Introduction to Bayesian Machine Learning Methods

Bayesian Machine Learning Methods: Overview

- Machine learning is a a very broad topic, involving a range of methods
- Widely used in many statistical and professional disciplines, beginning to be used in economics
- Broadly speaking, it is all about finding patterns in data in an automatic fashion (i.e. via the machine)
- Relates to data mining/artificial intelligence/data science
- We will cover a few machine learning methods which are Bayesian (many exist which are not Bayesian)
- Focus on those we have seen used in economics
- Focus on regression model (but they can be used with other models)

Bayesian Machine Learning Methods: Big Data

- Big Data is hot topic that may revolutionize empirical work and change the way we do econometrics
- "Big" Data may be "tall" or "fat"
- Tall Data = data with many observations
- Fat Data = data with many variables
- In macroeconomics, Fat Data is common
- "Big Data" in this chapter means "Fat Data"

Bayesian Machine Learning Methods Overview

- In this lecture will show some Big Data methods in context of regression, but they also can be used with other models
- To illustrate use classic cross-country growth regression data set (see Lecture 2 on Regression)
- Dependent variable: average growth in GDP per capita from 1960-1992
- $K = 41$ explanatory variables (all normalized by subtracting of mean and dividing by st. dev.)
- But data set has only $N = 72$ countries
- Big Data: large number of explanatory variables relative to number of observations
- In other Big Data applications can have $K > N$ (e.g. stock returns for large K companies observed only for a few months).

Bayesian Machine Learning Methods Overview

- Why not just use conventional methods?
- Intuition:
- N reflects amount of information in the data
- K reflects dimension of things trying to estimate with that data
- If K is large relative to N you are trying to do too much with too little information
- If $K < N$ a method such as least squares will produce numbers, but very imprecise estimation (e.g. wide confidence intervals)
- If $K > N$ least squares will fail
- Bayesian prior information (if you have it), gives you more information to surmount this problem
- E.g. $E(\beta|y)$ using natural conjugate prior will exist even if $K > N$ and $var(\beta|y)$ will be reduced through use of prior information

Bayesian Machine Learning Methods Overview

- Over-fitting: data typically contains measurement error (noise)
- Regression methods seek to find pattern in the data
- With large data sets, often not a problem (things average out over large number of observations)
- But with Fat Data, easy to "fit the noise" rather than pattern in the data
- Good in-sample fit, but bad out-of-sample forecasting

Summary: New Tricks for Econometrics

- Conventional statistical methods (least squares, maximum likelihood, hypothesis testing) do not work
- New methods are called for and many of these are Bayesian
- This lecture discusses two main ones:
- i) Stochastic search variable selection (SSVS)
- ii) Least absolute shrinkage and selection operator (LASSO)

Variable Selection and Shrinkage Using Hierarchical Priors

- Any sort of prior information can be used to overcome lack of data information with Big Data regression
- But what if researcher does not have such prior information?
- Hierarchical priors are a common alternative
- A simple example: g-prior but treat g as unknown parameter with its own prior
- Global-local shrinkage priors are growing in popularity (in many models, not only regression)
- I introduce two popular ones: LASSO and SSVS
- Many others (and not all Bayesian)

SSVS: Overview

- To show main ideas assume (for now) β is a scalar and remember degree of shrinkage controlled by prior variance
- SSVS prior:

$$\beta|\gamma \sim (1-\gamma) N\left(0, \tau_0^2\right) + \gamma N\left(0, \tau_1^2\right)$$

- τ_0 is small and τ_1 is large
- $\gamma = 0$ or 1.
- If $\gamma = 0$, tight prior shrinking coefficient to be near zero
- If $\gamma = 1$, non-informative prior and β estimated in a data- based fashion.
- SSVS treats γ as unknown and estimates it
- Data choose whether to select a variable or omit it (in the sense of shrinking its coefficient to be very near zero).

SSVS: Overview

- prior for β is hierarchical: depends on γ which has its own prior.
- Gibbs sampler takes draw of γ and, conditional on these, results for independent Normal-Gamma prior used to draw β and h.
- If $\gamma = 1$ use $N\left(0, \tau_1^2\right)$ prior, else use $N\left(0, \tau_0^2\right)$
- Output from this GIbbs sampler can be used to:
- Do something similar to BMA: averages over restricted (when $\gamma = 0$ is drawn) and unrestricted $(\gamma = 1)$ models
- Do BMS (variable selection):
- If $\Pr(\gamma = 1|y) > \frac{1}{2}$ choose unrestricted model, else choose restricted model
- Can use threshold other than $\frac{1}{2}$

