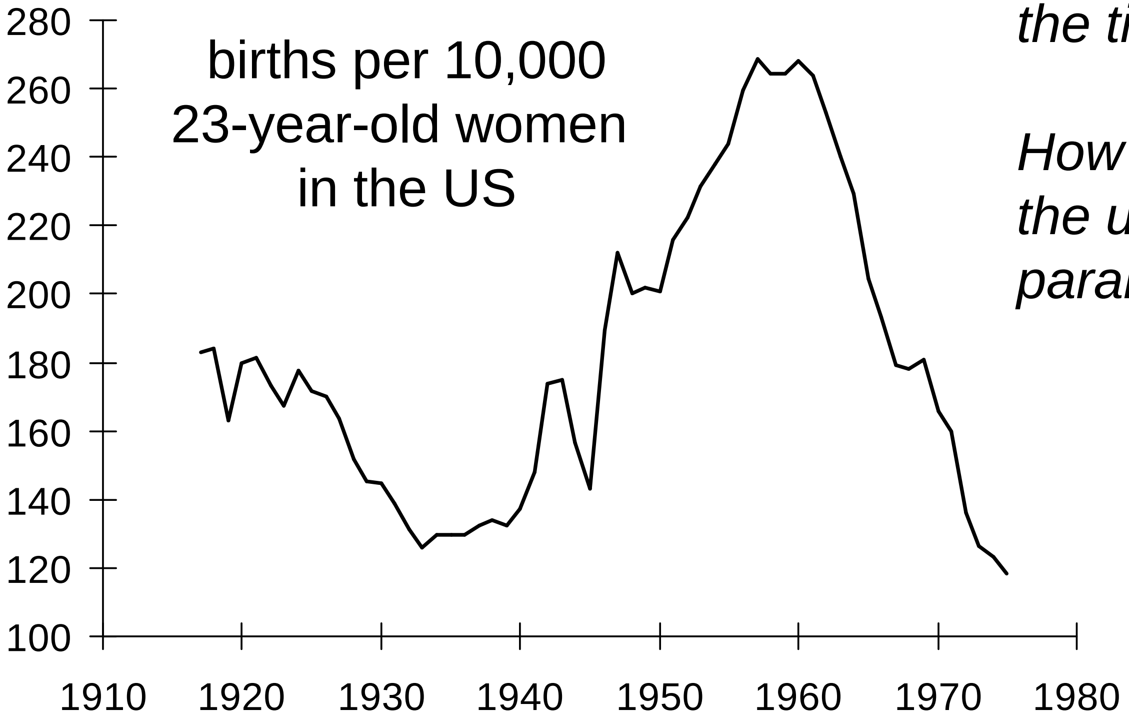
Recursive Bayesian Estimation of Non-linear/Non-Gaussian Dynamic Models

Rudolf Kulhavý

Honeywell Technology Center Prague &
Institute of Information Theory and Automation

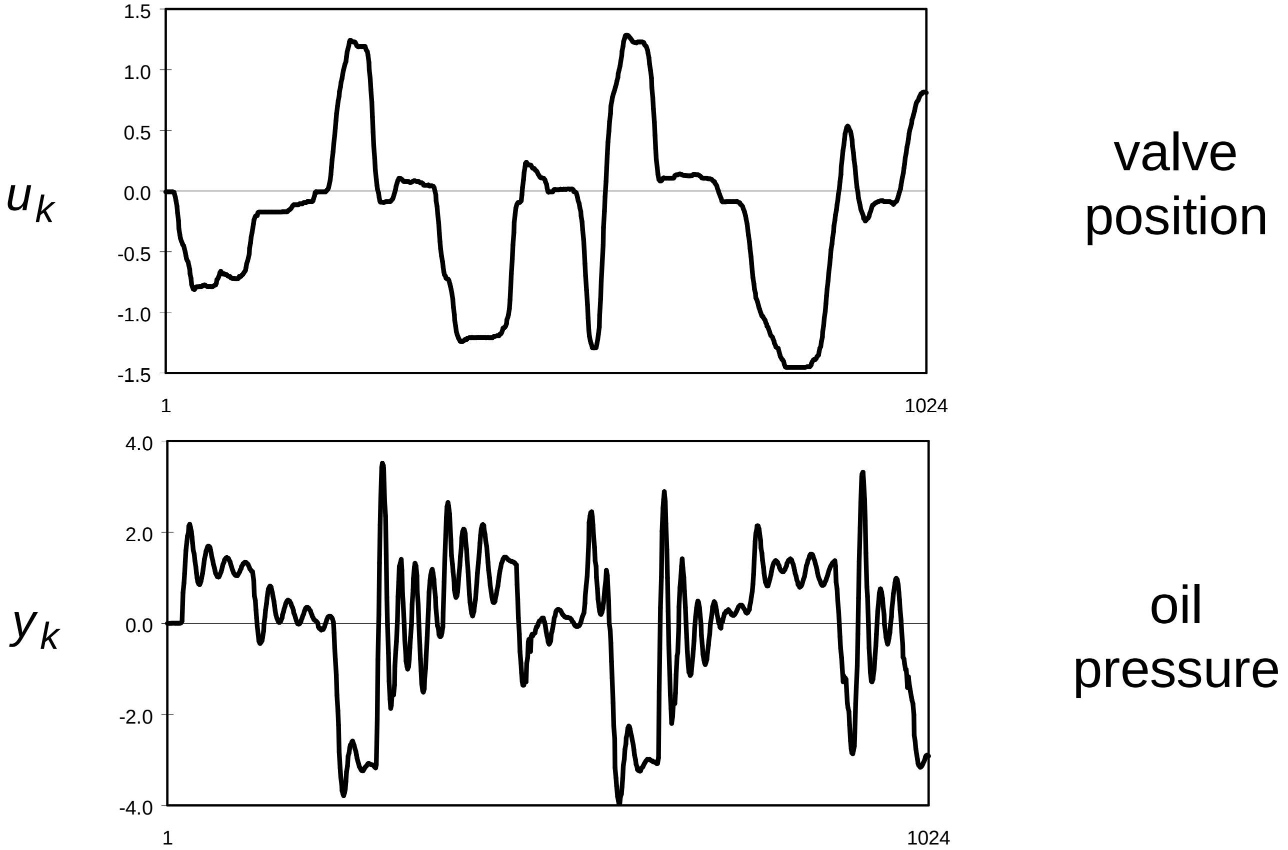
Example: Birth Rates



How to model the time series?

