

User Guide for the SSVS MATLAB Package

Sooraj Bhat

April 16, 2007

1 Introduction

Stochastic Search Variable Selection (SSVS) is a feature selection algorithm for linear regression problems. The algorithm is described by George and McCulloch [1]. The software is coded in MATLAB. It requires the Statistics Toolbox and has only been tested with MATLAB7.

2 Contents

The package contains the following components:

2.1 Code

These files contain the implementation of SSVS.

- `ssvs.m` – The main SSVS function.
- `count.m` – Helper function which computes the frequencies of the generated samples.

2.2 Examples

These files contain the implementation of various examples from the paper.

- `example1.m` – Example 4.1 (first part), a small synthetic example with independent features.
- `example2.m` – Example 4.1 (second part), the same small synthetic examples with a modified feature which creates a “strong proxy”.
- `example3.m` – Example 4.2, a large synthetic example.
- `example4.m` – Example 5.1, feature selection on the Hald dataset [2].

2.3 Sample Runs

These files contain the saved MATLAB workspace after running a particular example on a particular setting. The possible settings are $(\sigma_{\beta_i}/\tau_i, c_i) = (1, 5), (1, 10), (10, 100), (10, 500)$.

- `data1setting2.mat` – Example 4.1 (first part) with setting (1, 10).
- `data2setting1.mat` – Example 4.1 (second part) with setting (1, 5).

- `data3setting1.mat` – Example 4.2 with setting (1, 5).
- `data4setting1.mat` – Example 5.1 with setting (1, 5).
- `data4setting3.mat` – Example 5.1 with setting (10, 100).

3 Usage

To use SSVS, type the following at the MATLAB prompt:

```
samples = ssvs(X,Y,m,cfg)
```

The example files contain useful sample usages of the SSVS function. I strongly suggest looking at and running those examples.

3.1 Inputs

- **X** – An $n \times p$ matrix of observations, where n is the number of observations and p is the number of predictors/covariates/features.
- **Y** – An $n \times 1$ column vector of the dependent variable values.
- **m** – The number of samples to generate from the Gibbs sequence.
- **cfg** – Used to specify some settings on the prior on the coefficients β_i . The authors suggest the four semi-automatic settings [1 5], [1 10], [10 100], and [10 500].

3.2 Output

- **samples** – An $m \times p$ matrix whose rows are the Gibbs samples. Each sample represents a model, e.g. a row [1 0 0 1] would be the model in which features X_1 and X_4 should be included and features X_2 and X_3 should be excluded (X_1, X_2, X_3, X_4 form the columns of **X**).

References

- [1] George, Edward I.; McCulloch, Robert E. “Variable Selection Via Gibbs Sampling.” *Journal of American Statistical Association*; September 1993; 88, 423; pg 881.
- [2] Draper, N., and Smith, H. (1981), *Applied Regression Analysis* (2nd ed.), New York; John Wiley.