S610-Final_Project-sSINDy

ChunHsien Lu

12/11/2019

Introduction

In this project, we mainly follow the paper written by Boninsegna et. al. This paper focuses on the reconstruction of an SDE. It does well on the example they provide. In order to understand the ideas provide in the paper, here, we attempt to apply the same method to Brownian motions(BMs).

Suppose X_t is an Ito's stochastic process, i.e.

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t$$

where b and σ are functions, B_t is a BM. The following is two important properties that can help us to reconstruct b and σ:

$$b(x) = \lim_{s \to 0} \frac{1}{s} E[X_{t+s} - X_t | X_t = x]$$

and

$$\sigma(x) = \lim_{s \to 0} \frac{1}{s} E[(X_{t+s} - X_t)(X_{t+s} - X_t) | X_t = x].$$

Since what we want to do is BM, $b(x) = 0$ and $\sigma(x) = 1$. From the two properties, fixed some small $h > 0$,we try to fit $b(x)$ by using $[X_{(m+1)h} - X_{mh}]_m$ and $[X_{mh}]_m$. Furthermore, $\sigma(x)$ is via $[(X_{(m+1)h} - X_{mh})(X_{(m+1)h} - X_{mh})]_m$ and $[X_{mh}]_m$.

Intuitively, the algorithm proposed in the paper is just by using backward elimination and cross validation. The design matrix is $[M]_{ij} = [f_j(X_{ih})]_{ij}$ where f_j's are possible candidates in $b(x)$. Assume n functions f_j are chosen. The full model is fitted by

$$(X_{(i+1)h} - X_{ih}) \sim \sum_{j=1}^{n} \beta_j f_j(X_{ih}).$$

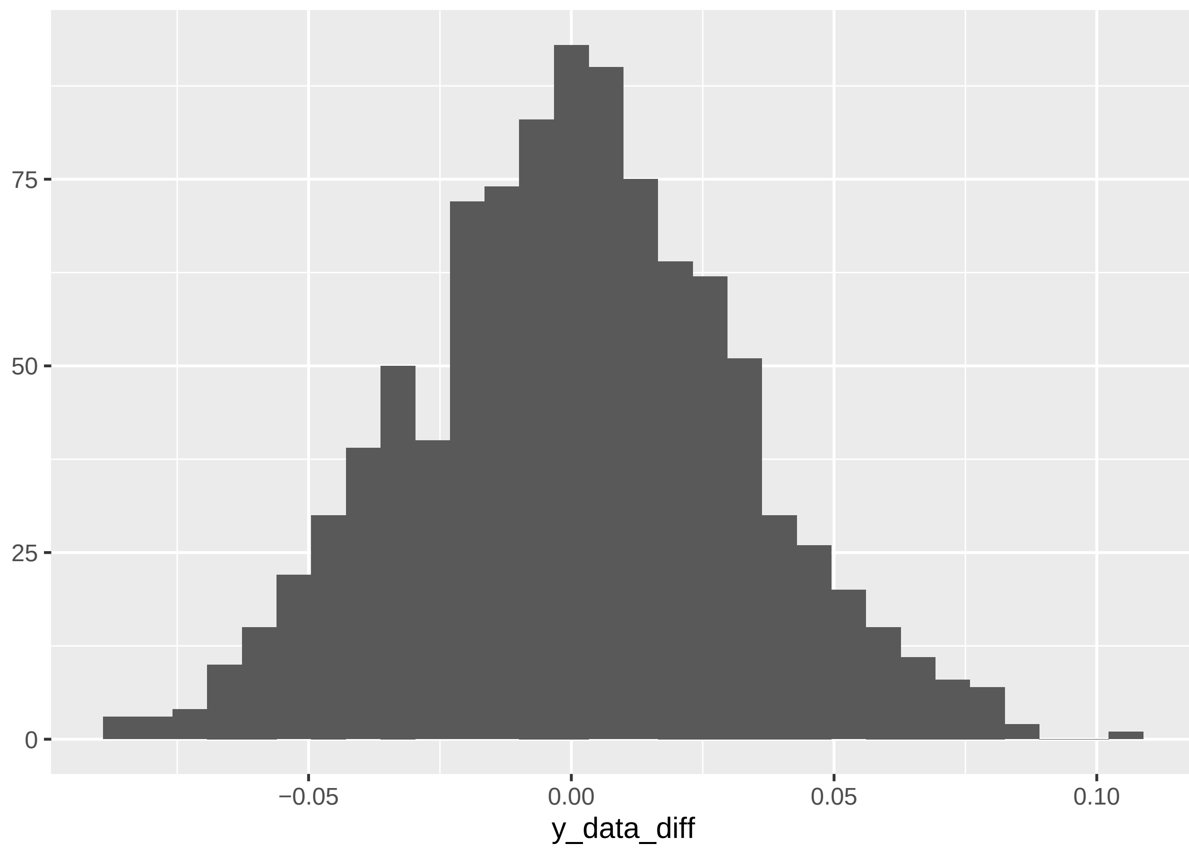
The smaller model with $k-1$ predicted variables is removed the one with lowest absolute value in the model with k variables. Via this way, we can have $n+1$ possible models.($y = 0$ is included) Now, CV comes into play to decide the best model.