How to estimate the unknown parameters?

Example: Hydraulic Actuator



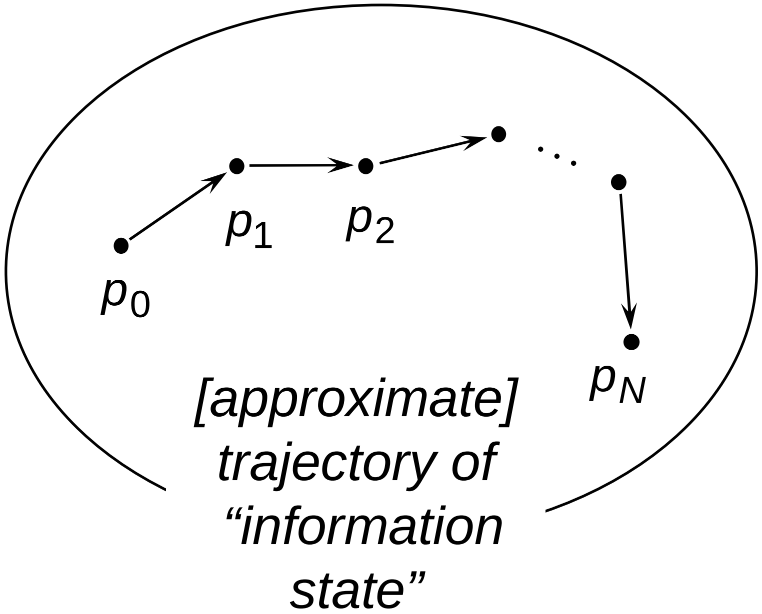
Approximation of Posterior Density

independent observations

$$p_N(\theta) = C\, p_0(\theta) \prod_{k=1}^{N} s_\theta(y_k)$$

controlled dynamic systems

$$p_N(\theta) = C\, p_0(\theta) \prod_{k=1}^{N} s_\theta(y_k | z_k)$$



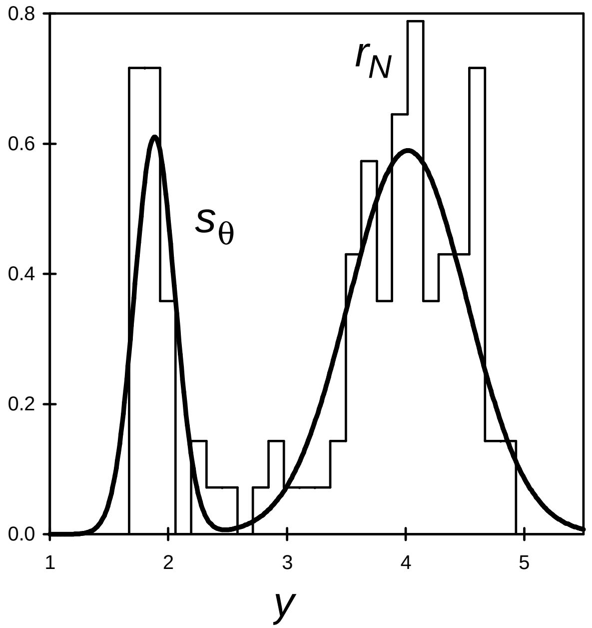
Typical Approaches

- **local approximation of model itself**
  - linearized dynamics
  - approximately Gaussian noise
- **global approximation of posterior distribution**
  - point-mass method [Bucy & Senne]
  - Gaussian sums [Sorenson & Alspach]
  - spline approximation [de Figueiredo]
- **multivariate numerical integration**
  - Markov chain simulation [Metropolis]
  - Langevin stochastic differential equation [Geman]

Fitting of Empirical Distribution

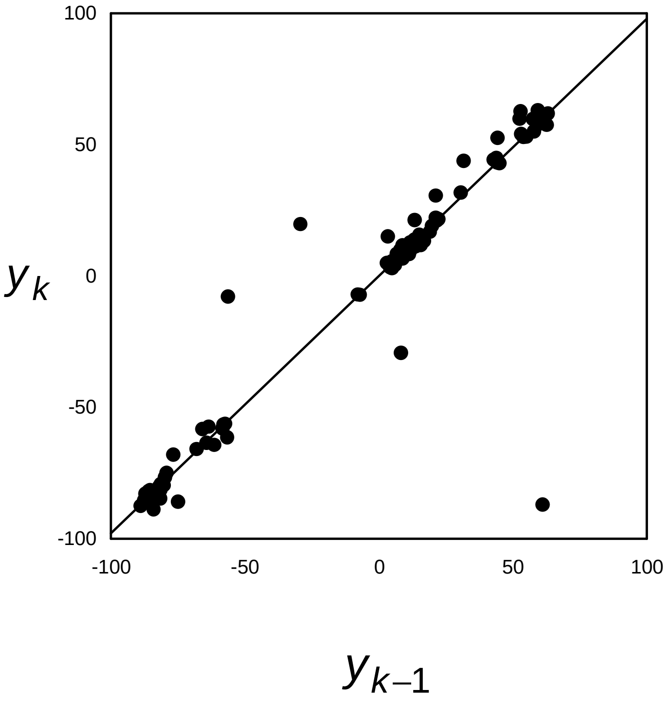
- **independent observations**

  'Old Faithful' geyser errruptions



- **controlled dynamic systems**

  non-Gaussian autoregression



From Data to Parameters ...

