University of Stuttgart
Germany

SimTech

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_b p(x)$$

$$h(X) = -\int_S f(x) \log_b f(x) \, dx$$

UNITE

EGU 2023
Supplemental Material

Manuel Álvarez Chaves

SimTech
Junior Research Group for Statistical Model-Data Integration

25.04.2023

Table of Contents

Part 1

Table of Contents

TOC

SimTech

$$H(X) \qquad H(X,Y) \qquad D_{KL}(p||q) \qquad I(X;Y)$$

(...) establishing of such applications is not a trivial matter of translating words to a new domain, but rather the slow tedious process of hypothesis and experimental verification.

Shannon (1956)

UNITE toolbox v0.1

SimTech

- Collection of methods to estimate information-theoric quantities from data.
  - binned frequencies, KDE, kNN.
- Inspired by and reliant on NumPy and SciPy.
- Simple and easy to use (with some care).

```
[1]: from scipy import stats

     dist = stats.norm(loc=0.0, scale=0.6577)
     samples = dist.rvs(size=(10_000, 1), random_state=42)
```

```
[2]: from unite_toolbox import knn_estimators

     est_h = knn_estimators.calc_knn_entropy(samples)
     print(f"Est. H = {est_h:.3f} nats")
     print(f"True H = {dist.entropy():.3f} nats")
```

```
Est. H = 1.002 nats
True H = 1.000 nats
```

UNITE toolbox v0.1

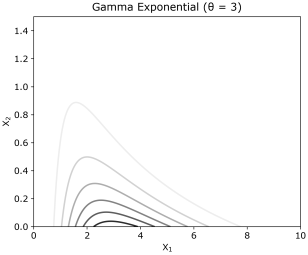
- Collection of methods to estimate information-theoric quantities from data.
  - binned frequencies, KDE, kNN.
- Inspired by and reliant on NumPy and SciPy.
- Simple and easy to use (with some care).

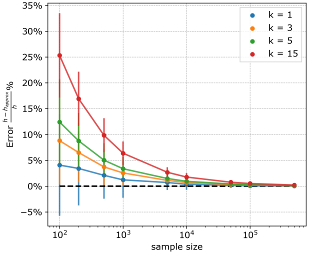
[1]:
```
from scipy import stats

dist = stats.norm(loc=0.0, scale=0.6577)
samples = dist.rvs(size=(10_000, 1), random_state=42)
```

[2]:
```
from unite_toolbox import knn_estimators

est_h = knn_estimators.calc_knn_entropy(samples)
print(f"Est. H = {est_h:.3f} nats")
print(f"True H = {dist.entropy():.3f} nats")
```

```
Est. H = 1.002 nats
True H = 1.000 nats
```

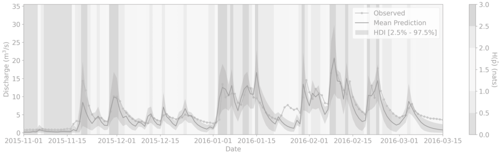
Estimator Assessment

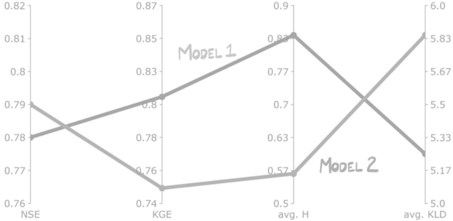




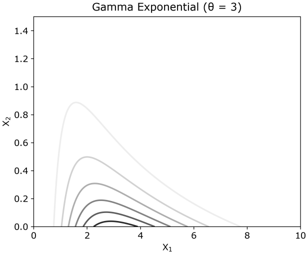
SimTech

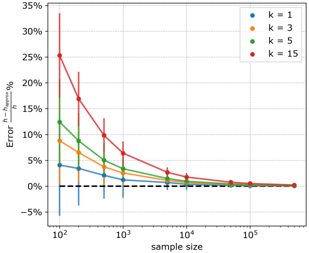
Applications in Model Evaluation



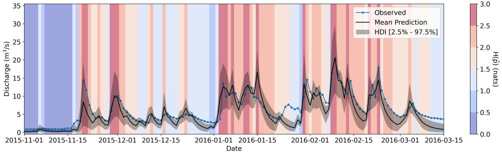


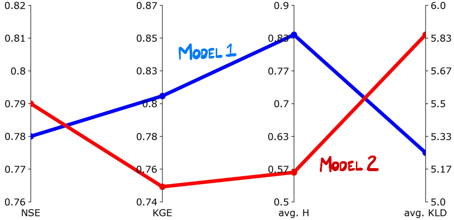
Estimator Assessment





Applications in Model Evaluation





Part 2

Table of Contents

TOC

Data

Attert catchment

Data available from the Catchments as Organized Systems (CAOS) project.

Area = 247.32 km^2

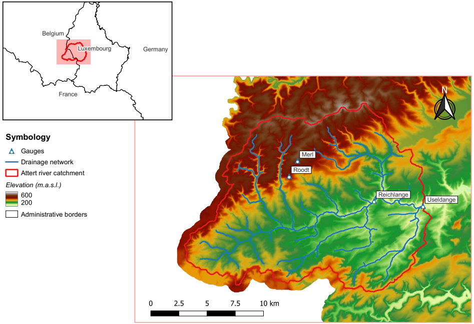
Precipitation inputs from three stations: Roodt (55%), Reichlange (36%) and Useldange (9%).[1]

Temperature (USL), humidity (USL), wind speed (ROD), net radiation (MER).

Discharge measured at Useldange

Periods available:

- Training: 01.11.2012 → 31.10.2015
- Testing: 01.11.2015 → 31.10.2016



[1] Area averaged precipitation when required.

