北京师范大學
Beijing Normal University
全球变化与地球系统科学研究院
College of Global Change and Earth System Science

On the Information in Models: quantification of model structure adequacy with information based metrics

Wei Gong, Junior Research Scientist
GCESS, Beijing Normal University
Email: gongwei2012@bnu.edu.cn

April-24-2016

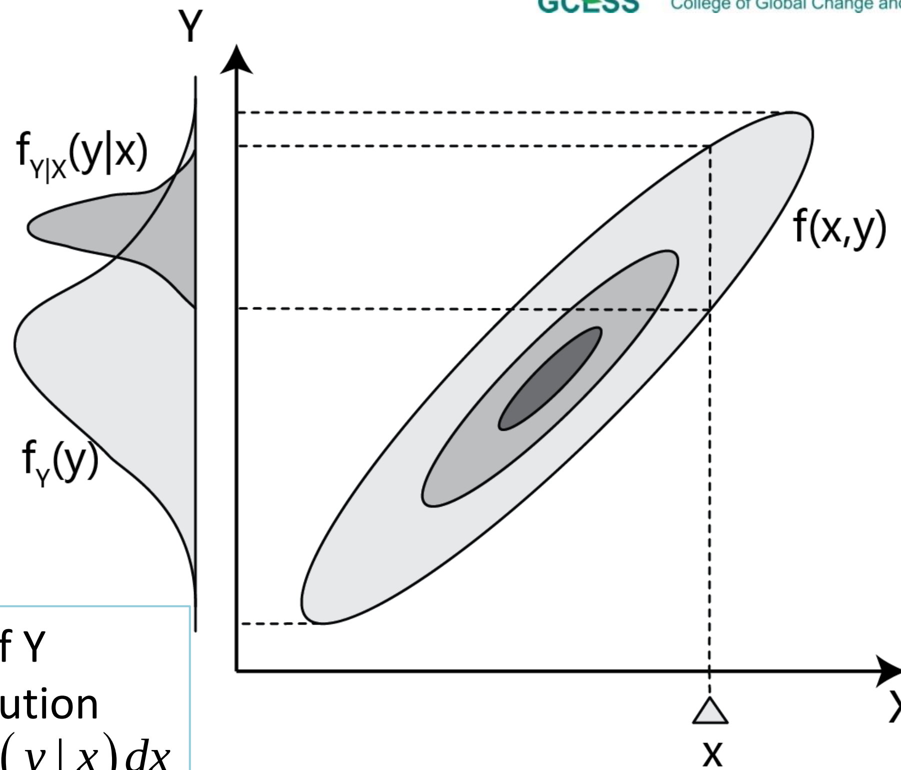
Outlines

- Information and model adequacy
- Information metrics
  - manifold learning based method
  - ICA (Independent Component Analysis)
- Estimating entropy for 1D case
- Estimating entropy for multi-dimensional case
- Discussions

Information and model adequacy

北京師範大學
Beijing Normal University
全球变化与地球系统科学研究院
College of Global Change and Earth System Science
GCESS

Random variables X， Y
Input X， Output Y
Marginal distribution $f_X(x)$
$f_Y(y)$
Joint distribution $f(x,y)$



Entropy of Y
Defined with marginal dist.

$$H(Y)=\int f_Y(y)\log f_Y(y)dx$$

Given X， conditional entropy of Y
Defined with conditional distribution

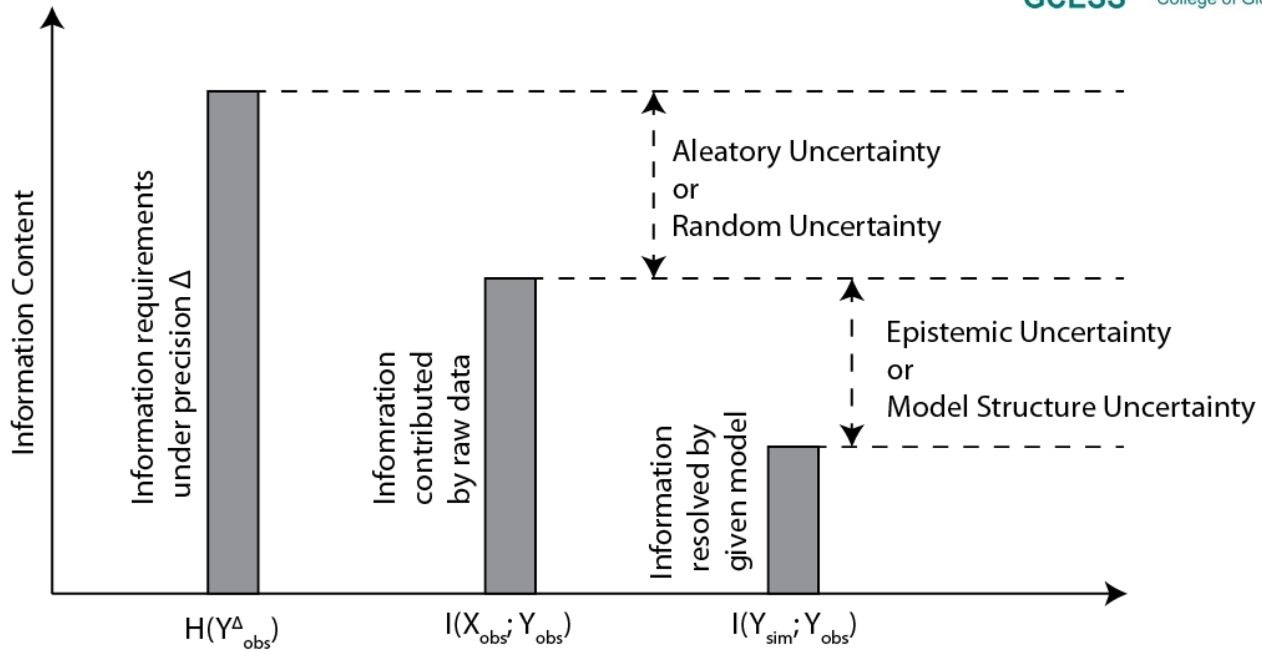
$$H(Y|X)=\int f_{Y|X}(y|x)\log f_{Y|X}(y|x)dx$$