- **independent observations**

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

empirical density

$$r_N(y) = \frac{1}{N} \sum_{k=1}^{N} \delta(y - y_k)$$

inaccuracy [Kerridge, 1960]

$$K(r_N : s_\theta) = \int r_N(y) \log \frac{1}{s_\theta(y)}\, \mathrm{d}y$$

- **controlled dynamic systems**

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

higher-order empirical density

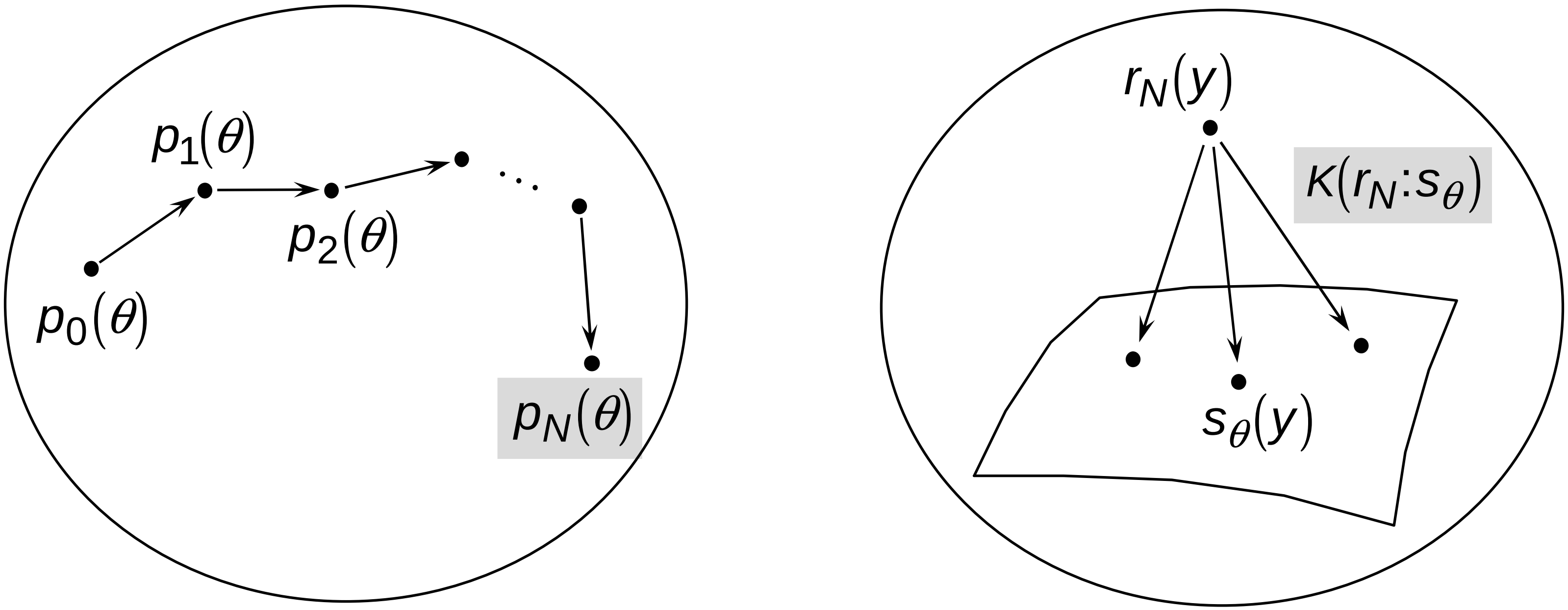
$$r_N(y,z) = \frac{1}{N} \sum_{k=1}^{N} \delta(y - y_k, z - z_k)$$

conditional inaccuracy

$$\overline{K}(r_N : s_\theta) = \iint r_N(y,z) \log \frac{1}{s_\theta(y|z)}\, \mathrm{d}y\, \mathrm{d}z$$

From Parameters to Data ...

probability of *parameters* ⇒ probability of *data*



Inaccuracy for Discrete Data

$$K(r_N : s_\theta) = \sum_y r_N(y) \log \frac{1}{s_\theta(y)}$$

$$= \sum_y r_N(y) \log \frac{r_N(y)}{s_\theta(y)} + \sum_y r_N(y) \log \frac{1}{r_N(y)}$$

Kullback-Leibler distance

Shannon entropy

Conditional Inaccuracy for Discrete Data

$$\bar{K}(r_N : s_\theta) = \sum_{y,z} r_N(y,z) \log \frac{1}{s_\theta(y|z)}$$

$$= \sum_{y,z} r_N(y,z) \log \frac{r_N(y|z)}{s_\theta(y|z)} + \sum_{y,z} r_N(y,z) \log \frac{1}{r_N(y|z)}$$

conditional Kullback-Leibler distance

conditional Shannon entropy

Example: Linear Normal Regression

- theoretical density

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

- conditional inaccuracy

$$\overline{K}(r_N : s_\theta) = \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2} V_N + \frac{1}{2\sigma^2}\left(\theta - \hat{\theta}_N\right)^T C_N \left(\theta - \hat{\theta}_N\right)$$

