Yule-Simon Processes

Asher A. Hensley, Ph.D.

Outline

- Background
- Application
- Experiments

Background

Some History…

George Udny **Yule**

Herbert **Simon**

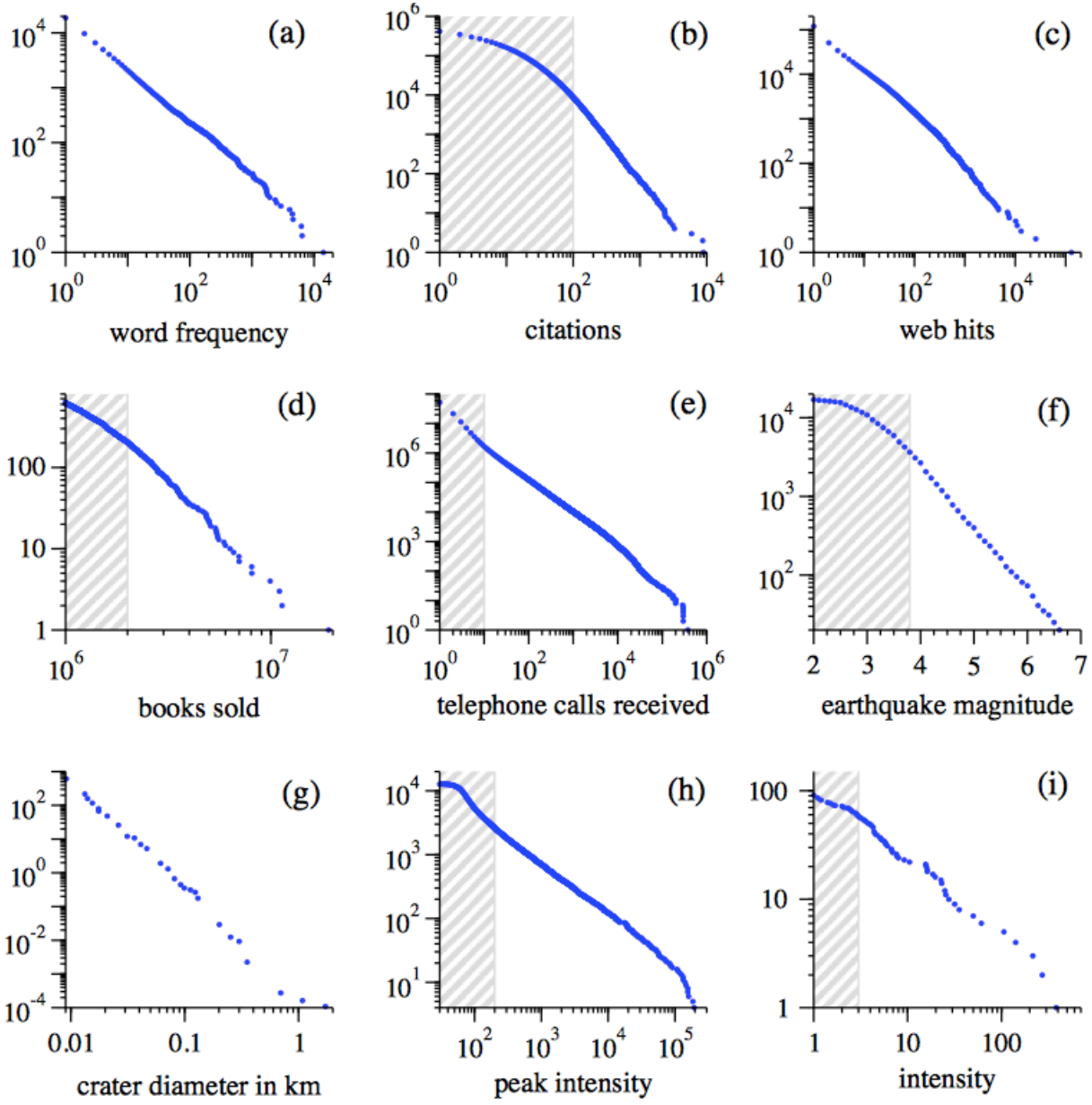
- Developed one of the first models to explain power-law occurrence.
- **Concept:** Success leads to future success!

Power Laws:

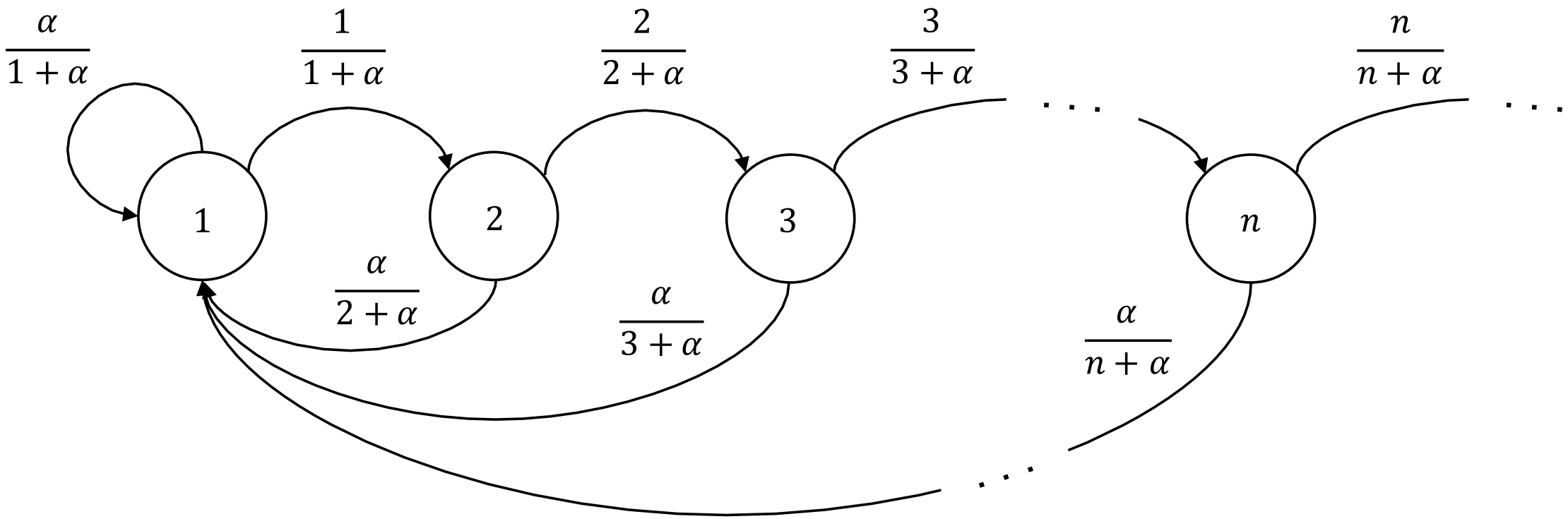
$$p(x|\alpha) = Cx^{-\alpha}$$

	quantity	minimum $x_{\min}$	exponent α
(a)	frequency of use of words	1	2.20(1)
(b)	number of citations to papers	100	3.04(2)
(c)	number of hits on web sites	1	2.40(1)
(d)	copies of books sold in the US	2 000 000	3.51(16)
(e)	telephone calls received	10	2.22(1)
(f)	magnitude of earthquakes	3.8	3.04(4)
(g)	diameter of moon craters	0.01	3.14(5)
(h)	intensity of solar flares	200	1.83(2)
(i)	intensity of wars	3	1.80(9)

Source:
Newman, Mark EJ. "Power laws, Pareto distributions and Zipf's law." *Contemporary physics* 46.5 (2005): 323-351.



The Yule-Simon Process:



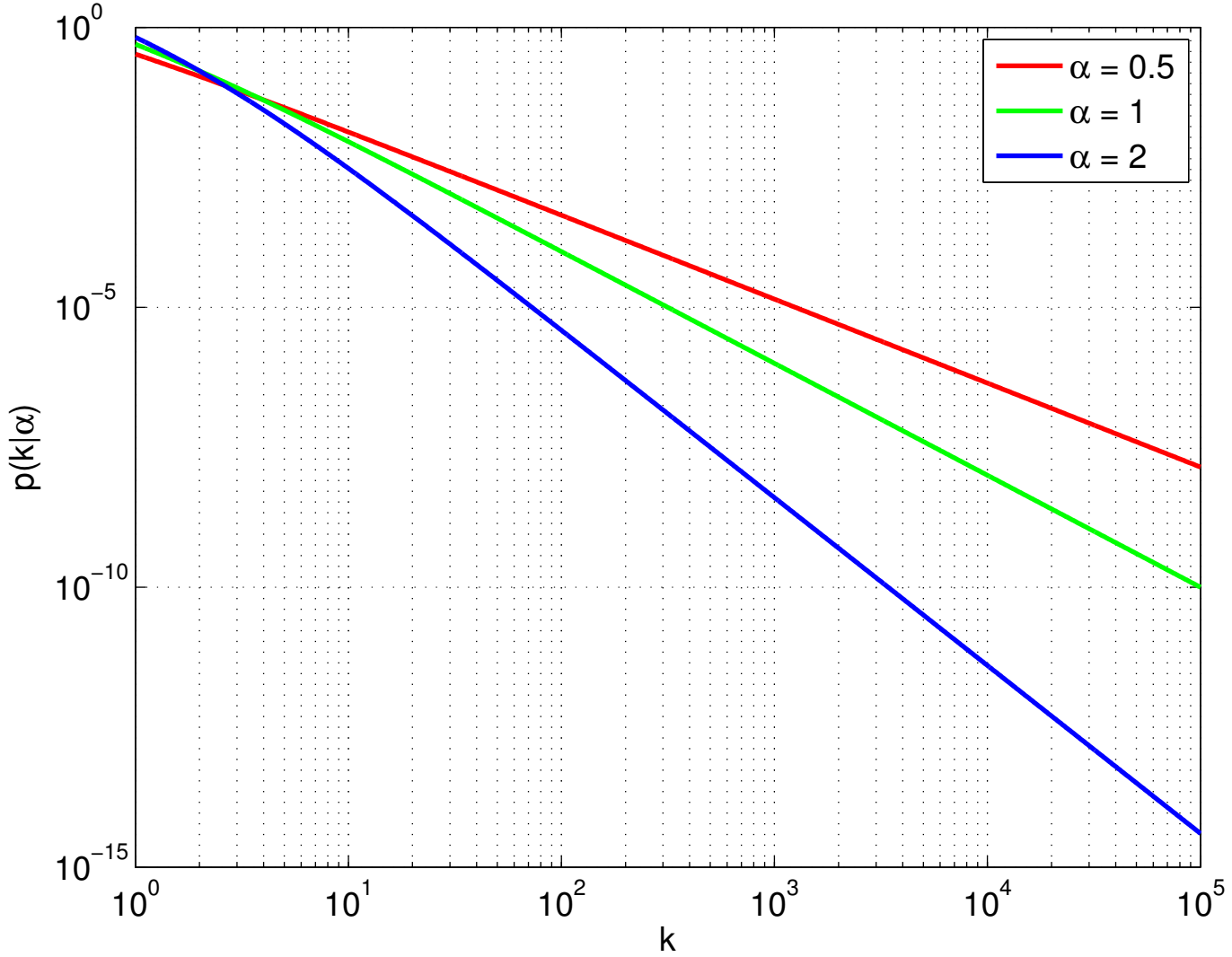
Example:

$$p(k = n|\alpha) = \frac{\alpha}{n+\alpha}\prod_{i=1}^{n-1}\frac{i}{i+\alpha}$$

$$= \frac{\alpha(n-1)!}{(n+\alpha)(n-1+\alpha)\cdots(1+\alpha)}$$

$$= \frac{\alpha\Gamma(n)\Gamma(\alpha+1)}{\Gamma(n+\alpha+1)} = \boxed{\alpha B(n, \alpha+1)}$$

Yule-Simon Distribution: $p(k|\alpha) = \alpha B(k, \alpha + 1)$



Yule-Simon Counting Process:

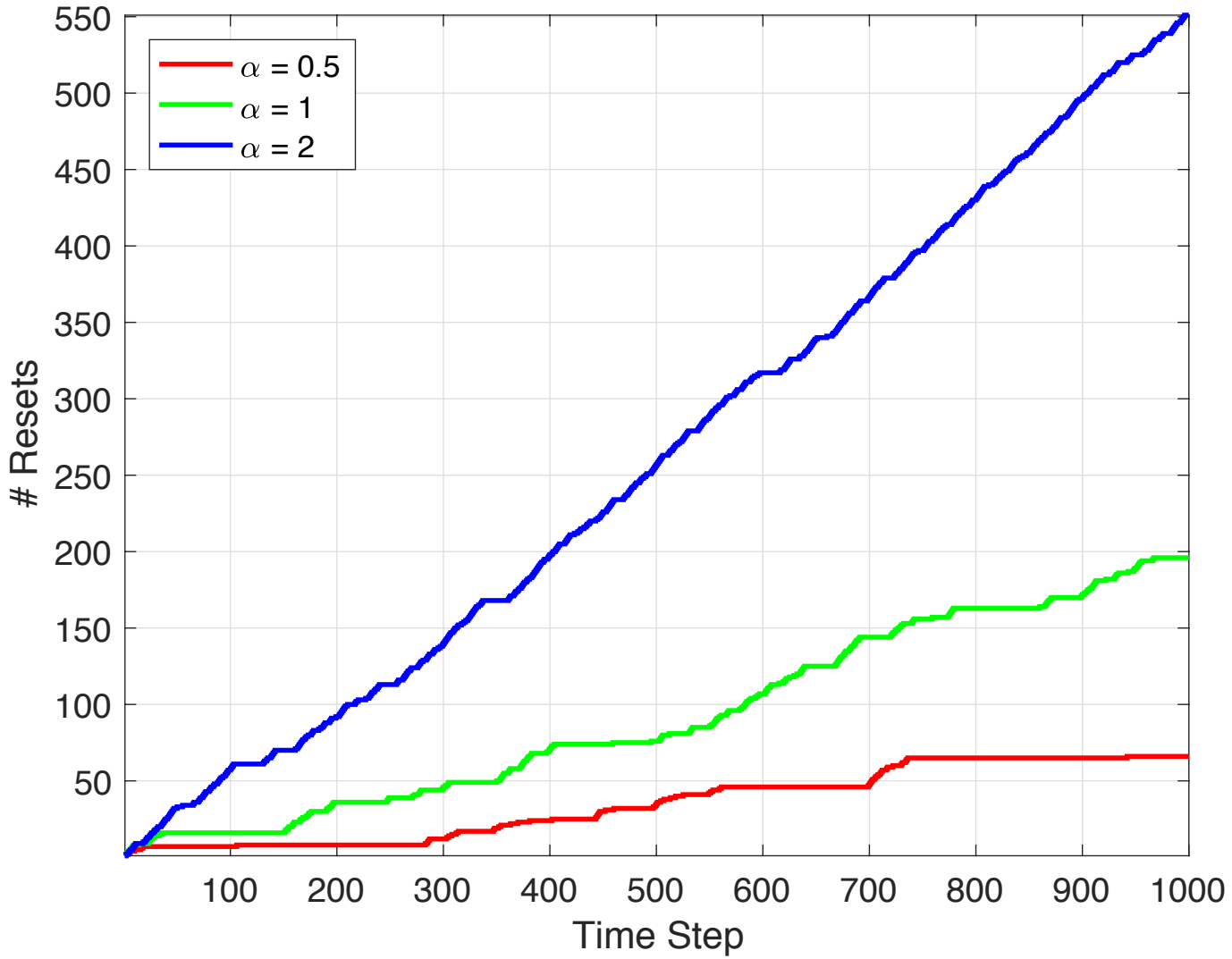
$$s_t \mid \alpha, \mathbf{x}_{1:t-1} \sim \text{Bernoulli}\left(\alpha\left(n_j + \alpha\right)^{-1}\right)$$

$$x_t = x_{t-1} + s_t$$

where, $\mathbf{x}_{1:t-1} = [x_1, \ldots, x_{t-1}]$

$$n_j = \#(\mathbf{x}_{1:t-1} == x_{t-1})$$

Example:


α = 0.5
α = 1
α = 2
Resets
Time Step
550
500
450
400
350
300
250
200
150
100
50
100
200
300
400
500
600
700
800
900
1000

Basic Idea(s):

- Power-law distributed waiting times
- Event clustering
- Long (or infinite) run-lengths

Application

Concept:

- Use Yule-Simon process as a prior for **volatility cluster*** formation

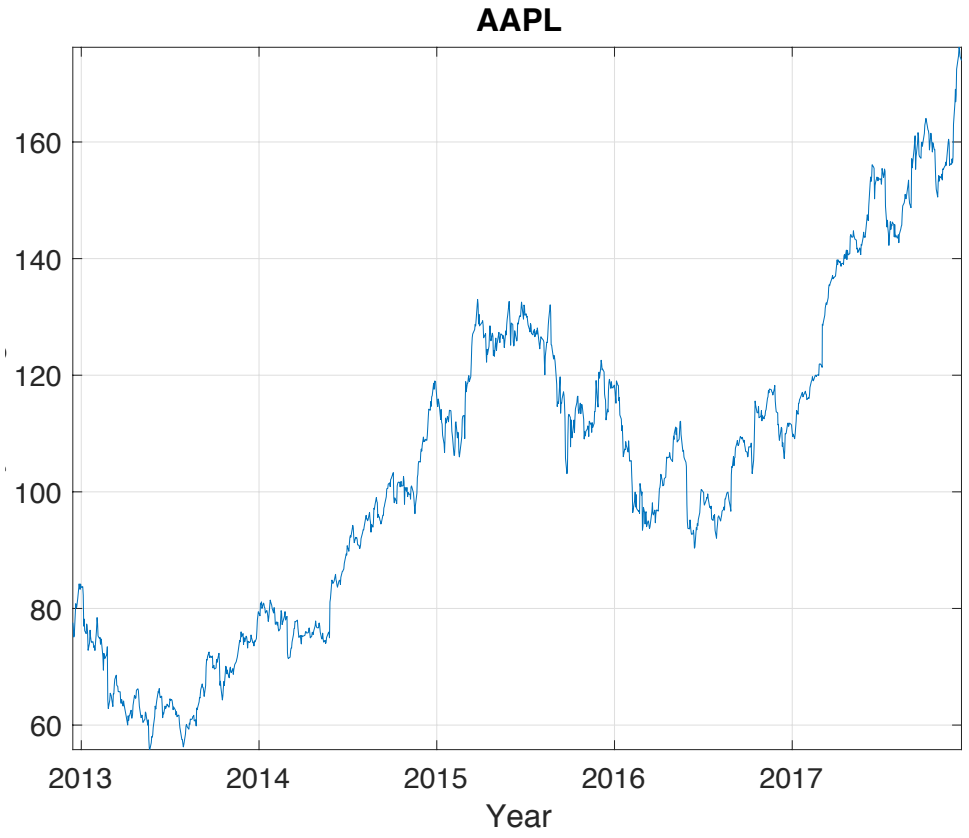
Concept:

- Use Yule-Simon process as a prior for **volatility cluster*** formation

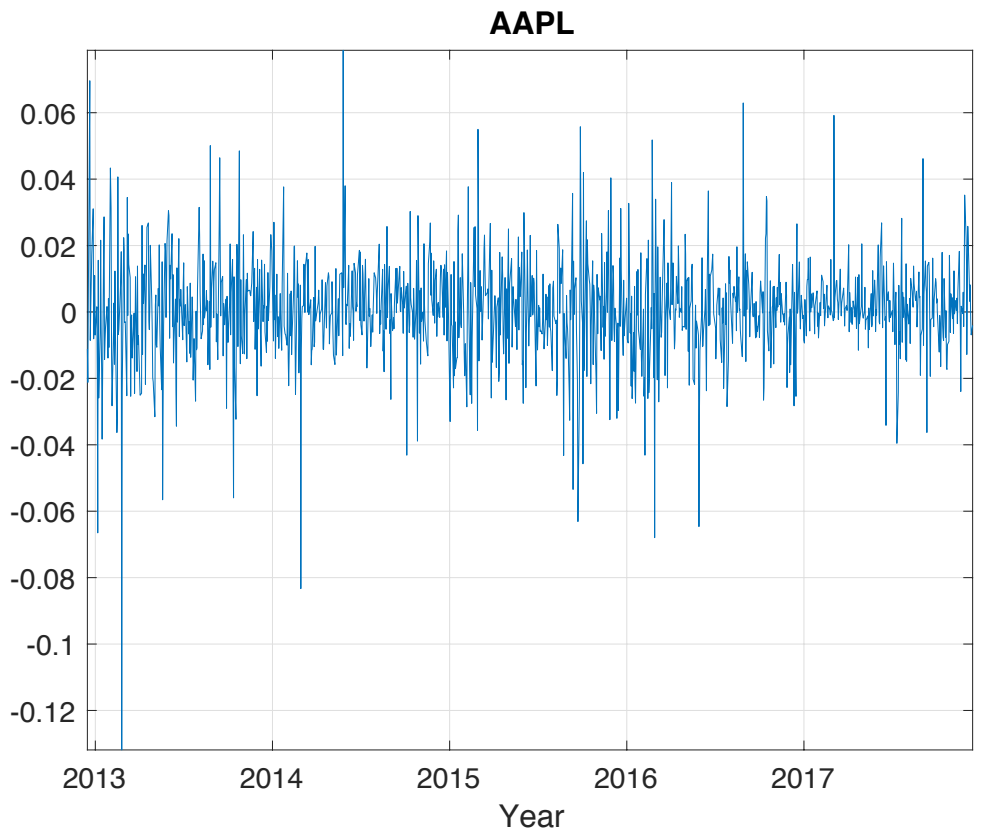
***volatility cluster:** segment of log return time series with constant volatility.

Log Returns: $r_t = \log P_t - \log P_{t-1}$

Daily Closing Price



Daily Log-Return



Generative Model:

$$s_t \mid \alpha, \mathbf{x}_{1:t-1} \sim \text{Bernoulli}\left(\alpha\left(n_j + \alpha\right)^{-1}\right)$$

$$x_t = x_{t-1} + s_t$$

Yule-Simon Counting Proc.

$$w_t \mid x_t, \boldsymbol{\lambda}_{1:\infty} \sim \text{Normal}\left(0, \lambda_{x_t}^{-1}\right)$$

$$\lambda_j \mid c, d \sim \text{Gamma}(c, d)$$

$$r_t = \mu_t + w_t$$ (Observed log return sequence)

$$\mu_t = \mu_{t-1} + v_t$$

$$v_t \mid \phi \sim \text{Normal}(0, \phi)$$

Hyper-Priors:

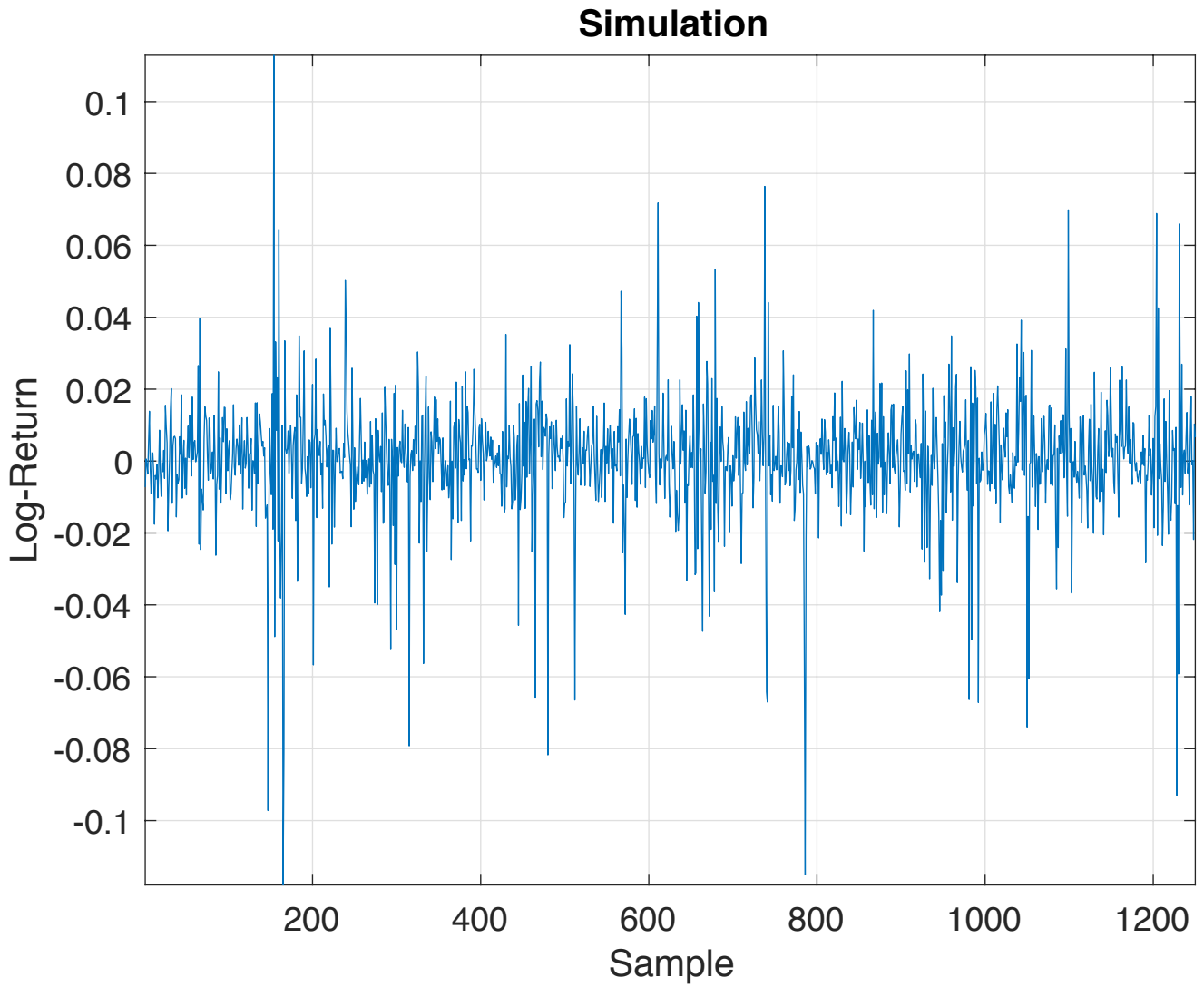
$$\alpha | a, b \sim \text{Gamma}(a, b)$$

$$c, d | p, q, t, s \sim \pi(p, q, t, s)$$

where,

$$\pi(p, q, t, s) \propto \frac{p^{c-1} e^{-dq}}{\Gamma(c)^t d^{-cs}}$$

Example:


Simulation
Log-Return
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
200
400
600
800
1000
1200
Sample

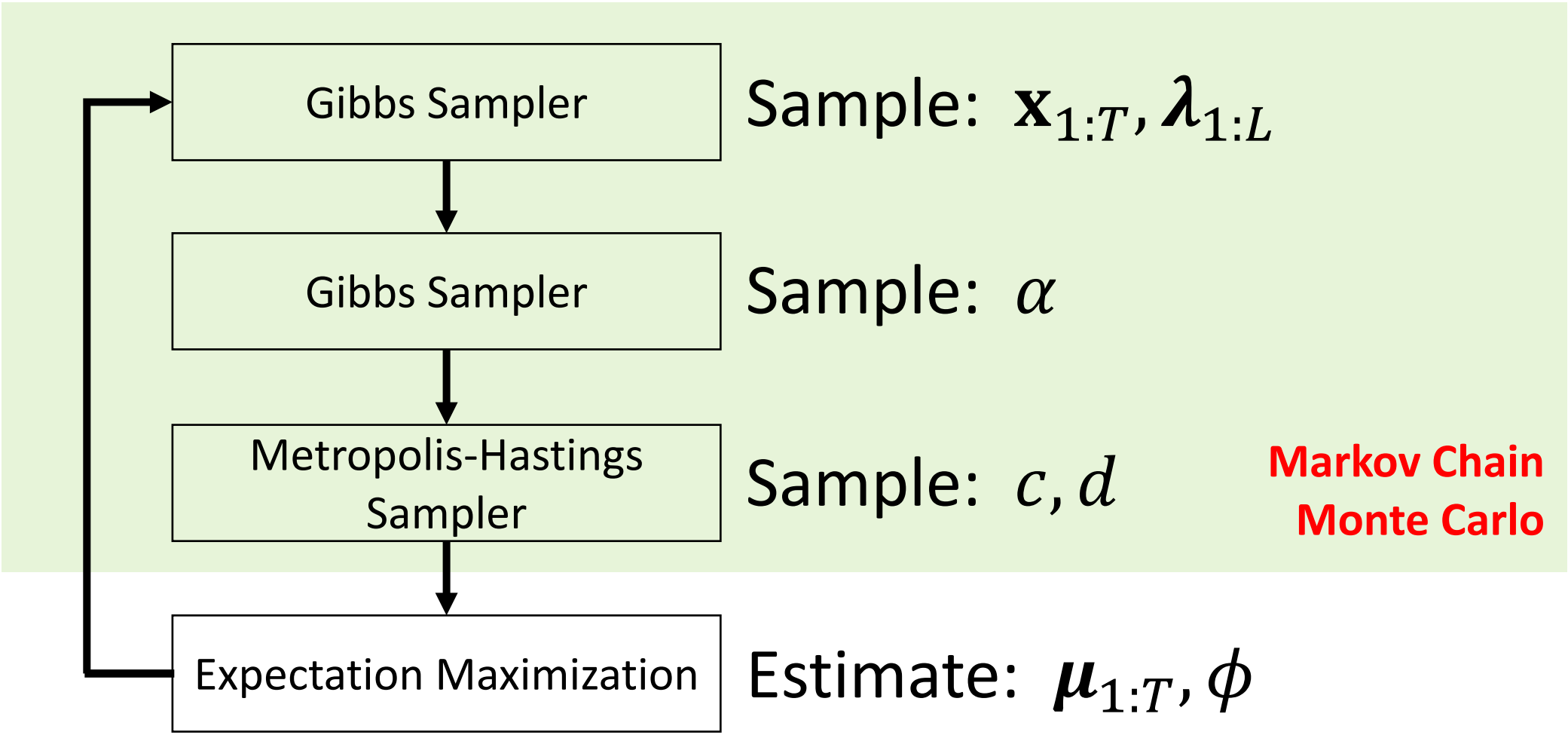
Problem:

Given: $\mathbf{r}_{1:T} = [r_1, \ldots, r_{T-1}, r_T]$ (i.e. the "Data")

Find: $p(\mathbf{x}_{1:T}, \boldsymbol{\lambda}_{1:L}, \boldsymbol{\mu}_{1:T}, \alpha, c, d | a, b, p, q, t, s, \phi, \mathbf{r}_{1:T})$

Model Posterior | "Data"

Approach:


Gibbs Sampler
Sample: $\mathbf{x}_{1:T}, \boldsymbol{\lambda}_{1:L}$
Gibbs Sampler
Sample: α
Metropolis-Hastings Sampler
Sample: c, d
Markov Chain Monte Carlo
Expectation Maximization
Estimate: $\boldsymbol{\mu}_{1:T}, \phi$

Markov Chain Monte Carlo (MCMC)

- Create Markov chain with same target distribution as the posterior
- Simulate Markov chain to draw samples

Expectation Maximization

- **E-Step:** Kalman Smoother to estimate $\boldsymbol{\mu}_{1:T}$
- **M-Step:** Maximum Likelihood to estimate ϕ

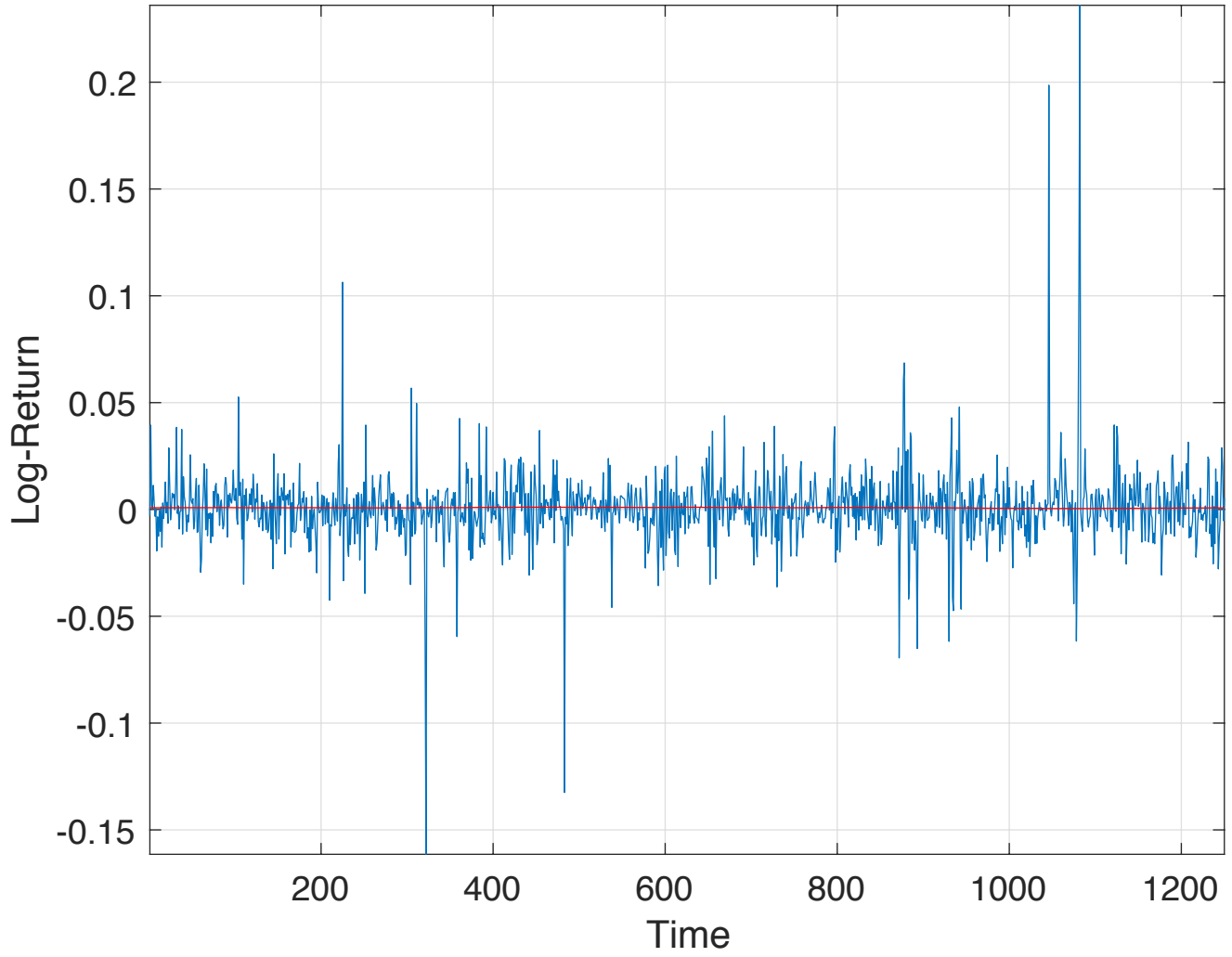
Experiments

E1: Simulation

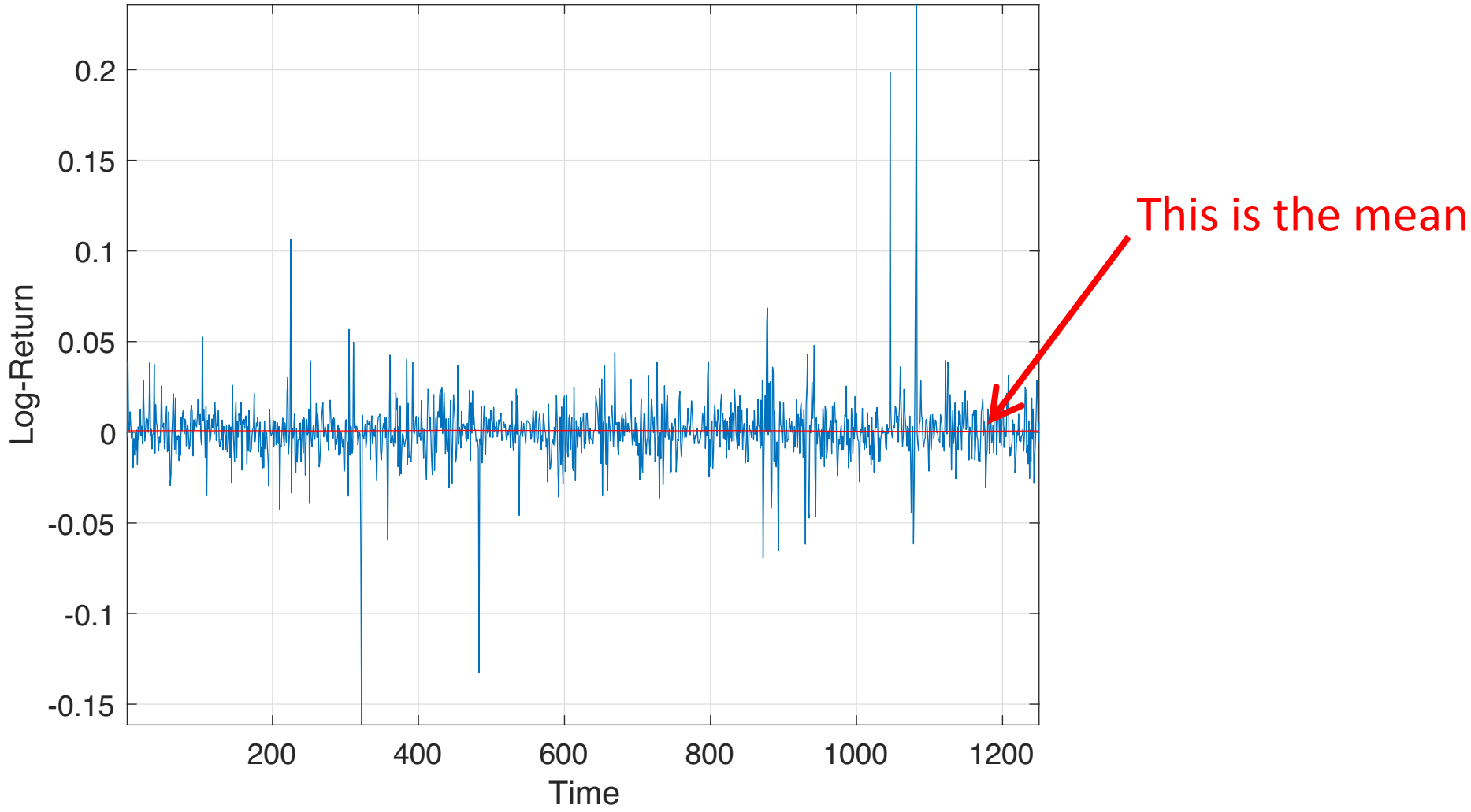
Strategy:

- Run Generative Model with Known Parameters
- Run Inference Algorithm
- Evaluate Performance

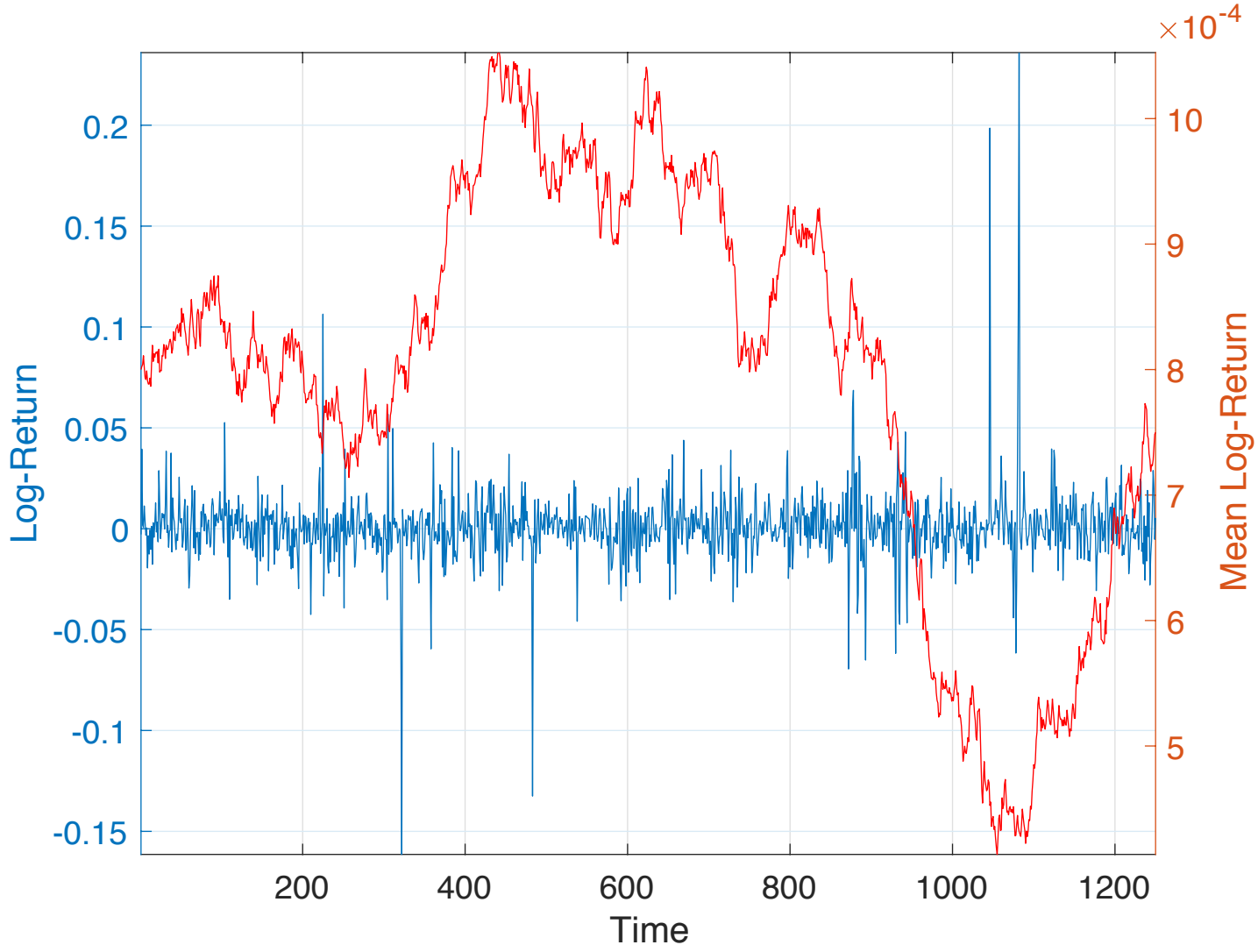
Simulated Log-Returns:



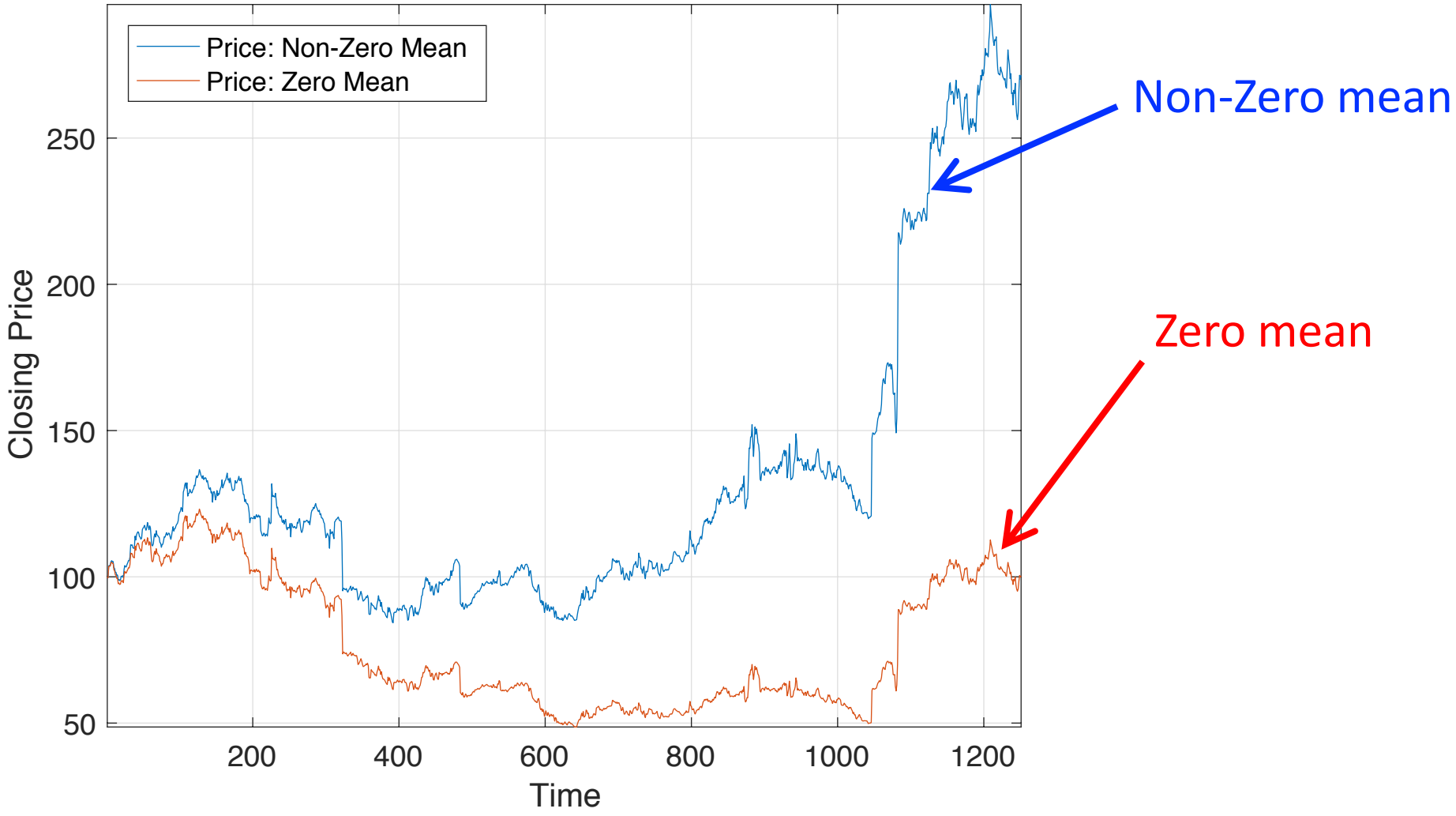
Simulated Log-Returns:



Simulated Log-Returns: $SNR \cong -67dB$



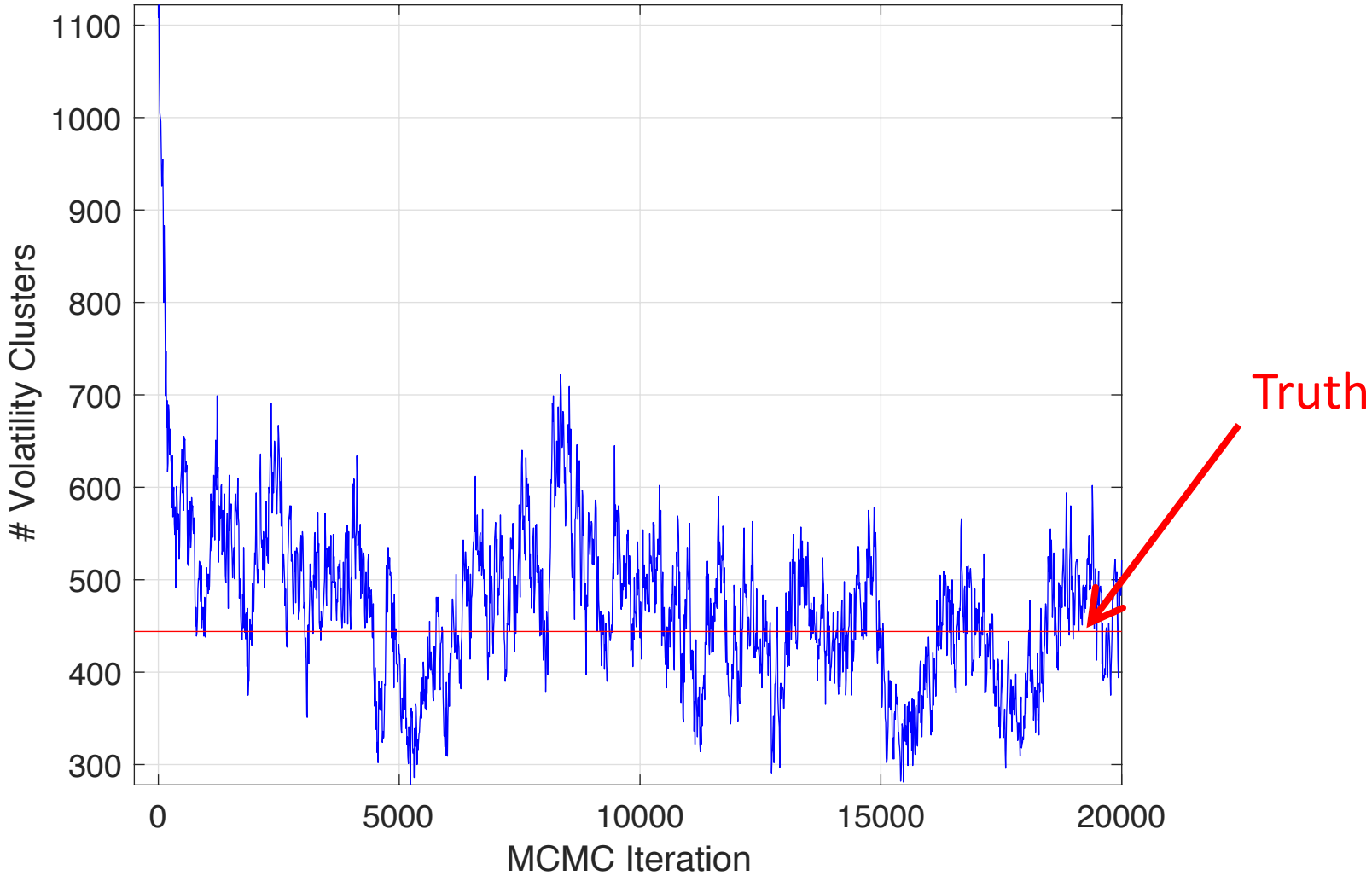
Why the Mean Matters:



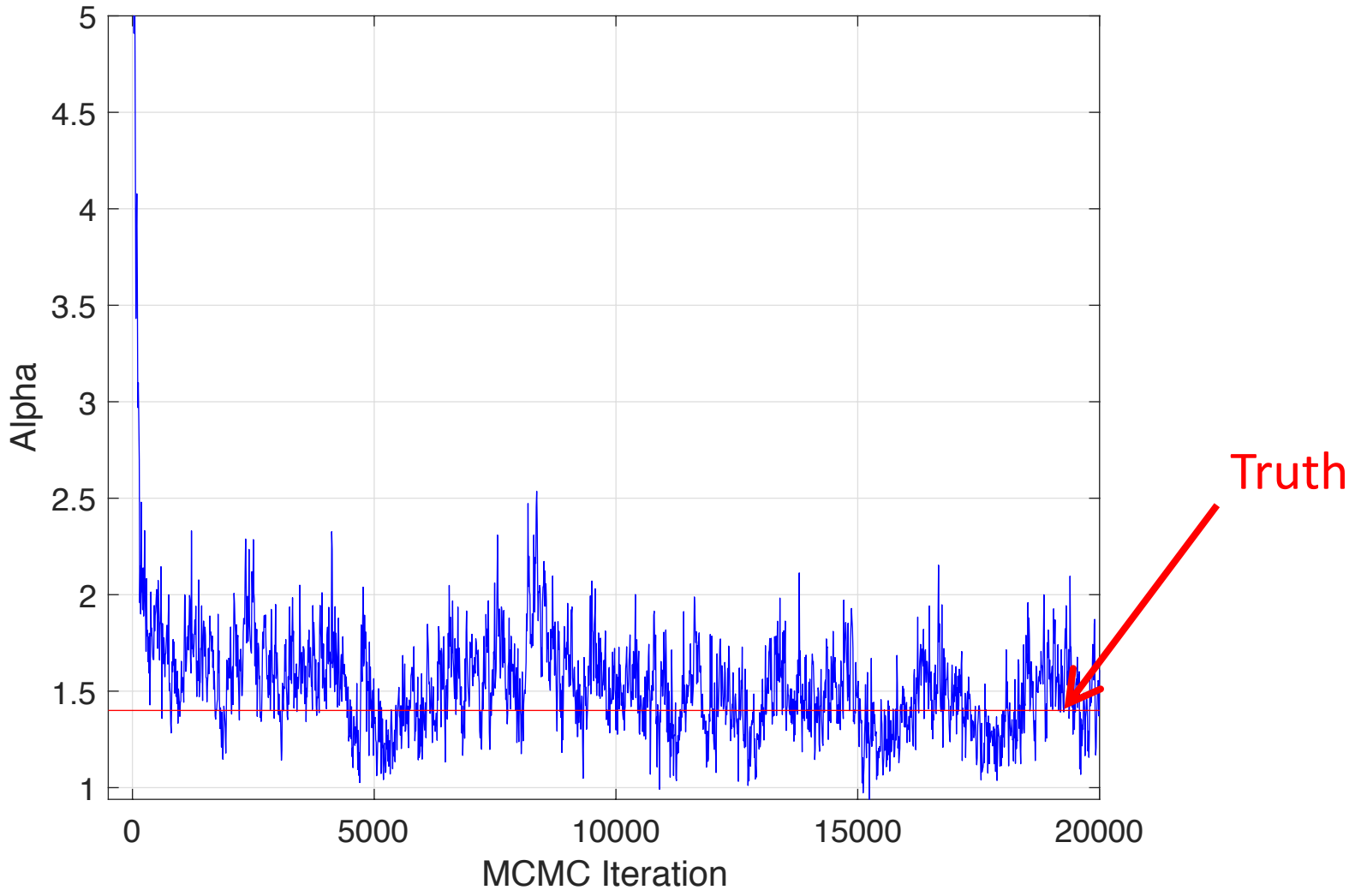
Evaluation:

- MCMC Sampler Convergence
- Model Adequacy Checking
- Mean Estimation Accuracy

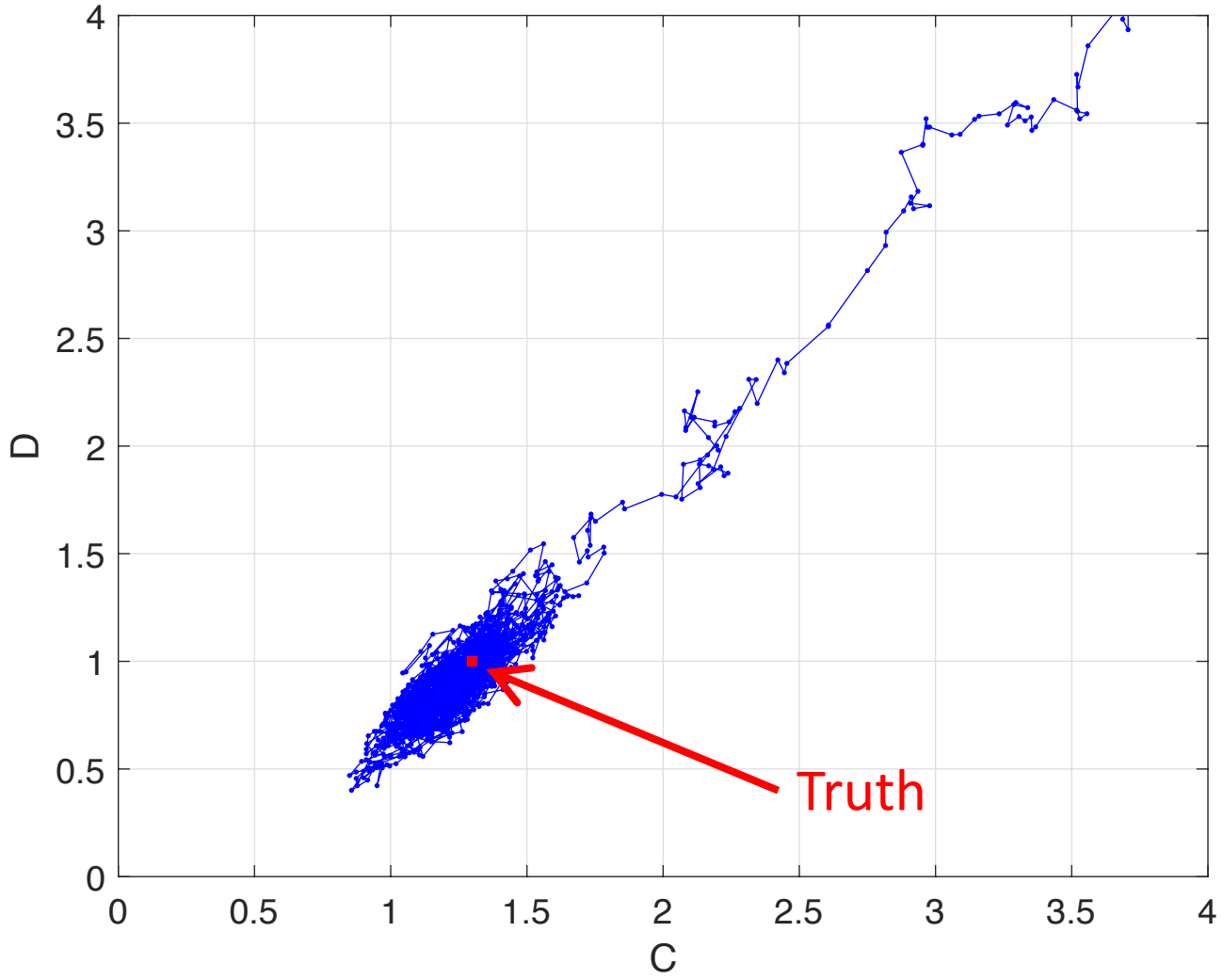
Gibbs Sampler 1: # Volatility Clusters



Gibbs Sampler 2: Yule-Simon Parameter



Metropolis-Hastings: Volatility Parameters



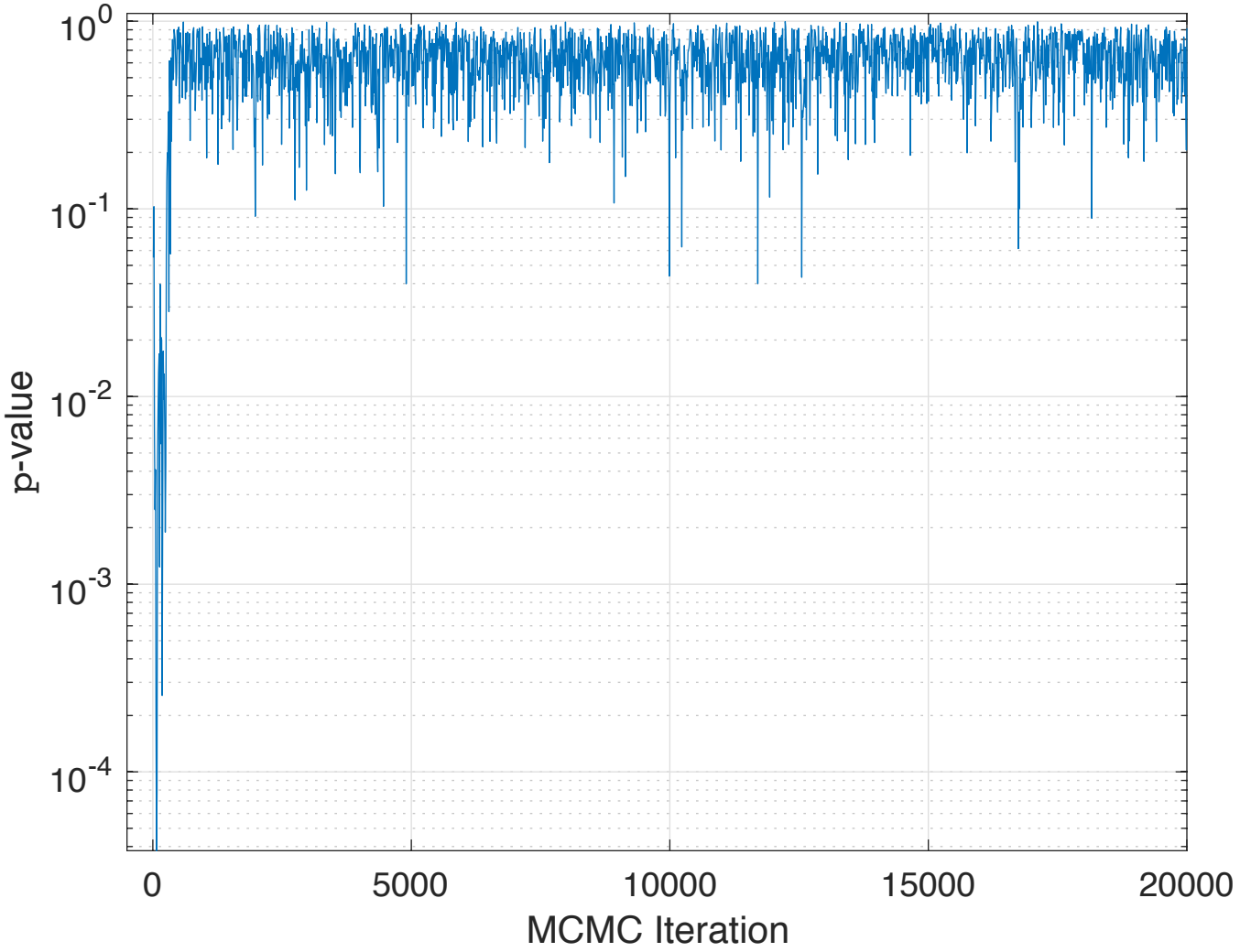
Model Adequacy Checking:

- For Each Iteration:
  - Standardize Log-Returns:

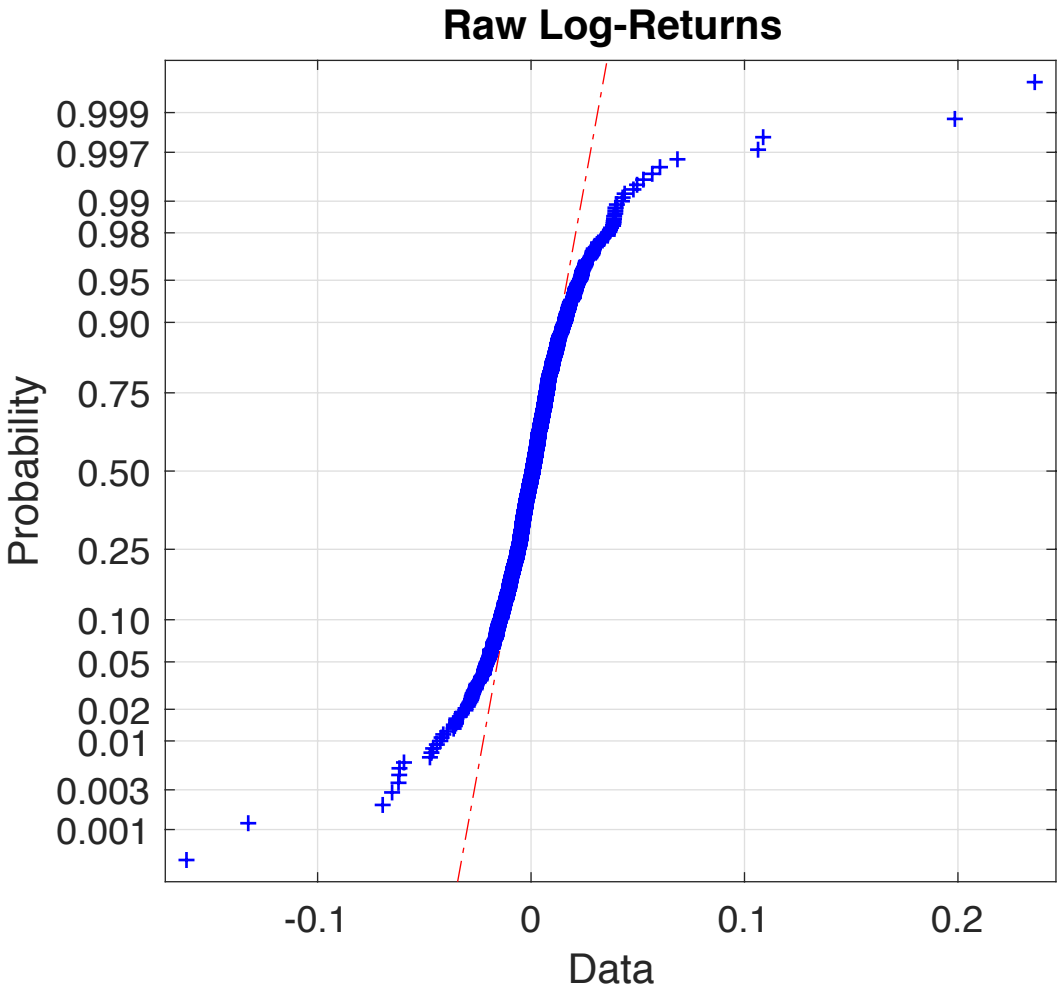
$$z_t = \frac{r_t - \mu_t}{\sigma_t} = (r_t - \mu_t)\lambda_{x_t}^{1/2}$$

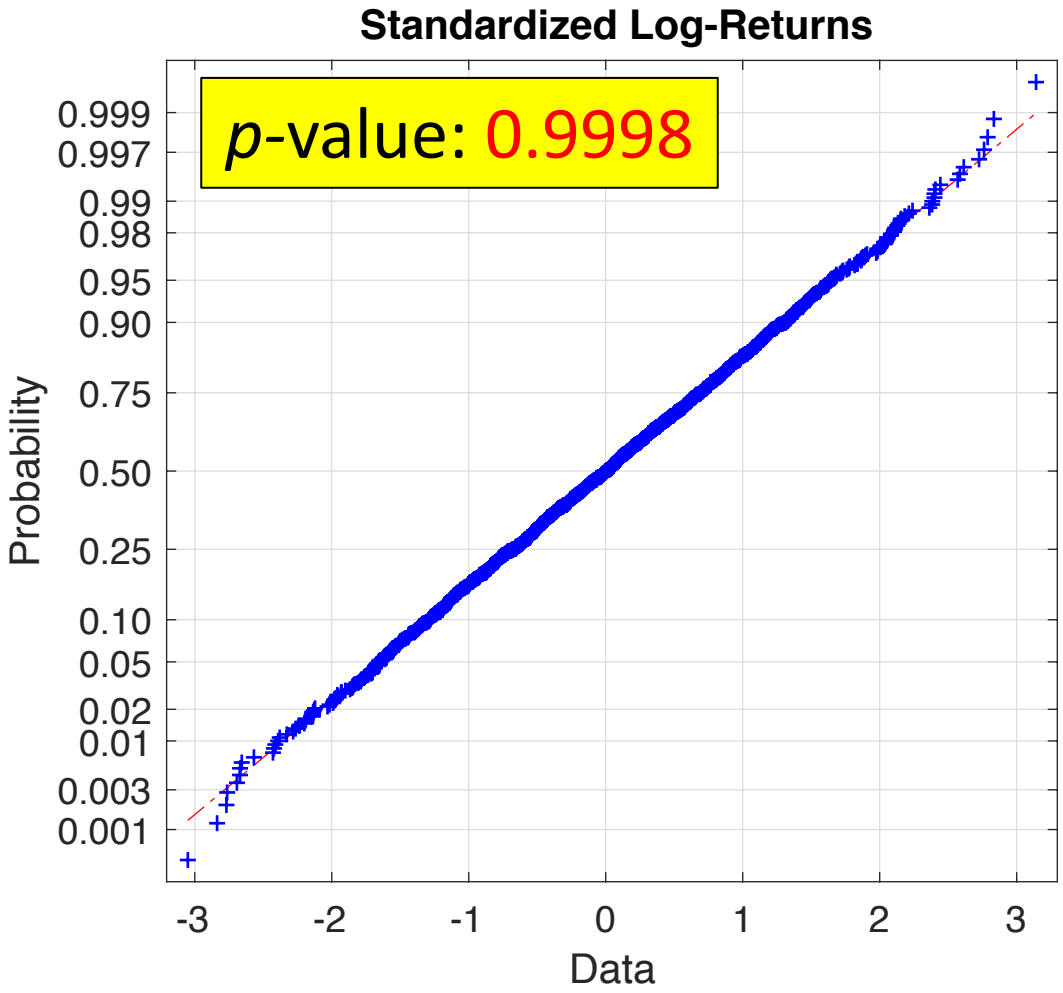
  - Test for Normality: χ^2-Test
- Evaluate Iteration w/ Largest p-value

Model Adequacy Checking: *p*-values

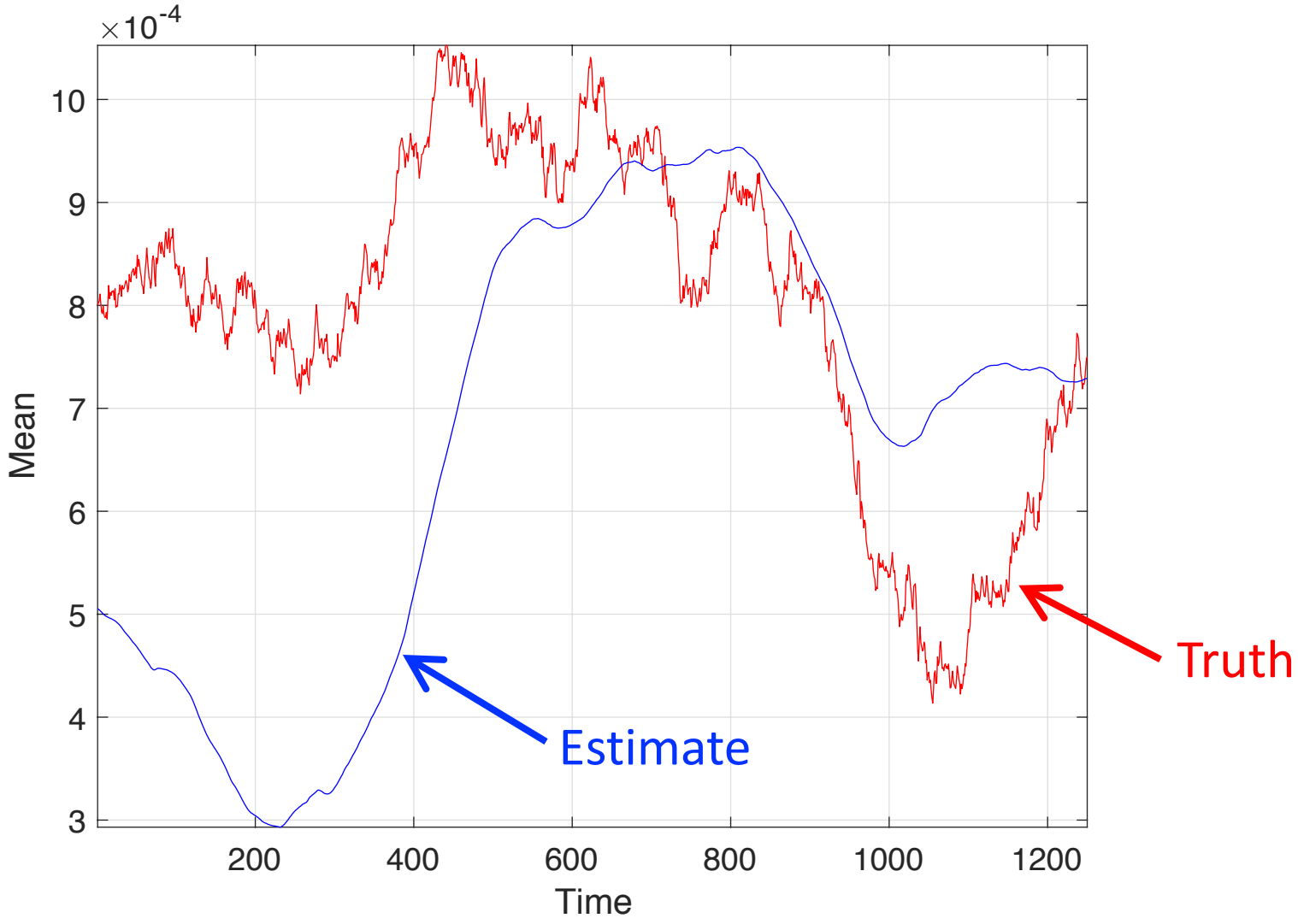


Model Adequacy Checking:





Mean Estimation:



E2: Dow Jones 30

Data Set: Dow Jones 30

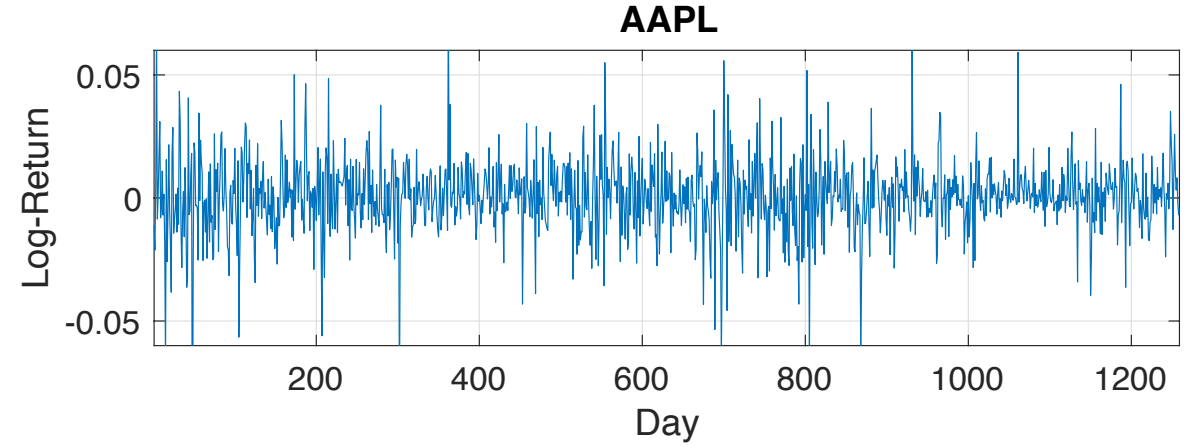
- Daily Closing Prices: 11/13/2012 – 11/13/2017