Information from X to Y (Mutual information)

$$I(X;Y)=\iint f(x,y)\log\frac{f(x,y)}{f_X(x)f_Y(y)}dxdy=H(Y)-H(Y|X)$$

Entropy： discrepancy of X —— similar with standard deviation
Mutual Information： relationship between X and Y

Information and model adequacy



- Input: X_{obs}; Output: Y_{obs}; Simulated: Y_{sim}
- Information requirement: $H(Y_{obs}^{\Delta})$

$$H\left(Y^{\Delta}\right) = -\sum_{-\infty}^{\infty} P_i \ln P_i = -\sum_{-\infty}^{\infty} f_y\left(y_{i+1/2}\right)\Delta_i \ln\left(f_y\left(y_{i+1/2}\right)\Delta_i\right)$$

- Information contributed by raw data: $I(X_{obs};Y_{obs})$

$$I(X_1, X_2, ..., X_m; Y) = H(X_1, X_2, ..., X_m) + H(Y) - H(X_1, X_2, ..., X_m, Y)$$

- Information resolved by model: $I(Y_{sim};Y_{obs})$

$$I(X;Y) = \int\int f_{x,y}(x,y) \ln \frac{f_{x,y}(x,y)}{f_x(x) f_y(y)} dx\, dy$$

Data processing inequality

- if X、Y、Z is a Markov chain $X \leftrightarrow Y \leftrightarrow Z$, then

$$I(X;Y) \geq I(X;Z)$$

- Y: input data, note as *Data*
- X: observed output, note as Q_{obs}
- Z: simulated output, note as $Q_{sim}=f(Data)$
- Data processing inequality：

$$I(Q_{obs};Data) \geq I(Q_{obs};Q_{sim})$$

There is no data processing method that can ‘create’ information, it can only use the information of data.

Information from raw data is always more than that in simulated output.

Epistemic uncertainty comes from data processing inequality.

Information and model adequacy

- Best Achievable Performance (BAP)
  - If a model can sufficiently use all the information offered by raw data, it achieves BAP.
- How to qualify the total information offered by observed input/output data?

Namely, mutual information I(input;output).

$$I(X_1, X_2, ..., X_m; Y) = H(X_1, X_2, ..., X_m) + H(Y) - H(X_1, X_2, ..., X_m, Y)$$

- Information metrics: manifold learning
- Information metrics: ICA

Information metrics: manifold learning

北京師範大學
Beijing Normal University
全球变化与地球系统科学研究院
College of Global Change and Earth System Science
GCESS

Find Nearest Neighbors [ANN]

Dimensionality Reduction [ISOMAP]

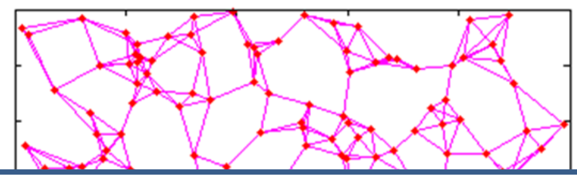
Estimate Shannon Information Entropy [kNNG]

Compute Multivariate Mutual Information

Approximate Nearest Neighbors (ANN)
[Arya et.al., 1994; 1998; 2009]

ISOmetric feature MAPping (ISOMAP)
[Tenenbaum et.al., 2000]

K-Nearest-Neighbor Graph (kNNG)
[Leonenko et al., 2008]

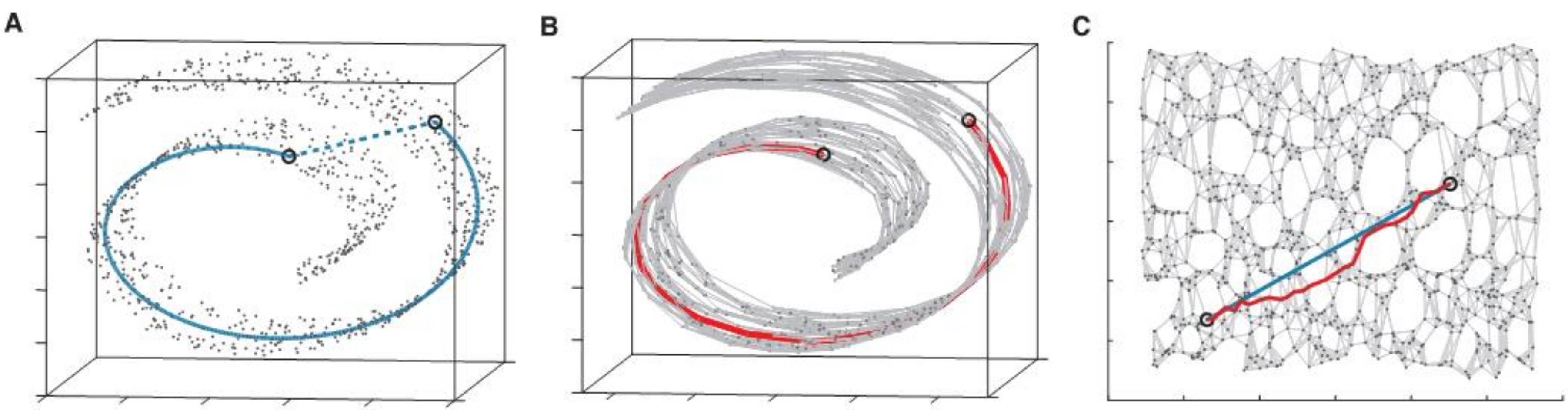


Chain Rule for
Multivariate Mutual Information

$$I\left(X_1, X_2, \ldots, X_m; Y\right) = H\left(X_1, X_2, \ldots, X_m\right) + H\left(Y\right) - H\left(X_1, X_2, \ldots, X_m, Y\right)$$

