On-line Nonlinear Estimation

Rudolf Kulhavý

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

Outline

- ➔ What's Wrong with 'Nonlinear Estimation'?
- ➔ Estimation as Probability Matching
- ➔ Three Approaches to Approximation
  - ➔ Locally Weighted Smoothing
  - ➔ Non-Iterative Monte Carlo Sampling
  - ➔ Iterative Monte Carlo Sampling
  - ➔ Restoration of Information Divergence
- ➔ Which Approximation?

‘Robust AR’ Example

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

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

μ is constant

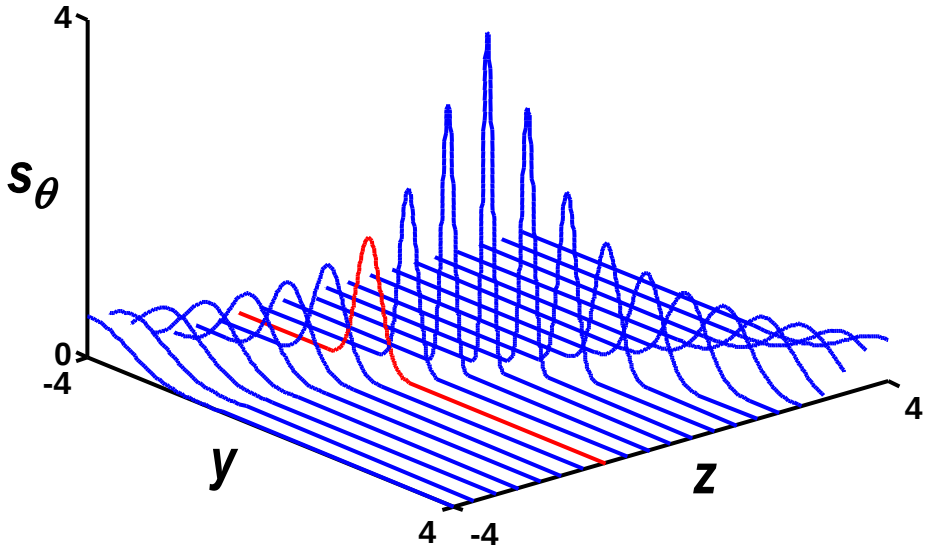
v_k is $N(0, \sigma_v^2)$ distributed

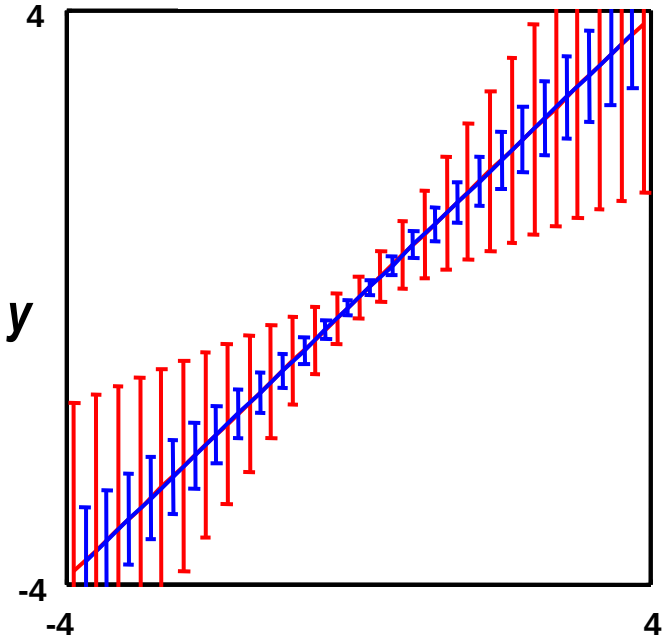
e_k is $N(0, \sigma_e^2)$ distributed

unknown parameters

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

Model Density





z-dependent variance: $\sigma^2(z) = \sigma_e^2 + z^2\,\sigma_v^2$

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

No Way to Compress Data!

likelihood

$$= \frac{1}{\sqrt{2\pi(\sigma_e^2 + z_1^2\sigma_v^2)}} \exp\left(-\frac{1}{2(\sigma_e^2 + z_1^2\sigma_v^2)}(y_1 - \mu(z_1))^2\right)$$

$$\times \frac{1}{\sqrt{2\pi(\sigma_e^2 + z_2^2\sigma_v^2)}} \exp\left(-\frac{1}{2(\sigma_e^2 + z_2^2\sigma_v^2)}(y_2 - \mu(z_2))^2\right)$$

$$\times \frac{1}{\sqrt{2\pi(\sigma_e^2 + z_3^2\sigma_v^2)}} \exp\left(-\frac{1}{2(\sigma_e^2 + z_3^2\sigma_v^2)}(y_3 - \mu(z_3))^2\right)$$

$$\Lambda$$

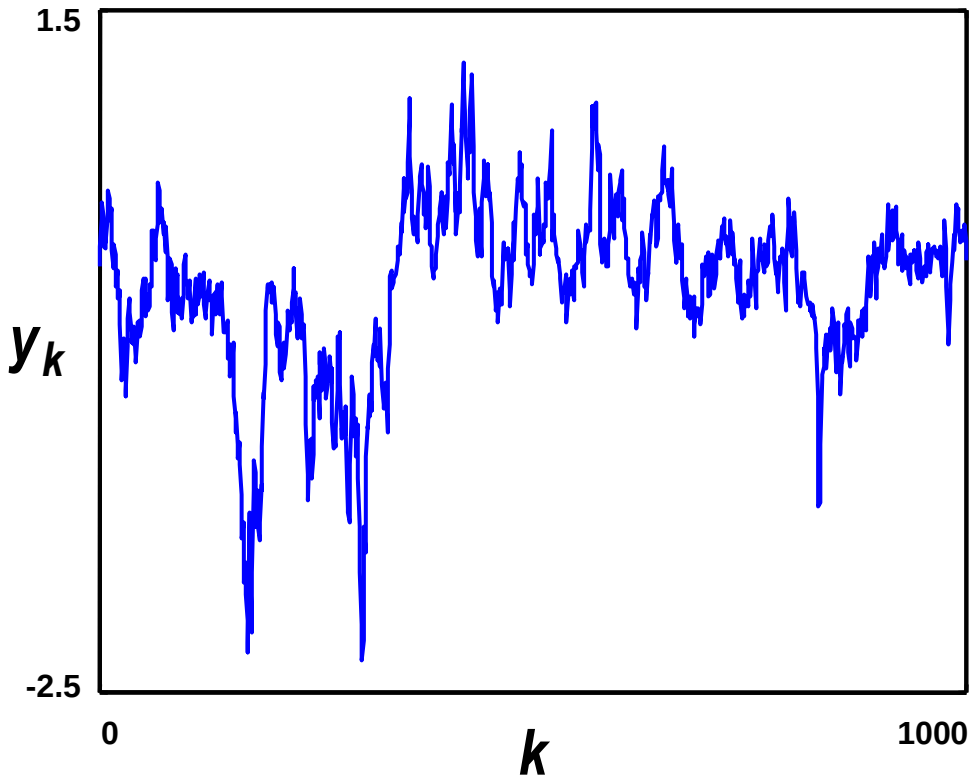
Problem

Statistical theory
does NOT count on
computational complexity

Outline

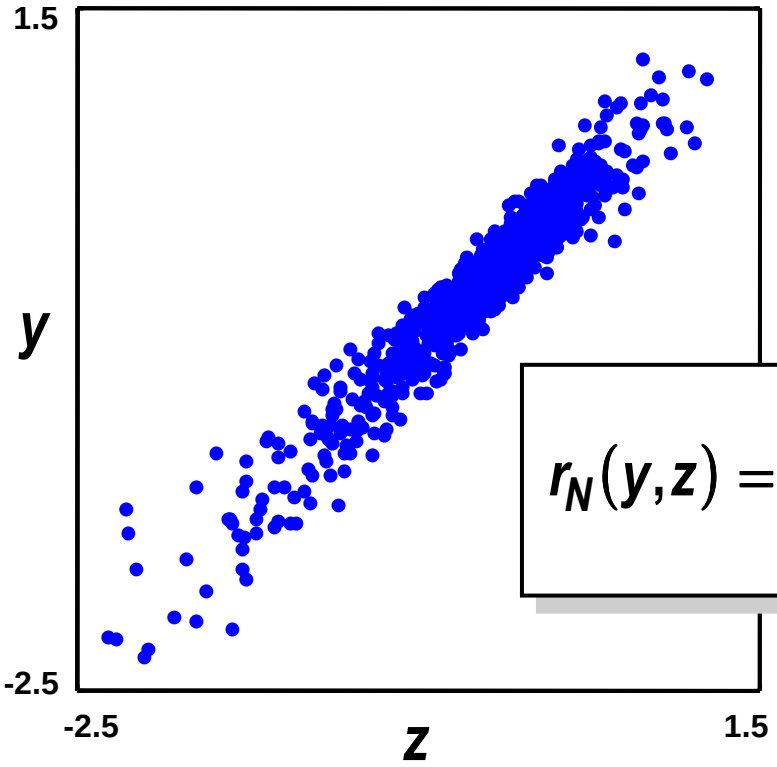
- ➔ What's Wrong with 'Nonlinear Estimation'?
- ➔ Estimation as Probability Matching
- ➔ Three Approaches to Approximation
  - ➔ Locally Weighted Smoothing
  - ➔ Non-Iterative Monte Carlo Sampling
  - ➔ Iterative Monte Carlo Sampling
  - ➔ Restoration of Information Divergence
- ➔ Which Approximation?

Sample of Data



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

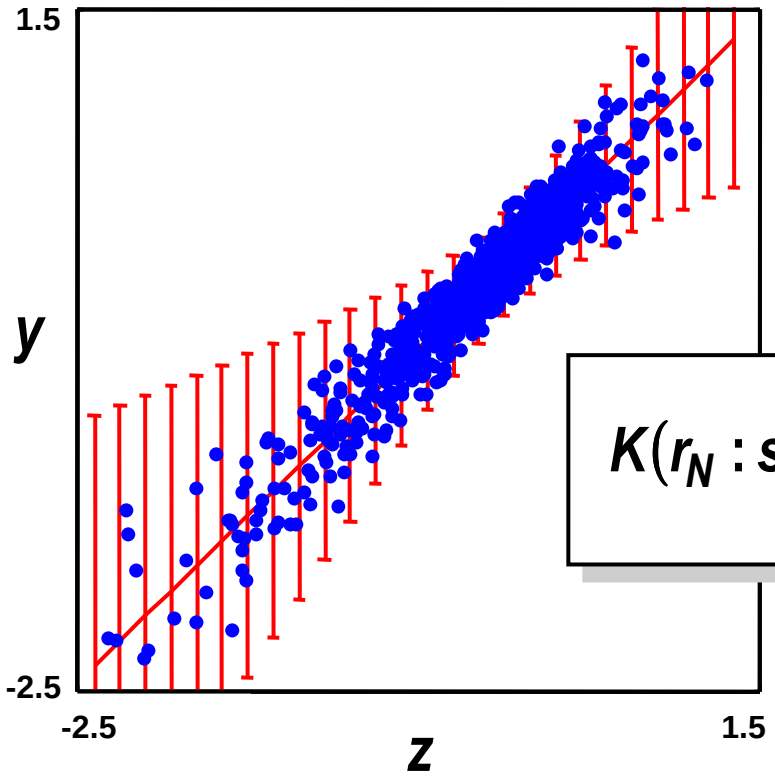
Empirical Density



Mixture of
Dirac functions

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

Probability Matching

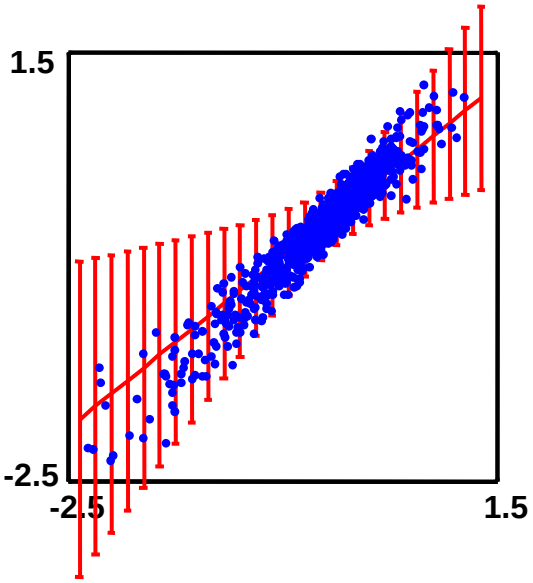


Conditional inaccuracy

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

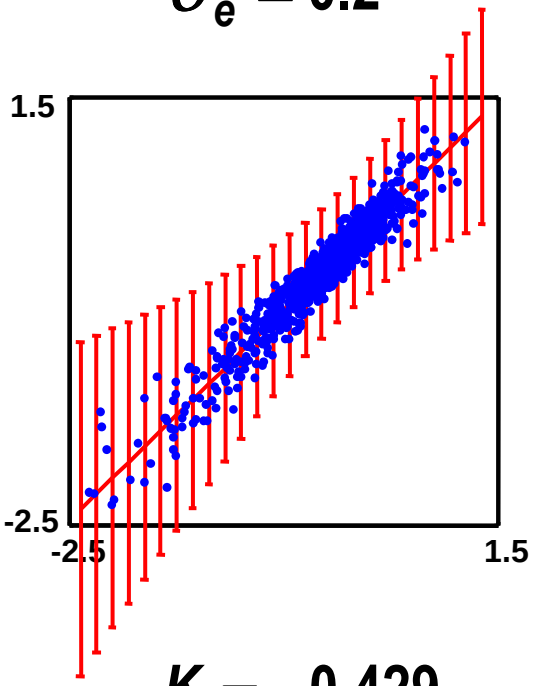
Minimum Inaccuracy Estimation

$\mu = 0.8$



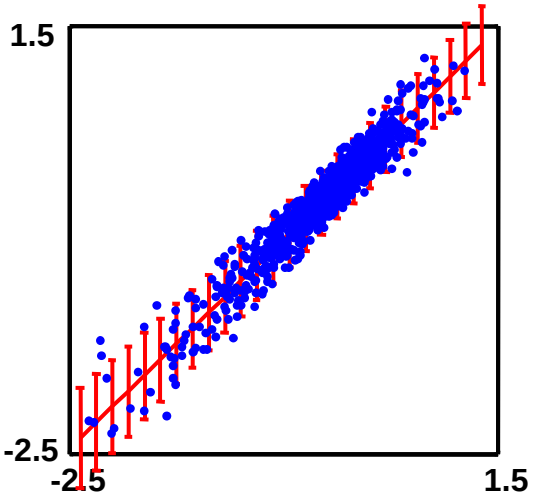
$K = -0.552$

$\sigma_e = 0.2$



$K = -0.429$

$\sigma_v = 0.05$



$K = -0.517$

Inaccuracy

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

Likelihood

$$I_N(\theta) = c \exp(-N K(r_N : s_\theta))$$

Posterior

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

Outline

- ➔ What's Wrong with 'Nonlinear Estimation'?
- ➔ Estimation as Probability Matching
- ➔ Three Approaches to Approximation
  - ➔ Locally Weighted Smoothing
  - ➔ Non-Iterative Monte Carlo Sampling
  - ➔ Iterative Monte Carlo Sampling
  - ➔ Restoration of Information Divergence
- ➔ Which Approximation?

