Bayesian comparison of SSVS versus hierarchical SSVS

Liuyi Hu and Wenhao Hu

Department of Statistics, NCSU

Objective

- Identify robust economic growth determinants using two methods:
  (1) stochastic search variable selection
  (2) stochastic search variable selection with hierarchical modeling
- Compare the performance of the two methods

Introduction

The multiplicity of possible regressors is one of the major difficulties faced by researchers trying to make sense of the empirical evidence on economic growth. When questions can be addressed with very large datasets it is routine practice to include every regressor that comes to mind and then report those that have significant coefficients. Often, however, we do not have the luxury of having a sample size that allows us to include all potential regressors. Especially for the cross-country growth regressions, the number of countries in the world is limited, rendering the all-inclusive regression computationally impossible. Therefore, we implement stochastic search variable selection to identify those significant predictors.

Data

The dataset is from Sala-i-Martin *et al.* (2004). The total number of explanatory variables is $K = 67$ with observations for $N = 88$ countries. The dependent dependent variable is ***average growth rate of GDP per capita between 1960-96***. Figure 1 shows the *average growth rate of GDP* across 88 countries.

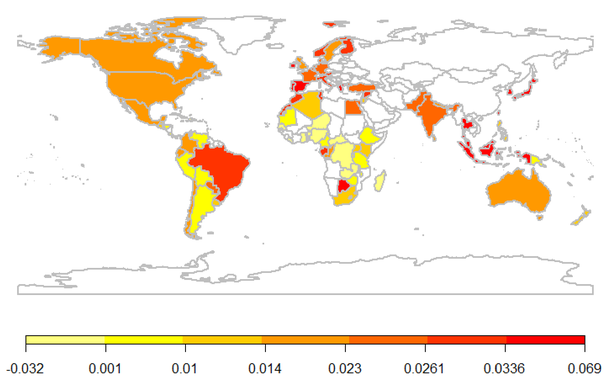

Figure 1: average growth rate of GDP per capita between 1960-96

Models

- **Model 1**

$$Y|\mu, \beta, \sigma^2 \sim N(\mu\mathbf{1_n} + \mathbf{X}\beta, \sigma^2\mathbf{I_n})$$

$$\beta = (\delta_1\alpha_1, \ldots, \delta_p\alpha_p)$$

Prior

$$\mu \sim flat \quad \frac{1}{\sigma^2} \sim Gamma(0.01, 0.01)$$

$$\delta_j \sim B(1, 0.5), \quad \alpha_j \sim N(0, 1)$$

- **Model 2**

$$Y_k|\mu_k, \beta_k, \sigma^2 \sim N(\mu_k\mathbf{1_{n_k}} + \mathbf{X_k}\beta_\mathbf{k}, \sigma^2\mathbf{I_{n_k}})$$

$$\beta_k = (\delta_{k1}\alpha_{k1}, \ldots, \delta_{kp}\alpha_{kp}) \quad \delta_{kj} \sim B(1, P_k)$$

Prior

$$\mu \sim flat \quad \frac{1}{\sigma^2} \sim Gamma(0.01, 0.01)$$

$$\alpha_{kj} \sim N(0, 1), P_k \sim Beta(1, 1)$$

In Model 2, we divided the countries into 4 groups based on their development status, and Y_i is the observations in the ith group.

Model Selection

For each model, Gibbs sampling is implemented to get posterior sample. We set burning in as 2000 and the total runs as 20000. The convergence of Gibbs sampling is confirmed by checking the trace plot. Bayesian p-value and Cross Validation are used as the criterion for model selection in this project. From Figure 2, we can see that the bayesian p-value of Model 2 is closer to 0.5 than the the bayesian p-value of Model 1. It suggests Model 2 fit the data better than Model 1.

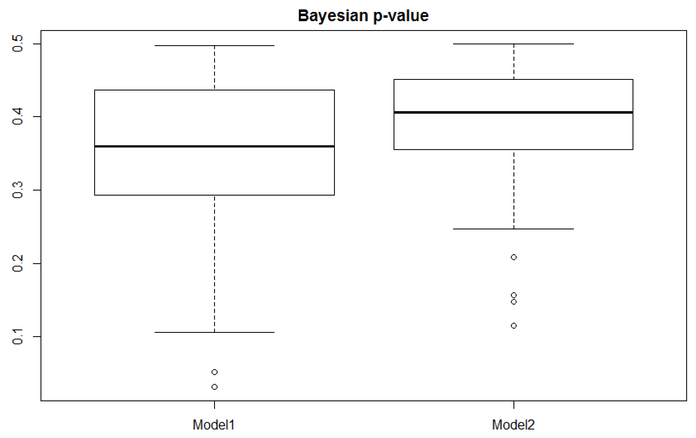

Figure 2: Bayesian p-value for each model

Table 1 shows the cross validation results, Model 2 has a smaller MSE and MAD. And threfore we choose Model 2 as our Final Model.

Table 1: 5-fold CV results

Criterion	Model 1	Model 2
MSE	2.16×10^{-4}	1.34×10^{-4}
MAD	1.11×10^{-2}	6.38×10^{-3}

Results

In model 2, we divide all countries to four groups according to their GDP growth rate. In Group 1, GDP growth rate is negative. GDP growth rate is in the range of [0, 0.014), [0.014, 0.026), and [0.026, $+\infty$) for Group 2, 3 and 4 respectively. Table 2 shows the top 10 variables relevant to GDP growth rate in each group. In different groups, the variables relevant to GDP growth rate may not be the same. For instance, *EAST* and *CONFUC* rank the top 2 in Group 4. And in Group 2, *CONFUC* and *HINDU00* are more relevant to the GDP growth rate than other variables.

Table 2: Top 10 variables relevant to GDP growth rate in each group

Rank	Group 1	Group 2	Group 3	Group 4
1	EAST	CONFUC	EAST	EAST
2	EUROPE	HINDU00	OIL	CONFUC
3	SOCIALIST	BUDDHA	CONFUC	HINDU00
4	SPAIN	EAST	SOCIALIST	SPAIN
5	HINDU00	POP60	SAFRICA	SCOUT
6	CONFUC	OIL	LANDLOCK	TOTIND
7	BUDDHA	EUROPE	SPAIN	EUROPE
8	ORTH00	GDE1	AVELF	SOCIALIST
9	H60	COLONY	LAAM	HERF00
10	ENGFRAC	LANDAREA	BRIT	LAAM

Conclusions

- We found the performance of stochastic search variable selection with hierarchical structure is better than stochastic search variable selection without hierarchical modeling in this dataset.
- Top variables relevant to GDP growth rate are identified in each group. In different groups, the factors relevant to GDP growth rate may not be the same.

Limitations and Future Works

- In this project, we separated the groups by using pre-defined cutoff. It may not appropriate if we set the pre-defined cutoff wrongly. Future works will explore whether we can set the cutoff more flexibly.
- For future work, we can take advantage of the geographical information to provide informative structure and prior for the model.