Bayesian Methods for Regression Models with Fat Data

Bayesian Methods for Regression with Fat Data Overview

- Reading: Handout "Bayesian Methods for Fat Data" on course website
- Big Data is hot topic that may revolutionize empirical work and change the way we do econometrics
- Hal Varian, "Big Data: New Tricks for Econometrics," Journal of Economic Perspectives, 2014
- "Big" Data may be "tall" or "fat"
- Tall Data = data with many observations
- Fat Data = data with many variables
- In macroeconomics, Fat Data is becoming common and this is what I will cover in this course

Bayesian Methods for Regression with Fat Data Overview

- For many countries can easily get data for over 100 variables
- With globalization may want to work with several countries (so even more variables)
- US: FRED-MD: A Monthly Database for Macroeconomic Research (Federal Reserve Bank of St. Louis)
- 134 variables (output, prices, consumption, interest rates, stock prices, money, housing, unemployment, wages, etc. etc. etc.)
- Why work with all of them?
- When forecasting (e.g. inflation, GDP growth, unemployment) the more information the better
- When estimating a model want to avoid omitted variables bias
- E.g. even if you have DSGE model with inflation, interest rates and unemployment do not model just these 3 variables
- If other variables have important explanatory power and you omit them, model is mis-specified

Bayesian Methods for Regression with Fat Data Overview

- In this lecture will show some Fat Data methods in context of regression, but they also can be used with other models
- To illustrate use a classic cross-country growth regression data set:
- Why do some countries grow faster than others?
- Numerous potential explanations (e.g. education, investment, governance, institutions, trade, colonialism, etc. etc.)
- 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
- Fat Data: large number of explanatory variables relative to number of observations
- In other Fat Data applications can have $K > N$ (e.g. stock returns for large K companies observed only for a few months).

Bayesian Methods for Regression with Fat Data 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 Methods for Regression with Fat Data Overview

- Why not do hypothesis testing to reduce K?
- Pre-test problem (also called multiple testing or multiple comparisons problem)
- The unrestricted regression will have $K = 41$
- There are 41 different restricted regressions which drop one of the explanatory variables
- There are $\frac{K(K-1)}{2}$ restricted regressions with drop two of the explanatory variables
- etc. etc. etc.
- In total there are $2^K = 2,199,023,255,552$ possible regression models involving some combination of the explanatory variables
- Jargon: this is the model space
- Which one to choose?

Bayesian Methods for Regression with Fat Data Overview

- Sequential hypothesis testing methods often used in smaller problems
- Let us suppose you can come up with a sequence of hypothesis tests to navigate through your huge model space
- E.g. do a hypothesis test to decide whether to drop a variable, then do a second hypothesis test using the restricted regression
- But significance levels no longer valid (or must be adjusted) when more than one test is done
- E.g. one t-test using standard critical value has 5% level of significance. But if you do two t-tests sequentially second one no longer has 5% level of significance
- Maybe minor issue in small data problems, but with Fat Data problems number of sequential hypothesis tests may be HUGE, true level of significance vastly different from nominal one (or necessary adjustments become huge)
- Bottom line: not easy to do hypothesis testing to select a more parsimonious model

Bayesian Methods for Regression with Fat Data 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 provides introduction to new methods including:
- i) Bayesian Model Averaging (BMA) and Bayesian Model Selection (BMS)
- ii) Stochastic search variable selection (SSVS)
- iii) Least absolute shrinkage and selection operator (LASSO)

Bayesian Model Averaging

Overview

- BMA can be used with any set of models
- Proved useful with Fat Data problems.
- Model selection: choose a single model and present estimates or forecasts based on it
- Model averaging: take a weighted average of estimates or forecasts from all models with weights given by $p(M_r|y)$
- Let M_r for $r = 1, .., R$ denote R models.
- If ϕ is a parameter to be estimated (or a function of parameters) or a variable to be forecast, then the rules of probability imply:

$$p(\phi|y) = \sum_{r=1}^{R} p(\phi|y, M_r)\, p(M_r|y)$$

- Allows for a formal treatment of model uncertainty.
- Model selection: choose a single model and act as though it were true
- BMA incorporates uncertainty about which model generated the data.

