Likelihood-free Estimation by Matching Random Features

Michael Wieck-Sosa

Joint work with Cosma Shalizi

Department of Statistics & Data Science, Carnegie Mellon University

April 3, 2026

Overview of presentation

Overview of presentation

Part I

Introduction

Time series data

Sim-based estimation

Choices of features

Overview of presentation

Part I

Introduction

Time series data

Sim-based estimation

Choices of features

Part II

Random features

Embeddings

Estimators

Example

Overview of presentation

Part I	Part II	Part III
Introduction	**Random features**	**Theory**
Time series data	Embeddings	Main results
Sim-based estimation	Estimators	Assumptions
Choices of features	Example	Future work

Time series setting

- We observe X_t^{obs}, $t = 1, \ldots, n$, for some sample size $n \in \mathbb{N}$

Time series setting

- We observe X_t^{obs}, $t = 1, \ldots, n$, for some sample size $n \in \mathbb{N}$
- Each observation X_t^{obs} takes values in $\mathbb{R}^d$ for some dimension $d \in \mathbb{N}$

Time series setting

- We observe X_t^{obs}, $t = 1, \ldots, n$, for some sample size $n \in \mathbb{N}$
- Each observation X_t^{obs} takes values in $\mathbb{R}^d$ for some dimension $d \in \mathbb{N}$
- **Temporal dependence:** X_t^{obs} may depend on the past $X_{t-1}^{\mathrm{obs}}, X_{t-2}^{\mathrm{obs}}, \ldots$

Time series setting

- We observe X_t^{obs}, $t = 1, \dots, n$, for some sample size $n \in \mathbb{N}$
- Each observation X_t^{obs} takes values in $\mathbb{R}^d$ for some dimension $d \in \mathbb{N}$
- **Temporal dependence:** X_t^{obs} may depend on the past $X_{t-1}^{\mathrm{obs}}, X_{t-2}^{\mathrm{obs}}, \dots$
- **Nonstationarity:** Joint distribution of $X_{t_1+\tau}^{\mathrm{obs}}, \dots, X_{t_k+\tau}^{\mathrm{obs}}$ may change with τ

Simulation-based estimation

- For some $\theta_0 \in \Theta \subset \mathbb{R}^p$, assume the time series is *generated by nature* as

$$X_t^{\text{obs}} = G_t\left(\varepsilon_t, \theta_0\right), \quad t = 1, \ldots, n$$

Simulation-based estimation

- For some $\theta_0 \in \Theta \subset \mathbb{R}^p$, assume the time series is *generated by nature* as

$$X_t^{\text{obs}} = G_t\left(\varepsilon_t, \theta_0\right), \quad t = 1, \ldots, n$$

- Want estimate $\hat{\theta}$ of θ_0, but likelihood is intractable, so we use simulation

Simulation-based estimation

- For some $\theta_0 \in \Theta \subset \mathbb{R}^p$, assume the time series is *generated by nature* as

$$X_t^{\text{obs}} = G_t\left(\varepsilon_t, \theta_0\right), \quad t = 1, \ldots, n$$

- Want estimate $\hat{\theta}$ of θ_0, but likelihood is intractable, so we use simulation
- **Key assumption:** Can simulate from an *approximation* $G_t^{(i)}$ to G_t at any $\theta \in \Theta$

Simulation-based estimation

- For some $\theta_0 \in \Theta \subset \mathbb{R}^p$, assume the time series is *generated by nature* as

$$X_t^{\text{obs}} = G_t\left(\varepsilon_t, \theta_0\right), \quad t = 1, \ldots, n$$

- Want estimate $\hat{\theta}$ of θ_0, but likelihood is intractable, so we use simulation
- **Key assumption:** Can simulate from an *approximation* $G_t^{(i)}$ to G_t at any $\theta \in \Theta$
  1. Approximation improves as $i \to \infty$

Simulation-based estimation

- For some $\theta_0 \in \Theta \subset \mathbb{R}^p$, assume the time series is *generated by nature* as

$$X_t^{\text{obs}} = G_t\left(\varepsilon_t, \theta_0\right), \quad t = 1, \ldots, n$$

- Want estimate $\hat{\theta}$ of θ_0, but likelihood is intractable, so we use simulation
- **Key assumption:** Can simulate from an *approximation* $G_t^{(i)}$ to G_t at any $\theta \in \Theta$
  1. Approximation improves as $i \to \infty$
  2. Growing burn-in periods, discretized ODEs, etc.

How simulation-based estimation is usually done

- **Rough idea:** Tune the parameters θ until some features F "look right"

How simulation-based estimation is usually done

- **Rough idea:** Tune the parameters θ until some features F "look right"
- For time series that are stationary, at least asymptotically, often use estimator

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \left\| F^{\mathrm{obs}} - \frac{1}{s} \sum_{r=1}^{s} F^{(r)}(\theta) \right\|$$

How simulation-based estimation is usually done

- **Rough idea:** Tune the parameters θ until some features F "look right"
- For time series that are stationary, at least asymptotically, often use estimator

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \left\| F^{\mathrm{obs}} - \frac{1}{s} \sum_{r=1}^{s} F^{(r)}(\theta) \right\|$$

- For consistency $\hat{\theta} \xrightarrow{p} \theta_0$, need $F(\theta) \xrightarrow{p} \mathbb{E}_{P_\theta}[F(\theta)] \equiv \Phi(\theta)$ for all θ (ULLN)

How simulation-based estimation is usually done

- **Rough idea:** Tune the parameters θ until some features F "look right"
- For time series that are stationary, at least asymptotically, often use estimator

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \left\| F^{\text{obs}} - \frac{1}{s} \sum_{r=1}^{s} F^{(r)}(\theta) \right\|$$

- For consistency $\hat{\theta} \xrightarrow{p} \theta_0$, need $F(\theta) \xrightarrow{p} \mathbb{E}_{P_\theta}[F(\theta)] \equiv \Phi(\theta)$ for all θ (ULLN)
- Hope Φ^{-1} exists and is nice, so if we knew $\Phi(\theta)$, then could get back θ

How simulation-based estimation is usually done

- **Rough idea:** Tune the parameters θ until some features F "look right"
- For time series that are stationary, at least asymptotically, often use estimator

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \left\| F^{\mathrm{obs}} - \frac{1}{s} \sum_{r=1}^{s} F^{(r)}(\theta) \right\|$$

