# Prediction-based Similarity Identification for Autoregressive Processes

**Hanwei Wu, Qiwen Wang and Markus Flierl**

*KTH Royal Institute of Technology, Stockholm*
*School of Electrical Engineering and Computer Science*

## Introduction

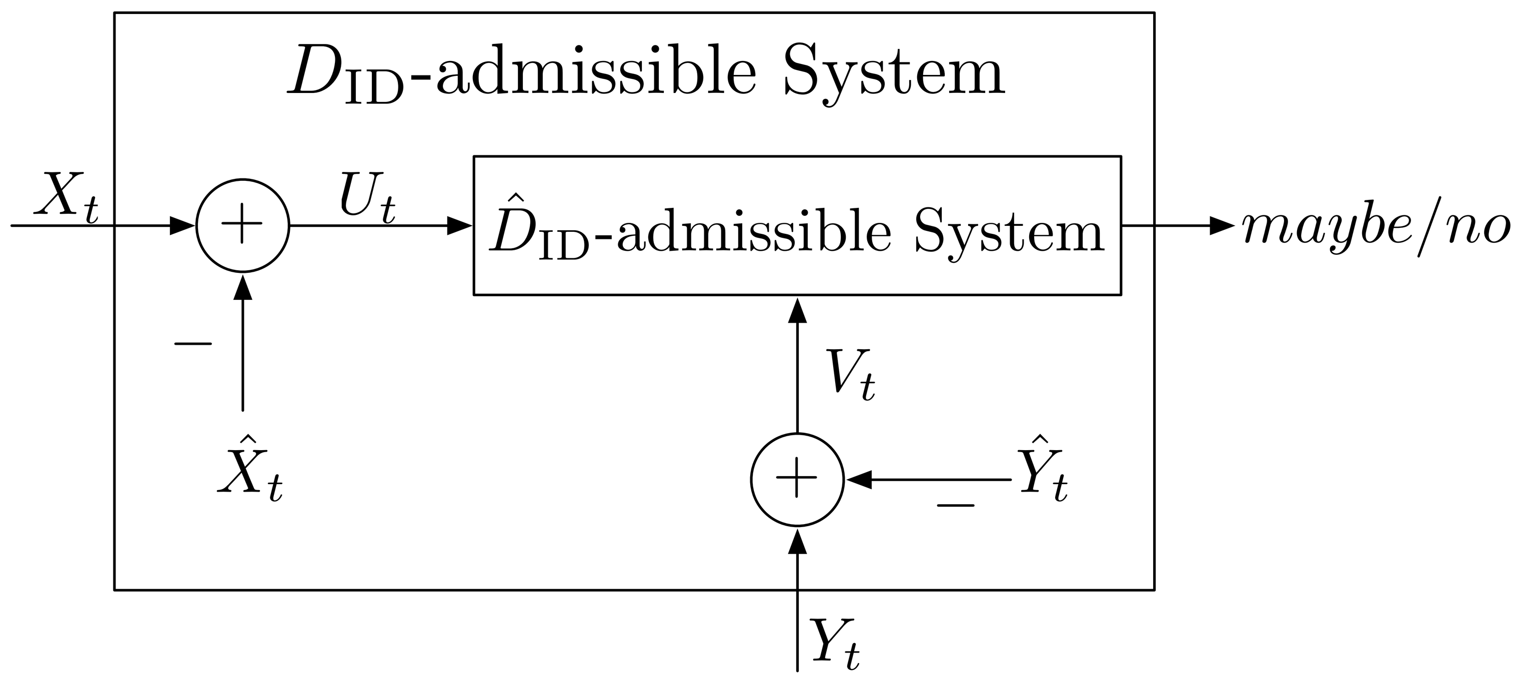
### Problem

- Compression of autoregressive processes for similarity identification [1].

### Goals

- Use prediction-based model to process autoregressive signals for similarity identification.
- Derive the identification rate of the autoregressive processes using prediction-based model.

## Prediction-based Approach




- Database and the query are zero-mean autoregressive processes: $X_t = U_t + \mathbf{a}_m^T \mathbf{X}_{t-1}^{(m)}$ and $Y_t = V_t + \mathbf{a}_m^T \mathbf{Y}_{t-1}^{(m)}$.
- Optimal predictor for autoregressive processes: $\hat{X}_t = \mathbf{a}_{N_p}^T \mathbf{X}_{t-1}^{(N_p)}$, where $\mathbf{a}_{N_p} = (a_1, \cdots, a_m, 0, \cdots, 0)^T$.
- Similarity identification is conducted in the embedded $\hat{D}_{\mathrm{ID}}$-admissible system.