SSVS in Multiple Regression

- We have posterior results for regression model with prior

$$N\left(\underline{\beta}, \underline{V}\right)$$

- SSVS prior makes specific choices for $\underline{\beta}$ and $\underline{V}$
- $\underline{\beta} = 0$ so as to shrink coefficients towards zero
- $$\underline{V} = DD$$
- D is diagonal matrix with elements

$$d_i = \begin{cases} \tau_{0i} \text{ if } \gamma_i = 0 \\ \tau_{1i} \text{ if } \gamma_i = 1 \end{cases}$$

- We now have $i = 1, .., K$
- $\gamma_i \in \{0, 1\}$ indicating whether each variable is excluded
- Small/large prior variances, τ_{0i}^2 and τ_{1i}^2, for each variable

SSVS: Gibbs Sampler

- Conditional on draw of γ we are in familiar world
- Use independent Normal-Gamma posterior for β and h
- What about γ?
- Needs a prior
- A simple choice is:

$$\begin{array}{l} \Pr\left(\gamma_i = 1\right) = \underline{q}_i \\ \Pr\left(\gamma_i = 0\right) = 1 - \underline{q}_i \end{array}$$

- Non-informative choice is $\underline{q}_i = \frac{1}{2}$ (each coefficient is *a priori* equally likely to be included as excluded)

SSVS: Gibbs Sampler

- Can show conditional posterior distribution is Bernoulli:

$$\Pr\left(\gamma_i = 1 | y, \gamma\right) = \overline{q}_i,$$
$$\Pr\left(\gamma_i = 0 | y, \gamma\right) = 1 - \overline{q}_j,$$

- where

$$\overline{q}_j = \frac{\frac{1}{\tau_{1j}} \exp\left(-\frac{\gamma_j^2}{2\tau_{1j}^2}\right) \underline{q}_j}{\frac{1}{\tau_{1j}} \exp\left(-\frac{\gamma_j^2}{2\tau_{1j}^2}\right) \underline{q}_j + \frac{1}{\tau_{0j}} \exp\left(-\frac{\gamma_j^2}{2\tau_{0j}^2}\right) \left(1 - \underline{q}_j\right)}.$$

SSVS: Choosing Small and Large Prior Variances

- Researcher must choose τ_{0i}^2 and τ_{1i}^2
- Want τ_{0i}^2 to imply virtually all of prior probability is attached to region where β_i is so small as to be negligible
- Approximate rule of thumb: 95% of the probability of a distribution lies within two standard deviations from its mean.
- E.g. is $\tau_{0i} = 0.01$ small?
- Expresses a prior belief that β_i is less than 0.02 in absolute value.
- Is $\beta_i = 0.02$ a "small" value or not?
- Depends on empirical application at hand and units dependent and explanatory variables are measured in
- Sometimes researcher can subjectively make good choices for τ_{0i}
- But often not, want a method of choosing them that does not require (much) prior input from researcher

SSVS: Choosing Small and Large Prior Variances

- Common to use “default semi-automatic approach”
- Choose τ_{0i}^2 and τ_{1i}^2 based on initial estimation procedure.
- Use initial estimates (e.g. OLS) from regression with all exp vars:
- produce $\widehat{\sigma}_i$ – the standard error of β_i.
- Set $\tau_{0i} = \frac{1}{c} \times \widehat{\sigma}_i$ and $\tau_{1i} = c \times \widehat{\sigma}_i$ for large value for c (e.g. $c = 10$ or 100).
- Basic idea: $\widehat{\sigma}_i$ is estimate of the standard deviation of β_i
- Question: how do we choose small value for prior variance of β_i?
- Answer: choose one which is small relative to its standard deviation

SSVS Application

- Use cross-country growth data set.
- Default semi-automatic prior elicitation approach with $c = 10$.
- $110,000$ draws of which first $10,000$ are discarded as the burn-in.
- Single Best Model results use SSVS but with γ_i not drawn, but fixed
- Set $\gamma_i = 1$ if $\Pr(\gamma_i = 1|y) > \frac{1}{2}$ and set $\gamma_i = 0$ otherwise.
- $\Pr(\gamma_i = 1|y)$ obtained using an initial run of MCMC algorithm.