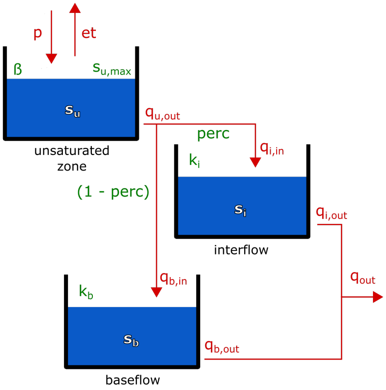
Models

sSHM - (simpler) Simple Hydrological Model

Lumped hydrological model adapted from Ehret et al. (2020).

sSHM was implemented in PyTorch (Paszke et al., 2019).

Three storage components [s_u, s_i, s_b] and five parameters [$s_{u,max}$, β, $perc$, k_i, k_b].



Optimization

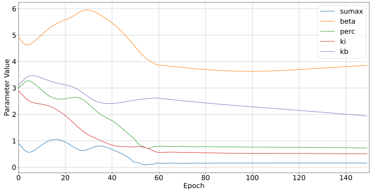
Dimensionality reduction

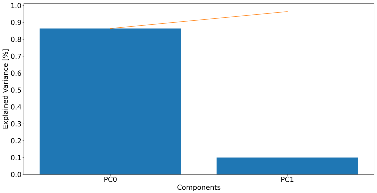
Principal component analysis (PCA) to reduce from parameter space (5-d) to PCA space (2-d).

Loadings: each PC is given by the linear combination of the original variables:

$$PC_1 = w_{11}X_1 + w_{12}X_2 + \cdots + w_{1p}X_p$$

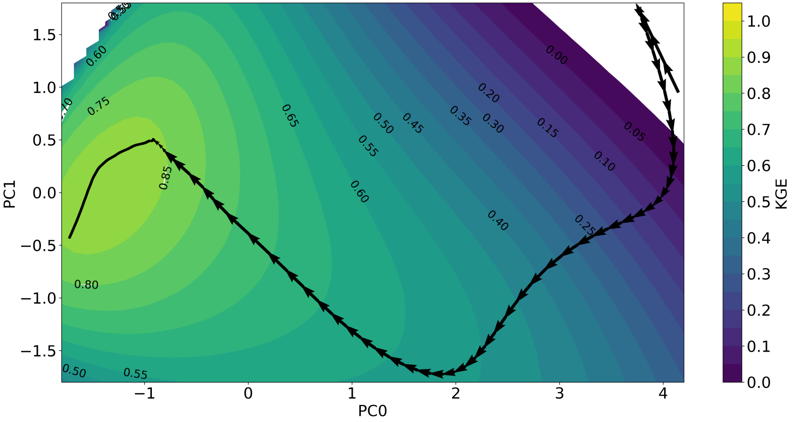
Parameter	PC0	PC1
su_{max}	0.459	-0.319
beta	0.416	-0.674
perc	0.474	0.036
k_i	0.455	0.393
k_b	0.430	0.538





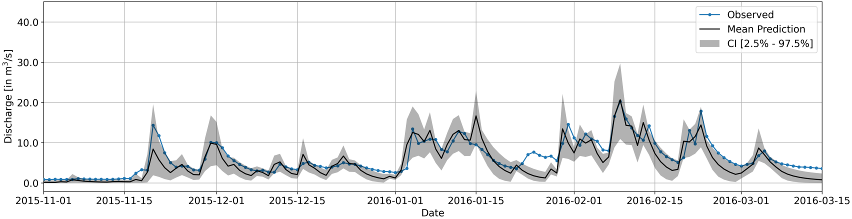
PCA

In the training set:



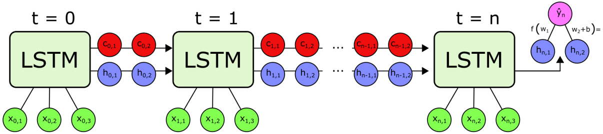
Probabilistic sSHM

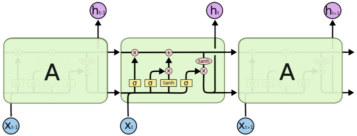
- Equifinality: rejection of an optimal model in favour of multiple possibilities for producing simulations that are acceptable (Beven and Freer, 2001).
- GLUE (generalized likehood uncertainty estimation) $\rightarrow$ ABC (approximate Bayesian computation).



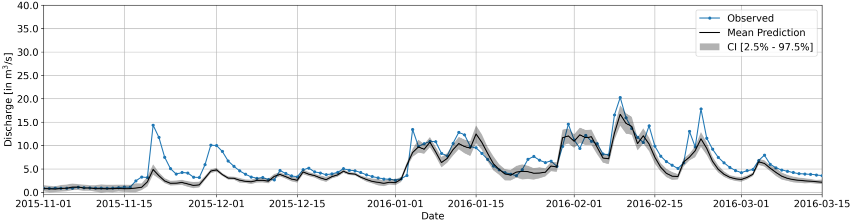
LSTM

- LSTM is a recurrent neural network (RNN) where the cell and hidden states of the network are able to store long time dependencies in the data.
- Monte Carlo dropout for uncertainty (Klotz et al., 2022).
- 64 internal states.





LSTM (cont.)



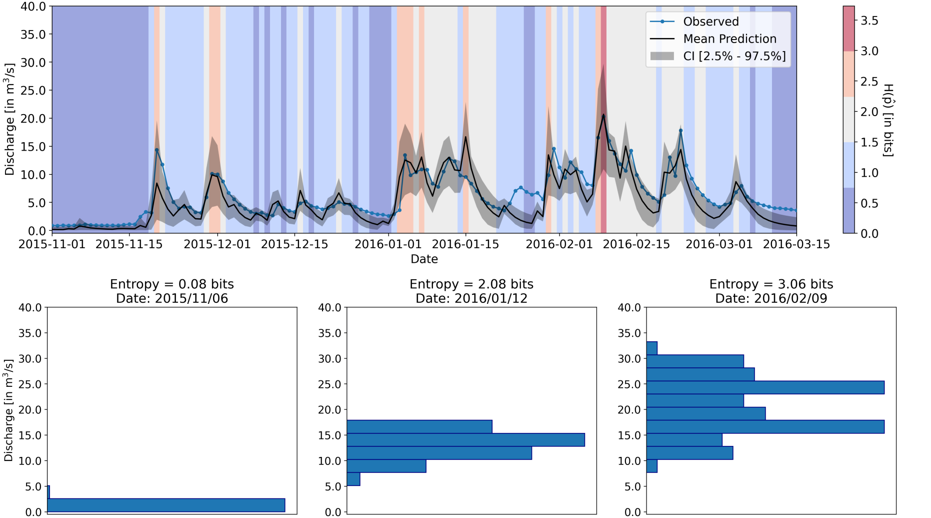
Observed
Mean Prediction
CI [2.5% - 97.5%]
Discharge [in m³/s]
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
2015-11-01
2015-11-15
2015-12-01
2015-12-15
2016-01-01
2016-01-15
2016-02-01
2016-02-15
2016-03-01
2016-03-15
Date

