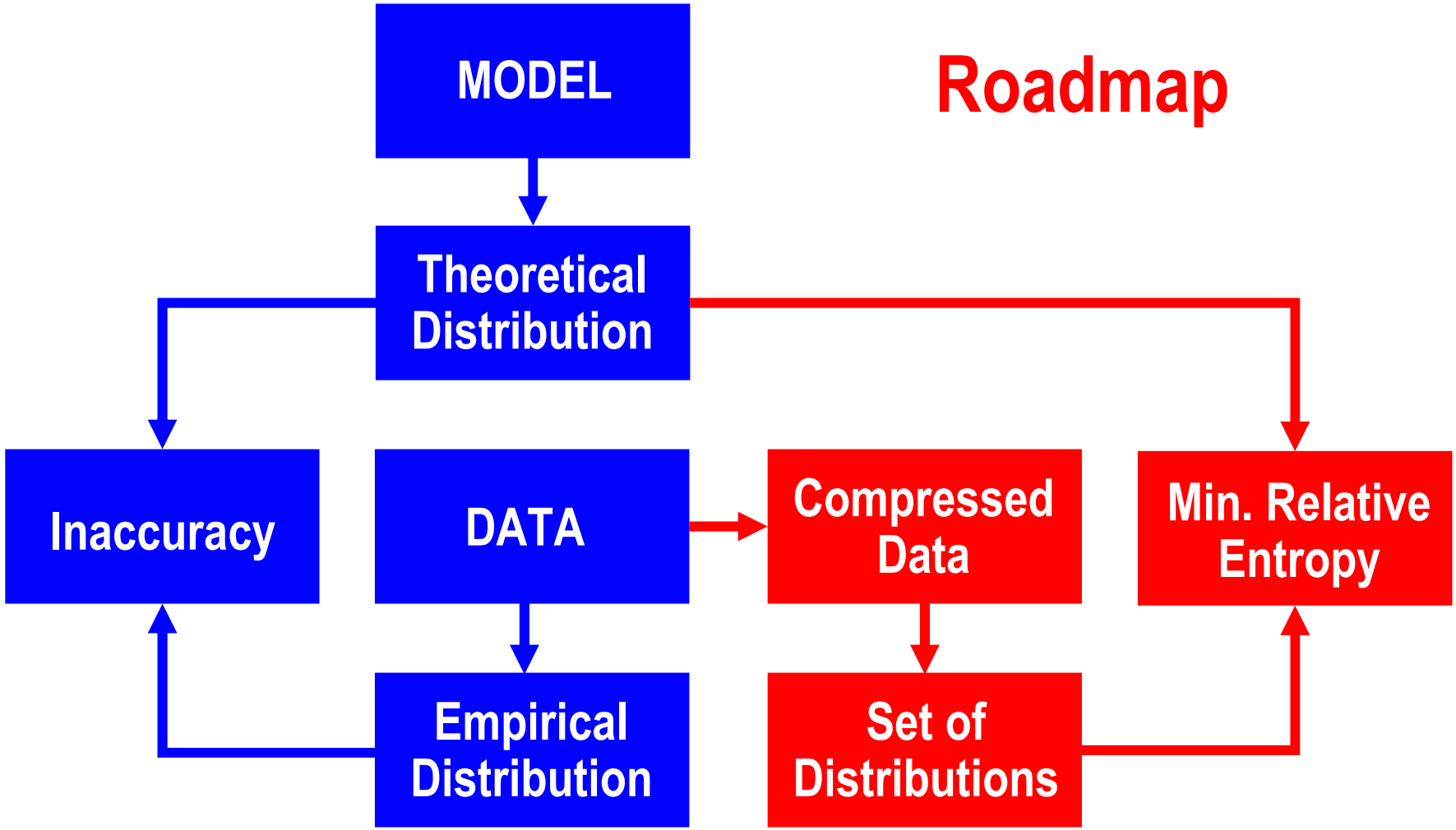
System Identification: From Matching Data to Matching Empirical Probabilities

Rudolf Kulhavý

Honeywell Technology Center &
Institute of Information Theory and Automation
Prague, Czech Republic

A new approach?

Again?!


MODEL
Roadmap
Theoretical Distribution
Inaccuracy
DATA
Compressed Data
Min. Relative Entropy
Empirical Distribution
Set of Distributions

Statistical Estimation Revisited

Information Viewpoint

Random AR(1) Coefficient

- **Model [K.J. Åström]**

$$y_k = (\mu + v_k)\, y_{k-1} + e_k$$

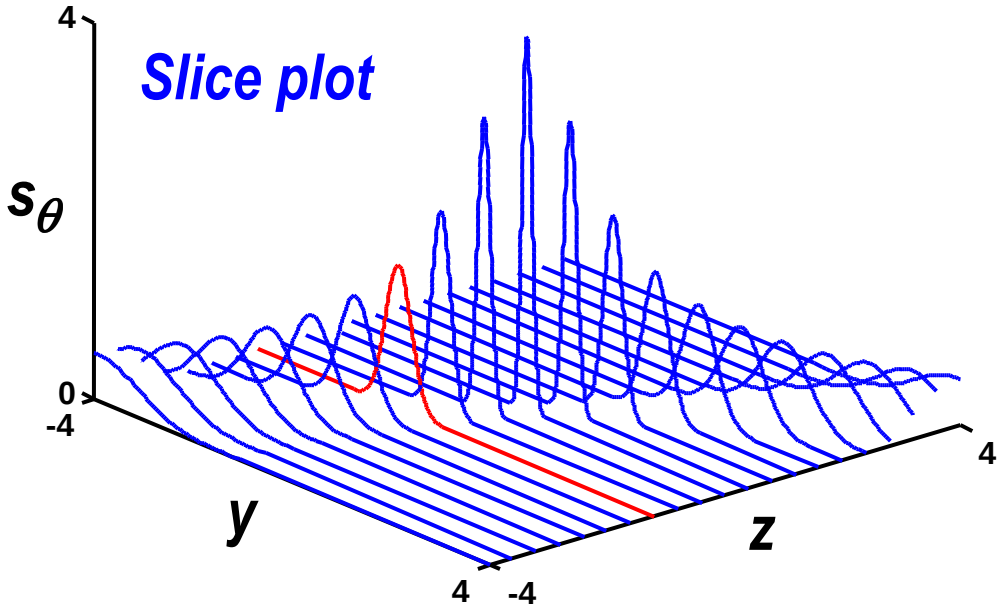
random fluctuation of AR(1) coefficient $= z_k$ *regressor*

- **Assumptions**
  - μ is constant
  - v_k is $N(0, \sigma_v^2)$ distributed
  - e_k is $N(0, \sigma_e^2)$ distributed

unknown parameters

$$\theta = (\mu, \sigma_e, \sigma_v)$$

Model Density

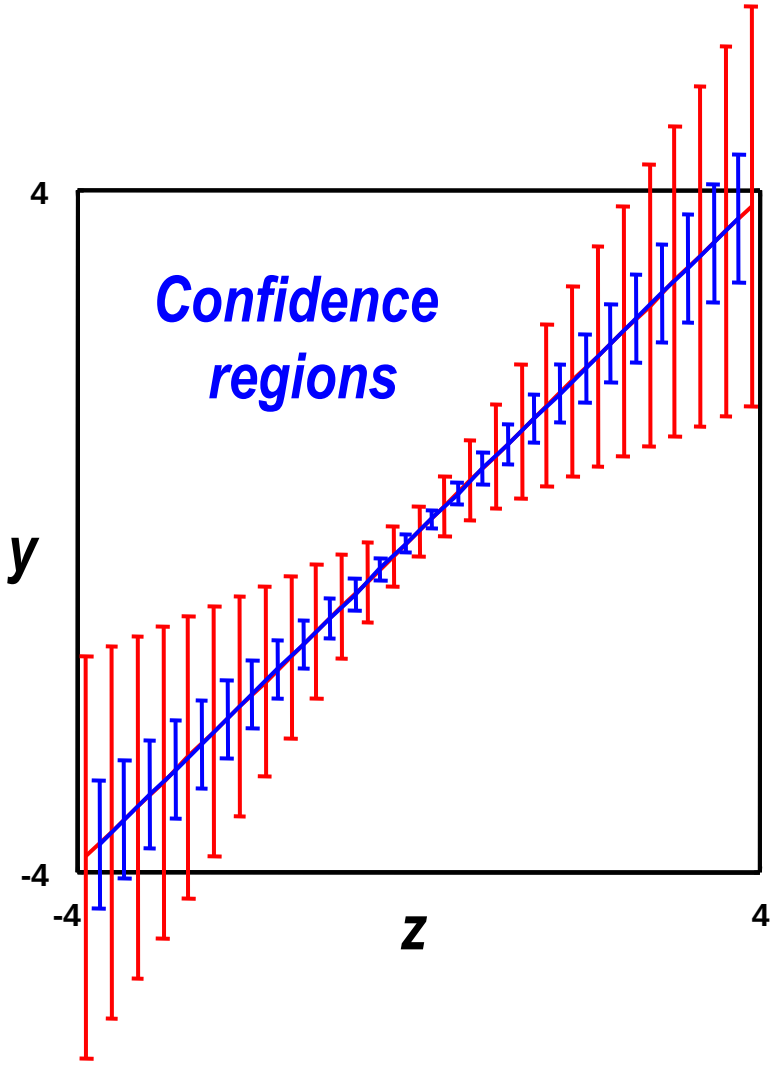


z-dependent variance

$$\sigma^2(z) = \sigma_e^2 + z^2 \sigma_v^2$$

$$s_\theta(y \mid z) = \frac{1}{\sqrt{2\pi\sigma^2(z)}} \exp\left(-\frac{1}{2\sigma^2(z)} (y - \mu z)^2 \right)$$

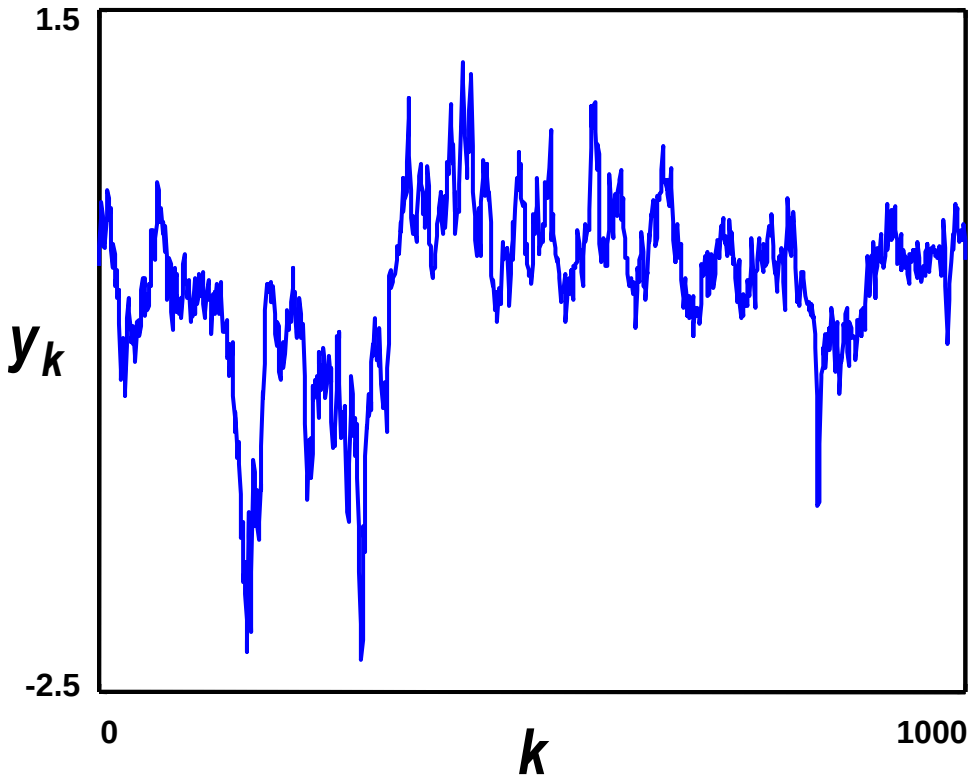
Model Density



$$\mu z \pm \sqrt{\sigma_e^2 + z^2 \sigma_v^2}$$

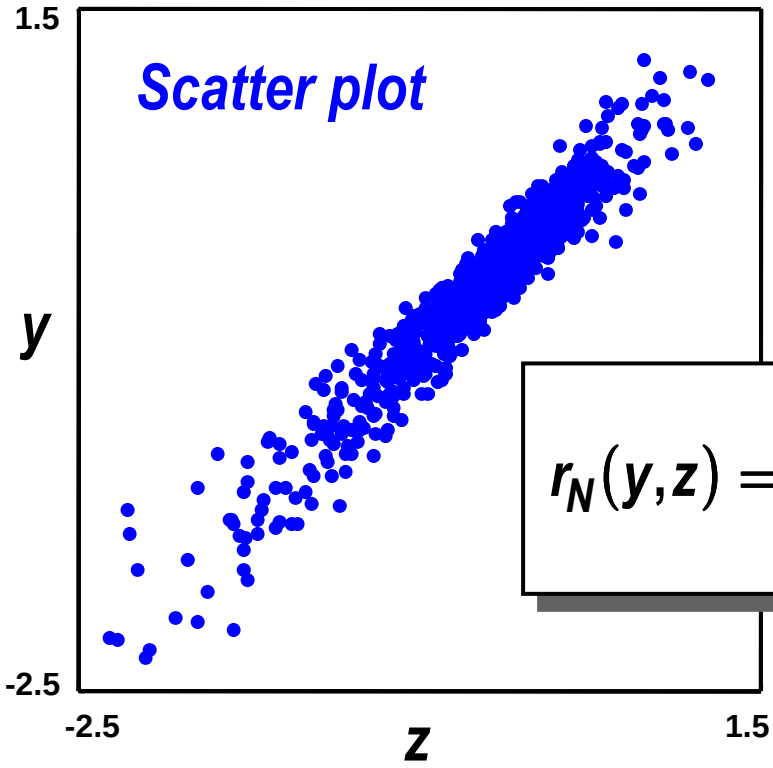
$$\mu z \pm 3 \sqrt{\sigma_e^2 + z^2 \sigma_v^2}$$

Sample of Data



$$\mu = 0.98$$
$$\sigma_e = 0.1$$
$$\sigma_v = 0.2$$

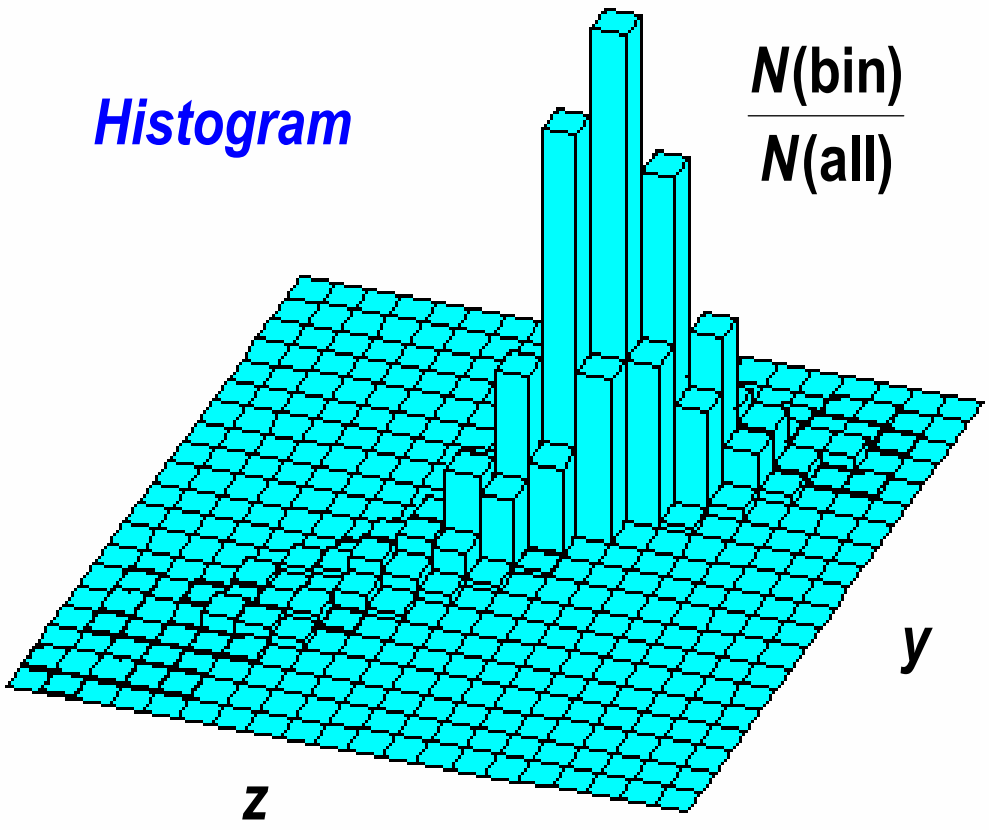
Empirical Density



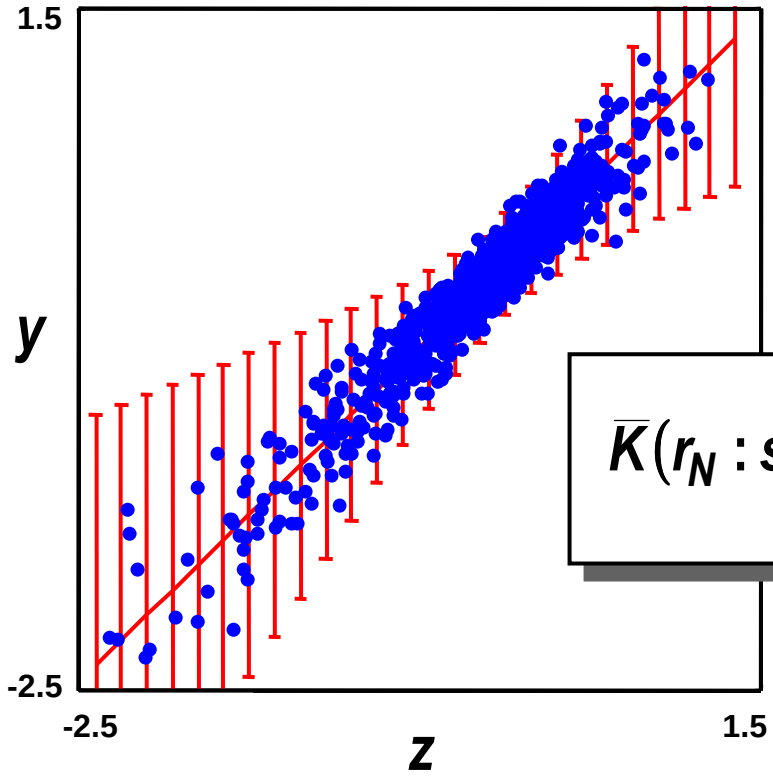
Mixture of
Dirac functions

$$r_N(y,z) = \frac{1}{N} \sum_{k=m+1}^{N+m} \delta(y - y_k, z - z_k)$$

Empirical Density

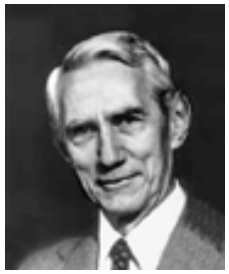


Model-Based Matching of Empirical Probabilities



Conditional inaccuracy

$$\overline{K}(r_N : s_\theta) = \iint r_N(y,z) \log \frac{1}{s_\theta(y \mid z)} \, dy\, dz$$