- For consistency $\hat{\theta} \xrightarrow{p} \theta_0$, need $F(\theta) \xrightarrow{p} \mathbb{E}_{P_\theta}[F(\theta)] \equiv \Phi(\theta)$ for all θ (ULLN)
- Hope Φ^{-1} exists and is nice, so if we knew $\Phi(\theta)$, then could get back θ
- Good features are sensitive to small changes in θ, but takes time to craft features

Many choices of features

- **Method of simulated moments:** F based on functions of moments

Many choices of features

- **Method of simulated moments:** F based on functions of moments
  - Sample mean, variance, skewness, autocorrelation, and so on

Many choices of features

- **Method of simulated moments:** F based on functions of moments
  - Sample mean, variance, skewness, autocorrelation, and so on
- **Indirect inference:** F from parameter estimates in auxiliary model

Many choices of features

- **Method of simulated moments:** F based on functions of moments
  - Sample mean, variance, skewness, autocorrelation, and so on
- **Indirect inference:** F from parameter estimates in auxiliary model
  - To get $\hat{\theta}$ for MA(1): $X_t = \varepsilon_t + \theta_0 \varepsilon_{t-1}$

Many choices of features

- **Method of simulated moments:** F based on functions of moments
  - Sample mean, variance, skewness, autocorrelation, and so on
- **Indirect inference:** F from parameter estimates in auxiliary model
  - To get $\hat{\theta}$ for MA(1): $X_t = \varepsilon_t + \theta_0 \varepsilon_{t-1}$
  - Use $F = \hat{\beta}$ from AR(1): $X_t = \beta X_{t-1} + \xi_t$

Many choices of features

- **Method of simulated moments:** F based on functions of moments
  - Sample mean, variance, skewness, autocorrelation, and so on
- **Indirect inference:** F from parameter estimates in auxiliary model
  - To get $\hat{\theta}$ for MA(1): $X_t = \varepsilon_t + \theta_0 \varepsilon_{t-1}$
  - Use $F = \hat{\beta}$ from AR(1): $X_t = \beta X_{t-1} + \xi_t$
- **Synthetic likelihood:** F from user-chosen summary statistics

Many choices of features

- **Method of simulated moments:** F based on functions of moments
  - Sample mean, variance, skewness, autocorrelation, and so on
- **Indirect inference:** F from parameter estimates in auxiliary model
  - To get $\hat{\theta}$ for MA(1): $X_t = \varepsilon_t + \theta_0 \varepsilon_{t-1}$
  - Use $F = \hat{\beta}$ from AR(1): $X_t = \beta X_{t-1} + \xi_t$
- **Synthetic likelihood:** F from user-chosen summary statistics
  - Coefficients from polynomial regression of X_t on $X_{t-1}, \ldots, X_{t-\ell}$

Many choices of features

- **Method of simulated moments:** F based on functions of moments
  - Sample mean, variance, skewness, autocorrelation, and so on
- **Indirect inference:** F from parameter estimates in auxiliary model
  - To get $\hat{\theta}$ for MA(1): $X_t = \varepsilon_t + \theta_0 \varepsilon_{t-1}$
  - Use $F = \hat{\beta}$ from AR(1): $X_t = \beta X_{t-1} + \xi_t$
- **Synthetic likelihood:** F from user-chosen summary statistics
  - Coefficients from polynomial regression of X_t on $X_{t-1}, \ldots, X_{t-\ell}$
- **Neural summary statistics:** F from neural network

Many choices of features

- **Method of simulated moments:** F based on functions of moments
  - Sample mean, variance, skewness, autocorrelation, and so on
- **Indirect inference:** F from parameter estimates in auxiliary model
  - To get $\hat{\theta}$ for MA(1): $X_t = \varepsilon_t + \theta_0 \varepsilon_{t-1}$
  - Use $F = \hat{\beta}$ from AR(1): $X_t = \beta X_{t-1} + \xi_t$
- **Synthetic likelihood:** F from user-chosen summary statistics
  - Coefficients from polynomial regression of X_t on $X_{t-1}, \ldots, X_{t-\ell}$
- **Neural summary statistics:** F from neural network
  - Extract early layer of neural network

Can we just use random features?

- **Benefits:** No feature crafting, efficient, little knowledge about model is needed

Can we just use random features?

- **Benefits:** No feature crafting, efficient, little knowledge about model is needed
- Typically, need $2p + 1$ random features for a p-dimensional parameter space Θ

Can we just use random features?

- **Benefits:** No feature crafting, efficient, little knowledge about model is needed
- Typically, need $2p + 1$ random features for a p-dimensional parameter space Θ
- Probabilistic embedding theorem due to Sauer, J. A. Yorke, and Casdagli (1991) states that, for "almost every" C^1 smooth function Φ from $\mathbb{R}^p$ to $\mathbb{R}^{2p+1}$:

Can we just use random features?

- **Benefits:** No feature crafting, efficient, little knowledge about model is needed
- Typically, need $2p+1$ random features for a p-dimensional parameter space Θ
- Probabilistic embedding theorem due to Sauer, J. A. Yorke, and Casdagli (1991) states that, for "almost every" C^1 smooth function Φ from $\mathbb{R}^p$ to $\mathbb{R}^{2p+1}$:
  1. Φ is **one-to-one** on compact subsets of $\mathbb{R}^p$

Can we just use random features?

- **Benefits:** No feature crafting, efficient, little knowledge about model is needed
- Typically, need $2p+1$ random features for a p-dimensional parameter space Θ
- Probabilistic embedding theorem due to Sauer, J. A. Yorke, and Casdagli (1991) states that, for "almost every" C^1 smooth function Φ from $\mathbb{R}^p$ to $\mathbb{R}^{2p+1}$:
  1. Φ is **one-to-one** on compact subsets of $\mathbb{R}^p$
  2. Φ has a **nice derivative** on certain compact subsets of $\mathbb{R}^p$
     – Jacobian of Φ has full rank on compact subsets of smooth manifolds in $\mathbb{R}^p$

What do we mean by "almost every" function? Prevalent subset

Definition (J. Yorke and Ott (2005))

Let V be a completely metrizable topological vector space. A Borel set $E \subset V$ is said to be **prevalent** if there exists some Borel measure μ on V such that:

What do we mean by "almost every" function? Prevalent subset

Definition (J. Yorke and Ott (2005))

