Lecture 3

3.1 Maximum likelihood estimation
3.2 Kinetic Ising models
3.3 Gradient free parameter optimization
3.4 Application to neuron firing
3.5 Application to hidden variables
3.6 Application to missing data

All work done with Junghyo Jo and others

We are going to study stochastic systems where the present state determines the probable next state with specific probabilities. Suppose there's a single spin $\sigma(t) = \pm 1$, flipping with

$$P(\sigma(t+1) = \rho | \sigma(t)) = \frac{\exp(\rho(b + w\sigma(t)))}{2\cosh(b + w\sigma(t))}.$$

How do we find b, w?
We could look at the mean value and covariance over a period of observation:

$$m \equiv \langle \sigma(t) \rangle, \quad \langle \sigma(t+1)\sigma(t) \rangle.$$

So, for example, we might expect

$$m = \tanh(b + wm)$$

for some kind of mean-field self-consistency.
We can be a lot more sophisticated. The expectation value of $\sigma(t+1)$ is $\tanh(b + w\sigma(t))$. So we can demand

$$\langle \sigma(t+1)\sigma(t) \rangle = \langle \tanh(b + w\sigma(t))\sigma(t) \rangle.$$

Maximum likelihood estimation

We want to maximize the likelihood of the observed time series:

$$ML = \prod_{t=1}^{L-1} P(\sigma(t+1)|\sigma(t)),$$

Then differentiating with respect to b, w gives

$$\partial_b \ln ML = \sum_t [\sigma(t+1) - \tanh(b + w\sigma(t))],$$

$$\partial_w \ln ML = \sum_t [\sigma(t+1)\sigma(t) - \sigma(t)\tanh(b + w\sigma(t))].$$

So our intuitive equations are the same as ML! Then

$$(\partial_w - m\partial_b) \ln ML = \sum_t [\sigma(t+1)\delta\sigma(t) - \tanh(b + w\sigma(t))\delta\sigma(t)],$$

where we have defined $\delta\sigma(t) \equiv \sigma(t) - \langle\sigma(t)\rangle$. $\tanh(x + \delta) \approx (1 - \tanh^2(x))\delta$, so

$$(\partial_w - m\partial_b) \ln ML \approx \langle\sigma(t+1)\delta\sigma(t)\rangle - w(1 - m^2)\langle\delta\sigma(t)\delta\sigma(t)\rangle.$$

So we finally arrive at the naïve mean field approximation:

$$w \approx \frac{1}{(1-m^2)} \frac{\langle \sigma(t+1)\delta\sigma(t)\rangle}{\langle \delta\sigma(t)\delta\sigma(t)\rangle} \ .$$

If you consider expanding tanh to second order, you get a better approximation: the Thouless-Anderson-Palmer approximation. If you treat fluctuations in $\sigma(t+1)$ as following a Gaussian distribution, then you get the so-called exact mean-field approximation.
We will do something different but I wanted to give you an idea of why all these expressions involve a ratio of covariances.

Now we do it with many indices!

Kinetic Ising model: N-spin state $\sigma = (\sigma_1, \cdots, \sigma_N)$ at time $t+1$ is stochastically determined from the current state $\sigma(t)$ at time t with the following conditional probability,

$$P(\sigma_i(t+1)|\sigma(t)) = \frac{\exp(\sigma_i(t+1)H_i(\sigma(t)))}{\exp(H_i(\sigma(t))) + \exp(-H_i(\sigma(t)))}, \tag{1}$$

for $i = 1, \cdots, N$. Local field:

$$H_i(\sigma(t))$$

is how the present state $\sigma(t)$ stochastically determines $\sigma_i(t+1)$.

$$H_i(\sigma(t)) = \sum_j W_{ij}\sigma_j(t),$$

and we aim to determine the weight matrix

$$W_{ij}.$$

Danh-Tai Hoang, Juyong Song, Vipul Periwal, and Junghyo Jo, PRE

Model expectation:

$$\langle \sigma_i(t+1) \rangle_{H_i(\sigma(t))} \equiv \sum_{\rho=\pm 1} \rho \; P(\sigma_i(t+1) = \rho | \sigma(t)) = \tanh(H_i(\sigma(t))).$$

Notice

$$\left| \frac{\langle \sigma_i(t+1) \rangle_{H_i(\sigma(t))}}{\sigma_i(t+1)} \right| = |\tanh(H_i(\sigma(t)))| \leq 1.$$

Define

$$H_i^{\text{new}}(\sigma(t)) \leftarrow \frac{\sigma_i(t+1)}{\langle \sigma_i(t+1) \rangle_{H_i(\sigma(t))}} H_i(\sigma(t)),$$

then

$$|\langle \sigma_i(t+1) \rangle_{H_i^{\text{new}}(\sigma(t))}| \geq |\langle \sigma_i(t+1) \rangle_{H_i(\sigma(t))}|$$

because $|H_i^{\text{new}}(\sigma(t))| \geq |H_i(\sigma(t))|$ and, therefore, $|\tanh(H_i^{\text{new}}(\sigma(t)))| \geq |\tanh(H_i(\sigma(t)))|$. So the model prediction for $\sigma_i(t+1)$ is closer to ± 1, and is therefore better IF it's actually correct.

Where did this update come from?

Start with a moment generating function:

$$Z_i(J,\beta) = \sum_t \exp(J\cdot\sigma(t) - \beta H_i^{\text{new}}(\sigma(t))), \tag{1}$$

which is a function of a vector parameter J, a scalar parameter β, and an observable $H_i^{\text{new}}(\sigma(t))$ of data $\sigma(t)$. Define a convex free energy $F_i \equiv \log Z_i$ so we can get expectation values of spin activities and observables by differentiation,

$$\begin{aligned}
\frac{\partial F_i}{\partial J_j} &= \frac{\sum_t \sigma_j(t)\exp(J\cdot\sigma(t) - \beta H_i^{\text{new}}(\sigma(t)))}{\sum_t \exp(J\cdot\sigma(t) - \beta H_i^{\text{new}}(\sigma(t)))} \\
&= \langle\sigma_j\rangle_J \equiv m_j(J), \qquad (2)\\
\frac{\partial F_i}{\partial \beta} &= -\frac{\sum_t H_i^{\text{new}}(\sigma(t))\exp(J\cdot\sigma(t) - \beta H_i^{\text{new}}(\sigma(t)))}{\sum_t \exp(J\cdot\sigma(t) - \beta H_i^{\text{new}}(\sigma(t)))} \\
&= -\langle H_i^{\text{new}}\rangle_J. \qquad (3)
\end{aligned}$$

Notice we have a sum over observed configurations, but we NEVER ever assume that there is some sort of equilibrium distribution or even try to deduce some sort of distribution for the observed configurations!!

Define a convex dual free energy G_i so that the expected activity vector m is the independent variable, and $J(m)$ the dependent vector:

$$F_i(J) + G_i(m) = J \cdot m.$$

Define a normalized probability, $P(\sigma(t)) \equiv \exp(J \cdot \sigma(t) - \beta H_i^{\text{new}}(\sigma(t)) - F_i)$ then G_i is a thermodynamic free energy,

$$G_i = \beta \langle H_i^{\text{new}} \rangle_J - S_i \tag{1}$$

with the expectation value of H_i^{new} taking the place of internal energy and the Shannon entropy of data,

$$S_i = -\sum_t P(\sigma(t)) \log P(\sigma(t)).$$

At $\beta = 0$, minimizing the free energy G_i is exactly maximizing the entropy S_i.

The usual Legendre transform duality gives

$$\frac{\partial G_i}{\partial m_j} = J_j, \tag{1}$$

$$\frac{\partial G_i}{\partial \beta} = -\frac{\partial F_i}{\partial \beta} = \langle H_i^{\text{new}} \rangle_m, \tag{2}$$

where we identify $\langle H_i^{\text{new}} \rangle_{J(m)} \equiv \langle H_i^{\text{new}} \rangle_m$.
With G_i, we can easily get $\langle H_i^{\text{new}} \rangle_m$, the expectation value of observable H_i^{new} conditioned on the expectation value $m = \langle \sigma \rangle$ of microstates σ. All we need is the derivatives of $G_i(m)$ at its minimum for $\beta = 0$.

The Taylor expansion of $G_i(m)$ up to second-order terms at $m = m^*$ is

$$G_i(m) = G_i(m^*) + \frac{1}{2} \sum_{j,k} \left[\frac{\partial^2 G_i}{\partial m_j \partial m_k} \right]^* (m_j - m_j^*)(m_k - m_k^*), \tag{1}$$

where the derivatives $[\cdot]^*$ are taken at $m = m^*$.
Differentiating the expanded $G_i(m)$ with respect to β leads to

$$\frac{\partial G_i(m)}{\partial \beta} = \frac{\partial G_i(m^*)}{\partial \beta} - \sum_{j,k} \frac{\partial m_k^*}{\partial \beta} \left[\frac{\partial^2 G_i}{\partial m_j \partial m_k} \right]^* (m_j - m_j^*). \tag{2}$$

Big Picture

$$\frac{\partial}{\partial m}\frac{\partial}{\partial \beta}G = \frac{\partial}{\partial \beta}\frac{\partial}{\partial m}G.$$

We are going to calculate one side directly and the other side using the Taylor expansion!