- statistics

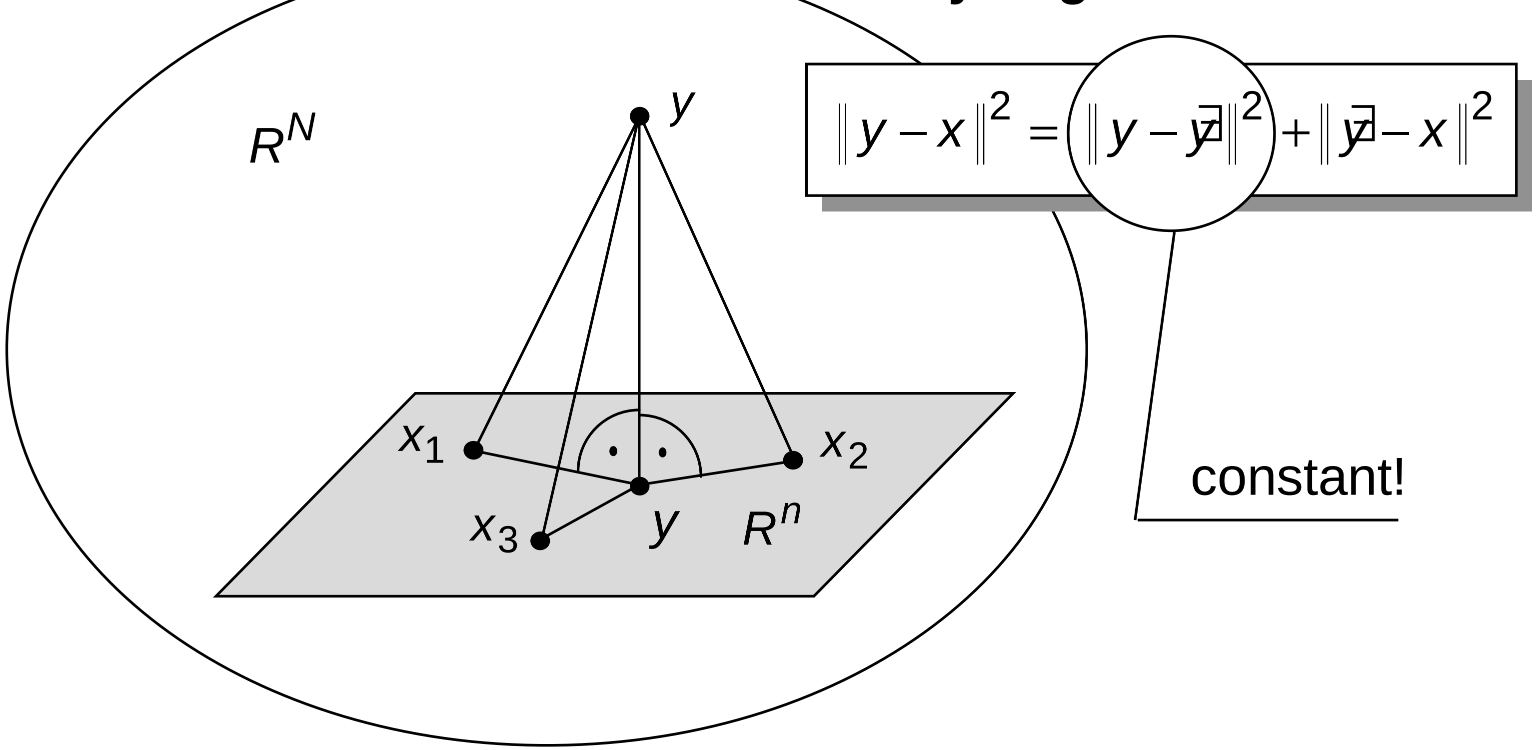
$$\hat{\theta}_N = C_N^{-1} E_N(ZY)$$

$$V_N = E_N\left(Y^2\right) - E_N\left(Y Z^T\right) C_N^{-1} E_N(ZY)$$

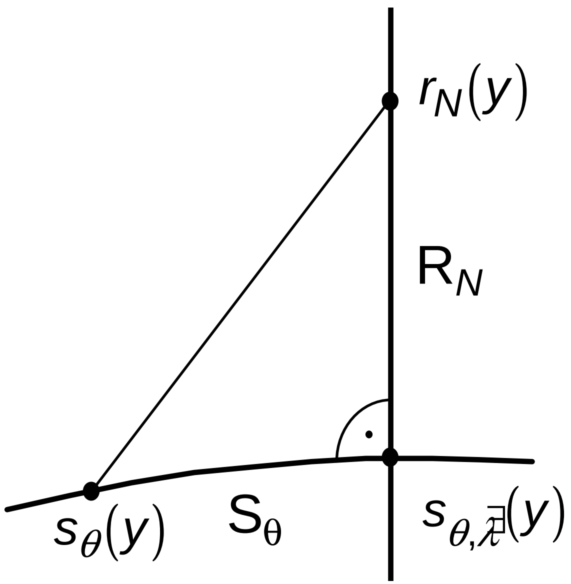
$$C_N = E_N\left(Z Z^T\right)$$

Euclidean Motivation

Pythagorean theorem



Pythagorean Geometry for *Independent* Data



exponential approximating family

$$S_\theta = \left\{ s_{\theta,\lambda}(y) = C\, s_\theta(y)\, \exp\left(\lambda^T h(y)\right) : \lambda \in R^n \right\}$$

partial information about r_N

$$R_N = \left\{ r(y) : \int r(y)\, h(y)\, dy = \frac{1}{N}\sum_{k=1}^{N} h(y_k) \right\}$$

$$= \bar{h}_N$$

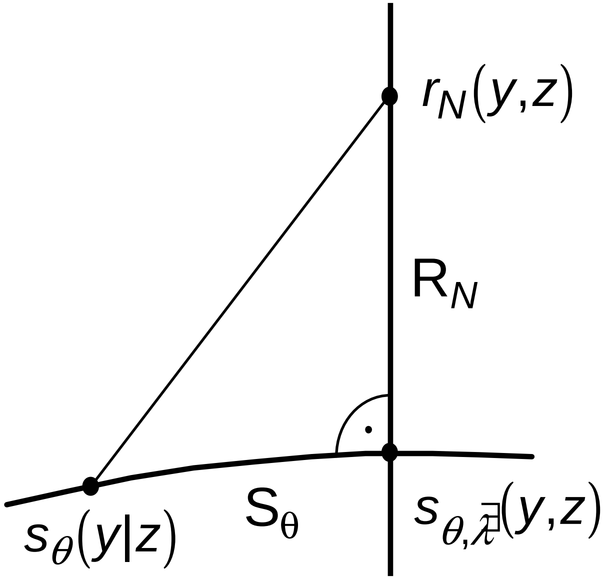
$$K(r_N : s_\theta) = K\left(r_N : s_{\theta,\tilde{\lambda}}\right) + D\left(s_{\theta,\tilde{\lambda}} \| s_\theta\right)$$

