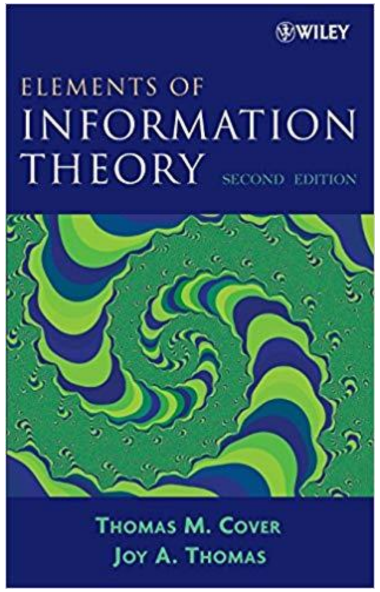
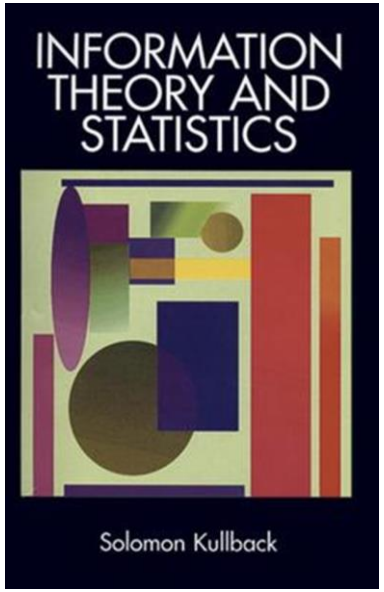
데이터 학습과 정보이론

Junghyo Jo

Department of Statistics
Keimyung University

Information theory everywhere





[Physics] physical entity of entropy

[Statistics] hypothesis testing

[Communication theory] minimal coding length, channel capacity, rate distortion

[Machine learning] distance measures between data and model

[Bioinformatics] information contents in DNA sequences

"it from bit" –Wheeler
"human body, mediator of genetic information" -Dawkins

How much "information" in data?

Prior probability

$$P_1 = P_2 = P_3 = P_4 = P_5 = P_6 = \frac{1}{6}$$

Data

$$\{1, 2, 1, 3, 4, 1, 5, 3, 4, 6\}$$

$$n_1 = 3,\ n_2 = 1,\ n_3 = 2,\ n_4 = 2,\ n_5 = 1,\ n_6 = 1$$

$$n_1 + n_2 + n_3 + n_4 + n_5 + n_6 = 10$$

$$P_1 = 1,\ P_2 = P_3 = P_4 = P_5 = P_6 = 0$$

$$\{1, 1, 1, 1, 1, 1, \cdots, 1\}$$

$$I(Q||P) = 0$$

$$\frac{n!}{n_1!\,n_2!\cdots n_6!} P_1^{n_1} P_2^{n_2} \cdots P_6^{n_6} \quad \left\{ \begin{array}{l} \{1, 3, 2, 1, 4, 1, 5, 4, 3, 6\} \\ \{6, 5, 2, 1, 3, 1, 3, 4, 4, 1\} \\ \vdots \end{array} \right.$$

Log-likelihood

$$\log \frac{n!}{n_1!\,n_2!\cdots n_6!} P_1^{n_1} P_2^{n_2} \cdots P_6^{n_6} = -\sum_X Q_X \log \frac{Q_X}{P_X} = -I(Q||P)$$

Information

$$Q_X = \frac{n_X}{n} \qquad \log n! \approx n \log n - n$$

How to extract features from data?

Given a prior probability P_X and a constraint $\sum_X T(X) Q_X = \theta$, obtain Q_X^* closest to P_X

$$I(Q||P) = \sum_X Q_X \log \frac{Q_X}{P_X} \qquad \mathcal{L}[Q_X] = I(Q||P) + \alpha\left(\sum_X Q_X - 1\right) + \beta\left(\sum_X T(X) Q_X - \theta\right)$$

$$I(Q||P) \geq I(Q^*||P)$$

$$Q_X^* = \frac{P_X e^{-\beta T(X)}}{Z}, \qquad Z = \sum_X P_X e^{-\beta T(X)}$$

$$I(Q^*||P) = -\beta\theta - \log Z$$

Statistical mechanics

$$T(X) = E(X), \qquad \theta = \bar{E}, \qquad \beta = \frac{1}{T}$$

$P_X = 1$ (prior distribution)

$$Q^*(E) = \frac{e^{-\beta E}}{Z} \qquad \mathcal{F} = \bar{E} - TS$$

One can extract features in data using Q_X^* or $I(Q^*||P)$

Recipe of extracting features from data

① Assign maximally-likely prior distribution P_X of data X

② Design appropriate transformation (estimator) $T(X)$ of data to extract features in your interest

③ Obtain a special data distribution Q_X^* minimally-distant from P_X constrained by the expectation of $T(X)$