The Taylor expansion of $G_i(m)$ up to second-order terms at $m = m^*$ is

$$G_i(m) = G_i(m^*) + \frac{1}{2}\sum_{j,k}\left[\frac{\partial^2 G_i}{\partial m_j \partial m_k}\right]^* (m_j - m_j^*)(m_k - m_k^*), \tag{1}$$

where the derivatives $[\cdot]^*$ are taken at $m = m^*$.
Differentiating the expanded $G_i(m)$ with respect to β leads to

$$\frac{\partial G_i(m)}{\partial \beta} = \frac{\partial G_i(m^*)}{\partial \beta} - \sum_{j,k}\frac{\partial m_k^*}{\partial \beta}\left[\frac{\partial^2 G_i}{\partial m_j \partial m_k}\right]^* (m_j - m_j^*). \tag{2}$$

Remember that differentiating the connected correlation/cumulant generating function gives us connected correlation functions. So if we define $\langle f\rangle^* \equiv \langle f\rangle_{J=0}$, and $\langle \delta f\rangle_m \equiv \langle f\rangle_m - \langle f\rangle^*$,

$$-\frac{\partial m_k}{\partial \beta} = \langle \delta H_i^{\text{new}} \delta\sigma_k\rangle$$

and

$$\frac{\partial^2 G_i}{\partial m_j \partial m_k} = \frac{\partial J_k}{\partial m_j} = \langle \delta\sigma_j \delta\sigma_k\rangle_m^{-1} \equiv C_{jk}^{-1}.$$

From $\partial_\beta G_i = \langle H_i \rangle_m$, we see that the quadratic expansion of $G_i(m, \beta)$ around m^* is just

$$\langle \delta H_i^{\text{new}} \rangle_m = \sum_{j,k} \langle \delta H_i^{\text{new}} \delta \sigma_k \rangle^* [C^{-1}]^*_{kj} \langle \delta \sigma_j \rangle_m .$$

Differentiating with respect to m, and matching terms, we get

$$W_{ij}^{\text{new}} \leftarrow \sum_k \langle \delta H_i^{\text{new}} \delta \sigma_k \rangle^* [C^{-1}]^*_{kj} .$$

This is the update that we intuitively suggested would work.
Note: We made no use of maximizing any quantity. All we did was use the minimization of the free energy.
We can do the Taylor expansion of the free energy to higher orders and get higher order terms in the local field H_i in the same way.

Not so fast!
When do we stop the iteration?

Define the discrepancy as

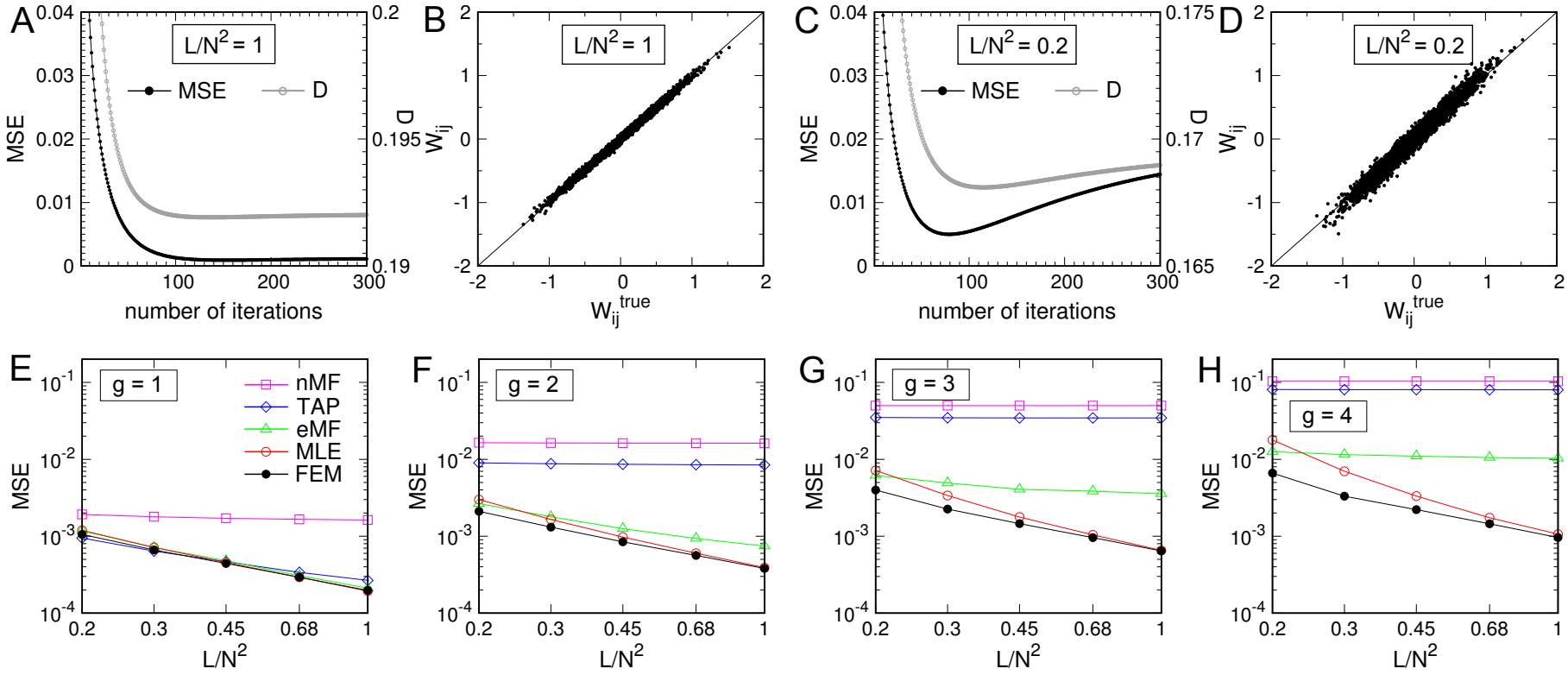
$$D_i \equiv \sum_t \left[\sigma_i(t+1) - \langle \sigma_i(t+1) \rangle_{H_i(\sigma(t))}\right]^2$$

which is a measure of how well the data for spin σ_i is being fit by the model. Note that $\sigma^2 = 1$ so

$$D_i \equiv \sum_t \left[1 - \frac{\langle \sigma_i(t+1) \rangle_{H_i(\sigma(t))}}{\sigma_i(t+1)}\right]^2,$$

in which each term shows that the individual update of each observation is trying to reduce D_i. Of course, we want the overall D_i to decrease and because W_{ij} is the same for all the data points, not every observation gets its wish. **We stop the iteration when D_i starts to increase.**

The Sherrington-Kirkpatrick (SK) model assumes W_{ij} are normally distributed with zero mean and variance equal to g^2/N.



(i) Compute $H_i(\sigma(t)) = \sum_j W_{ij}\sigma_j(t)$ (initialize with a random W_{ij});

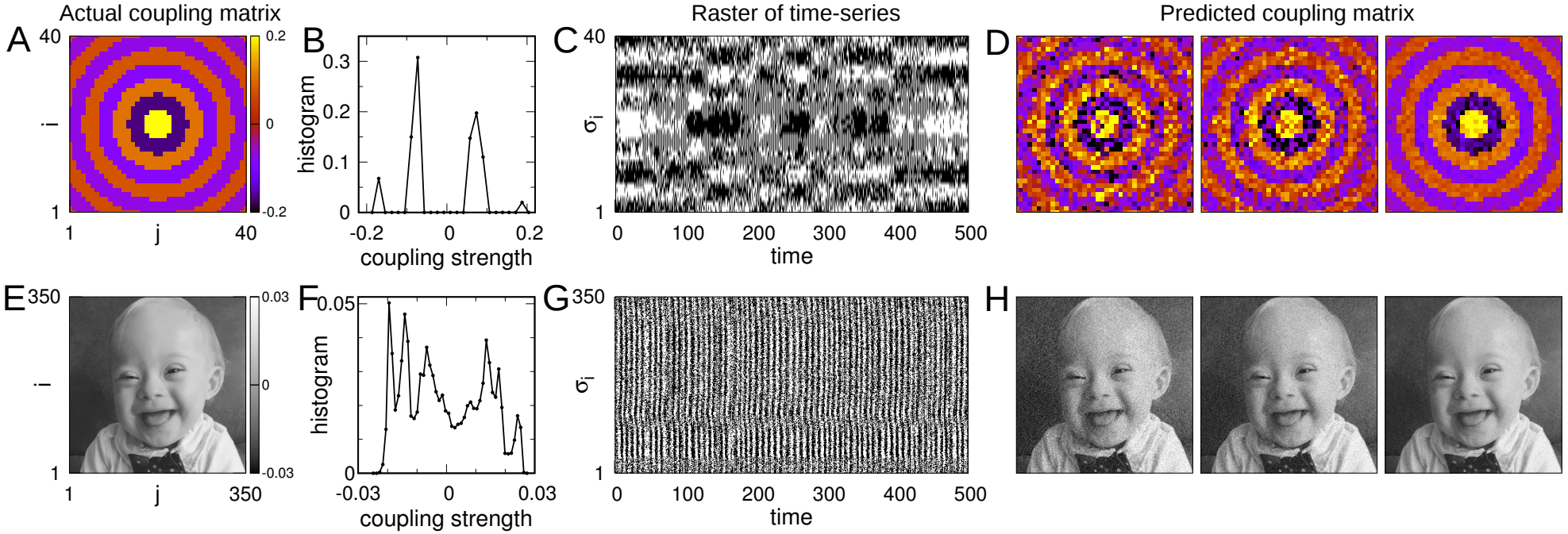
(ii) Compute $H_i^{\text{new}}(\sigma(t)) = \frac{\sigma_i(t+1)}{\langle\sigma_i(t+1)\rangle_{H_i(\sigma(t))}} H_i(\sigma(t))$;

(iii) Update $W_{ij} = W_{ij}^{\text{new}} \leftarrow \sum_k \langle \delta H_i^{\text{new}} \delta\sigma_k\rangle^* [C^{-1}]^*_{kj}$;

