Monotonicity of the Trace-Inverse of Covariance Submatrices and Two-Sided Prediction

Anatoly Khina

Joint work with Arie Yeredor & Ram Zamir

Tel Aviv University

ISIT 2022
Espoo, Finland

Outline

- Introduction: differential entropy as a measure of “memory strength”
- Trace–inverse of the precision matrix
  - Characterization via the eigenvalues of the covariance matrix
  - Characterization via estimation
- Monotonicity of the precision matrix trace–inverse
- Trace–inverse rate
  - Relation to two-sided prediction
- Spectral estimation: Max entropy principle vs. Min trace–inverse principle
- Example: Autoregressive processes

Introduction: Differential Entropy Measure

- $\{X_n\}$ is a **stationary** process with finite second moment

Normalized differential entropy: $\bar{h}_n \triangleq \frac{1}{n} h(X_1, X_2, \dots, X_n)$

Prediction gain: $D_n \triangleq \bar{h}_1 - \bar{h}_n = \frac{1}{n}\mathbb{D}\big(p(x_1, \dots, x_n) || p(x_1) \times \cdots \times p(x_n)\big) \geq 0$

- For a **Gaussian** process:

$$D_n^G = \frac{1}{2}\log\frac{\mathrm{Var}(X_1)}{|\mathbf{C}_n|^{1/n}} = \frac{1}{2}\log\frac{\frac{1}{n}\mathrm{tr}(\mathbf{C}_n)}{|\mathbf{C}_n|^{1/n}} = \frac{1}{2}\log\frac{\frac{1}{n}\sum_{i=1}^{n}\lambda_i}{\left(\prod_{i=1}^{n}\lambda_i\right)^{1/n}}$$

- $\mathbf{C}_n \triangleq \mathrm{Cov}(X_1, \dots, X_n)$—covariance of n consecutive samples
- $\{\lambda_i\}$—the eigenvalues of $\mathbf{C}_n = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, $\mathbf{U}$—orthogonal, $\boldsymbol{\Lambda}$—positive diagonal
- $D_n^G = 0$ iff all λ_i are equal $\Leftrightarrow$ $(X_1, X_2, \dots, X_n)$ is a white vector

Introduction: Differential Entropy Measure

- By the chain rule:

$$n\bar{h}_n = h(X_1) + h(X_2|X_1) + \cdots + h(X_n|X_{n-1}, X_{n-2}, \dots, X_1)$$

- For a Gaussian process:

$$D_n^G = \frac{1}{2}\log\frac{\mathrm{Var}(X_1)}{|\mathbf{C}_n|^{1/n}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\log\frac{\mathrm{Var}(X_i)}{\overline{\mathcal{E}^2}(X_i|X_{i-1}, X_{i-2}, \dots, X_1)}$$

- $\overline{\mathcal{E}^2}(X_i|X_{i-1}, X_{i-2}, \dots, X_1) \triangleq$ prediction (L)MMSE of X_i given $X_{i-1}, X_{i-2}, \dots, X_1$
- $D_n^G = 0$ iff $\overline{\mathcal{E}^2}(X_i|X_{i-1}, X_{i-2}, \dots, X_1) = \mathrm{Var}(X_i) \quad \forall i$
  $\Leftrightarrow (X_1, X_2, \dots, X_n)$ is a white vector

Trace–Inverse (Tin) of a Precision Matrix

- $\mathbf{C}_n \triangleq \mathrm{Cov}(X_1, \dots, X_n)$—covariance of n consecutive samples
- $\mathbf{C}_n^{-1}$—the inverse of $\mathbf{C}_n$ a.k.a. the *precision matrix* [Gauss 1809]

Normalized trace–inverse (Tin): $M_n \triangleq \frac{1}{n}\mathrm{tr}(\mathbf{C}_n^{-1})$

Trace–Inverse (Tin) of a Precision Matrix

- $\mathbf{C}_n \triangleq \mathrm{Cov}(X_1, \dots, X_n)$—covariance of n consecutive samples
- $\mathbf{C}_n^{-1}$—the inverse of $\mathbf{C}_n$ a.k.a. the *precision matrix* [Gauss 1809]

Normalized trace–Inverse (Tin): $M_n \triangleq \frac{1}{n}\mathrm{tr}(\mathbf{C}_n^{-1})$

- $\mathbf{C}_n = \mathbf{U\Lambda U}^T$, $\mathbf{U}$—orthogonal, $\mathbf{\Lambda}$—positive diagonal
- $\mathbf{C}_n^{-1} = \mathbf{U\Lambda}^{-1}\mathbf{U}^T \Rightarrow$ $M_n \triangleq \frac{1}{n}\mathrm{tr}(\mathbf{C}_n^{-1}) = \frac{1}{n}\mathrm{tr}\left(\mathbf{U\Lambda}^{-1}\mathbf{U}^T\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\lambda_n}$