The Model Space

- Let X_r is a $N \times k_r$ matrix containing some (or all) columns of X, then each model is

$$y = \alpha \iota_N + X_r \beta_r + \varepsilon$$

- ι_N is a $N \times 1$ vector of ones so as to say each model contains an intercept
- Other assumptions as for Normal linear regression model under classical assumptions.
- 2^K possible choices for X_r and, thus, the number of models, $R = 2^K$.
- Computational concerns: estimating every model will be impossible
- E.g. if each model could be estimated in 0.001 seconds, over 100 years to estimate them all
- Use natural conjugate prior to make estimation of each model as fast as possible

BMA Priors

- We want a prior for model r that is:
- Informative (so as to provide valid marginal likelihoods for model comparison)
- Objective (requiring minimal subjective input)
- Automatic (does not have to be individually chosen for each of the many models)
- g-prior is commonly used:
- Prior mean shrinks coefficients towards zero:

$$\underline{\beta}_r = 0$$

- Prior covariance matrix is $h^{-1}\underline{V}_r$ where

$$\underline{V}_r = \left(gX_r'X_r\right)^{-1}$$

- g is a scalar

The g-prior

- The g-prior was suggested in Zellner (1986)
- Justification:
- Under non-informative prior $h^{-1}\left(X_r'X_r\right)^{-1}$ is posterior covariance matrix
- Amount of information in data for estimating β_r (information matrix)
- Prior covariance matrix $h^{-1}\left(gX_r'X_r\right)^{-1}$ says:
- Prior information that $\beta_r = 0$ takes same form as data information
- g controls relative strengths of the prior and data information.
- $g = 1$: prior and data are given equal weight.
- $g = 0.01$: prior information receives one per cent of the weight as data
- There exist commonly-used rules of thumb for choosing g
- Or g can be treated as unknown parameter with own prior and estimated
- Noninformative prior for h typically used

BMA Posterior

- With natural conjugate prior, analytical results for M_r
- Posterior is Normal-Gamma
- Marginal likelihood (for producing posterior model probs) analytical
- Predictive density is t-distribution
- Exact formulae given in Handout
- Key thing: for each model, everything we need can be calculated quickly
- But even with this, doing BMA with 2^K models for $K > 20$ or so too computationally demanding

BMA Computation

- Previously we talked about posterior simulation as tool for learning about complicated posteriors
- For BMA can do model simulation
- A popular algorithm is Markov Chain Monte Carlo Model Composition (MC^3)
- Similar to a random walk Metropolis-Hastings algorithm, but models are drawn instead of parameters

MC-cubed

- $M^{(s)}$ for $s = 1, .., S$ are drawn models
- Averaging estimates/forecasts over drawn models will converge to the true BMA posterior or predictive estimates as $S \to \infty$.
- if ϕ is parameter of interest, then

$$\widehat{\phi} = \frac{1}{S} \sum_{s=1}^{S} E\left(\phi | y, M^{(s)}\right)$$

- will converge to $E\left(\phi|y\right)$.
- Frequencies with which models are drawn can be used to calculate Bayes factors.
- If MC^3 algorithm draws M_i A times and M_j B times, then $\frac{A}{B}$converges to Bayes factor comparing M_i to M_j.
- In practice, discard initial draws as burn-in

MC-cubed: How are models drawn?