Three Objects - Three Options

❶ Empirical density

$r_N(y,z)$

❷ Model family

$\{ s_\theta(y \mid z) : \theta \in \boldsymbol{\Gamma} \}$

❸ Inaccuracy

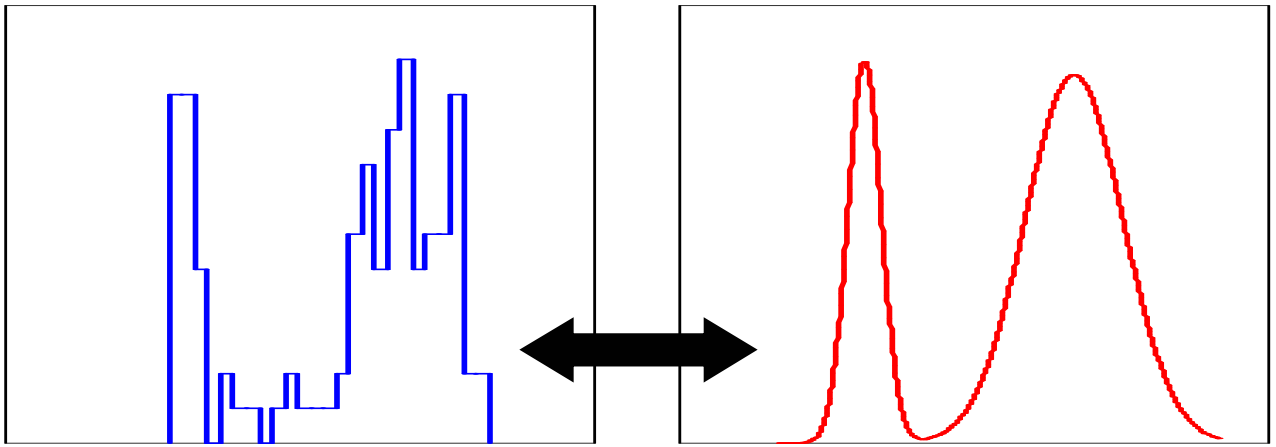
$K(r_N : s_\theta)$

- Replacement of the empirical distribution with its locally-weighted version
- Reduction of the model family to a finite model set
- Restoration of the information divergence between the empirical and model distributions from compressed data

‘Old Faithful Geyser’ Example

❶ Empirical density

❷ Model density



❸ Inaccuracy

Amount of data

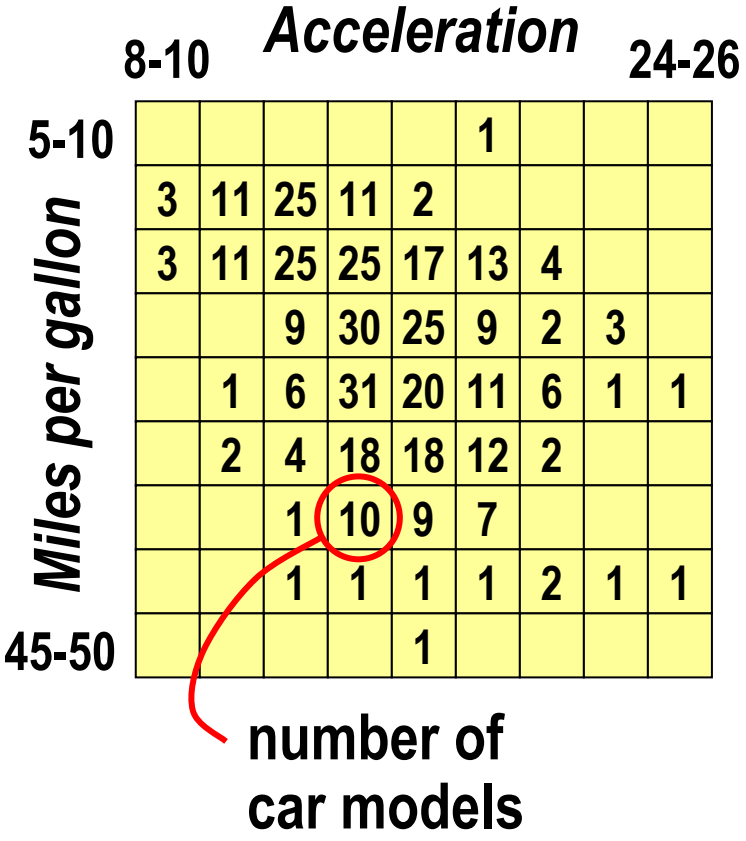
Computational complexity

Model complexity

Outline

- ➔ What's Wrong with 'Nonlinear Estimation'?
- ➔ Estimation as Probability Matching
- ➔ Three Approaches to Approximation
  - ➔ Locally Weighted Smoothing
  - ➔ Non-Iterative Monte Carlo Sampling
  - ➔ Iterative Monte Carlo Sampling
  - ➔ Restoration of Information Divergence
- ➔ Which Approximation?

‘Car Market’ Example

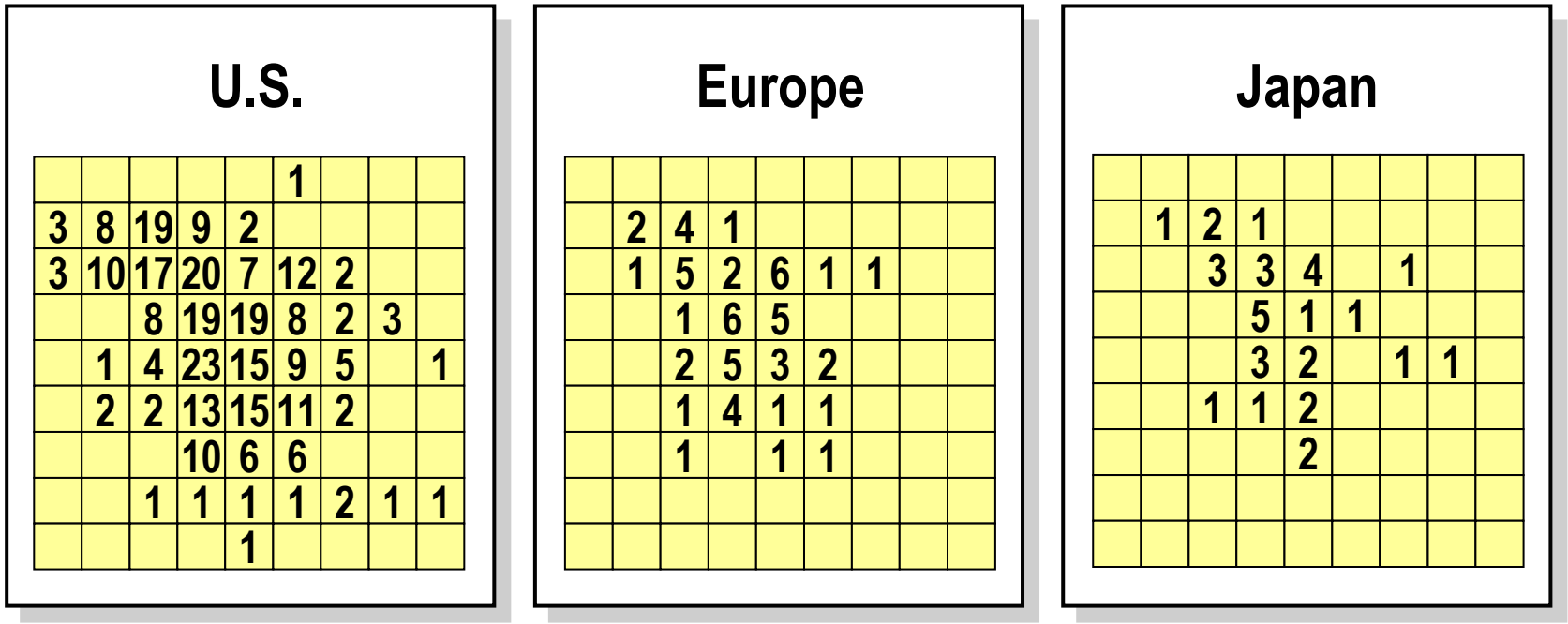


5-10						1			
	3	11	25	11	2				
	3	11	25	25	17	13	4		
			9	30	25	9	2	3	
		1	6	31	20	11	6	1	1
		2	4	18	18	12	2		
			1	10	9	7			
			1	1	1	1	2	1	1
45-50					1				

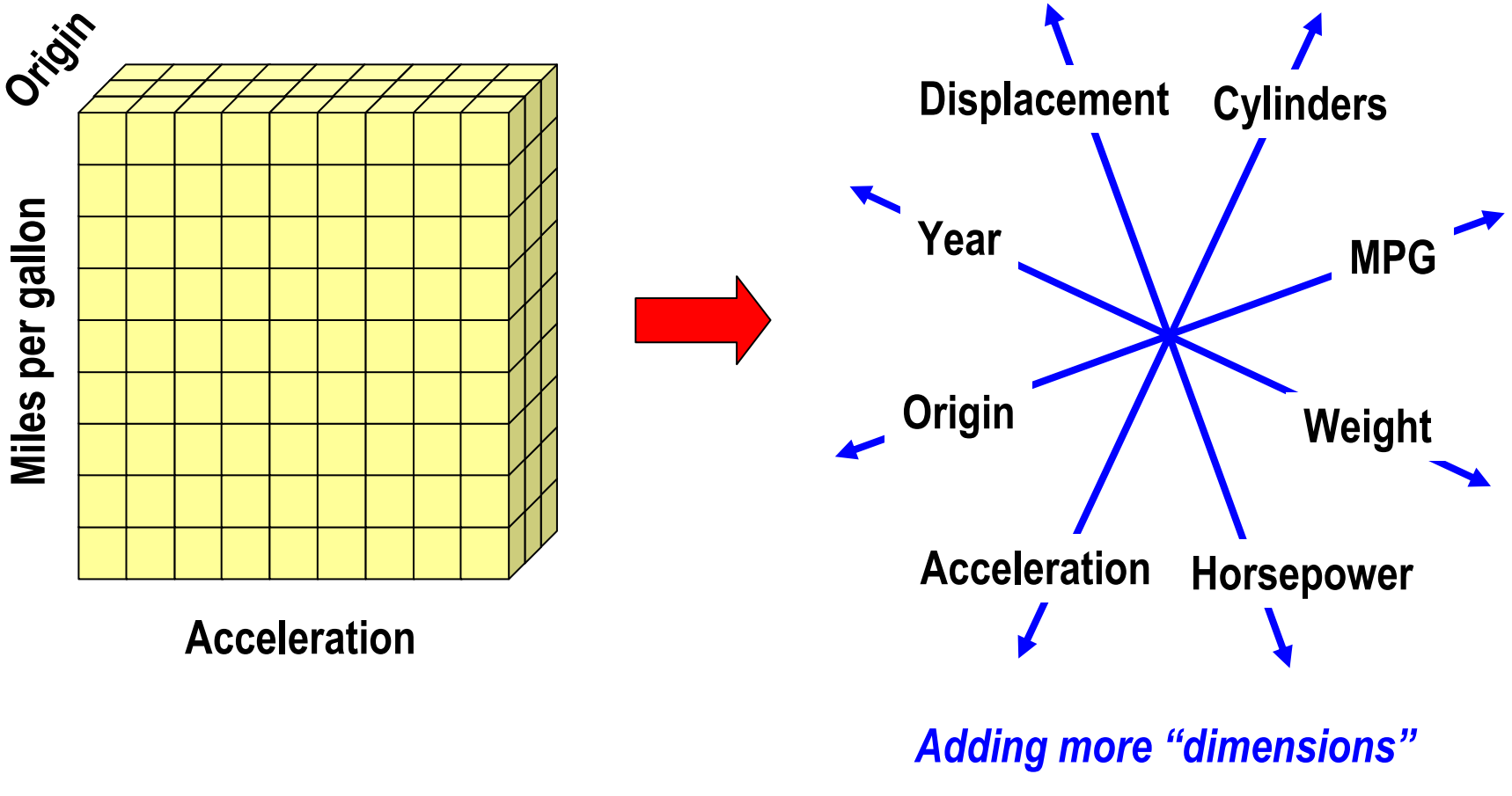
Data Cube

DBMS for storing multidimensional data and handling queries that aggregate over some dimensions

Disaggregating Data

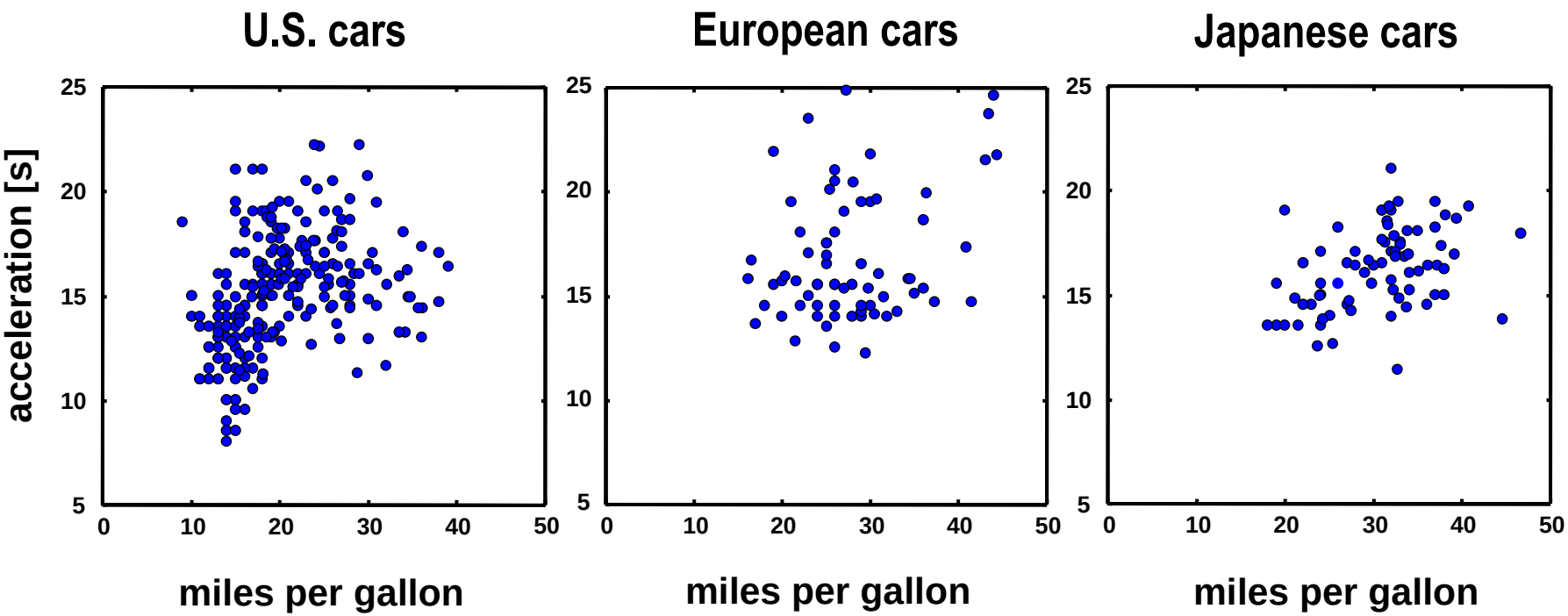


Building Up Data Cube

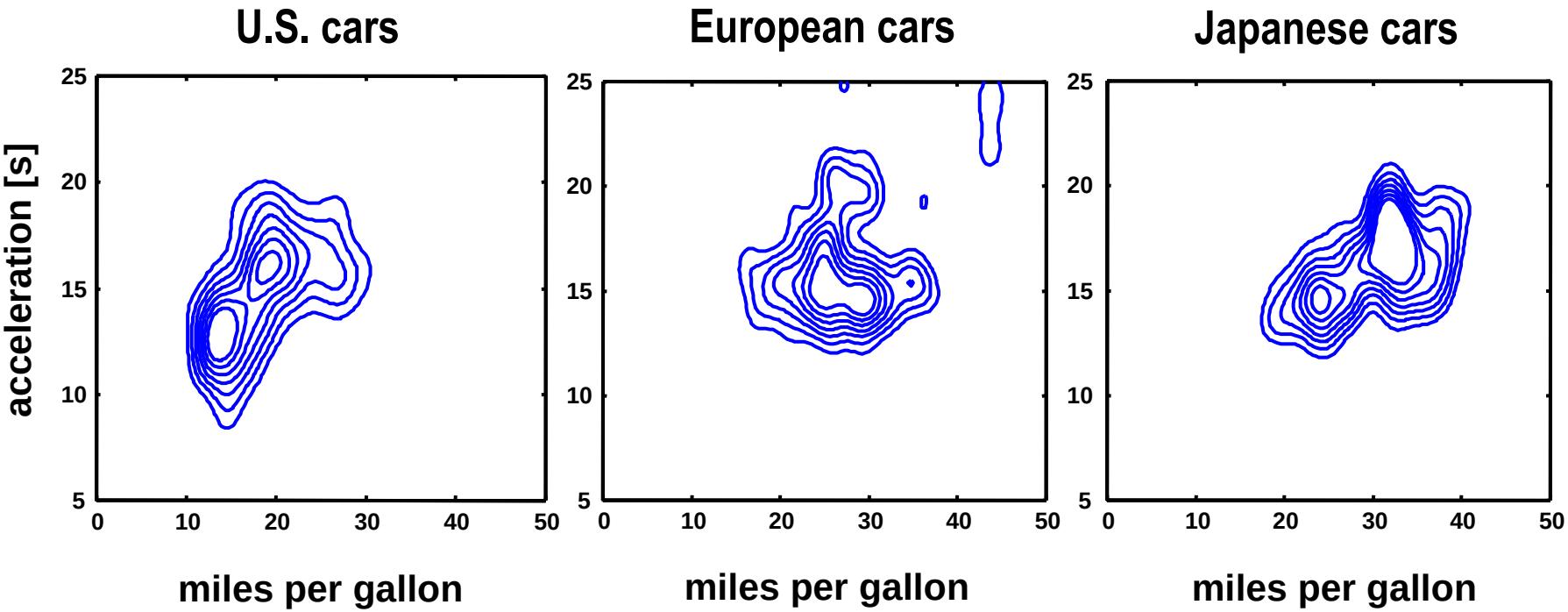


Adding more "dimensions"