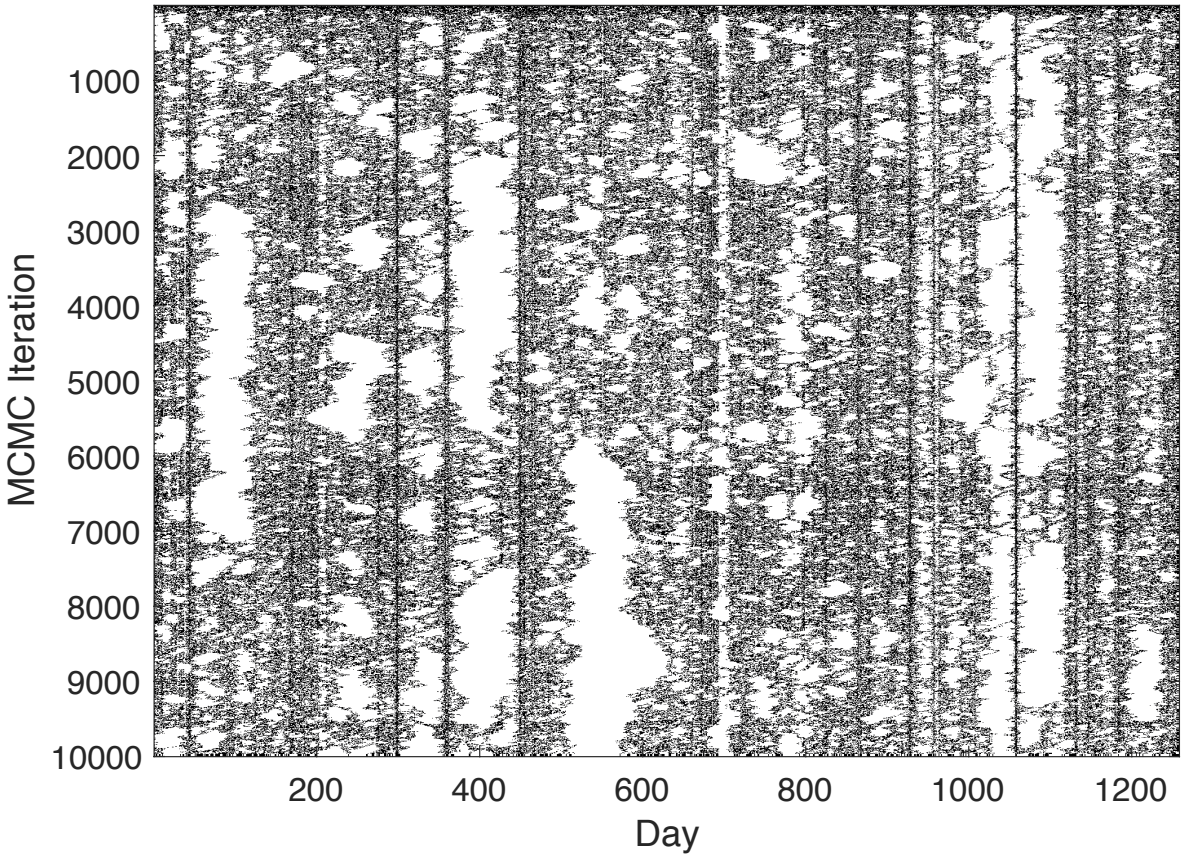
AAPL	Apple	**MCD**	McDonalds
AXP	American Express	**MMM**	3M
BA	Boeing	**MRK**	Merck
CAT	Caterpillar	**MSFT**	Microsoft
CSCO	Cisco	**NKE**	Nike
CVX	Chevron	**PFE**	Pfizer
DIS	Disney	**PG**	Procter & Gamble
DWDP	DowDuPont Inc	**TRV**	Travelers Companies Inc.
GS	Goldman Sachs	**UNH**	United Health
HD	Home Depot	**UTX**	United Technologies
IBM	IBM	**V**	Visa
INTC	Intel	**VZ**	Verizon
JNJ	Johnson & Johnson	**WBA**	Walgreen
JPM	JPMorgan Chase	**WMT**	Wal-Mart
KO	Coca Cola	**XOM**	Exxon Mobil

Analysis

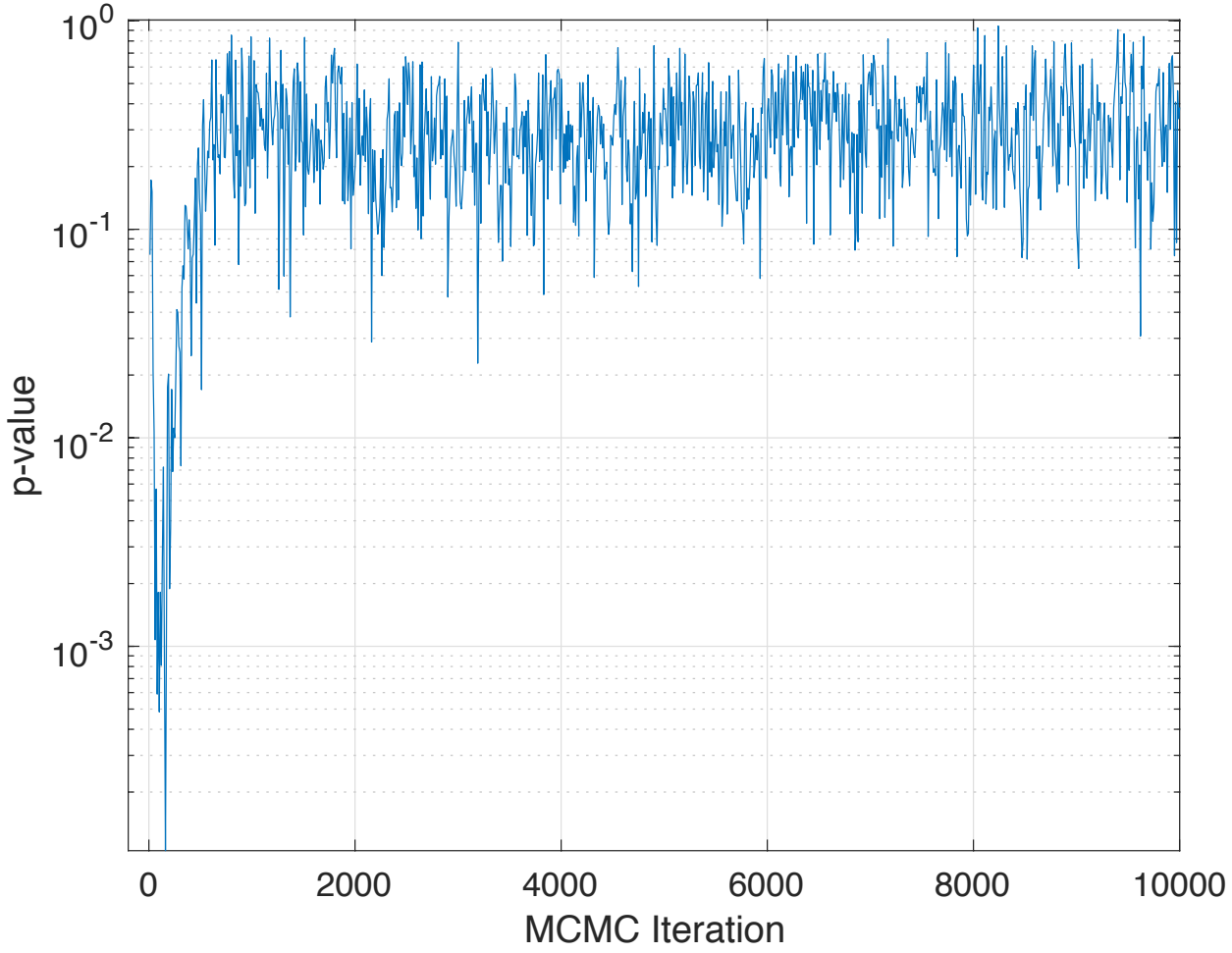
- **Case Study:** AAPL
- Ensemble Results

