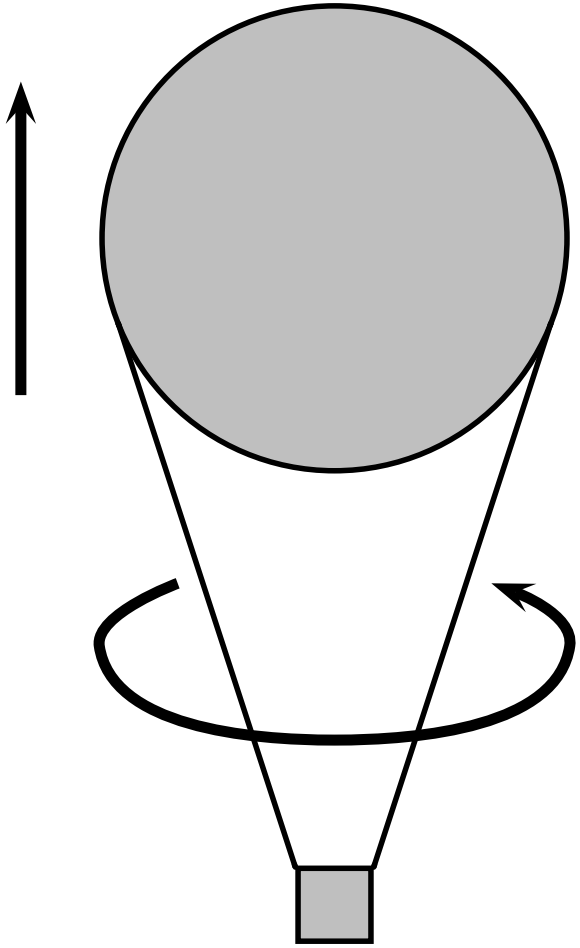
A Kullback-Leibler Distance Approach to System Identification

Rudolf Kulhavý

Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Prague

Example 1: BALLOON



Sun radiation measurements

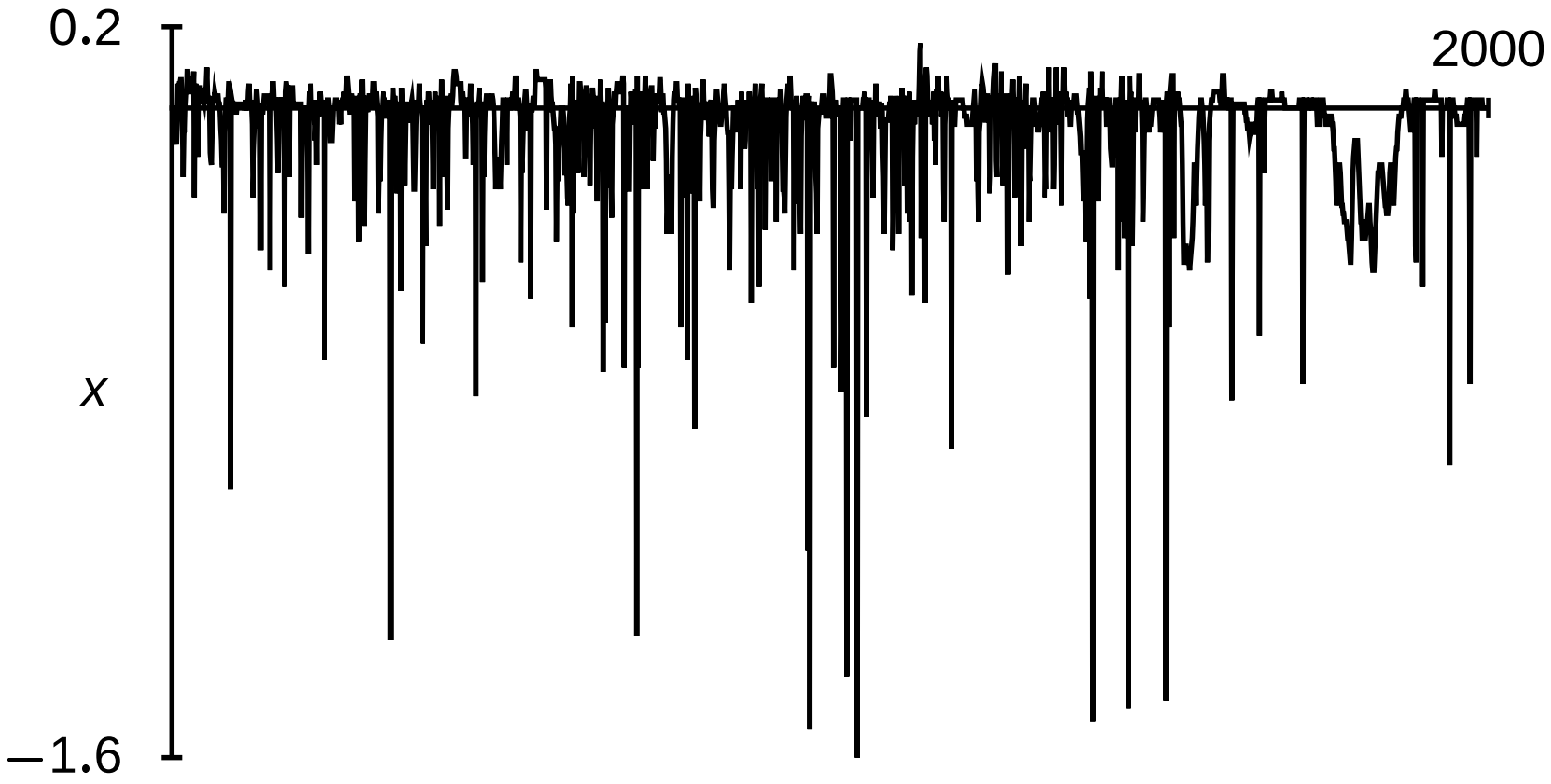
Source:

`http://lib.stat.cmu.edu`

`/Datasets/balloon`

L. Davis and U. Gather, JASA, 1993

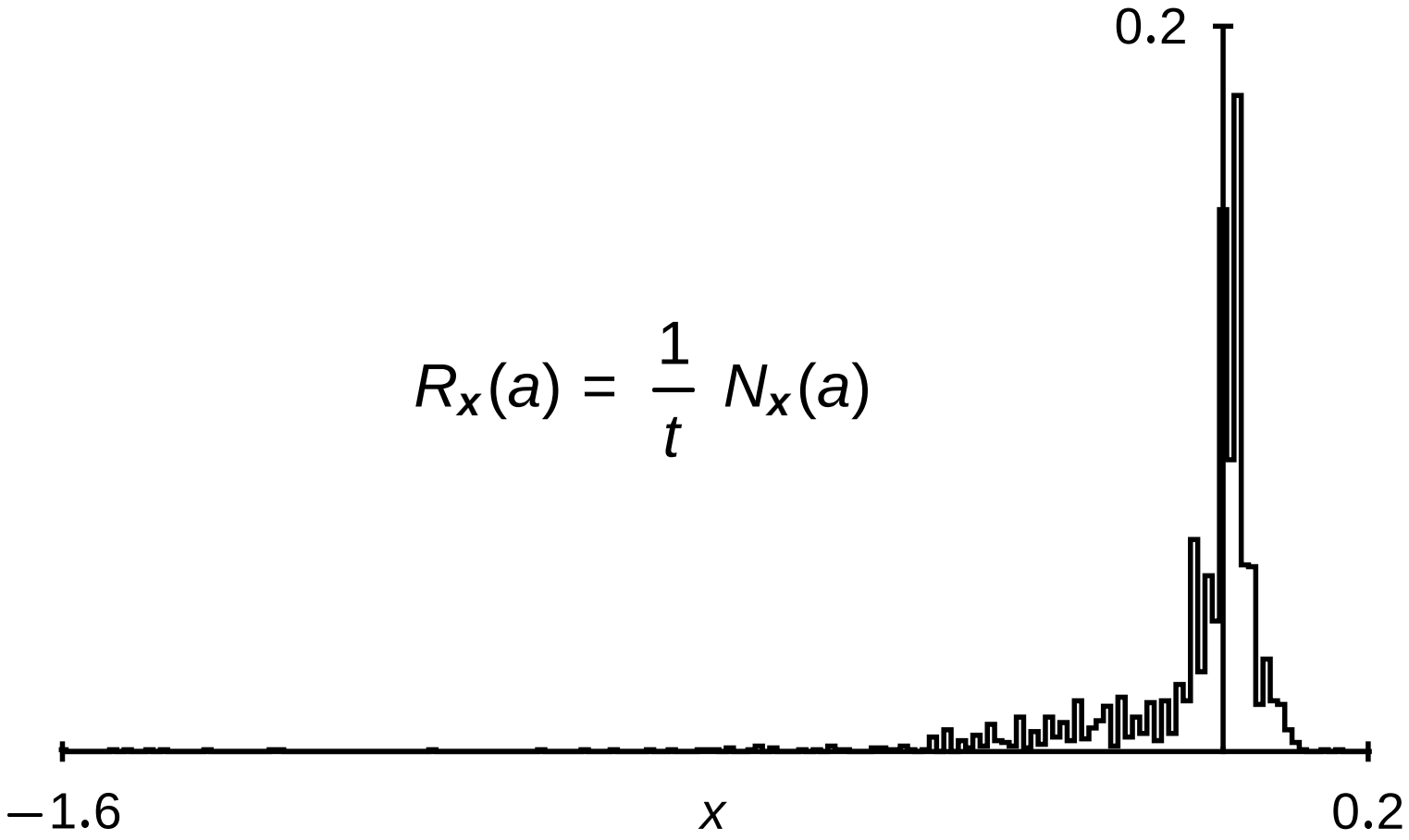
BALLOON: Residuals



What's the actual distribution of Data?