- $1/M_n$ = harmonic mean of the eigenvalues (spectrum) of $\mathbf{C}_n$
- $|\mathbf{C}_n|^{1/n}$ = geometric mean of the eigenvalues (spectrum) of $\mathbf{C}_n$

Trace–Inverse (Tin) of a Precision Matrix

Lemma: The i^{th} diagonal entry of $\mathbf{C}_n^{-1}$ equals

$$[\mathbf{C}_n^{-1}]_{i,i} = \frac{1}{\overline{\mathcal{E}^2}(X_i | X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)}$$

- Proved in [Kay TSP'83] using Lagrange multiplies
- New proof via the partition matrix inversion lemma / Schur's complement

Trace–Inverse (Tin) of a Precision Matrix

Lemma: The i^{th} diagonal entry of $\mathbf{C}_n^{-1}$ equals

$$[\mathbf{C}_n^{-1}]_{i,i} = \frac{1}{\overline{\mathcal{E}^2}(X_i|X_1, X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_n)}$$

- Proved in [Kay TSP’83] using Lagrange multiplies
- New proof via the partition matrix inversion lemma / Schur’s complement

Corollary: $M_n \triangleq \frac{1}{n}\mathrm{tr}(\mathbf{C}_n^{-1}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\overline{\mathcal{E}^2}(X_i|X_1, X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_n)}$

- $1/M_n \quad =$ harmonic mean of the MMSEs of each X_i given its **past & future**
- $|\mathbf{C}_n|^{1/n} \propto$ geometric mean of the MMSEs of each X_i given its **past**

Normalized-Tin Monotonicity

Normalized Tin: $M_n \triangleq \frac{1}{n}\mathrm{tr}(\mathbf{C}_n^{-1}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\overline{\varepsilon^2}(X_i|X_1, X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_n)}$

Theorem: The sequence $\{M_n | n \in \mathbb{N}\}$ is monotonically non-decreasing:

$$M_n \leq M_{n+1}$$

with equality iff one of the following holds:

- $(X_1, \dots, X_{n+1})$ is white $\Leftrightarrow$ $M_1 = M_2 = \cdots = M_n = M_{n+1}$
- $\mathbf{C}_n$ is singular $\Leftrightarrow$ $M_n = \infty$ and then also $M_{n+1} = \infty$

- **Proof 1:** Using last corollary + simple averaging and MMSE arguments
- **Proof 2:** Via AR modeling (even if $\{X_n\}$ not an AR process)

Tin Rate & Two-Sided Prediction

At present, the future is just as important as the past

Tin Rate: Infinite-Order Normalized Tin

- We have proved that the n^{th} order normalized Tin equals

$$M_n \triangleq \frac{1}{n}\operatorname{tr}(\boldsymbol{C}_n^{-1}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\overline{\mathcal{E}^2}(X_i|X_1, X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_n)} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\lambda_n}$$

Tin rate (infinite-order normalized Tin): $M_\infty = \lim_{n\to\infty} M_n$

Lemma: $M_\infty = 1 \,/\, \overleftrightarrow{\mathcal{E}^2} \overset{(\star)}{=} \int_{-1/2}^{1/2} df \,/ S(\mathrm{e}^{j2\pi f})$

- $\overleftrightarrow{\mathcal{E}^2} \triangleq \overline{\mathcal{E}^2}(X_0| \dots, X_{-2}, X_{-1}, X_1, X_2, \dots)$—two-sided prediction (TSP) LMMSE
- S—power spectral density of the process X

- (⋆) was previously proved by [Kolmogorov ’39, ‘41][Grenander–Szegö ‘58] [Rozanov ‘67][Kay TSP’83][Picinbono ’86]

Tin Rate: Infinite-Order Normalized Tin

Szegö–Kolmogorov Theorem: $\overleftarrow{\mathcal{E}^2} = \exp\left\{\int_{-1/2}^{1/2} \log S\left(\mathrm{e}^{j2\pi f}\right) df\right\}$

- $\overleftarrow{\mathcal{E}^2} \triangleq \mathrm{MMSE}(X_0 | \dots, X_{-2}, X_{-1})$—one-sided prediction (OSP) MMSE

Finite order:

- $1/M_n$ = harmonic mean of the eigenvalues (spectrum) of $\mathbf{C}_n$
- $|\mathbf{C}_n|^{1/n}$ = geometric mean of the eigenvalues (spectrum) of $\mathbf{C}_n$

Infinite order:

- $1/M_\infty = \overleftrightarrow{\mathcal{E}^2}$ = harmonic mean of the spectrum S
- $\frac{\exp\{2h(\mathcal{X})\}}{2\pi\mathrm{e}} = \overleftarrow{\mathcal{E}^2}$ = geometric mean of the spectrum S