Let V be a completely metrizable topological vector space. A Borel set $E \subset V$ is said to be **prevalent** if there exists some Borel measure μ on V such that:

1. $0 < \mu(C) < \infty$ for some compact subset C of V

What do we mean by "almost every" function? Prevalent subset

Definition (J. Yorke and Ott (2005))

Let V be a completely metrizable topological vector space. A Borel set $E \subset V$ is said to be **prevalent** if there exists some Borel measure μ on V such that:

1. $0 < \mu(C) < \infty$ for some compact subset C of V
2. The set $E + v$ has full μ-measure (complement has measure zero) for all $v \in V$

What do we mean by "almost every" function? Prevalent subset

Definition (J. Yorke and Ott (2005))

Let V be a completely metrizable topological vector space. A Borel set $E \subset V$ is said to be **prevalent** if there exists some Borel measure μ on V such that:

1. $0 < \mu(C) < \infty$ for some compact subset C of V
2. The set $E + v$ has full μ-measure (complement has measure zero) for all $v \in V$

- If $E \subset V$ contains a prevalent Borel set, then "almost every" element of V is in E

What do we mean by "almost every" function? Prevalent subset

Definition (J. Yorke and Ott (2005))

Let V be a completely metrizable topological vector space. A Borel set $E \subset V$ is said to be **prevalent** if there exists some Borel measure μ on V such that:

1. $0 < \mu(C) < \infty$ for some compact subset C of V
2. The set $E + v$ has full μ-measure (complement has measure zero) for all $v \in V$

- If $E \subset V$ contains a prevalent Borel set, then "almost every" element of V is in E
- In this sense, "almost every" C^1 smooth map from $\mathbb{R}^p$ to $\mathbb{R}^{2p+1}$ has the properties from the embedding theorem of Sauer, J. A. Yorke, and Casdagli (1991)

Main ideas for estimation via random features

- Let P_θ be the joint distribution of $[X_t(\theta), \ldots, X_{t-m}(\theta)]^\top$ for some $m \in \mathbb{N}$

Main ideas for estimation via random features

- Let P_θ be the joint distribution of $[X_t(\theta), \ldots, X_{t-m}(\theta)]^\top$ for some $m \in \mathbb{N}$
- Under smoothness conditions on $\theta \mapsto P_\theta$, the expectations of $2p+1$ "generic" continuous, bounded functions $\varphi : \mathbb{R}^{(m+1)\times d} \to \mathbb{R}^{2p+1}$ will be C^1 smooth in θ

Main ideas for estimation via random features

- Let P_θ be the joint distribution of $[X_t(\theta), \ldots, X_{t-m}(\theta)]^\top$ for some $m \in \mathbb{N}$
- Under smoothness conditions on $\theta \mapsto P_\theta$, the expectations of $2p+1$ "generic" continuous, bounded functions $\varphi : \mathbb{R}^{(m+1)\times d} \to \mathbb{R}^{2p+1}$ will be C^1 smooth in θ
- For some φ, define the C^1 smooth function $\Phi : \mathbb{R}^p \to \mathbb{R}^{2p+1}$ by extending the map

$$\theta \mapsto \mathbb{E}_{P_\theta}\left[\varphi\left([X_t(\theta), \ldots, X_{t-m}(\theta)]^\top\right)\right]$$

Main ideas for estimation via random features

- Let P_θ be the joint distribution of $[X_t(\theta), \ldots, X_{t-m}(\theta)]^\top$ for some $m \in \mathbb{N}$
- Under smoothness conditions on $\theta \mapsto P_\theta$, the expectations of $2p+1$ "generic" continuous, bounded functions $\varphi : \mathbb{R}^{(m+1)\times d} \to \mathbb{R}^{2p+1}$ will be C^1 smooth in θ
- For some φ, define the C^1 smooth function $\Phi : \mathbb{R}^p \to \mathbb{R}^{2p+1}$ by extending the map

$$\theta \mapsto \mathbb{E}_{P_\theta}\left[\varphi\left([X_t(\theta), \ldots, X_{t-m}(\theta)]^\top\right)\right]$$

- Unless we are unlucky in our selection of φ, then Φ will embed $\Theta \subset \mathbb{R}^p$ in $\mathbb{R}^{2p+1}$
  – That is, Φ will be one-to-one on Θ and have a nice derivative on certain subsets

Main ideas for estimation via random features

- Let P_θ be the joint distribution of $[X_t(\theta), \ldots, X_{t-m}(\theta)]^\top$ for some $m \in \mathbb{N}$
- Under smoothness conditions on $\theta \mapsto P_\theta$, the expectations of $2p+1$ "generic" continuous, bounded functions $\varphi : \mathbb{R}^{(m+1)\times d} \to \mathbb{R}^{2p+1}$ will be C^1 smooth in θ
- For some φ, define the C^1 smooth function $\Phi : \mathbb{R}^p \to \mathbb{R}^{2p+1}$ by extending the map

$$\theta \mapsto \mathbb{E}_{P_\theta}\left[\varphi\left([X_t(\theta), \ldots, X_{t-m}(\theta)]^\top\right)\right]$$

- Unless we are unlucky in our selection of φ, then Φ will embed $\Theta \subset \mathbb{R}^p$ in $\mathbb{R}^{2p+1}$ – That is, Φ will be one-to-one on Θ and have a nice derivative on certain subsets
- **Key idea:** Estimate θ_0 by *minimizing distance* of $\hat{\Phi}(\theta_0)$ and $\hat{\Phi}(\theta)$ over $\theta \in \Theta$, $\hat{\Phi}(\theta)$ is a time-average of $2p+1$ random features of a time series generated at θ

Random Fourier features

- Let $x_1, \ldots, x_{m+1}$ be vectors in $\mathbb{R}^d$, and define $x = (x_1, \ldots, x_{m+1})^\top \in \mathbb{R}^{(m+1)\times d}$

Random Fourier features

- Let $x_1, \ldots, x_{m+1}$ be vectors in $\mathbb{R}^d$, and define $x = (x_1, \ldots, x_{m+1})^\top \in \mathbb{R}^{(m+1)\times d}$
- We consider random features of the form

$$\varphi_i(x) = \cos\left(\sum_{j=1}^{m+1} \Omega_{i,j} \cdot x_j + \alpha_i\right)$$

where $\Omega_{i,j} \overset{\text{iid}}{\sim} N(0, I_d)$ and $\alpha_i \overset{\text{iid}}{\sim} U(-\pi, \pi)$ for $i = 1, \ldots, k$ and $j = 1, \ldots, m+1$