④ Extract features from the optimized distribution Q_X^* and the minimum discrimination information $I(Q^*||P)$

Dynamics and time series

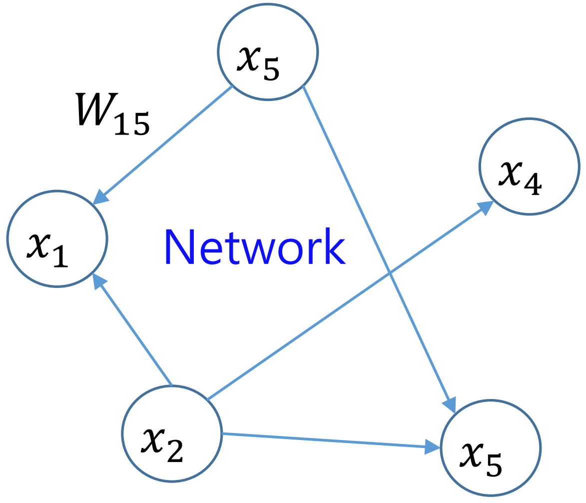
$$\boldsymbol{x} = (x_1, x_2, x_3, \cdots, x_n)$$

$$H_i(t) = \sum_{j=1}^{n} W_{ij} x_j(t)$$

- Brain activities
- Gene/protein expressions
- Biochemical reactions
- Population dynamics
- Currency exchange rates
- Stock trading prices
- Human activities
- …

time

$$\begin{array}{l} x_1(1), x_2(1), x_3(1), \cdots, x_n(1) \\ x_1(2), x_2(2), x_3(2), \cdots, x_n(2) \\ x_1(3), x_2(3), x_3(3), \cdots, x_n(3) \\ \vdots \quad \vdots \quad \vdots \quad \cdots \quad \vdots \\ x_1(L), x_2(L), x_3(L), \cdots, x_n(L) \end{array}$$



Deterministic dynamics

$$x_i(t+1) = x_i(t) + H_i(t)$$

$$\dot{x}_i(t) = \frac{x_i(t+dt) - x_i(t)}{dt} = \sum_{j=1}^{n} W_{ij} x_j(t)$$

Stochastic dynamics

$$P[x_i(t+1) = \pm 1 | \boldsymbol{x}(t)] = \frac{\exp(\pm H_i(t))}{\exp(H_i(t)) + \exp(-H_i(t))}$$

Minimum discrimination information

$$x_1(1), x_2(1),\ x_3(1), \cdots, x_n(1)$$
$$x_1(2), x_2(2),\ x_3(2), \cdots, x_n(2)$$
$$x_1(3), x_2(3),\ x_3(3), \cdots, x_n(3)$$
$$\vdots \quad \vdots \quad \vdots \quad \cdots \quad \vdots$$
$$x_1(L), x_2(L),\ x_3(L), \cdots, x_n(L)$$

$$P_0(\boldsymbol{x}) = \frac{1}{L}\sum_{t=1}^{L} \delta(\boldsymbol{x} - \boldsymbol{x}(t))$$

$$m_j^0 = \sum_{\boldsymbol{x}} x_j P_0(\boldsymbol{x}) = \frac{1}{L}\sum_{t=1}^{L} x_j(t)$$

$$h_i^0 = \sum_{\boldsymbol{x}} H_i(\boldsymbol{x}) P_0(\boldsymbol{x}) = \frac{1}{L}\sum_{t=1}^{L} H_i(\boldsymbol{x}(t))$$

Generation probability

$$\boldsymbol{x} = (x_1, x_2,\ x_3, \cdots, x_n)$$

$$P(\boldsymbol{x})?$$

$$m_j = \sum_{\boldsymbol{x}} x_j P(\boldsymbol{x})$$

$$h_i = \sum_{\boldsymbol{x}} H_i(\boldsymbol{x}) P(\boldsymbol{x})$$

$$T(x) = (\boldsymbol{x}, \boldsymbol{H}(\boldsymbol{x}))$$

$$\theta = (\boldsymbol{m}, \boldsymbol{h})$$

$$\beta = (\boldsymbol{J}, \boldsymbol{\beta})$$

$$I(P\!:\!P_0) = \sum_{\boldsymbol{x}} P(\boldsymbol{x}) \log \frac{P(\boldsymbol{x})}{P_0(\boldsymbol{x})}$$

$$I(P\!:\!P_0) \geq I(P^*\!:\!P_0)$$

$$P^*(\boldsymbol{x}) = \frac{P_0(\boldsymbol{x}) \exp(\boldsymbol{J}\cdot\boldsymbol{x} + \boldsymbol{\beta}\cdot\boldsymbol{H})}{Z}$$

$$Z(\boldsymbol{J}, \boldsymbol{\beta}) = \sum_{\boldsymbol{x}} P_0(\boldsymbol{x}) \exp(\boldsymbol{J}\cdot\boldsymbol{x} + \boldsymbol{\beta}\cdot\boldsymbol{H})$$

$$I(P^*\!:\!P_0) = G(\boldsymbol{m}, \boldsymbol{h}) = \boldsymbol{J}\cdot\boldsymbol{m} + \boldsymbol{\beta}\cdot\boldsymbol{h} - \log Z(\boldsymbol{J}, \boldsymbol{\beta})$$

$$\frac{\partial \log Z}{\partial \boldsymbol{J}} = \boldsymbol{m} \qquad \frac{\partial \log Z}{\partial \boldsymbol{\beta}} = \boldsymbol{h}$$