ISOMAP

- complete ISOmetric feature MAPing [Tenenbaum, et al. 2000]
- 3 steps
  - 1. Construct neighborhood graph
    (Original: compute directly. We use ANN query)
  - 2. Compute shortest geodesic paths
    (Dijkstra algorithm)
  - 3. Construct m-dimensional embedding
    (Multidimensional Scaling, MDS)



Compute N-dimensional Entropy directly from kNNG

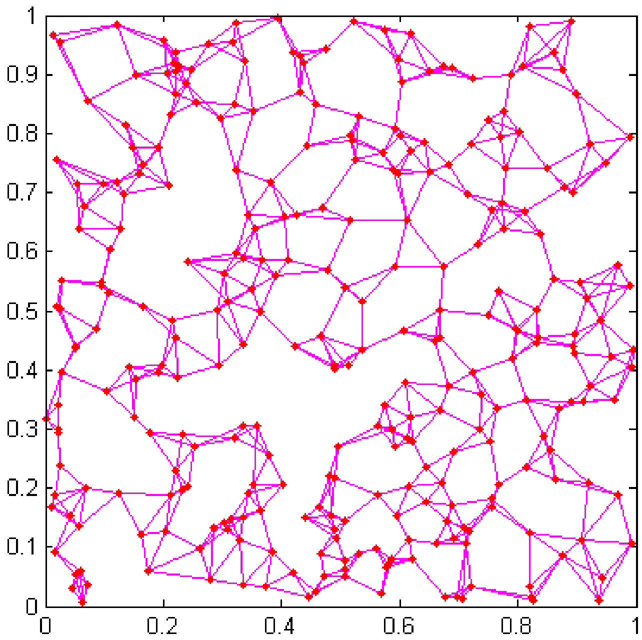
- Use Leonenko's method to compute Shannon entropy directly.
- [Leonenko et.al. 2008]

$$H_{n,k} = \frac{1}{n}\sum_{i=1}^{n}\log \xi_{n,i,k}$$

$$\xi_{n,i,k} = (n-1)\exp[-\Psi(k)]V_m\left(e_{k,n-1}^{(i)}\right)^m$$

$$\Psi(k) = \Gamma'(k)/\Gamma(k)$$

$$V_m = \pi^{m/2}/\Gamma(m/2+1)$$



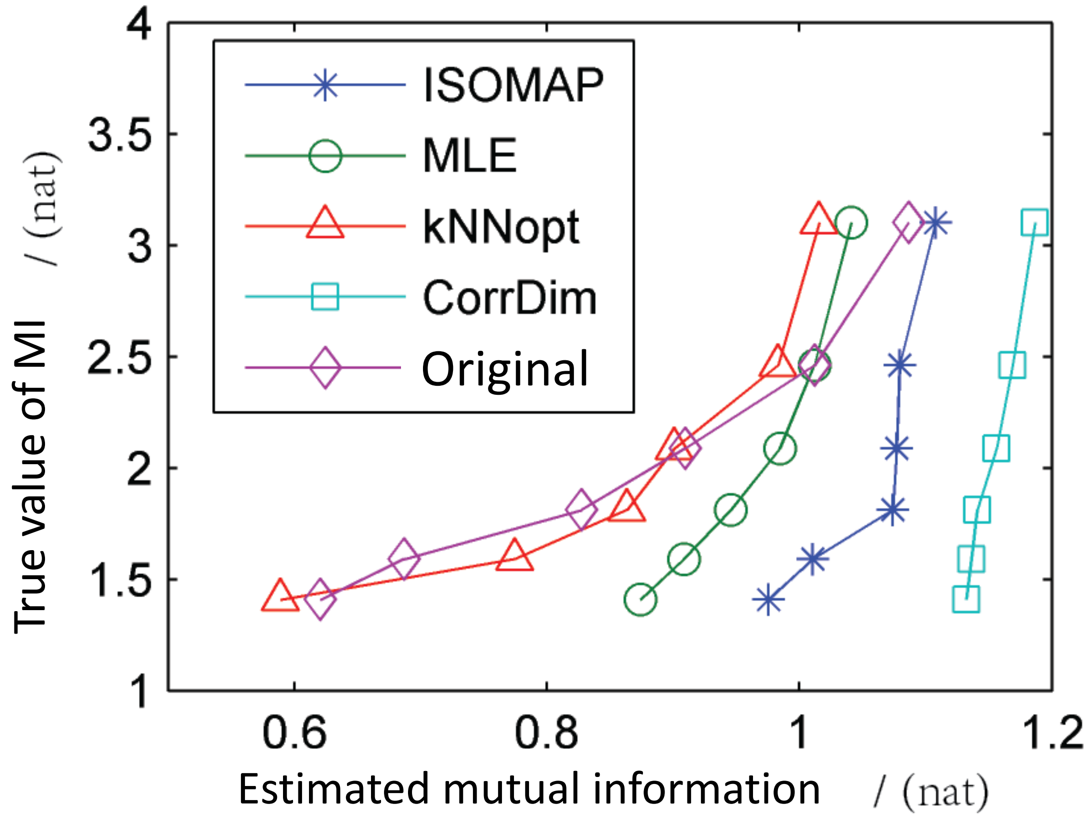
- Where: k is the number of nearest neighbor

  m is the intrinsic dimension given by ISOMAP

  $e_{k,n-1}^{(i)}$ is the distance between point $\boldsymbol{x}_i$ and its i-th nearest neighbor

Information metrics: manifold learning

北京师范大學
Beijing Normal University
全球变化与地球系统科学研究院
College of Global Change and Earth System Science



For hydrology data (have obs error), the estimated MI has similar trend, but significantly biased.