Random Fourier features

- Let $x_1, \ldots, x_{m+1}$ be vectors in $\mathbb{R}^d$, and define $x = (x_1, \ldots, x_{m+1})^\top \in \mathbb{R}^{(m+1)\times d}$
- We consider random features of the form

$$\varphi_i(x) = \cos\left(\sum_{j=1}^{m+1} \Omega_{i,j} \cdot x_j + \alpha_i\right)$$

  where $\Omega_{i,j} \overset{\text{iid}}{\sim} N(0, I_d)$ and $\alpha_i \overset{\text{iid}}{\sim} U(-\pi, \pi)$ for $i = 1, \ldots, k$ and $j = 1, \ldots, m+1$
- Denote all k of these functions by

$$\varphi = (\varphi_1, \ldots, \varphi_k)$$

  so that $\varphi : \mathbb{R}^{(m+1)\times d} \to \mathbb{R}^k$

Random Fourier features

- Define the i-th random feature at time $t = m+1, \ldots, n$ as

$$f_{t,i}^{\mathrm{obs}} = \varphi_i(X_{t-m:t}^{\mathrm{obs}}), \quad f_{t,i}^{(r)}(\theta) = \varphi_i\left(X_{t-m:t}^{(r)}(\theta)\right)$$

Random Fourier features

- Define the i-th random feature at time $t = m+1, \ldots, n$ as

$$f_{t,i}^{\text{obs}} = \varphi_i(X_{t-m:t}^{\text{obs}}), \quad f_{t,i}^{(r)}(\theta) = \varphi_i\left(X_{t-m:t}^{(r)}(\theta)\right)$$

- Write all k random features at time t as

$$f_t^{\text{obs}} = \left(f_{t,1}^{\text{obs}}, \ldots, f_{t,k}^{\text{obs}}\right), \quad f_t^{(r)}(\theta) = \left(f_{t,1}^{(r)}(\theta), \ldots, f_{t,k}^{(r)}(\theta)\right)$$

Time-average estimator

- Define the observed and simulated time-averages of the random features by

$$F^{\mathrm{obs}} = \frac{1}{n-m} \sum_{t=m+1}^{n} f_t^{\mathrm{obs}}, \quad \bar{F}^{\mathrm{sim}}(\theta) = \frac{1}{n-m} \sum_{t=m+1}^{n} \bar{f}_t^{\mathrm{sim}}(\theta)$$

where $\bar{f}_t^{\mathrm{sim}}(\theta) = \frac{1}{s} \sum_{r=1}^{s} f_t^{(r)}(\theta)$

Time-average estimator

- Define the observed and simulated time-averages of the random features by

$$F^{\text{obs}} = \frac{1}{n-m} \sum_{t=m+1}^{n} f_t^{\text{obs}}, \quad \bar{F}^{\text{sim}}(\theta) = \frac{1}{n-m} \sum_{t=m+1}^{n} \bar{f}_t^{\text{sim}}(\theta)$$

where $\bar{f}_t^{\text{sim}}(\theta) = \frac{1}{s} \sum_{r=1}^{s} f_t^{(r)}(\theta)$

- The time-average estimator is given by

$$\hat{\theta}^{\text{TA}} = \underset{\theta \in \Theta}{\text{argmin}} \left\| \hat{Q}_n^{\text{TA}}(\theta) \right\|$$

where the sample discrepancy function is defined as

$$\hat{Q}_n^{\text{TA}}(\theta) = F^{\text{obs}} - \bar{F}^{\text{sim}}(\theta)$$

Time-average estimator

- Define the observed and simulated time-averages of the random features by

$$F^{\mathrm{obs}} = \frac{1}{n-m} \sum_{t=m+1}^{n} f_t^{\mathrm{obs}}, \quad \bar{F}^{\mathrm{sim}}(\theta) = \frac{1}{n-m} \sum_{t=m+1}^{n} \bar{f}_t^{\mathrm{sim}}(\theta)$$

where $\bar{f}_t^{\mathrm{sim}}(\theta) = \frac{1}{s} \sum_{r=1}^{s} f_t^{(r)}(\theta)$

- The time-average estimator is given by

$$\hat{\theta}^{\mathrm{TA}} = \operatorname*{argmin}_{\theta \in \Theta} \left\| \hat{Q}_n^{\mathrm{TA}}(\theta) \right\|$$

where the sample discrepancy function is defined as

$$\hat{Q}_n^{\mathrm{TA}}(\theta) = F^{\mathrm{obs}} - \bar{F}^{\mathrm{sim}}(\theta)$$

- In practice, an optimization procedure is used to find the minimizer of $\left\| \hat{Q}_n^{\mathrm{TA}}(\cdot) \right\|$

Rolling-window estimator

- Define the observed and simulated rolling-window random features at time t by

$$F_t^{\mathrm{obs}} = \frac{1}{w \wedge t} \sum_{j=(t-w)\vee 1}^{t} f_j^{\mathrm{obs}}, \quad \bar{F}_t^{\mathrm{sim}}(\theta) = \frac{1}{w \wedge t} \sum_{j=(t-w)\vee 1}^{t} \bar{f}_j^{\mathrm{sim}}(\theta),$$

where $\bar{f}_t^{\mathrm{sim}}(\theta) = \frac{1}{s} \sum_{r=1}^{s} f_t^{(r)}(\theta)$ and $w \in \mathbb{N}$ is the window size

Rolling-window estimator

- The rolling-window estimator is given by

$$\hat{\theta}^{\mathrm{RW}} = \operatorname*{argmin}_{\theta \in \Theta} \left| \hat{Q}_n^{\mathrm{RW}}(\theta) \right|,$$

where the sample discrepancy function is defined as

$$\hat{Q}_n^{\mathrm{RW}}(\theta) = \frac{1}{n-m} \sum_{t=m+\tau+L}^{n} \left[\left\| F_t^{\mathrm{obs}} - \bar{F}_t^{\mathrm{sim}}(\theta) \right\|^2 + K_t(\theta) \right],$$

$$K_t(\theta) = 2 \left(F_{t-L}^{\mathrm{obs}} - \bar{F}_{t-L}^{\mathrm{sim}}(\theta) \right)^\top \left(\left[f_t^{\mathrm{obs}} - \bar{f}_t^{\mathrm{sim}}(\theta) \right] - \left[F_{t-L}^{\mathrm{obs}} - \bar{F}_{t-L}^{\mathrm{sim}}(\theta) \right] \right),$$

where $\tau \in \mathbb{N}$ is an initial time-offset and $L \in \mathbb{N}$ is a lag