OSP vs. TSP Criteria: Autoregressive Processes

Autoregressive (AR) process:

$$\overbrace{a_0}^{=1} X_i = -\sum_{\ell=1}^{p} a_\ell X_{i-\ell} + W_i \quad \Leftrightarrow \quad \sum_{\ell=1}^{p} a_\ell X_{i-\ell} + W_i = 0$$

- p is the order of the AR process if $a_p \neq 0$ ($a_0 = 1$)
- $\{W_i\}$ is white

Spectrum: $$S\left(\mathrm{e}^{j2\pi f}\right) = \frac{1}{\sum_{\ell=0}^{m-1} \lambda_\ell \cos(2\pi\ell f)} = \frac{\gamma}{\prod_{k=1}^{m-1} \left|1 - \xi_k \mathrm{e}^{j2\pi f}\right|^2}$$

$\sigma_W^2, \{a_\ell\} \quad \Leftrightarrow \quad \{\lambda_\ell\}$ (equivalently, γ, $\{\xi_\ell\}$)

Yule–Walker Theorem: Let $\mathbf{C}_{n+1}$ be some covariance (of dim. $n+1$).

$\Rightarrow$ There exists an AR process of order up to n that is consistent with $\mathbf{C}_{n+1}$.

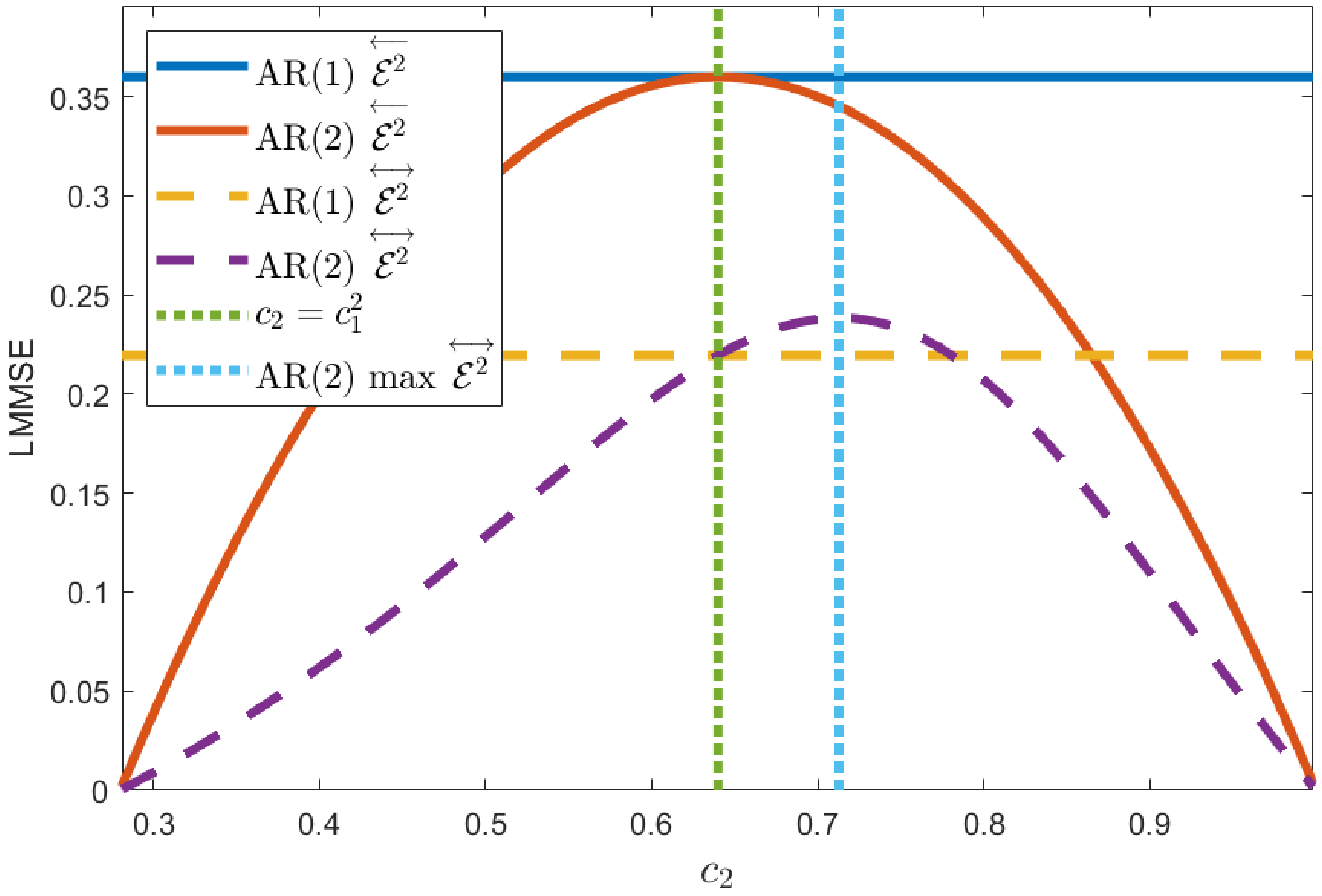
- Yule–Walker equations: $c_0, \dots, c_n \quad \Leftrightarrow \quad \sigma_W^2$ and $a_1, \dots, a_n$ where $c_i \triangleq \mathrm{Cov}(X_0, X_i)$