Inaccuracy

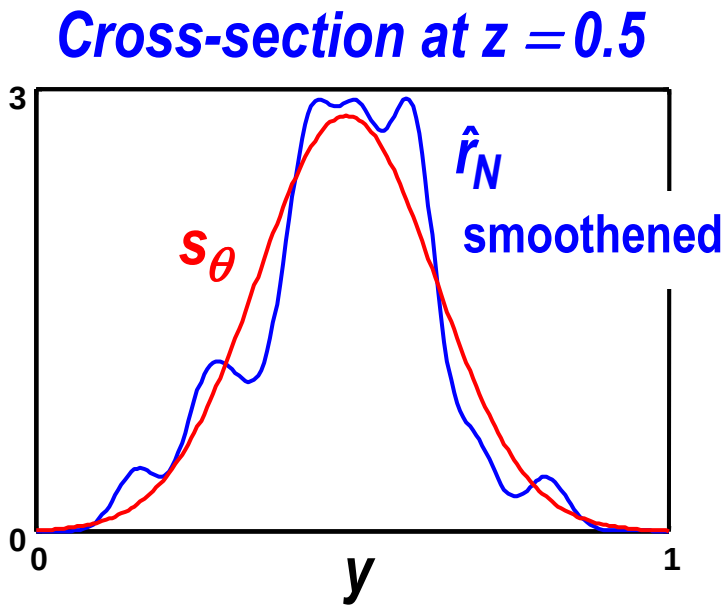
$$\overline{K}(r_N : s_\theta) = -\frac{1}{N} \log \prod_{k=m+1}^{N+m} s_\theta(y_k \mid z_k)$$

Likelihood

$$I_N(\theta) = c \exp\left(-N \overline{K}(r_N : s_\theta)\right)$$

Posterior

$$p_N(\theta) = c\, p_0(\theta) \exp\left(-N \overline{K}(r_N : s_\theta)\right)$$



Measure of Total Uncertainty

$$\bar{K}(r:s) = \bar{H}(r) + \bar{D}(r \,\|\, s)$$

$$\bar{K}(r:s) = \iint r(y,z) \log \frac{1}{s(y \mid z)} \, dy\, dz$$

Inaccuracy

[Kerridge 1963]

$$\bar{H}(r) = \iint r(y,z) \log \frac{1}{r(y \mid z)} \, dy\, dz$$

Entropy

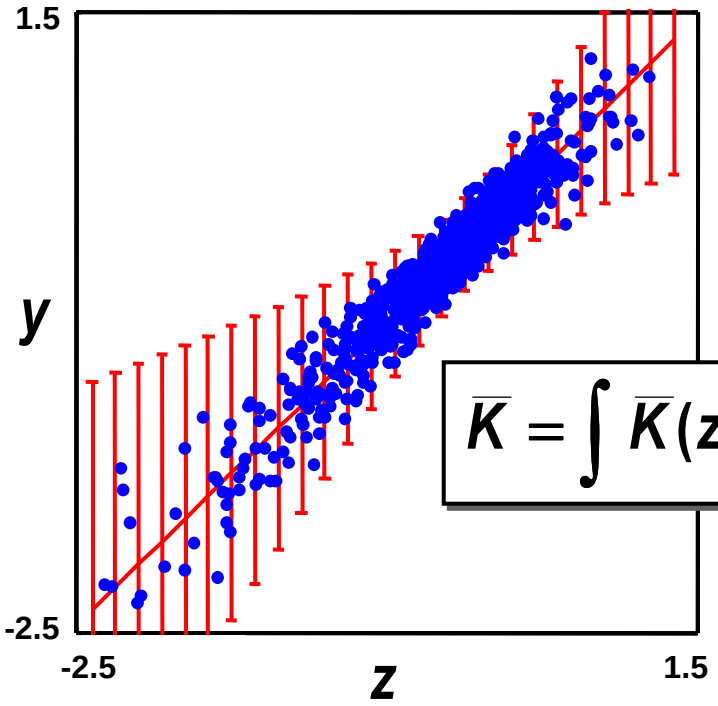
[Shannon 1948]

$$\bar{D}(r \,\|\, s) = \iint r(y,z) \log \frac{r(y \mid z)}{s(y \mid z)} \, dy\, dz$$

Relative entropy

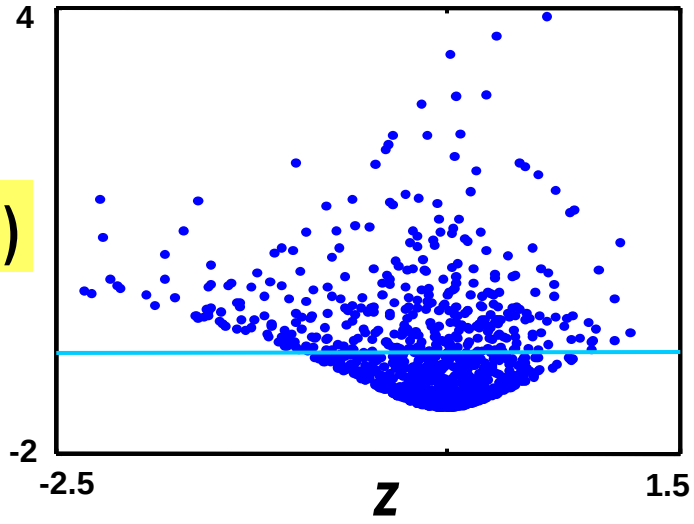
[Kullback and Leibler 1951]

$$\overline{K}(r_N : s_\theta) = \int r_N(z) \int r_N(y \mid z) \log \frac{1}{s_\theta(y \mid z)} \, dy \, dz$$