Rolling-window estimator

- The rolling-window estimator is given by

$$\hat{\theta}^{\mathrm{RW}} = \operatorname*{argmin}_{\theta \in \Theta} \left| \hat{Q}_n^{\mathrm{RW}}(\theta) \right|,$$

where the sample discrepancy function is defined as

$$\hat{Q}_n^{\mathrm{RW}}(\theta) = \frac{1}{n-m} \sum_{t=m+\tau+L}^{n} \left[\left\| F_t^{\mathrm{obs}} - \bar{F}_t^{\mathrm{sim}}(\theta) \right\|^2 + K_t(\theta) \right],$$

$$K_t(\theta) = 2 \left(F_{t-L}^{\mathrm{obs}} - \bar{F}_{t-L}^{\mathrm{sim}}(\theta) \right)^{\top} \left(\left[f_t^{\mathrm{obs}} - \bar{f}_t^{\mathrm{sim}}(\theta) \right] - \left[F_{t-L}^{\mathrm{obs}} - \bar{F}_{t-L}^{\mathrm{sim}}(\theta) \right] \right),$$

where $\tau \in \mathbb{N}$ is an initial time-offset and $L \in \mathbb{N}$ is a lag

- In practice, an optimization procedure is used to find the minimizer of $\left| \hat{Q}_n^{\mathrm{RW}}(\cdot) \right|$

Example: Mean of Gaussian

- **Goal:** Estimate $\theta_0 = \mu_0 = 0.5$ given $X_1^{\mathrm{obs}}, \ldots, X_n^{\mathrm{obs}} \overset{\mathrm{iid}}{\sim} \mathcal{N}(\theta_0, 1)$ with $\Theta = [-3, 3]$

Example: Mean of Gaussian

- **Goal:** Estimate $\theta_0 = \mu_0 = 0.5$ given $X_1^{\text{obs}}, \ldots, X_n^{\text{obs}} \overset{\text{iid}}{\sim} \mathcal{N}(\theta_0, 1)$ with $\Theta = [-3, 3]$
- Use time-averages of $2p + 1 = 3$ random Fourier features with $m = 0$ lags

Example: Mean of Gaussian

- **Goal:** Estimate $\theta_0 = \mu_0 = 0.5$ given $X_1^{\text{obs}}, \ldots, X_n^{\text{obs}} \overset{\text{iid}}{\sim} \mathcal{N}(\theta_0, 1)$ with $\Theta = [-3, 3]$
- Use time-averages of $2p + 1 = 3$ random Fourier features with $m = 0$ lags
- For $i = 1, \ldots, 2p + 1$, draw frequencies $w_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$, phases $b_i \overset{\text{iid}}{\sim} \mathrm{Unif}(-\pi, \pi)$

Example: Mean of Gaussian

- **Goal:** Estimate $\theta_0 = \mu_0 = 0.5$ given $X_1^{\text{obs}}, \ldots, X_n^{\text{obs}} \overset{\text{iid}}{\sim} \mathcal{N}(\theta_0, 1)$ with $\Theta = [-3, 3]$
- Use time-averages of $2p + 1 = 3$ random Fourier features with $m = 0$ lags
- For $i = 1, \ldots, 2p + 1$, draw frequencies $w_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$, phases $b_i \overset{\text{iid}}{\sim} \text{Unif}(-\pi, \pi)$
- Define the i-th feature as $\varphi_i(x) = \cos(w_i x + b_i)$

Example: Mean of Gaussian

- **Goal:** Estimate $\theta_0 = \mu_0 = 0.5$ given $X_1^{\text{obs}}, \ldots, X_n^{\text{obs}} \overset{\text{iid}}{\sim} \mathcal{N}(\theta_0, 1)$ with $\Theta = [-3, 3]$
- Use time-averages of $2p + 1 = 3$ random Fourier features with $m = 0$ lags
- For $i = 1, \ldots, 2p + 1$, draw frequencies $w_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$, phases $b_i \overset{\text{iid}}{\sim} \mathrm{Unif}(-\pi, \pi)$
- Define the i-th feature as $\varphi_i(x) = \cos(w_i x + b_i)$
- Denoting $\varphi = (\varphi_1, \ldots, \varphi_{2p+1})$, our estimator is given by

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \left\| \frac{1}{n} \sum_{t=1}^{n} \varphi(X_t^{\text{obs}}) - \frac{1}{s} \sum_{r=1}^{s} \left(\frac{1}{n} \sum_{t=1}^{n} \varphi\left(X_t^{(r)}(\theta) \right) \right) \right\|$$

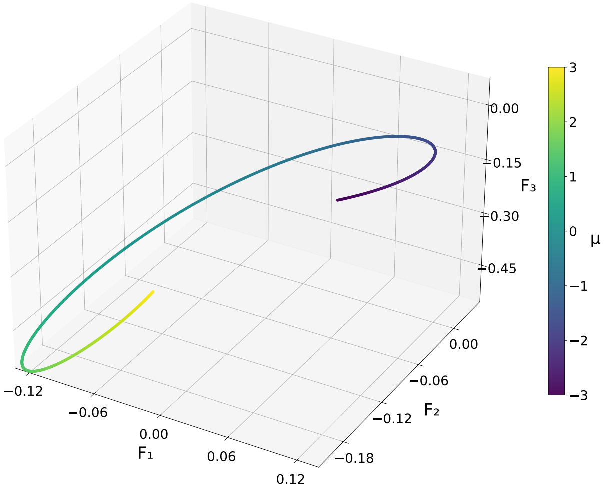
Example: Mean of Gaussian

- **Goal:** Estimate $\theta_0 = \mu_0 = 0.5$ given $X_1^{\mathrm{obs}}, \dots, X_n^{\mathrm{obs}} \overset{\mathrm{iid}}{\sim} \mathcal{N}(\theta_0, 1)$ with $\Theta = [-3, 3]$
- Use time-averages of $2p + 1 = 3$ random Fourier features with $m = 0$ lags
- For $i = 1, \dots, 2p+1$, draw frequencies $w_i \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,1)$, phases $b_i \overset{\mathrm{iid}}{\sim} \mathrm{Unif}(-\pi, \pi)$
- Define the i-th feature as $\varphi_i(x) = \cos(w_i x + b_i)$
- Denoting $\varphi = (\varphi_1, \dots, \varphi_{2p+1})$, our estimator is given by