Information metrics: ICA

- For 1D H(X): plug-in estimators (PDF => entropy)
  bin-counting, kernel, ASH …
- For high-dimension (>3D), curse of dimensionality
  To hard to get PDF!
- How to get $H(X_1, X_2, \ldots, X_m)$ without computing high-dimensional PDF?
- If $X_1, X_2, \ldots, X_m$ are independent, things becomes easy!

$$H(X_1, X_2, \ldots X_m) = \sum_{i=1}^{m} H(X_i | X_{i-1}, X_{i-2}, \ldots X_1) \leq \sum_{i=1}^{m} H(X_i)$$

How to compute $H(X_1, X_2, \dots, X_m)$?

- If we can transform $\mathbf{X} = [X_1, X_2, \dots, X_m\ ]$ into independent signals $\mathbf{S} = [S_1, S_2, \dots, S_m\ ]$:

$$
\begin{aligned}
H(\mathbf{X}) = H(\mathbf{AS}) &= H(\mathbf{S}) + \log|\det(\mathbf{A})| \\
&= \sum_{i=1}^{m} H(S_i) + \log|\det(\mathbf{A})|
\end{aligned}
$$

- Only use 1D estimator, very Easy!

How to transform **X** to **S**

北京师范大学
Beijing Normal University
GCESS
全球变化与地球系统科学研究院
College of Global Change and Earth System Science

- For Gaussian distributions – PCA (Principle Component Analysis)
- For non-Gaussian distributions – ICA (Independent Component Analysis)

What is ICA ?

- 1) Assume a linear transformation: **X = AS**
- 2) Inversely, **S = WX**
- 3) Due to **CLT**, sum of **X** (namely $y = \boldsymbol{w}^T\boldsymbol{x}$) is more closer to Gaussian, if **X** is non-Gaussian.
- 4) The metric of nongaussianity is **Negentropy**.
- 5) **ICA** is find the optimal **A** that can maximize nongaussianity of $y = \boldsymbol{w}^T\boldsymbol{x}$.

The great leap

北京师范大学
Beijing Normal University
全球变化与地球系统科学研究院
College of Global Change and Earth System Science
GCESS

- **Negentropy.**

$$J(\boldsymbol{y}) = H(\boldsymbol{y}_{gauss}) - H(\boldsymbol{y})$$

- But we can use approximation of J base on nonquadratic function G (v is std Gaussian)

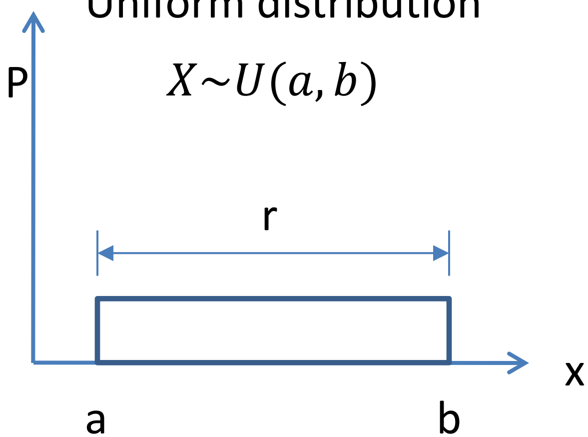
$$J(y) \propto [E\{G(y)\} - E\{G(v)\}]^2$$

- By estimating entropy with ICA, we got a more **accurate** estimation from a **coarse** estimation.

Synthetic study

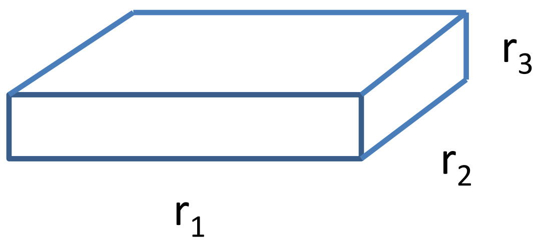
Uniform distribution

$X \sim U(a, b)$



$$H(X) = -\int_a^b \frac{1}{b-a} log \frac{1}{b-a} dx = \log(b-a) = \log(r)$$

Similarly



$$H(\boldsymbol{X}) = \log(r_1 r_2 r_3)$$

For high-dimensional uniform distribution $H(\boldsymbol{X}) = \log(r_1 r_2 r_3 \ldots r_m)$

Where m is the dimension

$H(\boldsymbol{X})$ doesn't change under affine transformation.

Synthetic study

北京師範大學 Beijing Normal University
GCESS 全球变化与地球系统科学研究院 College of Global Change and Earth System Science

$[r_1, r_2, \dots, r_m]$ =[1,2,3,...,10] $H(\boldsymbol{X}) = \log(10!) \approx 15.1044$

