Statistical Evaluation of Long Memory in Recurrent Neural Networks

Alec Greaves-Tunnell
PhD Student, UW Statistics

SDSS 2019

UNIVERSITY of WASHINGTON

NSF

Plan

Long Memory Processes: Motivation & Background

Semiparametric Estimation

Long Memory in Language, Music, and RNNs

Problem

Given a sequence model trained on data with long-range dependencies, how can we **evaluate** whether these have been successfully learned?

Problem

Given a sequence model trained on data with long-range dependencies, how can we **evaluate** whether these have been successfully learned?

How can we **identify** these long-range dependencies in the first place?

Contributions

- Introduce a framework for evaluation of long memory as a **statistical property** with an existing literature.

Contributions

- Introduce a framework for evaluation of long memory as a **statistical property** with an existing literature.
- Demonstrate practical tools for **estimation** and **hypothesis testing**.

Contributions

- Introduce a framework for evaluation of long memory as a **statistical property** with an existing literature.
- Demonstrate practical tools for **estimation** and **hypothesis testing**.
- Establish **criteria** for long memory in trained RNN models.

Long Memory in the Time Domain

Long Memory in the Time Domain

Stochastic process $X_t \in \mathbb{R}$, $t \in \mathbb{Z}$ has *long memory* if

$$\gamma(k) \triangleq \mathrm{Cov}(X_t, X_{t+k}) = \underbrace{L_\gamma(k)}_{\textbf{slowly varying}} \overbrace{|k|^{-(1-2d)}}^{\textbf{slow decay}}, \quad \text{as } k \to \infty$$

for some $d \in (0, 1/2)$.

Long Memory in the Time Domain

Stochastic process $X_t \in \mathbb{R}$, $t \in \mathbb{Z}$ has *long memory* if

$$\gamma(k) \triangleq \mathrm{Cov}(X_t, X_{t+k}) = \underbrace{L_\gamma(k)}_{\textbf{slowly varying}} \overbrace{|k|^{-(1-2d)}}^{\textbf{slow decay}}, \quad \text{as } k \to \infty$$

for some $d \in (0, 1/2)$.

Slowly varying at infinity:

$$L(xu) \sim L(u) \quad \text{as } x \to \infty \;\; (\gamma(k) \text{ asymptotically } |k|^{-(1-2d)})$$

Long Memory in the Time Domain

Stochastic process $X_t \in \mathbb{R}$, $t \in \mathbb{Z}$ has *long memory* if

$$\gamma(k) \triangleq \mathrm{Cov}(X_t, X_{t+k}) = \underbrace{L_\gamma(k)}_{\text{slowly varying}} \overbrace{|k|^{-(1-2d)}}^{\text{slow decay}}, \quad \text{as } k \to \infty$$

for some $d \in (0, 1/2)$.

Slowly varying at infinity:

$$L(xu) \sim L(u) \quad \text{as } x \to \infty \;\; (\gamma(k) \text{ asymptotically } |k|^{-(1-2d)})$$

Slow decay of autocovariance:

$$\sum_{k=0}^{n} |\gamma(k)| \to \infty \quad \text{as } n \to \infty \;\; (\text{vs. } \gamma(k) \text{ abs. summable})$$

Long Memory in the Frequency Domain

Definitions:

Spectral density function

$$f_X(\lambda) \triangleq \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma(k) e^{-ik\lambda}$$

Long Memory in the Frequency Domain

Definitions:

Spectral density function

$$f_X(\lambda) \triangleq \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma(k) e^{-ik\lambda}$$

Given observations $X_{1:T} = (X_1, ..., X_T)$, define:

Periodogram

$$I(\lambda) \triangleq \frac{1}{2\pi} \sum_{|k|<T}^{\infty} \widehat{\gamma}(k) e^{ik\lambda} = \frac{1}{2\pi T} \left| \sum_{t=1}^{T} X_t e^{-it\lambda} \right|^2,$$

with the second equality holding only at *Fourier frequencies*

$$\lambda_j = 2\pi j/T, \quad j = 1, ..., T.$$

Long Memory in the Frequency Domain

Key idea: $\gamma(k)$ as $k \to \infty \iff f_X(\lambda)$ as $\lambda \to 0$

Long Memory in the Frequency Domain

Key idea: $\gamma(k)$ as $k \to \infty \iff f_X(\lambda)$ as $\lambda \to 0$

Stochastic process $X_t \in \mathbb{R}$, $t \in \mathbb{Z}$ with spectral density function satisfying

$$f_X(\lambda) = L_f(\lambda)|\lambda|^{-2d}$$

$$\text{has } \begin{cases} \textbf{long memory} & \text{if } d \in (0, 1/2) \\ \textbf{short memory} & \text{if } d = 0 \end{cases}.$$

