High-dimensional probability estimation for empirical data

January 15, 2021

1 Finite-dimensional marginal distribution

The finite-dimensional distribution of X are the joint distribution of $X_{t_1}, X_{t_2}, \ldots X_{t_n}, t_1, \ldots, t_n \in T, n \in \mathbb{N}$ It determines whether two process has the same distribution.

Theorem 1.1. *Let X_1 and X_2 be two value-process from T with paths in U. Then, X_1 and X_2 have the same distribution iff all their finite-dimensional distributions agree.*

A theorem due to Kolmogorov extension theorem [1], established the existence of unique finite-dimensional marginals.

Theorem 1.2. *Consider a family of probability measures*

$$\{P_{t_1}, \ldots P_{t_n}, t_1, \ldots, t_n \in T, n \in \mathbb{N}\}$$

such that: $P_{t_1}, \ldots P_{t_n}$ is the probability of $\mathbb{R}^n$ and if

$$\{t_{k_1} < \cdots < t_{k_m}\} \subset t_1 < \cdots < t_n$$

then $P_{t_{k_1}, \ldots, t_{k_n}}$ is the marginal density of $P_{t_1, \ldots, t_n}$, corresponding to the indexes $k_1, \ldots, k_m$. Then, there exists a unique probability on F which has the family as $\{P_{t_1, \ldots, t_n}\}$ as finite-dimensional marginal densities.

Studying the joint distribution directly in the observed space presents computational difficulty. Instead, we will utilize Theorem 1.2 to determine the two processes by marginal density. This simplifies the multivariate problem into a univariate one.

2 Approach

2.1 Map high-dimensional empirical data to Dirichlet space

Since the finite-dimensional distributions of a stochastic process form a projective family, we applied the method [2] to construct bijective transformation. We aim to map empirical output into simple latent distributions space $Y \subset \mathbb{R}^n$ to explore the structure in high-dimensional data sets. The transformation of data is approximately factorized and has identical and known marginal densities.

To build the bijective map, We choose sigmoid function to be the invertible function g

$$\sigma(x) = \frac{1}{1 + e^{-x}}, x \in X$$

and applied the function for the final output. The invertible function map

$$g : X \to Y \tag{1}$$

such that the projective data under

$$g^{-1} : Y \to X \tag{2}$$

analytically available.

2.1.1 Compute marginal density distribution for each dimension

Since the projective space is a unit hypercube $Y = [0,1]^n$, we chose our objective distribution to be Dirichlet distribution. Given the data projection $\{y_n\}_{n=1}^N$, the marginal density of Dirichlet distribution can be computed by Beta distribution:

$$q(y_n;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y_n^{\alpha-1}(1-y_n)^{\beta-1} \tag{3}$$

The parameters of Beta distribution for each dimension is estimated by

$$\begin{aligned} \hat{\alpha_n} &= \hat{\mu_n}\left[\frac{\hat{\mu_n}(1-\hat{\mu_n})}{\hat{\sigma_n^2}} - 1\right] \\ \hat{\beta_k} &= (1-\hat{\mu_n})\left[\frac{\hat{\mu_n}(1-\hat{\mu_n})}{\hat{\sigma_n^2}} - 1\right] \end{aligned} \tag{4}$$

where $\hat{\mu_n}$ and $\hat{\sigma_n^2}$ is the sample mean and sample variance.

2.2 Measure theoritical marginal distribution density

The distribution of a Gaussian random n-vector X is uniquely determined by its mean vector μ and covariance matrix Γ. Assume Γ is invertible, the joint probability density can be computed by

$$f(x_1,\ldots,x_n) = f(\overrightarrow{x}) = \frac{1}{\sqrt{2\pi}^n\sqrt{\det(\Gamma)}}\exp\left(-\frac{1}{2}(\overrightarrow{x}-\overrightarrow{\mu})^T\Gamma^{-1}(\overrightarrow{x}-\overrightarrow{\mu})\right) \tag{5}$$

the term on the right are the marginal distribution of X_i's.

$$f(x_1,\ldots,x_n) = f_{x_1}(x_1)\ldots f_{x_n}(x_n) \tag{6}$$

The univariate marginal distribution for x_n is

$$f_{x_n}(x_n) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x_1,\ldots,x_n)dx_1\ldots d_{x_{n-1}} \tag{7}$$

Since the multivariate marginal density (7) requires the indenpendent data, we implemented two steps to turn a geometric Brownian motion into standard Brownian motion. First, transform the geometric Brownian motion into the increments of independent sequence X by

$$X_t = \log\left(\frac{S_t}{S_0}\right) = \log(S_t) - \log(S_0) \tag{8}$$

Then, standardize (8) by

$$W_t = \frac{X_t - rt + \frac{\sigma^2 t}{2}}{\sigma} \tag{9}$$

where r is the risk free interest rate and volatility σ.

After we get the marginal distribution function from (7), we computed the multi-intergration (6) by Monte Carlo method.

3 Result

After obtaining the empirical results, we selected the data from each dimension to fill in the formula (3) and (6), (5) to calculate the marginal density of the point for each dimension. Then, took the average of the Root mean square error(RMSE) for each dimension.

The RMSE for each dimension form 1 to 10 is

Dimension	1	2	3	4	5
RMSE	2.99e-261	2.19e-229	1.928e-278	0.00e+000	0.00e+000

Dimension	6	7	8	9	10
RMSE	0.00e+000	3.61e-320	0.00e+000	9.68e-317	3.16e-322

The average of it is 0.003.

Therefore, since the RMSE of marginal density at each dimension of two processes is extremely low, it showed that the two distribution comes from same distribution.

References

[1] Cosma Rohilla Shalizi. Almost none of the theory of stochastic processes, 2010.

[2] Oren Rippel and Ryan Prescott Adams. High-dimensional probability estimation with deep density models, 2013.