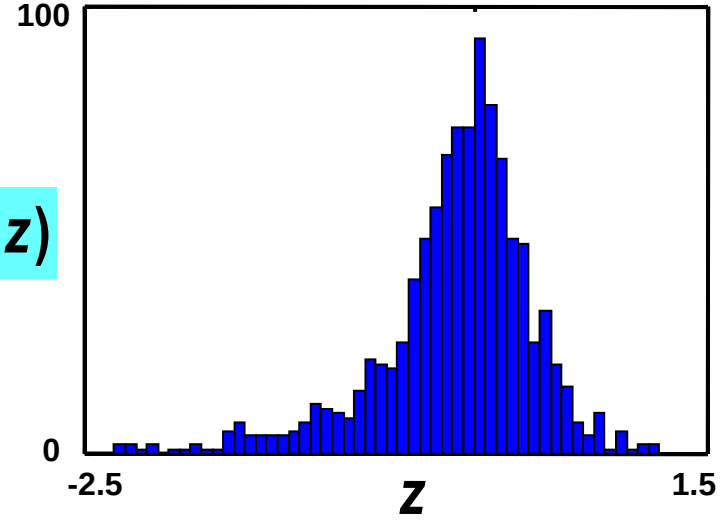


$$\overline{K} = \int \overline{K}(z) \, r_N(z) \, dz$$

$\overline{K}(z)$



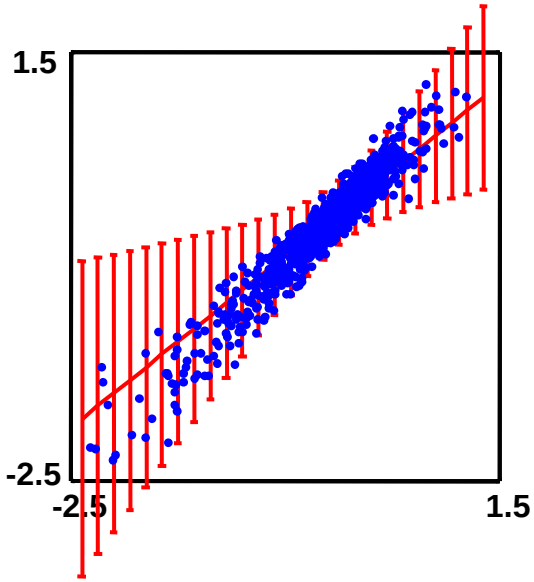
$r_N(z)$



Two-Step Evaluation

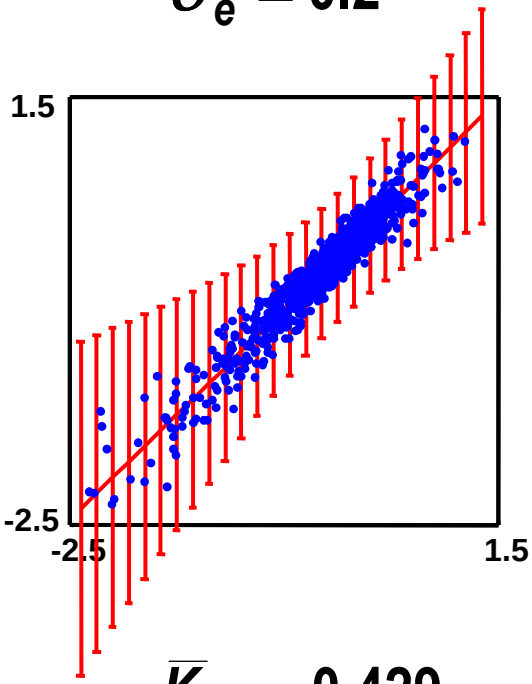
Estimation via Inaccuracy

$\mu = 0.8$



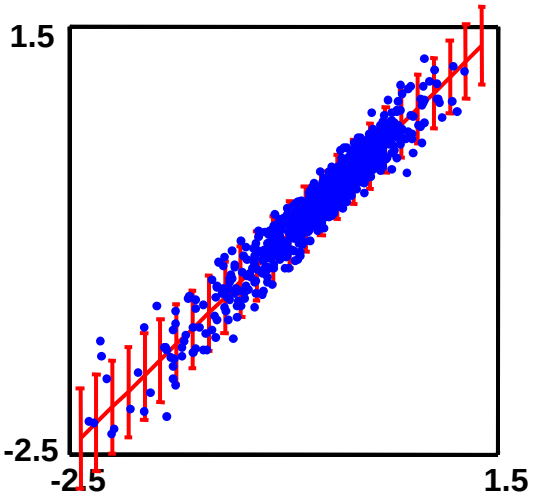
$\overline{K} = -0.552$

$\sigma_e = 0.2$



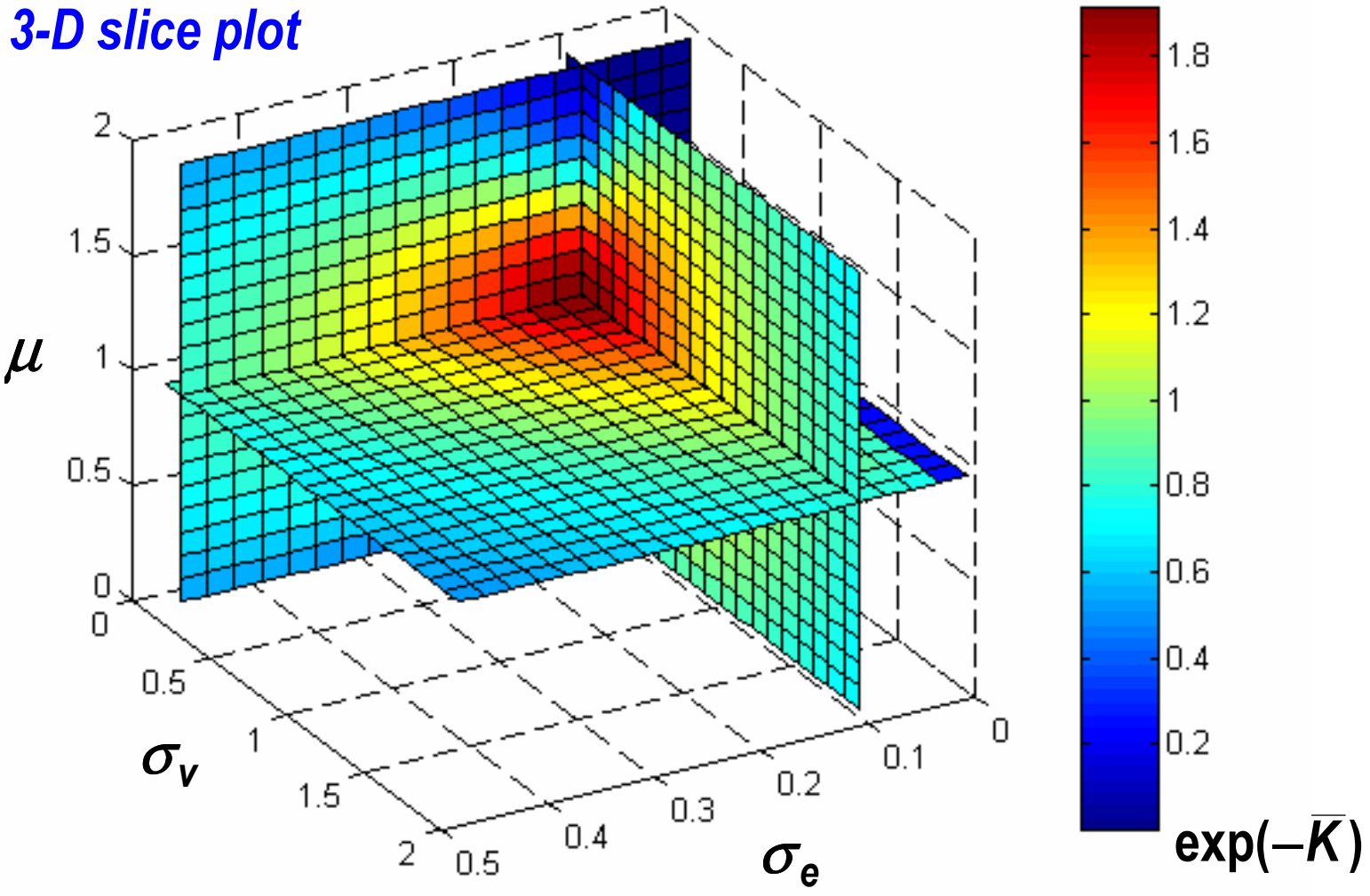
$\overline{K} = -0.429$

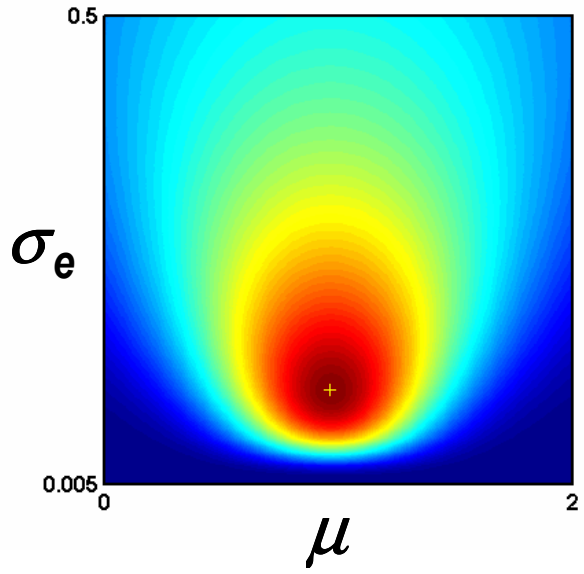
$\sigma_v = 0.05$



$\overline{K} = -0.517$

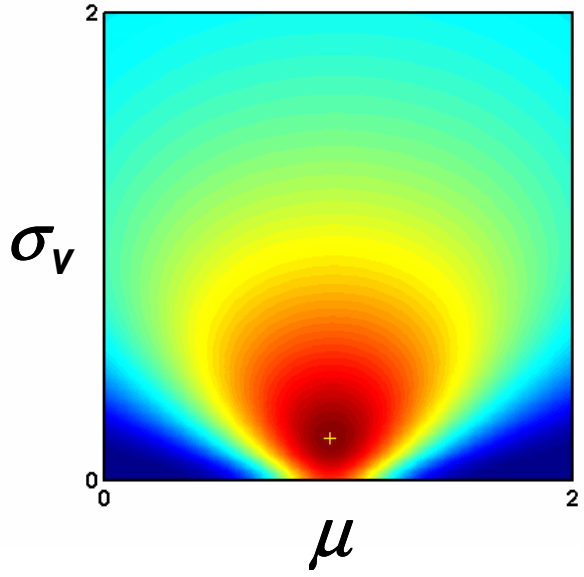
Inaccuracy over Parameter Grid

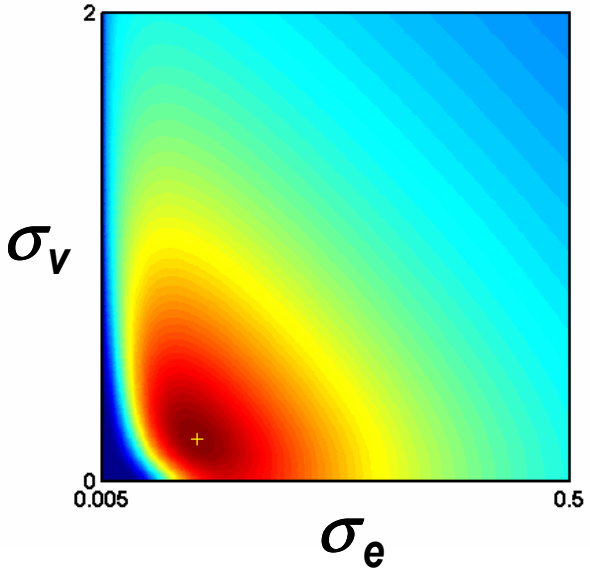




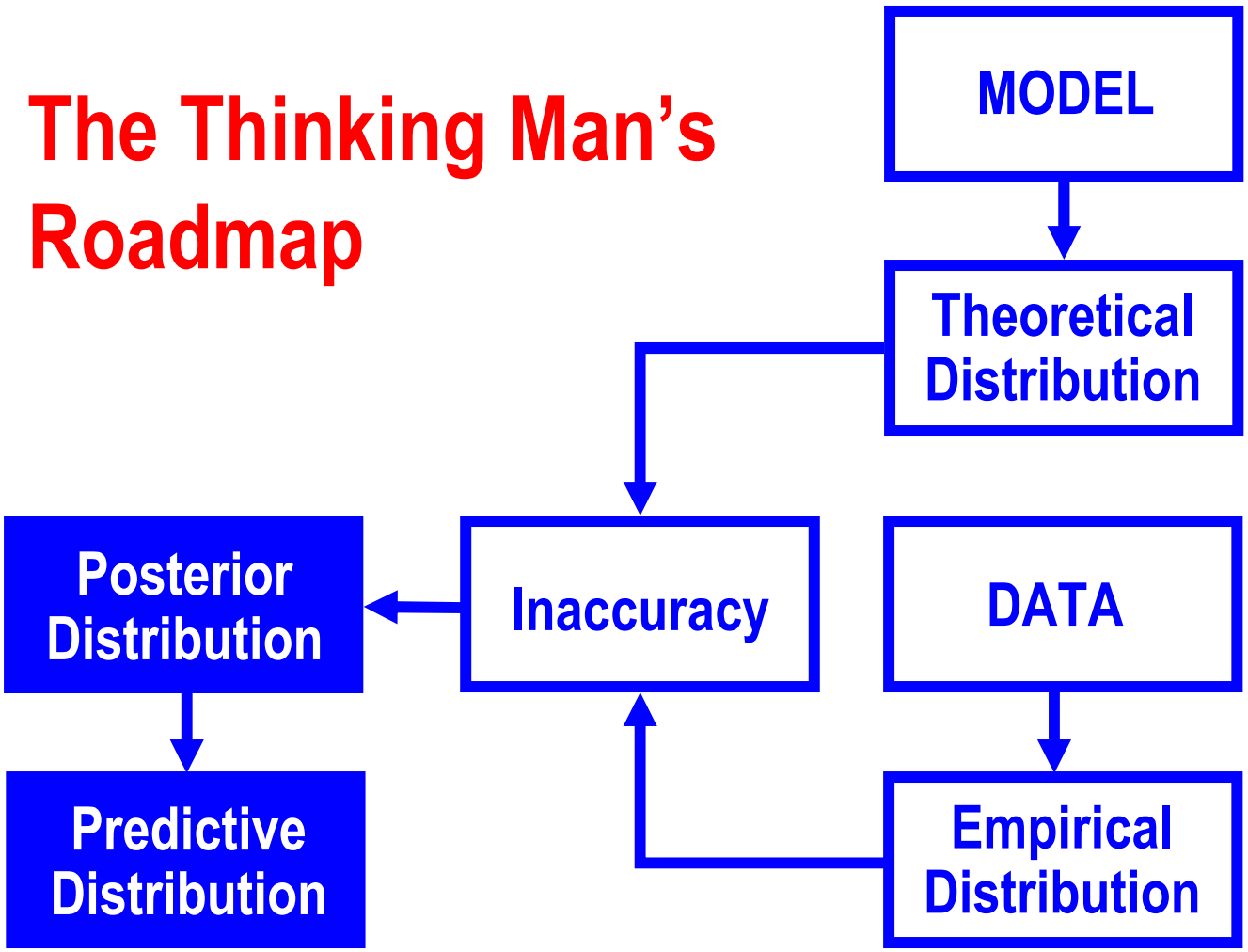
Inaccuracy Interpolated over Parameter Grid

2-D slice matrix

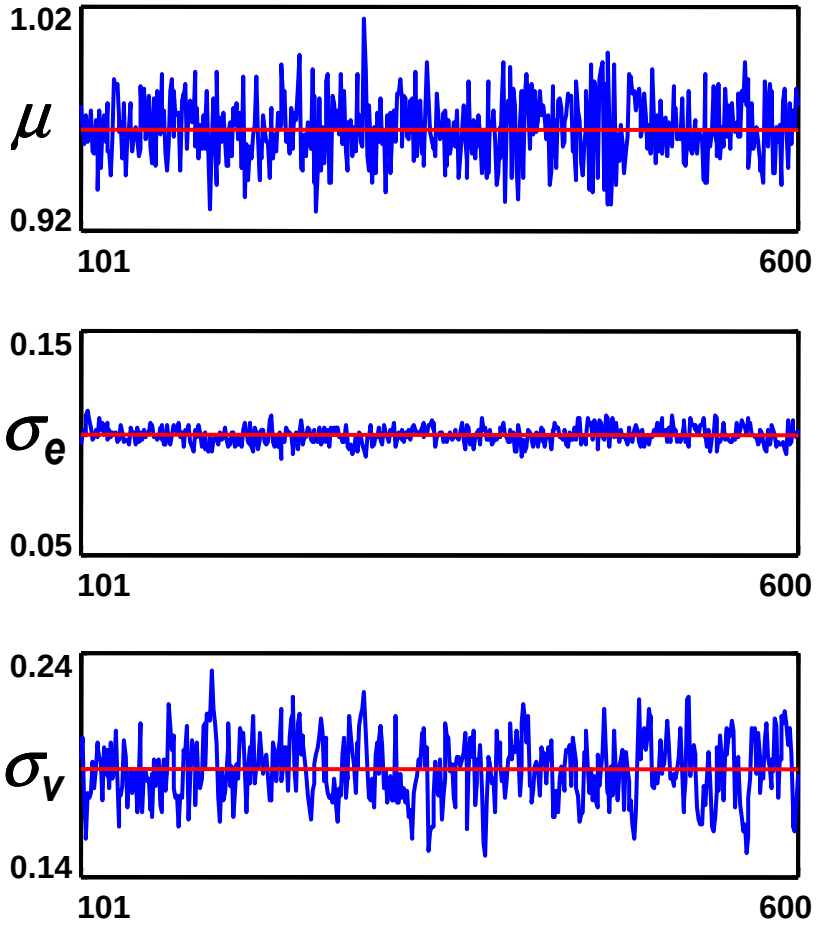




The Thinking Man's Roadmap



Sampling from Posterior Distribution



ALGORITHM

1. Generate a sample

$$\mu^* = \mu^{(i)} + 0.1 * \text{randn}$$

$$\sigma_e^* = \sigma_e^{(i)} * \exp(0.1 * \text{randn})$$

$$\sigma_v^* = \sigma_v^{(i)} * \exp(0.1 * \text{randn})$$

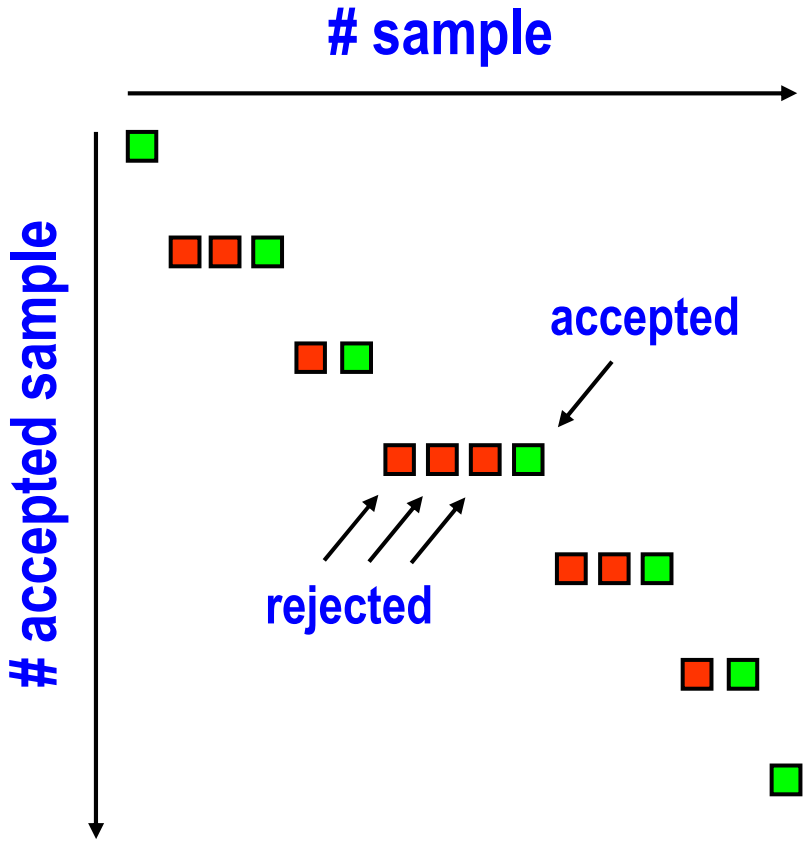
2. Accept the sample

$$\theta^{(i+1)} = \theta^*$$

with the probability

$$\min\left\{\frac{p_N(\theta^*)}{p_N(\theta^{(i)})}, 1\right\}$$

Metropolis Algorithm



1. Generate a candidate sample from symmetric Markov chain

$$\theta^* \text{ from } q(\theta \mid \theta^{(i)})$$

2. Accept the sample

$$\theta^{(i+1)} = \theta^*$$

with the probability

$$\min\left\{\frac{p_N(\theta^*)}{p_N(\theta^{(i)})}, 1\right\}$$

Markov Chain Monte Carlo (MCMC)

Statistical physics
[Metropolis 1953]

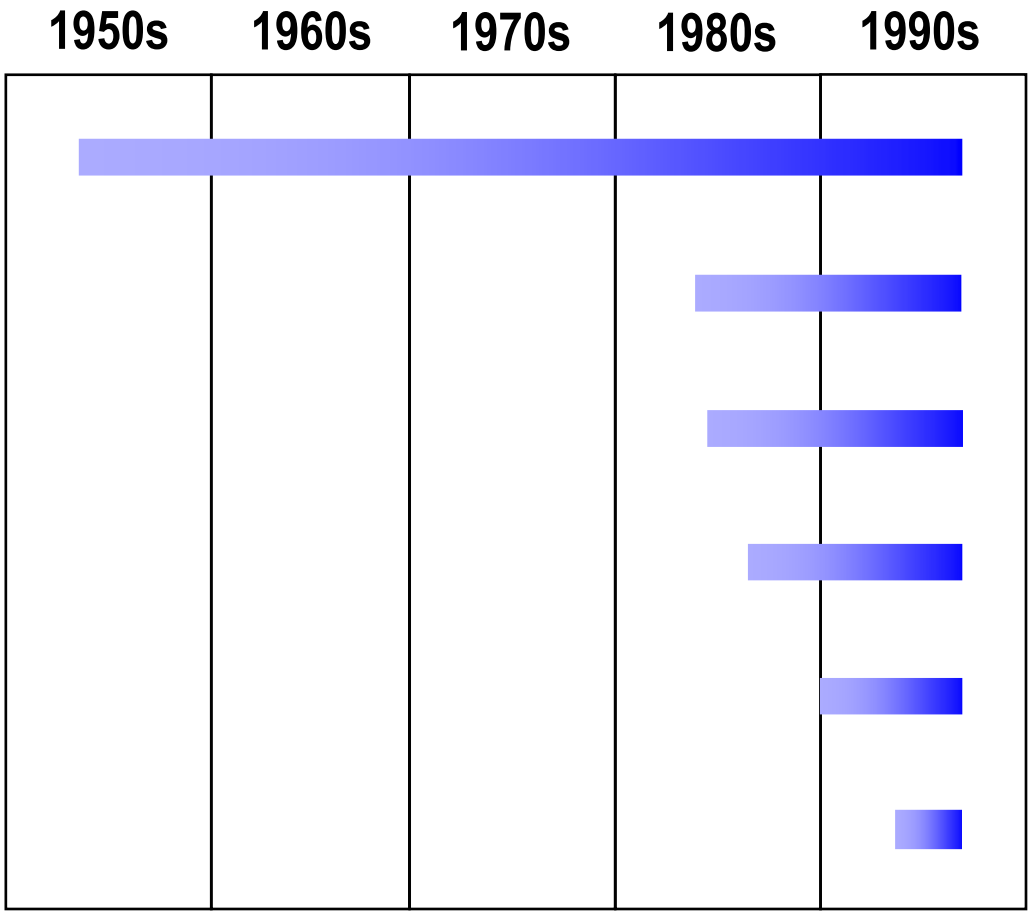
Simulated annealing
[Kirkpatrick et al 1983]

Image processing
[Geman & Geman 1984]

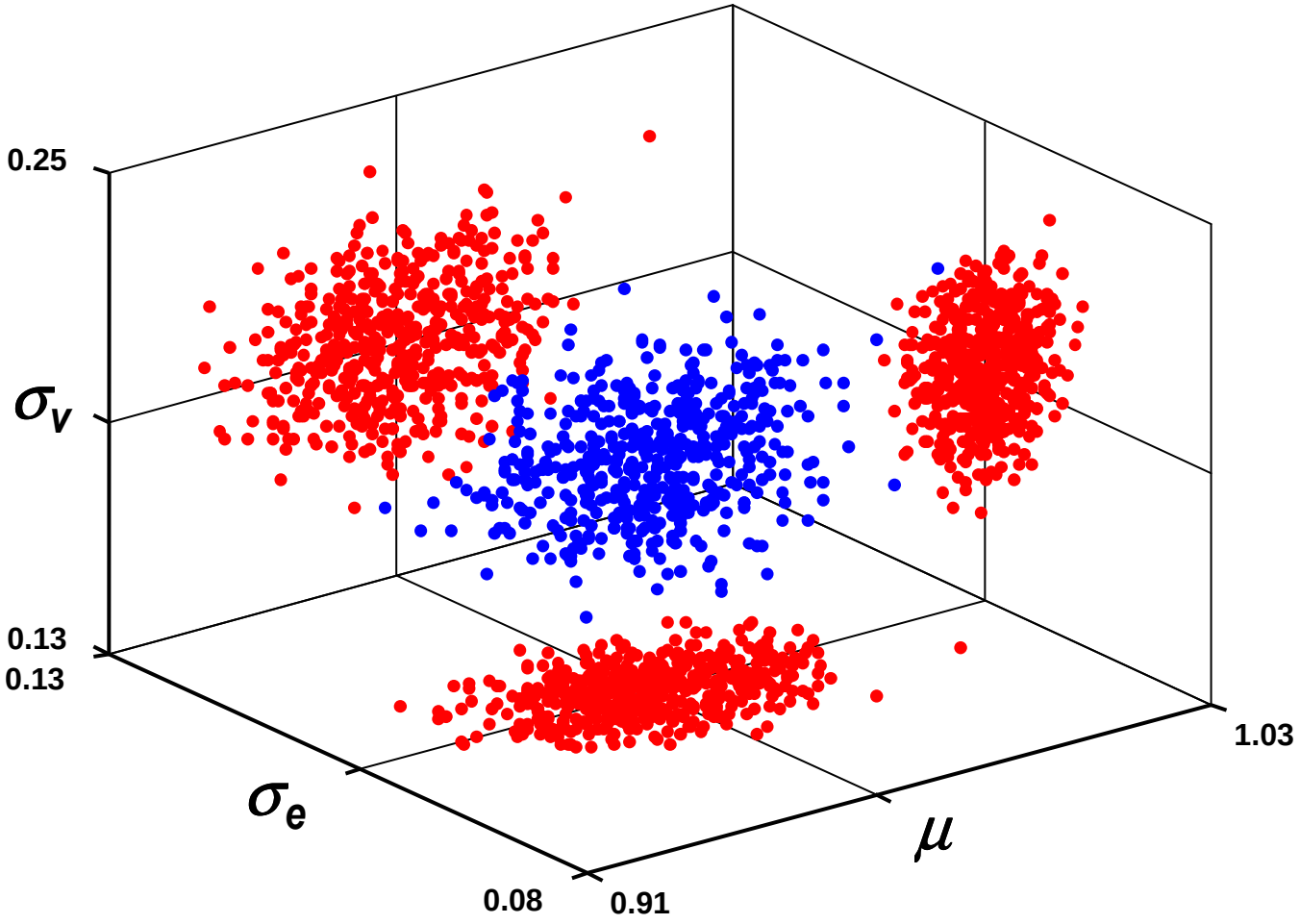
Belief networks
[Pearl 1987]