OSP vs. TSP Criteria: Autoregressive Processes

- AR(2): $X_i = -a_1 X_{i-1} - a_2 X_{i-2} + W_i$
- AR(1): $X_i = -a_1 X_{i-1} + W_i$
- $c_i \triangleq \mathrm{Cov}(X_0, X_i)$
- For $c_2 = c_1^2 \;\Rightarrow\; a_2 = 0$

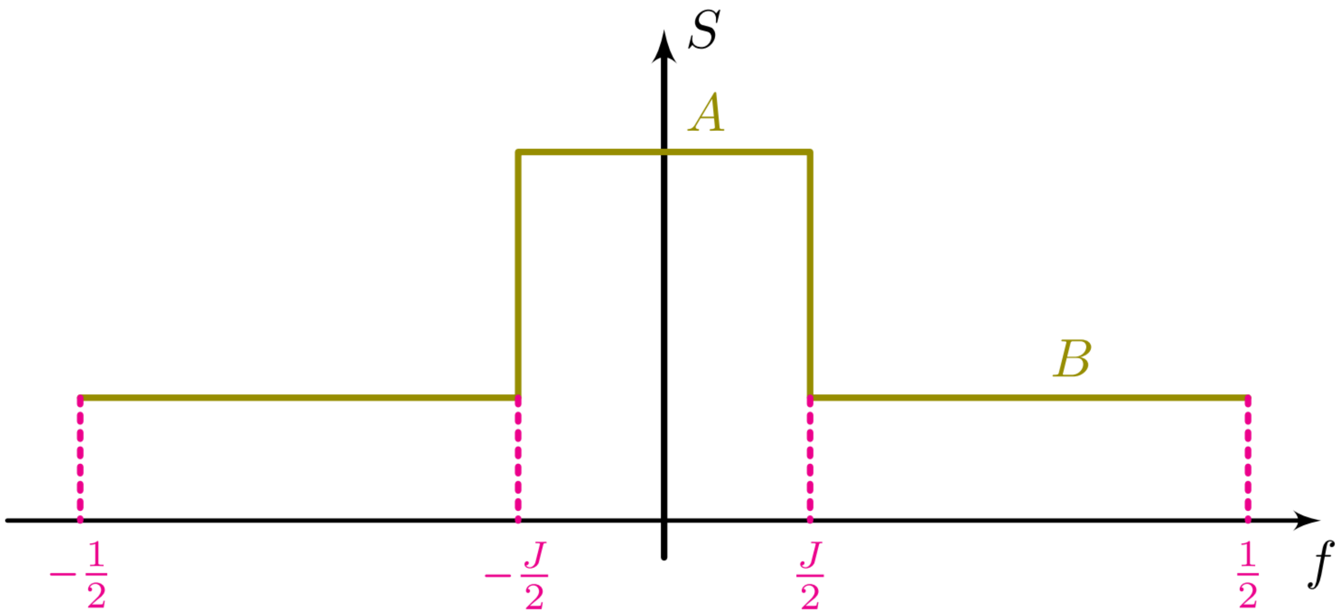
Opposing results of OSP and TSP criteria