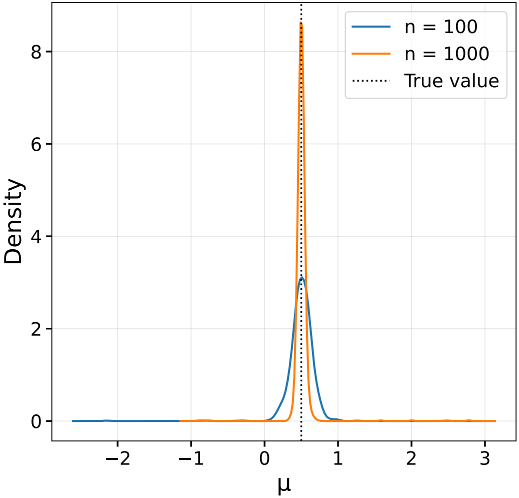
$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \left\| \frac{1}{n} \sum_{t=1}^{n} \varphi(X_t^{\mathrm{obs}}) - \frac{1}{s} \sum_{r=1}^{s} \left(\frac{1}{n} \sum_{t=1}^{n} \varphi\left(X_t^{(r)}(\theta) \right) \right) \right\|$$

- Let's visualize an embedding and check the estimation results

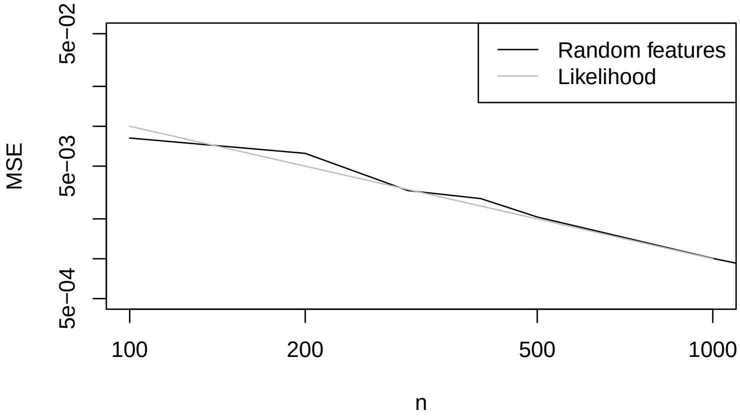
Each point is a time-average of 3 features, colored by value of parameter μ



Density plot from 1,000 independent estimates (different random features)



Theoretical MSE of $\hat{\theta}^{\mathrm{MLE}}$ ($1/n$) vs MSE of our $\hat{\theta}$ (100 replicates per n)



Theoretical guarantees for our estimators

- Guarantees for *time-average estimator* and *rolling-window estimator*

Theoretical guarantees for our estimators

- Guarantees for *time-average estimator* and *rolling-window estimator*
- Consistency $\hat{\theta} \xrightarrow{p} \theta_0$ and asymptotic normality $\sqrt{n}\left(\hat{\theta} - \theta_0\right) \Rightarrow N\left(0, \Sigma\right)$

Theoretical guarantees for our estimators

- Guarantees for *time-average estimator* and *rolling-window estimator*
- Consistency $\hat{\theta} \xrightarrow{p} \theta_0$ and asymptotic normality $\sqrt{n}\left(\hat{\theta} - \theta_0\right) \Rightarrow N\left(0, \Sigma\right)$
- **Key assumptions:** Decay of temporal dependence and control of nonstationarity

Representation in terms of noise inputs (Wu 2005)

- Let $(\varepsilon_i)_{i\in\mathbb{Z}}$ be an iid sequence of random variables and denote $\boldsymbol{\varepsilon}_t = (\varepsilon_t, \varepsilon_{t-1}, \ldots)$

Representation in terms of noise inputs (Wu 2005)

- Let $(\varepsilon_i)_{i\in\mathbb{Z}}$ be an iid sequence of random variables and denote $\boldsymbol{\varepsilon}_t = (\varepsilon_t, \varepsilon_{t-1}, \ldots)$
- Let $\Theta \subset \mathbb{R}^p$ be a parameter space

Representation in terms of noise inputs (Wu 2005)

- Let $(\varepsilon_i)_{i\in\mathbb{Z}}$ be an iid sequence of random variables and denote $\boldsymbol{\varepsilon}_t = (\varepsilon_t, \varepsilon_{t-1}, \ldots)$
- Let $\Theta \subset \mathbb{R}^p$ be a parameter space
- For each $n \in \mathbb{N}$, let $G_t^{(n)} : \mathbb{R}^\infty \times \Theta \to \mathbb{R}^d$, $t = 1, \ldots, n$, define a *generative model*

Representation in terms of noise inputs (Wu 2005)

- Let $(\varepsilon_i)_{i\in\mathbb{Z}}$ be an iid sequence of random variables and denote $\boldsymbol{\varepsilon}_t = (\varepsilon_t, \varepsilon_{t-1}, \ldots)$
- Let $\Theta \subset \mathbb{R}^p$ be a parameter space
- For each $n \in \mathbb{N}$, let $G_t^{(n)} : \mathbb{R}^\infty \times \Theta \to \mathbb{R}^d$, $t = 1, \ldots, n$, define a *generative model*

Assumption

Given a parameter $\theta \in \Theta$, sample size $n \in \mathbb{N}$, and noise inputs $\boldsymbol{\varepsilon}_t$, a time series $X_{1:n}$ is generated as

$$X_t = G_t^{(n)}(\boldsymbol{\varepsilon}_t, \theta), \quad t = 1, \ldots, n$$

Decaying influence of noise inputs (Wu 2005)

- For all $j \in \mathbb{N}$, denote $\boldsymbol{\varepsilon}_t = (\varepsilon_t, \varepsilon_{t-1}, \ldots)$ with ε_{t-j} replaced by an iid copy ε^*_{t-j} as

$$\tilde{\boldsymbol{\varepsilon}}_{t,j} = \left(\varepsilon_t, \ldots, \varepsilon_{t-j+1}, \varepsilon^*_{t-j}, \varepsilon_{t-j-1}, \ldots\right)$$