## Identification rate $\mathbf{R}^{\mathbf{P}}_{\mathbf{ID}}$

- Vector representation of the autoregressive process:

$$\mathbf{x} = \mathbf{M}_t \mathbf{u} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_t \end{bmatrix} = \begin{bmatrix} 1 & & & & \\ m_1 & 1 & & & \\ & m_1 & 1 & & \\ \vdots & & & \ddots & \\ m_{t-1} & & \cdots & m_1 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_t \end{bmatrix}.$$

- Relation of the similarity measures : data space *v.s.* residual space

$$d(\mathbf{x}, \mathbf{y}) = \frac{1}{t} \sum_{i=1}^{t} \lambda_{t,i} d(\tilde{u}_i, \tilde{v}_i),$$

where $\tilde{u} = \mathbf{Q}\mathbf{u}$, $\tilde{v} = \mathbf{Q}\mathbf{v}$, $\mathbf{Q}$ and $\lambda_{t,i}$ are the eigenmatrix and eigenvalues of $\mathbf{P}_t = \mathbf{M}^T \mathbf{M}_t$.

- Assume for each time step an ideal identification system for Gaussian data, i.e., $d(\tilde{u}_i, \tilde{v}_i) \le \tilde{D}_{\mathrm{ID}}^{(i)}$ with identification rate $\tilde{R}_{\mathrm{ID}}^{(i)}$. The identification rate based on the prediction-model is obtained by the following constrained optimization:

$$\max_{\tilde{D}_{\mathrm{ID}}^{(1)}, \ldots, \tilde{D}_{\mathrm{ID}}^{(t)}} D_{\mathrm{ID}} = \frac{1}{t} \sum_{i=1}^{t} \lambda_{t,i} \tilde{D}_{\mathrm{ID}}^{(i)}$$

$$\text{s.t.} \quad \frac{1}{t} \sum_{i=1}^{t} \tilde{R}_{\mathrm{ID}}^{(i)} \le R_{\mathrm{ID}},$$

$$\text{s.t.} \quad \tilde{R}_{\mathrm{ID}}^{(i)} \ge 0.$$

- Identification rate of the prediction-based model for autoregressive Gaussian processes is given by

$$R_{\mathrm{ID}}^{P} = \frac{1}{t} \sum_{i=1}^{t} \max \left\{ \log_2 \left( \frac{2 \ln(2) \lambda_{t,i}}{v} \right), 0 \right\}$$

$$D_{\mathrm{ID}}^{P} = \frac{1}{t} \sum_{i=1}^{t} \lambda_{t,i} 2 \left( 1 - 2^{-R_{\mathrm{ID}}^{(i)}} \right).$$

## Special Case

### $R_{\mathrm{ID}}^{\mathrm{PS}}$ for Autoregressive Processes

- Only the smallest eigenvalue of $\mathbf{P}_t = \mathbf{M}_t^T \mathbf{M}_t$ is known.
- Relation of the similarity measures : data space *v.s.* residual space

$$d(\mathbf{x}, \mathbf{y}) \ge \lambda_{\min} d(\mathbf{u}, \mathbf{v})$$

- Similarity threshold for embedded similarity identification system

$$d(\mathbf{u}, \mathbf{v}) \le \frac{d(\mathbf{x}, \mathbf{y})}{\lambda_{\min}} \le \frac{D_{\mathrm{ID}}}{\lambda_{\min}} := \hat{D}_{\mathrm{ID}}.$$

- Identification rate of Gaussian autoregressive processes for the special case

$$R_{\mathrm{ID}}^{\mathrm{PS}} = \log_2 \left( \frac{2\lambda_{\min}}{2\lambda_{\min} - D_{\mathrm{ID}}} \right).$$

### Asymptotic Upper Bound of $R_{\mathrm{ID}}^{\mathrm{PS}}$

- $\mathbf{P}_t$ is asymptotically equivalent to a Toeplitz matrix $T_t(g)$, where $g(\omega)$ is

$$g(\omega) = \sum_{k=-\infty}^{\infty} e^{-\mathbf{j}k\omega} \left\{ \sum_{i=0}^{\infty} m_i m_{i+k} \right\} = \left| \sum_{k=0}^{\infty} m_k e^{-\mathbf{j}k\omega} \right|^2.$$