Replicate 10 times for each sample size

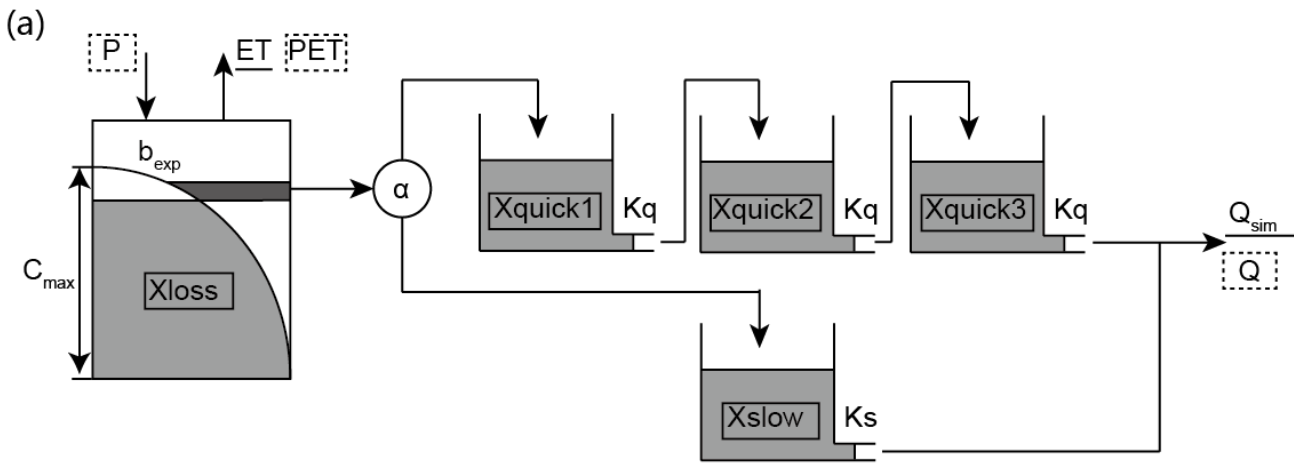
Sample size N	1000	10000	100000	1000000
1	15.4921	15.3116	15.2003	15.1495
2	15.5636	15.3058	15.209	15.1511
3	15.5431	15.2917	15.2043	15.1493
4	15.5372	15.3166	15.1983	15.1519
5	15.4783	15.3154	15.2041	15.1501
6	15.4550	15.3157	15.1999	15.1496
7	15.5668	15.3032	15.2018	15.1500
8	15.5533	15.3236	15.2009	15.1506
9	15.5415	15.3116	15.1949	15.1499
10	15.4868	15.334	15.2041	15.1495
mean	15.5218	15.3129	15.2018	15.1502
var	0.00159	0.00013	1.5E-05	6.8E-07
mean error	0.41737	0.20852	0.09736	0.04575

1) About 2.5% relative error when N = 1000, m = 10
2) Error and variance decrease with increasing sample size N

Case study

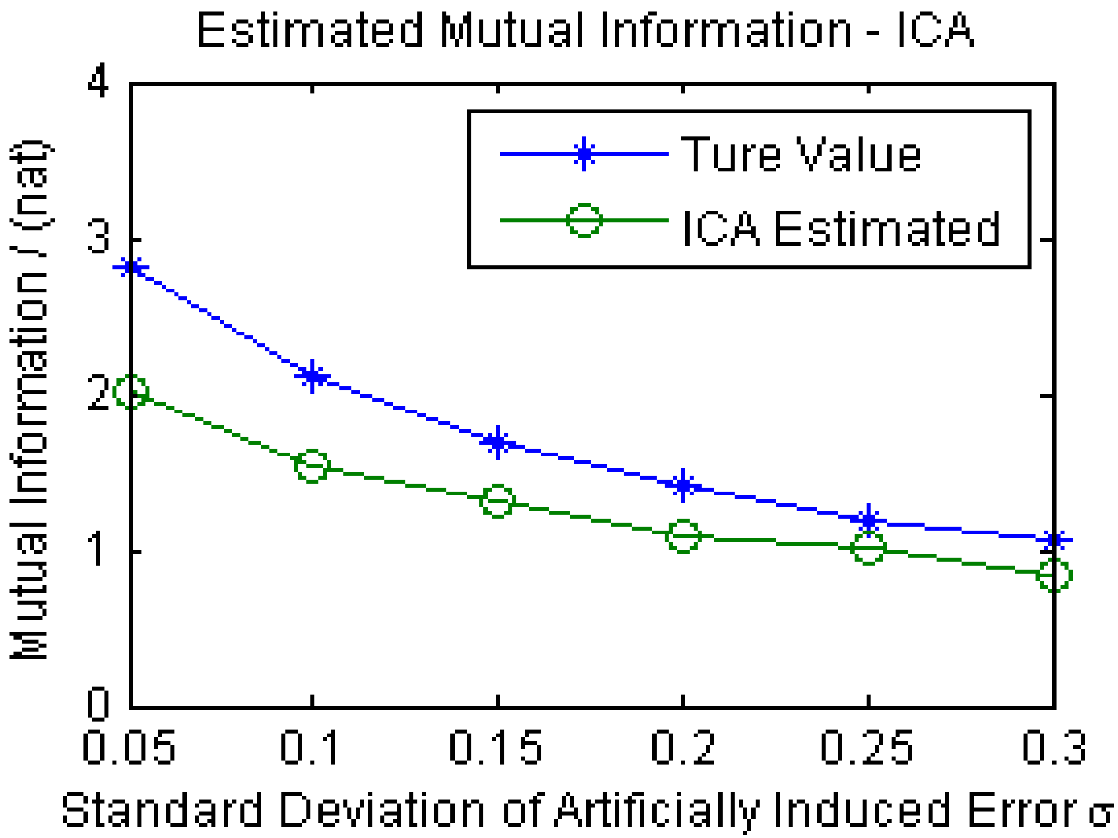
- 1) Simulation study
  - Leaf River, 1948~1978, HyMod
- 2) Inter-comparison of 3 catchments, 3 models
  - Leaf River, 1948~1978, HyMod and SAC
  - Chunky River, 1948~1978, HyMod and SAC
  - Chuzhou, 1980~2000, HyMod and X3M (Xinanjiang Model)

simulation study



- Assuming input variables P and PET are known.
- Assuming the hydrological processes can be completely expressed by HyMod, namely, the true value of streamflow Q_{true} is the simulation result of HyMod.
- Artificially induce (heteroscedastic) error in P, PET and Q_{true}, we get "observed" precipitation P_{obs}, pan evapotranspiration PET_{obs}, and streamflow Q_{obs}.
- Q_{sim} is the simulation result of HyMod with input P_{obs} and PET_{obs}.
- Because HyMod is the "true" model, $I(Q_{sim};Q_{obs}) = I(P_{obs},PET_{obs},\ldots;Q_{obs})$
- Compare $I(Q_{sim};Q_{obs})$ and $I(P_{obs},PET_{obs},\ldots;Q_{obs})$