Table of Contents

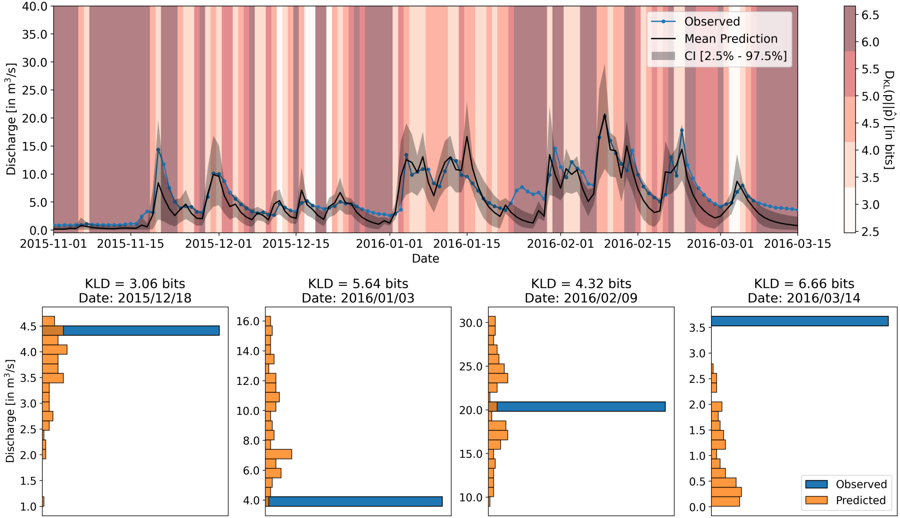
TOC

Evaluation

sSHM - Entropy

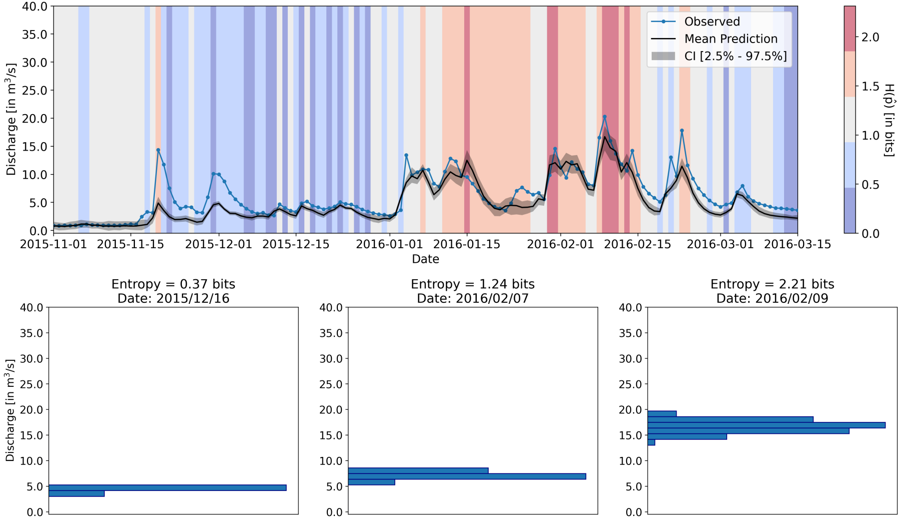


sSHM - KLD

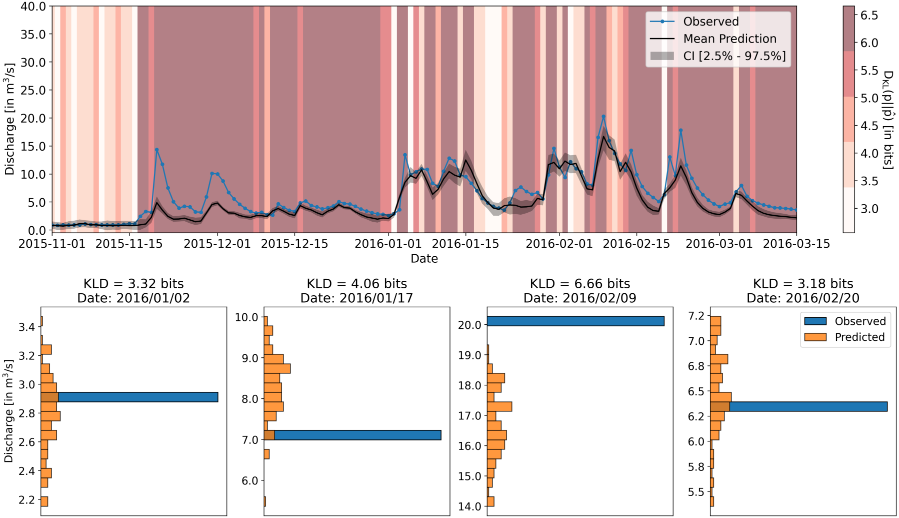


Observed
Mean Prediction
CI [2.5% - 97.5%]
Discharge [in m³/s]
Date
D_KL(p||p̂) [in bits]
2015-11-01
2015-11-15
2015-12-01
2015-12-15
2016-01-01
2016-01-15
2016-02-01
2016-02-15
2016-03-01
2016-03-15
KLD = 3.06 bits
Date: 2015/12/18
KLD = 5.64 bits
Date: 2016/01/03
KLD = 4.32 bits
Date: 2016/02/09
KLD = 6.66 bits
Date: 2016/03/14
Observed
Predicted