“Pythagorean” theorem

K-L distance [Kullback & Leibler]

$$D(\tilde{s} \| s) = \int \tilde{s}(y) \log \frac{\tilde{s}(y)}{s(y)}\, dy$$

Pythagorean Geometry for *Dependent* Data



exponential approximating family

$$s_{\theta,\lambda}(y,z) = C\, s_\theta(y|z)\, \exp\left(\lambda^T h(y,z)\right)$$

partial information about r_N

$$\iint r(y,z)\, h(y,z)\, \mathrm{d}y\, \mathrm{d}z = \frac{1}{N}\sum_{k=1}^{N} h(y_k, z_k)$$

$$= \bar{h}_N$$

$$K(r_N : s_\theta) = K\left(r_N : s_{\theta,\lambda^*}\right) + D\left(s_{\theta,\lambda^*} \| s_\theta\right)$$

"Pythagorean" theorem

"unnormalized" K-L distance

$$D(\tilde{s}\|s) = \iint \tilde{s}(y,z)\, \log \frac{\tilde{s}(y,z)}{s(y|z)}\, \mathrm{d}y\, \mathrm{d}z$$

What Makes the Dynamic Case So Difficult?

Asymmetry!

joint empirical density	conditional theoretical density
$r_N(y,z)$	$s_\theta(y\|z)$

The distribution of z is <u>NOT</u> modelled!

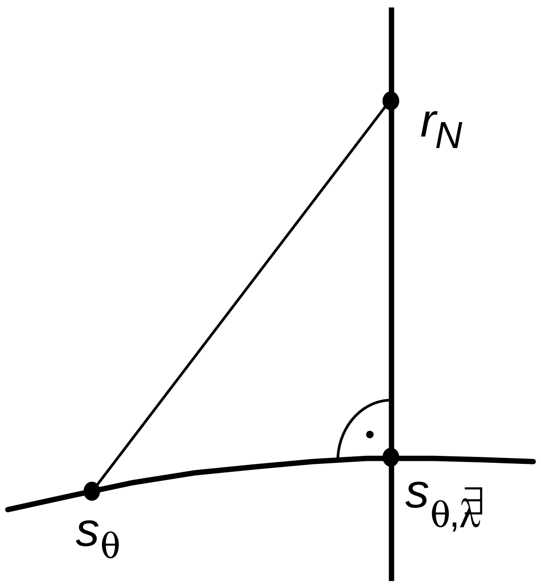
Unnormalized K-L Distance

$$D(\tilde{s}\|s) = \iint \tilde{s}(y,z) \log \frac{\tilde{s}(y,z)}{s(y|z)} \, \mathrm{d}y \, \mathrm{d}z$$

$$= \int \tilde{s}(z) \int \tilde{s}(y|z) \log \frac{\tilde{s}(y|z)}{s(y|z)} \, \mathrm{d}y \, \mathrm{d}z \; - \; \int \tilde{s}(z) \log \frac{1}{\tilde{s}(z)} \, \mathrm{d}z$$

conditional K-L distance — *marginal* Shannon entropy

Key Steps of Approximation



1 choose $h(y)$ or $h(y,z)$
so as to ensure

$$K_N(\theta) = K\left(r_N : s_{\theta,\lambda}\right) \approx \text{const.}$$

for expected values of θ

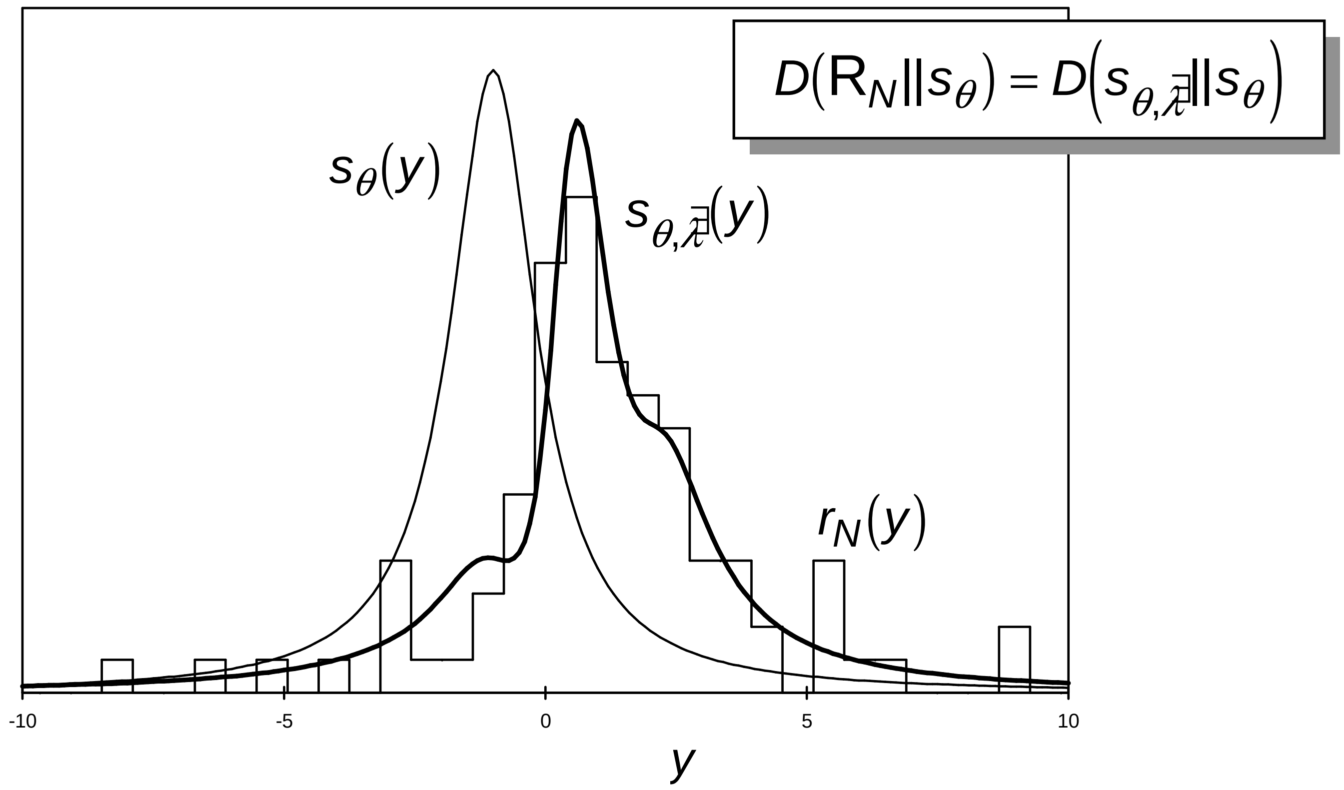
2 approximate $K(r_N \| s_\theta)$
by minimum K-L distance