Bayesian analysis
[Gelfand & Smith 1990]

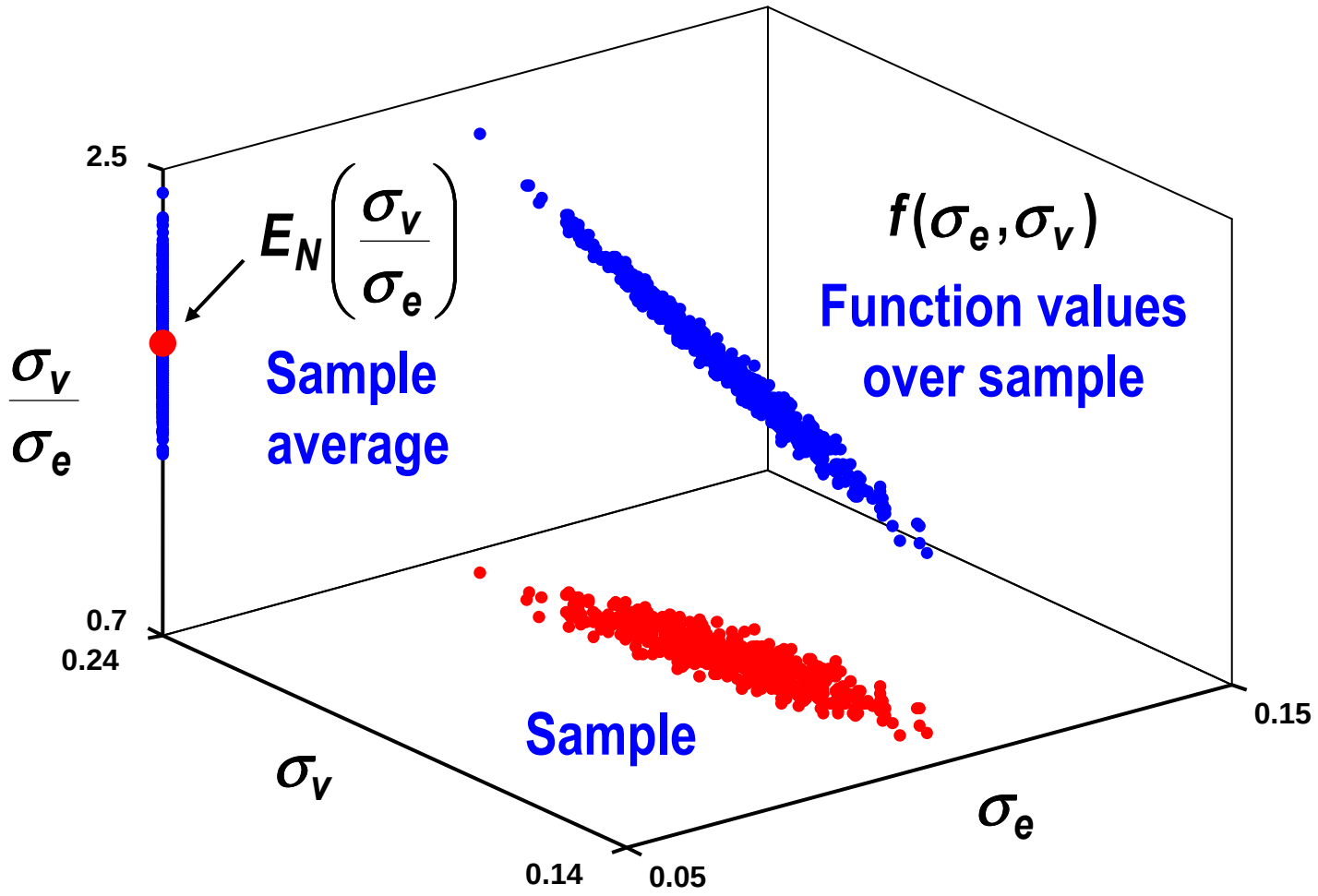
Engineering



Marginal Density via Sample Projection

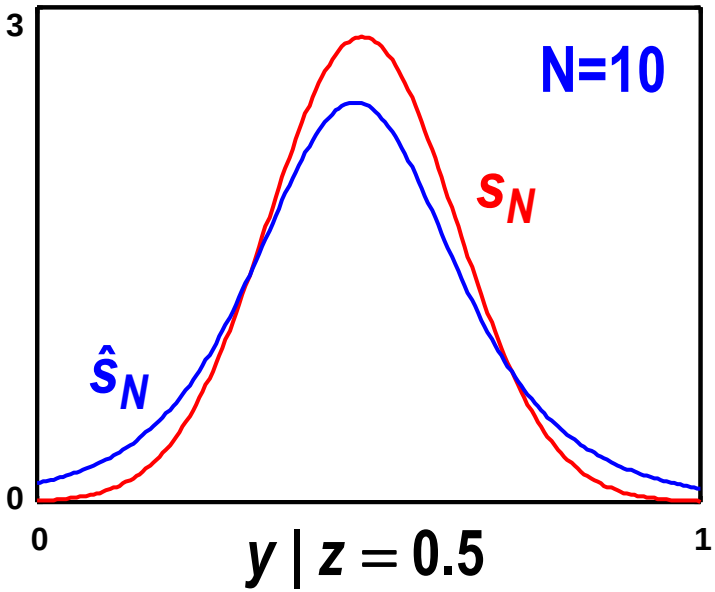


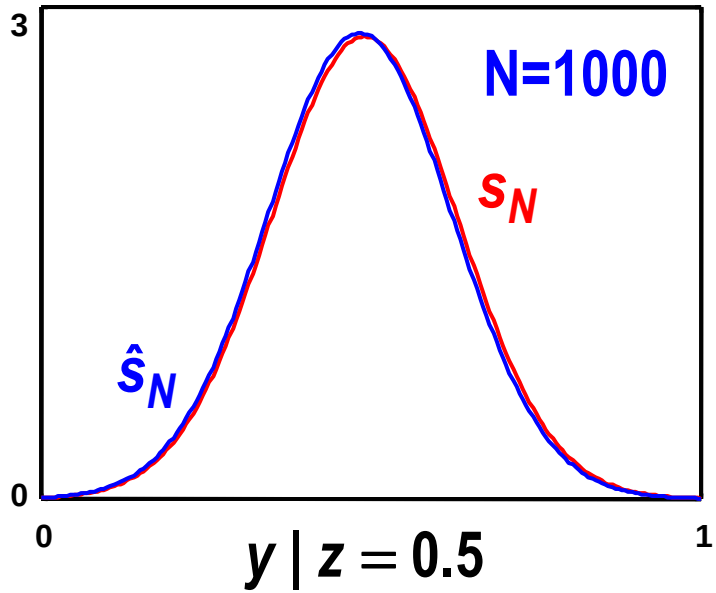
Expectation via Sample Average



Predictive Density via Sample Mixture

- **Predictive density** $s_N(y \mid z) = \iint s_\theta(y \mid z)\, p_N(\theta)\, \mathrm{d}\theta$
- **Rao-Blackwellized estimate** $\hat{s}_N(y \mid z) = \frac{1}{M} \sum_{i=1}^{M} s_{\theta(i)}(y \mid z)$





Finite Memory Estimation

Minimum Relative Entropy (MRE) Method

Ideal Compression of Data

- If $\{\log s_\theta(y \mid z)\}$ belongs to an affine space

$$\log s_\theta(y \mid z) = \log s_0(y \mid z) + \lambda^T(\theta)\, h(y,z) - \psi(\lambda(\theta)),$$

fixed $\neq f(z)$!

- then the inaccuracy

$$\bar{K}(r_N : s_\theta) = \bar{K}(r_N : s_0) - \lambda^T(\theta)\, \bar{h}_N + \psi(\lambda(\theta))$$

- depends on data only through the statistic

$$\bar{h}_N = \frac{1}{N} \sum_{k=m+1}^{N+m} h(y_k, z_k).$$

➔ *The statistic carries all essential information.*

Typical Construction of *h*-Statistic

Parameter grid
$\theta_1, \mathrm{K}, \theta_{n+1}$

- **Logarithm of density ratio**

$$h_i(y,z) = \log \frac{s_{\theta_i}(y \mid z)}{s_{\theta_{i+1}}(y \mid z)}, \quad i = 1, \mathrm{K}, n$$

- **Inaccuracy difference**

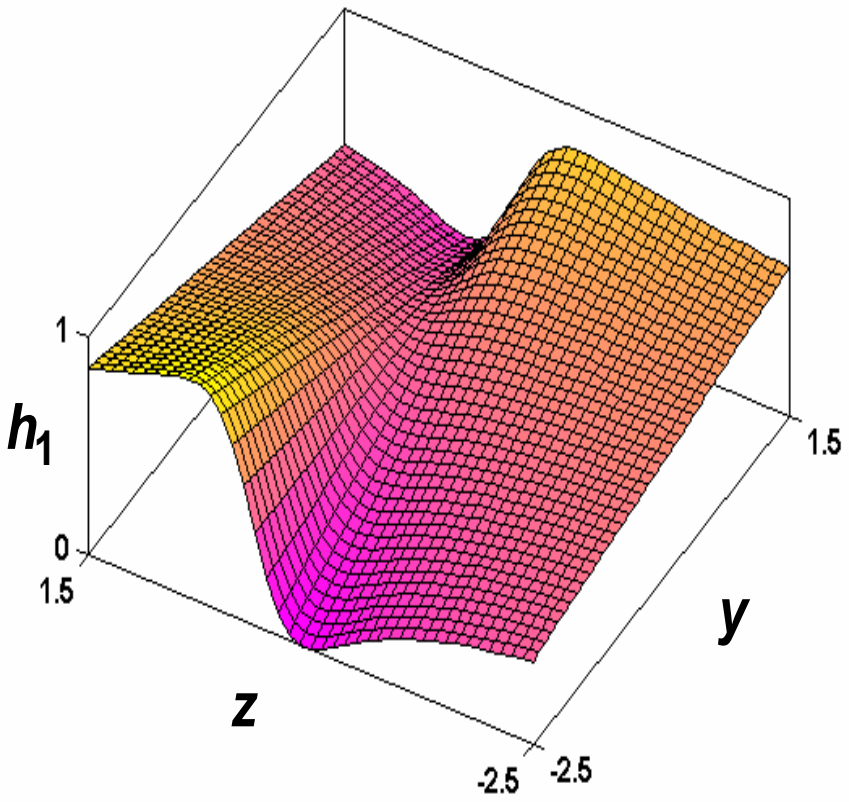
$$\bar{h}_{i,N} = -\bar{K}(r_N : s_{\theta_i}) + \bar{K}(r_N : s_{\theta_{i+1}}), \quad i = 1, \mathrm{K}, n$$

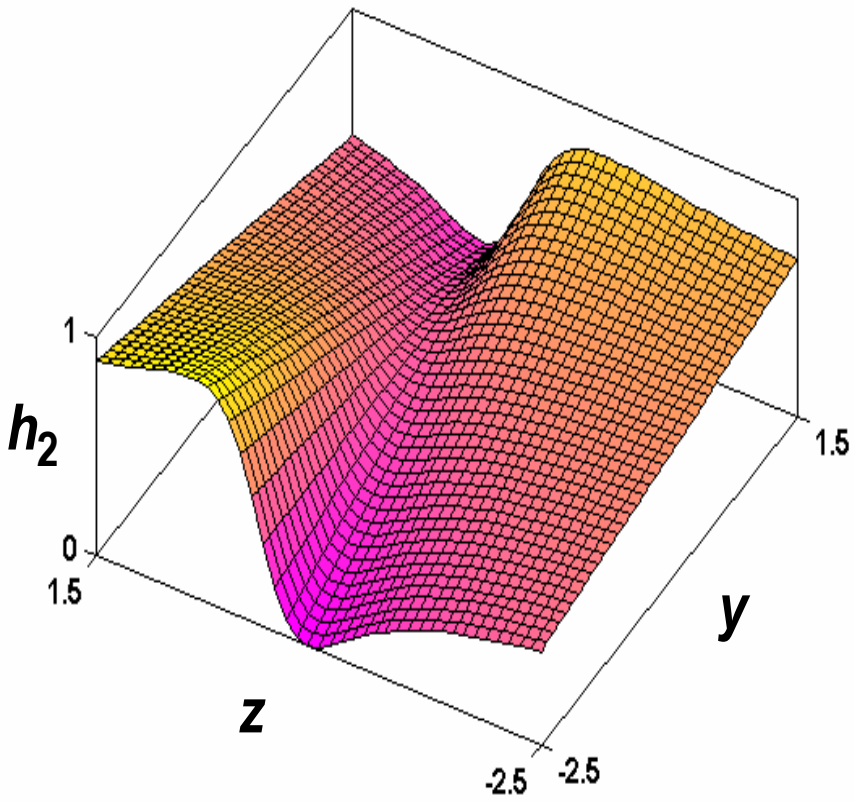
- **Normalized logarithm of likelihood ratio**

$$\bar{h}_{i,N} = \frac{1}{N} \log \frac{I_N(\theta_i)}{I_N(\theta_{i+1})}, \quad i = 1, \mathrm{K}, n$$

AR(1) Coefficient

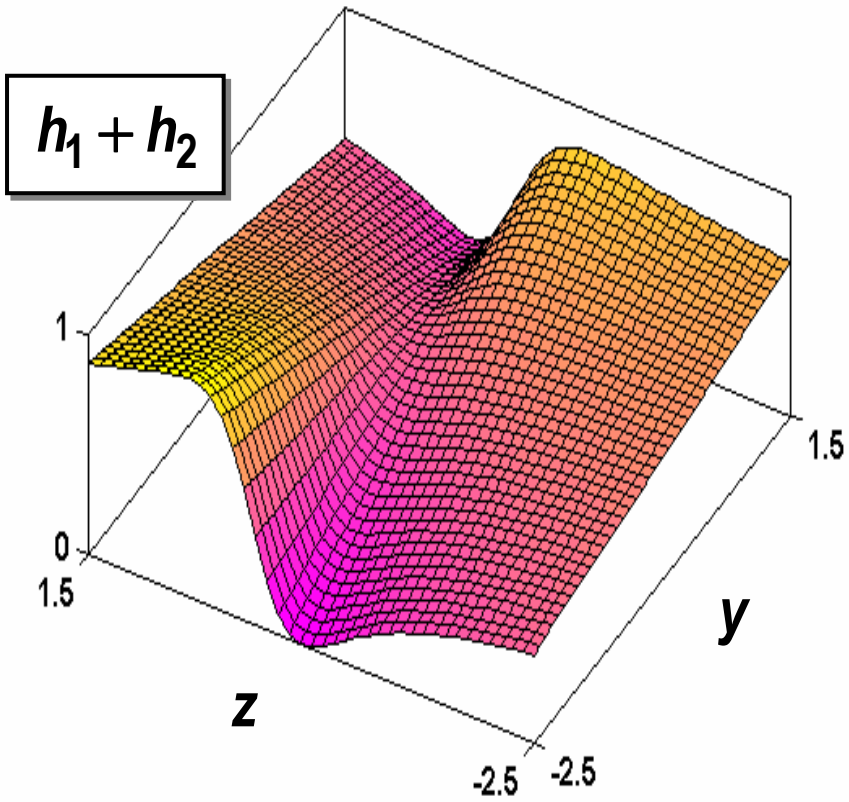
$\mu_1 = 0.5 \quad \mu_2 = 1.0 \quad \mu_3 = 1.5$