LSTM - Entropy



LSTM - KLD



Observed
Mean Prediction
CI [2.5% - 97.5%]
Discharge [in m³/s]
Date
2015-11-01
2015-11-15
2015-12-01
2015-12-15
2016-01-01
2016-01-15
2016-02-01
2016-02-15
2016-03-01
2016-03-15
D_KL(p||p̂) [in bits]
KLD = 3.32 bits
Date: 2016/01/02
KLD = 4.06 bits
Date: 2016/01/17
KLD = 6.66 bits
Date: 2016/02/09
KLD = 3.18 bits
Date: 2016/02/20
Observed
Predicted

Metrics - Summary

sSHM

- NSE = 0.78
- KGE = 0.81
- avg. H = 0.84 bits
- avg. KLD = 5.24 bits

LSTM

- NSE = 0.79
- KGE = 0.75
- avg. H = 0.57 bits
- avg. KLD = 5.87 bits

Metrics - Summary (cont.)

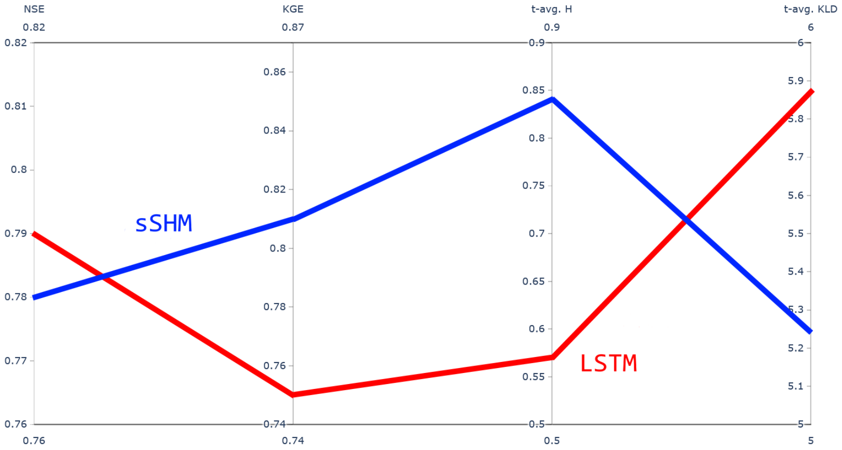


Table of Contents

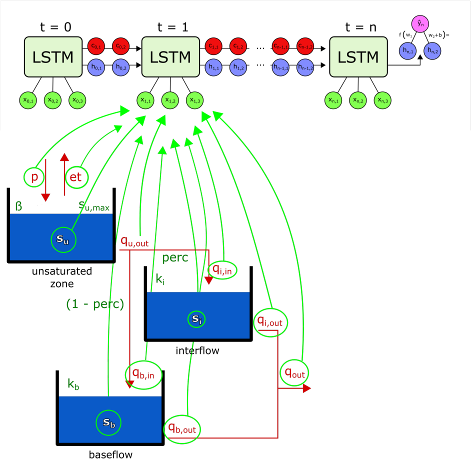
SimTech

TOC

Hybrids

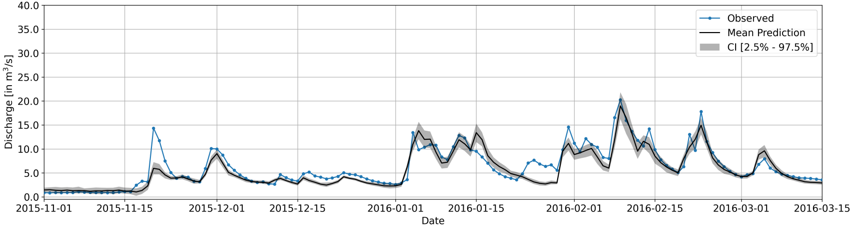
sSHM → LSTM

- Concept adapted from Chabok (2022) and Frame et al. (2021).
- Post-processing of the conceptual model to boost performance.

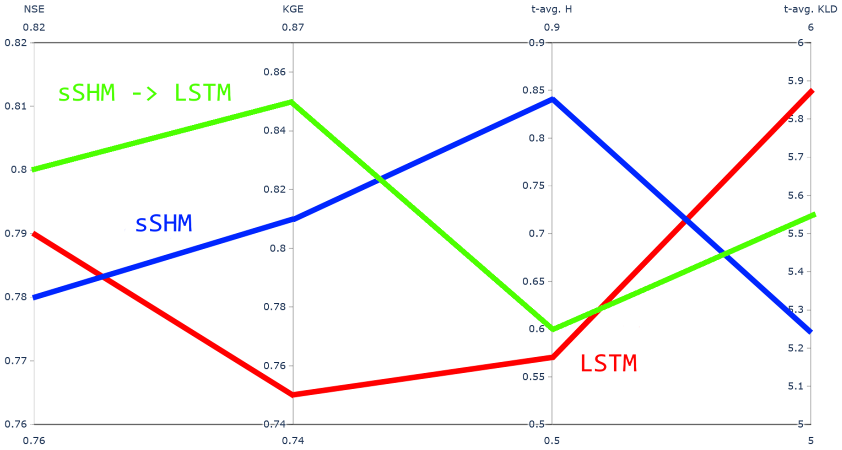


sSHM → LSTM

- NSE = 0.80
- KGE = 0.85
- avg. H = 0.60 bits
- avg. KLD = 5.55 bits

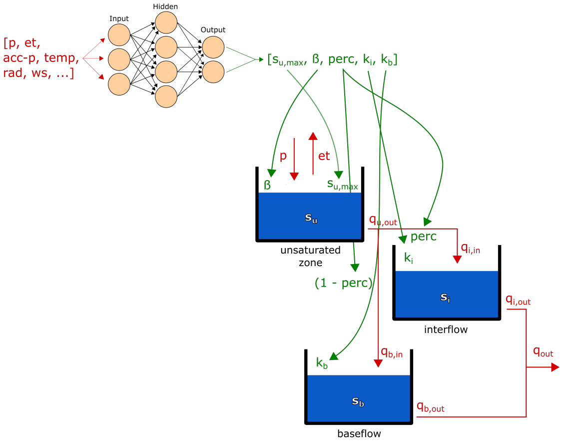


sSHM → LSTM (cont.)