Long Memory in the Frequency Domain

Key idea: $\gamma(k)$ as $k \to \infty \iff f_X(\lambda)$ as $\lambda \to 0$

Stochastic process $X_t \in \mathbb{R}$, $t \in \mathbb{Z}$ with spectral density function satisfying

$$f_X(\lambda) = L_f(\lambda)|\lambda|^{-2d}$$

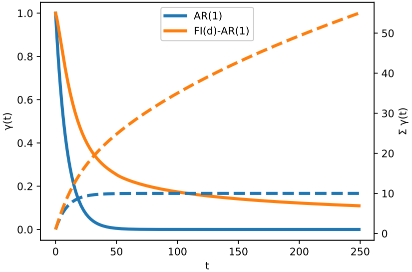
$$\text{has} \begin{cases} \textbf{long memory} & \text{if } d \in (0, 1/2) \\ \textbf{short memory} & \text{if } d = 0 \end{cases}.$$

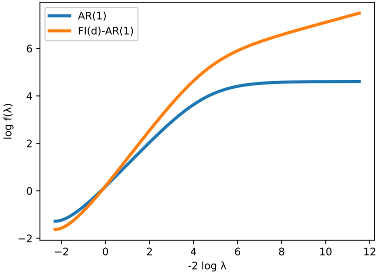
Note: long memory parameter d is slope of log-log plot:

$$\log f_X(\lambda) \approx -2d \log \lambda \quad \text{as } \lambda \to 0 \ \ (\text{i.e.} -2\log\lambda \to \infty).$$

Simple Illustration

Short memory AR process vs **long memory** FI-AR process:





Left: time domain. **Right**: frequency domain.

Plan

Long Memory Processes: Motivation & Background

Semiparametric Estimation

Long Memory in Language, Music, and RNNs

Semiparametric estimation

What:

$\rightarrow$ Investigate some feature in data without fully specifying joint distribution

$\rightarrow$ Finite parameter of interest (**long memory parameter**), infinite-dimensional nuisance parameter (**full spectral density**)

Semiparametric estimation

What:

$\rightarrow$ Investigate some feature in data without fully specifying joint distribution

$\rightarrow$ Finite parameter of interest (**long memory parameter**), infinite-dimensional nuisance parameter (**full spectral density**)

Why:

$\rightarrow$ Robust to misspecification of short-term behavior

$\rightarrow$ Computationally efficient even for very long sequences

Semiparametric estimation

How:
Spectral approx. near zero frequency:

$$f_X(\lambda) = \Lambda(d) G \Lambda(d)^*, \quad \Lambda(d) \triangleq \mathrm{diag}(\lambda^{-d} e^{i(\pi-\lambda)/2})$$

with *long run covariance* G real, symmetric, positive definite.

Semiparametric estimation

How:
Spectral approx. near zero frequency:

$$f_X(\lambda) = \Lambda(d) G \Lambda(d)^*, \quad \Lambda(d) \triangleq \operatorname{diag}(\lambda^{-d} e^{i(\pi-\lambda)/2})$$

with *long run covariance* G real, symmetric, positive definite.

Maximize **local** **Whittle** **profile** likelihood

$$\mathcal{L}_{\mathbf{m}}(\widehat{G}(d), d) = \frac{1}{\mathbf{m}} \sum_{j=1}^{\mathbf{m}} \Big[\log\det \Lambda_j(d) \widehat{G}(d) \Lambda_j^*(d) + \mathbf{Tr} \left[\left(\Lambda_j(d) \widehat{G}(d) \Lambda_j^*(d) \right)^{-1} I(\lambda_j) \right] \Big].$$

Gaussian Semiparametric Estimator

The *Gaussian semiparametric estimator* is

$$\hat{d}_{\mathsf{GSE}} = \arg\min_{d \in \Theta} \mathcal{L}_m(d)$$

with $\Theta = (-1/2, 1/2)^p$.

Gaussian Semiparametric Estimator

The *Gaussian semiparametric estimator* is

$$\hat{d}_{\mathrm{GSE}} = \arg\min_{d \in \Theta} \mathcal{L}_m(d)$$

with $\Theta = (-1/2, 1/2)^p$.

Asymptotic normality [Shimotsu, 2007]: Let $X_t \in \mathbb{R}^p$ have long memory d_0 and long-run covariance G, and define

$$\Omega = 2\left[I_p + G \odot G^{-1} + \frac{\pi^2}{4}(G \odot G^{-1} - I_p) \right].$$

Then

$$\sqrt{m}(\hat{d}_{\mathrm{GSE}} - d_0) \to_d \mathcal{N}(0, \Omega^{-1}).$$

Plan

Long Memory Processes: Motivation & Background

Semiparametric Estimation

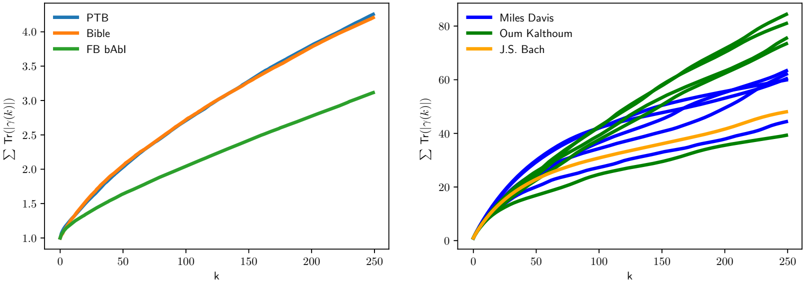
Long Memory in Language, Music, and RNNs

Estimating long memory of sequence data

Do language and music data have long memory?