- Want to draw $s = 1, .., S$ and suppose you have drawn $M^{(s-1)}$
- Candidate model, M^*, is proposed drawn randomly (with equal probability) from a set of models including:
- i) $M^{(s-1)}$
- ii) all models which delete one explanatory variable from $M^{(s-1)}$
- iii) all models which add one explanatory variable to $M^{(s-1)}$.
- Candidate model accepted with probability:

$$\alpha\left(M^{(s-1)}, M^*\right) = \min\left[\frac{p(y|M^*)p(M^*)}{p(y|M^{(s-1)})p(M^{(s-1)})}, 1\right]$$

- If M^* is accepted then $M^{(s)} = M^*$, else $M^{(s)} = M^{(s-1)}$.
- Can prove MC-cubed will converge to true BMA posterior/predictive

BMA Application

- Cross-country growth regression data set with $N = 72$ and $K = 41$
- Use common recommendation to set $g = \frac{1}{N}$ if $N > K^2$ or $g = \frac{1}{K^2}$ if $N \leq K^2$
- Run MC-cubed algorithm for $2,200,000$ draws, discarding first $200,000$ as burn-in
- Is this enough draws?
- Convergence diagnostic: calculate posterior model probabilities analytically and using MC^3 and compare
- Next table indicates convergence
- Note that best model receives less than 1% of posterior model
- Model selection puts all weight on this single model — ignoring huge amount of model uncertainty

Posterior Model Probabilities for Top 10 Models		
	$p(M_r\|y)$ Analytical	$p(M_r\|y)$ MC^3 estimate
1	0.0087	0.0089
2	0.0076	0.0077
3	0.0051	0.0050
4	0.0034	0.0035
5	0.0031	0.0032
6	0.0029	0.0029
7	0.0027	0.0025
8	0.0027	0.0027
9	0.0027	0.0026
10	0.0024	0.0022

BMA Application

- Next table presents results:
- Posterior mean and standard deviation for each explanatory variable using BMA and BMS
- Rule of thumb: if an estimate (posterior mean) more than two standard deviations from zero likely to be important
- Column labelled "Prob." = probability that the corresponding explanatory variable should be included.
- = proportion of models drawn by MC^3 which contain the corresponding explanatory variable
- BMS ensures parsimony by choosing 14 variables
- By ignoring model uncertainty estimates are more precise (smaller st. dev.)
- BMA ensures parsimony by averaging over many small models
- Average number of exp. vars in a model drawn by MC^3 is 11.4

Point Estimates and Standard Devs of Regression Coefficients (Mean and standard deviations multiplied by 100)					
	BMA			BMS	
Explanatory Variable	Prob.	Mean.	St. Dev.	Mean	St. Dev.
Primary School Enrolment	0.207	0.104	0.234	0.048	0.018
Life expectancy	0.933	0.961	0.392	0.090	0.020
GDP level in 1960	0.999	-1.425	0.278	-1.463	0.193
Fraction GDP in Mining	0.459	0.147	0.181	0.322	0.108
Degree of Capitalism	0.457	0.151	0.183	0.387	0.094
No. Years Open Economy	0.513	0.260	0.283	0.557	0.138
% Pop. Speaking English	0.069	-0.011	0.047	–	–
% Pop. Speak. For. Lang.	0.068	0.012	0.059	–	–
Exchange Rate Distortions	0.082	-0.017	0.070	–	–
Equipment Investment	0.923	0.552	0.236	0.548	0.128
Non-equipment Investment	0.434	0.136	0.174	0.347	0.099
St. Dev. of Black Mkt. Prem.	0.048	-0.006	0.037	–	–
Outward Orientation	0.037	-0.003	0.029	–	–

Point Estimates and Standard Devs of Regression Coefficients (Mean and standard deviations multiplied by 100)					
	BMA			BMS	
Explanatory Variable	Prob.	Mean.	St. Dev.	Mean	St. Dev.
Black Market Premium	0.179	-0.040	0.097	–	–
Area	0.030	-0.001	0.021	–	–
Latin America	0.215	-0.082	0.191	–	–
Sub-Saharan Africa	0.738	-0.473	0.347	-0.543	0.124
Higher Education Enrolment	0.046	-0.008	0.056	–	–
Public Education Share	0.032	-0.001	0.024	–	–
Revolutions and Coups	0.031	-0.001	0.023	–	–
War	0.075	-0.014	0.062	–	–