Case Study:
AAPL

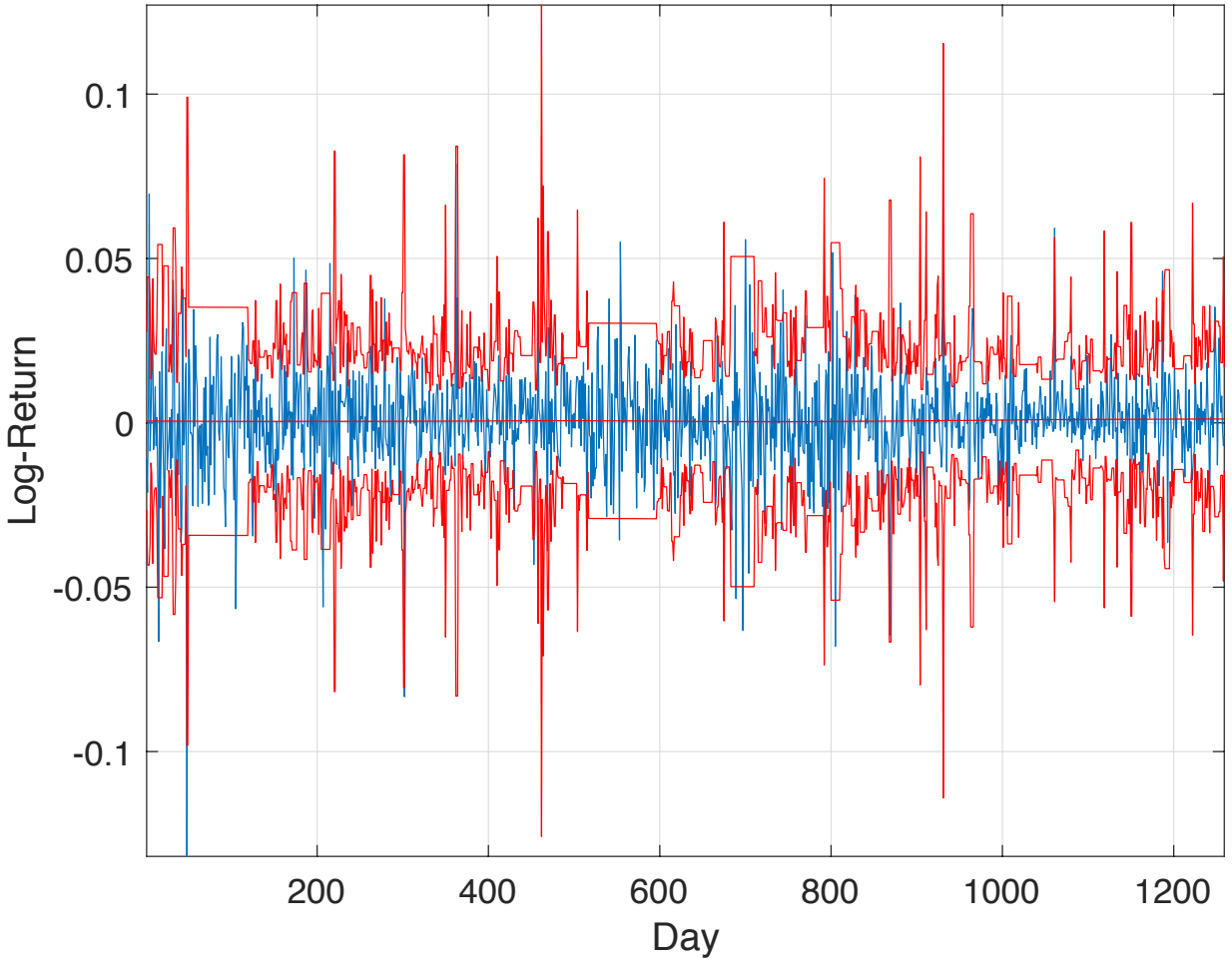




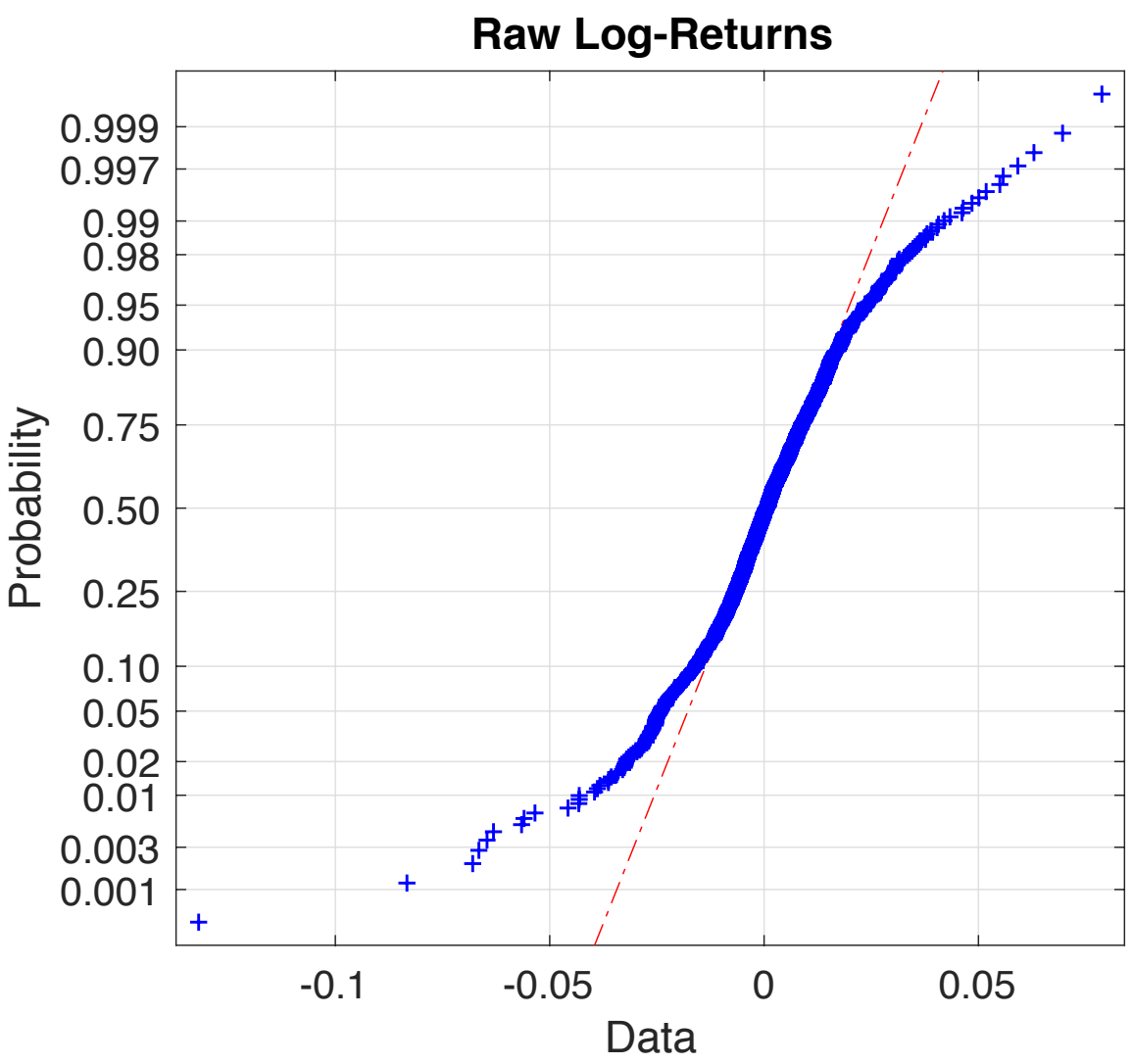
Model Adequacy Checking: *p*-values

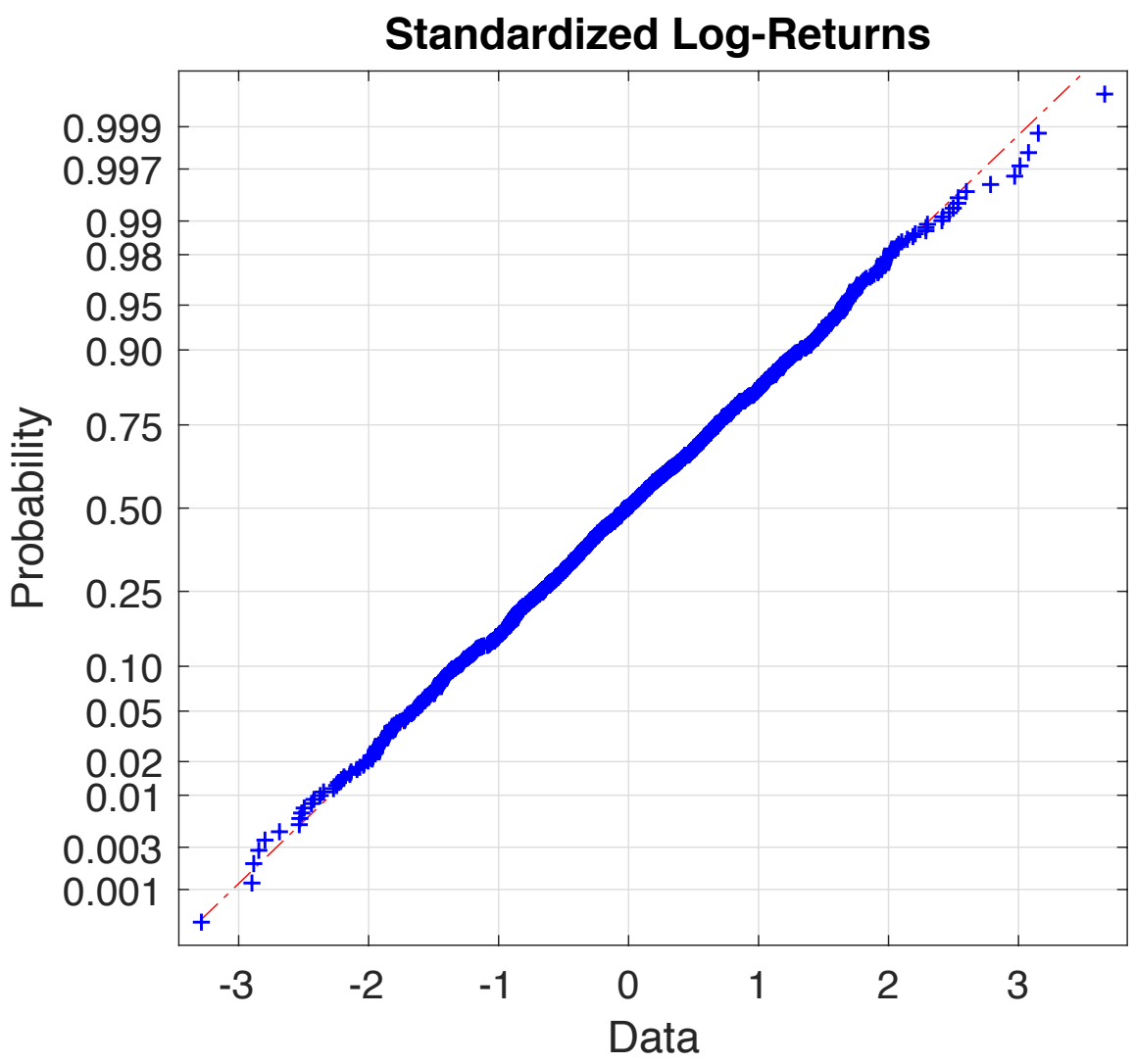


Best Fit: 2σ Bounds

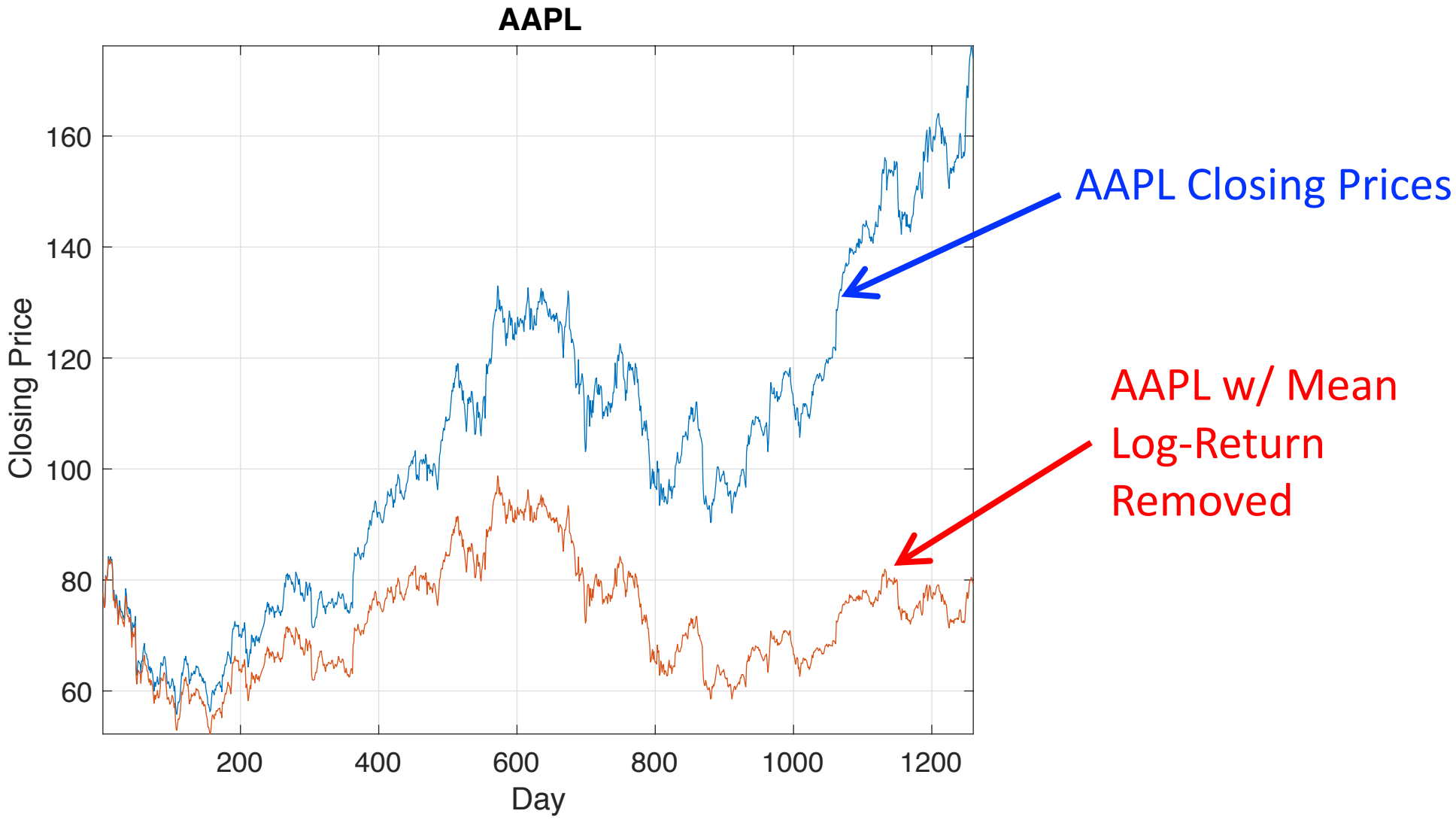


p-value: 0.9558

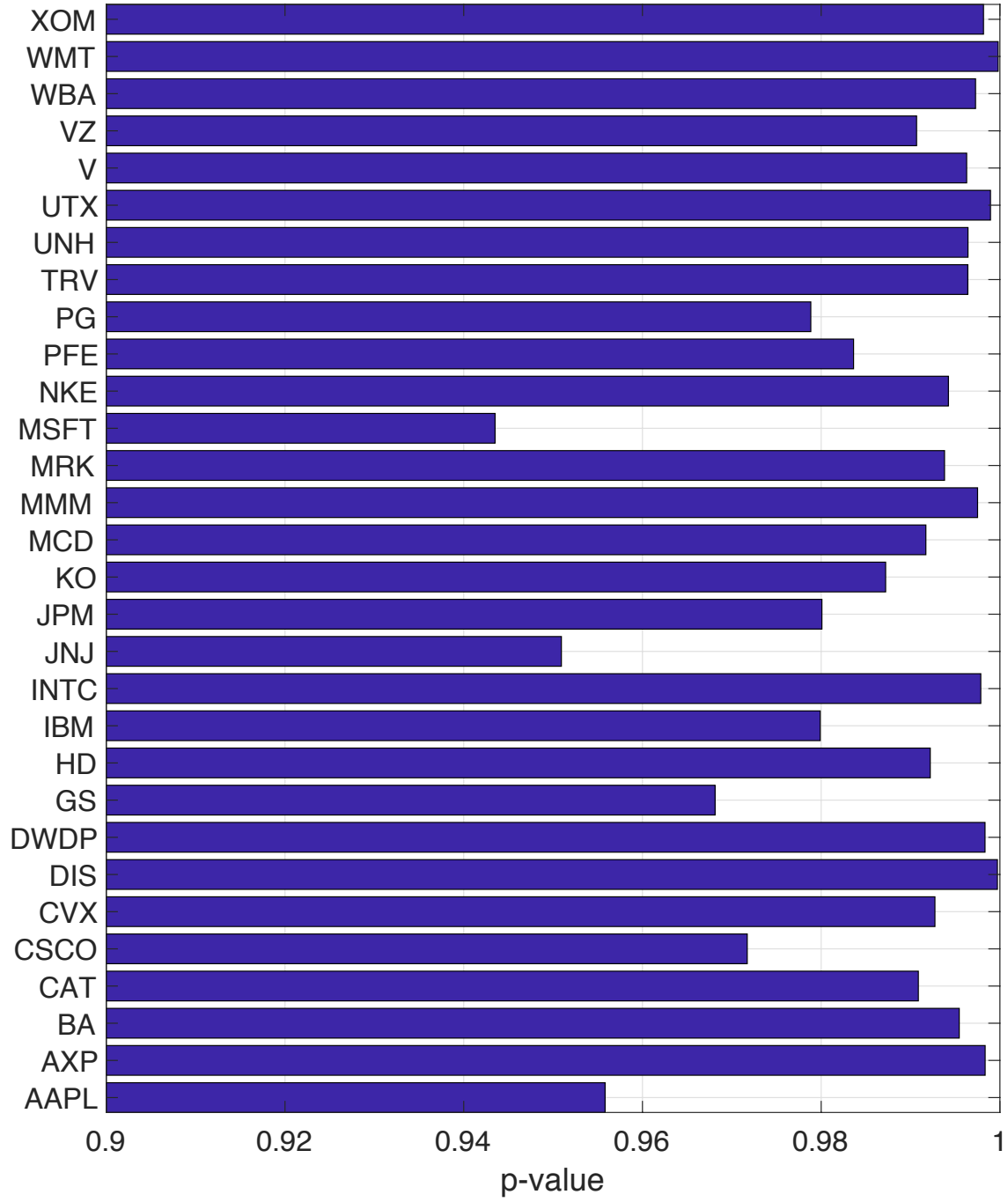

Raw Log-Returns
Probability
0.999
0.997
0.99
0.98
0.95
0.90
0.75
0.50
0.25
0.10
0.05
0.02
0.01
0.003
0.001
-0.1
-0.05
0
0.05
Data


Standardized Log-Returns
Probability
0.999
0.997
0.99
0.98
0.95
0.90
0.75
0.50
0.25
0.10
0.05
0.02
0.01
0.003
0.001
-3
-2
-1
0
1
2
3
Data

Mean Removal:



Ensemble Results



Thank You!