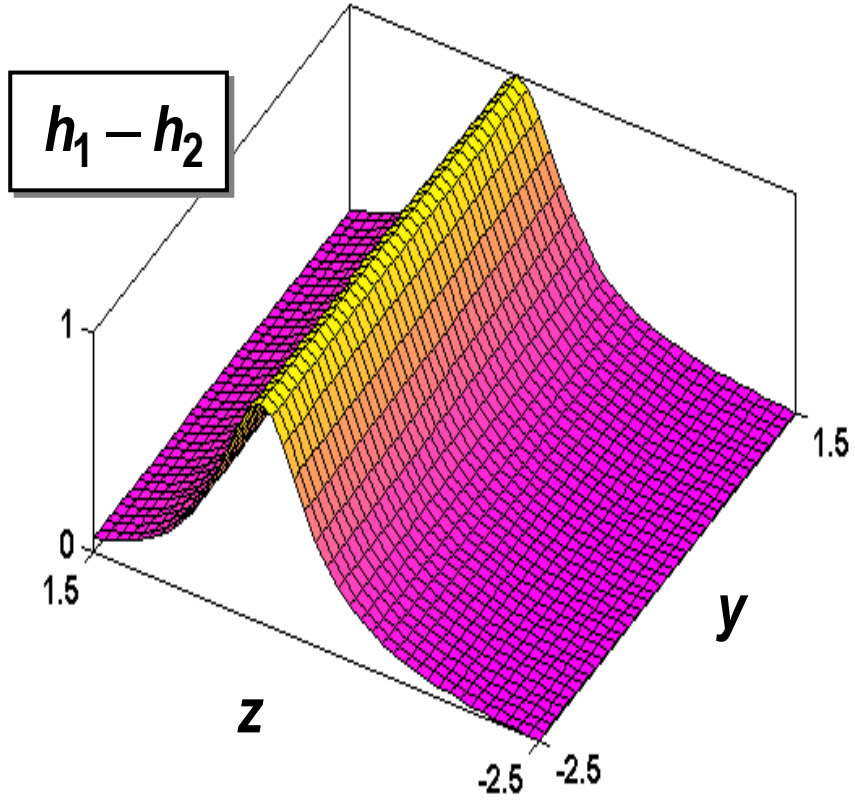




AR(1) Coefficient

$\mu_1 = 0.5 \quad \mu_2 = 1.0 \quad \mu_3 = 1.5$

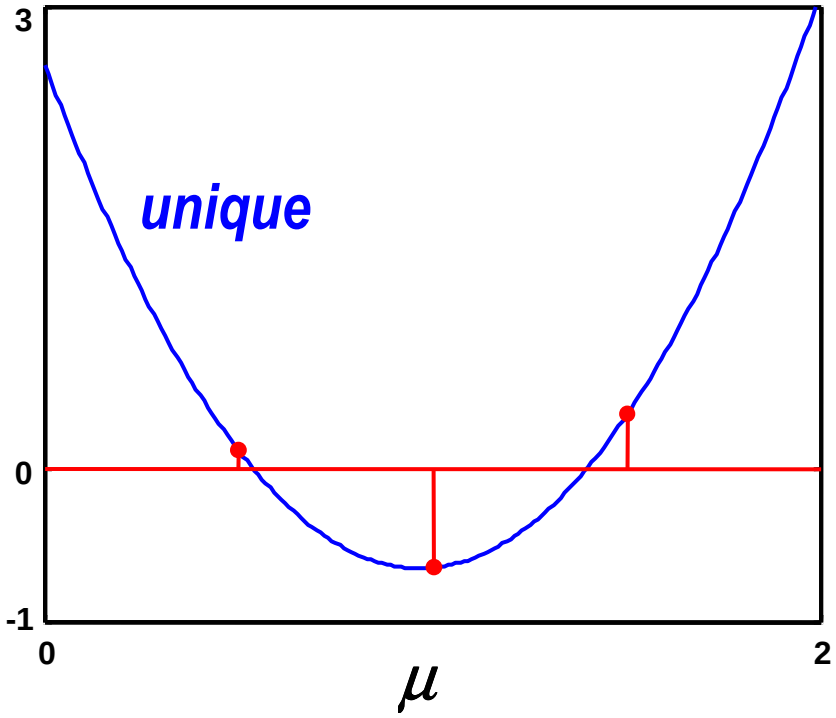




Three Points Are Sufficient

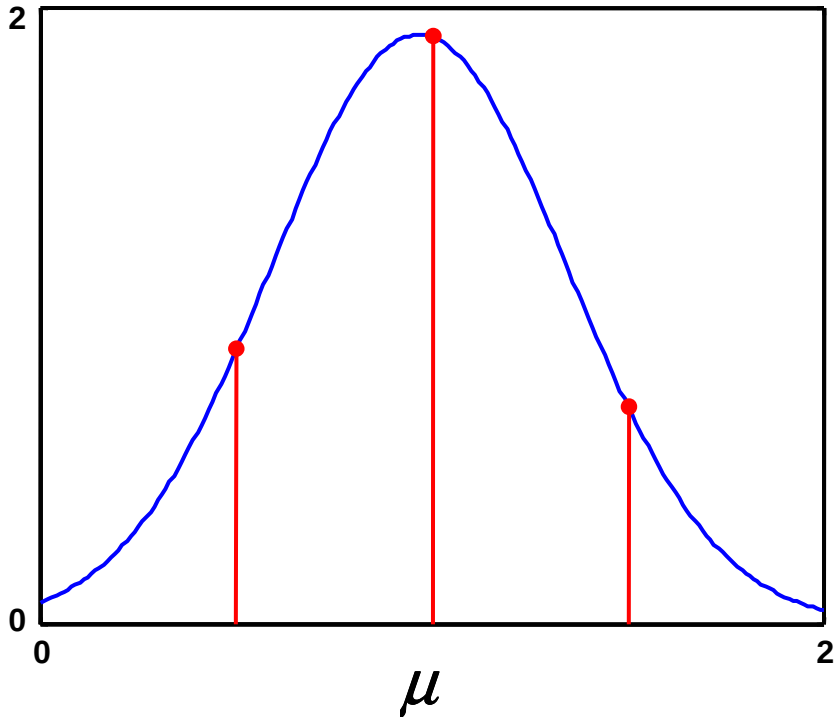
Inaccuracy

$\overline{K}(r_N : s_\mu)$



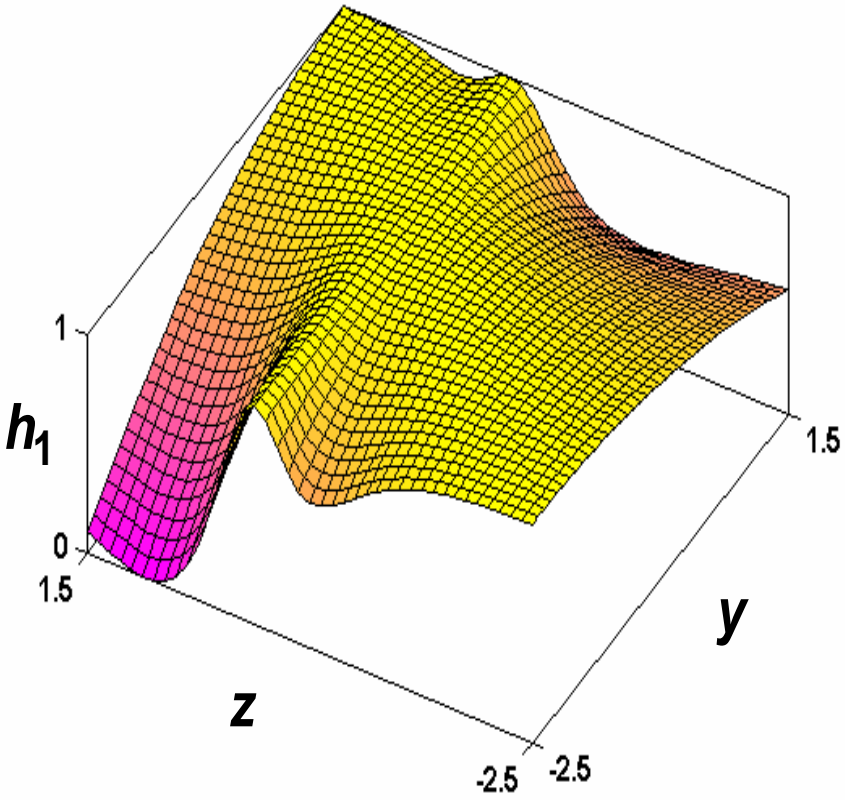
Likelihood

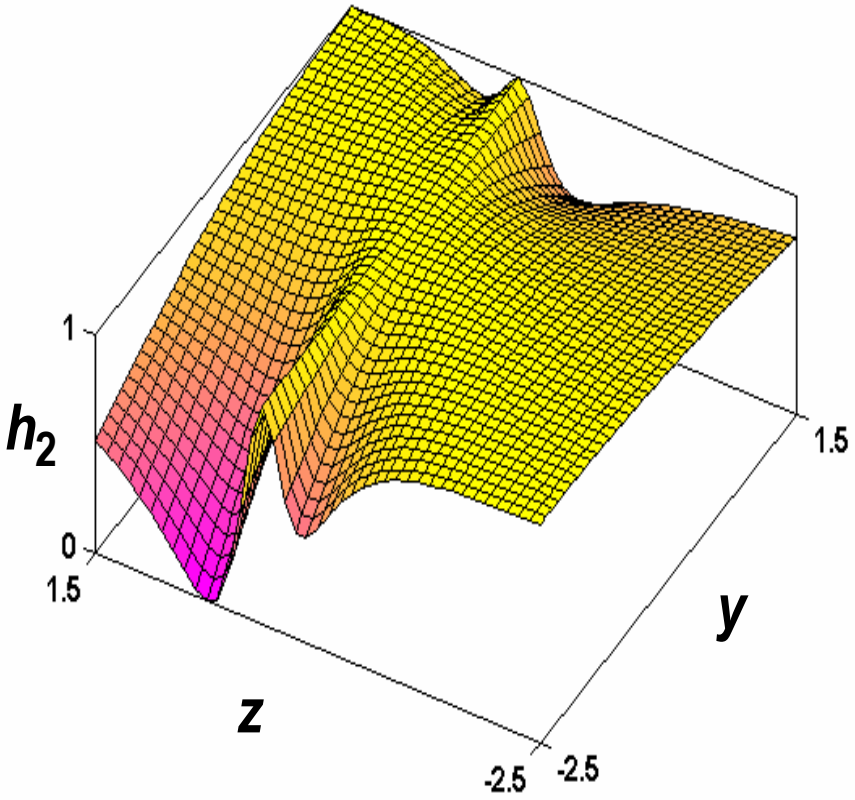
$[\, I_N(\mu) \,]^{1/N}$



Standard Deviation of AR(1) Fluctuation

$$\sigma_{v1} = 0.1 \quad \sigma_{v2} = 0.2 \quad \sigma_{v3} = 0.3$$

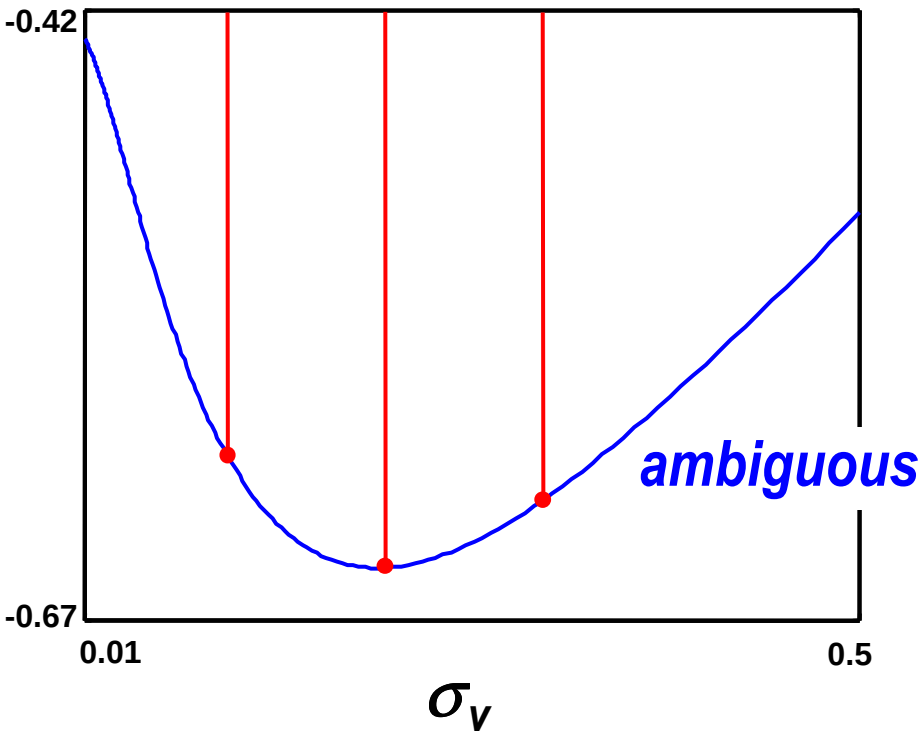




Three Points Are NOT Enough

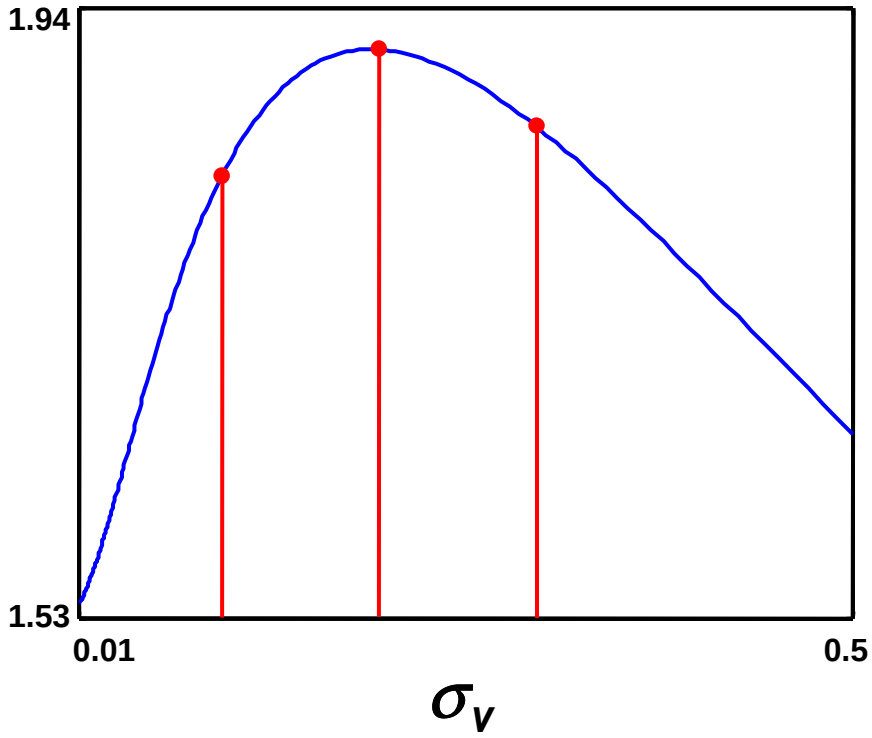
Inaccuracy

$\overline{K}(r_N : s_{\sigma_v})$

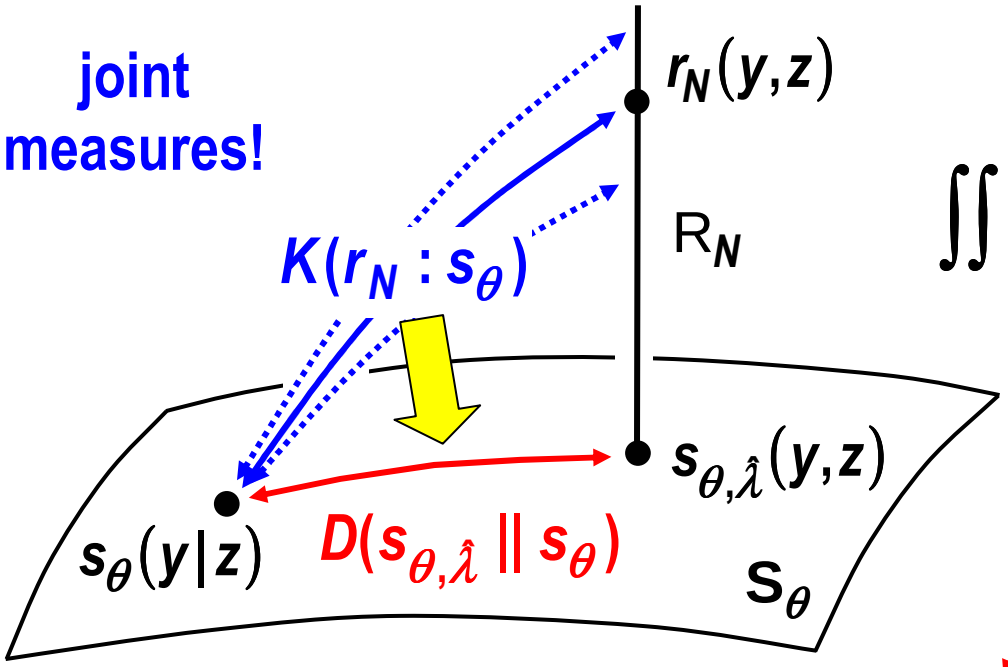


Likelihood

$[\,l_N(\sigma_v)\,]^{1/N}$



How to Cope with Ambiguity?



$$\iint r(y,z)\, h(y,z)\, dy\, dz = \bar{h}_N$$

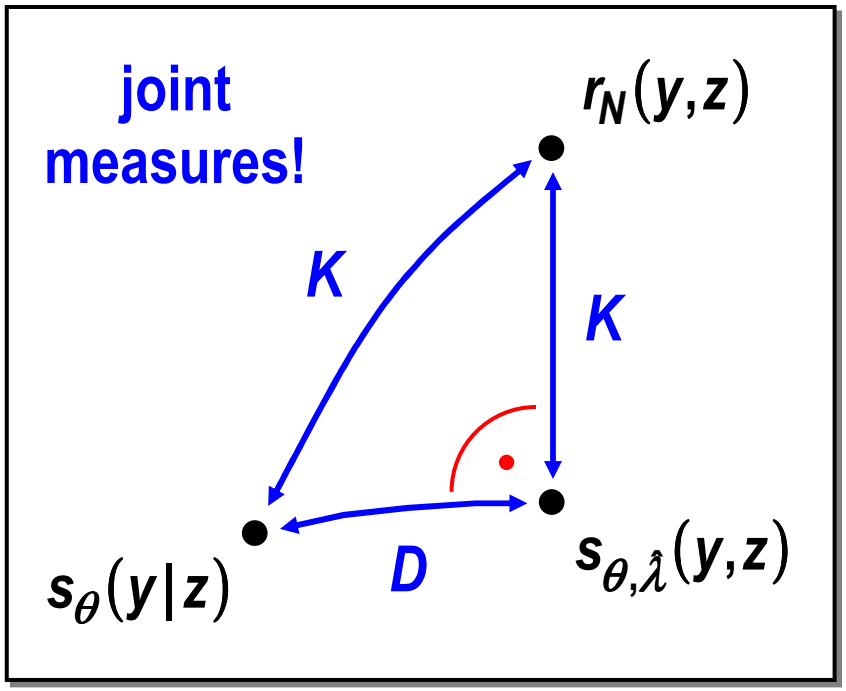
$$s_{\theta,\lambda}(y,z) = s_\theta(y|z) \exp(\lambda' h(y,z) - \psi(\theta,\lambda))$$

Pythagorean Relationship

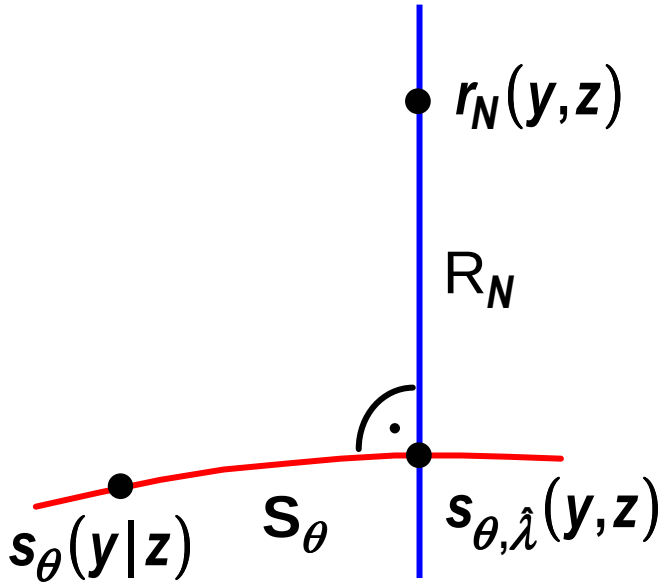
$$K(r_N : s_\theta) = K(r_N : s_{\theta,\hat{\lambda}}) + D(s_{\theta,\hat{\lambda}} \,\|\, s_\theta)$$

$\min_{\lambda \in R^n} K(r_N : s_{\theta,\lambda})$

$\min_{r \in \mathrm{R}_N} D(r \,\|\, s_\theta)$



Minimum Relative Entropy (MRE) Inference



$$D(\mathbf{R}_N \,\|\, s_\theta) = \min_{r \in \mathbf{R}_N} D(r \,\|\, s_\theta)$$

1 choose *h*-statistic so that

$$K(r_N : s_{\theta,\hat{\lambda}}) \approx \text{const.}$$

2 approximate inaccuracy

$$K(r_N : s_\theta) \approx D(\mathbf{R}_N \,\|\, s_\theta) + \text{const.}$$

3 approximate likelihood

$$\hat{I}_N(\theta) = c \exp(-N D(\mathbf{R}_N \,\|\, s_\theta))$$

How to Interpret Relative Entropy?

$$D(r \,\|\, s) = \iint r(y,z) \log \frac{r(y,z)}{s(y \mid z)} \, dy \, dz$$

Conditional!