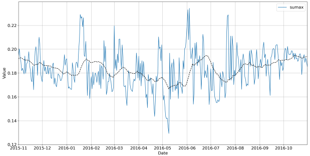
ANN → sSHM

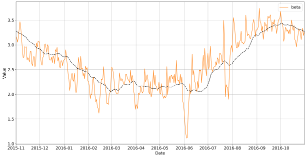
- Predicting sSHM model parameters through a simple artificial neural network (ANN).
- Dynamic model.

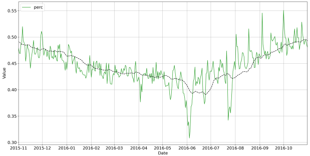


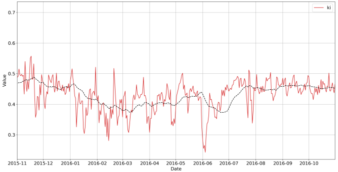
University of Stuttgart 24

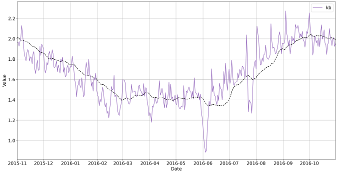
ANN $\to$ sSHM (cont.)



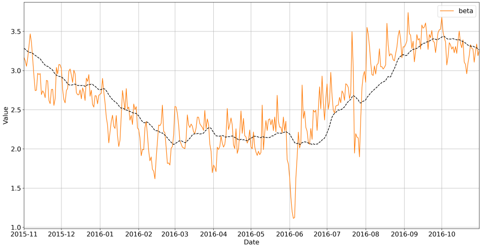


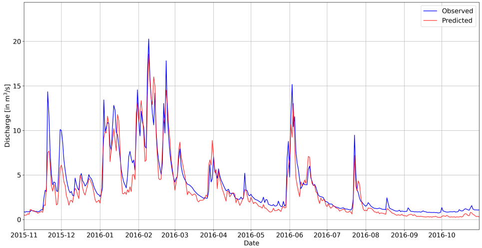






ANN → sSHM (cont.)





Part 3

Table of Contents

TOC

Information Theory

Entropy

$$h(X) = -\int_S f(x)\ \log(x)\, dx$$

Entropy as a measure of variability.

Kullback–Leibler Divergence

$$D(f||g) = \int_G f\ \log\left(\frac{f}{g}\right)$$

KLD as a measure of the predictive capacity.

Mutual Information

$$I(X;Y) = H(Y) - H(Y|X)$$

Reduction in entropy by conditioning.
Mutual dependence between variables.

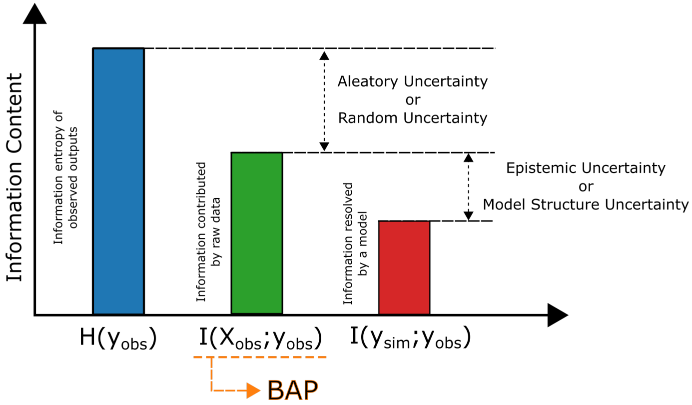
$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

Table of Contents

TOC

Framework

- Initial proposal by Gong et al. (2013).
- BAP: best achievable performance.



Joint Mutual Information

- $I(x_i; Y)$: relevance of a single input on the target.
- $I(X; Y)$: relevance of a set of input features on the target.
- Estimation of high-dimensional mutual information, from samples, using non-parametric density estimators.
- Plug-in estimators: quantization (binning), kernel density estimates, ***k*-nearest neighbours**.

Table of Contents

TOC

*k*NN Estimators

If $p(x)$ is an unknown probability mass function, the *k*NN density of p at x_i is:

$$\hat{p}_k(x_i) = \frac{k}{n-1} \cdot \frac{1}{c_1(d) \cdot \rho_k^d(i)}$$

Where:

- $c_1(d)$: is the volume of a d-dimensional unit ball.
- $\rho_k^d(i)$: is the distance between x_i and its kth nearest neighbour.

kNN Estimators (cont.)

SimTech

Entropy

$$h(X) = -\int_S f(x)\ \log(x)\, dx$$

Kullback–Leibler Divergence

$$D(f||g) = \int_G f\ \log\left(\frac{f}{g}\right)$$

Mutual Information

$$I(X;Y) = \int_{\mathcal{Y}}\int_{\mathcal{X}} P_{(X,Y)}(x,y) \log\left(\frac{P_{(X,Y)}(x,y)}{P_X(x)P_Y(y)}\right) dx\, dy$$

*k*NN Estimators (cont.)

