Sub-sampling in Parametric Estimation of Effective Stochastic Models from Discrete Data

with R. Azencott, A. Beri

University of Houston

Motivation

Goal: Parametric estimation of Effective Stocahstic Models from Discrete Data

- Develop data-driven Parametrizations for Various Physical Processes

- Develop data-driven techniques for parametric fitting of effective stochastic models for large-scale structures in PDEs

Data Source: Observations or Numerical Simulations; Discrete Time-Series of the Large-Scale Structures; no knowledge about the small-scale data

Most of the work: “Fixed” time-step

Realistic Situation: Data comes from a deterministic model (smooth trajectories); can be sampled with an arbitrary time-step

In this talk: Role of the sampling time-step in parametric estimation

Sub-Sampling: Data is not approximated well by a stochastic model for small time-steps; Effective Model is an approximation which is valid only on larger times

Outline of the Problem

Available Data $\mathbf{U} = \{U_k\} = \{Y(k\Delta)\}$ sampled from a continuous trajectory Y_t with arbitrary time-step Δ

Propose an Effective SDE Model

$$dX_t = b(X_t, \theta)dt + \sigma(X_t, \theta)dW$$

Estimate Parameters θ using the Max. Likelihood Approach

In this Talk

Consider multiscale Fast-Slow systems Y_t^ϵ - slow variable

$$Y_t^\epsilon \to X_t \text{ as } \epsilon \to 0$$

- Understand the performance of the estimators as $\epsilon, \Delta \to 0$

- Can access the performance of parametric fitting by comparing Max. Likelihood Estimators with Analytical formulas

Prototype Example

Data is generated by Smoothed OU Process Y_t:

$$Y_t^\epsilon = \frac{1}{\epsilon} \int_{t-\epsilon}^{t} X_s ds$$

$$dX_t = -\gamma X_t dt + \sigma dW_t$$

Utilize Discrete Data $\{Y_{k\Delta}^\epsilon\}$ to Estimate Effective Model

$$dZ_t = -g Z_t dt + s dB_t$$

Since

$$Y_t^\epsilon \longrightarrow X_t, \text{ as } \epsilon \longrightarrow 0$$

Question: For which Δ estimates are consistent as $\epsilon \longrightarrow 0$? i.e.

Let $\Delta = \epsilon^\alpha$ what are the conditions for α such that $(\hat{g}, \hat{s}) \longrightarrow (\gamma, \sigma)$ as $\epsilon \longrightarrow 0$

Likelihood Function for SDEs

$$dZ = D(\mathbf{a}, Z)dt + G(\mathbf{a}, Z)dW,$$

Euler Discretization

$$Z_{n+1} = Z_n + D(\mathbf{a}, Z_n)\Delta t + G(\mathbf{a}, Z_n)\Delta W_n$$

Gaussian Random Variable

$$G(\mathbf{a}, Z_n)\Delta W_n = [Z_{n+1} - Z_n - D(\mathbf{a}, Z_n)\Delta t]$$

Likelihood Function

$$L(\mathbf{a}|Z_{obs}) = \frac{1}{(2\Delta t)^{(N-1)/2} \prod G(\mathbf{a}, Z_n)} e^{-\frac{1}{2\Delta t} \sum \frac{(Z_{n+1} - Z_n - D(\mathbf{a}, Z_n)\Delta t)^2}{G^2(\mathbf{a}, Z_n)}}$$

Ornstein-Uhlenbeck Process and Likelihood Function

$$dZ_t = -gZ_t dt + s dW_t$$

Given Discrete sample with time-step Δ, i.e. $U_k = Z_{k\Delta}$

$$\hat{g}(N) = \frac{1}{\Delta} Ln\left(\frac{\hat{r}_1(N)}{\hat{r}_0(N)}\right), \quad \hat{s}(N) = 2\hat{g}(N)\hat{r}_0(N)$$

$$\hat{r}_0(N) = \frac{1}{N}\sum_0^{N-1} U_n^2, \quad \hat{r}_1(N) = \frac{1}{N}\sum_0^{N-1} U_{n+1}U_n$$

Interpretation of $\hat{g}(N)$: slope of the log of the correlation function at lag Δ