Minimum discrimination information

$$I(P:P_0) = \sum_{\boldsymbol{x}} P(\boldsymbol{x}) \log \frac{P(\boldsymbol{x})}{P_0(\boldsymbol{x})}$$

$$I(P:P_0) \geq I(P^*:P_0)$$

$$P^*(\boldsymbol{x}) = \frac{P_0(\boldsymbol{x}) \exp(\boldsymbol{J}\cdot\boldsymbol{x} + \boldsymbol{\beta}\cdot\boldsymbol{H})}{Z}$$

$$Z(\boldsymbol{J},\boldsymbol{\beta}) = \sum_{\boldsymbol{x}} P_0(\boldsymbol{x}) \exp(\boldsymbol{J}\cdot\boldsymbol{x} + \boldsymbol{\beta}\cdot\boldsymbol{H})$$

$$I(P^*:P_0) = G(\boldsymbol{m},\boldsymbol{h}) = \boldsymbol{J}\cdot\boldsymbol{m} + \boldsymbol{\beta}\cdot\boldsymbol{h} - \log Z(\boldsymbol{J},\boldsymbol{\beta})$$

$$\frac{\partial \log Z}{\partial \boldsymbol{J}} = \boldsymbol{m} \qquad \frac{\partial \log Z}{\partial \boldsymbol{\beta}} = \boldsymbol{h}$$

$\underline{\boldsymbol{J} = \boldsymbol{\beta} = \boldsymbol{0}}$

$$P^*(\boldsymbol{x}) = P_0(\boldsymbol{x})$$

$$I(P^*:P_0) = G(\boldsymbol{m}^0,\boldsymbol{h}^0) = 0$$

$$\frac{\partial G}{\partial \boldsymbol{m}} = \boldsymbol{J} = \boldsymbol{0} \qquad \frac{\partial G}{\partial \boldsymbol{h}} = \boldsymbol{\beta} = \boldsymbol{0}$$

$$G(\boldsymbol{m},\boldsymbol{h}) \approx \frac{1}{2}\sum_{j,k} \frac{\partial^2 G}{\partial m_j m_k}(m_j - m_j^0)(m_k - m_k^0) + \frac{1}{2}\sum_{j,k} \frac{\partial^2 G}{\partial m_j h_k}(m_j - m_j^0)(h_k - h_k^0) + \frac{1}{2}\sum_{j,k} \frac{\partial^2 G}{\partial h_j h_k}(h_j - h_j^0)(h_k - h_k^0)$$

Minimum discrimination information

$$H_i(\boldsymbol{x}) = \sum_{j=1}^{n} W_{ij} x_j$$

$$W_{ij} = \frac{\partial h_i}{\partial m_j} = \sum_k \left[\frac{\partial^2 G}{\partial h_i \partial m_k}\right]^{-1} \frac{\partial^2 G}{\partial m_k m_j}$$

$$E[H_i(\boldsymbol{x})] = \sum_{j=1}^{n} W_{ij} E[x_j]$$

$$= \sum_k COV(H_i, x_k)\, COV^{-1}(x_k, x_j)$$

Linear regression!

$$h_i = \sum_{j=1}^{n} W_{ij} m_j$$

$$H_i(\boldsymbol{x}) = \sum_{j=1}^{n} W_{ij} x_j + \frac{1}{2} \sum_{j,k=1}^{n} Q_{ijk} x_j x_k$$

$$G(\boldsymbol{m}, \boldsymbol{h}) \approx \frac{1}{2} \sum_{j,k} \frac{\partial^2 G}{\partial m_j m_k} (m_j - m_j^0)(m_k - m_k^0)$$
$$+ \frac{1}{2} \sum_{j,k} \frac{\partial^2 G}{\partial m_j h_k} (m_j - m_j^0)(h_k - h_k^0)$$
$$+ \frac{1}{2} \sum_{j,k} \frac{\partial^2 G}{\partial h_j h_k} (h_j - h_j^0)(h_k - h_k^0)$$
$$+ \cdots$$