simulation study

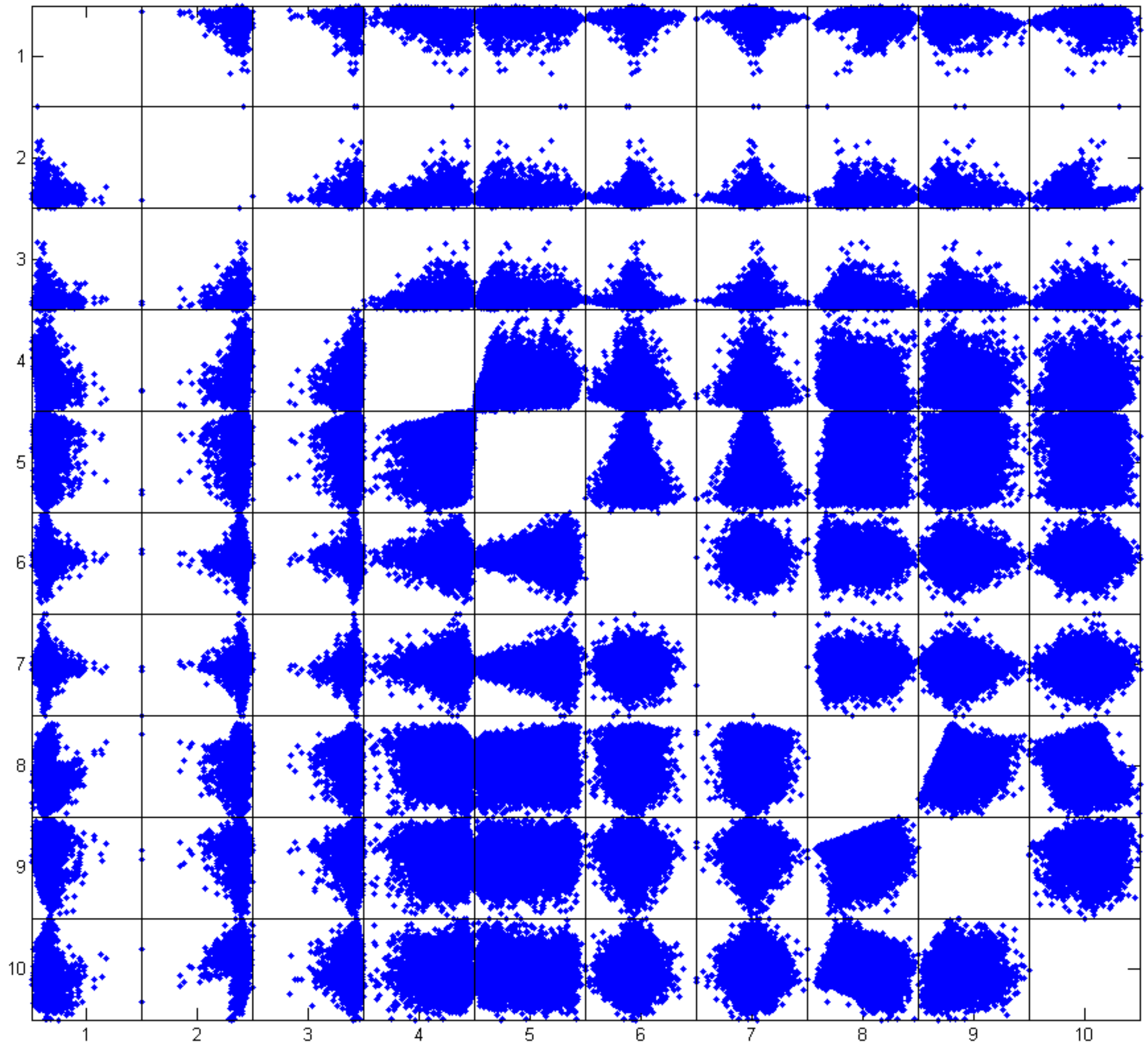


Time delay = 3 days

$I(Q_{sim};Q_{obs})$ vs.

$I(P_{obs}(t-1), P_{obs}(t-2), P_{obs}(t-3), PET_{obs}(t-1), PET_{obs}(t-2), PET_{obs}(t-3), Q_{obs}(t-1), Q_{obs}(t-2), Q_{obs}(t-3); Q_{obs}(t))$

