational architectures already flourish in industry. Techniques like high performance computing will be more widely used in statistics and will be aware to young statisticians eventually.

(*I would like to thank Professor Mattias Villani that introduced me to this exciting area.*)