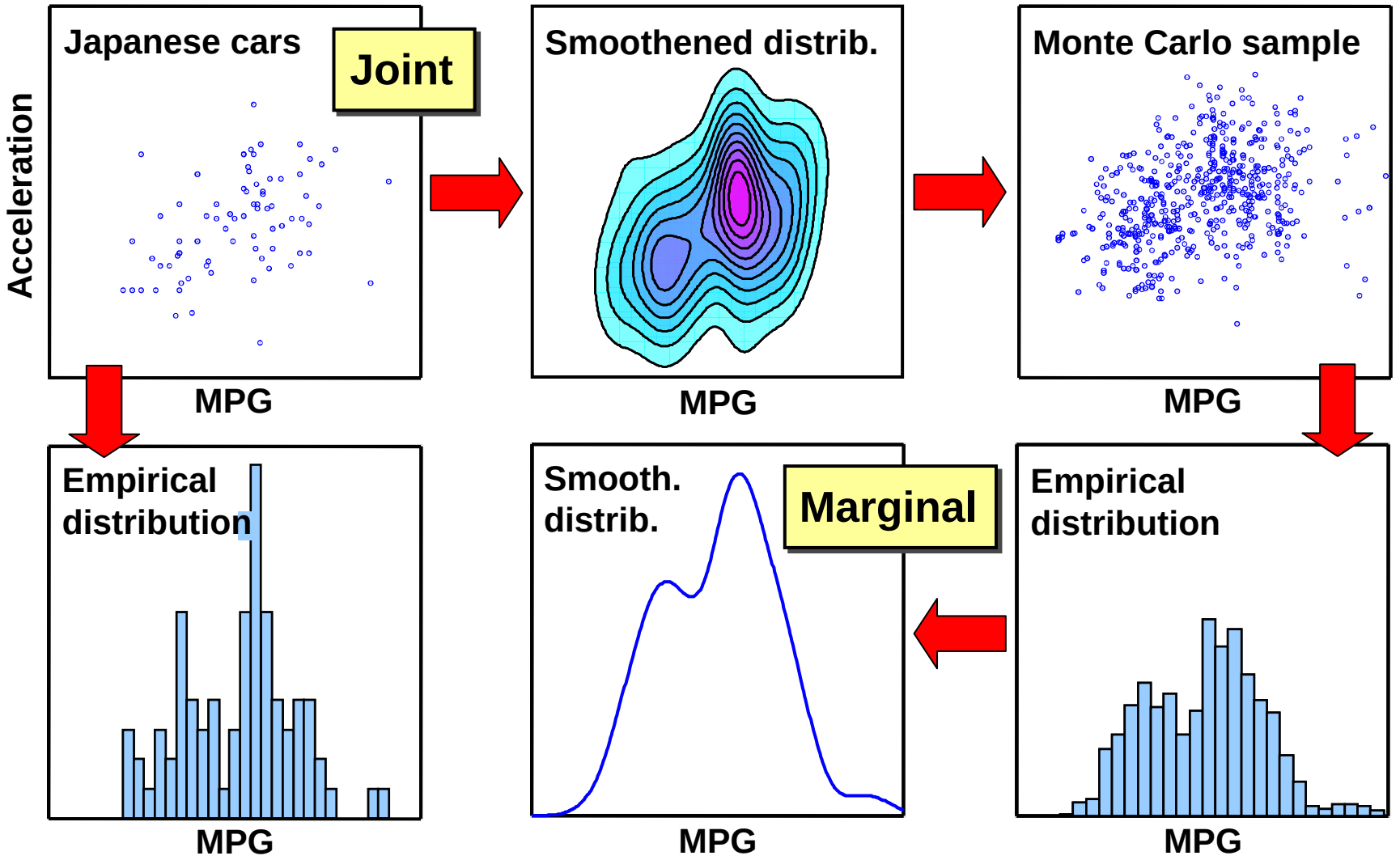
Empirical Distribution of Data



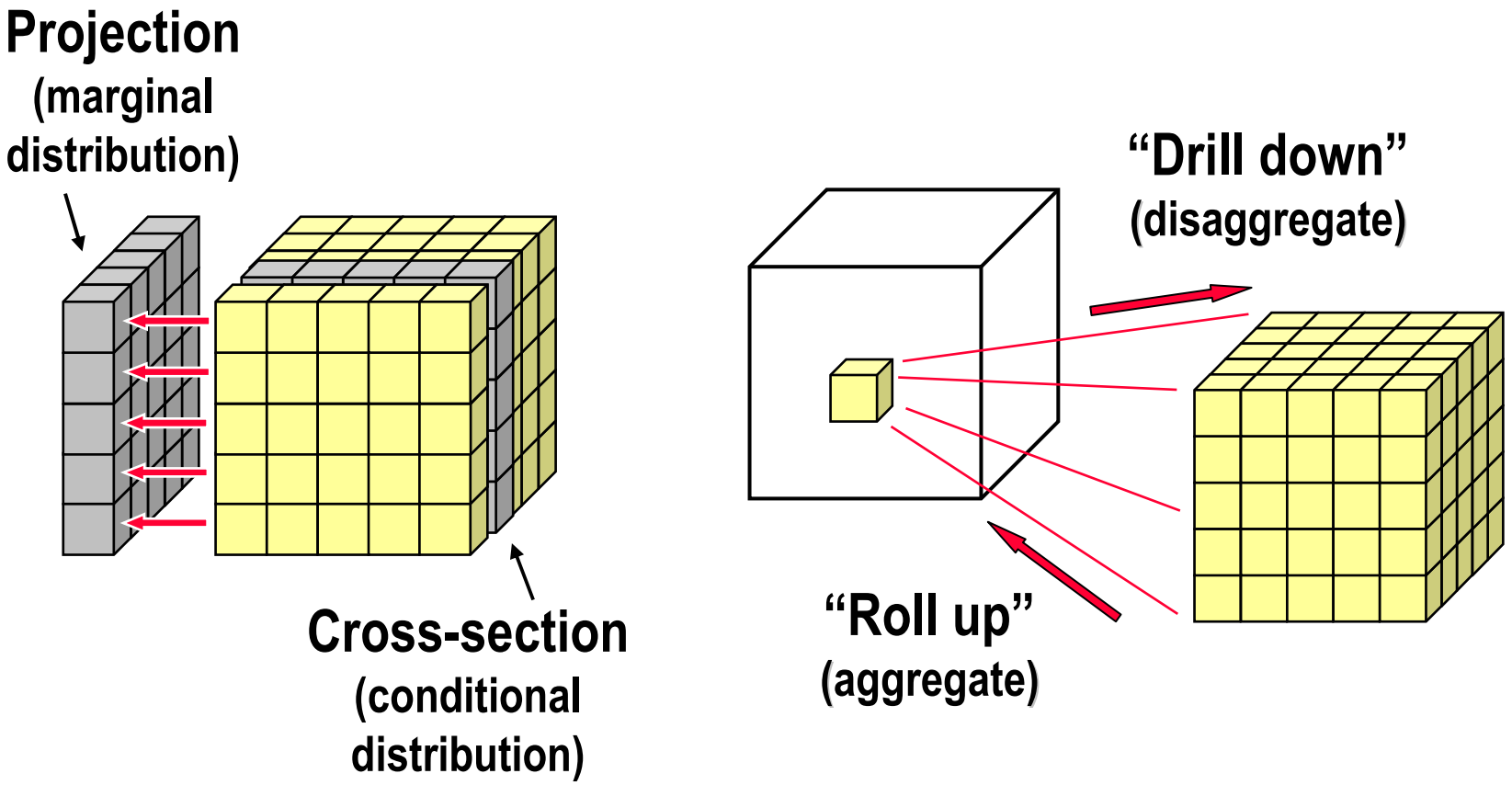
Smoothened Distribution of Data



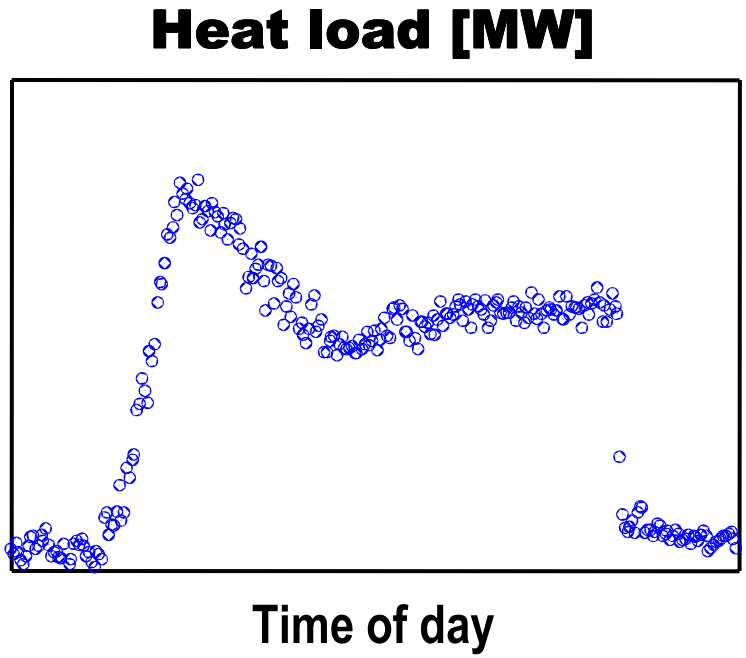
Monte Carlo Computations

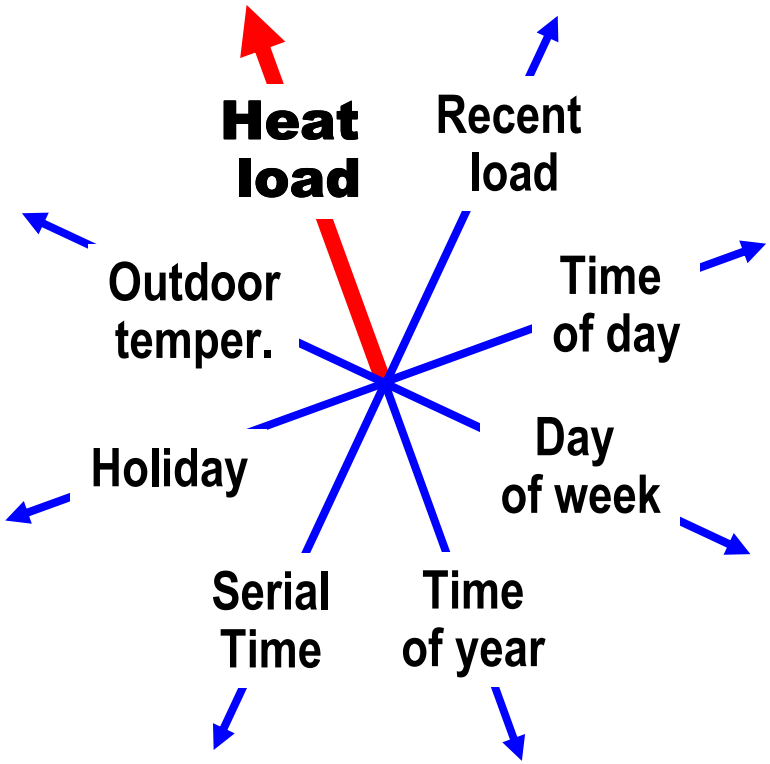


Data Cube ⇔ Empirical Prob.