$c_0 = 1, c_1 = 0.8$

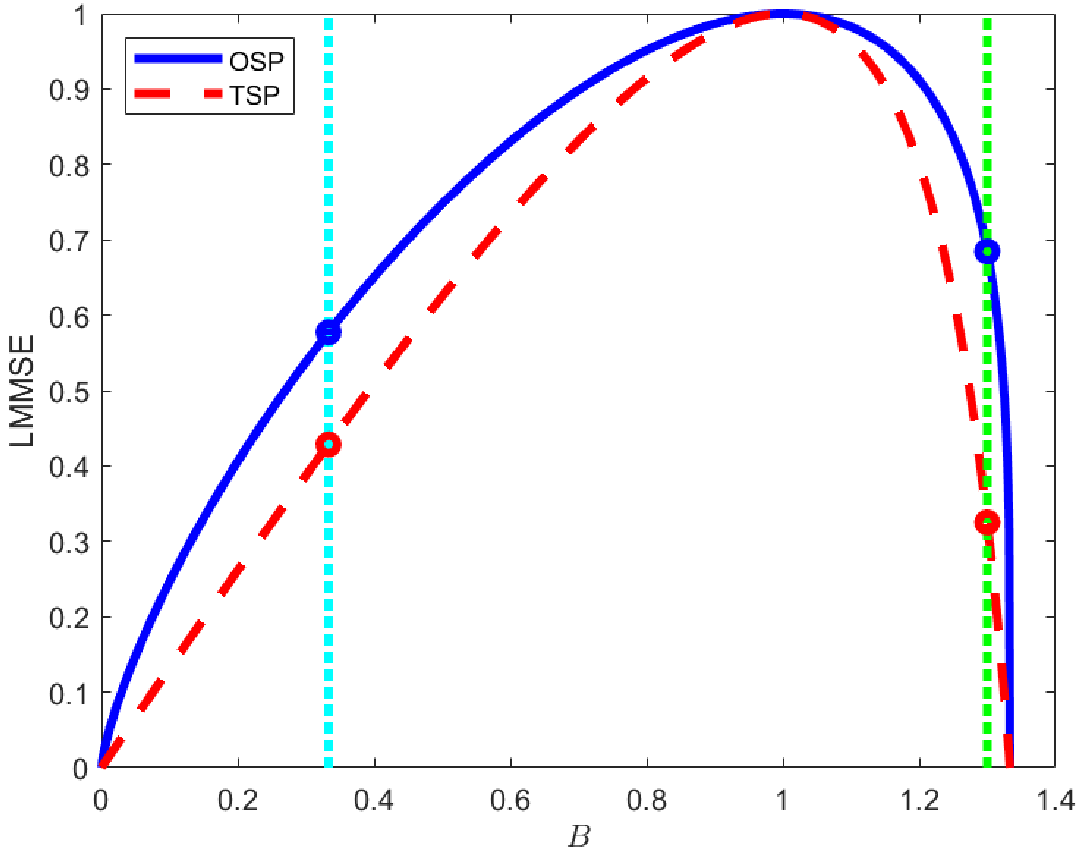


OSP vs. TSP Criteria: Step Spectra

- $JA + (1 - J)B = c_0 = 1$



Opposing results of OSP and TSP criteria



Spectrum Completion via Maximum Entropy

- **Constraints: $\mathbf{C}_m$** $\Leftrightarrow$ $c_0, c_1, \dots, c_{m-1}$
- How to complete the covariance/spectrum function?

Burg's maximum entropy principle: Use the maximum entropy rate stochastic process that is consistent with the constraints.

- The MaxEnt process is the m^{th}-order Gaussian AR process with $\mathbf{C}_m$
- Its spectrum is

$$S\left(\mathrm{e}^{j2\pi f}\right) = \frac{1}{\sum_{\ell=0}^{m-1} \lambda_\ell \cos(2\pi\ell f)} = \frac{\gamma}{\prod_{k=1}^{m-1} \left|1 - \xi_k \mathrm{e}^{j2\pi f}\right|^2}$$

consistent with $\int_{1/2}^{1/2} S\left(\mathrm{e}^{j2\pi f}\right) \cos(2\pi\ell f) = c_\ell\,, \quad \ell = 0, \dots, m-1 \;\Rightarrow\; \{\lambda_\ell\}$

- Proof relies on simple information-theoretic properties

Spectrum Completion via Minimum Tin

- **Constraints:** $\mathbf{C}_m \quad \Leftrightarrow \quad c_0, c_1, \dots, c_{m-1}$
- How to complete covariance/spectrum function?

Minimum normalized Tin principle: Use the minimum Tin rate stochastic process that is consistent with the constraints.

- The MinTin process is the m^{th}-order Gaussian Root AR (RAR) process with $\mathbf{C}_m$
- Its spectrum is

$$S\left(\mathrm{e}^{j2\pi f}\right) = \frac{1}{\sqrt{\sum_{\ell=0}^{m-1} \lambda_\ell \cos(2\pi\ell f)}} = \frac{\gamma}{\prod_{k=1}^{m-1} \left|1 - \xi_k \mathrm{e}^{j2\pi f}\right|}$$

consistent with $\int_{1/2}^{1/2} S\left(\mathrm{e}^{j2\pi f}\right) \cos(2\pi\ell f) = c_\ell\,, \quad \ell = 0, \dots, m-1 \;\Rightarrow\; \{\lambda_\ell\}$

- $\mathbf{C}_m$ doesn't determine say $\mathrm{Cov}(X_{m-1}, X_{-m+1})$ $\Rightarrow$ Proof via calculus of variations

AR Processes

Gaussian AR Process of order p

- $M_n \triangleq \frac{1}{n}\operatorname{tr}(\boldsymbol{C}_n^{-1}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\overline{\mathcal{E}^2}(X_i | X_1, X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_n)}$

- Spectrum of an AR process of order p: $S\left(\mathrm{e}^{j2\pi f}\right) = 1 / \sum_{\ell=0}^{p-1} \lambda_\ell \cos(2\pi \ell f)$