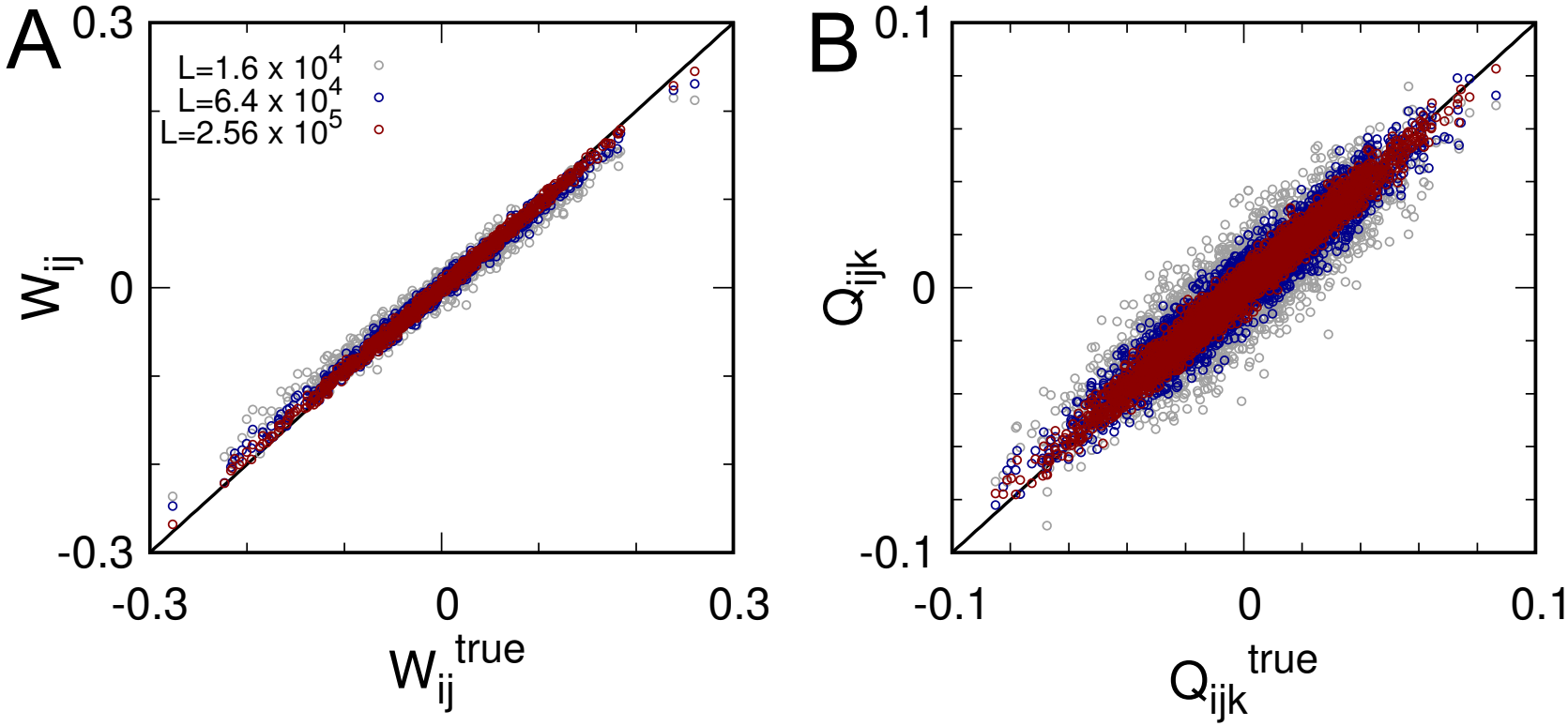
(iv) Repeat (i)-(iii) until D_i starts to increase;

(v) Compute (i)-(iv) in parallel for every index $i \in \{1, 2, \cdots, N\}$.

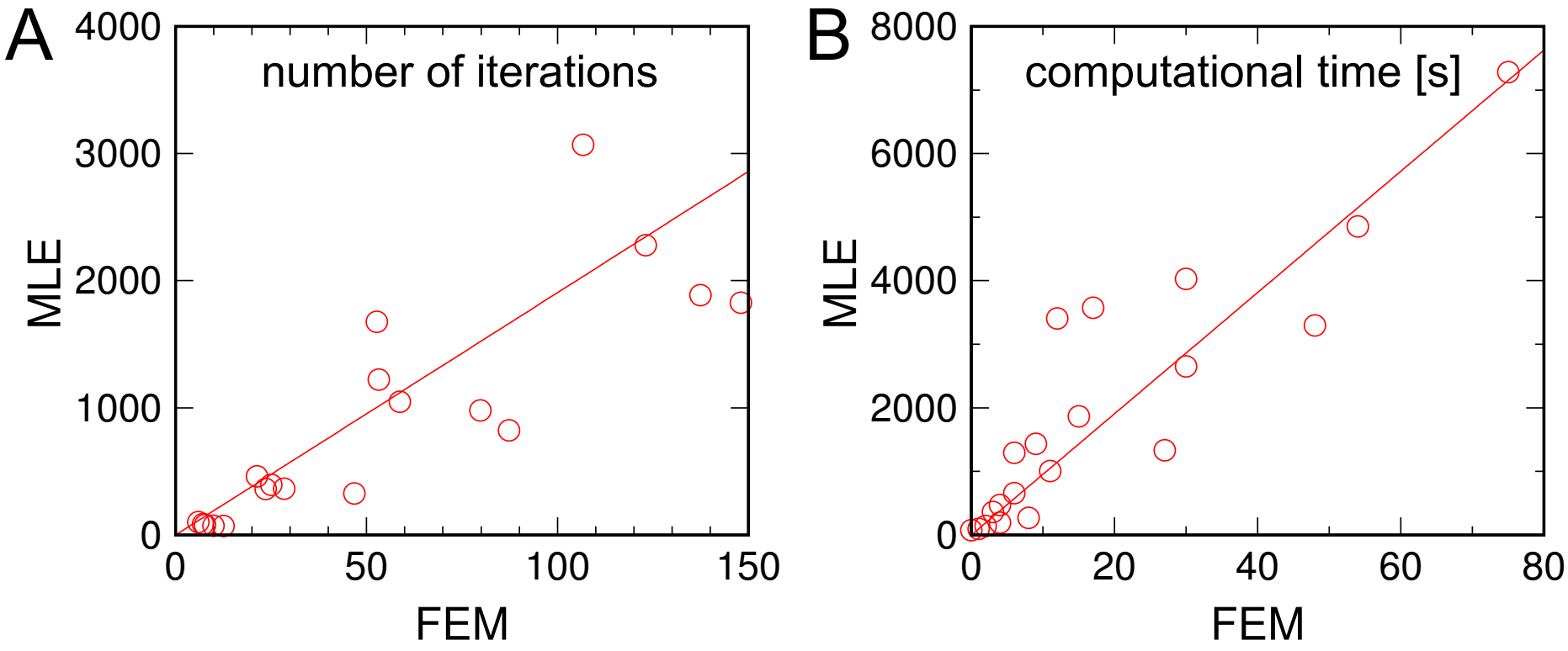
Other coupling distributions?



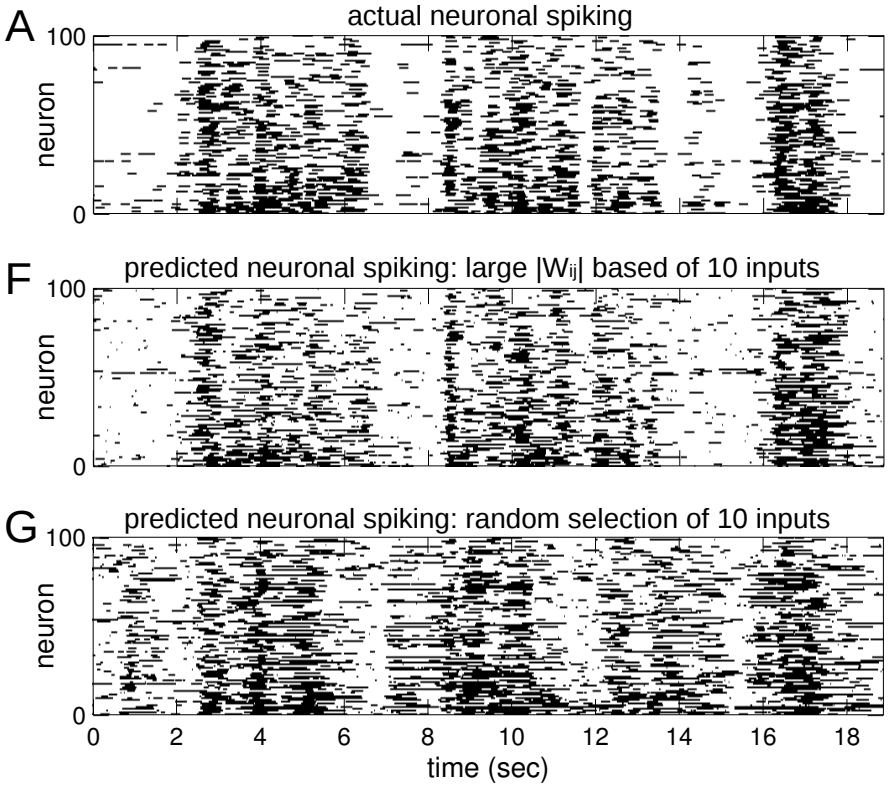
Higher order interactions?