Iterative inference algorithm

$$H_i(t) = \sum_j W_{ij} x_j(t)$$

observation

$$H_i'(t) = \frac{x_i(t+1)}{E[x_i(t+1)]} H_i(t)$$

estimation

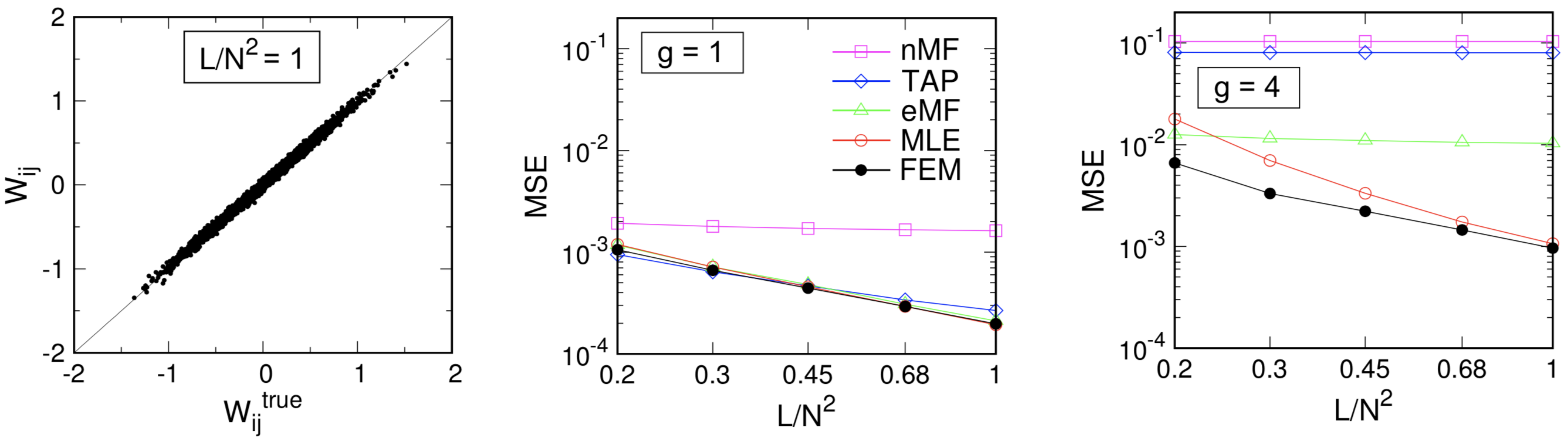
$$W_{ij} = \sum_k COV(H_i', x_k) COV^{-1}(x_k, x_j)$$

$$D_i = \sum_t \| x_i(t+1) - E[x_i(t+1)] \|^2$$

Benchmark: Kinetic Ising model (N=100, gaussian W)