SimTech

Entropy (Kozachenko and Leonenko, 1987)

$$\hat{H}(X) = \psi(N) - \psi(k) + log(c_d) + \frac{d}{N}\sum_{i=1}^{N} \log(\epsilon(i))$$

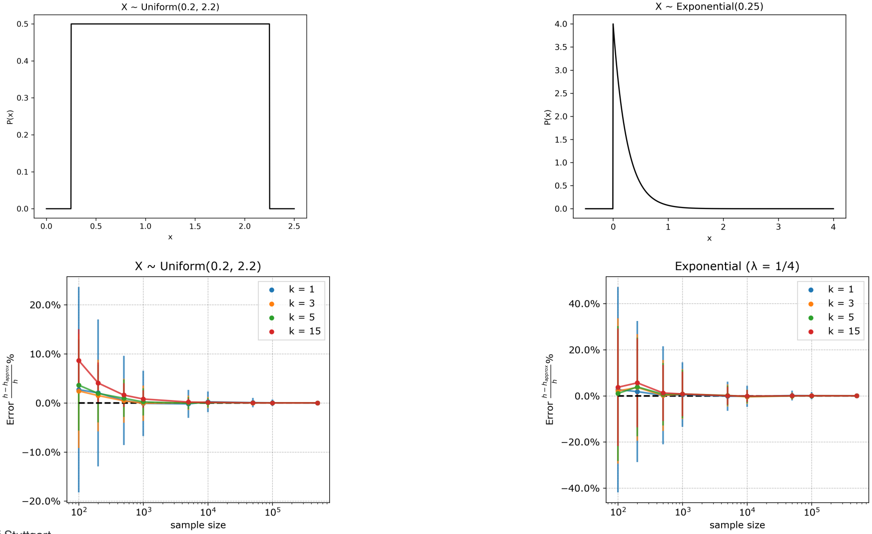
Kullback–Leibler Divergence (Wang et al., 2009)

$$\hat{D}(f||g) = \frac{d}{N}\sum_{i=1}^{n} \log\left(\frac{\nu(i)}{\rho(i)}\right) + \log\left(\frac{M}{N-1}\right)$$

Mutual Information (Kraskov et al., 2004)

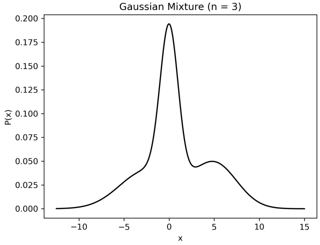
$$\hat{I}(X;Y) = \psi(k) - \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\psi(n_{i,x}) + \psi(n_{i,y})\right] + \psi(N)$$

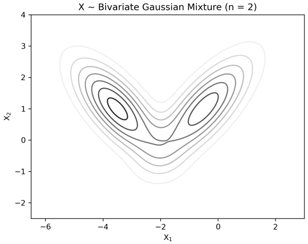
*k*NN Entropy

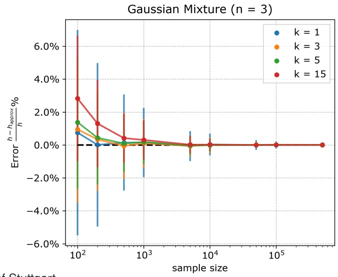


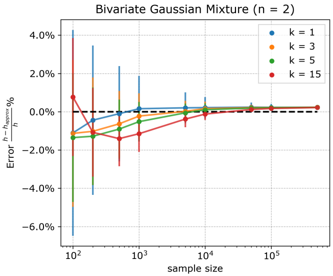
X ~ Uniform(0.2, 2.2)
P(x)
x
X ~ Exponential(0.25)
P(x)
x
X ~ Uniform(0.2, 2.2)
Error $\frac{h - h_{approx}}{h}$%
sample size
k = 1
k = 3
k = 5
k = 15
Exponential (λ = 1/4)
Error $\frac{h - h_{approx}}{h}$%
sample size
k = 1
k = 3
k = 5
k = 15

*k*NN Entropy (cont.)

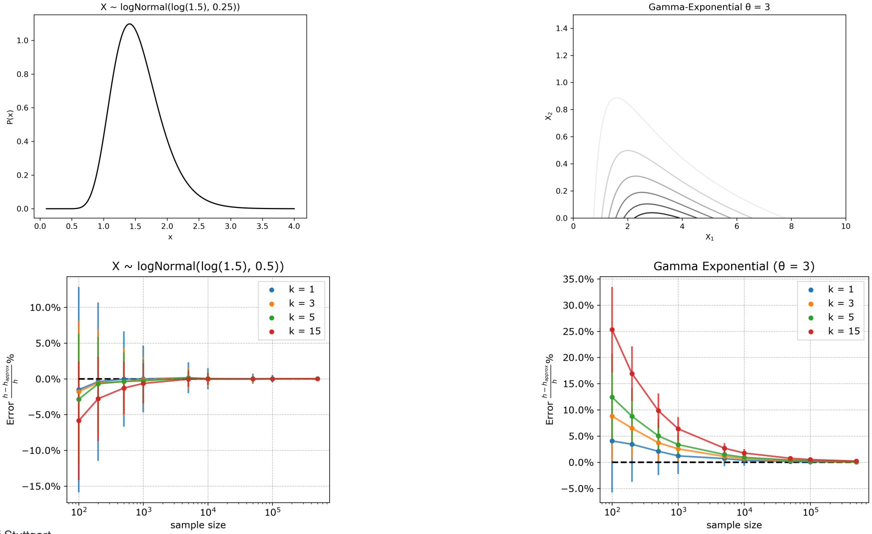




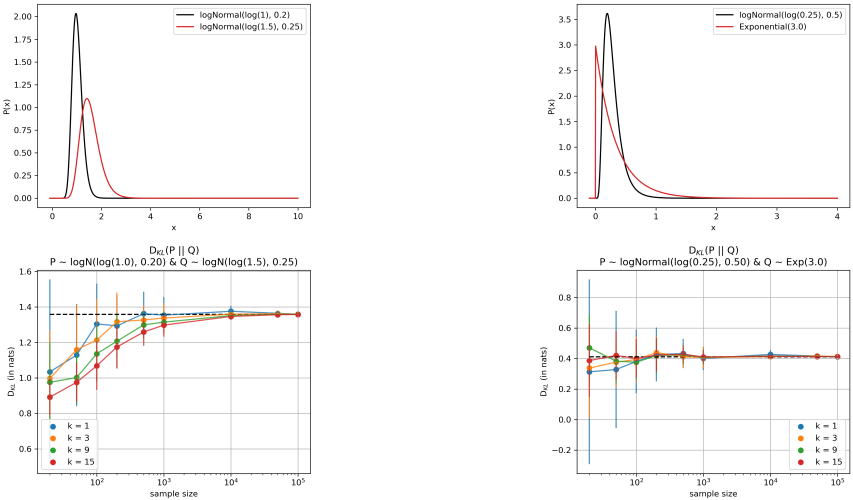




*k*NN Entropy (cont.)