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

3 approximate the posterior
density as follows

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

Illustration

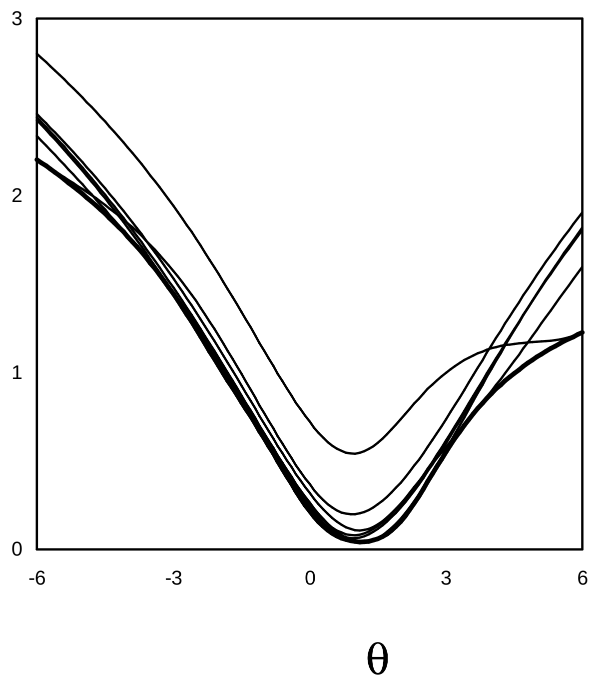


Example: Heavy-Tailed Noise

independent observations

$$y_k = \theta + e_k$$

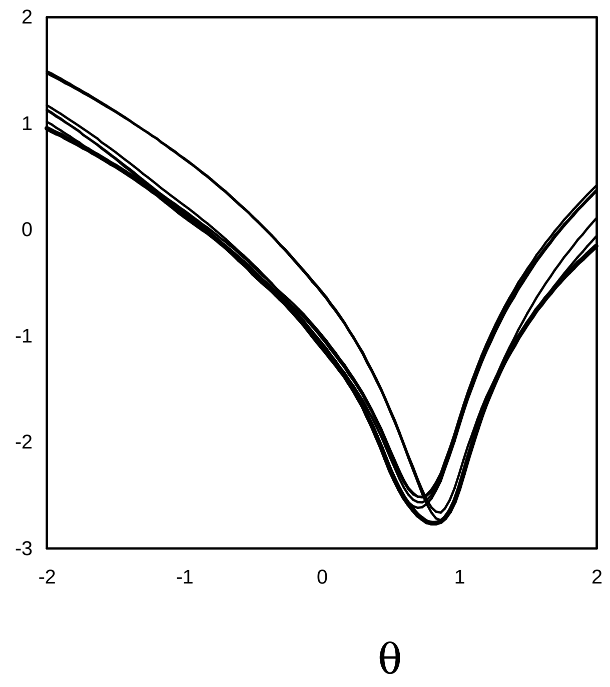
$$0 \le D(\mathrm{R}_N \| s_\theta) \le D(r \| s_\theta)$$



controlled dynamic systems

$$y_k = \theta y_{k-1} + e_k$$

$$-\max_{r \in \mathrm{R}_N} H(\tilde{r}) \le D(\mathrm{R}_N \| s_\theta) \le D(r \| s_\theta)$$



Choice of Statistic

- the choice

$$h_i(y) = L_i(\log s_\theta(y))$$

$$h_i(y,z) = L_i(\log s_\theta(y|z))$$

with $L_i(\cdot)$ linear, $L_i(1) = 0$

ensures

$$L_i(\log p_N(\theta)) = L_i(\log \tilde{p}_N(\theta))$$

- **in particular**, setting $L_i(\log p_N(\theta)) = \log \dfrac{p_N(\theta_i)}{p_N(\theta_{i+1})}, \quad i = 1, \ldots, n$

ensures the same grid profile

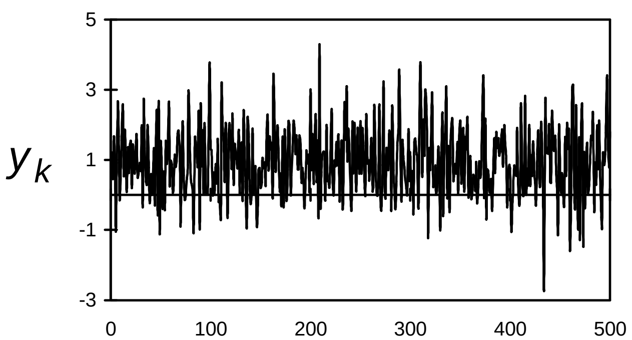
$$p_N(\theta_1), p_N(\theta_2), \ldots, p_N(\theta_{n+1})$$