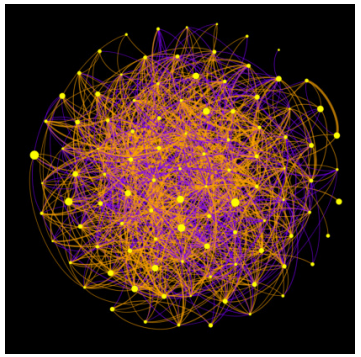


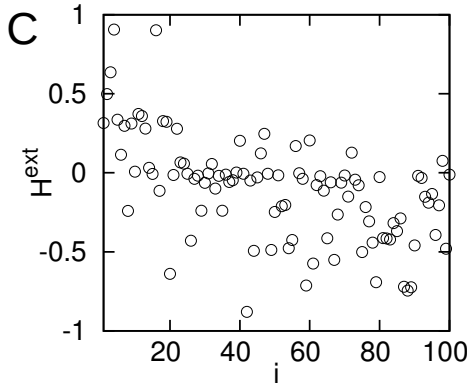
Benefit of a multiplicative update

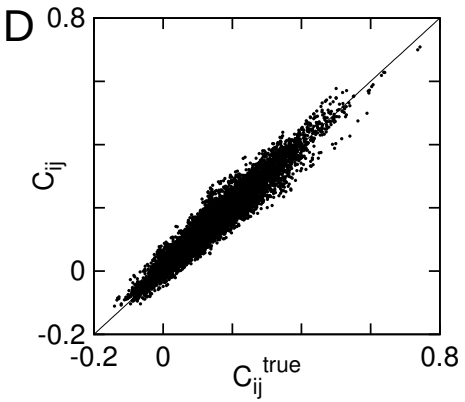


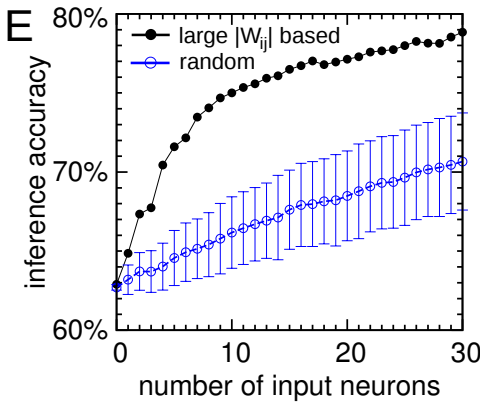
What happens when a salamander sees a movie of a fish?



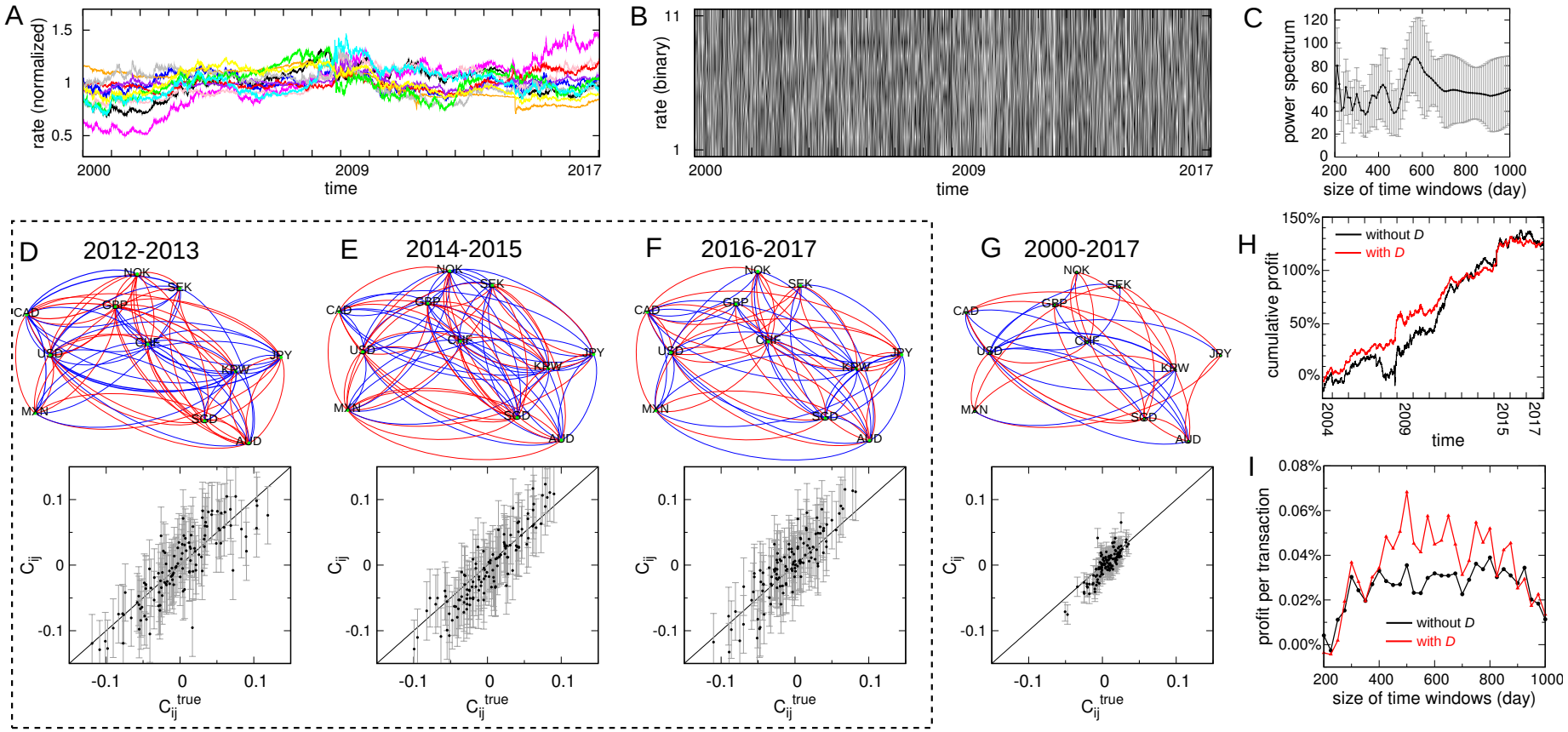








Money, money, money



Finding Hidden Variables

Expectation Maximization (EM) algorithm for hidden variables

Two alternating steps:

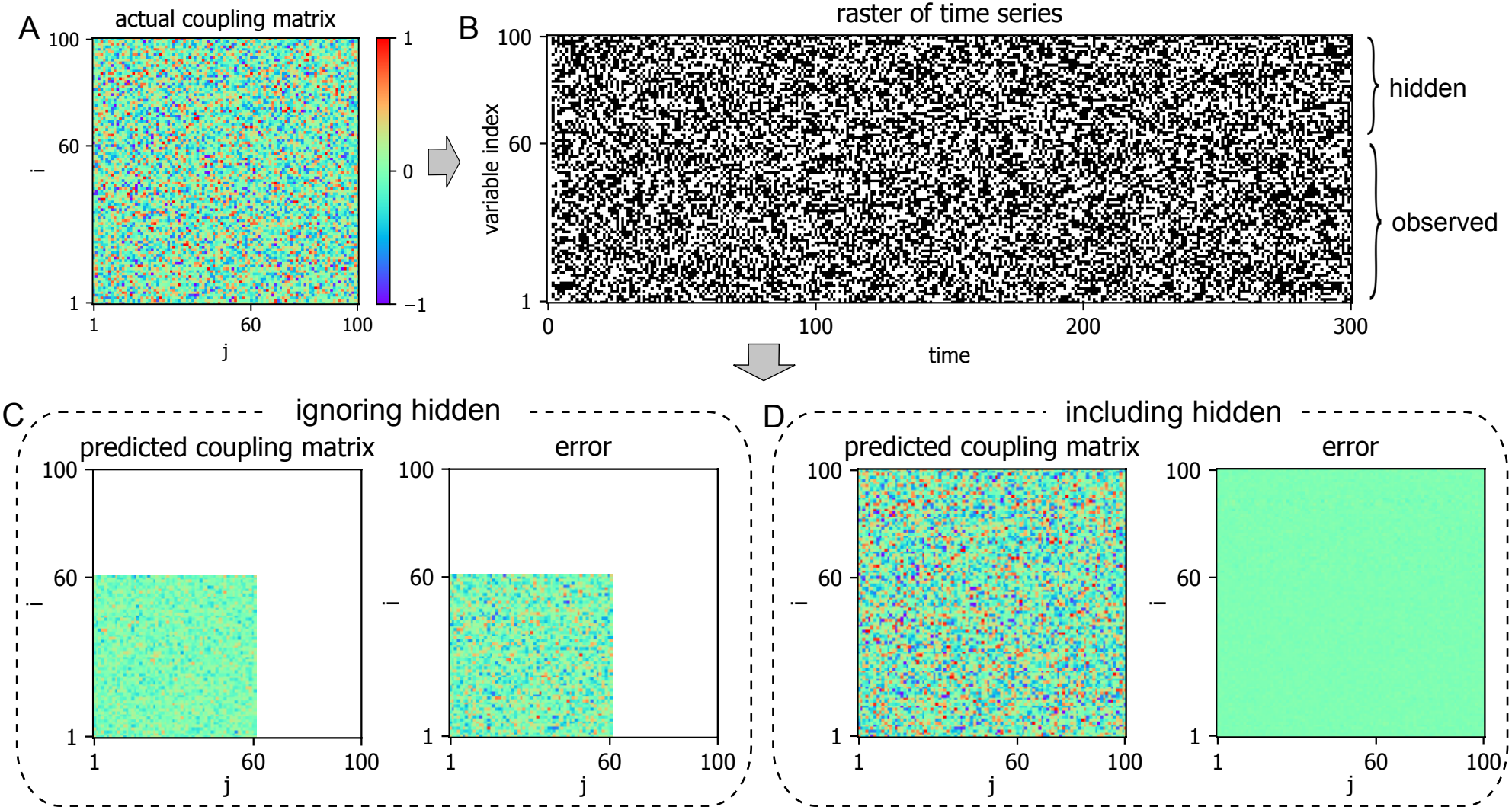
1. Inferring all interactions of observed and hidden variables from configurations of observed variables,
2. Reconstructing the configurations of hidden variables consistent with these inferred interactions.