Max Likelihood Estimates for the Smoothed OU Process

Understand Estimates

$$\hat{g}^{\epsilon}(N) = \frac{1}{\Delta} Ln\left(\frac{\hat{r}_1^{\epsilon}(N)}{\hat{r}_0^{\epsilon}(N)}\right), \quad \hat{s}^{\epsilon}(N) = 2\hat{g}^{\epsilon}(N)\hat{r}_0^{\epsilon}(N)$$

when data is generated by the Smoothed Ornstein-Uhlenbeck Process

Consistency = Equilibrium Values

$$\hat{r}_0^{\epsilon}(N) \rightarrow E[(Y_t^{\epsilon})^2] = \frac{\sigma^2}{2\gamma}\frac{2(\gamma\epsilon + e^{-\gamma\epsilon} - 1)}{\gamma^2\epsilon^2}$$

$$\hat{r}_1^{\epsilon}(N) \rightarrow E[Y_t^{\epsilon} Y_{t+\Delta}^{\epsilon}] = \frac{\sigma^2}{2\gamma} e^{-\gamma\Delta}\left[A_1(\epsilon) \text{ if } \Delta \geq \epsilon;\ A_2(\epsilon, \Delta) \text{ if } 0 < \Delta < \epsilon\right]$$

$$\hat{g}(N) \rightarrow -\frac{1}{\Delta} Ln\left(\frac{E[Y_t^{\epsilon} Y_{t+\Delta}^{\epsilon}]}{E[(Y_t^{\epsilon})^2]}\right) = g_{lim} \text{ as } N \rightarrow \infty$$

Question: $g_{lim} \rightarrow \gamma$? as $\epsilon \rightarrow 0$

Expansion for small ϵ, Δ

CASE $\Delta \geq \epsilon$

$$\text{Bias}: \; g_{lim} - \gamma \approx -\frac{1}{\Delta}\left[\frac{\gamma\epsilon}{3} + \frac{5\gamma^2\epsilon^2}{36}\right]$$

In particular:

$$\text{When } \; \Delta = \epsilon : g_{lim} - \gamma \approx -\left[\frac{\gamma}{3} + \frac{5\gamma^2\epsilon}{36}\right]$$

Constant Bias for any finite Δ, ϵ

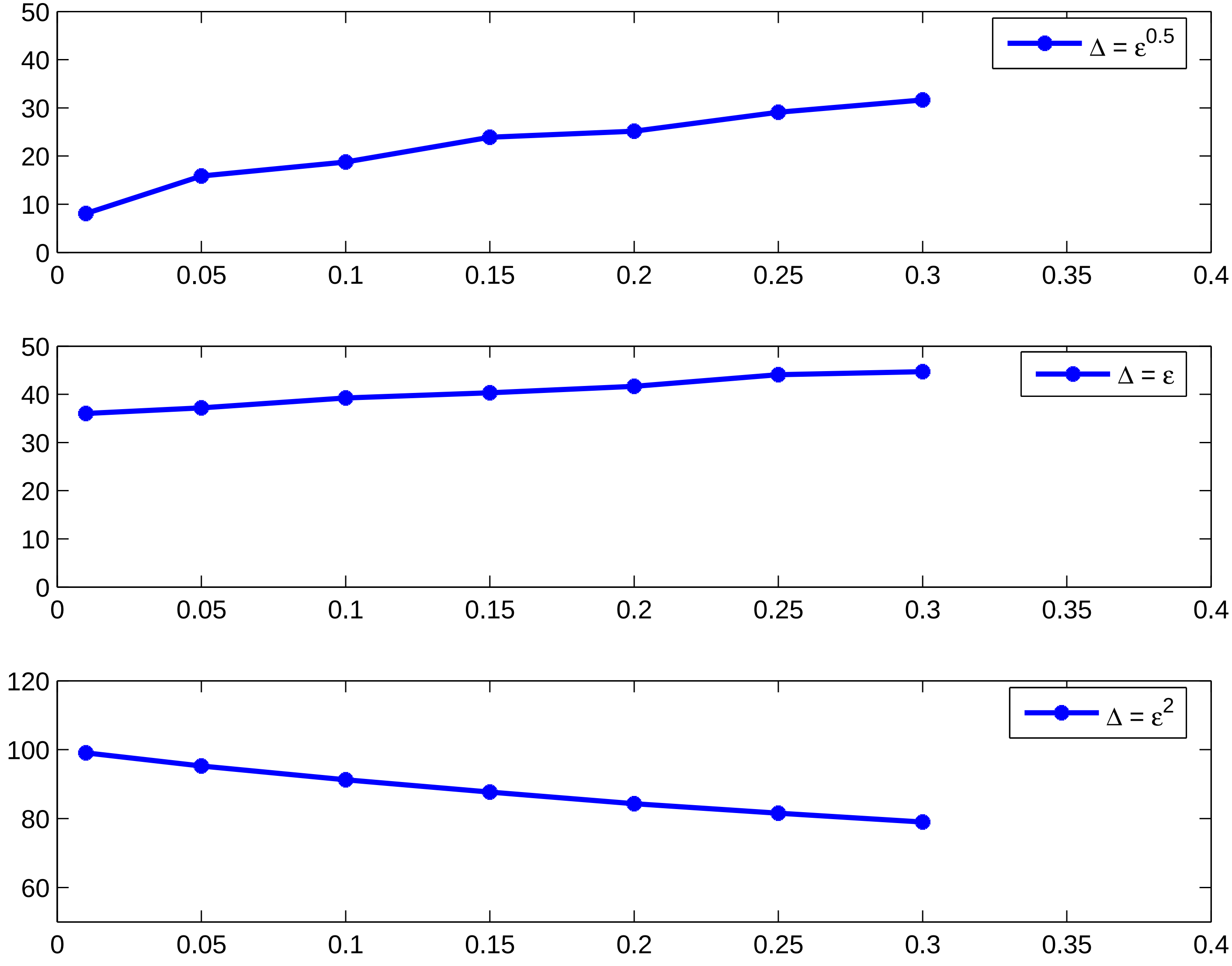
Consistency: $\epsilon, \Delta \rightarrow 0$