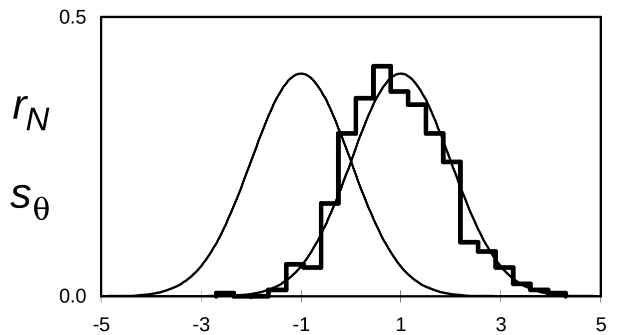
$$\tilde{p}_N(\theta_1), \tilde{p}_N(\theta_2), \ldots, \tilde{p}_N(\theta_{n+1})$$

⇒ **global approach**

Shift of Paradigm

1. from matching data
   to matching probabilities





2. Pythagorean geometry
   of estimation

$$K(r_N : s_\theta)$$

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

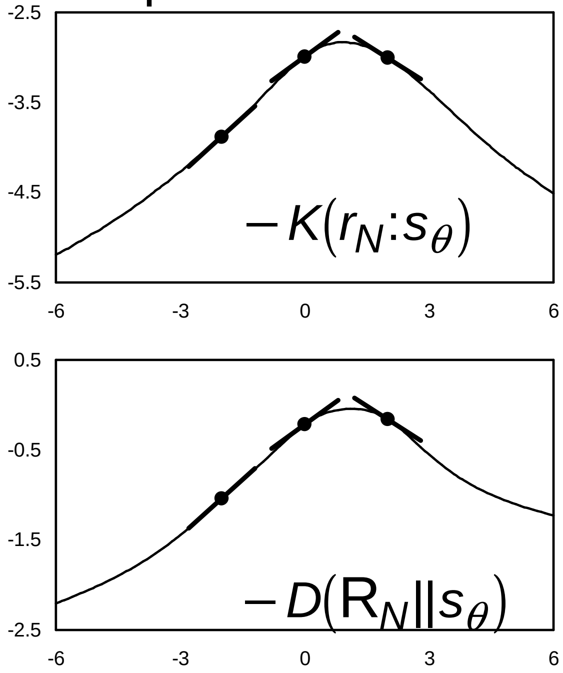
⇨ construction of statistic

⇨ design of approximation

- both discrete and *continuous* data
- both indep. observations and *contr. dynam. systems*

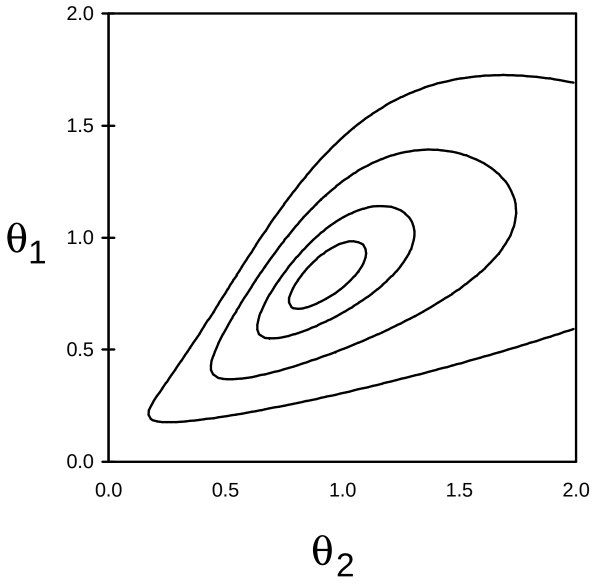
Practical Implications

1. sophisticated
   "interpolation"



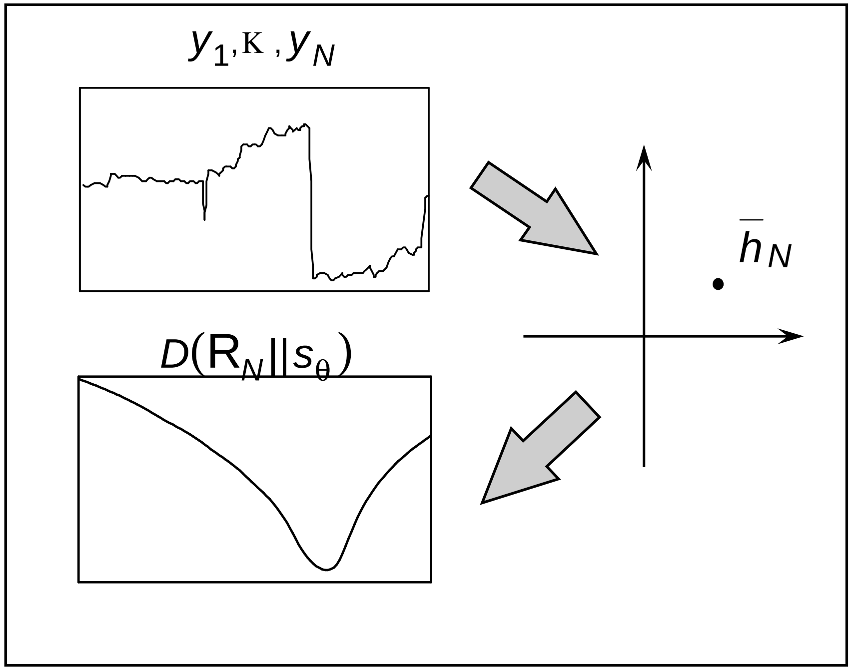
2. exploration of "cost function"

$$J_N(\theta) = D(\mathrm{R}_N \| s_\theta)$$



Practical Implications (cont.)