Posteror Estimates and Standard Devs of Regression Coefficients					
	Bayesian Model Averaging			Single Best Model	
Explanatory Variable	Prob.	Mean	St. Dev.	Mean	St. Dev.
Political Rights	0.094	−0.028	0.107	–	–
Civil Liberties	0.131	−0.050	0.015	−0.284	0.176
Latitude	0.041	0.001	0.052	–	–
Age	0.085	−0.015	0.058	–	–
British Colony	0.041	−0.003	0.032	–	–
Fraction Buddhist	0.196	0.047	0.109	–	–
Fraction Catholic	0.128	−0.011	0.121	–	–
Fraction Confucian	0.990	0.493	0.127	0.503	0.090
Ethnolinguistic Fractionalization	0.060	0.010	0.056	–	–
French Colony	0.049	0.007	0.040	–	–

Posteror Estimates and Standard Devs of Regression Coefficients					
	Bayesian Model Averaging			Single Best Model	
Explanatory Variable	Prob.	Mean	St. Dev.	Mean	St. Dev.
Fraction Hindu	0.126	−0.035	0.120	–	–
Fraction Jewish	0.037	−0.002	0.028	–	–
Fraction Muslim	0.640	0.025	0.023	0.295	0.093
Primary Exports	0.100	−0.029	0.105	−0.352	0.136
Fraction Protestant	0.455	−0.143	0.178	−0.277	0.098
Rule of Law	0.489	0.244	0.279	0.563	0.134
Spanish Colony	0.058	0.010	0.068	–	–
Population Growth	0.037	0.005	0.048	–	–
Ratio Workers to Population	0.045	−0.005	0.043	–	–
Size of Labor Force	0.075	0.018	0.097	–	–

Variable Selection and Shrinkage Using Hierarchical Priors

- Any sort of prior information can be used to overcome lack of data information with Fat 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
- But more sophisticated methods are growing in popularity (in many models, not only regression)
- I introduce two popular ones: LASSO and SSVS
- Many others (and not all Bayesian)
- Korobilis, D. (2013). Hierarchical shrinkage priors for dynamic regressions with many predictors. International Journal of Forecasting 29, 43-59.

SSVS: Overview

- To show main ideas assume (for now) β is a scalar
- Remember prior shrinkage can be done through prior variance: $\beta \sim N(0, \underline{V})$
- If $\underline{V}$ is small, then strong prior information β is near 0.
- E.g. $\underline{V} = 0.0001$ then $\Pr\left(-0.0196 \leq \beta \leq 0.0196\right) = 0.95$
- If $\underline{V}$ is big then prior becomes more non-informative
- If $\underline{V} = 100$ then $\Pr\left(-19.6 \leq \beta \leq 19.6\right) = 0.95$
- Note: exactly what "small" and "large" means depends on the empirical application and units of measurement of data

SSVS: Overview

- 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 $\Pr\left(\gamma = 1|y\right) > \frac{1}{2}$ 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^2_{0i} and τ^2_{1i}, 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:

$$\Pr(\gamma_i = 1) = \underline{q}_i$$
$$\Pr(\gamma_i = 0) = 1 - \underline{q}_i$$

- 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	$\Pr(\gamma = 1 \mid y)$	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

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

- adds penalty term which depends on magnitude of the regression coefficients
- Bigger values for $\left|\beta_j\right|$ 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 $\tau_i^2 =$ 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

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

-

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

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 $E(\sigma|y) = 0.0071$

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	$E(\tau_i \vert y)$	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 Fat Data are proliferating in economics
- We have shown 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
- Many other approaches 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, but sometimes used with VARs