- For (Gaussian) AR process of order p is Markov of order p: $\overleftarrow{\mathcal{E}^2} = \sigma_W^2, \quad \overleftrightarrow{\mathcal{E}^2} = \frac{\sigma_W^2}{\sum_{\ell=0}^{p} a_\ell^2}$

One-step ("greedy") covariance completion: Given $\mathbf{C}_{p+1} \quad \Leftrightarrow \quad c_0, c_1, \dots, c_p$

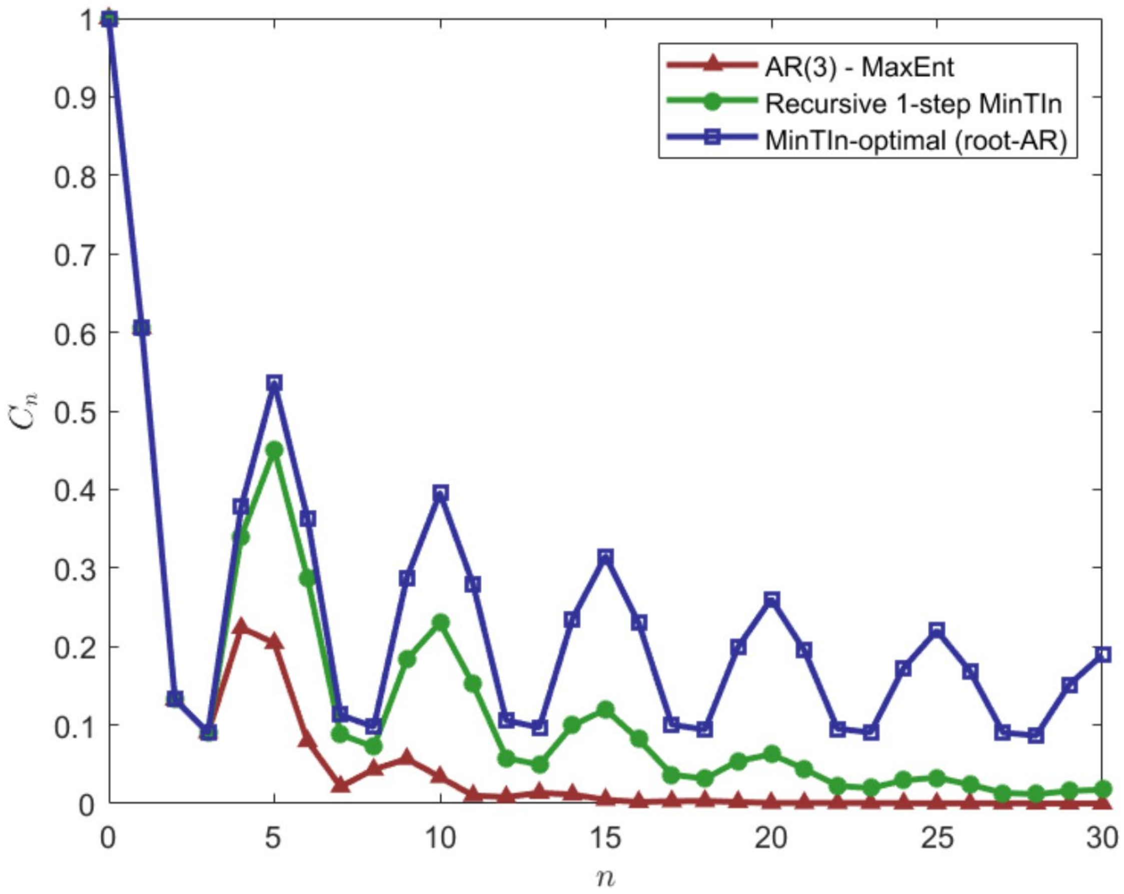
$$c_{p+1}^{\text{MaxEnt}} = c_{p+1}^{\text{MaxOSP}} = -\sum_{\ell=1}^{p} a_\ell c_{p+1-\ell} \quad \text{(Yule–Walker eqs.)}$$