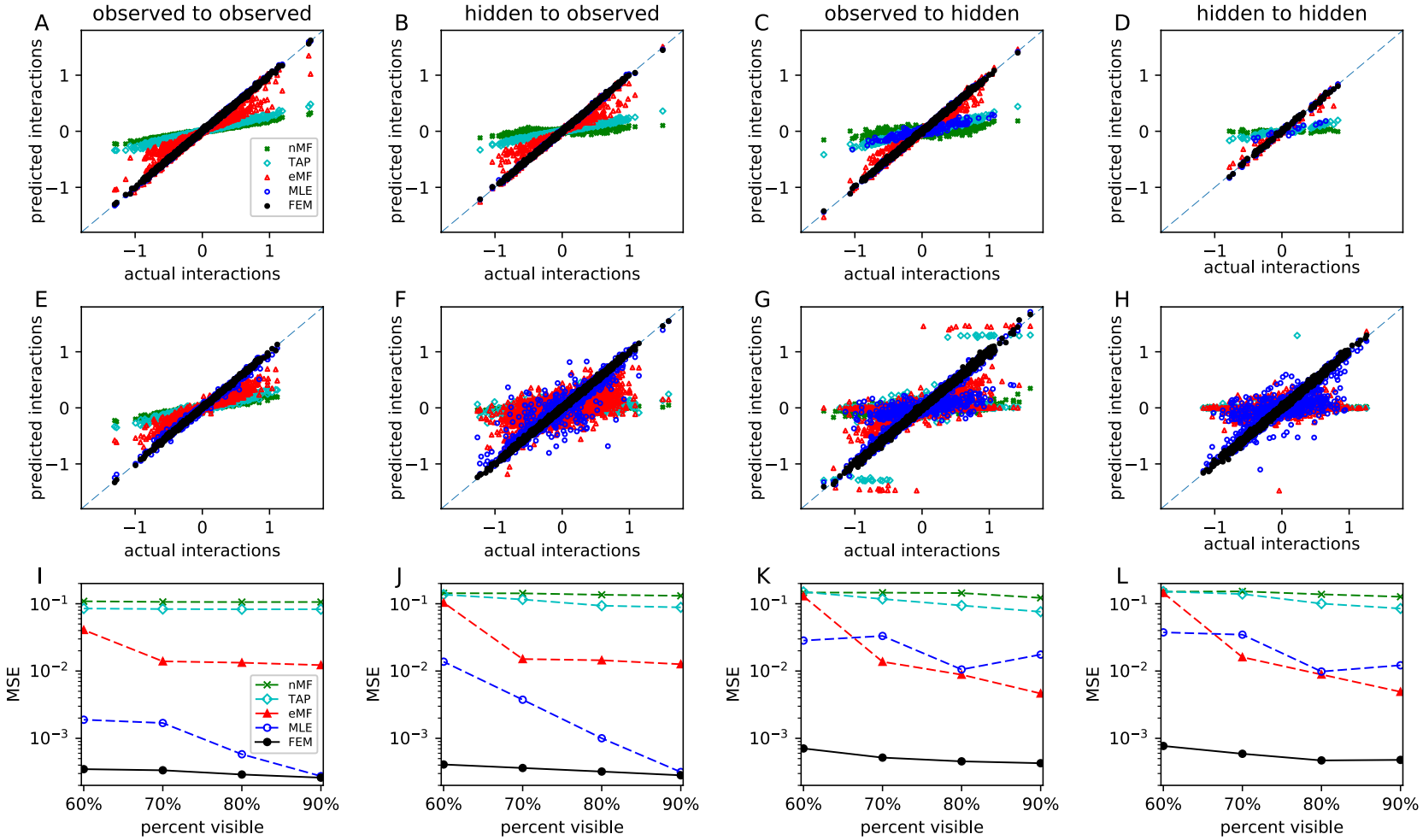
Hoang, Jo, Periwal, PRE

What is a hidden variable?

The problem is only well-posed if we **restrict the form of interactions** between the observed variables and the hidden variables.

Important part of the problem: **How many hidden variables?** You can keep adding hidden variables and make the observed data fit arbitrarily well — So there's always a problem of overfitting and not enough data.


A
actual coupling matrix
100
60
1
i
1
60
100
j
1
0
−1
B
raster of time series
100
60
1
variable index
0
100
200
300
time
hidden
observed
C
ignoring hidden
predicted coupling matrix
error
D
including hidden
predicted coupling matrix
error


A
observed to observed
B
hidden to observed
C
observed to hidden
D
hidden to hidden
predicted interactions
actual interactions
nMF
TAP
eMF
MLE
FEM
E
F
G
H
I
J
K
L
MSE
percent visible
60%
70%
80%
90%

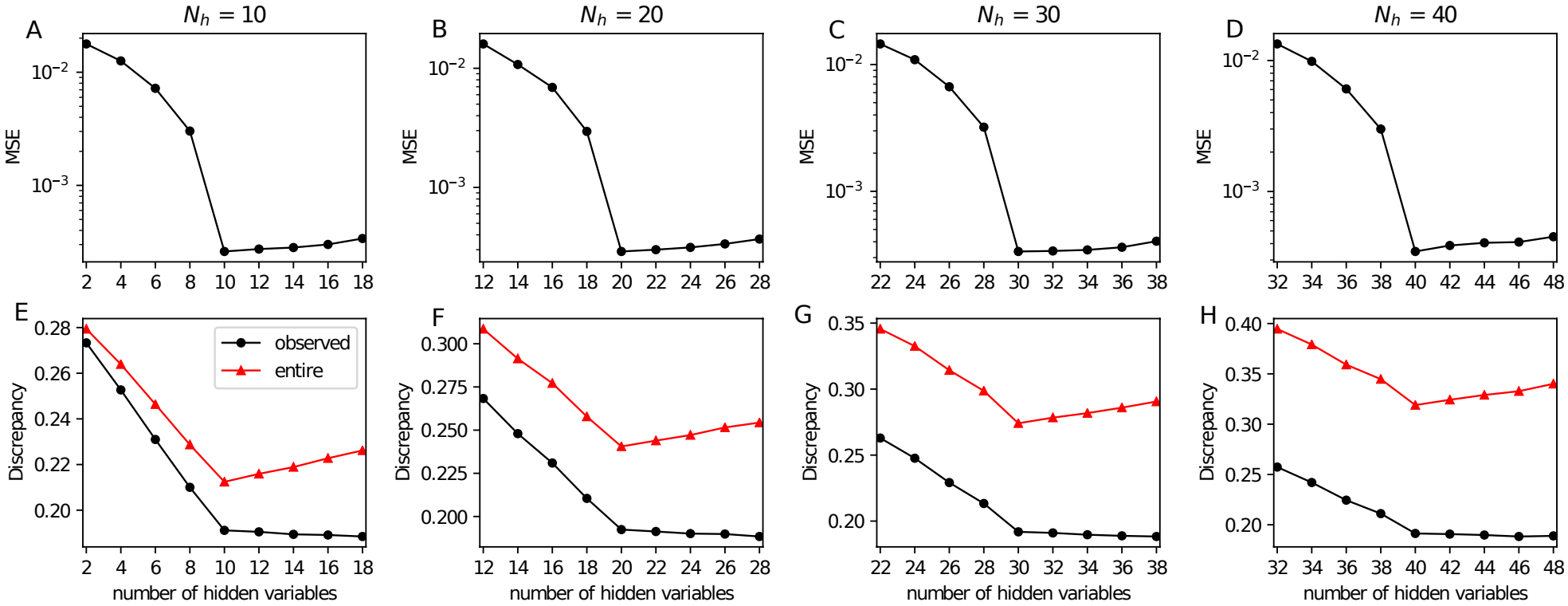
How many hidden variables?

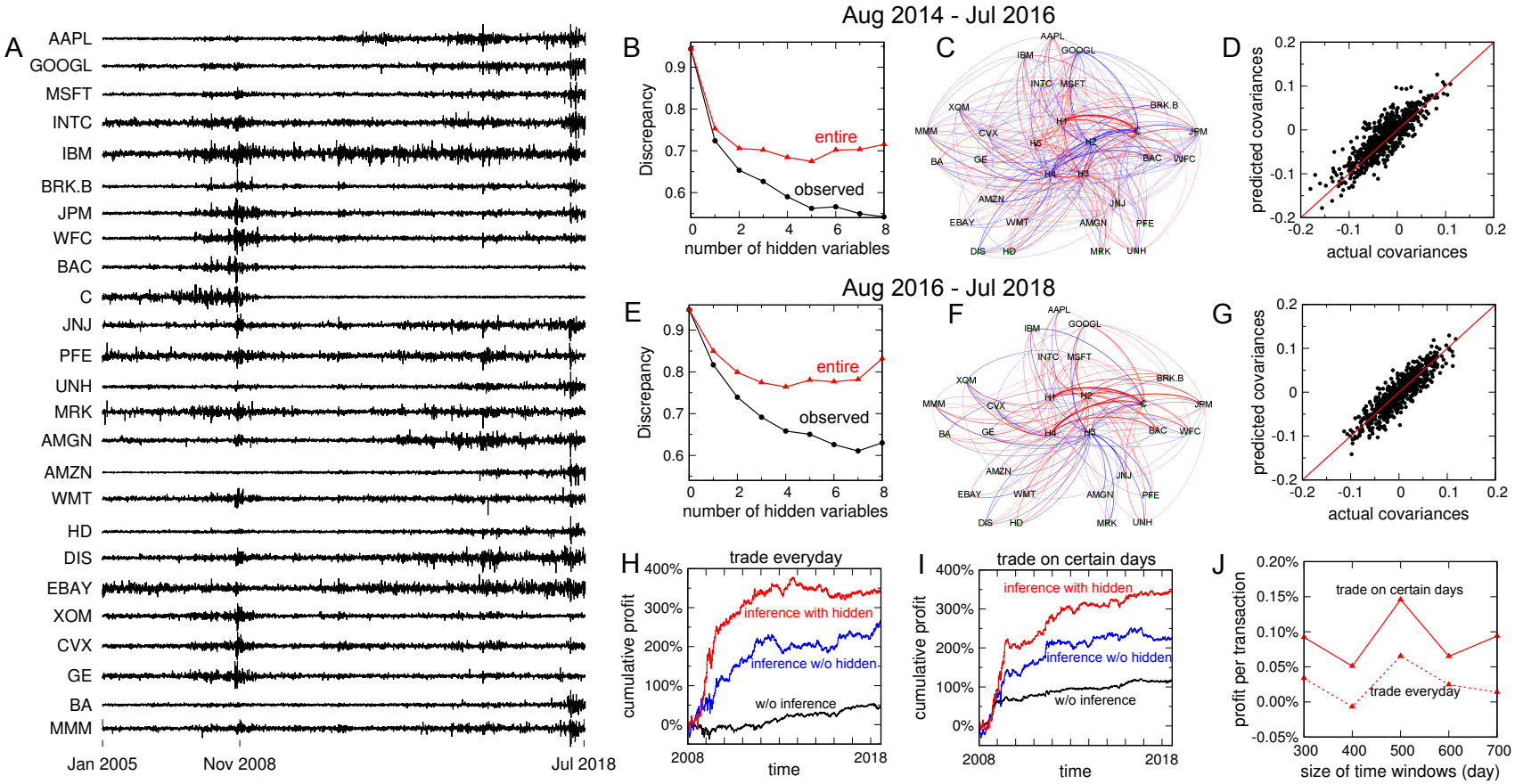
For a range of numbers of hidden variables, in parallel and independently,
(i) Assign configurations of hidden variables at random;
(ii) Infer interaction weights W_{ij} including observed-to-observed, hidden-to-observed, observed-to-hidden, and hidden-to-hidden from the configurations of observed and hidden variables using FEM;
(iii) Flip the states of hidden variables with probability $\mathcal{L}_{\text{after flip}}/(\mathcal{L}_{\text{before flip}} + \mathcal{L}_{\text{after flip}})$ (see below).
(iv) Repeat steps (ii) and (iii) until the discrepancy of observed variables is minimized. The final values of W_{ij} and hidden states are the inferred coupling weights and configurations of hidden spins, respectively.

