On Supervised and Unsupervised Classifications of General Stochastic Volatility Models

In this project we consider supervised and unsupervised classifications for the case where each point is a continuous-time stochastic volatility model. This statistical setting is nowadays motivated by the applications to many research and industrial areas, such as finance and marketing, geology, biology and medical research, signal processing, etc. The model is often considered when only a sample of N time series (for instance, asset prices of N companies, number of daily sold items of N products in a supermarket, number of hourly visits to each of the N webpages, ect) are observed, but little is known on the patterns among these time series. Our goal is to improve the existing approaches in literature and to provide new supervised and unsupervised classification approaches for clustering stochastic processes. It mainly involves working on the models based on Cadre (2013).

Cadre (2013) studied supervised classification of Brownian diffusions $\{X_t\}_t$, solved from quite a general stochastic volatility model:

$$\mathrm{d}X_t = b(t, X_t)\,\mathrm{d}t + \sigma(t, X_t)\,\mathrm{d}B_t,$$

where b, σ are unknown smooth deterministic functions and $\{B_t\}_t$ is a standard Brownian motion. Recall that Cadre (2013) has set a problem of classifying $(\{X_t\}, Y)$ for which

$$\left\{\begin{array}{ll} \mathrm{d}X_t = b(t, X_t)\,\mathrm{d}t + \sigma(t, X_t)\,\mathrm{d}B_t & \text{if } Y = 0; \\ \mathrm{d}X_t = (b(t, X_t) + (f\sigma)(t, X_t))\,\mathrm{d}t + \sigma(t, X_t)\,\mathrm{d}B_t & \text{if } Y = 1, \end{array}\right. \tag{1}$$

where b, σ, f are unknown Borel functions. Based on an explicit computation of the Bayes rule, Cadre (2013) constructed an empirical classi?cation rule drawn from an i.i.d. sample of copies of $(\{X_t\}, Y)$ and proved that $\hat{g}$ is a consistent rule with some rate of convergence. More precisely, if one denotes the Bayes rule by g^*:

$$g^*(x) = \left\{\begin{array}{ll} 1 & \text{if } \mathbb{P}(Y = 1|X = x) > \mathbb{P}(Y = 0|X = x); \\ 0 & \text{if else.} \end{array}\right.$$

Then Cadre (2013) has constructed some estimate of g^*, denoted by $\hat{g}$, such that

$$\mathbb{E}[L(\hat{g})] - L(g^*) \leq Cn^{-u}, \text{ for some } u > 0,$$

where a constant $C > 0$ represents optimal rate of convergence, and L denotes the Bayes risk function: $L(g) = \mathbb{P}(g(X) \neq Y)$.

Note that there are at least two conveniences in the model (1):

Inconvenience 1 In practice, it is unrealistic that any observed Brownian diffusion has equal stochastic volatility σ. (Ran Zhao, it would be great if you can elaborate this inconvenience by giving an example in finance.)

Inconvenience 2 In practice, the labels Y for each observed Brownian diffusion $\{X_t\}$ are often unavailable. (Ran Zhao, it would be great if you can elaborate this inconvenience by giving an example in finance.)

The desire of remedying the above two inconveniences motivates our framework in this project. To overcome the first problem, we consider an extended model of (1), namely,

$$\begin{cases} \mathrm{d}X_t = b(t,X_t)\,\mathrm{d}t + \sigma(t,X_t)\,\mathrm{d}B_t & \text{if } Y=0; \\ \mathrm{d}X_t = (1+\epsilon_1(t,X_t))b(t,X_t)\,\mathrm{d}t + (1+\epsilon_2(t,X_t))\sigma(t,X_t)\,\mathrm{d}B_t & \text{if } Y=1, \end{cases} \tag{2}$$

where $b, \sigma, \epsilon_1, \epsilon_2$ are unknown Borel functions. we would provide a new supervised classification algorithm which is nonparameteric and has a satisfying rate of convergence (Guangliang, would you expand this part to suggest some approaches in lieu of the naïve Bayes rule?).

To overcome Inconvenience 2 to make the learning approach more realistic, we will also study the unsupervised classification problem of (2). Recall that Khaleghi et al. (2016) has suggested a nonparametric consistent algorithm to cluster time series into $k \geq 2$ groups. The crucial idea is to propose a convenient similarity measure, which reflects the difference between the underlying distributions of time series. However, their algorithm only works for stationary ergodic time series. The framework on clustering martingales is still open. Hence our second main result will devote to designing an algorithm to cluster a continuous-time martingale, using an idea based on Khaleghi et al. (2016).

Our team consists of Guangliang Chen (San José State University), Xuemei Cheng (University of San Francisco), Qidi Peng (Claremont Graduate Unviersity), Yi Wang (Syracuse University) and Ran Zhao (American International Group). Guangliang Chen, Xuemei Cheng and Yi Wang's expertise in clustering analysis will help to solve the martingale-clustering problem; Qidi Peng, whose research field is stochastic calculus, can help to theoretically build up consistency of the algorithms. Ran Zhao, an expert of financial engineering, can help to motivate our model and approaches from an industrial point of view. He will also provide real data source for an empirical test of our algorithms. (My friends, the last paragraph is just a sample, so please feel free to modify them).

References

[1] Cadre B., Supervised classification of Diffusion Paths, Mathematical Methods of Statistics, Vol. 22, No. 3, pp. 213-225, (2013).

[2] Khaleghi A., Ryabko D., Mary J. and Preux P., Consistent algorithms for clustering time series, Journal of Machine Learning Research 17, pp. 1-32, (2016).