$$\Delta = \epsilon^{\alpha}, \;\; \alpha \in (0,1)$$

Numerical Simulations Data is generated by the SOU process

Error in $\hat{g}$ vs ϵ for different sub-sampling strategies

Top: $\Delta = \epsilon^{0.5}$, Middle: $\Delta = \epsilon$, Bottom: $\Delta = \epsilon^2$



Triad Model and Effective Dynamics

Consider Triad Model

$$\begin{aligned} dx &= \frac{1}{\epsilon} A_1 yz dt \\ dy &= \frac{1}{\epsilon} A_2 xz dt - \frac{1}{\epsilon^2} \gamma_1 y dt + \frac{1}{\epsilon} \sigma_1 dW_1 \\ dz &= \frac{1}{\epsilon} A_3 xy dt - \frac{1}{\epsilon^2} \gamma_2 z dt + \frac{1}{\epsilon} \sigma_2 dW_2 \end{aligned}$$

Homogenezation: $x \longrightarrow X$ as $\epsilon \longrightarrow 0$

Effective System

$$dX = -\Gamma X dt + \Sigma dW$$

with

$$\Gamma = \frac{-A_1}{2(\gamma_1 + \gamma_2)} \left(\frac{A_2 \sigma_2^2}{\gamma_2} + \frac{A_3 \sigma_1^2}{\gamma_1} \right), \quad \Sigma = \frac{A_1 \sigma_1 \sigma_2}{\sqrt{2\gamma_1 \gamma_2 (\gamma_1 + \gamma_2)}}$$

Triad Model vs Smoothed OU Process

Compare Correlation Functions for small lags

Smoothed OU Process $0 < \Delta < \epsilon$

$$CF_{SOU}(\Delta) \approx 1 - \frac{C\Delta^2}{\epsilon}$$

Triad Process

$$CF_{Triad}(\Delta) \approx 1 - \frac{(\gamma_1 + \gamma_2)C}{2\epsilon^2}\Delta^2$$