- The minimum eigenvalue of a Toeplitz matrix converges to the essential infimum $N_l$ of $g(\omega)$

$$\lim_{t \to \infty} \min_i \tau_{t,i} = N_l.$$

## Simulation Results

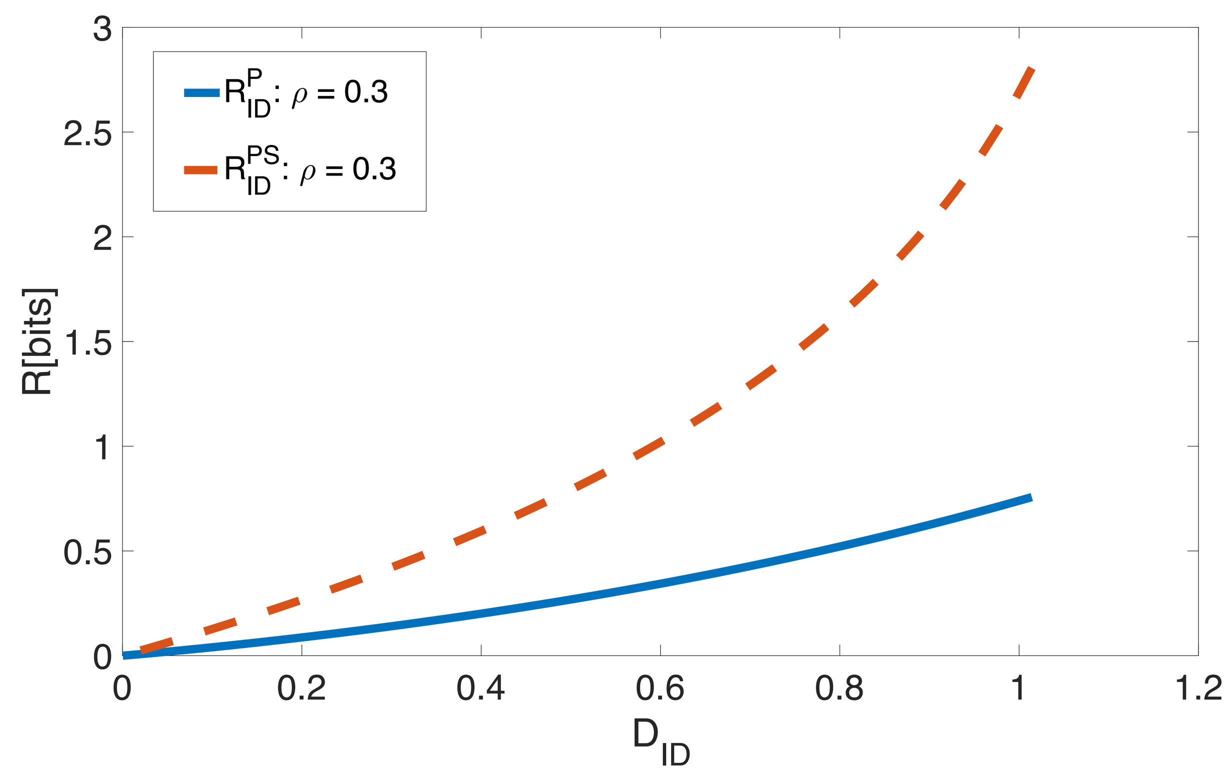



**Figure 1:** $R_{\mathrm{ID}}^{P}$ and $R_{\mathrm{ID}}^{\mathrm{PS}}$ for AR(1) sequences with $\rho = 0.3$, and varaince $\sigma_X^2 = \frac{1}{1-\rho^2}$

## Conclusions

- Propose a prediction-based model for computing the identification rate of Gaussian autoregressive processes.
- The identification rate depends on a sequence of eigenvalues that we derive from our prediction model.
- The identification rate for the special case depends only on minimum eigenvalue of the Toeplitz matrix.

## References

[1] A. Ingber, T. Courtade, and T. Weissman, ”Compression for quadratic similarity queries”, *IEEE Trans. on Information Theory*, vol. 61, no. 5, pp. 2729-2747, May 2015.