simulation study



Relationships between different independent components

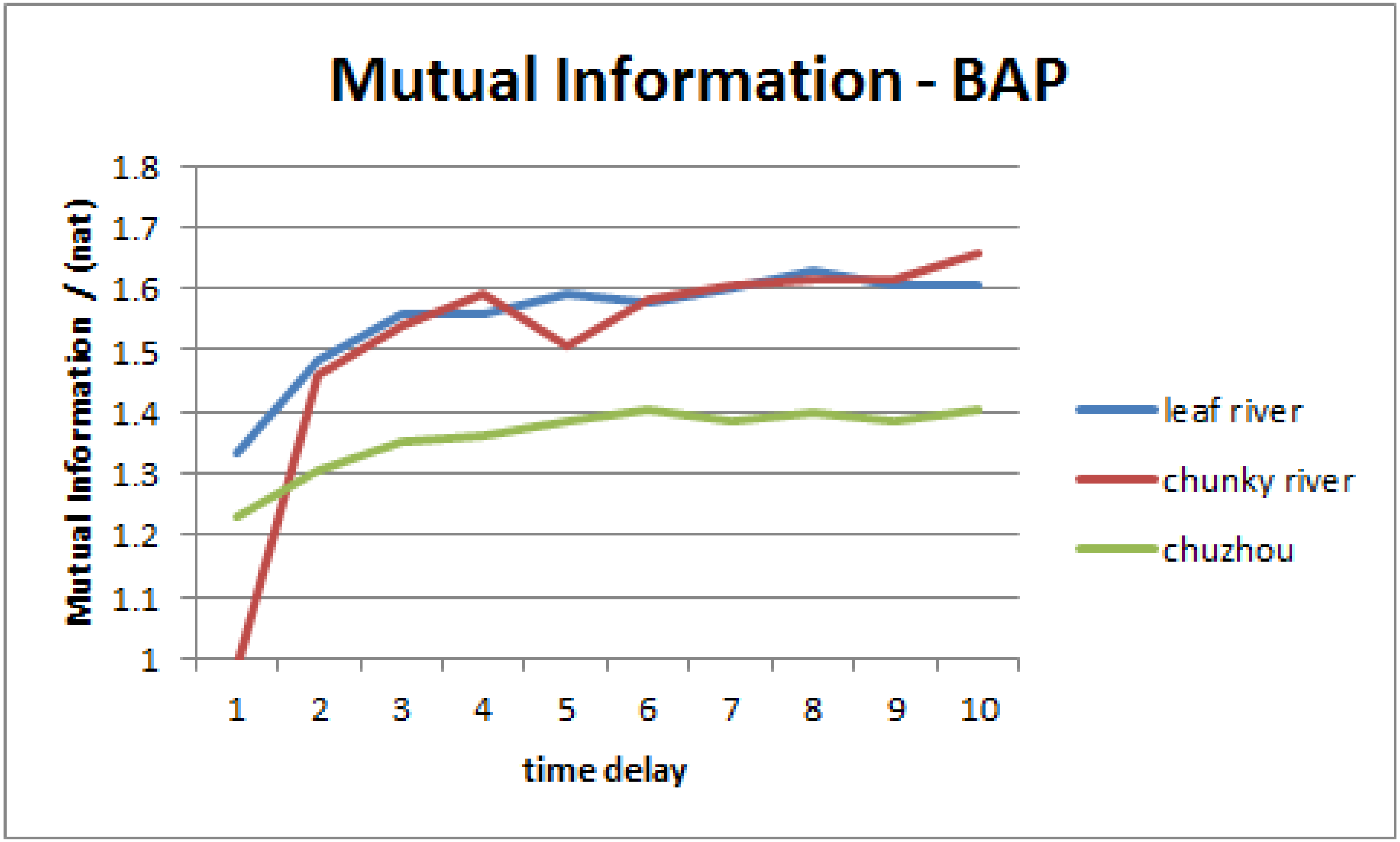
1) They are independent
2) Some interesting nongaussian patterns
3) Outliers

Dirty hack: add a small random error (1e-3mm), because of the zero values of precipitation.

Case study of 3 catchments, 3 models

- Leaf River, 1948~1978, HyMod and SAC
- Chunky River, 1948~1978, HyMod and SAC
- Chuzhou, 1980~2000, HyMod and Xinanjiang Model

Information from data I(Input;Q_{obs})



Case study of 3 catchments, 3 models

	Leaf River	Chunky River	Chuzhou
HyMod $I(Q_{sim};Q_{obs})$	0.7591	0.8921	1.0583
SAC $I(Q_{sim};Q_{obs})$	0.9158	0.9268	-
X3M $I(Q_{sim};Q_{obs})$	-	-	1.2212
BAP $I(Input;Q_{obs})$	1.5956	1.5960	1.3895

BAP is the mean mutual information of 4-10 days timedelay (the stable section of MI)

Gong, W., H. V. Gupta, D. W. Yang, K. Sricharan, and A. O. Hero (2013), Estimating Epistemic & Aleatory Uncertainty During Hydrologic Modeling: An Information Theoretic Approach, *Water Resour. Res.*, *49*(4), 2253–2273, doi:10.1002/wrcr.20161.

Estimating Entropy: 1D case

北京師範大學 Beijing Normal University
全球变化与地球系统科学研究院 College of Global Change and Earth System Science
GCESS

- Difficulties for practical dataset: precipitation and river discharge
  - Zero effect: Many zero values
  - Optimal bin width: balance between bias and variance
  - Measurement effect: heteroscedastic observation error
  - Skewness effect: long tail distribution
- How to fix?
  - 1D entropy equation for Hybrid PDF

$$H(X) = -(1-k_x)\log(1-k_x) - k_x\log(k_x) - k_x\sum_{i=1}^{N_{Bin}} f(x_i)\log\big(f(x_i)\big)\,h_i - \sum_{i=1}^{N_{Bin}} k_x f(x_i)\,h_i\log(h_i)$$

  - Optimal bin width using Scott's equation

$$h^* = 2\times 3^{1/3}\pi^{1/6}\sigma {N_{Data}}^{-1/3} \approx 3.49\sigma {N_{Data}}^{-1/3}$$

  - Box-Cox transformation to fix measurement and skewness effect
  - Extending to high-dimensional cases
- Potential applications
  - Predictability
  - Ensemble forecast pre/post-processor and skill score
  - Uncertainty qualification

Gong, W., D. Yang, H. V. Gupta, and G. Nearing (2014), Estimating information entropy for hydrological data: One-dimensional case, *Water Resour. Res.*, *50*(6), 5003–5018, doi:10.1002/2014wr015874.