Define the scaled discrepancy of the entire system based on the observed part as

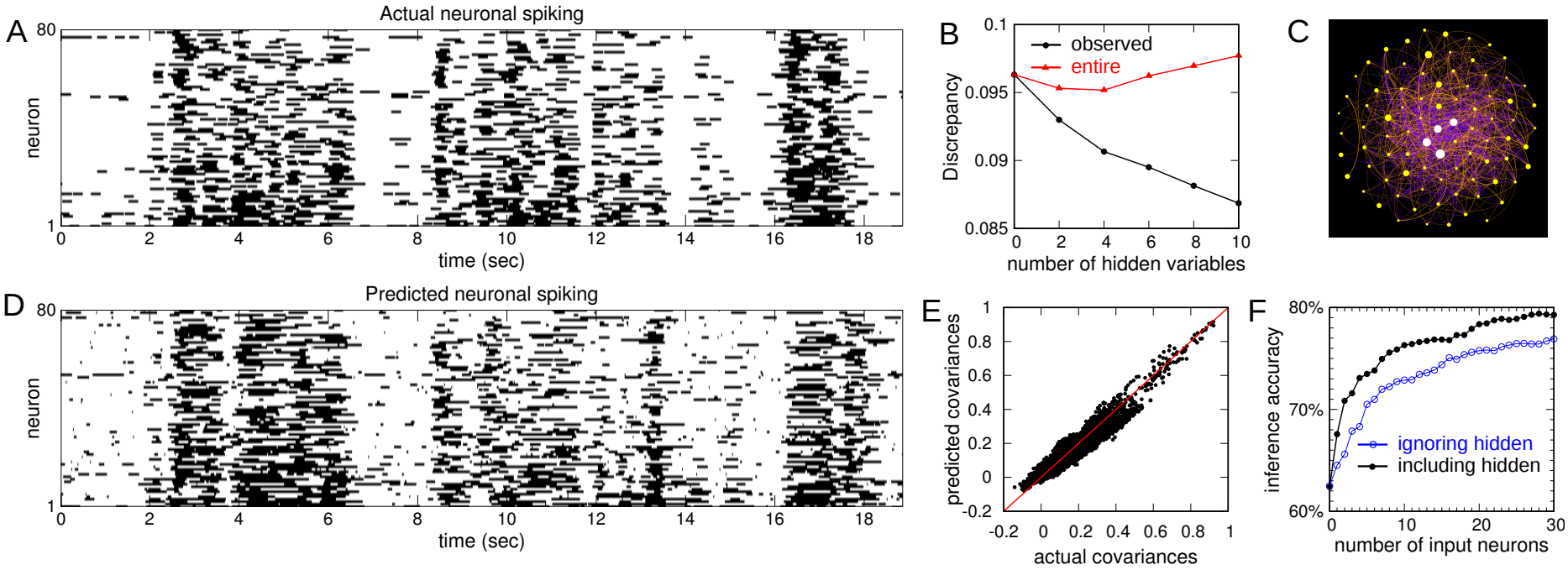
$$D \equiv D_{\text{visible}} \left(1 + \frac{N_h}{N_v} \right). \tag{1}$$

because the discrepancy in the hidden variables cannot be less than the discrepancy in the visible variables. You cannot fit unobserved variables better than your model fits the observed variables!
Note $D_{\text{visible}} \approx \ln \mathcal{L}_{\text{visible}}$ so this is a little like AIC/BIC.
We predict the right number of hidden variables by minimizing D.

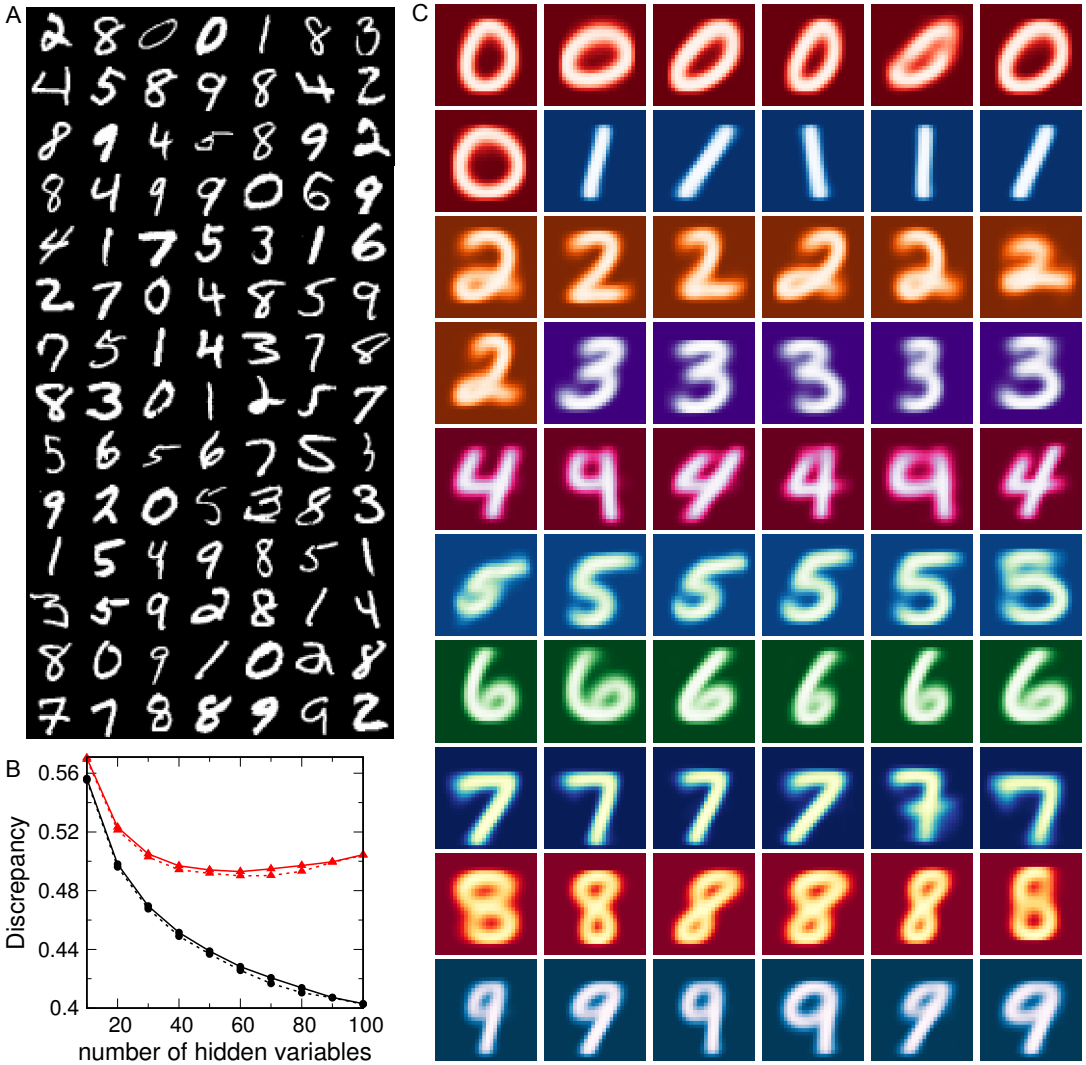

A
$N_h = 10$
B
$N_h = 20$
C
$N_h = 30$
D
$N_h = 40$
MSE
10^{-2}
10^{-3}
E
F
G
H
Discrepancy
observed
entire
number of hidden variables


A
AAPL
GOOGL
MSFT
INTC
IBM
BRK.B
JPM
WFC
BAC
C
JNJ
PFE
UNH
MRK
AMGN
AMZN
WMT
HD
DIS
EBAY
XOM
CVX
GE
BA
MMM
Jan 2005
Nov 2008
Jul 2018
Aug 2014 - Jul 2016
B
Discrepancy
entire
observed
number of hidden variables
C
D
predicted covariances
actual covariances
Aug 2016 - Jul 2018
E
Discrepancy
entire
observed
number of hidden variables
F
G
predicted covariances
actual covariances
H
trade everyday
cumulative profit
inference with hidden
inference w/o hidden
w/o inference
time
I
trade on certain days
cumulative profit
inference with hidden
inference w/o hidden
w/o inference
time
J
profit per transaction
trade on certain days
trade everyday
size of time windows (day)

Salamander and fish movies again!