SSVS Application

- Following tables show SSVS results similar to BMA results
- Similar estimates and standard deviations for β.
- Variable selection results also show high degree of similarity.
- SSVS is selecting 11 variables which is slightly more parsimonious than the 14 selected by BMS.
- Note: in Single Best Model results posterior means of variables not selected very near to zero and st devs very small
- Default semi-automatic approach's "small" prior variance is shrinking to zero
- Note: variable selection (which ignores model uncertainty) leads to estimates which are usually larger in absolute value and are more precise

SSVS Point Estimates and Standard Devs of Regression Coefficients (Mean and standard deviations multiplied by 100)					
	SSVS			Single Best Model	
Explanatory Variable	$\Pr(\gamma = 1 \mid y)$	Mean	St. Dev.	Mean	St. Dev.
Primary School Enrolment	0.256	0.111	0.204	2×10^{-5}	0.002
Life expectancy	0.956	0.991	0.365	1.124	0.236
GDP level in 1960	1.000	-1.410	0.286	-1.299	0.202
Fraction GDP in Mining	0.664	0.204	0.179	0.258	0.107
Degree of Capitalism	0.575	0.170	0.176	0.240	0.108
No. Years Open Economy	0.553	0.248	0.267	0.459	0.141
% Pop. Speaking English	0.171	-0.024	0.071	-2×10^{-5}	0.001
% Pop. Speak. For. Lang.	0.174	0.024	0.086	7×10^{-6}	0.001
Exchange Rate Distortions	0.215	-0.038	0.103	-3×10^{-5}	0.001
Equipment Investment	0.917	0.486	0.230	0.538	0.141
Non-equipment Investment	0.584	0.171	0.175	0.282	0.109

SSVS Point Estimates and Standard Devs of Regression Coefficients (Mean and standard deviations multiplied by 100)					
	SSVS			Single Best Model	
Explanatory Variable	$\Pr(\gamma = 1 \vert y)$	Mean	St. Dev.	Mean	St. Dev.
St. Dev. of Black Mkt. Prem.	0.138	-0.012	0.054	-2×10^{-5}	0.001
Outward Orientation	0.129	-0.013	0.055	-7×10^{-6}	0.001
Black Market Premium	0.340	-0.068	0.116	-1×10^{-5}	0.001
Area	0.080	-0.001	0.035	3×10^{-6}	0.001
Latin America	0.285	-0.105	0.205	-6×10^{-5}	0.003
Sub-Saharan Africa	0.699	-0.447	0.362	-0.378	0.135
Higher Education Enrolment	0.120	-0.022	0.100	-9×10^{-6}	0.002
Public Education Share	0.119	0.005	0.047	1×10^{-6}	0.001
Revolutions and Coups	0.110	0.002	0.047	-9×10^{-6}	0.001
War	0.204	-0.034	0.094	-2×10^{-5}	0.001

SSVS Posteror Estimates and Standard Devs of Regression Coefficients					
	SSVS			Single Best Model	
Explanatory Variable	$\Pr(\gamma = 1 \vert y)$	Mean	St. Dev.	Mean	St. Dev.
Political Rights	0.130	−0.033	0.121	-1×10^{-4}	0.004
Civil Liberties	0.187	−0.070	0.181	-2×10^{-4}	0.004
Latitude	0.104	0.006	0.086	3×10^{-5}	0.002
Age	0.237	−0.041	0.093	-2×10^{-5}	0.001
British Colony	0.084	−0.005	0.051	-5×10^{-5}	0.002
Fraction Buddhist	0.324	0.076	0.132	3×10^{-5}	0.001
Fraction Catholic	0.216	−0.023	0.158	-2×10^{-5}	0.002
Fraction Confucian	0.972	0.483	0.154	0.542	0.098
Ethnolinguistic Fractionalization	0.141	0.023	0.085	1×10^{-5}	0.002
French Colony	0.138	0.017	0.067	3×10^{-5}	0.001

SSVS Posteror Estimates and Standard Devs of Regression Coefficients					
	SSVS			Single Best Model	
Explanatory Variable	$\Pr(\gamma = 1 \vert y)$	Mean	St. Dev.	Mean	St. Dev.
Fraction Hindu	0.193	-0.068	0.184	-5×10^{-6}	0.003
Fraction Jewish	0.135	-0.008	0.052	-1×10^{-5}	0.001
Fraction Muslim	0.624	0.255	0.241	0.318	0.101
Primary Exports	0.243	-0.073	0.164	-7×10^{-5}	0.002
Fraction Protestant	0.603	-0.189	0.187	-0.276	0.107
Rule of Law	0.485	0.215	0.264	8×10^{-5}	0.002
Spanish Colony	0.129	0.024	0.109	-2×10^{-5}	0.002
Population Growth	0.116	0.017	0.096	3×10^{-6}	0.002
Ratio Workers to Population	0.132	-0.013	0.071	2×10^{-5}	0.001
Size of Labor Force	0.141	0.046	0.167	9×10^{-5}	0.003

LASSO: Theory

- LASSO = Least absolute shrinkage and selection operator
- Developed as a frequentist shrinkage and variable selection method for Fat Data regression models
- Frequentist intuition: OLS estimates minimize sum of squared residuals

$$(y - X\beta)'(y - X\beta)$$

- LASSO minimizes

$$(y - X\beta)'(y - X\beta) + \lambda \sum_{j=1}^{k} |\beta_j|$$

- adds penalty term which depends on magnitude of the regression coefficients
- Bigger values for $|\beta_j|$ penalized (shrink towards zero)
- λ is shrinkage parameter.