3. separation of data compression and likelihood restoration



Ramifications:

- the map

  $$\left(\bar{h}_N, \theta\right) \alpha \;\; D\left(\mathrm{R}_N || s_\theta\right)$$

  can be analyzed a priori
- models can be parametrized by the statistic itself
  - ⇨ *just-in-time modelling*
  - ⇨ *model-based fault detection and isolation*

Algorithm

- **independent observations**

$$D(\mathrm{R}_N || s_\theta) = \max_\lambda \left[\lambda^T \bar{h}_N - \log \int s_\theta(y) \exp\left(\lambda^T h(y)\right) \mathrm{d}y \right]$$

- **controlled dynamic systems**

$$D(\mathrm{R}_N || s_\theta) = \max_\lambda \left[\lambda^T \bar{h}_N - \log \iint s_\theta(y|z) \exp\left(\lambda^T h(y,z)\right) \mathrm{d}y \, \mathrm{d}z \right]$$

Numerical Implementation

- multivariate integration

$$\iint s_{\theta}(y|z)\,\exp\left(\lambda^{T} h(y,z)\right)\,\mathrm{d}y\,\mathrm{d}z$$

Monte Carlo methods

- approximation of model

$$s_{\theta}(y|z) = \exp\left(\lambda^{T} h(y,z) - \psi(\theta,\lambda)\right)$$

exponential family

- a priori computation

$$\left(\bar{h}_{N},\theta\right)\ \alpha\ \ D\left(\mathrm{R}_{N}\|s_{\theta}\right)$$

nonparametric regression

- parallel implementation

statistic: $\bar{h}_{i,N} = E_{N}\left(h_{i}(y,z)\right)$ for $i = 1,\mathrm{K},n$

K-L distance: $D\left(\mathrm{R}_{N}\|s_{\theta}\right)$ for $\theta \in \{\theta_{1},\mathrm{K},\theta_{M}\}$

Complexity Dropped Significantly But ...

$$D(R_N \| s_\theta) = \max_{\lambda \in R^n} \left[\lambda^T \bar{h}_N - \log \iint s_\theta(y|z) \exp\left(\lambda^T h(y,z)\right) dy\, dz \right]$$

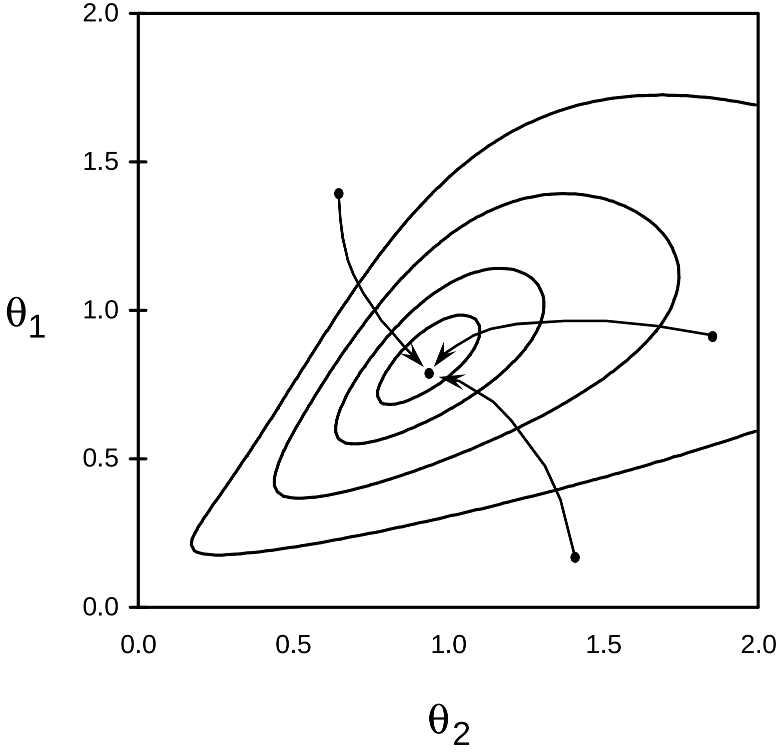
optimization over space of dimension of dim(h)=n

integration over space of dimension dim(y)+dim(z)

The optimization problem needs to be solved for every θ separately.

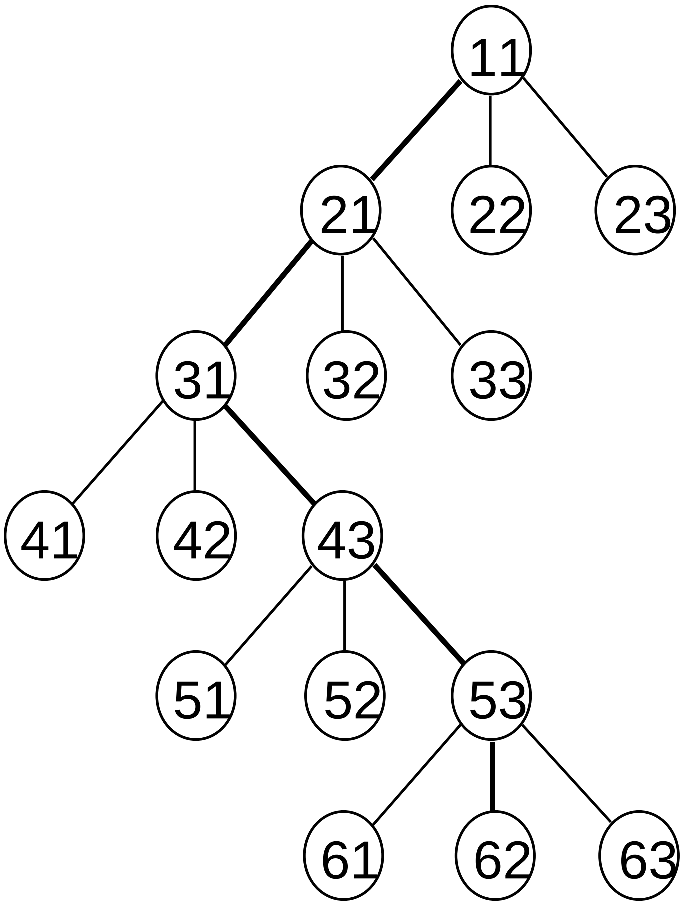
Sampling of ‘Cost Function’

$$J_N(\theta) = D(\mathrm{R}_N \| s_\theta)$$



taking samples
of the ‘cost function’
over a specific set
of points in the
parameter space

Evolutionary Modelling



step-by-step
building of
a model density
$s_{ij}(y|z)$

Only models
with small enough
$D(R_N||s_{ij})$
'survive'.