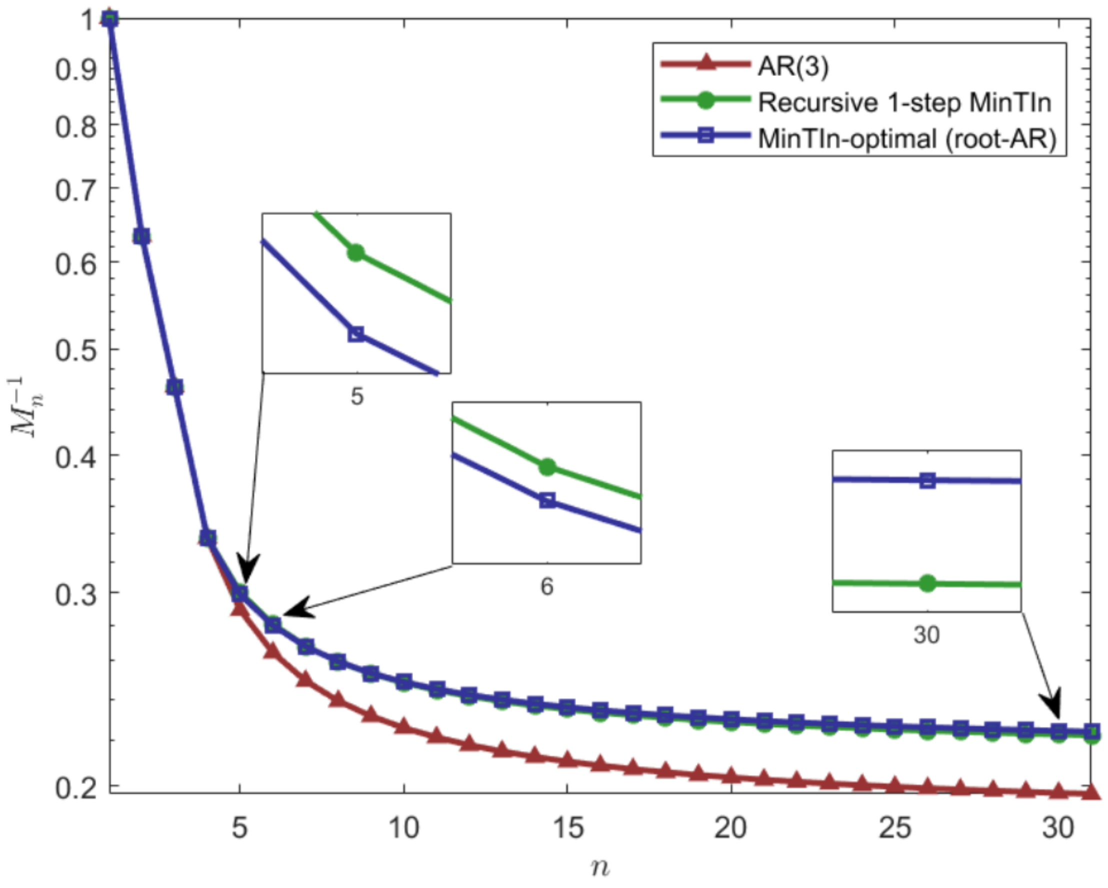
$$c_{p+1}^{\text{MinTin}} = c_{p+1}^{\text{MaxTSP}} = c_{p+1}^{\text{MaxEnt}} + \left(\alpha - \operatorname{sign}\{\alpha\}\sqrt{\alpha^2 - 1}\right)\sigma_W^2$$

where $\alpha \triangleq \frac{\sum_{\ell=0}^{p} a_\ell^2}{\sum_{\ell=1}^{p} a_\ell a_{p+1-\ell}}$

Gaussian AR Process of order p

- $c_0 = 1, c_1 \approx 0.6054, c_2 \approx 0.1324 \Rightarrow$ MinTin yields a RAR process with
- $S(e^{j2\pi f}) = \frac{\gamma}{|1-\xi_1 e^{-j2\pi f}||1-\xi_1^* e^{-j2\pi f}||1-\xi_2 e^{-j2\pi f}|}, \xi_1 = 0.97e^{j0.4\pi},\ \xi_2 = 0.99,\ \gamma \approx 0.4062$





Summary

- Normalized Tin is intimately related to TSP
  - Normalized diff entropy relates to OSP
- Normalized Tin is monotonic, similarly to normalized diff entropy
- Alternative measure to memory strength
- May be used for spectrum estimation/completion

Summary

- Normalized Tin is intimately related to TSP
  - Normalized diff entropy relates to OSP
- Normalized Tin is monotonic, similarly to normalized diff entropy
- Alternative measure to memory strength
- May be used for spectrum estimation/completion
- May be used as oracle in online/causal scenarios with regret
- Would be interesting to generalize to “partially-observable” setup:

  Two-sided estimation given noisy measurements
- Generalization of Tin beyond second-order statistics

Backup Slides

Trace–Inverse (Tin) of a Precision Matrix

Lemma: The i^{th} diagonal entry of $\mathbf{C}_n^{-1}$ equals

$$[\mathbf{C}_n^{-1}]_{i,i} = \frac{1}{\overline{\mathcal{E}^2}(X_i | X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)}$$

- Proved in [Kay TSP'83] using Lagrange multipliers

Alternative proof: Follows from the partition matrix inversion Lemma:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \quad \Rightarrow \quad \mathbf{A}^{-1} = \begin{bmatrix} (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}\mathbf{A}_{21})^{-1} & * \\ * & (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}\mathbf{A}_{12})^{-1} \end{bmatrix}$$

- Plugging in $\mathbf{A} = \mathbf{C}_n$ and $\mathbf{A}_{11} = [\mathbf{C}_n]_{1,1}$ proves the lemma for $i = 1$
- By rearranging the entries of the random vector $(X_1, \ldots, X_n)$ yields the lemma $\forall i$