LASSO: Theory

- LASSO estimate can be given a Bayesian interpretation:
- equivalent to Bayesian posterior modes if Laplace prior used for β
- I will not define Laplace distribution since will not work with it directly due to following:
- Laplace distribution can be written as scale mixture of Normals (i.e. a mixture of Normal distributions with different variances):

$$
\begin{aligned}
\beta_i &\sim N\left(0, h^{-1}\tau_i^2\right) \\
\tau_i^2 &\sim Exp\left(\frac{\lambda^2}{2}\right)
\end{aligned}
$$

- $Exp\left(.\right)$ is exponential distribution (special case of Gamma)
- Hierarchical prior: depends on τ_i^2 (parameters to be estimated) which have own prior
- Note: smaller τ_i^2 = stronger shrinkage of β_i
- Can show λ plays same role as frequentist λ above

LASSO: Theory

- Bayesian inference can be done using MCMC
- Main idea: conditional on τ_i^2, prior is Normal prior
- Can use standard results for Normal linear regression to obtain $p(\beta|y, h, \tau)$ and $p(h|y, \beta, \tau)$ where $\tau = (\tau_1, .., \tau_K)'$
- All we need is new blocks in MCMC algorithm for drawing τ and λ
- Details given in next slide, but note basic strategy same as for SSVS:
- Use hierarchical Normal prior for β
- Conditional on some other parameters (here τ, with SSVS it was γ) obtain Normal linear regression model
- So just need to work out conditional posterior for these other parameters
- Note: many variants on LASSO (elastic net LASSO) adopt similar strategy

LASSO: Theory

- Write LASSO prior covariance matrix of β as

$$\underline{V} = h^{-1} DD$$

- D is diagonal matrix with diagonal elements τ_i for $i = 1, .., K$
- Then $\beta | y, h, \tau$ is $N\left(\overline{\beta}, \overline{V}\right)$ where

$$\overline{\beta} = \left(X'X + (DD)^{-1}\right)^{-1} X'y$$

- $$\overline{V} = h^{-1}\left(X'X + (DD)^{-1}\right)^{-1}$$

- $h | y, \beta, \tau$ is $G(\overline{s}^{-2}, \overline{\nu})$ with

$$\overline{\nu} = N + K$$

- $$\overline{s}^2 = \frac{(y - X\beta)'(y - X\beta) + \beta'(DD)^{-1}\beta}{\overline{\nu}}$$

LASSO: Theory

- Easier to draw from $\frac{1}{\tau_i^2}$ for $i = 1, .., K$ as posterior conditionals are independent of one another and with inverse Gaussian distributions.
- Inverse Gaussian, $IG(.,.)$, is rarely used in econometrics.
- Standard ways for drawing from IG exist (all we need for MCMC)
- $p\left(\frac{1}{\tau_i^2}|y,\beta,h,\lambda\right)$ is $IG(\overline{c}_i, \overline{d}_i)$ with $\overline{d} = \lambda^2$

$$\overline{c}_i = \sqrt{\frac{\lambda^2}{h\beta_i^2}}$$

- Need prior for λ, convenient to use $\lambda^2 \sim G\left(\underline{\mu}_\lambda, \underline{\nu}_\lambda\right)$
- With this $p\left(\lambda^2|y,\tau\right)$ is $G(\overline{\mu}_\lambda, \overline{\nu}_\lambda)$ with

$$\overline{\nu}_\lambda = \underline{\nu}_\lambda + 2K$$

-

$$\overline{\mu_\lambda} = \frac{\underline{\nu}_\lambda + 2K}{2\sum_{i=1}^K \tau_i^2 + \frac{\underline{\nu}_\lambda}{\underline{\mu}_\lambda}}$$

