

Stochastic Search Variable Selection

Sooraj Bhat
ISyE 6416 Final Presentation

- Method: Stochastic Search Variable Selection (SSVS)
- Paper: George, Edward I.; McCulloch, Robert E. **Variable Selection Via Gibbs Sampling.** *Journal of American Statistical Association*; September 1993; 88, 423; pg 881.

- **SSVS is for feature selection in linear regression problems.**
- Given predictors $X = [X_1, \dots, X_p]$ and a dependent variable Y , find and fit the “best” model of the form

$$Y = X_1^* \beta_1^* + \dots + X_q^* \beta_q^* + \epsilon$$

where X_1^*, \dots, X_q^* is a selected subset of X_1, \dots, X_p .

- Goal: Avoid comparing all 2^p models.
- Solution: Hierarchical model + Gibbs sampling.

- Start with the standard regression setup:

$$Y|\beta, \sigma^2 \sim N_n(X\beta, \sigma^2 I)$$

- Add indicator variables $\gamma = [\gamma_1, \dots, \gamma_p]$, which dictate participation. The priors for β and σ^2 now depend on γ .
- Example: $\gamma = [1, 0, 0, 1]$ represents the model

$$Y = X_1\beta_1 + X_4\beta_4 + \epsilon$$

- We want the model with maximum posterior probability:

$$\gamma^* = \arg \max_{\gamma} f(\gamma|Y)$$

- Naive method: “enumerate and score”
- Better method: sampling
- Construct an auxiliary Gibbs sequence in which $f(\gamma|Y)$ is embedded:

$$\beta_0, \sigma_0^2, \gamma_0, \beta_1, \sigma_1^2, \gamma_1, \dots$$

- The sampling is simple and efficient.
- The most promising models show up frequently.
- The parameters for the prior on β require some tweaking.

- No general implementations of SSVS on the web!?
- Fortunately, a MATLAB implementation from scratch isn't too difficult.
- `ssvs.m`, ~100 LOC, pedagogical implementation, requires the Statistics Toolbox, only tested with MATLAB7.
- ~3x slower than the authors' (C++?) implementation.
- <http://www.cc.gatech.edu/~sooraj/ssvs/>

- Experiments: synthetic and real-world datasets.
- Many parameters have reasonable defaults.
- Exception: parameters for the prior on β_i
- The authors suggest some semi-automatic defaults.

- Let $Y = X_4 + 1.2X_5 + \epsilon$
- $1 \leq i \leq 5, X_i \sim N_{60}(0, I), \epsilon \sim N_{60}(0, \sigma^2 I), \sigma = 2.5$

Model	%
5	31.4
4,5	15.6
–	15.0
4	6.3
1,5	5.4
3,5	4.1
2,5	3.9
1,4,5	2.6
3,4,5	2.2
1	2.0

- Now, replace X_3 with $X_3^* = X_5 + 0.15Z$ where $Z \sim N_{60}(0, I)$
- X_3^* is a “strong proxy” for X_5 .

Model	%
4	24.8
–	10.3
3,4	9.1
2,4	7.5
4,5	6.9
1,4	5.8
2	3.3
3,4,5	3.3
3	3.2
5	2.8

- Hald dataset: $n = 13$ observations on Y (heat evolved during a chemical reaction) and $p = 4$ variables X_1, X_2, X_3, X_4 (inputs to the reaction).
- Other techniques favor the models (1,2), (1,4) and (1,2,4).

Model ¹	%	Model ²	%
—	31.4	1,2	66.7
1	21.8	1,4	24.2
4	8.3	1,2,4	3.4
3	6.7	1,2,3	2.9
2	6.6	1,3,4	2.6
1,2	5.4	1,2,3,4	(3)
1,3	4.9	1	(1)
1,4	4.9		
3,4	2.1		
2,3	1.8		

- SSVS is an efficient sampling-based method for feature selection in linear regression models.
- ...but still requires some exploratory analysis by the user (should be used as a tool for locating promising models).
- These results and implementation seem faithful to the original paper.

Thanks!