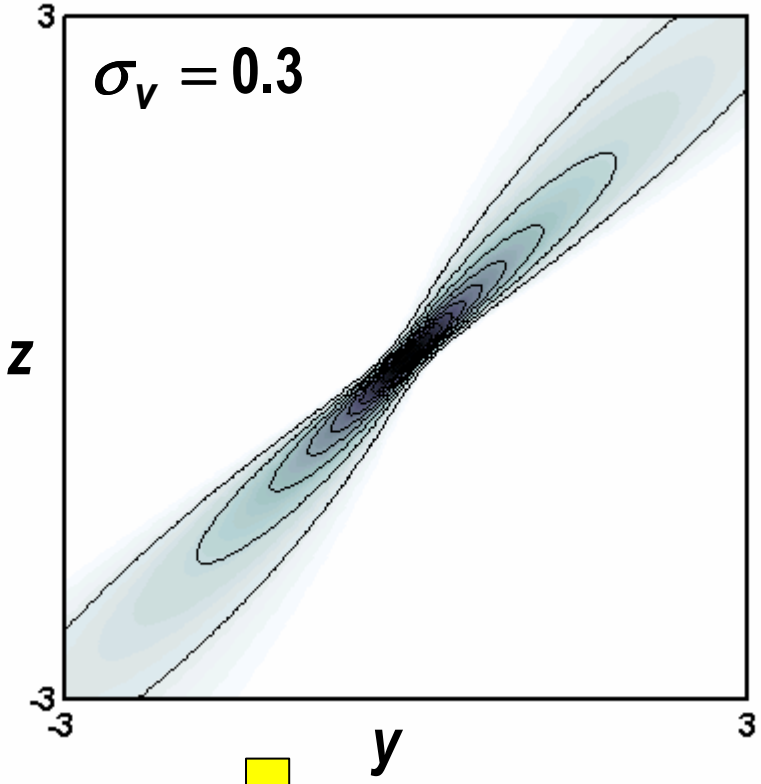
$$= \int r(z) \int r(y \mid z) \log \frac{r(y \mid z)}{s(y \mid z)} \, dy \, dz \; - \; \int r(z) \log \frac{1}{r(z)} \, dz$$

Conditional relative entropy

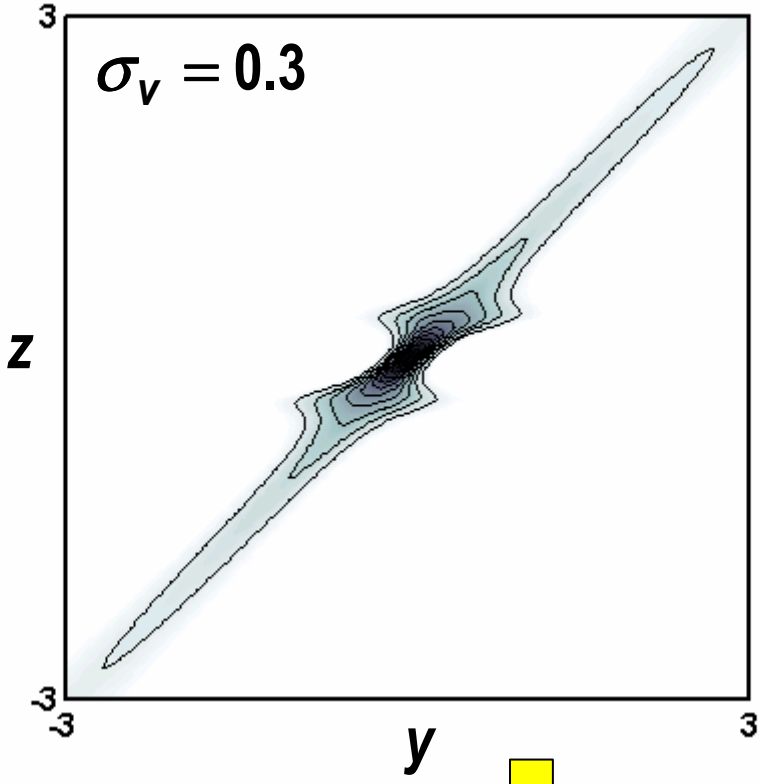
Marginal entropy

MRE Indirect Computation

$s_{\sigma_v}(y \mid z)$



$s_{\sigma_v,\hat{\lambda}}(y,z)$



$$D(\mathbf{R}_N \,\|\, s_{\sigma_v}) = D(s_{\sigma_v,\hat{\lambda}} \,\|\, s_{\sigma_v})$$

MRE Direct Computation

- **Convex optimization problem**

$$D(\mathrm{R}_N \,\|\, s_\theta) = \min_{r \in \mathrm{R}_N} D(r \,\|\, s_\theta)$$

$$= \max_{\lambda \in R^n} \left(\lambda' \bar{h}_N - \psi(\theta, \lambda) \right)$$

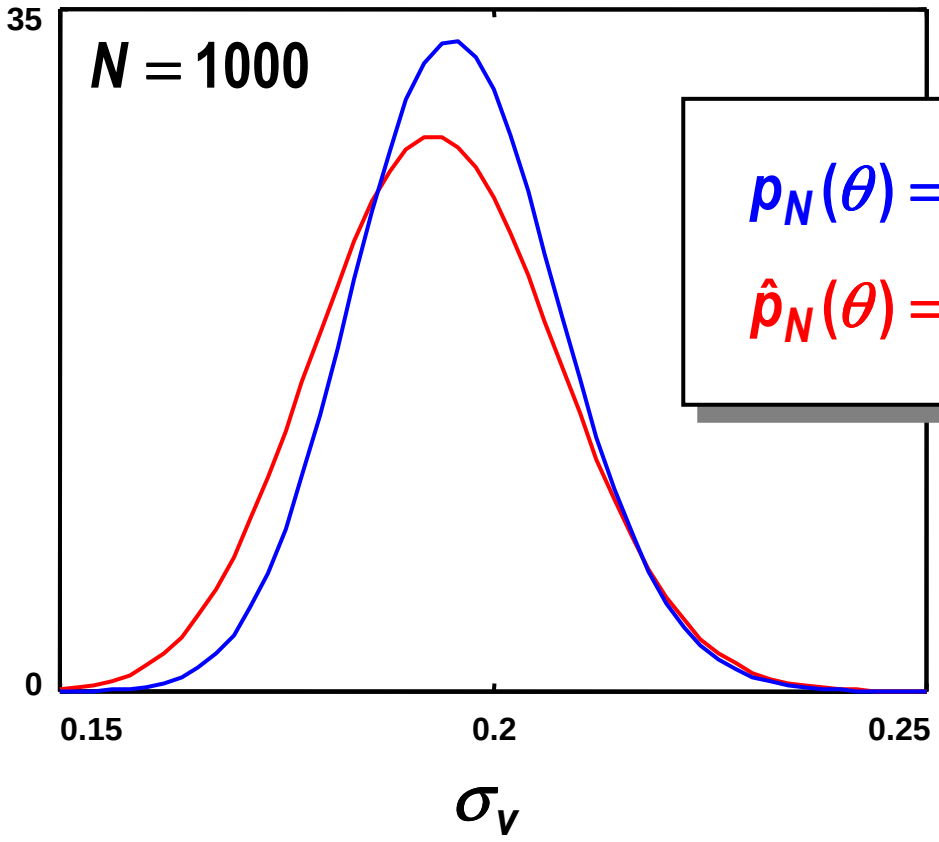
Statistic

Logarithm of normalizing divisor

- **Entails multivariate integration**

$$\psi(\theta, \lambda) = \log \iint s_\theta(y|z) \exp(\lambda' h(y,z)) \, dy \, dz$$

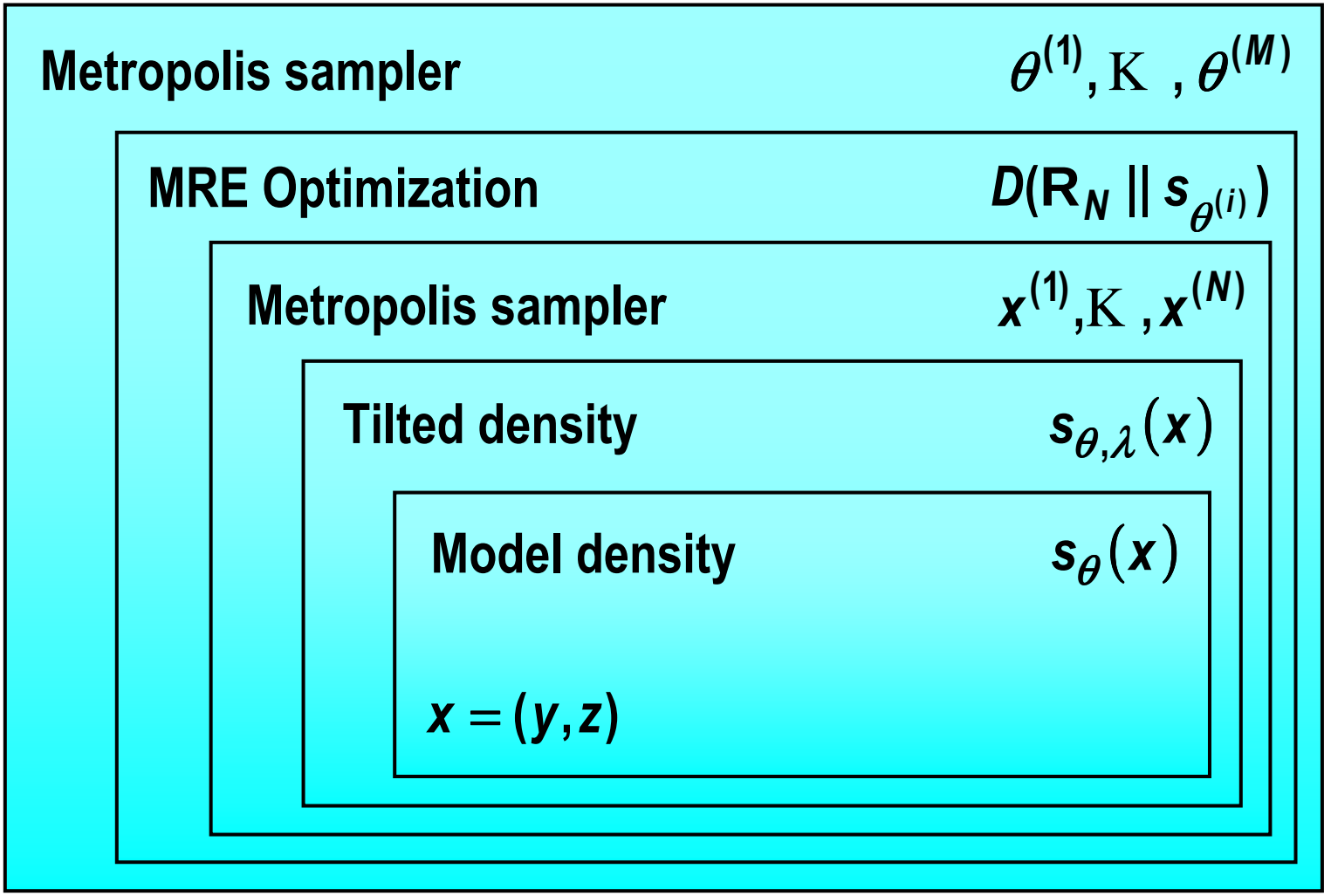
Posterior Approximation



$$p_N(\theta) = c\, p_0(\theta) \exp\left(-N K(r_N : s_\theta)\right)$$

$$\hat{p}_N(\theta) = c\, p_0(\theta) \exp\left(-N D(\mathbf{R}_N \,\|\, s_\theta)\right)$$

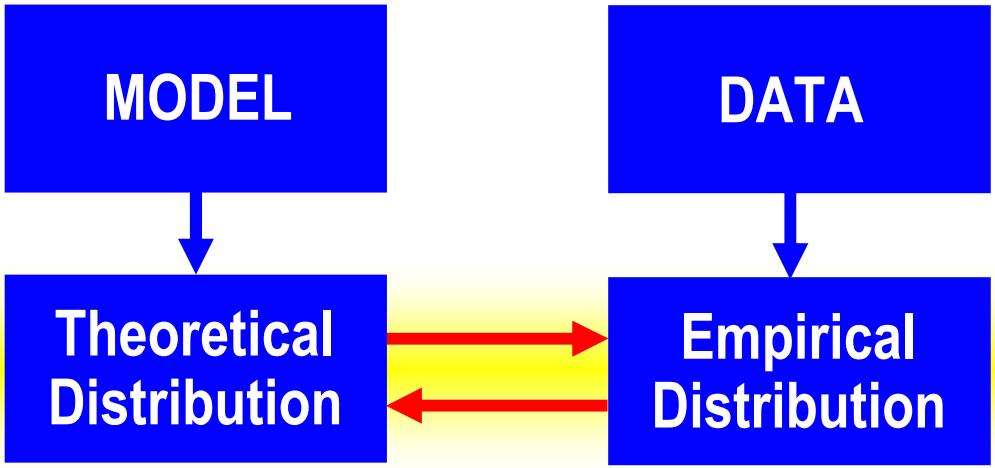
MCMC Implementation



Reinforcement

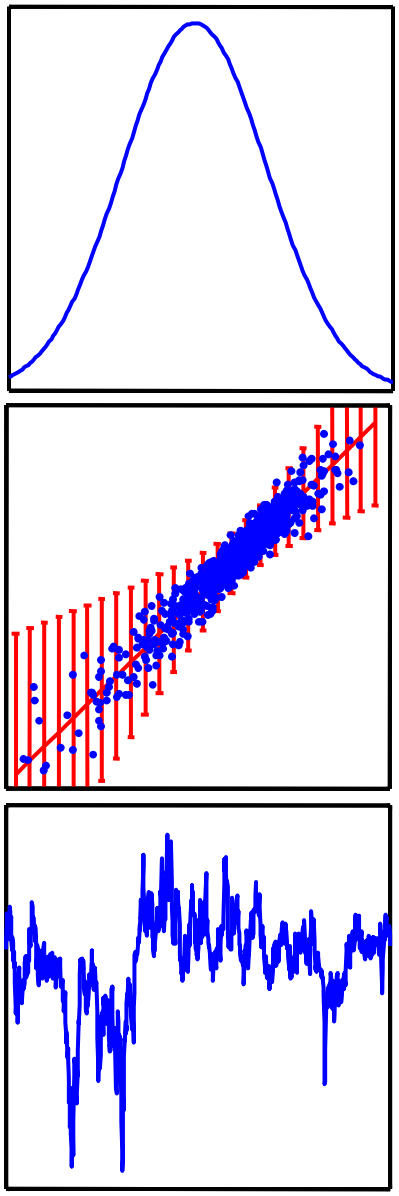
Highlights, Benefits, Limits

MODELLING:
Sample → Empirical Distribution



- *Data and model on the same level.*
- *A unified language of probability calculus.*