Decaying influence of noise inputs (Wu 2005)

- For all $j \in \mathbb{N}$, denote $\boldsymbol{\varepsilon}_t = (\varepsilon_t, \varepsilon_{t-1}, \ldots)$ with ε_{t-j} replaced by an iid copy ε^*_{t-j} as

$$\tilde{\boldsymbol{\varepsilon}}_{t,j} = \left(\varepsilon_t, \ldots, \varepsilon_{t-j+1}, \varepsilon^*_{t-j}, \varepsilon_{t-j-1}, \ldots\right)$$

- We assume the influence of the j-th noise input in the past *decays* as j grows

Decaying influence of noise inputs (Wu 2005)

- For all $j \in \mathbb{N}$, denote $\boldsymbol{\varepsilon}_t = (\varepsilon_t, \varepsilon_{t-1}, \dots)$ with ε_{t-j} replaced by an iid copy ε^*_{t-j} as

$$\tilde{\boldsymbol{\varepsilon}}_{t,j} = \left(\varepsilon_t, \dots, \varepsilon_{t-j+1}, \varepsilon^*_{t-j}, \varepsilon_{t-j-1}, \dots\right)$$

- We assume the influence of the j-th noise input in the past *decays* as j grows

Assumption

There exist constants $C > 0$, $\rho > 1$ such that, for some $q > 2$, we have

$$\sup_{n,t,\theta} \left\| G_t^{(n)}(\boldsymbol{\varepsilon}_t, \theta) - G_t^{(n)}(\tilde{\boldsymbol{\varepsilon}}_{t,j}, \theta) \right\|_{L^q(\theta)} \leq C\, j^{-\rho}, \quad j \in \mathbb{N}$$

where $\|\cdot\|_{L^q(\theta)} = \mathbb{E}_{P_\theta}\left(\|\cdot\|_2^q\right)^{1/q}$

Nonstationarity

- When model is not asymptotically stationary, need control of nonstationarity

Nonstationarity

- When model is not asymptotically stationary, need control of nonstationarity
- **Infill asymptotics:** More observations at each local structure as $n \to \infty$

Nonstationarity

- When model is not asymptotically stationary, need control of nonstationarity
- **Infill asymptotics:** More observations at each local structure as $n \to \infty$
- Let $G_t^{(n)} : \mathbb{R}^\infty \times \Theta \to \mathbb{R}^d$, $t = 1, \ldots, n$, and $\tilde{G}_u^{(n)} : \mathbb{R}^\infty \times \Theta \to \mathbb{R}^d$, $u \in [0, 1]$, define a *generative model*

Nonstationarity

- When model is not asymptotically stationary, need control of nonstationarity
- **Infill asymptotics:** More observations at each local structure as $n \to \infty$
- Let $G_t^{(n)} : \mathbb{R}^\infty \times \Theta \to \mathbb{R}^d$, $t = 1, \ldots, n$, and $\tilde{G}_u^{(n)} : \mathbb{R}^\infty \times \Theta \to \mathbb{R}^d$, $u \in [0, 1]$, define a *generative model*

Assumption

Given a parameter $\theta \in \Theta$, sample size $n \in \mathbb{N}$, and noise inputs ε_t, a time series $X_{1:n}$ is generated as

$$X_t = G_t^{(n)}(\varepsilon_t, \theta) = \tilde{G}_{t/n}^{(n)}(\varepsilon_t, \theta), \quad t = 1, \ldots, n$$

Control of nonstationarity

- We measure the nonstationarity using a p-variation-type quantity, define

$$\left\| \left(\tilde{G}_u^{(i)}(\varepsilon_0, \theta) \right)_u \right\|_{p\text{-var}, \mathcal{L}^q(\theta)} = \sup_{\substack{0 = u_0 < \ldots < u_\ell = 1 \\ \ell \in \mathbb{N}}} \left(\sum_{j=1}^{\ell} \left\| \tilde{G}_{u_j}^{(i)}(\varepsilon_0, \theta) - \tilde{G}_{u_{j-1}}^{(i)}(\varepsilon_0, \theta) \right\|_{\mathcal{L}^q(\theta)}^p \right)^{\frac{1}{p}}$$

Control of nonstationarity

- We measure the nonstationarity using a p-variation-type quantity, define

$$\left\| \left(\tilde{G}_u^{(i)}(\varepsilon_0, \theta) \right)_u \right\|_{p\text{-var}, \mathcal{L}^q(\theta)} = \sup_{\substack{0 = u_0 < \ldots < u_\ell = 1 \\ \ell \in \mathbb{N}}} \left(\sum_{j=1}^{\ell} \left\| \tilde{G}_{u_j}^{(i)}(\varepsilon_0, \theta) - \tilde{G}_{u_{j-1}}^{(i)}(\varepsilon_0, \theta) \right\|_{\mathcal{L}^q(\theta)}^p \right)^{\frac{1}{p}}$$

- We impose the following regularity condition to control the nonstationarity

Control of nonstationarity

- We measure the nonstationarity using a p-variation-type quantity, define

$$\left\| \left(\tilde{G}_u^{(i)}(\varepsilon_0, \theta) \right)_u \right\|_{p\text{-var},\mathcal{L}^q(\theta)} = \sup_{\substack{0=u_0<\ldots<u_\ell=1 \\ \ell\in\mathbb{N}}} \left(\sum_{j=1}^{\ell} \left\| \tilde{G}_{u_j}^{(i)}(\varepsilon_0, \theta) - \tilde{G}_{u_{j-1}}^{(i)}(\varepsilon_0, \theta) \right\|_{\mathcal{L}^q(\theta)}^p \right)^{\frac{1}{p}}$$

- We impose the following regularity condition to control the nonstationarity

Assumption

For some $q > 2$, some $p \in [1, 2)$, and all $i \in \mathbb{N}$, there exists a constant $\Lambda > 0$ such that

$$\sup_{\theta\in\Theta} \sup_{u\in[0,1]} \left\| \tilde{G}_u^{(i)}(\varepsilon_0, \theta) \right\|_{\mathcal{L}^q(\theta)} + \sup_{\theta\in\Theta} \left\| \left(\tilde{G}_u^{(i)}(\varepsilon_0, \theta) \right)_u \right\|_{p\text{-var},\mathcal{L}^q(\theta)} \le \Lambda$$

How restrictive are these assumptions?