Normalized-Tin Monotonicity via MMSE Estimation

Lemma: Since more observations can only reduce the MMSE:

$$\overline{\mathcal{E}^2}(X_i|X_1,X_2,\ldots,X_{i-1},X_{i+1},X_{i+2},\ldots,X_n) \geq \overline{\mathcal{E}^2}(X_i|X_1,X_2,\ldots,X_{i-1},X_{i+1},X_{i+2},\ldots,X_n,X_{n+1})$$

$$\overline{\mathcal{E}^2}(X_i|X_1,X_2,\ldots,X_{i-1},X_{i+1},X_{i+2},\ldots,X_n) \geq \overline{\mathcal{E}^2}(X_i|X_1,X_2,\ldots,X_{i-1},X_i,X_{i+2},\ldots,X_n,X_{n+1})$$

Normalized-Tin Monotonicity via MMSE Estimation

Lemma: Since more observations can only reduce the MMSE:

$$\overline{\mathcal{E}^2}(X_i|X_1,X_2,\dots,X_{i-1},X_{i+1},X_{i+2},\dots,X_n) \geq \overline{\mathcal{E}^2}(X_i|X_1,X_2,\dots,X_{i-1},X_{i+1},X_{i+2},\dots,X_n,X_{n+1})$$

$$\overline{\mathcal{E}^2}(X_i|X_1,X_2,\dots,X_{i-1},X_{i+1},X_{i+2},\dots,X_n) \geq \overline{\mathcal{E}^2}(X_i|X_1,X_2,\dots,X_{i-1},X_i,X_{i+2},\dots,X_n,X_{n+1})$$

Proof of M_n monotonicity: $M_n = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\overline{\mathcal{E}^2}(X_i|X_1,X_2,\dots,X_{i-1},X_{i+1},\dots,X_n)}$

- M_n equals the mean of $1/\overline{\mathcal{E}^2}(X_i|X_1,X_2,\dots,X_{i-1},X_{i+1},\dots,X_n)$

$\Rightarrow$ one of the elements is at least as large as the mean:

$$\exists i: \quad 1/\overline{\mathcal{E}^2}(X_i|X_1,X_2,\dots,X_{i-1},X_{i+1},\dots,X_n) \geq M_n$$

$$\Rightarrow \quad (n+1)M_{n+1} \geq nM_n + 1/\overline{\mathcal{E}^2}(X_i|X_1,X_2,\dots,X_{i-1},X_{i+1},\dots,X_n) \geq (n+1)M_n$$

Normalized-Tin Monotonicity via Autoregressive Modeling

Autoregressive (AR) process:

$$\overbrace{a_0}^{=1} X_i = -\sum_{\ell=1}^{p} a_\ell X_{i-\ell} + W_i \quad \Leftrightarrow \quad \sum_{\ell=1}^{p} a_\ell X_{i-\ell} + W_i = 0$$

- $a_0 = 1$
- p is the order of the AR process if $a_p \neq 0$
- $\{W_i\}$ is white

Yule–Walker Theorem: Let $\mathbf{C}_{n+1}$ be some covariance (of dim. $n + 1$).

$\Rightarrow$ There exists an AR process of order up to n that is consistent with $\mathbf{C}_{n+1}$.

- σ_W^2 and $a_1, \ldots, a_n$ can be found via the Yule–Walker equations

Normalized-Tin Monotonicity via Autoregressive Modeling

Theorem [Siddiqui ’58][Galbraith–Galbraith ‘74][Wise ‘55][Champernowne ‘48]:

For an AR process of order $p \le n$:

$$\sigma_W^2 \cdot [\mathbf{C}_n^{-1}]_{i,j} = \sum_{\ell=0}^{i-1} a_\ell a_{\ell+j-i} - \sum_{\ell=n+1-j}^{n+i-j} a_\ell a_{\ell+j-i}\,, \qquad 1 \le i \le j \le n$$

$$[\mathbf{C}_n^{-1}]_{j,i} = [\mathbf{C}_n^{-1}]_{i,j}$$

- $a_\ell = 0$ for $p < \ell \le n$

✓A new proof via the Gohberg–Semençul formula

Corollary: For an AR process of order $p \le n$ $M_n = \frac{1}{\sigma_W^2} \sum_{\ell=0}^{n} \left(1 - \frac{2\ell}{n}\right) a_\ell^2$

Proof of M_n monotonicity: $M_{n+1} - M_n = \frac{2}{n(n+1)\sigma_W^2} \sum_{\ell=1}^{n} \ell a_\ell^2 \ge 0$

- Equality iff $a_\ell = 0$ for all $\ell \ge 1$ $\Leftrightarrow$ $(X_1, \ldots, X_{n+1})$ is white