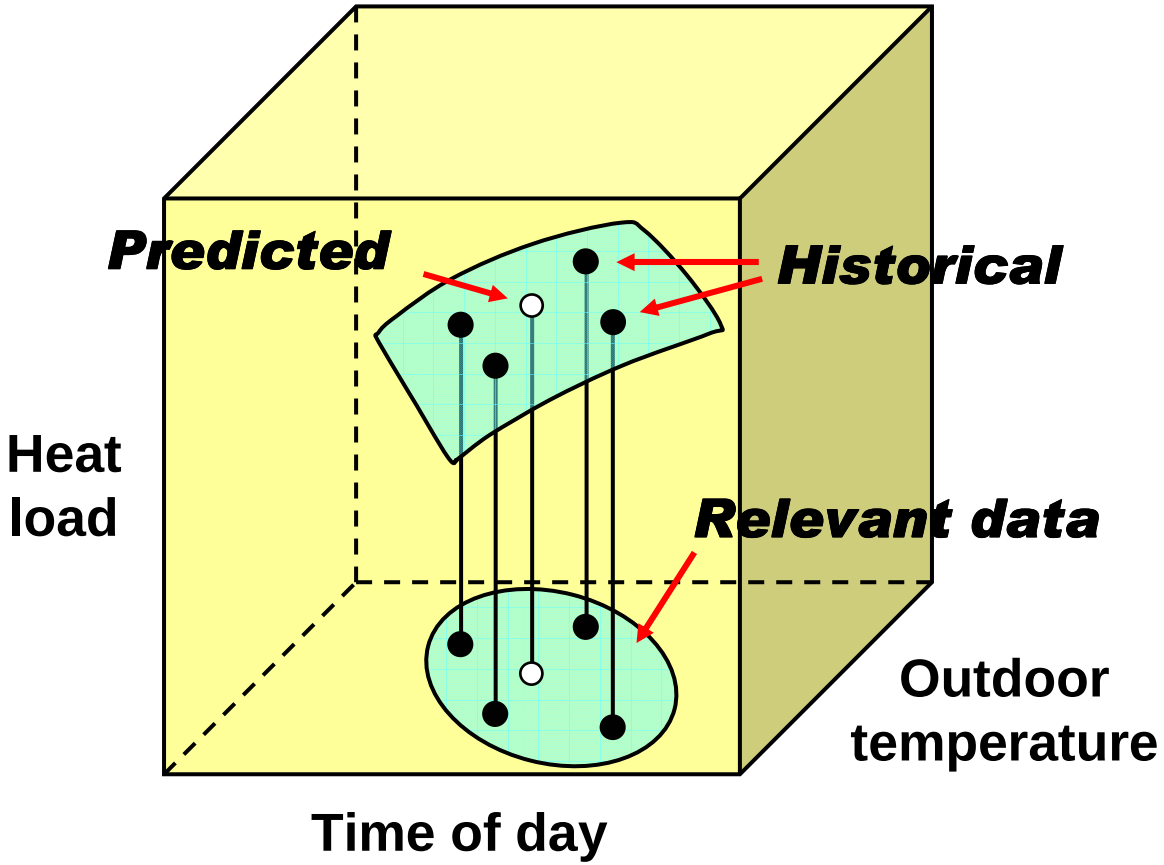


‘Heat Load’ Example





Locally-Weighted Smoothing



Weight on Data

$$K(\|z_k - z^*\|_H)$$

Euclidean distance

Kernel function

$$K(x) = c\ \exp(-x'x)$$

$$K(x) = c\ \min\{1 - x'x, 0\}$$

Weighted Empirical Density

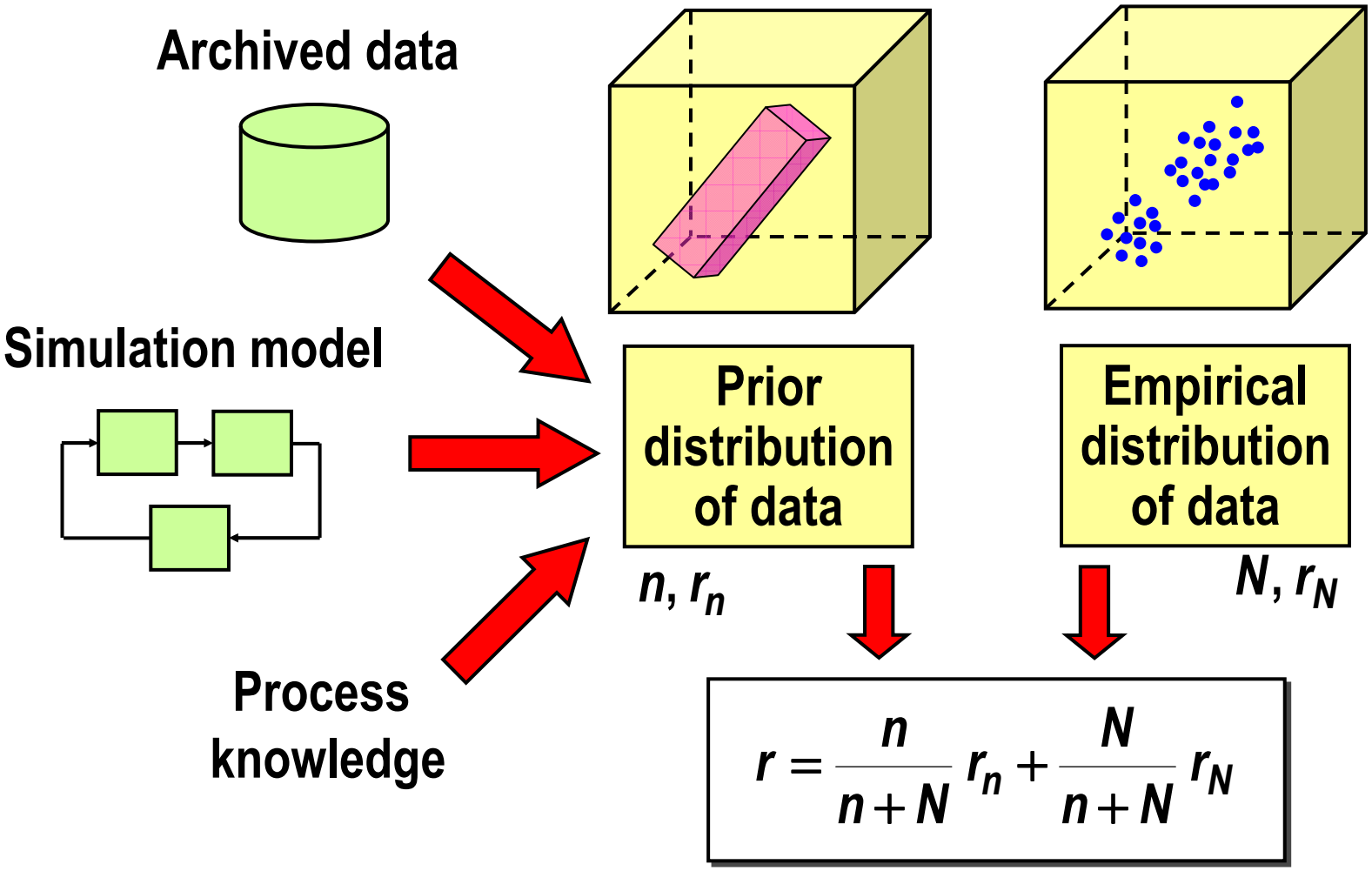
$$\nu_N = \sum_{k=m+1}^{N+m} w_k$$

$$w_k = K(\| z_k - z^* \|_H)$$

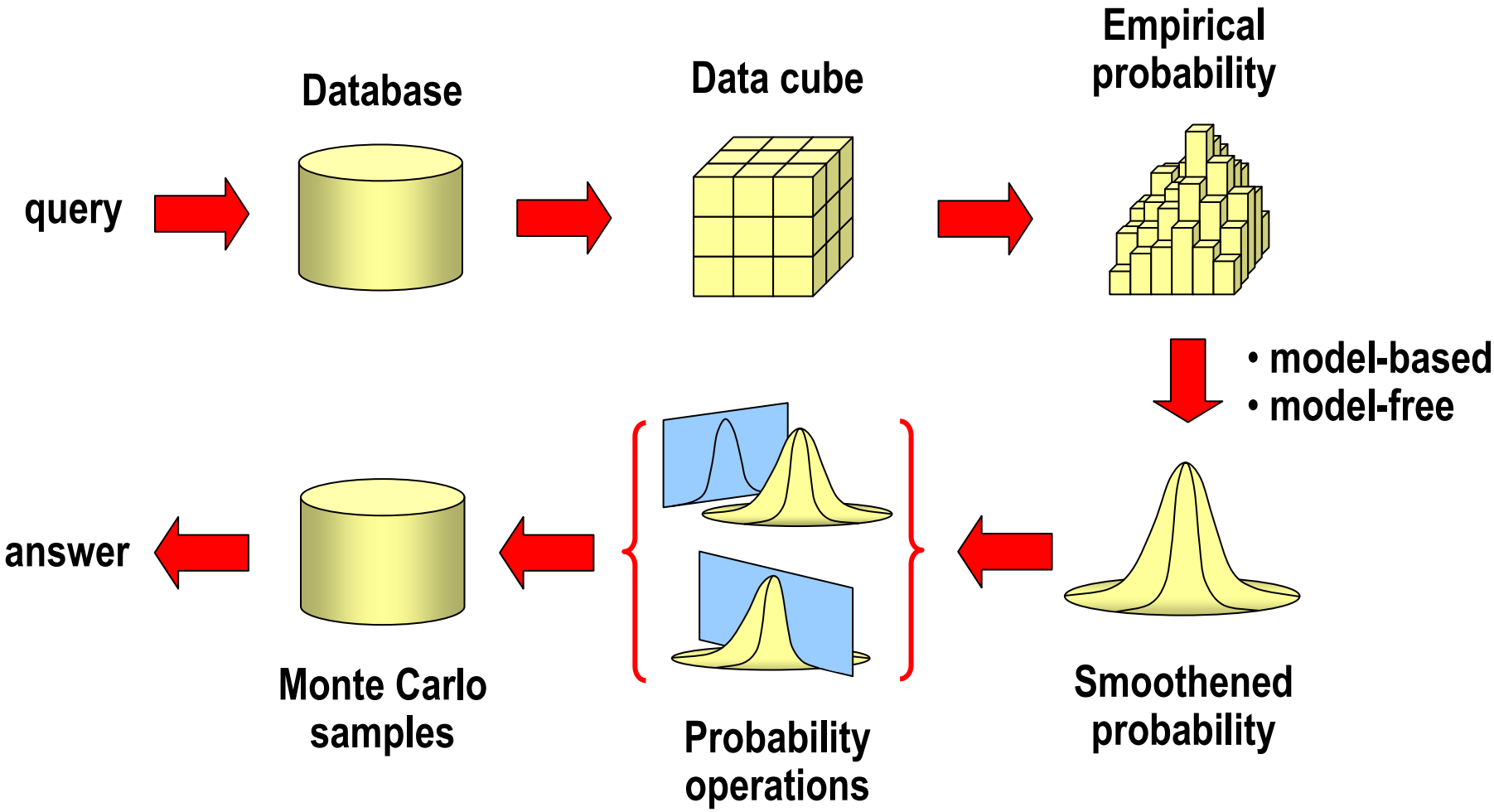
$$p_N(\theta) = c\, p_0(\theta) \exp(-\nu_N K(\rho_N : s_\theta))$$

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

Use of Prior Knowledge



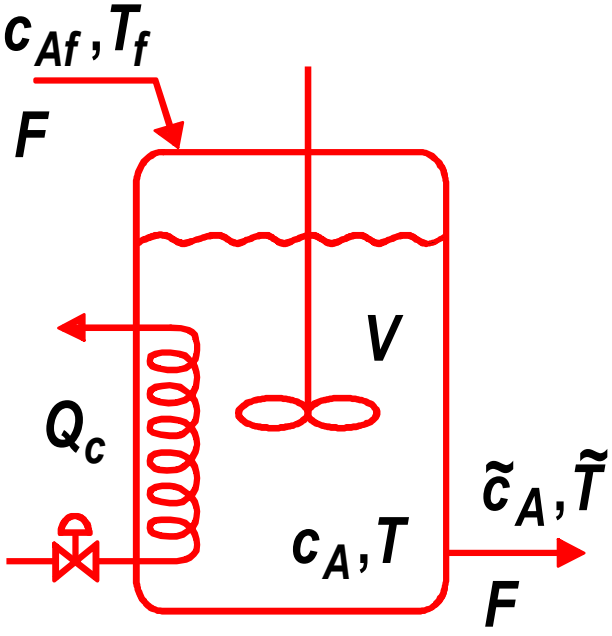
Probabilistic Data Mining



Outline

- ➔ What's Wrong with 'Nonlinear Estimation'?
- ➔ Estimation as Probability Matching
- ➔ Three Approaches to Approximation
  - ➔ Locally Weighted Smoothing
  - ➔ Non-Iterative Monte Carlo Sampling
  - ➔ Iterative Monte Carlo Sampling
  - ➔ Restoration of Information Divergence
- ➔ Which Approximation?

‘Chemical Reactor’ Example



Can we estimate c_A and T from the measurements?

Nonlinear CSTR model

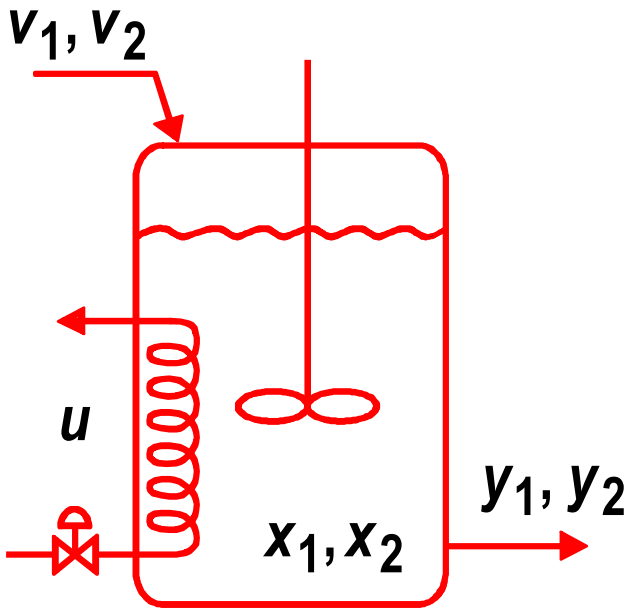
$$\frac{dc_A}{dt} = -\frac{1}{\theta} c_A - k(T)\, c_A + \frac{1}{\theta} c_{Af}$$

$$\frac{dT}{dt} = -\frac{1}{\theta} T - \beta\, k(T)\, c_A + \frac{1}{\theta} T_f - \chi$$

Reaction rate (Arrhenius relation)

$$k(T) = k_0 \exp\left(-\frac{E}{RT}\right)$$

State-Space Model



$$x_{k+1} = f_k(x_k, u_k, v_k)$$
$$y_k = g_k(x_k, e_k)$$

State equations (process model)

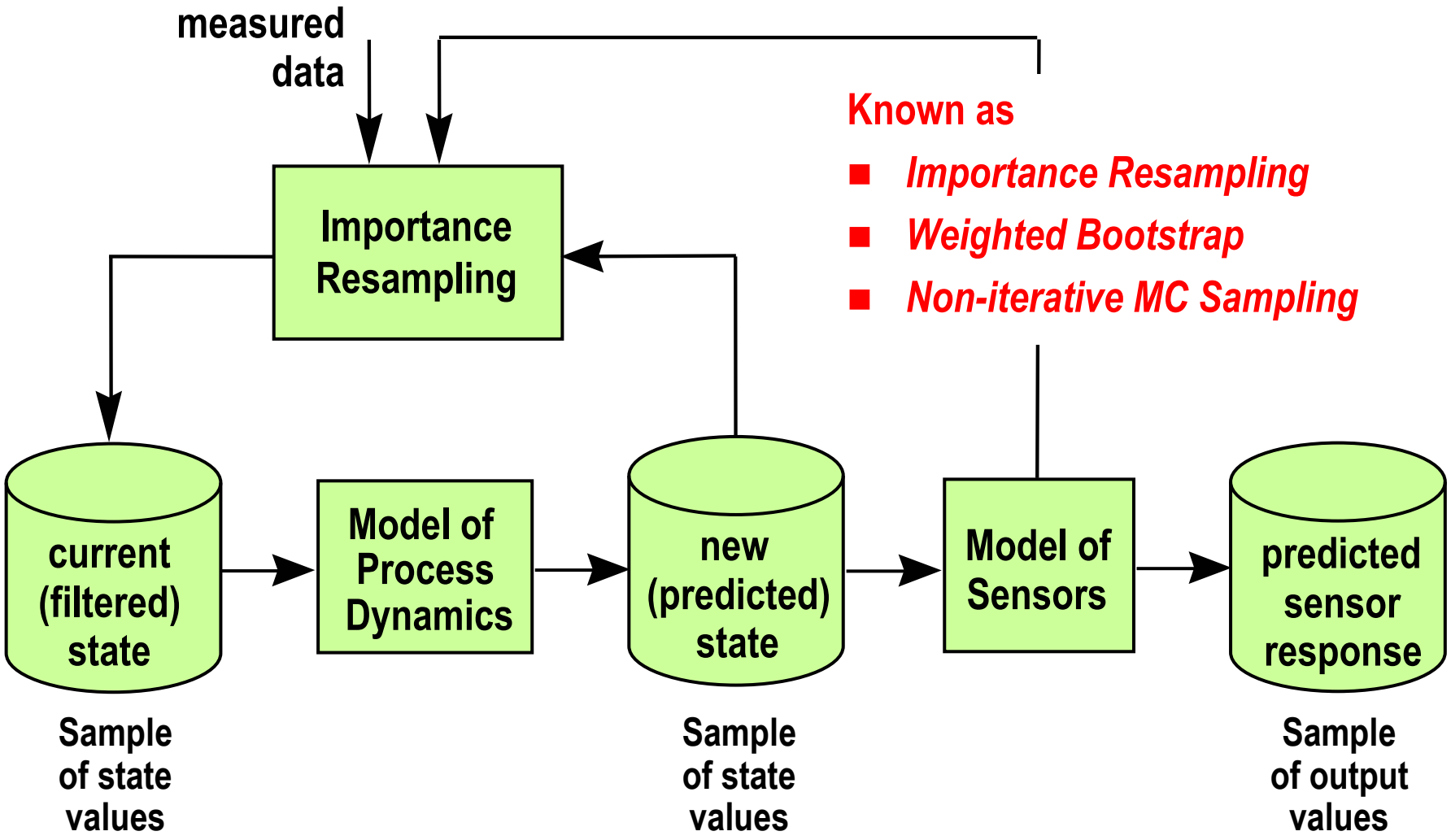
$$\frac{dx_1}{dt} = -\frac{1}{\theta}x_1 - k(x_2)\,x_1 + \frac{1}{\theta}d_1$$

$$\frac{dx_2}{dt} = -\frac{1}{\theta}x_2 - \beta\,k(x_2)\,x_1 + \frac{1}{\theta}d_2 + f(u)$$

Output equations (sensors model)

$$\left.\begin{aligned} y_1 &= x_1 + e_1 \\ y_2 &= x_2 + e_2 \end{aligned}\right\}$$ Measurement noises

Idea of Stochastic Simulation



Known as

- *Importance Resampling*
- *Weighted Bootstrap*
- *Non-iterative MC Sampling*

A. Data Update

A1. Calculate normalized weights

$$w_i = \frac{p(y_k \mid x_k^{(i)})}{\sum_{j=1}^{M} p(y_k \mid x_k^{(j)})}$$

A2. Resample *M*-times from

$$x_k^{(1)}, x_k^{(2)}, \mathrm{K}\ , x_k^{(M)}$$

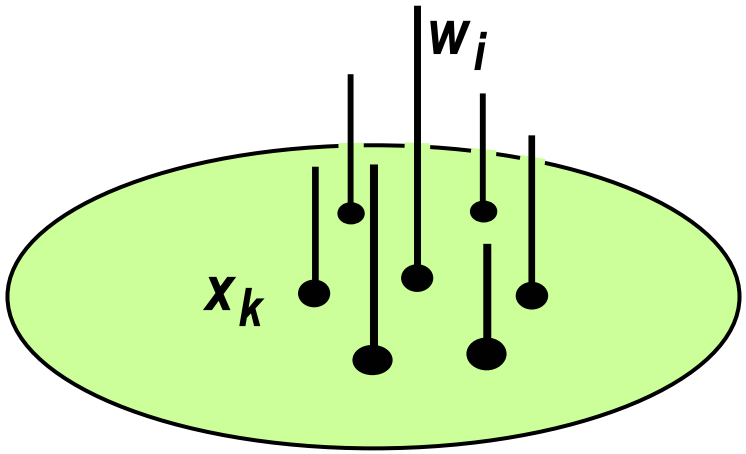
with probabilities

$$w_1, w_2, \mathrm{K}\ , w_M$$

A3. Perturb the points

$$x_{k+1}^{(i)} = x_k^{(j)} + \text{noise}$$

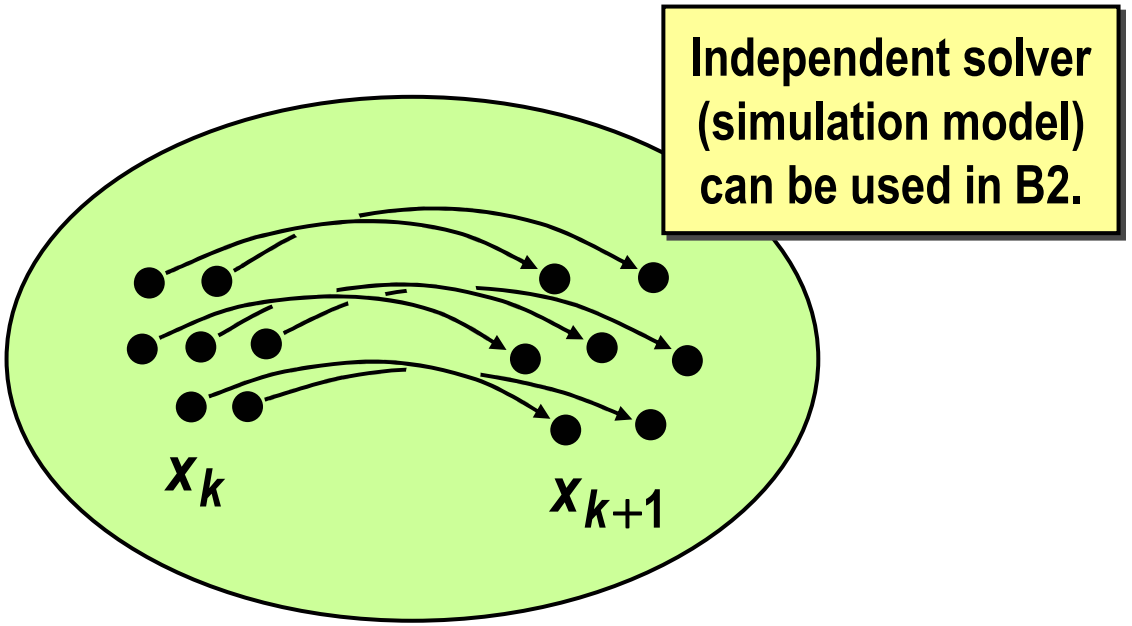
The state values $x_k^{(i)}$ that are not likely to yield the measurement y_k have little chance to appear again in the sample.