Shift of Paradigm



A unique "distance"

model ↑↓ data

Plenty of possible distances

Distributions of parameters

- prior $p_0(\theta)$
- posterior $p_N(\theta)$

Distributions of data

- empirical $r_N(y,z)$
- theoretical $s_\theta(y \mid z)$

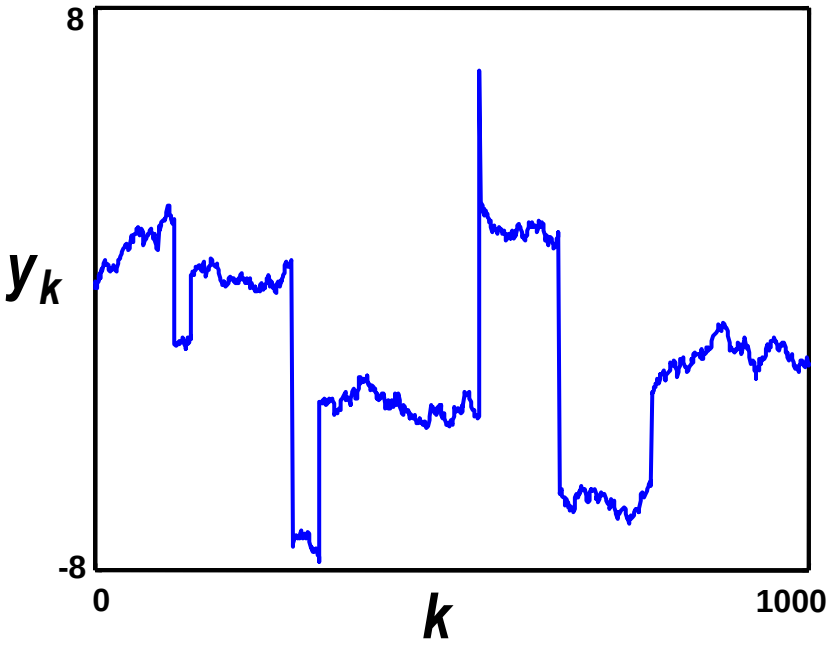
Measured data

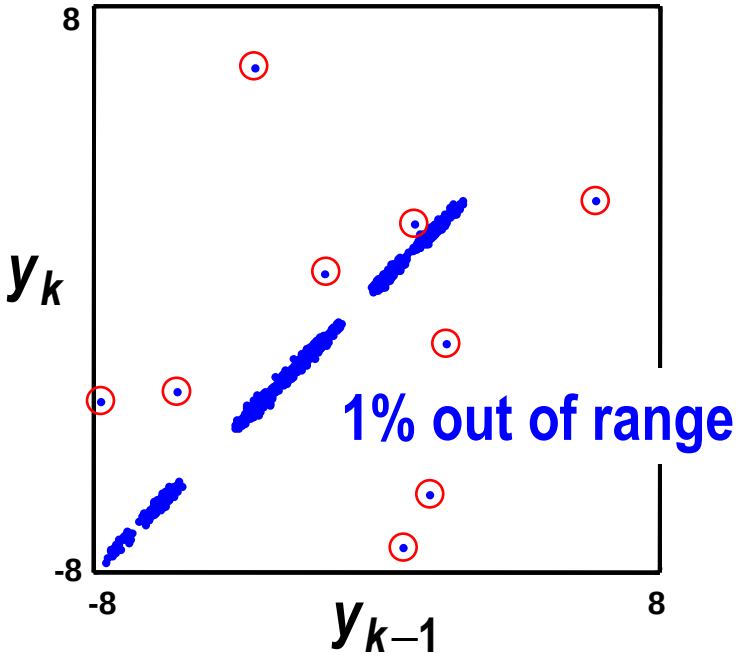
- output $y_1, \mathrm{K}\ , y_N$
- regressor $z_1, \mathrm{K}\ , z_N$

Example 1: Random Walk + Abrupt Jumps

ε-contaminated normal distribution

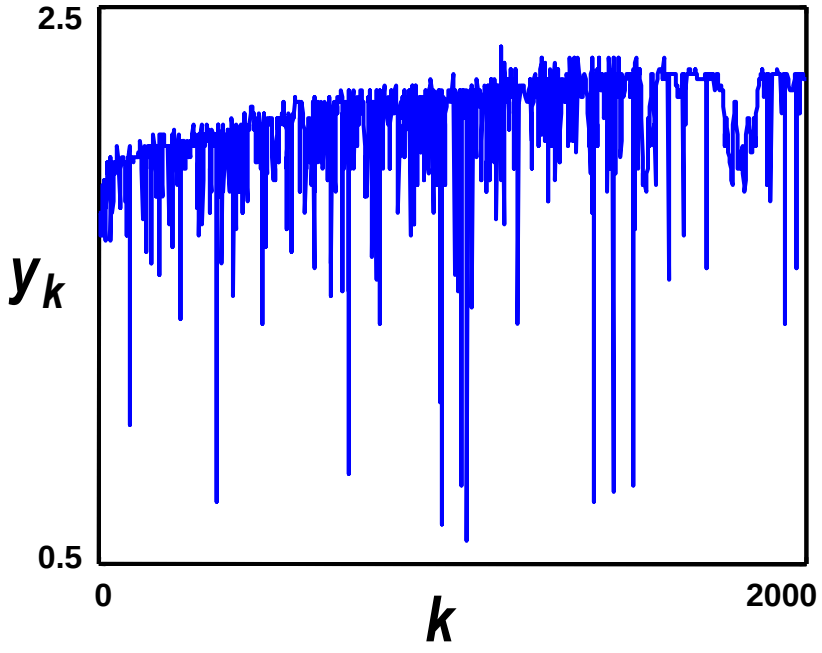
$$y_{k+1} = y_k + v_k, \quad v_k \sim (1-\varepsilon)\, N(0,0.01) + \varepsilon\, N(0,25), \quad \varepsilon = 0.01$$

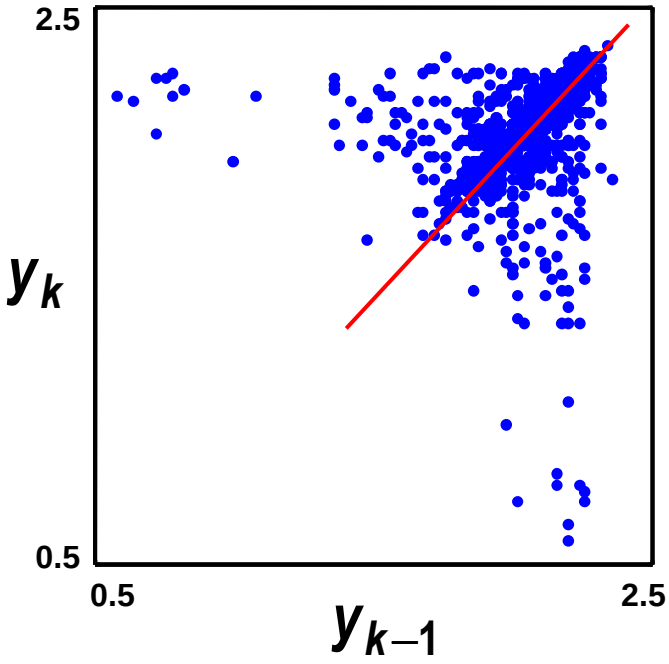




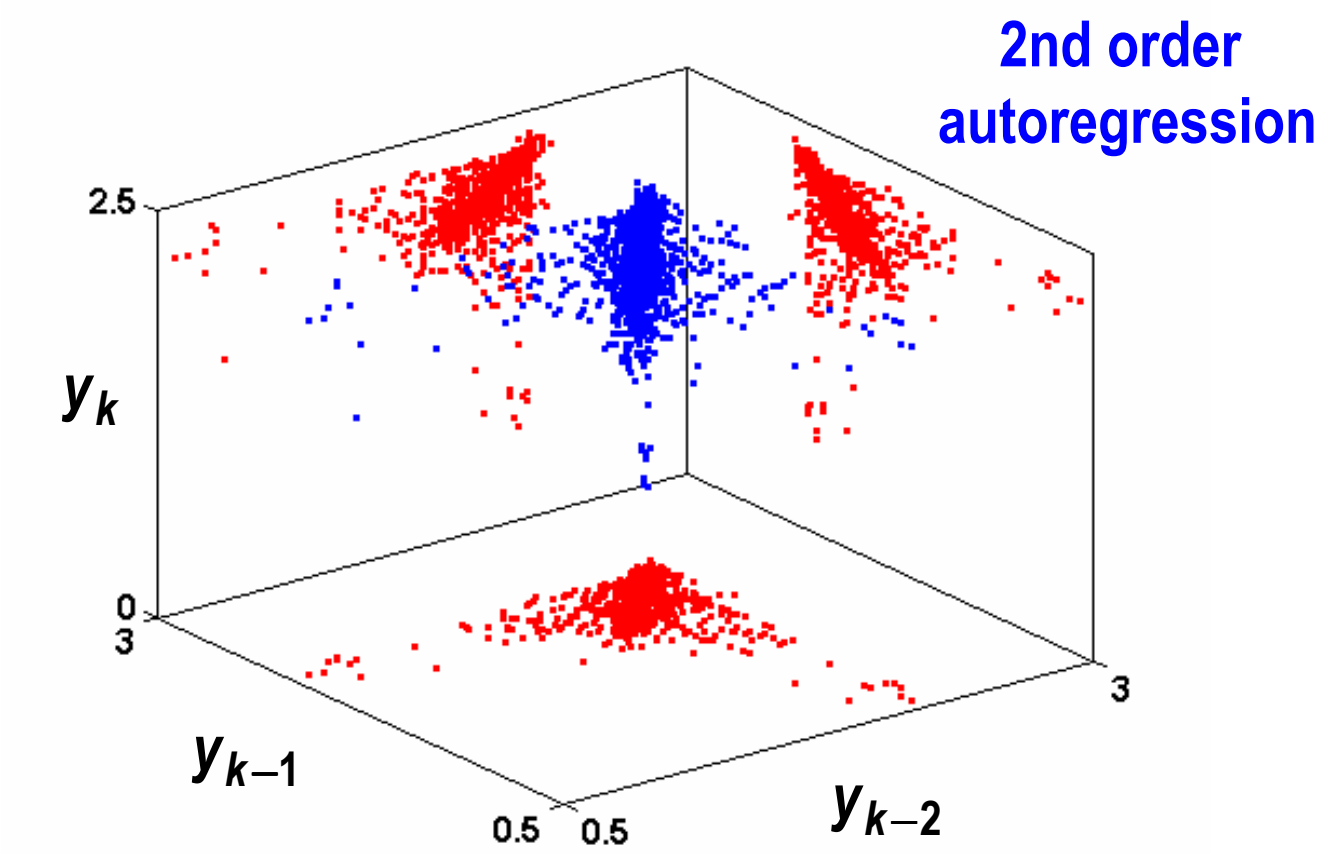
Example 2: One-Sided Outliers

Sun radiation measurements taken from a balloon
[Davis and Gather, JASA, 1993]

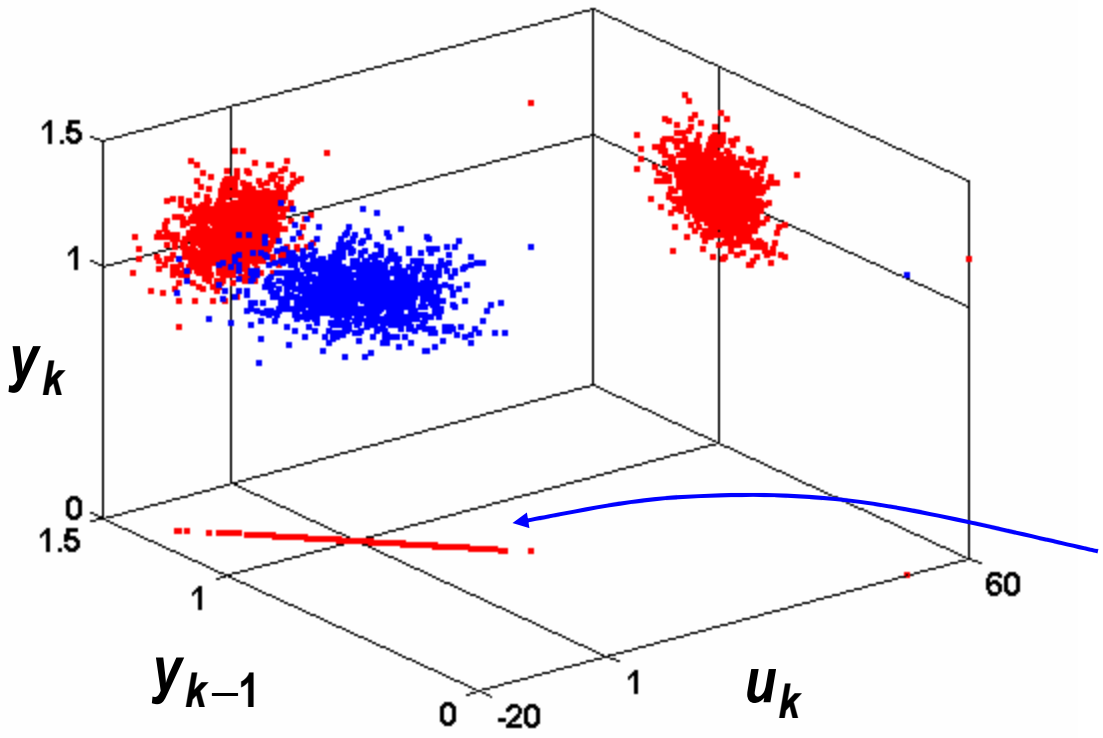




Example 2: One-Sided Outliers



Example 3: ARX Model + Poor Excitation



ARX model

$$y_k = \theta_1 y_{k-1} + \theta_2 u_k + e_k$$

$$\theta_1 = 0.98, \quad \theta_2 = 0.02$$

$$e_k \sim N(0, 0.01)$$

Linear feedback

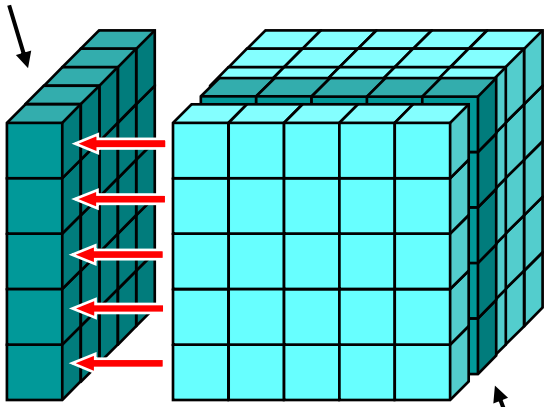
$$u_k = \frac{1}{\theta_2} - \frac{\theta_1}{\theta_2} y_{k-1}$$

Empirical Distribution ⇔ Data Hypercube

$$x = (y, z)$$

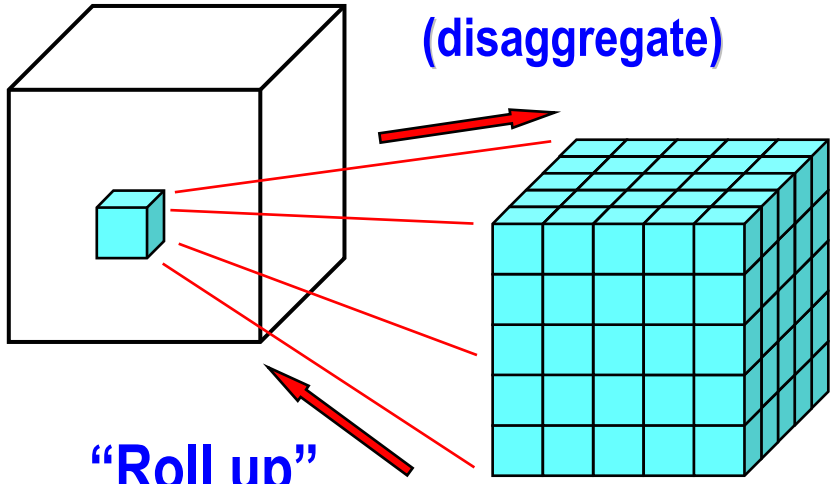
$$x = (x_1, x_2, \mathrm{K}, x_m)$$

Projection
(marginal distribution)