- **Examples:** Discrete-time dynamical systems and ODEs with observational noise, certain SDEs, linear and nonlinear time series models, and state-space models

How restrictive are these assumptions?

- **Examples:** Discrete-time dynamical systems and ODEs with observational noise, certain SDEs, linear and nonlinear time series models, and state-space models
- Equivalent to β-mixing assumption under certain conditions

How restrictive are these assumptions?

- **Examples:** Discrete-time dynamical systems and ODEs with observational noise, certain SDEs, linear and nonlinear time series models, and state-space models
- Equivalent to β-mixing assumption under certain conditions
- In fact, we can take the random variables $(\varepsilon_\ell)_{\ell \in \mathbb{Z}}$ to be $U[0,1]$

How restrictive are these assumptions?

- **Examples:** Discrete-time dynamical systems and ODEs with observational noise, certain SDEs, linear and nonlinear time series models, and state-space models
- Equivalent to β-mixing assumption under certain conditions
- In fact, we can take the random variables $(\varepsilon_\ell)_{\ell \in \mathbb{Z}}$ to be $U[0,1]$
- **Replication:** Let $G_t^{(n)}$ include compositions with measurable functions to replicate original $U[0,1]$, yielding $U_1, \ldots, U_d \overset{\text{iid}}{\sim} U[0,1]$ (see Ch. 4 Kallenberg (2021))

How restrictive are these assumptions?

- **Examples:** Discrete-time dynamical systems and ODEs with observational noise, certain SDEs, linear and nonlinear time series models, and state-space models
- Equivalent to β-mixing assumption under certain conditions
- In fact, we can take the random variables $(\varepsilon_\ell)_{\ell \in \mathbb{Z}}$ to be $U[0,1]$
- **Replication:** Let $G_t^{(n)}$ include compositions with measurable functions to replicate original $U[0,1]$, yielding $U_1, \ldots, U_d \overset{\text{iid}}{\sim} U[0,1]$ (see Ch. 4 Kallenberg (2021))
- **Rosenblatt transform:** For each $j \in [d]$, sample $X_{t,j}$ from the conditional quantile given $X_{t,1}, \ldots, X_{t,j-1}$ by transforming U_j (see Wu and Mielniczuk (2010))

Ongoing and future work

- Simulation-based *inference* (confidence sets, goodness-of-fit, etc.)

Ongoing and future work

- Simulation-based *inference* (confidence sets, goodness-of-fit, etc.)
- Extensions to spatiotemporal data, network data, etc.

Ongoing and future work

- Simulation-based *inference* (confidence sets, goodness-of-fit, etc.)
- Extensions to spatiotemporal data, network data, etc.
- In what other areas of statistics can random features be used?

Thank you!

Questions and comments are welcome

You can reach me at: mwiecksosa@cmu.edu

Michael Wieck-Sosa
Department of Statistics & Data Science
Carnegie Mellon University

Definition of smooth manifold

Definition

A C^∞-manifold (i.e. smooth manifold or differentiable manifold) is a Hausdorff topological space M, with countable basis, together with a C^∞-atlas

- The next slides explain the definitions needed to understand this
- Source: Lecture notes from Eckhard Meinrenken at the University of Toronto `https://www.math.utoronto.ca/mein/teaching/LectureNotes/Man.pdf`

Definition of smooth manifold

- An *n*-dimensional manifold is a space that locally looks like $\mathbb{R}^n$. To give a precise meaning to this idea, our space first of all has to come equipped with some topology (so that the word "local" makes sense)
- Recall that a *topological space* is a set M, together with a collection of subsets of M, called *open subsets*, satisfying the following three axioms: (i) the empty set $\emptyset$ and the space M itself are both open, (ii) the intersection of any finite collection of open subsets is open, (iii) the union of any collection of open subsets is open
- The collection of open subsets of M is also called the topology of M. A map $f : M_1 \to M_2$ between topological spaces is called *continuous* if the pre-image of any open subset in M_2 is open in M_1. A continuous map with a continuous inverse is called a *homeomorphism*

Definition of smooth manifold

- One ingredient in the definition of a manifold is that our topological space comes equipped with a covering by open sets which are homeomorphic to open subsets of $\mathbb{R}^n$
- Let M be a topological space. An n-dimensional *chart* for M is a pair (U, ϕ) consisting of an open subset $U \subset M$ and a continuous map $\phi : U \to \mathbb{R}^n$ such that ϕ is a homeomorphism onto its image $\phi(U)$. Two such charts (U_α, ϕ_α), (U_β, ϕ_β) are C^∞-compatible if the transition map

$$\phi_\beta \circ \phi_\alpha^{-1} : \phi_\alpha(U_\alpha \cap U_\beta) \to \phi_\beta(U_\alpha \cap U_\beta)$$

is a diffeomorphism (a smooth map with smooth inverse). A covering $\mathcal{A} = (U_\alpha)_{\alpha \in A}$ of M by pairwise C^∞-compatible charts is called a C^∞-atlas

Definition of smooth manifold

- We recall that a topological space is called *Hausdorff* if any two points have disjoint open neighborhoods.
- A *basis* for a topological space M is a collection $\mathcal{B}$ of open subsets of M such that every open subset of M is a union of open subsets in the collection $\mathcal{B}$.
- A topological space with countable basis is also called *second countable*
- For example, the collection of open balls $B_\varepsilon(x)$ in $\mathbb{R}^n$ define a basis. But one already has a basis if one takes only all balls $B_\varepsilon(x)$ with $x \in \mathbb{Q}^n$ and $\varepsilon \in \mathbb{Q}_{>0}$; this defines a countable basis.

References I

- Kallenberg, Olav (2021). *Foundations of modern probability*. Third edition. Springer.
- Sauer, Tim, James A. Yorke, and Martin Casdagli (1991). "Embedology". In: *Journal of Statistical Physics* 65.3, pp. 579–616.
- Wu, Wei Biao (2005). "Nonlinear system theory: another look at dependence". In: *Proceedings of the National Academy of Sciences* 102.40, pp. 14150–14154.
- Wu, Wei Biao and Jan Mielniczuk (2010). "A new look at measuring dependence". In: *Dependence in Probability and Statistics*, pp. 123–142.
- Yorke, James and William Ott (2005). "Prevalence". In: *Bulletin of the American Mathematical Society* 42.3, pp. 263–290.