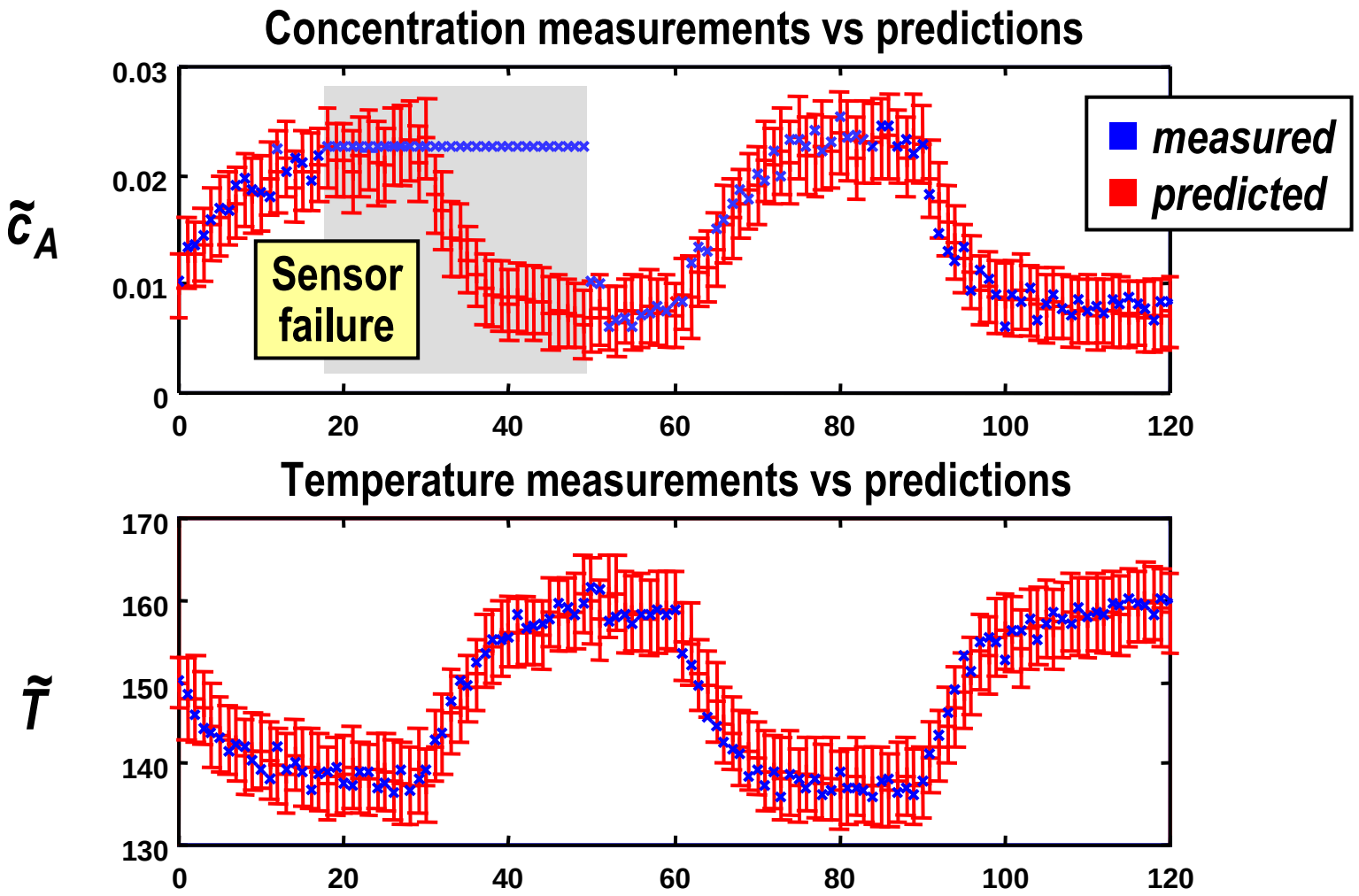
B. Time Update

B1. Generate noise samples $v_k^{(i)}$ for $i = 1, \mathrm{K}\ , M$

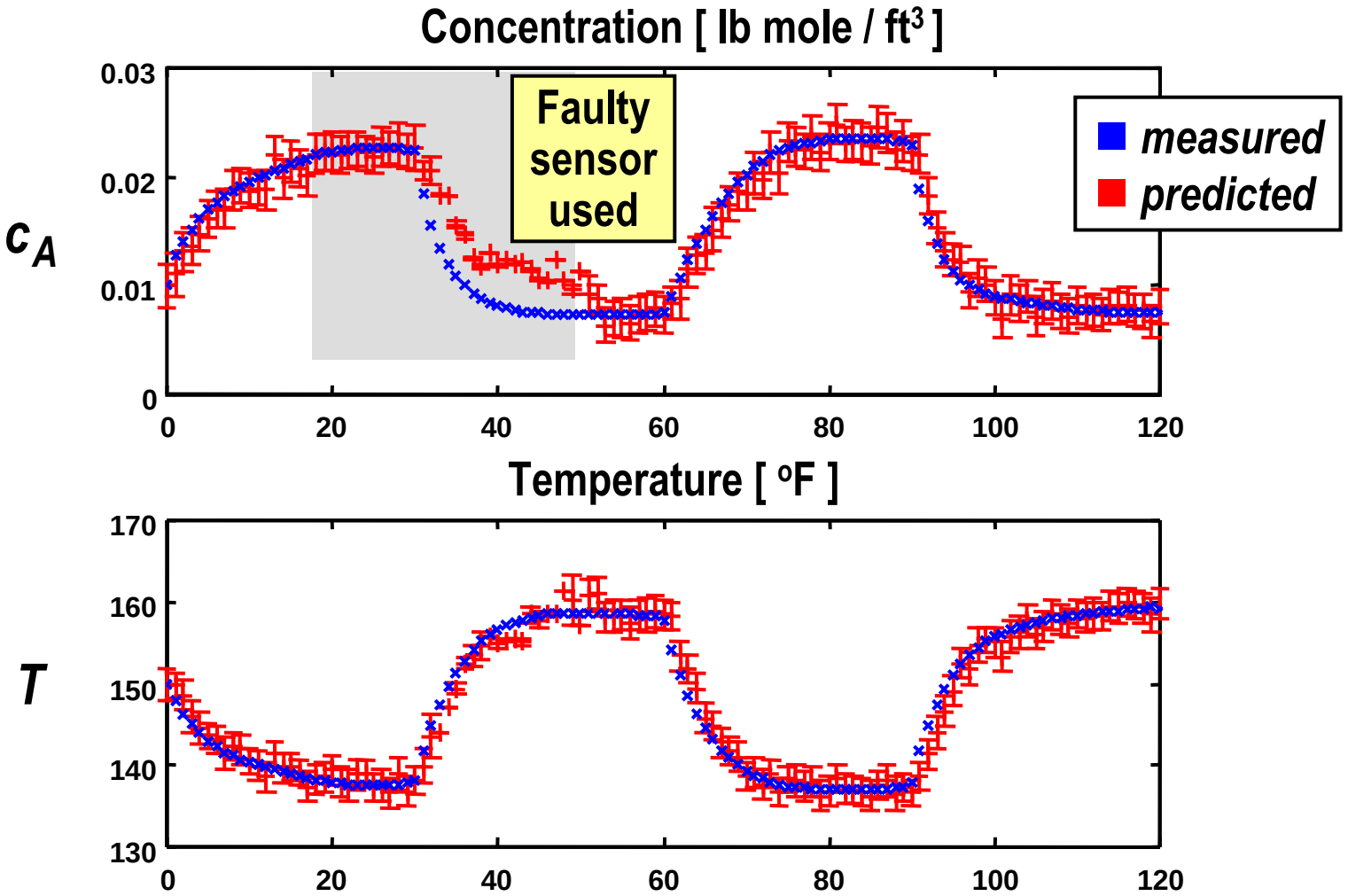
B2. Propagate $x_{k+1}^{(i)} = f_k(x_k^{(i)}, u_k, v_k^{(i)})$ for $i = 1, \mathrm{K}\ , M$



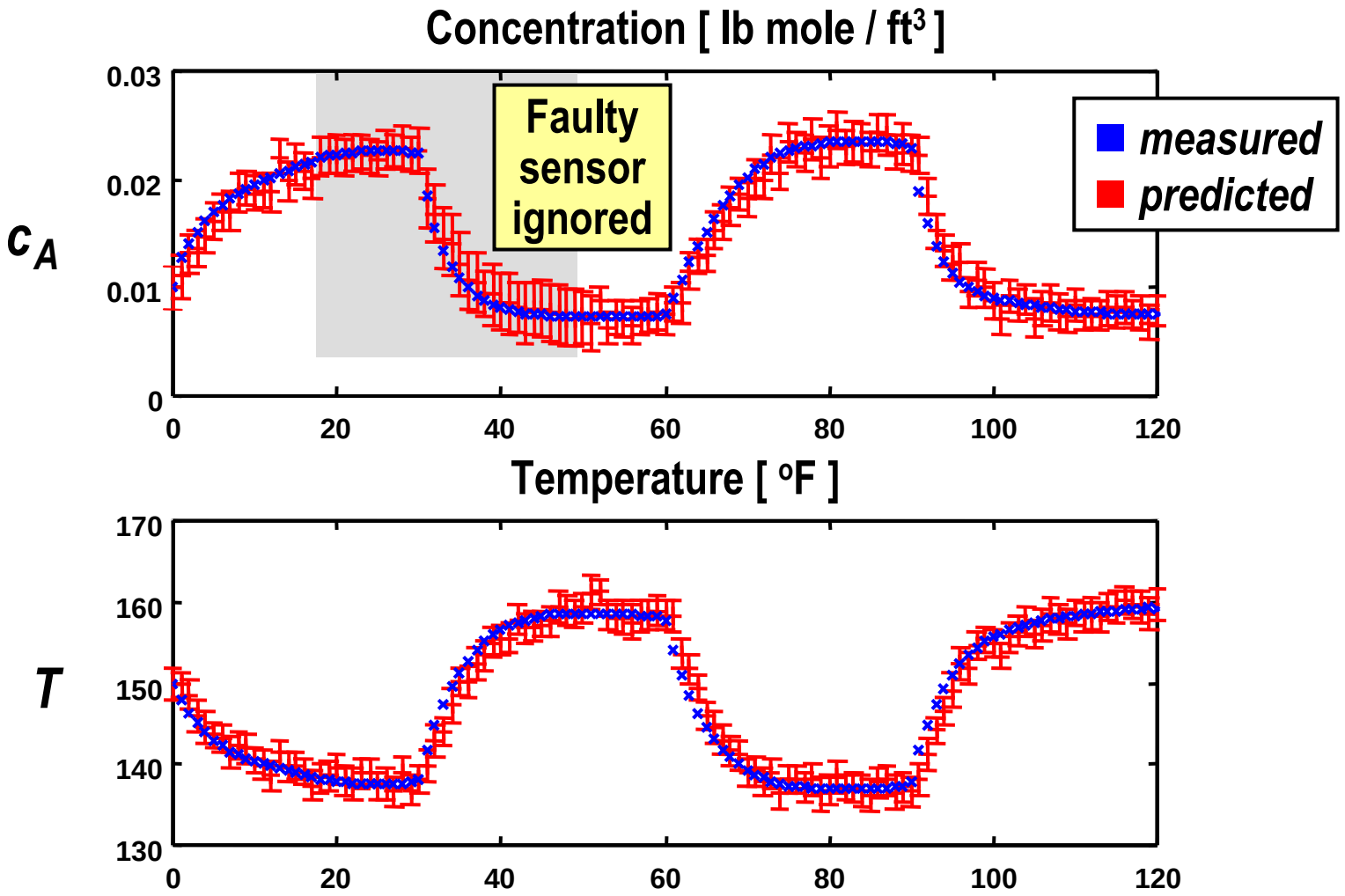
Measurement Prediction



Sensor Validation Off



Sensor Validation On



Outline

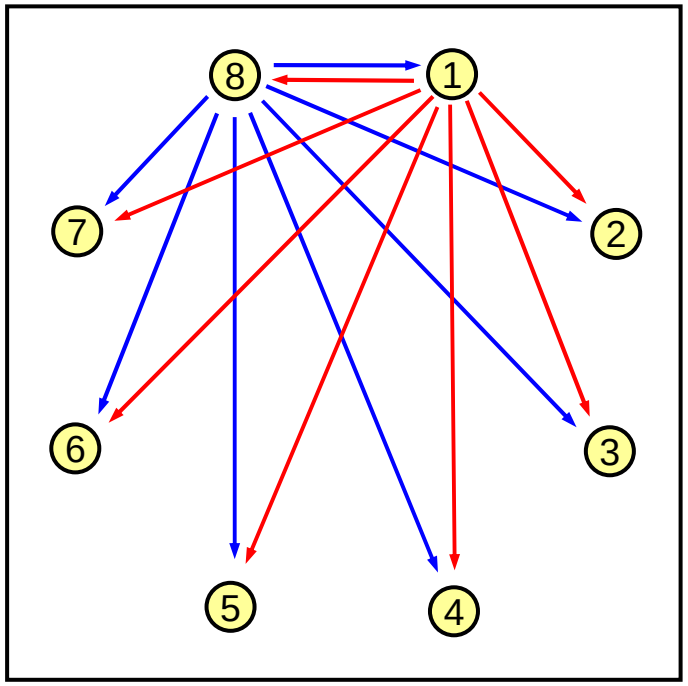
- ➔ What's Wrong with 'Nonlinear Estimation'?
- ➔ Estimation as Probability Matching
- ➔ Three Approaches to Approximation
  - ➔ Locally Weighted Smoothing
  - ➔ Non-Iterative Monte Carlo Sampling
  - ➔ Iterative Monte Carlo Sampling
  - ➔ Restoration of Information Divergence
- ➔ Which Approximation?

‘Markov Chain’ Monte Carlo

Markov chain simulation

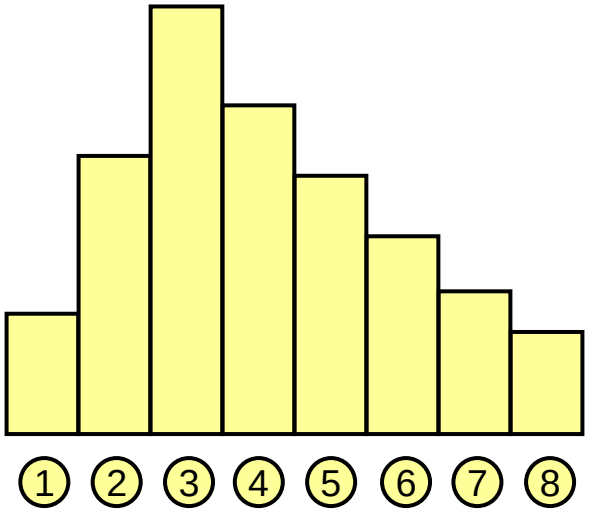
replaces

direct sampling



Design Markov chain (transition probabilities)

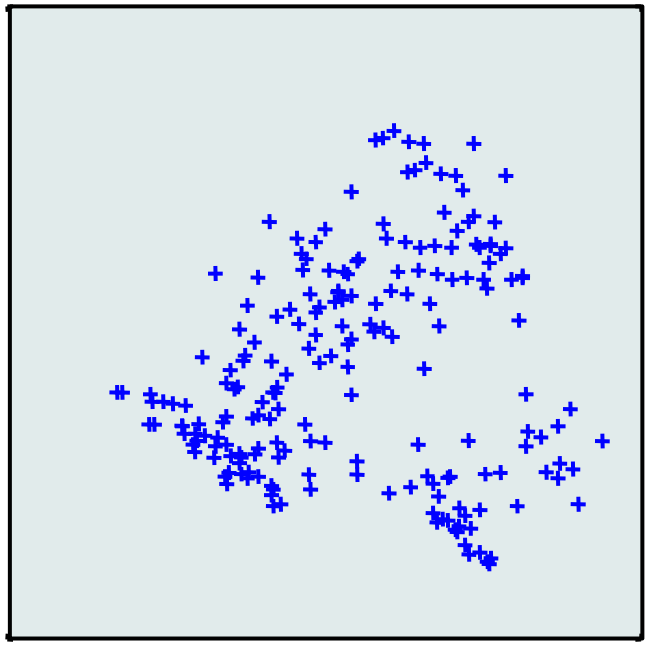
so that



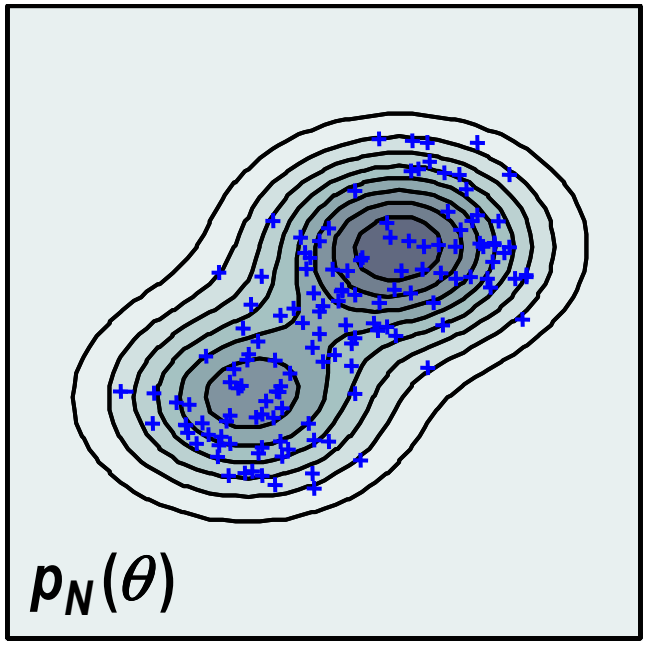
its stationary distribution coincides with the target distribution.