MNIST digit classification



How about missing values in data?

1. This is different from hidden variables because there are random (we assume!) places where the data is missing values — happens in biological data sets all the time.

2. We still use the EM algorithm, with a stochastic approximation for the Expectation step (just like for hidden variables) — Stochastic Approximation EM (SAEM).

3. We do need a different stopping criterion because there is no simple scale argument giving us a discrepancy for stopping iteration.

Once missing data is properly restored, there should not be any distinction between observed and restored data. We stop when

$$D_{\text{observed}} = D_{\text{missing}}.$$

Lee, Periwal, Jo, PRE

As before, we want to find the coefficients of H_i using:

$$P(\sigma_i(t+1) = \pm 1 | \vec{\sigma}(t), \theta) = \frac{\exp[\pm H_i(t)]}{\exp[H_i(t)] + \exp[-H_i(t)]},$$

which can be written as a logistic regression problem as well.

$$P(\sigma_i(t+1) = 1 | \vec{\sigma}(t), \theta) = \frac{1}{1 + \exp[-2\sum_j W_{ij}\sigma_j(t) - 2b_i]}.$$

and solving this by logistic regression is equivalent to maximizing the total likelihood of the data.

For missing data, first assign random values to the missing entries, then find W_{ij}, then stochastically update the missing values as follows. Define

$$\mathcal{L}^{\pm}_{i,t} \equiv P(\sigma^m_i(t) = \pm 1 | \vec{\sigma}(t-1), \theta) \prod_{j=1}^{N} P(\sigma_j(t+1) | F^{\pm}_i(\vec{\sigma}(t), \theta),$$

where $F^{\pm}_i(\vec{\sigma}(t)) = (\sigma_1(t), \cdots, \sigma^m_i(t) = \pm 1, \cdots, \sigma_N(t))$ for $(i,t) \in \mathcal{M}$.

For the end times, $t = 0, L-1$:

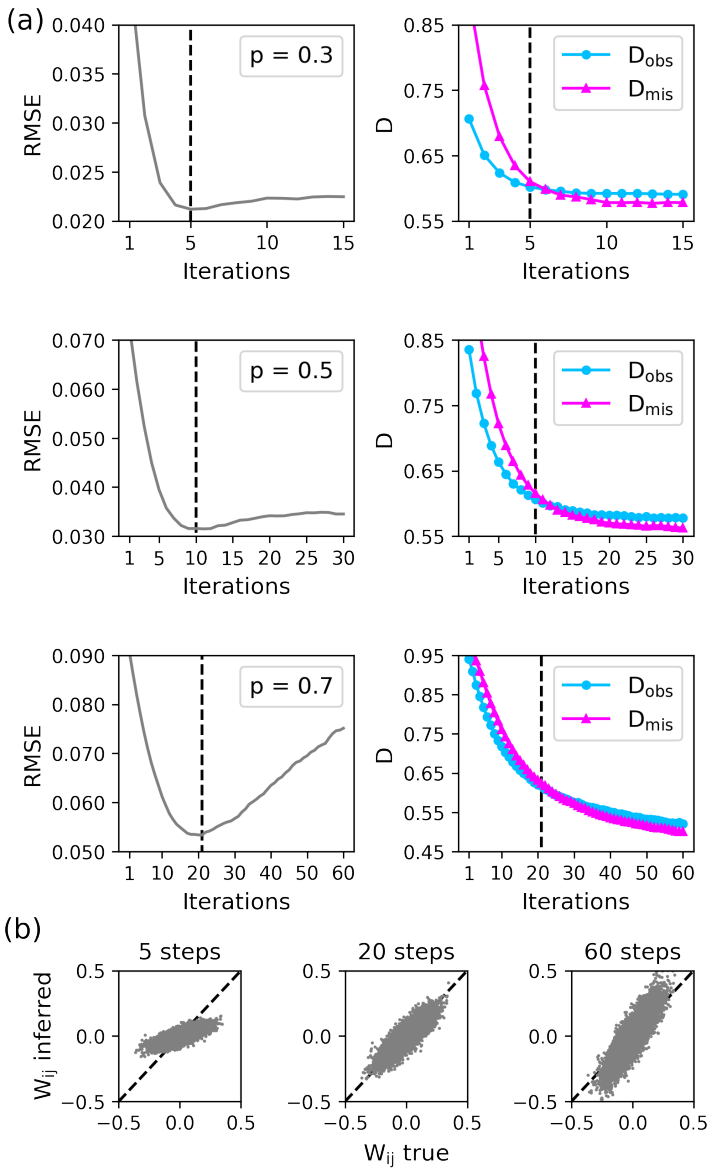
$$\mathcal{L}^{\pm}_{i,0} \equiv \prod_{j=1}^{N} P(\sigma_j(1) | F^{\pm}_i(\vec{\sigma}(0)), \theta)$$

and

$$\mathcal{L}^{\pm}_{i,L} \equiv P(\sigma^m_i(L) = \pm 1 | \vec{\sigma}(L-1), \theta).$$

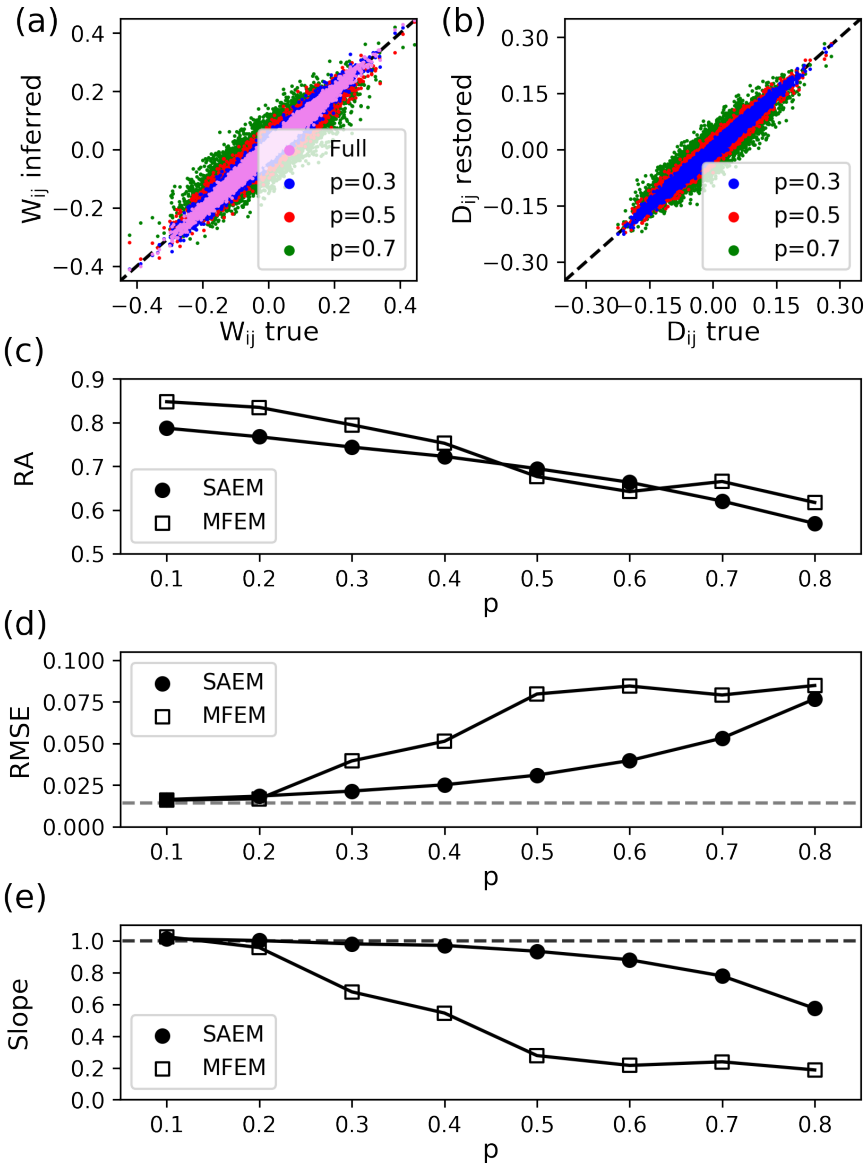
Stochastically re-assign ± 1 to $\sigma^m_i(t)$ with a probability of $\mathcal{L}^{\pm}_{i,t}/(\mathcal{L}^{+}_{i,t} + \mathcal{L}^{-}_{i,t})$ for every missing data point of $(i,t) \in \mathcal{M}$ with random order.