$W_{ij} \sim \mathcal{N}(0, \frac{g^2}{N})$

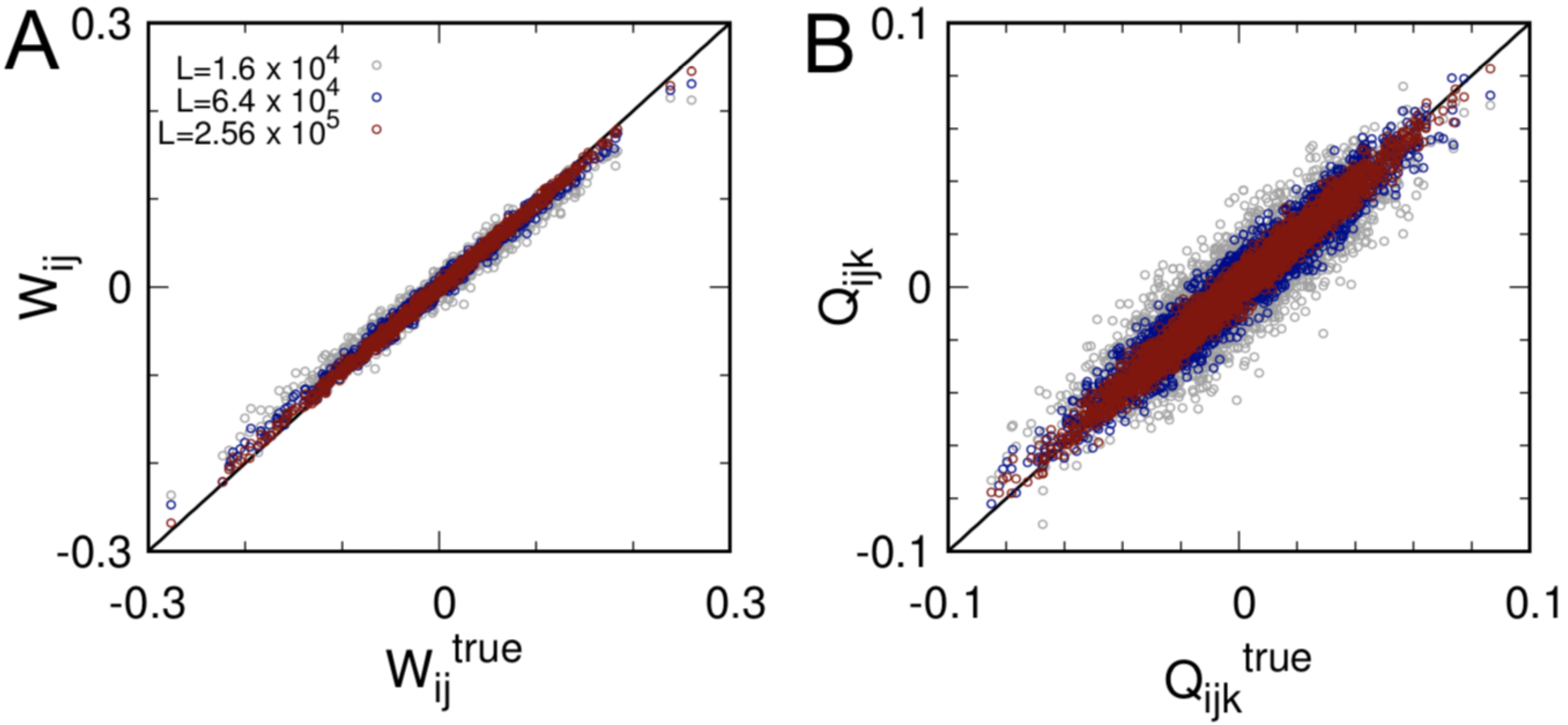
L/N^2 samples



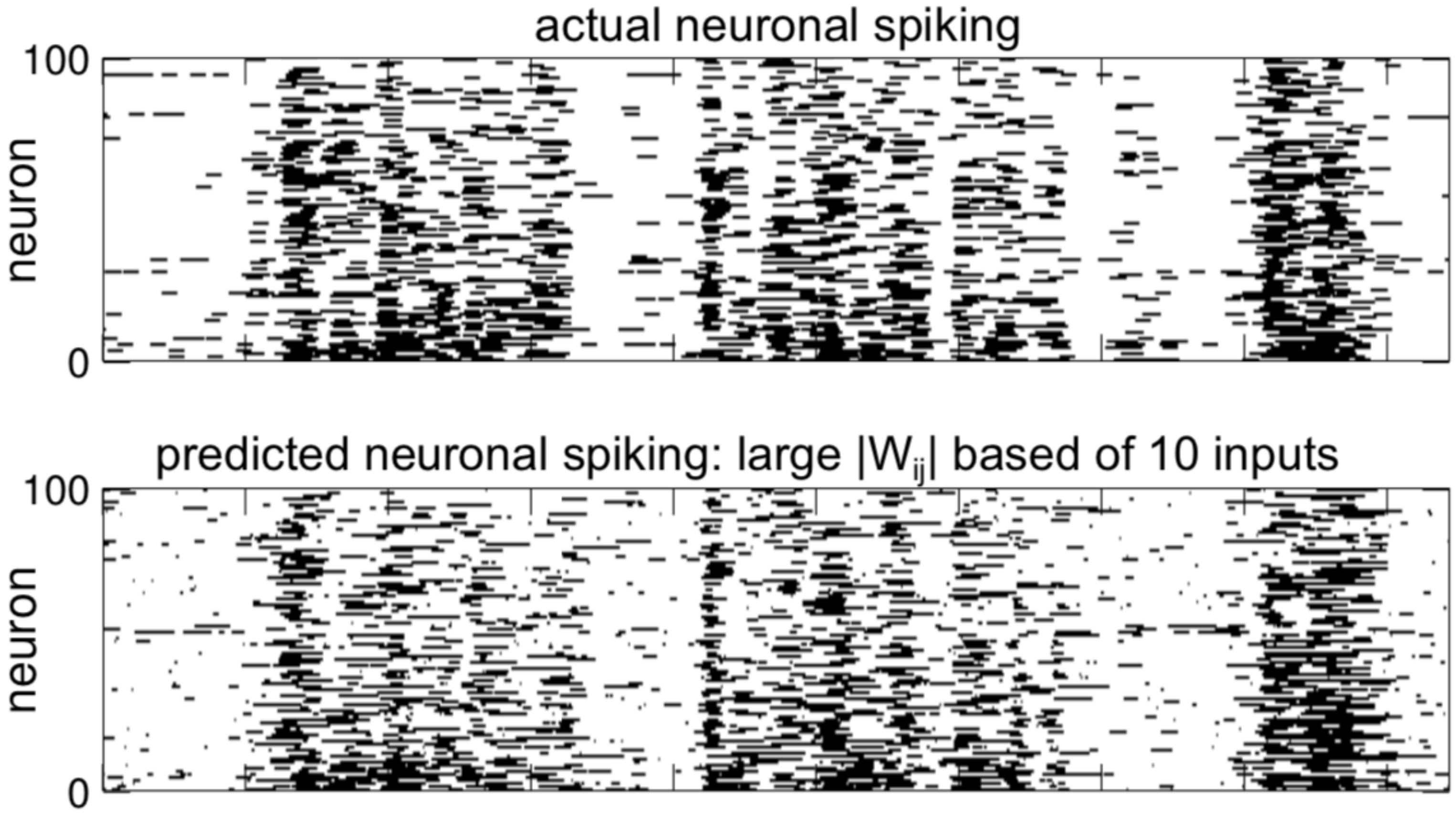
FEM outperforms for strong coupling and little data regimes