Metropolis Sampler

Random walk $\theta^* = \theta^{(i)} + v$



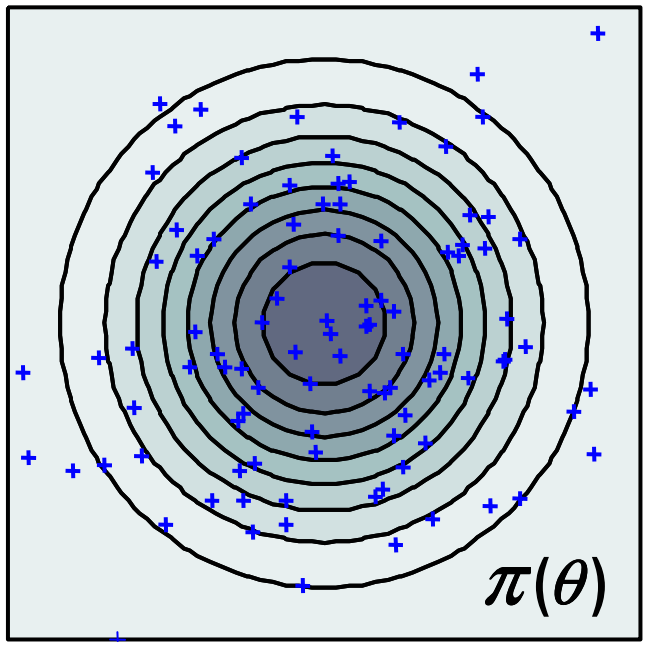
Accept $\theta^{(i+1)} = \theta^*$ w.p. α



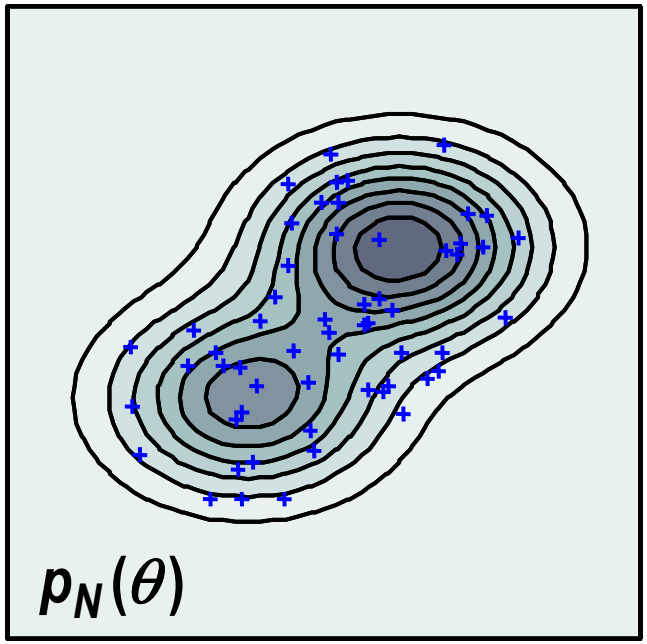
$$\alpha = \min\left\{\frac{p_N(\theta^*)}{p_N(\theta^{(i)})}, 1\right\}$$

Metropolis-Hastings Sampler

Sample θ^* from $\pi(\theta)$



Accept $\theta^{(i+1)} = \theta^*$ w.p. α



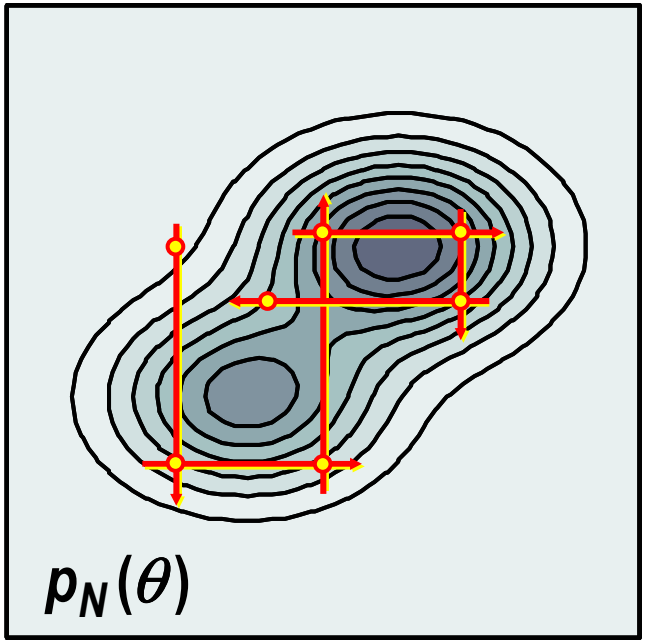
$$\alpha = \min\left\{ \frac{p_N(\theta^*)/\pi(\theta^*)}{p_N(\theta^{(i)})/\pi(\theta^{(i)})}, 1 \right\}$$

Gibbs Sampler

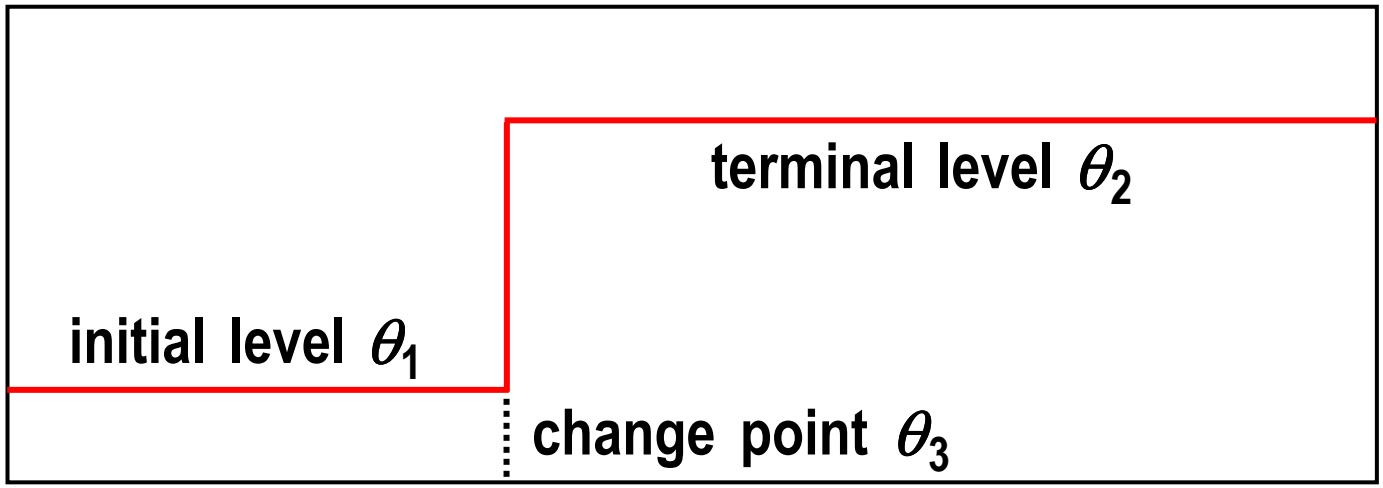
Alternating sampling from full conditionals

For $i = 1, 2, \mathrm{K}$

1. Draw $\theta_1^{(i)}$ from $p_N(\theta_1 \mid \theta_2^{(i-1)})$
2. Draw $\theta_2^{(i)}$ from $p_N(\theta_2 \mid \theta_1^{(i-1)})$



‘Change Point’ Example



Gibbs algorithm

1. Draw $\theta_1^{(i)}$ from $p_N(\theta_1 \mid \theta_2^{(i-1)}, \theta_3^{(i-1)})$
2. Draw $\theta_2^{(i)}$ from $p_N(\theta_2 \mid \theta_1^{(i)}, \theta_3^{(i-1)})$
3. Draw $\theta_3^{(i)}$ from $p_N(\theta_3 \mid \theta_1^{(i)}, \theta_2^{(i)})$

Markov Chain Monte Carlo

Statistical physics
[Metropolis 1953]

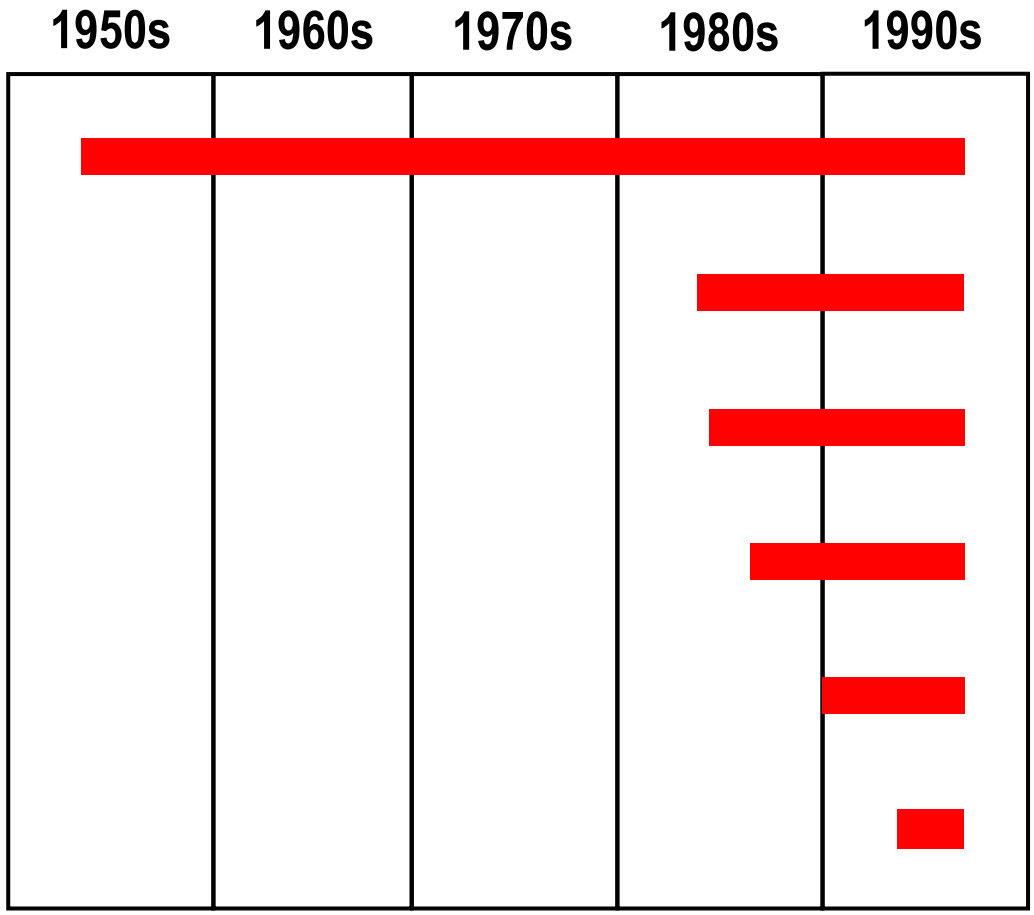
Simulated annealing
[Kirkpatrick et al 1983]

Image processing
[Geman & Geman 1984]

Belief networks
[Pearl 1987]

Bayesian analysis
[Gelfand & Smith 1990]

Engineering



Outline

- ➔ What's Wrong with 'Nonlinear Estimation'?
- ➔ Estimation as Probability Matching
- ➔ Three Approaches to Approximation
  - ➔ Locally Weighted Smoothing
  - ➔ Non-Iterative Monte Carlo Sampling
  - ➔ Iterative Monte Carlo Sampling
  - ➔ Restoration of Information Divergence
- ➔ Which Approximation?

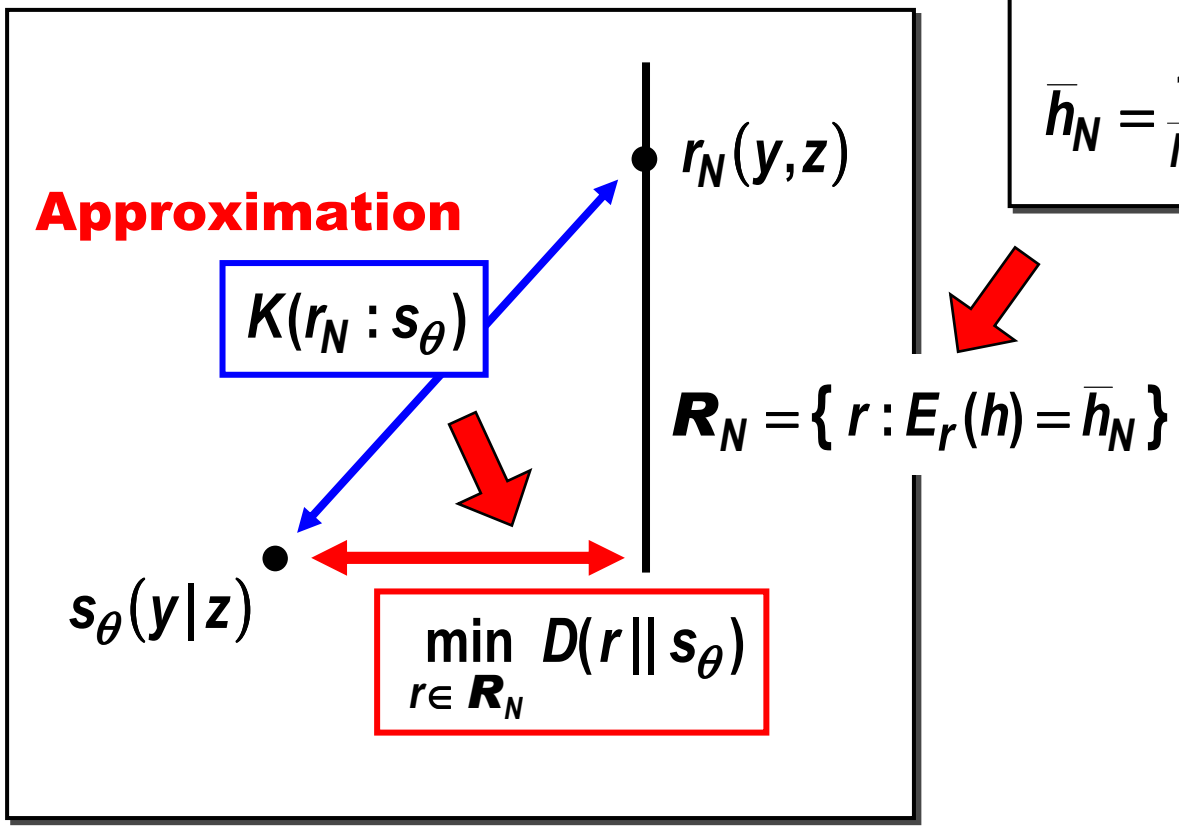
Minimum "Distance" Approximation

Data Compression

$$(y_{m+i}, z_{m+i}),\ i = 1, \mathrm{K}\ , N$$

$$\downarrow$$

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



Minimum Relative Entropy

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

Posterior approximation

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

Minimum relative entropy

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

Relative entropy

$$\boldsymbol{R}_N = \left\{ r(y,z) : \iint r(y,z)\, h(y,z)\, dy\, dz = \frac{1}{N} \sum_{k=m+1}^{N+m} h(y_k, z_k) \right\}$$

Equivalence class

Unnormalized Relative Entropy

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

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

conditional relative entropy

marginal Shannon entropy

MRE Direct Computation

- Convex optimization problem

$$D(\boldsymbol{R}_N \,\|\, s_\theta) = \min_{r \in \boldsymbol{R}_N} D(r \,\|\, s_\theta)$$
$$= \max_{\lambda \in R^n} \left(\lambda' \bar{h}_N - \psi(\theta, \lambda)\right)$$