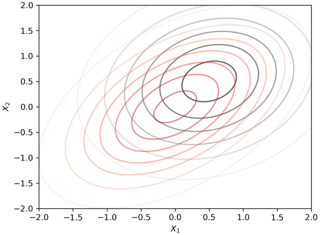


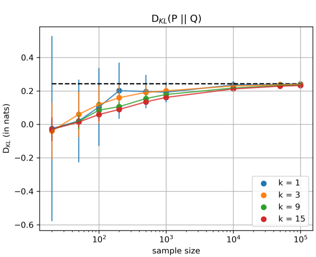
*k*NN KLD

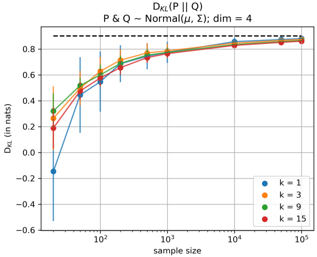


logNormal(log(1), 0.2)
logNormal(log(1.5), 0.25)
P(x)
x
logNormal(log(0.25), 0.5)
Exponential(3.0)
P(x)
x
D_KL(P || Q)
P ~ logN(log(1.0), 0.20) & Q ~ logN(log(1.5), 0.25)
D_KL (in nats)
sample size
k = 1
k = 3
k = 9
k = 15
D_KL(P || Q)
P ~ logNormal(log(0.25), 0.50) & Q ~ Exp(3.0)
D_KL (in nats)
sample size
k = 1
k = 3
k = 9
k = 15

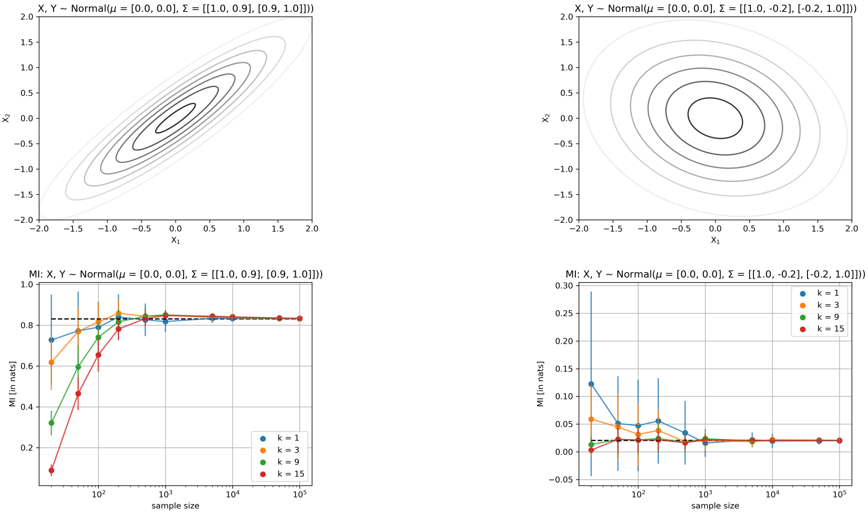
*k*NN KLD (cont.)


X₂
X₁


D_KL(P || Q)
D_KL (in nats)
sample size
k = 1
k = 3
k = 9
k = 15

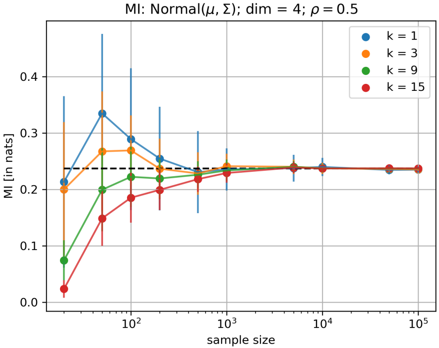

D_KL(P || Q)
P & Q ~ Normal(μ, Σ); dim = 4
D_KL (in nats)
sample size
k = 1
k = 3
k = 9
k = 15

*k*NN MI


X, Y ~ Normal(μ = [0.0, 0.0], Σ = [[1.0, 0.9], [0.9, 1.0]])
X, Y ~ Normal(μ = [0.0, 0.0], Σ = [[1.0, -0.2], [-0.2, 1.0]])
X₁
X₂
MI: X, Y ~ Normal(μ = [0.0, 0.0], Σ = [[1.0, 0.9], [0.9, 1.0]])
MI: X, Y ~ Normal(μ = [0.0, 0.0], Σ = [[1.0, -0.2], [-0.2, 1.0]])
MI [in nats]
sample size
k = 1
k = 3
k = 9
k = 15

*k*NN MI (cont.)

$$I\left(X_1, X_2 \ldots X_{m-1}; X_m\right) = -\frac{1}{2}\log\left(1 - \rho^2\right)$$