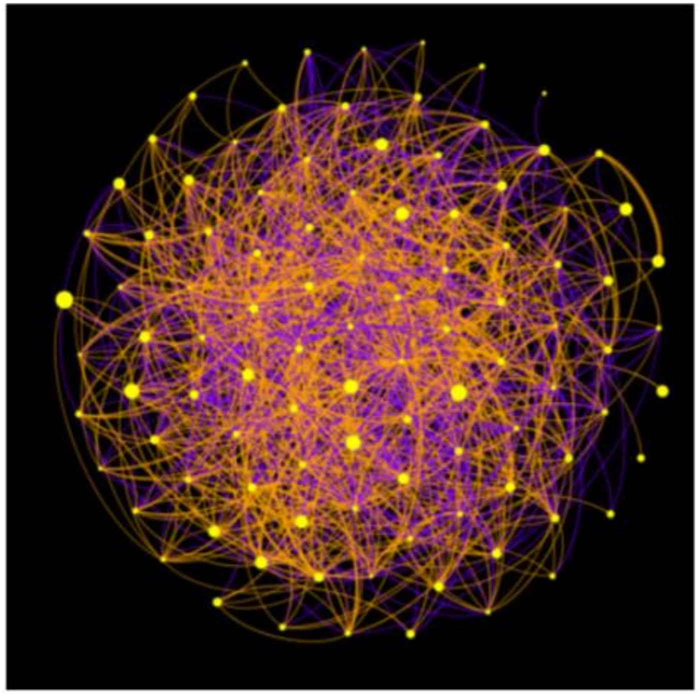
Quadratic interaction inference

$$H_i(t) = \sum_j W_{ij}\, x_j(t) + \frac{1}{2}\sum_{j,k} Q_{ijk} x_j(t) x_k(t)$$