Synthetic Data

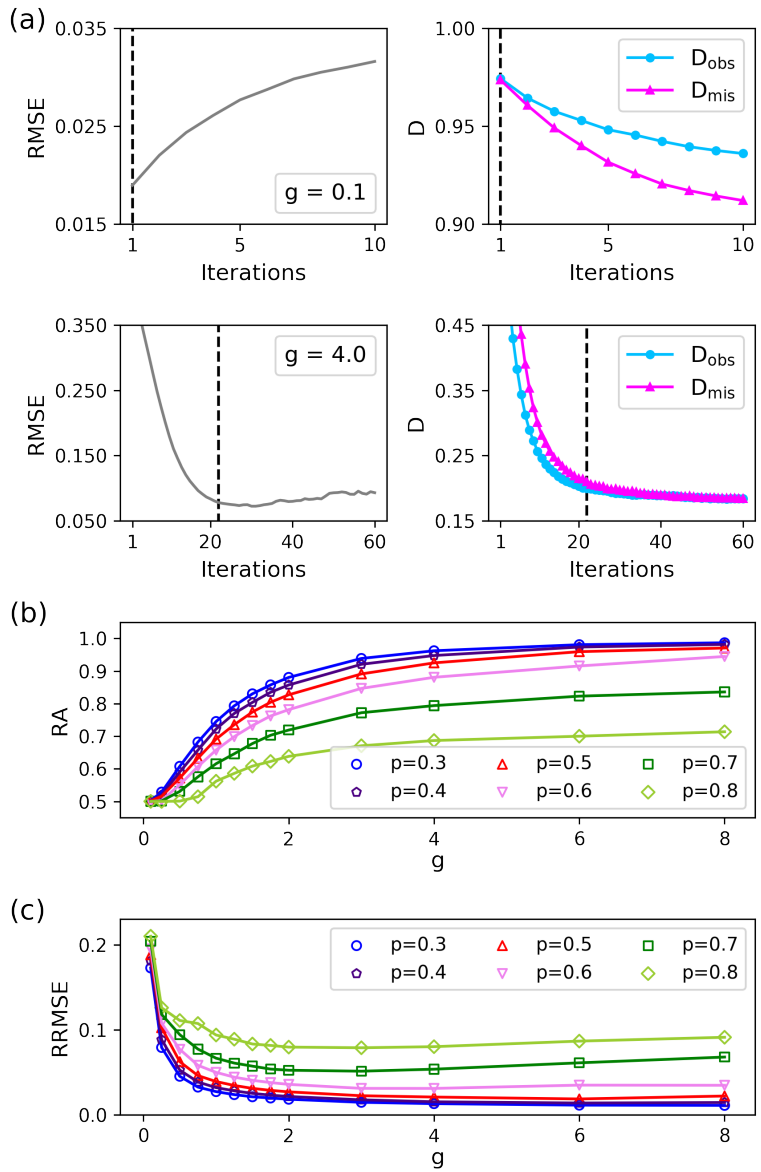


Mean field EM vs. SAEM

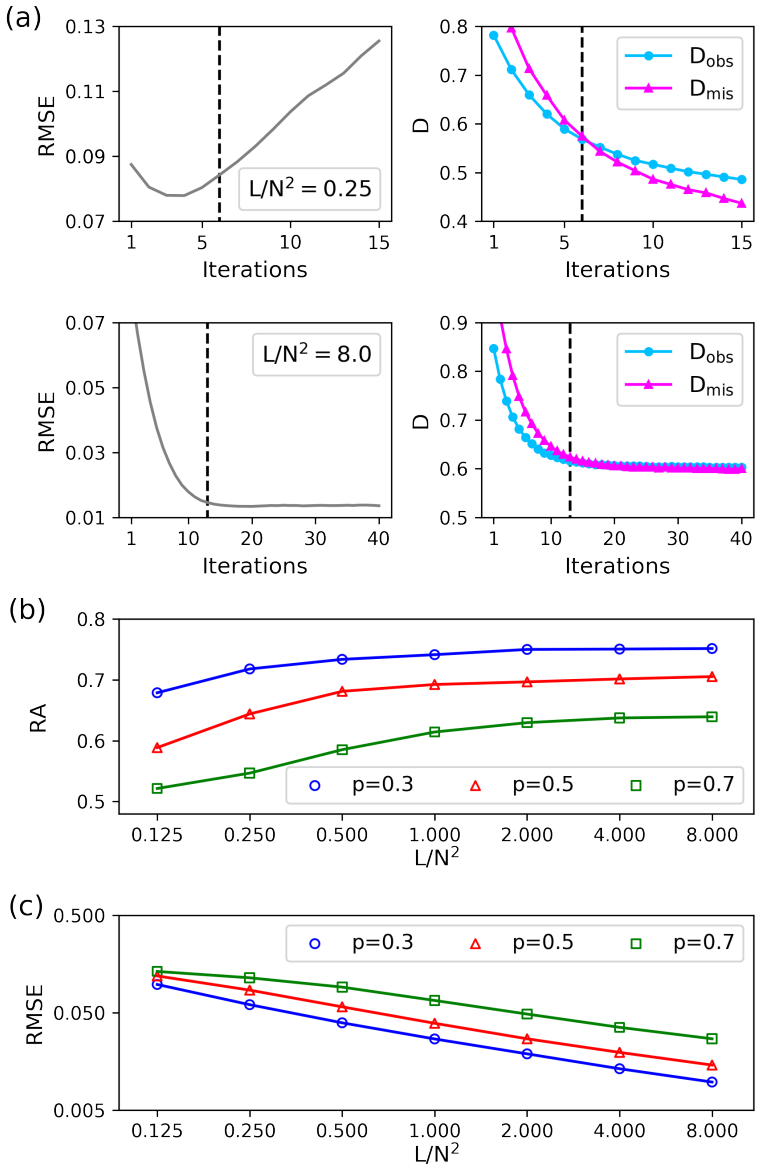
$$D_{ij} = \langle \sigma_i(t+1)\sigma_j(t) \rangle$$



Very strong and very weak coupling limits

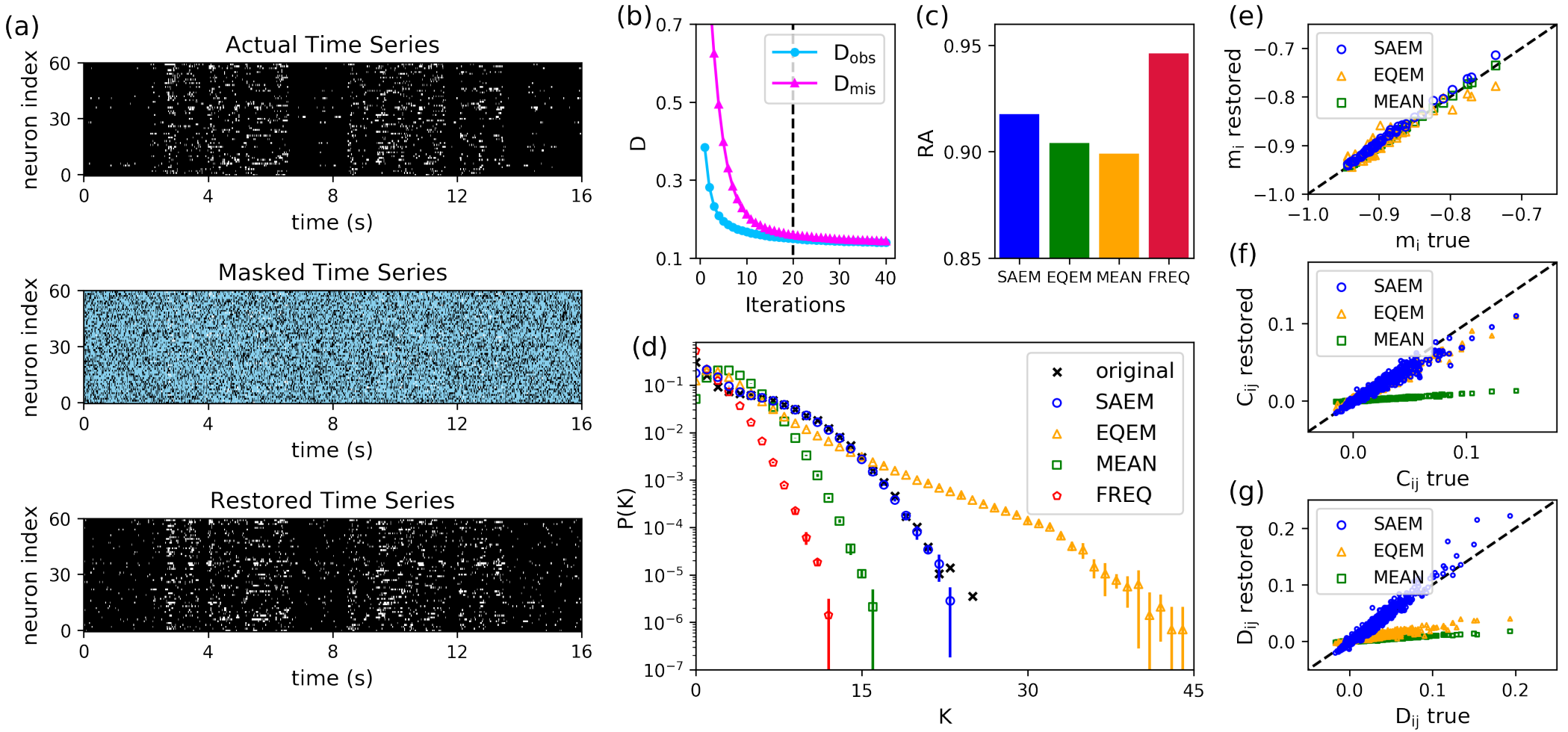


Dependence on the amount of data



$$K(t) = \frac{1}{2}\sum_{i=0}^{N}(\sigma_i(t)+1)$$

is the number of simultaneous spikes of N neurons.
$P(K)$ is the probability of K simultaneous spikes.


(a)
Actual Time Series
Masked Time Series
Restored Time Series
neuron index
time (s)
(b)
D
Iterations
D_{obs}
D_{mis}
(c)
RA
SAEM EQEM MEAN FREQ
(d)
P(K)
K
original
SAEM
EQEM
MEAN
FREQ
(e)
m_i restored
m_i true
(f)
C_{ij} restored
C_{ij} true
(g)
D_{ij} restored
D_{ij} true