Therefore

$$\epsilon \ \ (\mathrm{SOU}) \ \ \sim \epsilon^2 \ \ (\mathrm{Triad})$$

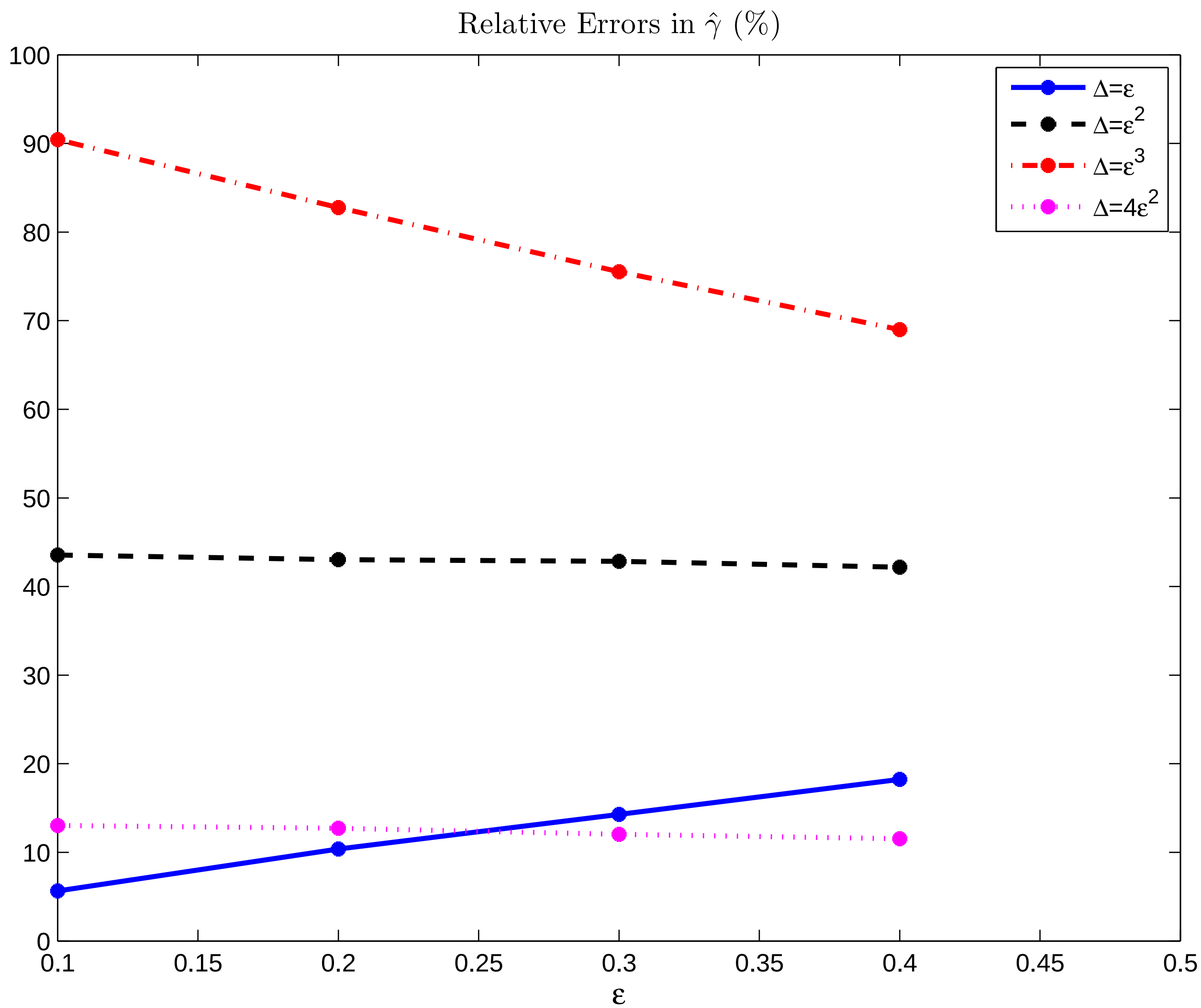
Consistency for the Triad $\Delta = \epsilon^{2\alpha}, \alpha \in (0, 1)$

Numerical Simulations Data generated by the Triad Model, i.e. $Y_t^\epsilon = x(t)$

Numerical Error in $\hat{g}$ vs ϵ Bias: $\hat{g} - \gamma \approx -\frac{\gamma\epsilon^2}{3\Delta}$