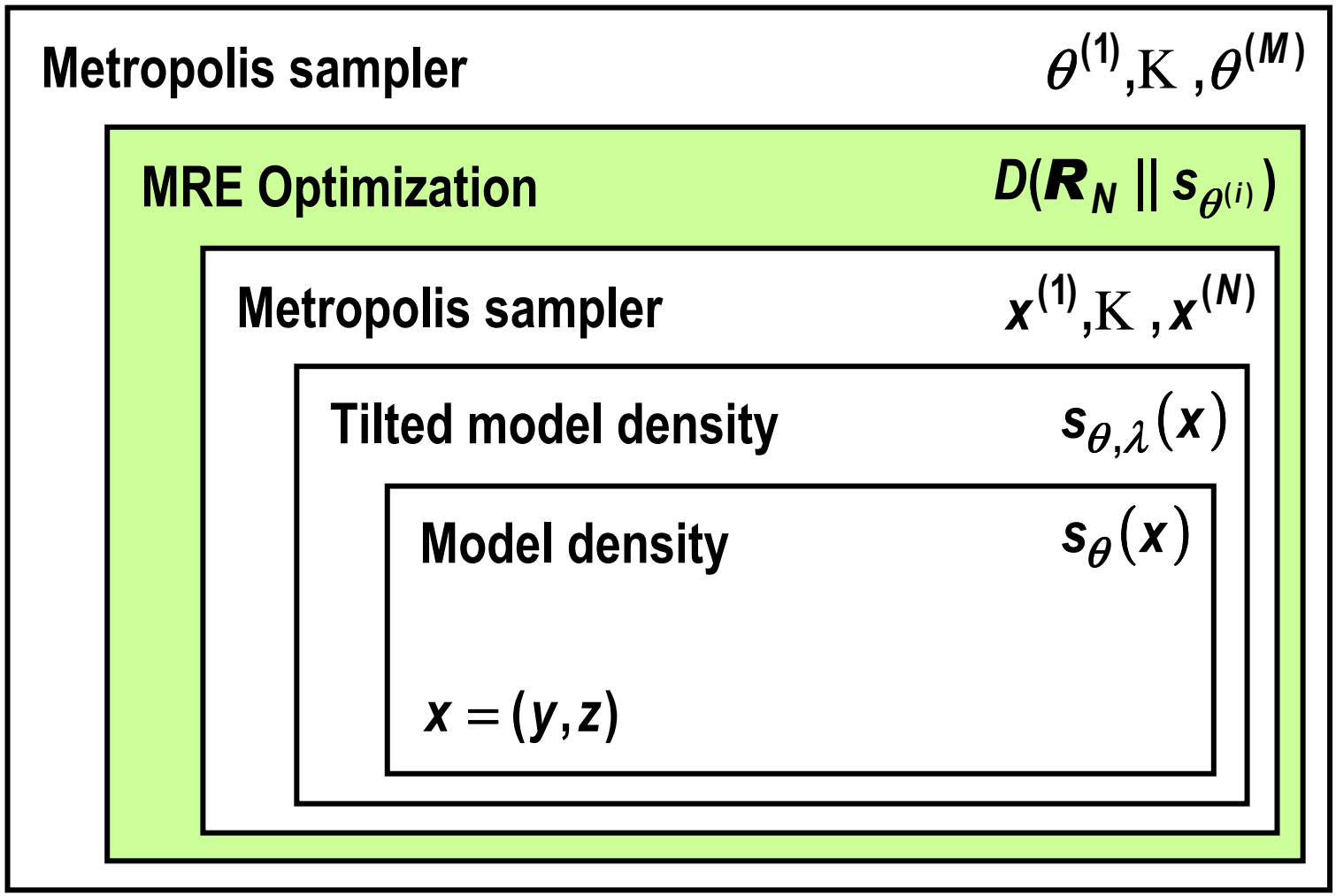
Statistic

Logarithm of normalizing divisor

- Entails multivariate integration

$$\psi(\theta, \lambda) = \log \underbrace{\iint s_\theta(y|z) \exp(\lambda' h(y,z))\, dy\, dz}_{\text{Tilted model density}}$$