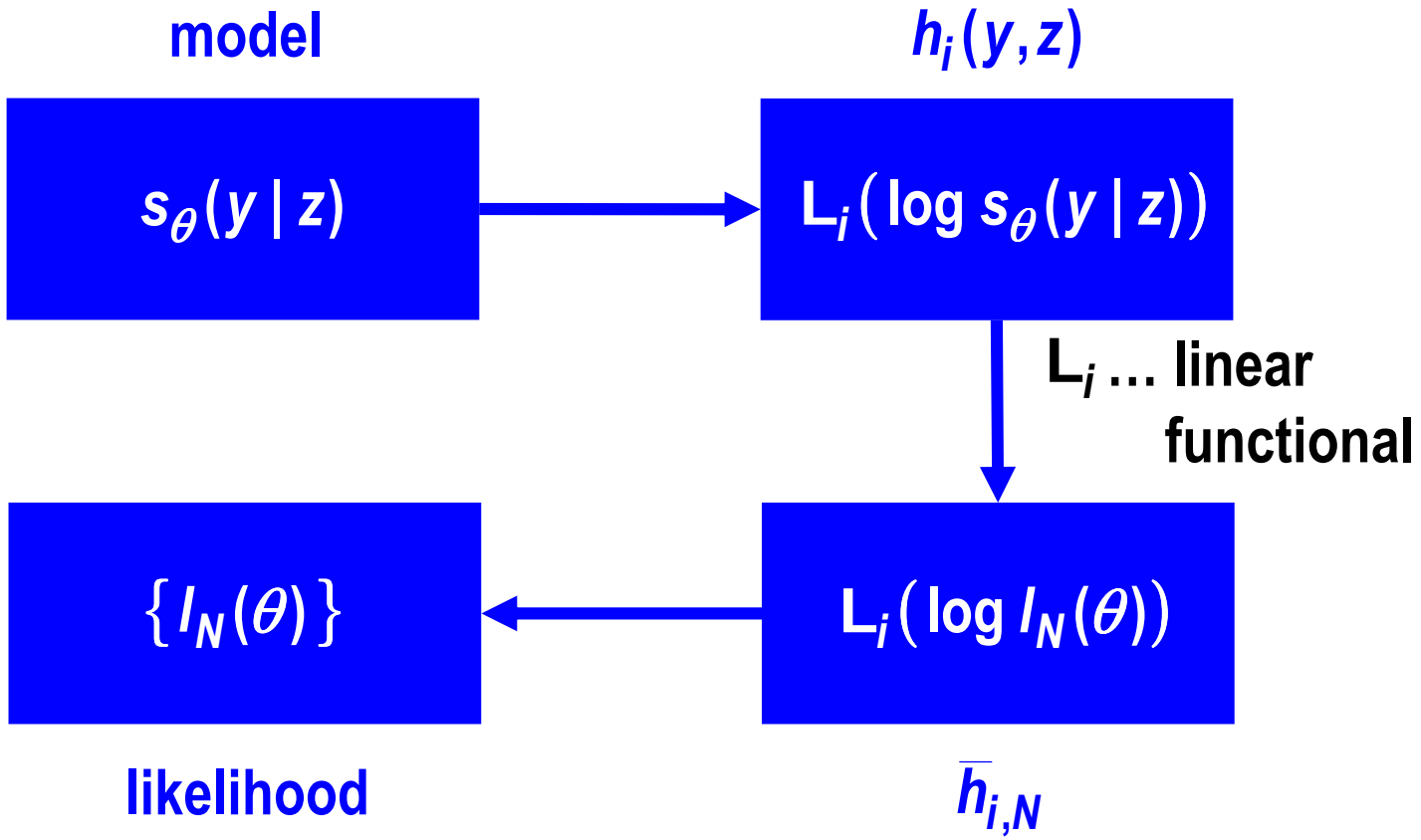
Cross-section
(conditional distribution)

"Drill down"
(disaggregate)



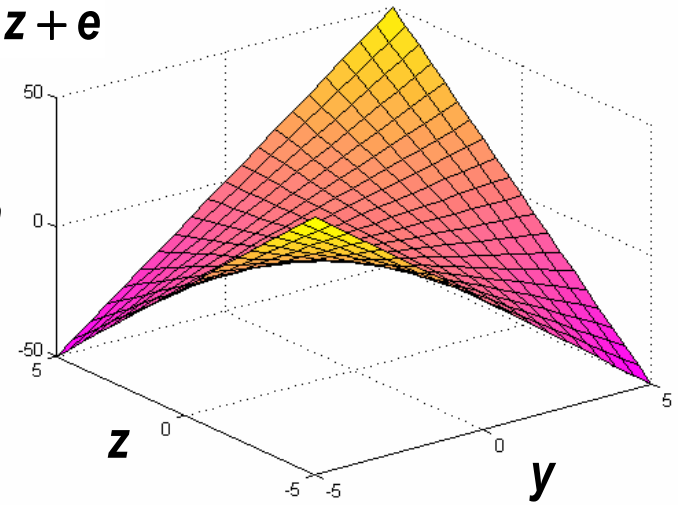
"Roll up"
(aggregate)

FINITE MEMORY ESTIMATION: Sample ➔ Statistic



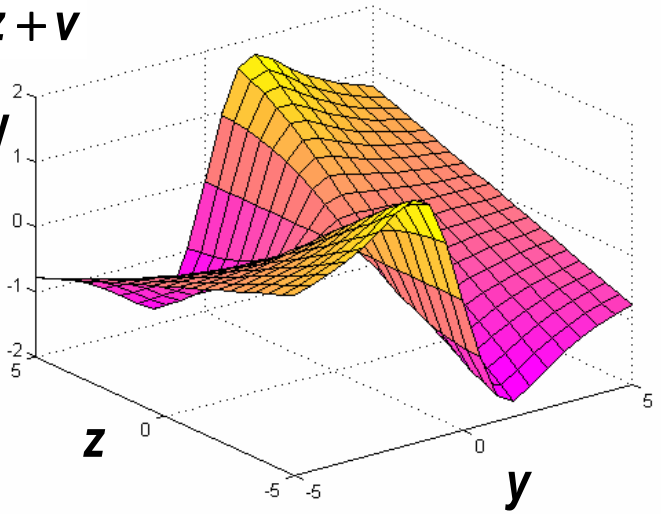
$$h_i(y,z) = \log s_{\theta_i}(y \mid z) - \log s_{\theta_{i+1}}(y \mid z)$$

$y = \theta z + e$

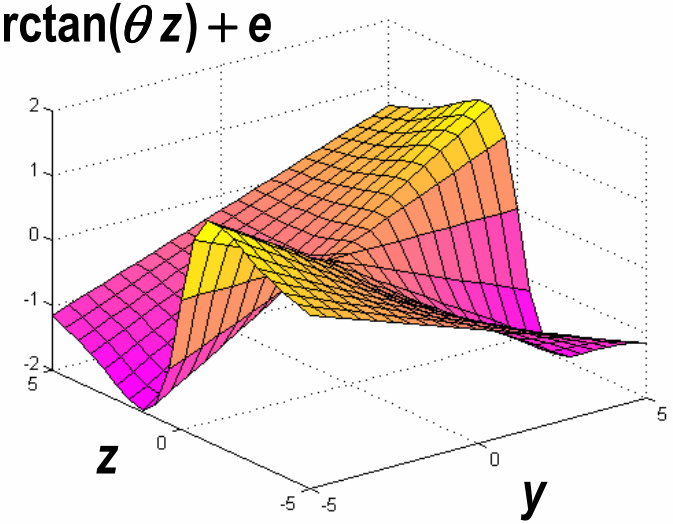


$y = \theta z + v$

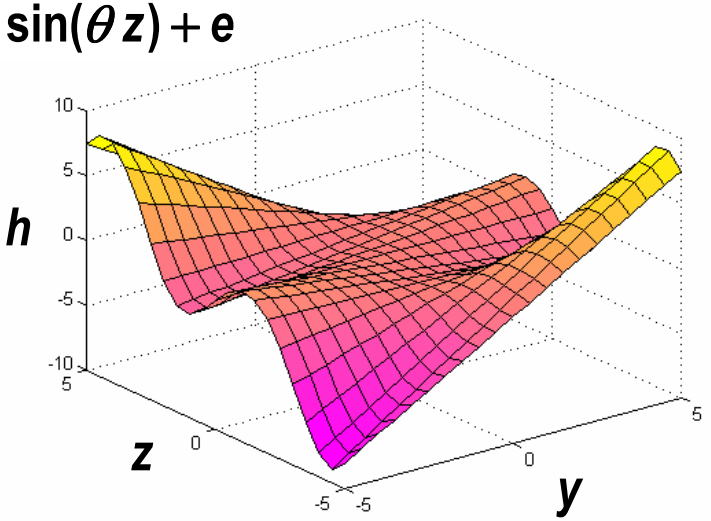
Cauchy noise



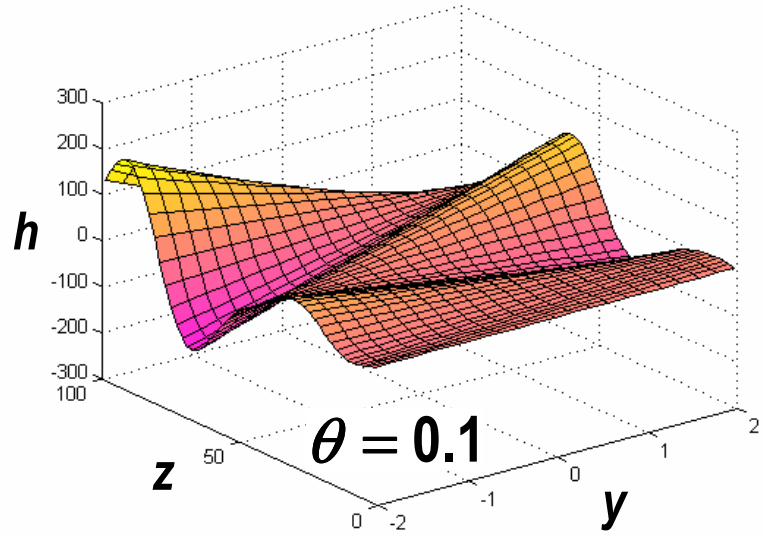
$y = \arctan(\theta z) + e$

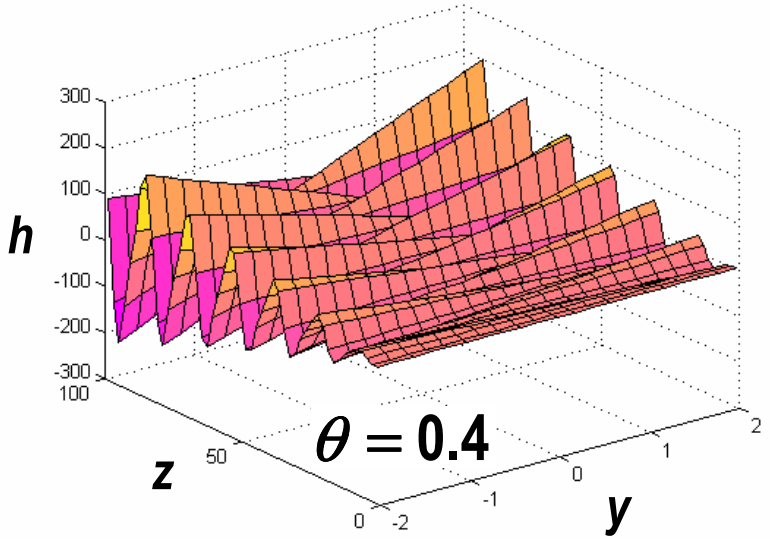


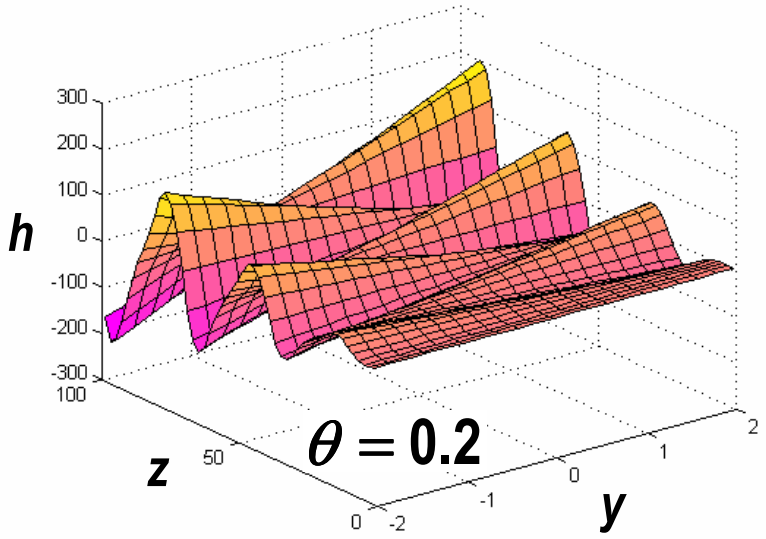
$y = \sin(\theta z) + e$



$$h_i(y,z) = \partial \log s_{\theta_i}(y \mid z) / \partial \theta$$

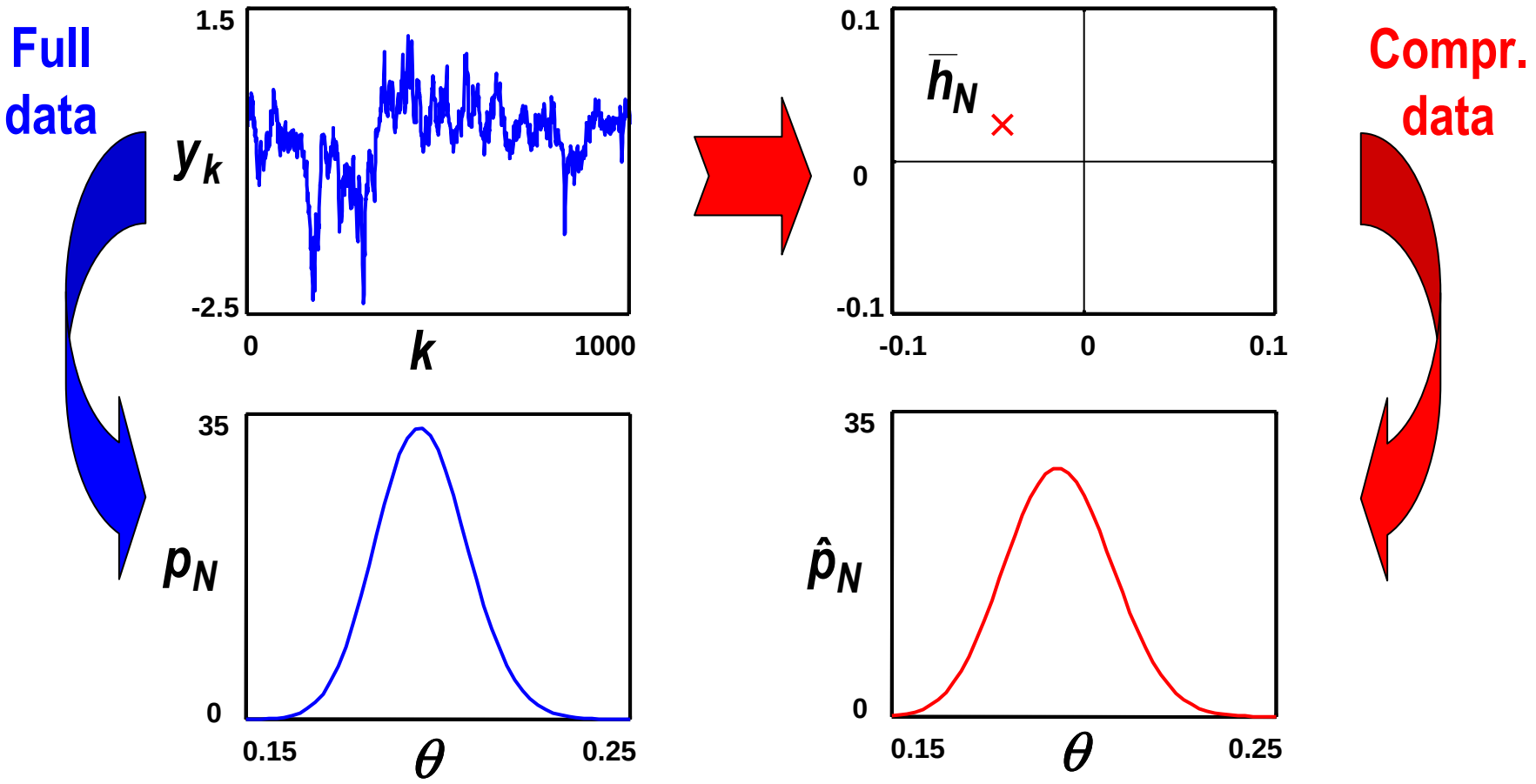




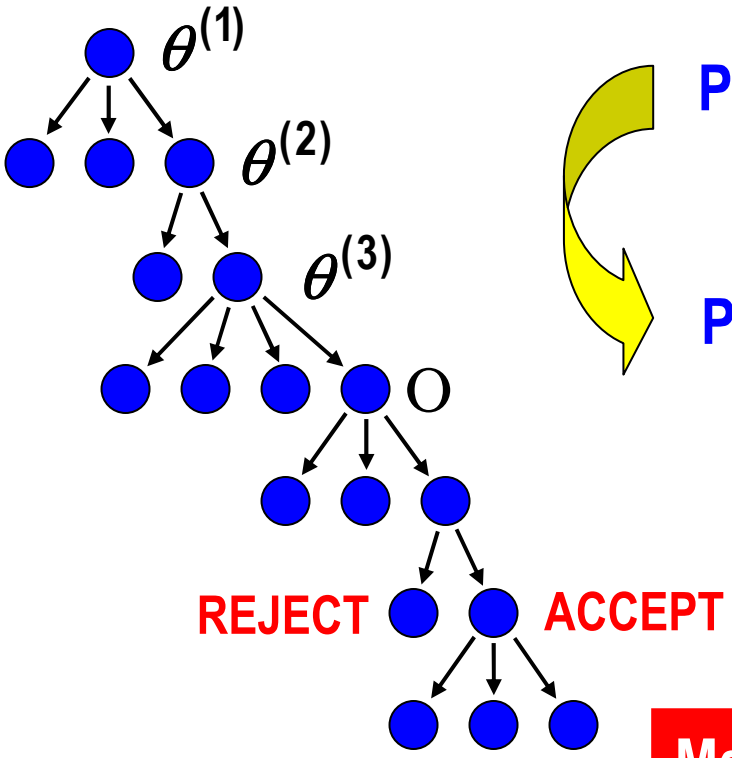


$$y = \sin(\theta z) + e$$

FINITE MEMORY ESTIMATION: Inaccuracy ➔ MRE



FINITE TIME ESTIMATION: Posterior Distribution → Sample



Posterior approximation

$$\hat{p}_N(\theta) = c\, p_0(\theta) \exp\left(-N D(\mathrm{R}_N \,\|\, s_\theta)\right)$$

Posterior sample

$$\theta^{(1)}, \theta^{(2)}, \mathrm{K}\ , \theta^{(M)} \sim \hat{p}_N(\theta)$$

Markov chain Monte Carlo

"There is little need for the traditional mathematical studies of properties of statistical techniques that dominate statistical journals.

There is a great need for better understanding of the principles of model construction, criticism, and revision, and for studies of sensitivity of inferences to model change as a function of features of the data."

– A. P. Dempster