Red: $\Delta = \epsilon^3$ Blue: $\Delta = \epsilon$

Black: $\Delta = \epsilon^2$ Magenta: $\Delta = 4\epsilon^2$

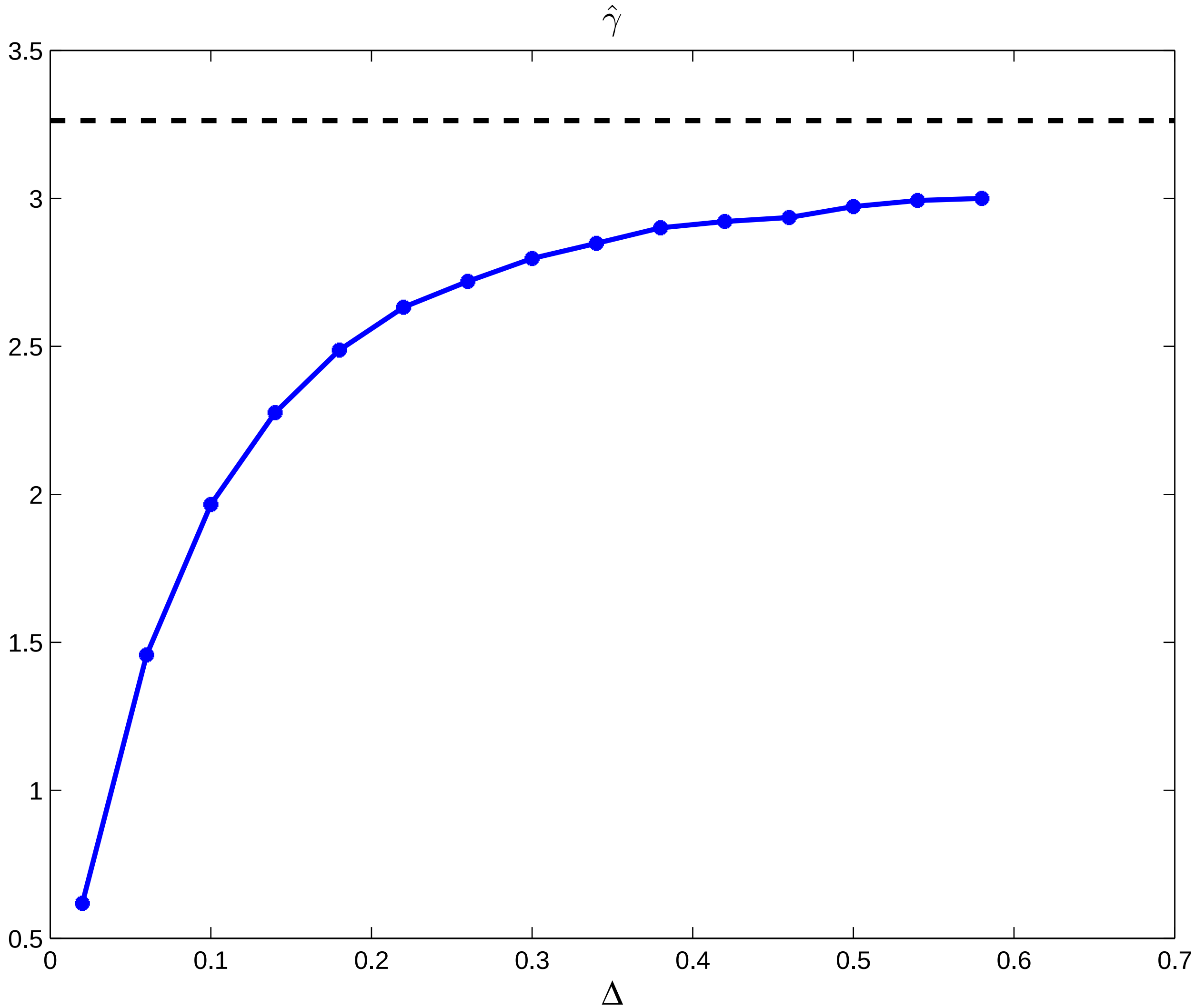


Finite ϵ

Consider a Particular Triad Dataset with a fixed value of $\epsilon = 0.3$

Consider $\hat{g}(\Delta)$ vs Δ

Bias: $\hat{g} - \gamma \approx -\frac{\gamma\epsilon^2}{3\Delta}$

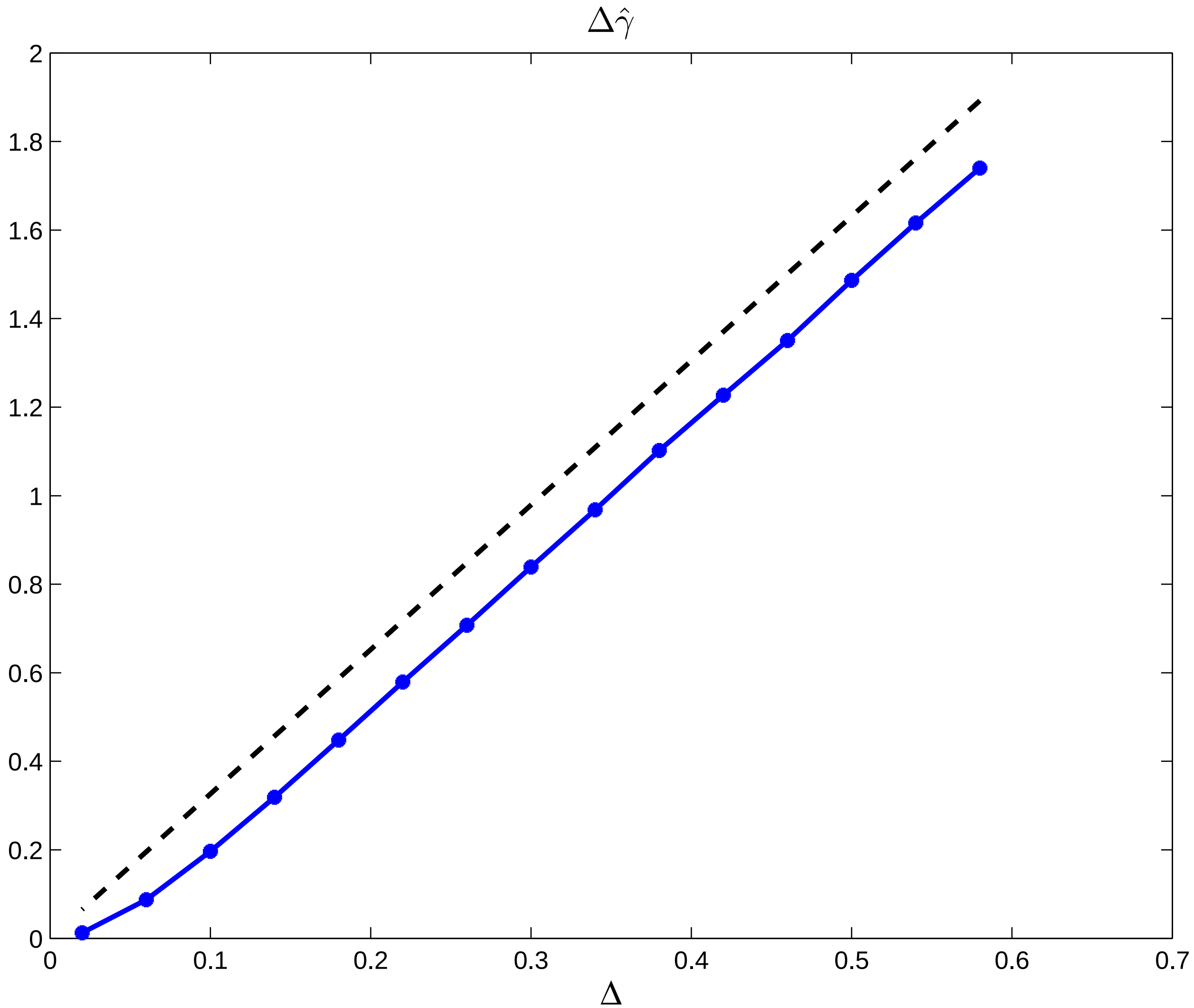


Finite ϵ