Estimating long memory of sequence data

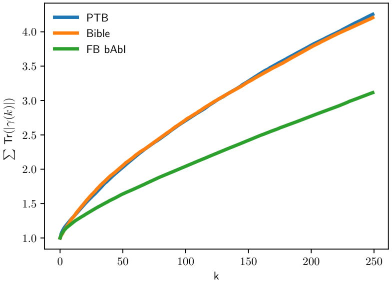
Do language and music data have long memory?

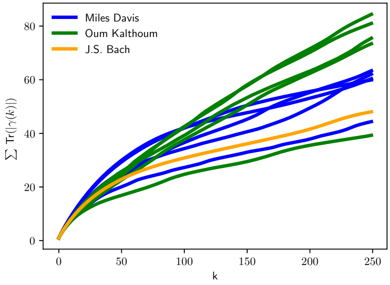


Left: language data. **Right**: music data.

Estimating long memory of sequence data

Do language and music data have long memory?





Left: language data. **Right**: music data.

Semiparametric estimation and testing confirm long memory suggested by visual heuristic.

Long memory criterion for RNNs

ARFIMA model:
Represent long memory X_t via **linear filtering** and **fractional integration** of white noise Z_t

$$X_t = \overbrace{(1-B)^{-d}}^{\textbf{long memory}} \underbrace{\Phi^{-1}(B)\Theta(B)}_{\textbf{linear features}} Z_t$$

Long memory criterion for RNNs

ARFIMA model:

Represent long memory X_t via **linear filtering** and **fractional integration** of white noise Z_t

$$X_t = \overbrace{(1 - B)^{-d}}^{\text{long memory}} \underbrace{\Phi^{-1}(B)\Theta(B)}_{\text{linear features}} Z_t$$

RNN: Study the stochastic process

$$X_t = \Psi(Z_t)$$

with $\Psi(\cdot)$ the learned RNN transformation of inputs to hidden features.

Long memory criterion for RNNs

A simple criterion:

Do we have

$$X_t = \Psi(Z_t) = \overbrace{(1-B)^{-d}}^{\textbf{long memory}} \underbrace{\tilde{\Psi}(Z_t)}_{\textbf{nonlinear, short memory}}$$

for some $d \neq 0$ and short memory $\tilde{\Psi}(Z_t)$?

Long memory criterion for RNNs

A simple criterion:
Do we have

$$X_t = \Psi(Z_t) = \overbrace{(1-B)^{-d}}^{\text{long memory}} \underbrace{\tilde{\Psi}(Z_t)}_{\text{nonlinear, short memory}}$$

for some $d \neq 0$ and short memory $\tilde{\Psi}(Z_t)$?
How to evaluate:

1. Train RNN model(s) to benchmark accuracy on long memory data
2. Generate from $X_t = \Psi(Z_t)$ by computing RNN hidden representation of white noise
3. Estimate and test for long memory with GSE

Results: RNN models

Hypothesis test for long memory:

$$\mathcal{H}_0 : \bar{d} = 0 \quad \text{vs.} \quad \overbrace{\mathcal{H}_1 : \bar{d} > 0}^{\textbf{expected result}} .$$

Results: RNN models

Hypothesis test for long memory:

$$\overbrace{\mathcal{H}_0 : \bar{d} = 0}^{\textbf{observed result}} \quad \text{vs. } \mathcal{H}_1 : \bar{d} > 0.$$

Total Memory in RNN Representations of White Noise Input.

Model	**Norm. total memory**	**p-value**	**Reject $\mathcal{H}_0$?**
LSTM (trained)	-8.59×10^{-4}	0.583	X
LSTM (untrained)	-4.17×10^{-4}	0.572	X
Memory cell	-5.96×10^{-4}	0.552	X
SCRN	2.37×10^{-3}	0.324	X

References

Paper:

Greaves-Tunnell, A, and Harchaoui, H. "A Statistical Investigation of Long Memory in Language and Music." In *ICML*. 2019.

Further references:

Beran, J., Feng, Y., Ghosh, S., and Kulik, R. *Long-Memory Processes: Probabilistic Properties and Statistical Methods*. Springer, 2013.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 9(8):1735−1780, 1997.

Levy, O., Lee, K., FitzGerald, N., and Zettlemoyer, L. Long short-term memory as a dynamically computed elementwise weighted sum. In *ACL*, 2018.

Mikolov, T., Joulin, A., Chopra, S., Mathieu, M., and Ranzato, M. Learning longer memory in recurrent neural networks. In *ICLR*, 2015.

Shimotsu, K. Gaussian semiparametric estimation of multivariate fractionally integrated processes. *Journal of Econometrics*, 137(2):277−310, 2007.

Thanks!