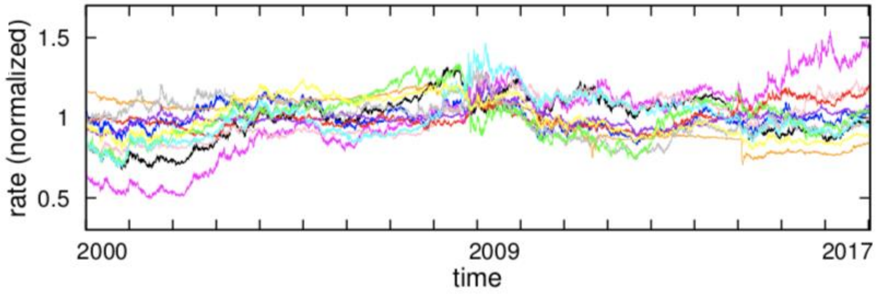
Neural network

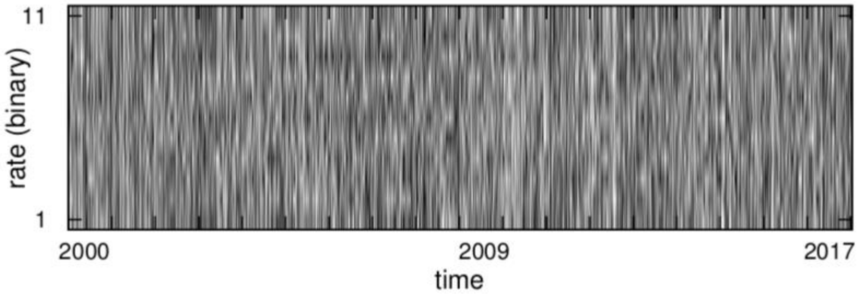


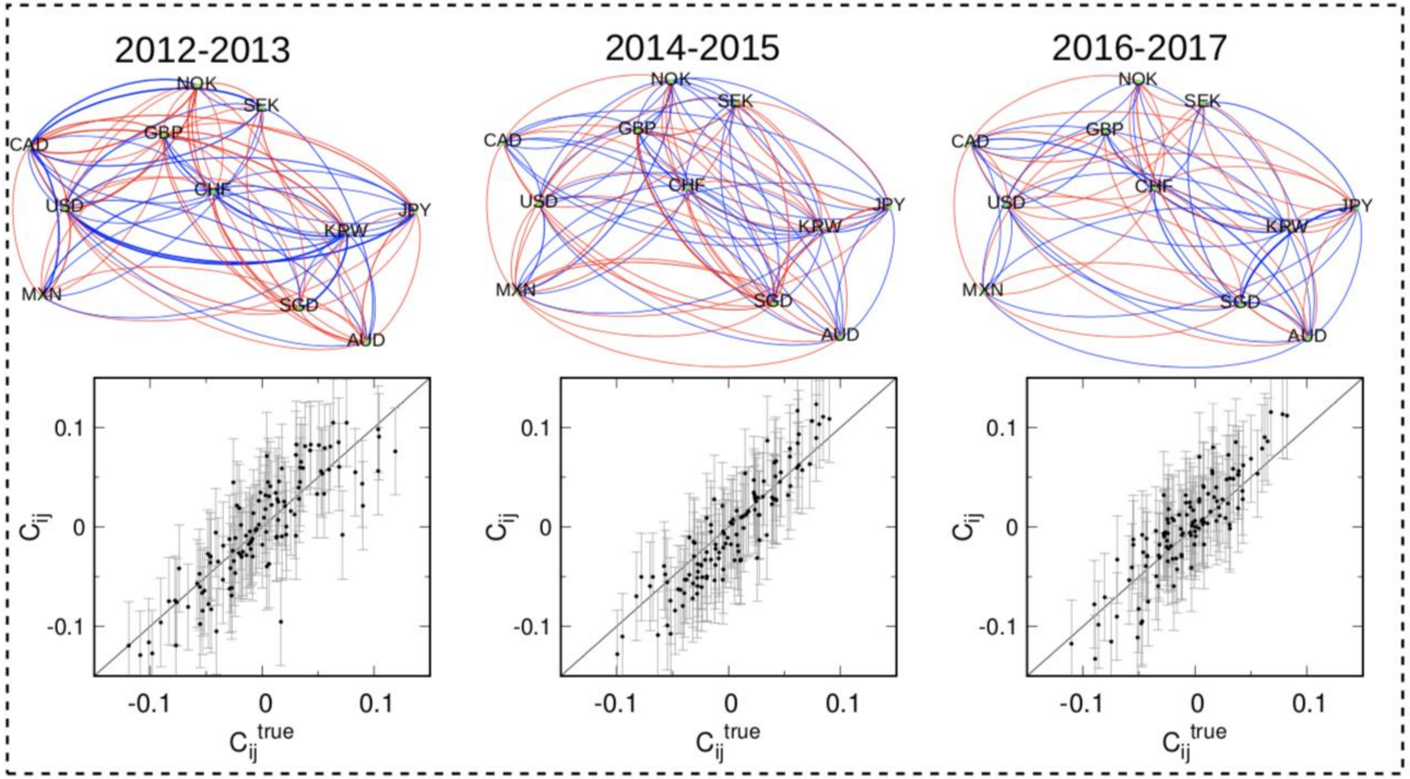


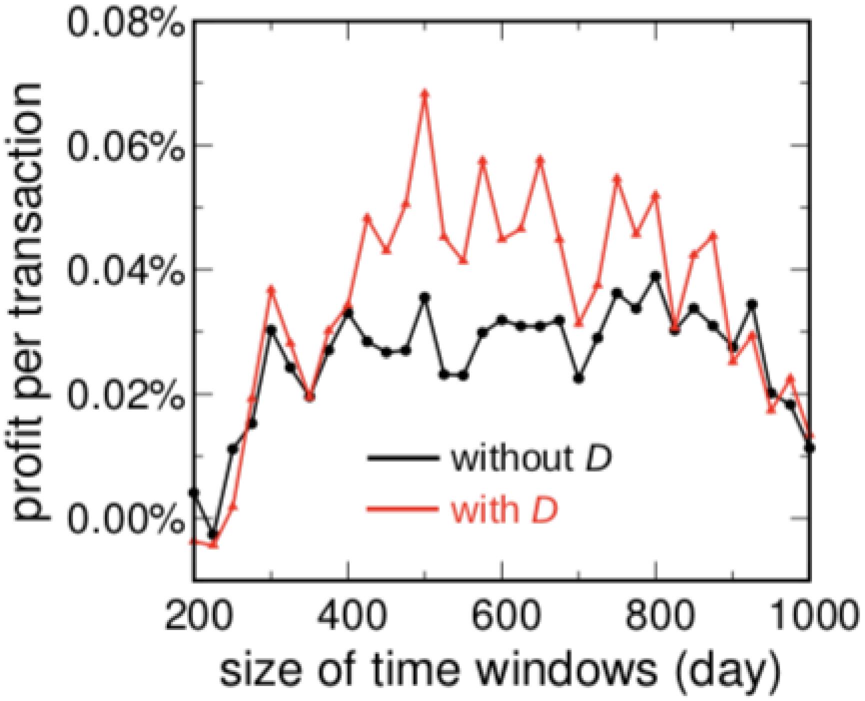
Prediction

Currency network









Time-dependent coupling W_{ij}

Hidden variables

$\boldsymbol{x} = (x_1, x_2, x_3, \cdots, x_n)$

$$\begin{matrix} x_1(1), x_2(1), x_3(1), \cdots, x_n(1) \\ x_1(2), x_2(2), x_3(2), \cdots, x_n(2) \\ x_1(3), x_2(3), x_3(3), \cdots, x_n(3) \\ \vdots \quad \vdots \quad \vdots \quad \cdots \quad \vdots \\ x_1(L), x_2(L), x_3(L), \cdots, x_n(L) \end{matrix}$$