Triad Dataset with $\epsilon = 0.3$

Consider $\Delta\hat{g}(\Delta)$ vs Δ

Conjecture $\hat{g}\Delta = \gamma\Delta + C$

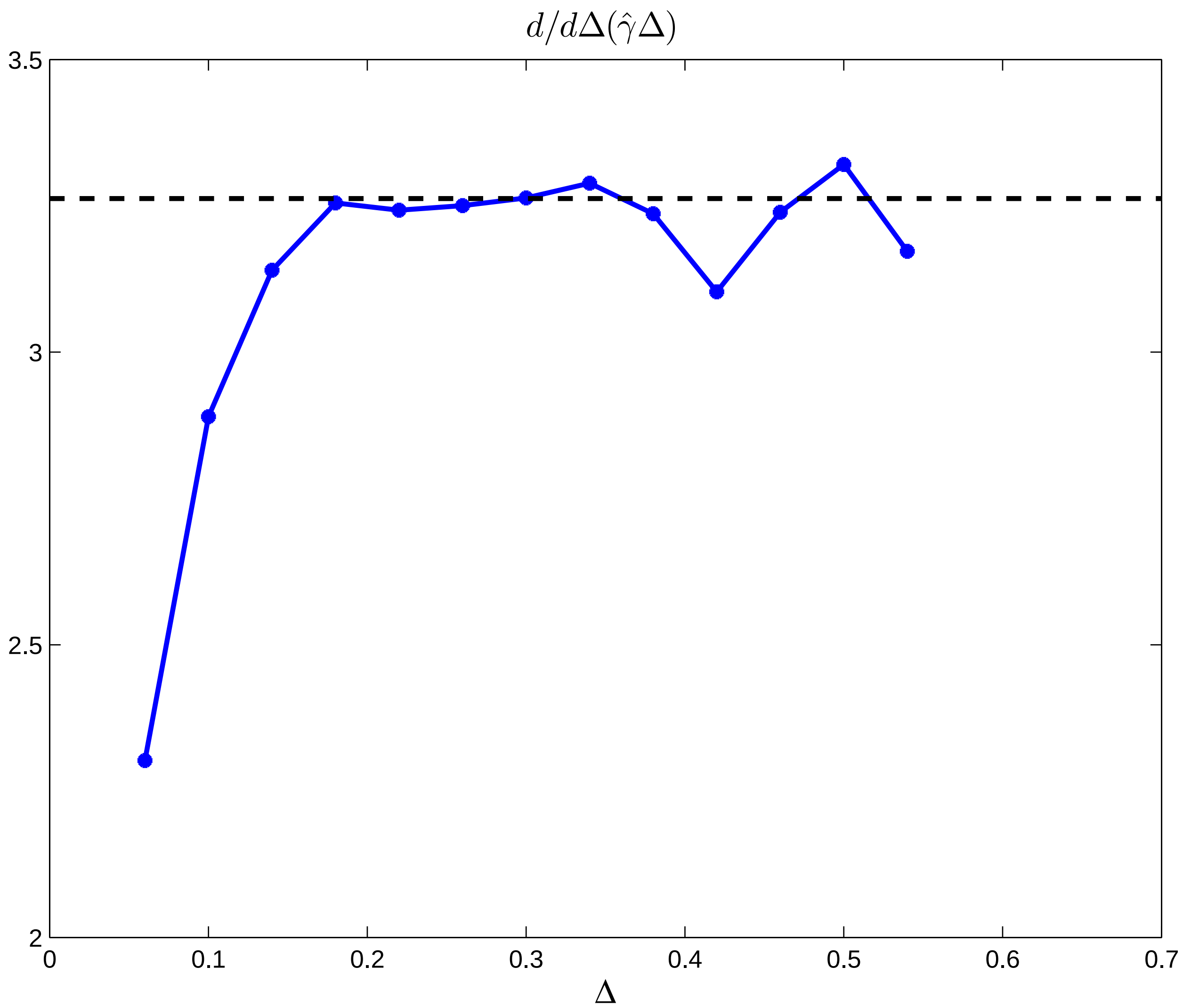


Finite ϵ

Triad Dataset with $\epsilon = 0.3$

Consider $\frac{d}{d\Delta}[\Delta \hat{g}(\Delta)]$ vs Δ

Conjecture $\frac{d}{d\Delta}[\Delta \hat{g}] = \gamma$



Conclusions

Essentially, all models are wrong, but some are useful

– George E. P. Box

- Time-step can be viewed as another parameter to be optimized
- Data cannot be approximated by a stochastic process for small Δ
- Sub-sampling: Determine critical time-step for which SDE is valid (on longer time-scales)
- Behavior of the correlation function of the large scales near $lag = 0$ is crucial for understanding sub-sampling
- Estimators from the data with small time-step underestimate the damping term