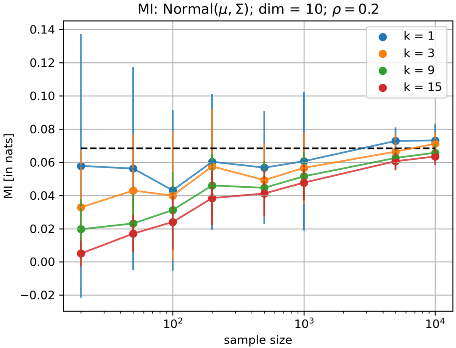


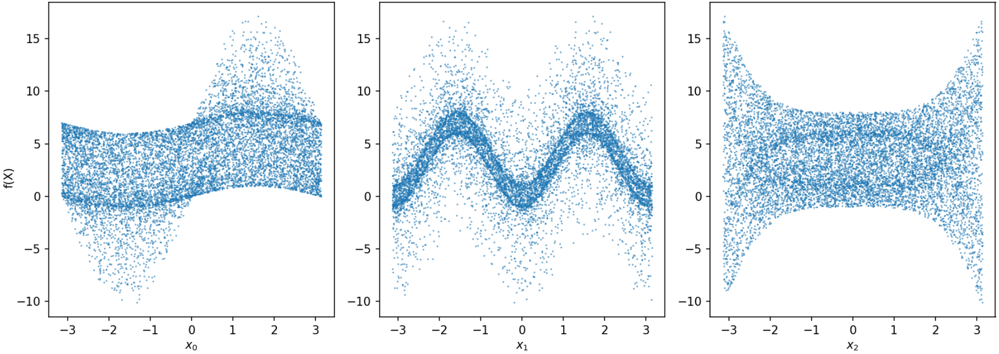
Table of Contents

TOC

kNN MI → Ishigami Function

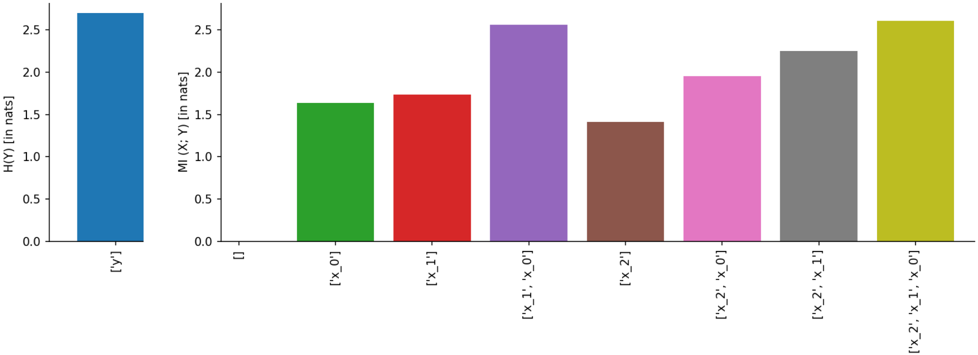
SimTech

$$f(X) = \sin(x_0) + 7 \cdot \sin^2(x_1) + 0.1 \cdot x_2^4 \cdot \sin(x_0)$$

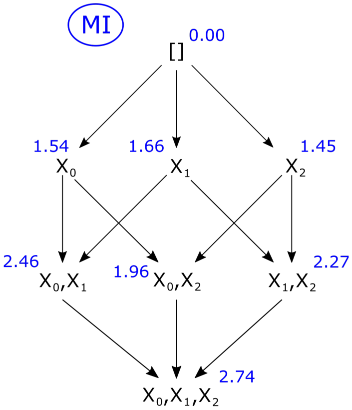


kNN MI → Ishigami Function (cont.)

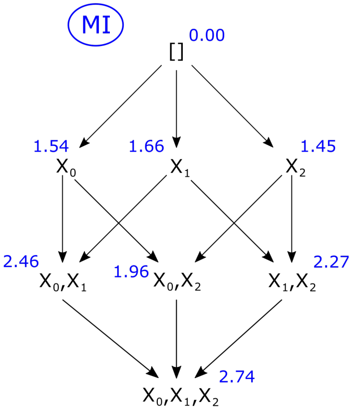
$$f(X) = \sin(x_0) + 7 \cdot \sin^2(x_1) + 0.1 \cdot x_2^4 \cdot \sin(x_0)$$

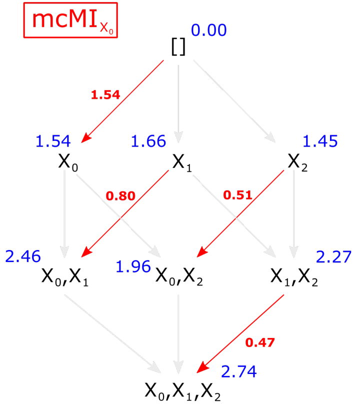


kNN MI → Ishigami Function (cont.)



*k*NN MI → Ishigami Function (cont.)

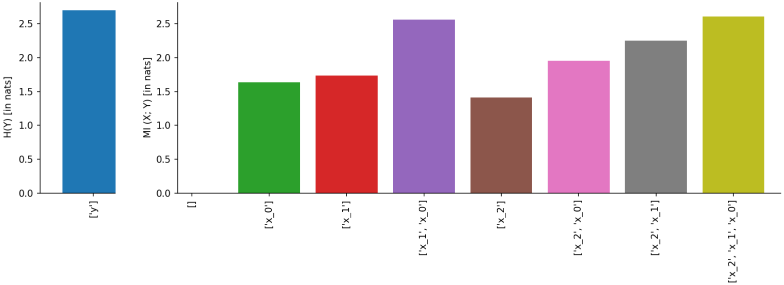


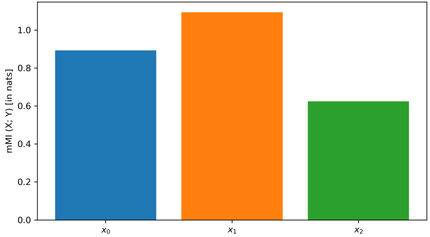


$$mcMI_{X_0} = \tfrac{1}{3}(1.54) + \tfrac{1}{6}(0.80) + \tfrac{1}{6}(0.51) + \tfrac{1}{3}(0.47)$$

$$= \mathbf{0.89}$$

*k*NN MI → Ishigami Function (cont.)

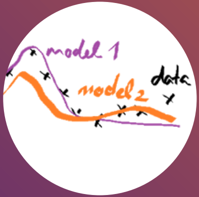




University of Stuttgart
Germany

SimTech

Thank you!



Manuel Álvarez Chaves

manuel.alvarez-chaves@simtech.uni-stuttgart.de

+49 711 685-60130

University of Stuttgart