$\boldsymbol{y} = (y_1, y_2, \cdots, y_h)$

$$\begin{matrix} y_1(1), y_2(1), \cdots, y_h(1) \\ y_1(2), y_2(2), \cdots, y_h(2) \\ y_1(3), y_2(3), \cdots, y_h(3) \\ \vdots \quad \vdots \quad \cdots \quad \vdots \\ y_1(L), y_2(L), \cdots, y_n(L) \end{matrix}$$

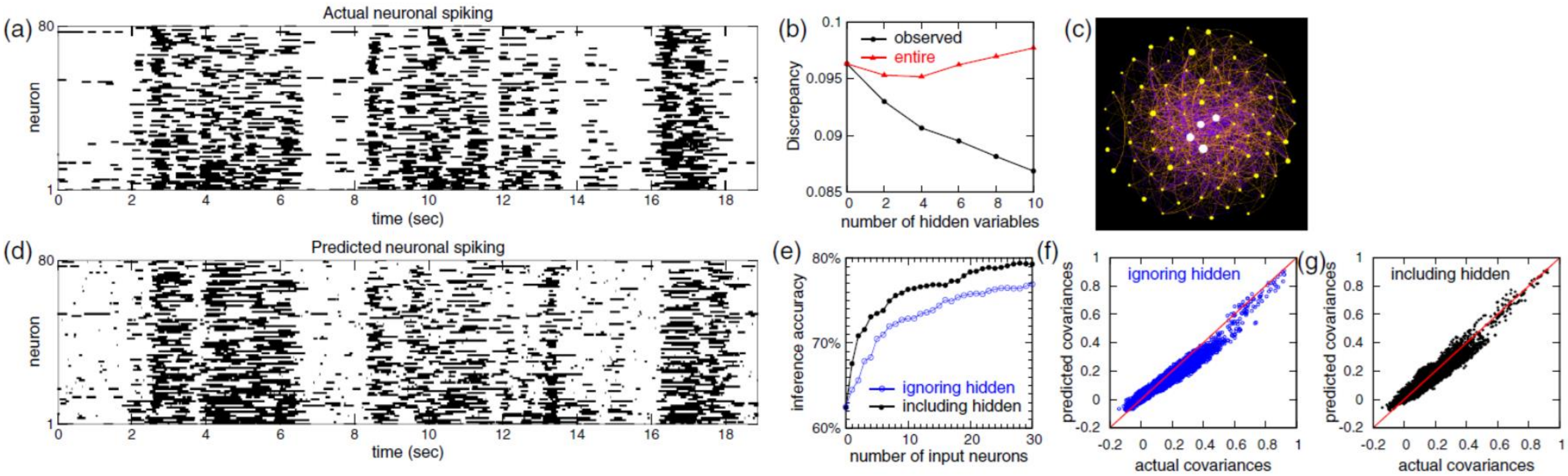
$\boldsymbol{z} = (x_1, x_2, x_3, \cdots, x_n, y_1, y_2, \cdots, y_h)$

$$H_i(t) = \sum_{j=1}^{n+h} W_{ij} z_j(t)$$

Expectation Maximization

M-step $\boldsymbol{z} = (x_1, x_2, x_3, \cdots, x_n, y_1, y_2, \cdots, y_h) \rightarrow W_{ij}$

E-step $W_{ij} \rightarrow \boldsymbol{y} = (y_1, y_2, \cdots, y_h)$



References

PHYSICAL REVIEW E **99**, 023311 (2019)

Network inference in stochastic systems from neurons to currencies: Improved performance at small sample size

Danh-Tai Hoang,[1,2] Juyong Song,[3,4,5] Vipul Periwal,[1,*] and Junghyo Jo[6,7,†]

PHYSICAL REVIEW E **99**, 042114 (2019)

Editors' Suggestion

Data-driven inference of hidden nodes in networks

Danh-Tai Hoang,[1,2] Junghyo Jo,[3,4,*] and Vipul Periwal[1,†]

Applications

c. elegans brain imaging https://www.youtube.com/watch?v=llHrk7RR4GE

Zebrafish brain imaging https://www.youtube.com/watch?v=YLVdRPVj-XM

https://www.youtube.com/watch?v=eKkaYDTOauQ

Acknowledgement

- Vipul Periwal (NIH)
- Danh-Tai Hoang (NIH)
- Juyong Song (Samsung Research)
- Matteo Marsili (ICTP)