MCMC Implementation



'Sensor Validation' Example

- Monitoring of signal differences

$$e_k = y_k - y_{k-1}$$

- Model = mixture of 3 normal distributions

$$(1-\theta_f-\theta_g)\,N(0,Q)+\theta_f\,N(0,0.01*Q)+\theta_g\,N(0,100*Q)$$

normal operation

"frozen" sensor

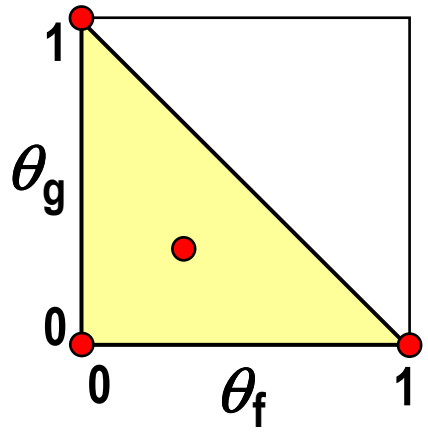
gross errors

- Unknown parameters

  probabilities θ_f, θ_g

- Statistic chosen

$$h_i(e) = \log \frac{s_{\theta_i}(e)}{s_{\theta_0}(e)}$$



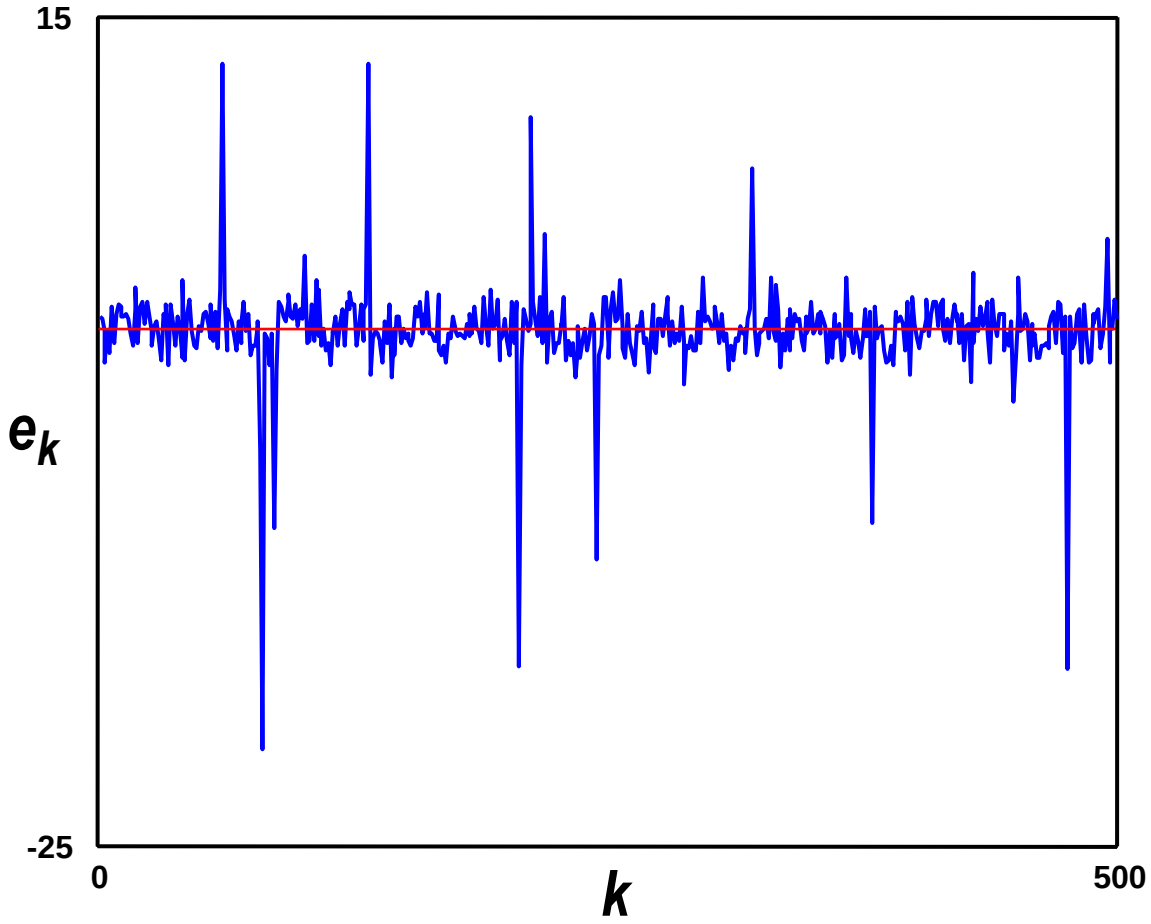
$\theta_0 = [0,0]$

$\theta_1 = [1,0]$

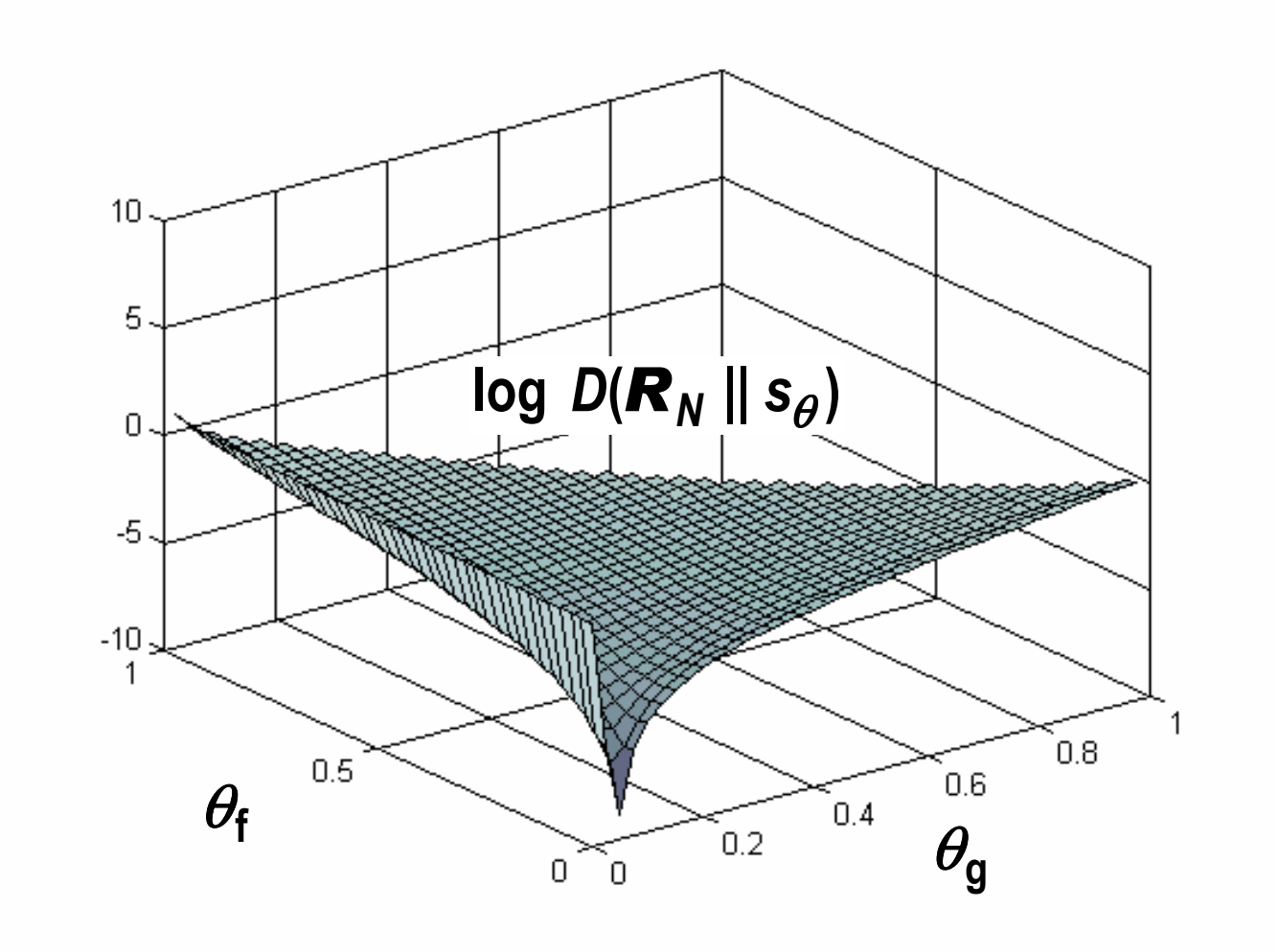
$\theta_2 = [0,1]$

$\theta_3 = [1/3,1/3]$

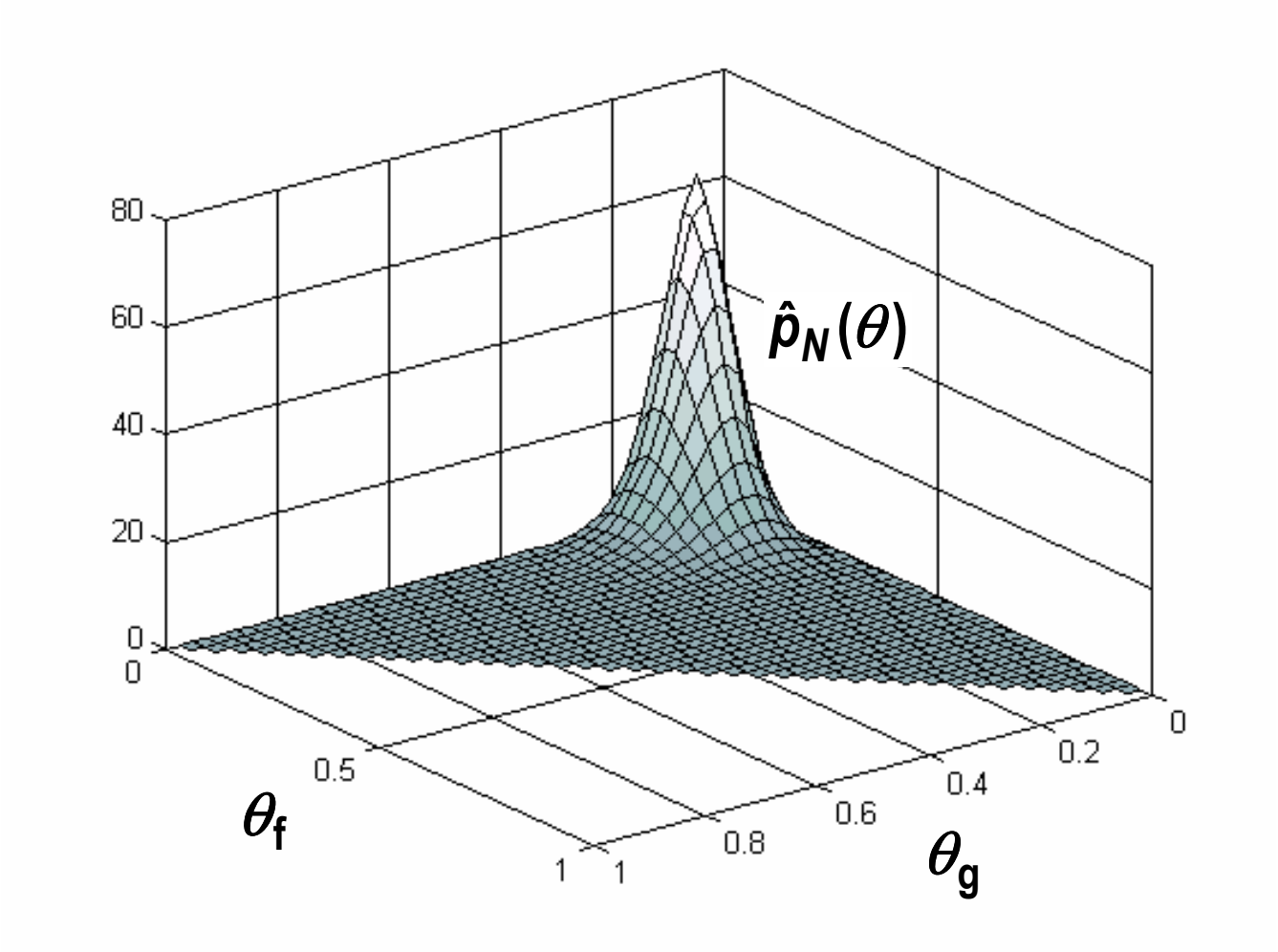
Signal Difference



Relative Entropy

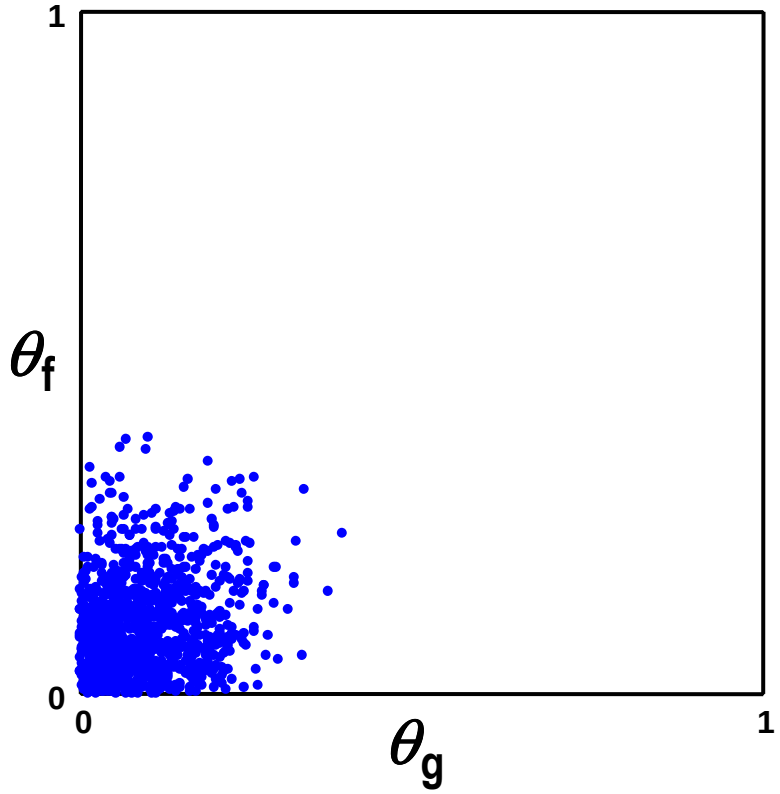


Posterior Density

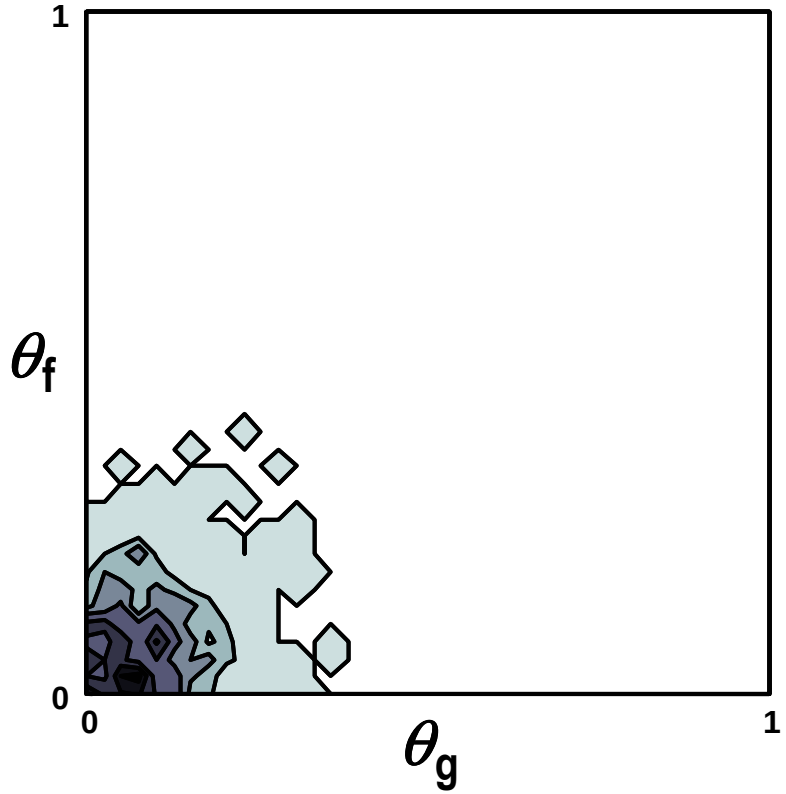


Metropolis Sampler

scatter plot



histogram



Extension to Filtering

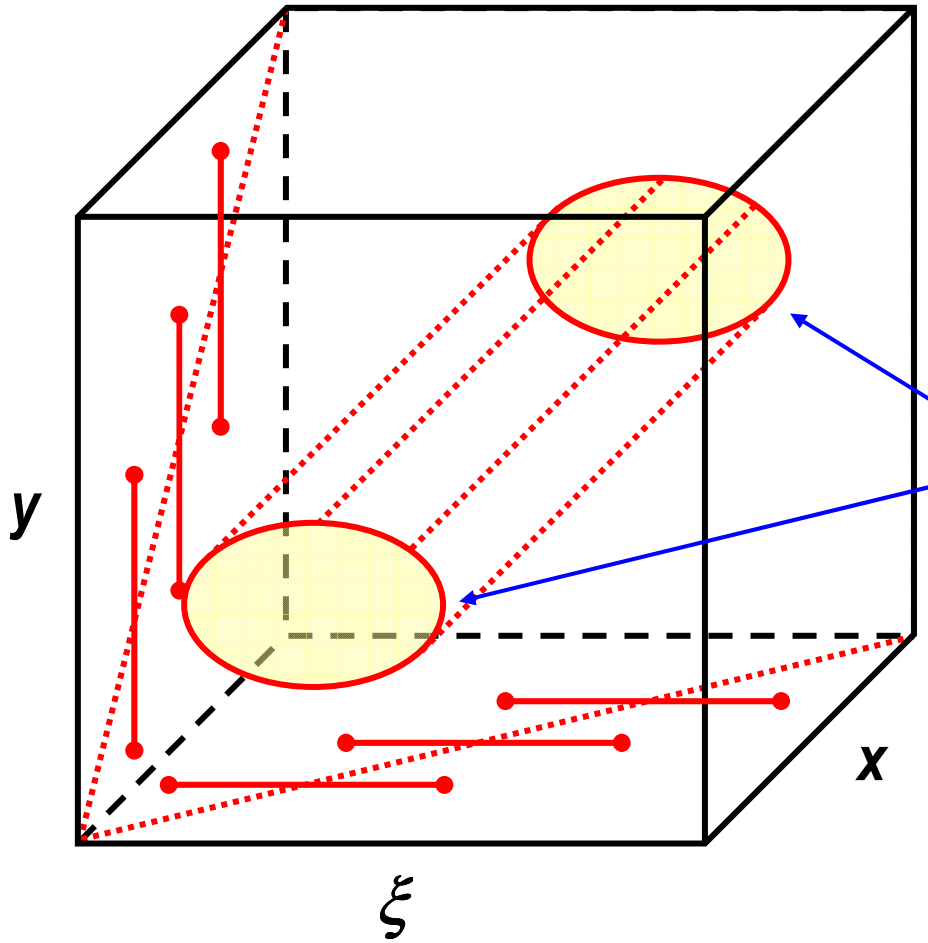
- **State transition density**

  $q(x_{k+1} \mid x_k)$ if controlled: $q(x_{k+1} \mid x_k, u_k)$

- **System output density**

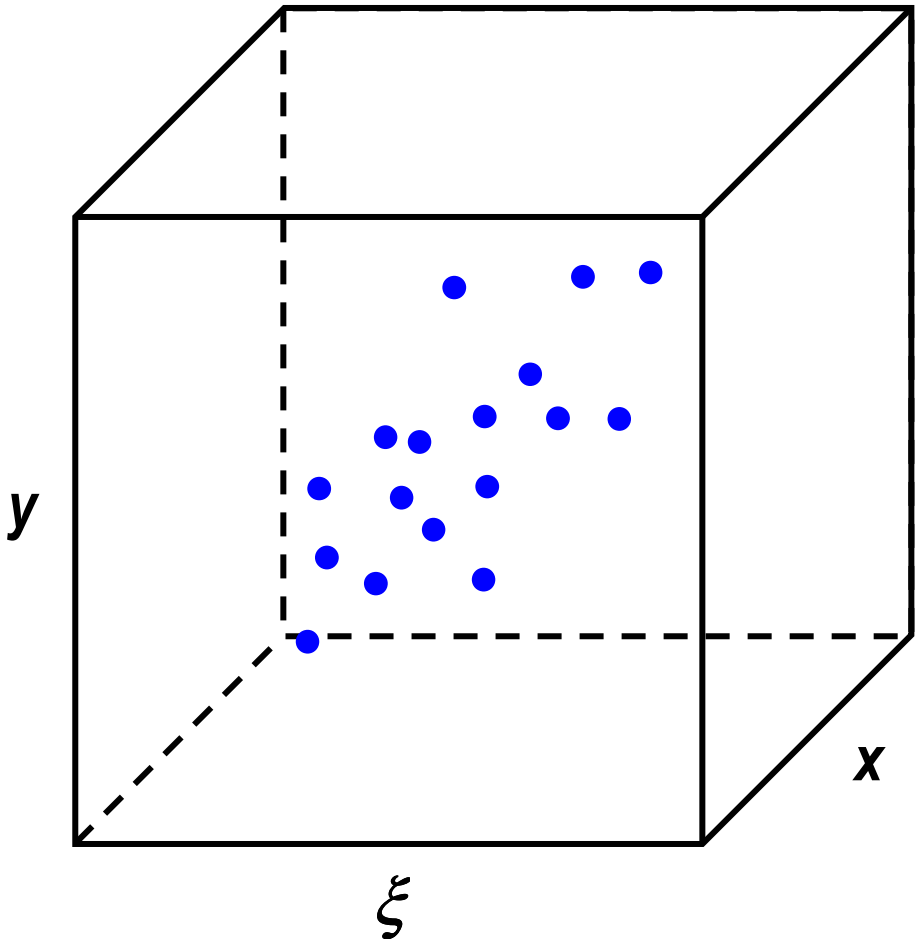
  $s(y_k \mid x_k)$ if controlled: $s(y_k \mid x_k, u_k)$

Model Density



$$\xi_k = x_{k+1}$$

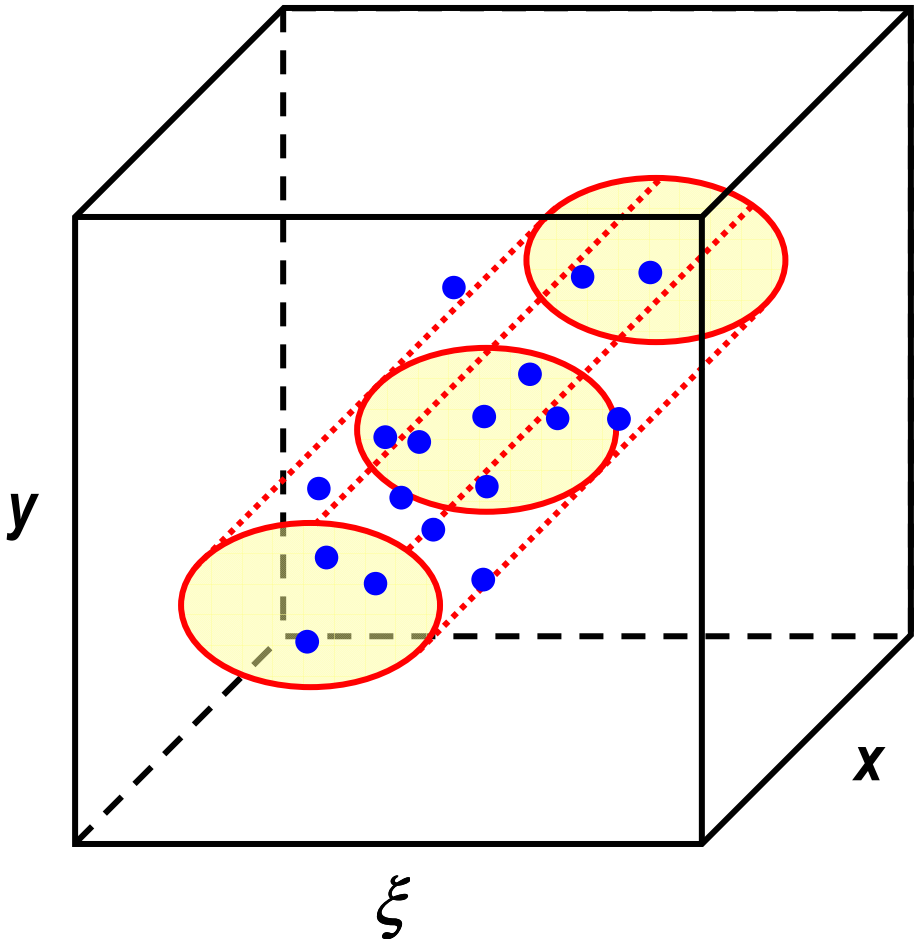
$$q(\xi \mid x)\, s(y \mid x)$$



Empirical Density

Mixture of Dirac functions

$$r_N(\xi,x,y)=\frac{1}{N}\sum_{k=1}^{N}\delta(\xi-\xi_k,x-x_k,y-y_k)$$



Probability Matching

Conditional inaccuracy

$$K(r_N : qs) = \iiint r_N(\xi, x, y) \log \frac{1}{q(\xi \mid x)\, s(y \mid x)} \, d\xi\, dx\, dy$$

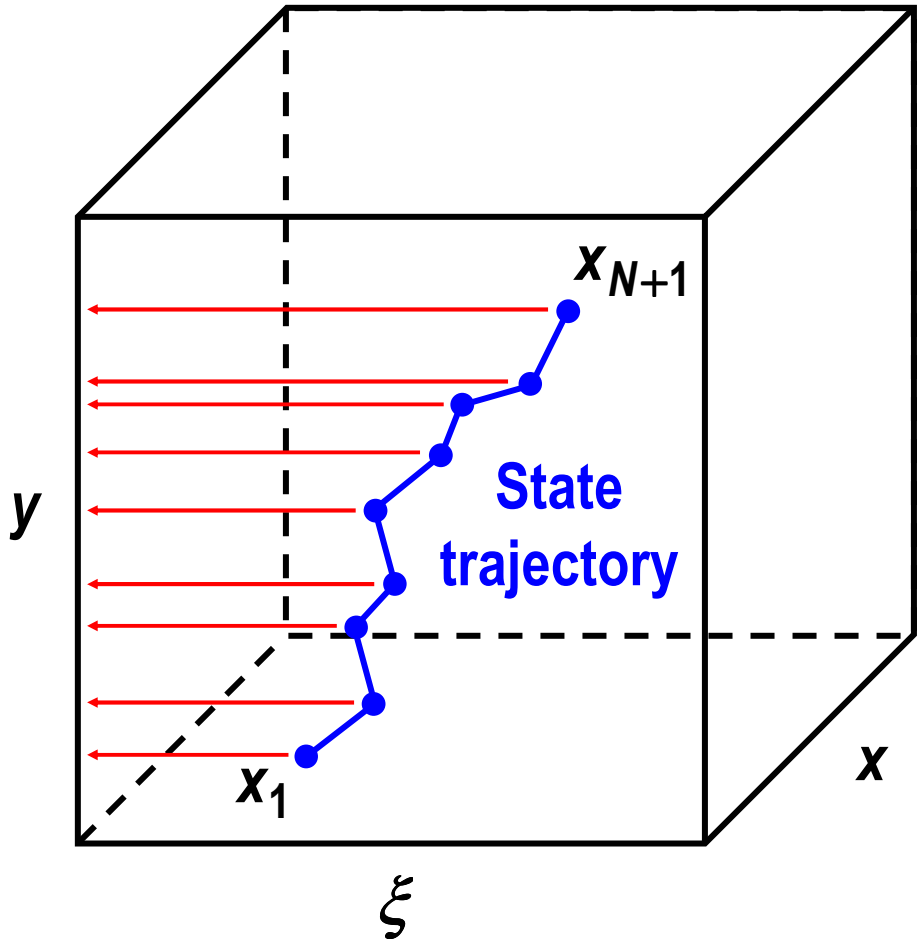
Inaccuracy

$$K(r_N : qs) = -\frac{1}{N} \log \prod_{k=1}^{N} q(x_{k+1} \mid x_k)\, s(y_k \mid x_k)$$

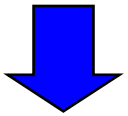
Posterior

$$p_N(x_1, \mathrm{K}, x_{N+1}) = c\, p_0(x_1) \exp(-N K(r_N : qs))$$

Estimation of X_{N+1}



- Initial state fixed
- Terminal state fixed
- Measurements known



- Uncertainty $r_N \in \boldsymbol{R}_N$

$$K(r_N : qs) \rightarrow D(\boldsymbol{R}_N \,||\, qs)$$

Outline

- ➔ What's Wrong with 'Nonlinear Estimation'?
- ➔ Estimation as Probability Matching
- ➔ Three Approaches to Approximation
  - ➔ Locally Weighted Smoothing
  - ➔ Non-Iterative Monte Carlo Sampling
  - ➔ Iterative Monte Carlo Sampling
  - ➔ Restoration of Information Divergence
- ➔ Which Approximation?

Which Approximation?

	'Simple' models	'Complex' models
Small samples	trivial	MCMC
Compress. data	common	IR / MRE
Huge datasets	Local weighting of data	???

MCMC = Computational Engine