LASSO: Application

- Again we will use our cross-country growth data set
- All we need to choice are prior hyperparameters: $\underline{\mu}_\lambda = 0.05$ and $\underline{\nu}_\lambda = 1$.
- Relatively non-informative choice
- MCMC algorithm is run for $10,000$ burn in draws followed by $100,000$ included draws.
- In addition to regression coefficient results, tables present results for τ_i for $i = 1, .., K$.
- To gauge degree of shrinkage in LASSO prior, remember:
- prior standard deviation for a regression coefficient is $\sigma\tau_i$
- We find $E\left(\sigma|y\right) = 0.0071$

LASSO: Application

- We find similar results to SSVS and BMA
- Using rule of thumb where we select variables with posterior means two posterior standard deviations from zero select nine explanatory variables.
- These variables are also selected by SSVS and BMS.
- LASSO is doing a very good job at shrinking unimportant variables

Posterior Results for Regression Coefficients with LASSO Prior (Means and standard deviations of regression coeffs multiplied by 100)			
Explanatory Variable	$E\left(\tau_i \middle| y\right)$	Posterior Mean	St. Dev.
Primary School Enrolment	0.293	0.237	0.215
Life expectancy	0.932	1.218	0.182
GDP level in 1960	0.901	-1.144	0.109
Fraction GDP in Mining	0.429	0.303	0.058
Degree of Capitalism	0.158	0.094	0.110
No. Years Open Economy	0.578	0.509	0.084
% Pop. Speaking English	4×10^{-4}	-6×10^{-5}	0.003
% Pop. Speak. For. Lang.	0.122	0.069	0.093
Exchange Rate Distortions	6×10^{-4}	-1×10^{-4}	0.004
Equipment Investment	0.581	0.511	0.081
Non-equipment Investment	0.190	0.118	0.124

Posterior Results for Regression Coefficients with LASSO Prior (Means and standard deviations of regression coeffs multiplied by 100)			
Explanatory Variable	$E\left(\tau_i \middle| y\right)$	Posterior Mean	St. Dev.
St. Dev. of Black Mkt. Prem.	5×10^{-4}	-9×10^{-5}	0.003
Outward Orientation	5×10^{-4}	-9×10^{-4}	0.004
Black Market Premium	6×10^{-4}	-9×10^{-5}	0.004
Area	3×10^{-4}	4×10^{-5}	0.001
Latin America	0.005	0.002	0.017
Sub-Saharan Africa	3×10^{-4}	-1×10^{-5}	0.002
Higher Education Enrolment	6×10^{-4}	-1×104	0.005
Public Education Share	3×10^{-4}	2×10^{-5}	0.001
Revolutions and Coups	0.001	3×10^{-4}	0.047
War	5×10^{-4}	1×10^{-4}	0.002

Posterior Results for Regression Coefficients with LASSO Prior			
Explanatory Variable	τ_i	Posterior Mean	St. Dev.
Political Rights	5×10^{-4}	3×10^{-5}	0.002
Civil Liberties	3×10^{-4}	5×10^{-5}	0.002
Latitude	7×10^{-4}	2×10^{-4}	0.003
Age	3×10^{-4}	1×10^{-5}	0.001
British Colony	4×10^{-4}	2×10^{-5}	0.001
Fraction Buddhist	0.436	0.314	0.077
Fraction Catholic	0.373	0.253	0.130
Fraction Confucian	0.645	0.617	0.062
Ethnolinguistic Fractionalization	0.001	4×10^{-4}	0.004
French Colony	0.075	0.039	0.071

Posterior Results for Regression Coefficients with LASSO Prior			
Explanatory Variable	τ_i	Posterior Mean	St. Dev.
Fraction Hindu	8×10^{-4}	2×10^{-4}	0.004
Fraction Jewish	6×10^{-4}	1×10^{-4}	0.002
Fraction Muslim	0.671	0.662	0.087
Primary Exports	6×10^{-4}	-6×10^{-5}	0.004
Fraction Protestant	0.002	-9×10^{-4}	0.013
Rule of Law	0.002	8×10^{-4}	0.009
Spanish Colony	0.007	0.003	0.021
Population Growth	0.002	5×10^{-4}	0.007
Ratio Workers to Population	0.001	1×10^{-4}	0.002
Size of Labor Force	0.349	0.217	0.057

Summary

- Applications involving Big Data are proliferating in economics
- In the lecture on regression, we showed how BMA can be used to surmount over-parameterization problems
- Challenges with BMA largely computational: How do we handle 2^K models?
- An answer was MC^3
- The approaches in this lecture turn model space problem (involving marginal likelihoods, etc.) into estimation problem
- SSVS and LASSO are two important such methods
- Estimate one model (using hierarchical prior of particular form) and let it do model selection or model averaging
- These are just two of many such methods (hot area of literature)
- Here we have used them with regression, later we will return to them with VARs