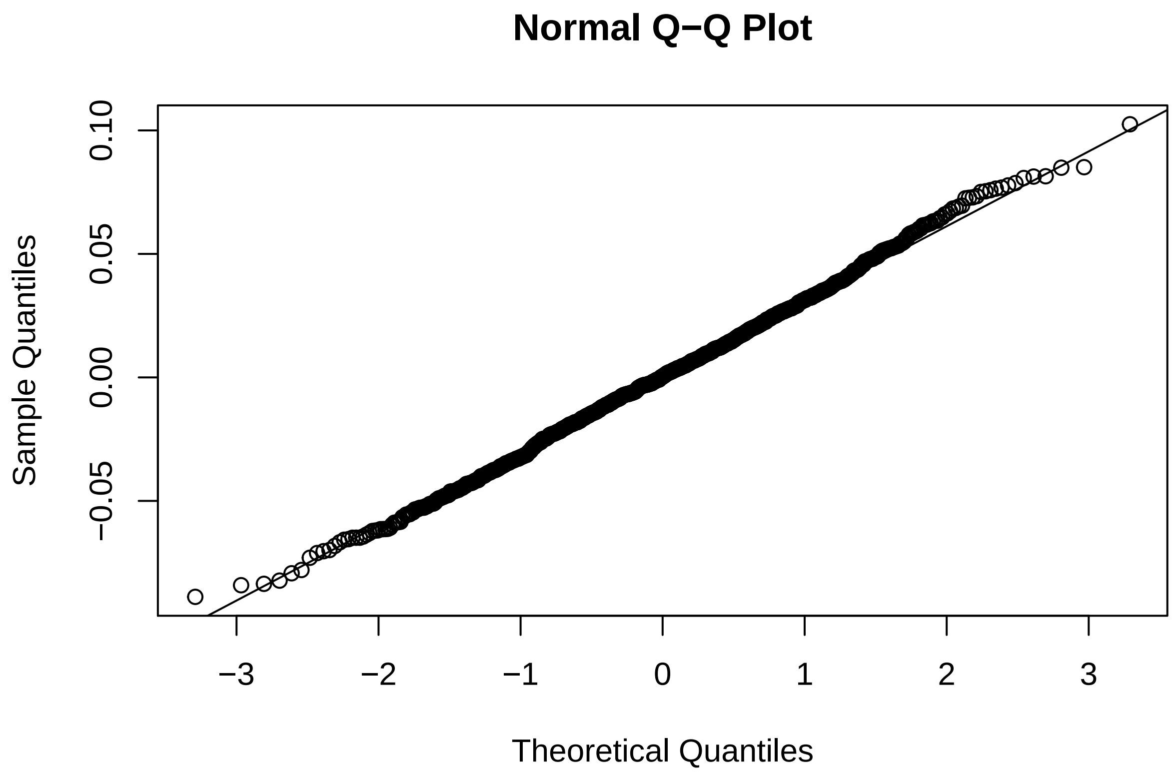
The codes can be found in the following link:
https://github.com/Pielann/sto_SINDy.git

Simulation Design

Fit the drift term and explain the output of sSINDy:

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```


75
50
25
0
−0.05
0.00
0.05
0.10
y_data_diff




```
##  mean: 0.0005100079
##  variance: 0.0009834589
```

Since X_t is a BM, then $[X_{(m+1)h} - X_{mh}]_m$ should follow $N(0, h)$. We choose $h = 0.001$ here. As shown above, the variance is 0.0009835, which is really closed to h.

```
## Warning: package 'magrittr' was built under R version 3.5.3
```

```
## $coef_order
## [1] 4 3 2 1
##
## $cv_error
## [1] 8844620 9102773 9097858 9202561 9332238
##
## $`min#`
## [1] 0
##
## $coef
## [1] 0 0 0 0
```

We did it by 10-fold cross-validation(CV) to compare the models which are up to 3rd order. More precisely, we can compare the models

$$y = 0,\ y = \beta_{i_1} x^{i_1},\ y = \beta_{i_1} x^{i_1} + \beta_{i_2} x^{i_2},\ y = \beta_{i_1} x^{i_1} + \beta_{i_2} x^{i_2} + \beta_{i_3} x^{i_3},\ y = \beta_{i_1} x^{i_1} + \beta_{i_2} x^{i_2} + \beta_{i_3} x^{i_3} + \beta_{i_4} x^{i_4},$$

wehre $i_1, i_2, i_3, i_4 \in \{0, 1, 2, 3\}$ are distinct. The output "coef_order" express the order of the coefficients to be removed. 1 means the first variable to be removed. Hence, the model with three variables is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

since the fouth term is the first one to be thrown. "cv_error" denotes the total errors via CV for models with variables from 0 to 3. "min#" means the model with least errors. Moreover, the "coef" is the coefficients from the best choice. Surprisingly, the true function for drift term is 0. It shows that, in this example, we do really rebuild the drift term.

Fit the diffussion term:

```
## $coef_order
## [1] 4 3 2 1
##
## $cv_error
## [1] 25467.62 17055.15 17182.80 17418.07 17619.68
##
## $`min#`
## [1] 1
##
## $coef
## [1] 0.9827356 0.0000000 0.0000000 0.0000000
```

The true diffusion term is 1.011784 which is also really closed to 1. Hence, the diffusion is rebuilt.

Confidence Intervals

CI for Drift Terms

```
##       [,1] [,2] [,3] [,4]
## 2.5%     0    0    0    0
## 97.5%    0    0    0    0
```

This simulation shows that the 95% CI are 0 for 1000 samples which has really good performance.

CI for Diffussion Terms

```
##             [,1]      [,2] [,3] [,4]
## 2.5%  0.01665203 0.0000000    0    0
## 97.5% 1.08543059 0.1014428    0    0
```

The diffusion term for standard BM is 1. However, even the true values are within the CI, the CI for the constant term is wide. This tells us that the variance is harder to reconstruct.

References

Boninsegna et. al, Sparse learning of stochastic dynamical equations, J. Chem. Phys. 148, 241723 (2018).