Estimating Entropy: high-D cases
Considering zero values

北京师范大学
Beijing Normal University
全球变化与地球系统科学研究院
GCESS
College of Global Change and Earth System Science

3D entropy

For $X_1 \geq 0, X_2 \geq 0, X_3 \geq 0$

k_{111}: proportion of $X_1 > 0, X_2 > 0, X_3 > 0$

k_{011}: proportion of $X_1 = 0, X_2 > 0, X_3 > 0$

k_{001}: proportion of $X_1 = 0, X_2 = 0, X_3 > 0$

k_{101}: proportion of $X_1 > 0, X_2 = 0, X_3 > 0$

k_{110}: proportion of $X_1 > 0, X_2 > 0, X_3 = 0$

k_{010}: proportion of $X_1 = 0, X_2 > 0, X_3 = 0$

k_{000}: proportion of $X_1 = 0, X_2 = 0, X_3 = 0$

k_{100}: proportion of $X_1 > 0, X_2 = 0, X_3 = 0$

$$
\begin{aligned}
-H(X_1, X_2, X_3) = {} & k_{111}\log k_{111} + k_{111} \iiint f(x_1, x_2, x_3) \log[f(x_1, x_2, x_3)]\, dx_1 dx_2 dx_3 + k_{111} \log(h_1 h_2 h_3) \\
& + k_{011}\log k_{011} + k_{011} \iint f(x_2, x_3) \log[f(x_2, x_3)]\, dx_2 dx_3 + k_{011} \log(h_2 h_3) \\
& + k_{001}\log k_{001} + k_{001} \int f(x_3) \log[f(x_3)]\, dx_3 + k_{001} \log(h_3) \\
& + k_{101}\log k_{101} + k_{101} \iint f(x_1, x_3) \log[f(x_1, x_3)]\, dx_1 dx_3 + k_{101} \log(h_1 h_3) \\
& + k_{110}\log k_{110} + k_{110} \iint f(x_1, x_2) \log[f(x_1, x_2)]\, dx_1 dx_2 + k_{110} \log(h_1 h_2) \\
& + k_{010}\log k_{010} + k_{010} \int f(x_2) \log[f(x_2)]\, dx_2 + k_{010} \log(h_2) \\
& + k_{000}\log k_{000} \\
& + k_{100}\log k_{100} + k_{100} \int f(x_1) \log[f(x_1)]\, dx_1 + k_{100} \log(h_1)
\end{aligned}
$$

Try to remove the dirty hack, but failed.
Try to deal with the nonlinear case,
not successful yet!

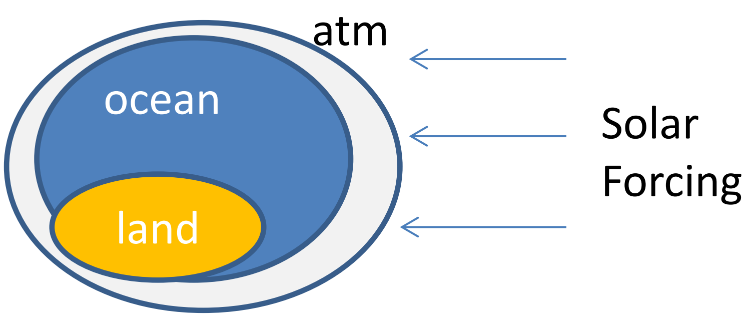
Discussion 1: data processing inequality

$$I\left(Q_{obs};Data\right) \geq I\left(Q_{obs};Q_{sim}\right)$$

- All information from (forcing) data

  No information provided by model

- Earth system model "paradox":
  - Almost constant forcing: solar radiation
  - A lot of information: atm, land, ocean, ice
- Why?



Discussion 1: data processing inequality

$$I\left(Q_{obs}; ForcingData, ModelInfo\right) \geq I\left(Q_{obs}; Q_{sim}\right)$$

- All information from:

Forcing data + information from model

Simple Hydro model → Modern Hydro Model → Land Surface Model → Weather Forecast Model → Earth System Model

More complex modeling structure →

More information from model structure!

Question: How to quantify the information from model?

Discussion2: Information and toy model

北京師範大學
Beijing Normal University
GCESS
全球变化与地球系统科学研究院
College of Global Change and Earth System Science

- Toy models in hydrology
  - Tank, sixpar, hymod, …
- Toy models in atmosphere/climate
  - Lorenz95 (the 3D Lorenz is not good)
- Toy models for fun:
  - The three bodies problem
- Question:

Let’s play with the toy models!

A toy model library.

Discussion3: Information in data

- Estimating entropy for 1D case:
- Estimating entropy and information for highD case:
  - Manifold learning: can identify nonlinearity, but biased !
  - ICA: strong linear assumption, how to extend to nonlinear?
  - Other possible methods?
- Copula function and copula entropy
  - Becoming more and more popular in hydrology
  - Good for long tail distribution: precipitation, etc.
  - Good for flood/drought analysis
  - Ensemble forecast/multi-model averaging

Discussion4: information based performance metrics

- A growing demand of performance metrics!
- RMSE of observables: only a snapshot
- How to measure the consistence of dynamic processes?
- How to deal with a huge amount of variables, and get the key dynamic features?
- Information theory based metrics may have a great opportunity to do so!

Discussion4: information in network?

北京师范大学 Beijing Normal University
全球变化与地球系统科学研究院 College of Global Change and Earth System Science

- Big BOSS: Earth System Model
- How to archiving good performance with the correct way?
- ESMs have many outputs, constrain some of them may degrade others.
- Simultaneously optimize all outputs? Too many!
- New objective functions:

From ERROR of observables to CONSISTENCY of networks

- Question:

How to quantify the consistency of networks?

Or, the similarity of networks?

Discussion5: information and optimization

- Equifinality: due to not enough information?
- The information maybe enough:
  - Evidence 1: data driven models are better than dynamic hydrological models
  - Evidence 2: quantify the information in data
- So why equifinality?
  - Maybe due to the inconsistency of model structure
  - In optimization, model is used as a regression function

  Multiple regression curves fit the data

  But not all of them are physically correct

Joke ^-^

What's the most significant improvement of hydrological models in the recent 30 years?

Understanding more about hydrologic processes without improving the model performance!

Stop joking

- Information provider:
  - More complex model structure and more input data in the recent 30 years.
  - Distributed hydrological model
- Information consumer
  - More output variables rather than streamflow
  - Higher temporal/spatial resolution
- Question:

Performance not improved because there are more “information consumer”?

Thank You!

Comments and questions