BALLOON: Empirical Distribution

$$R_x(a) = \frac{1}{t} N_x(a)$$

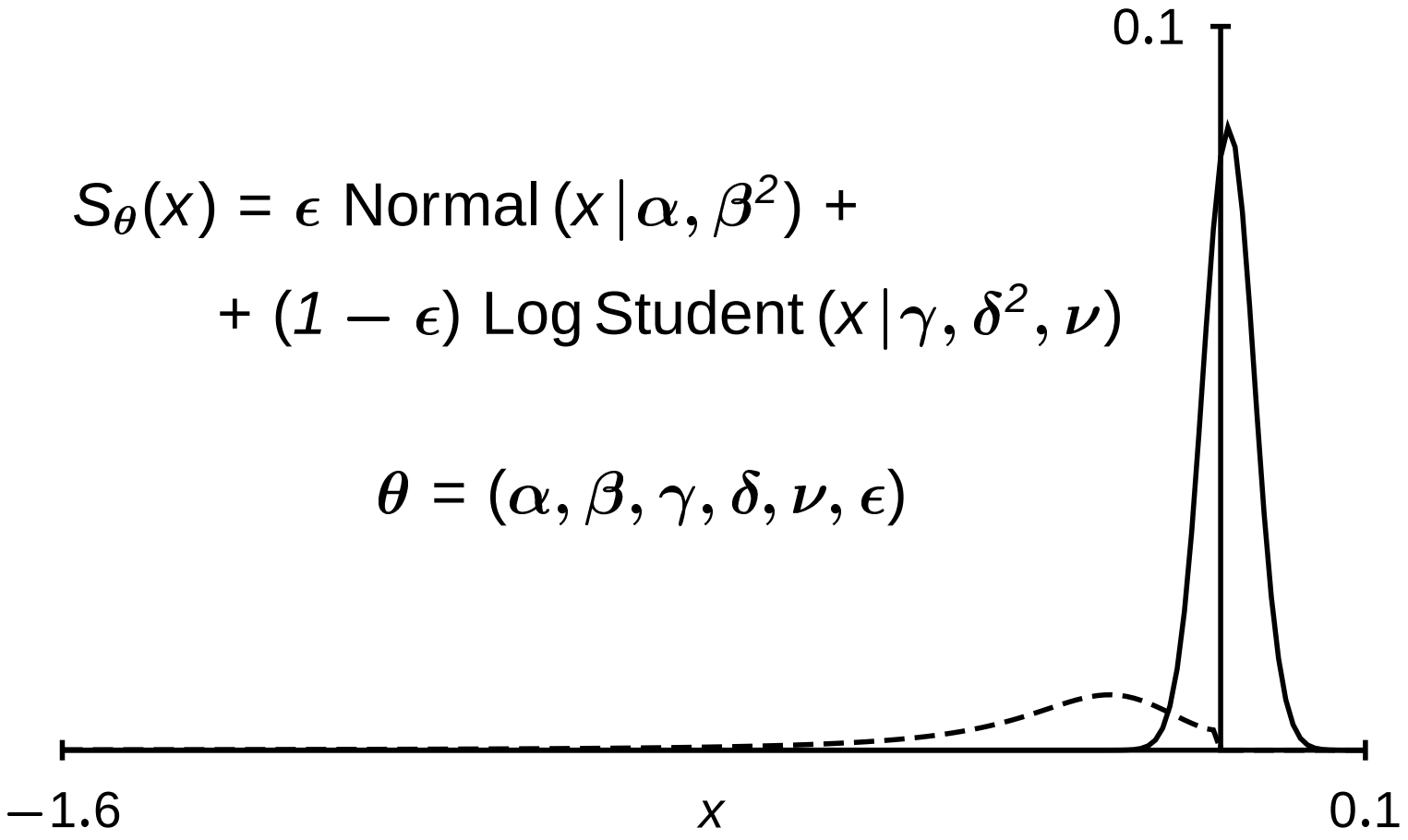


BALLOON: Model Distribution

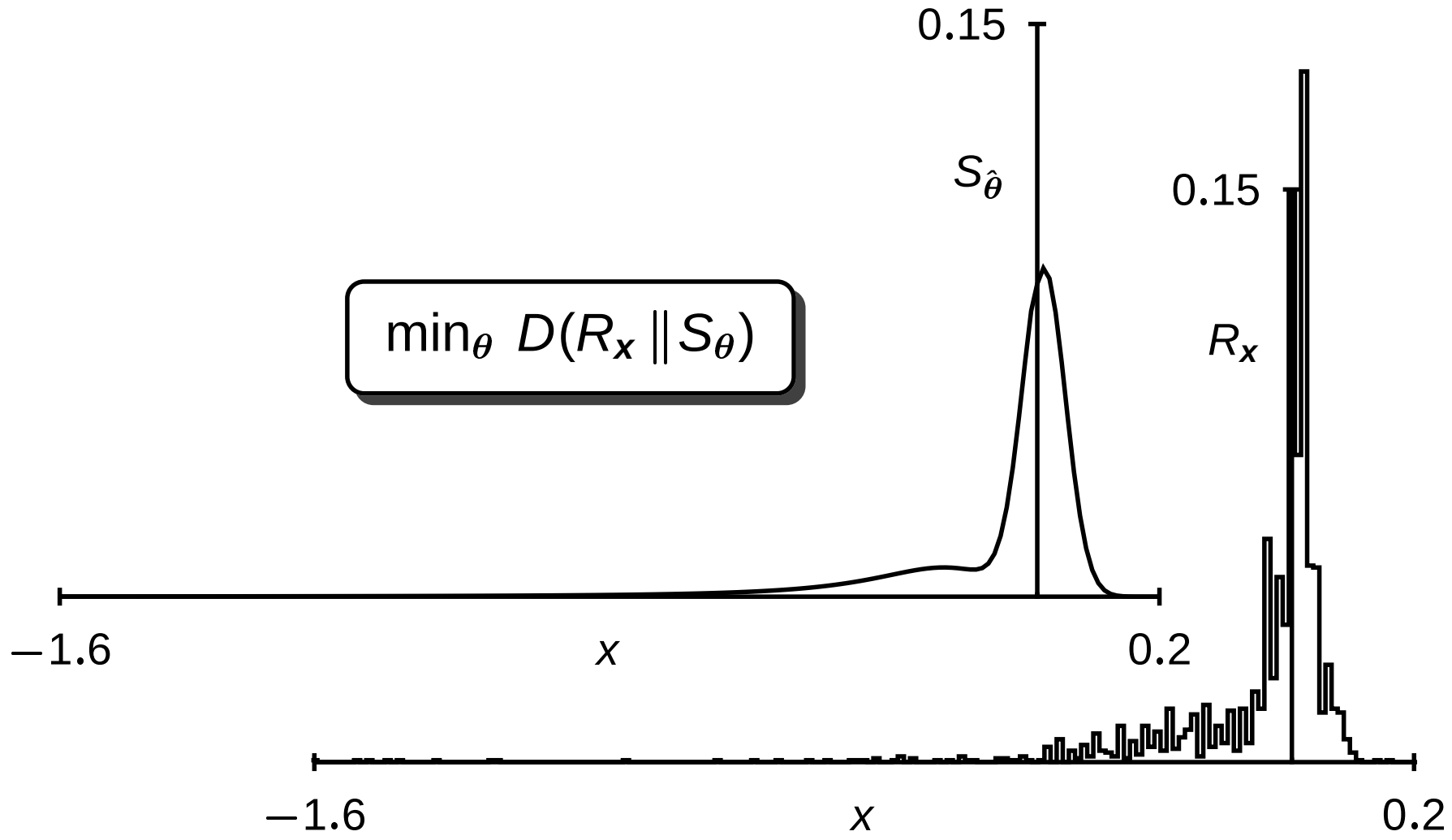
$$S_\theta(x) = \epsilon \text{ Normal}(x \mid \alpha, \beta^2) +$$

$$+ (1 - \epsilon) \text{ Log Student}(x \mid \gamma, \delta^2, \nu)$$

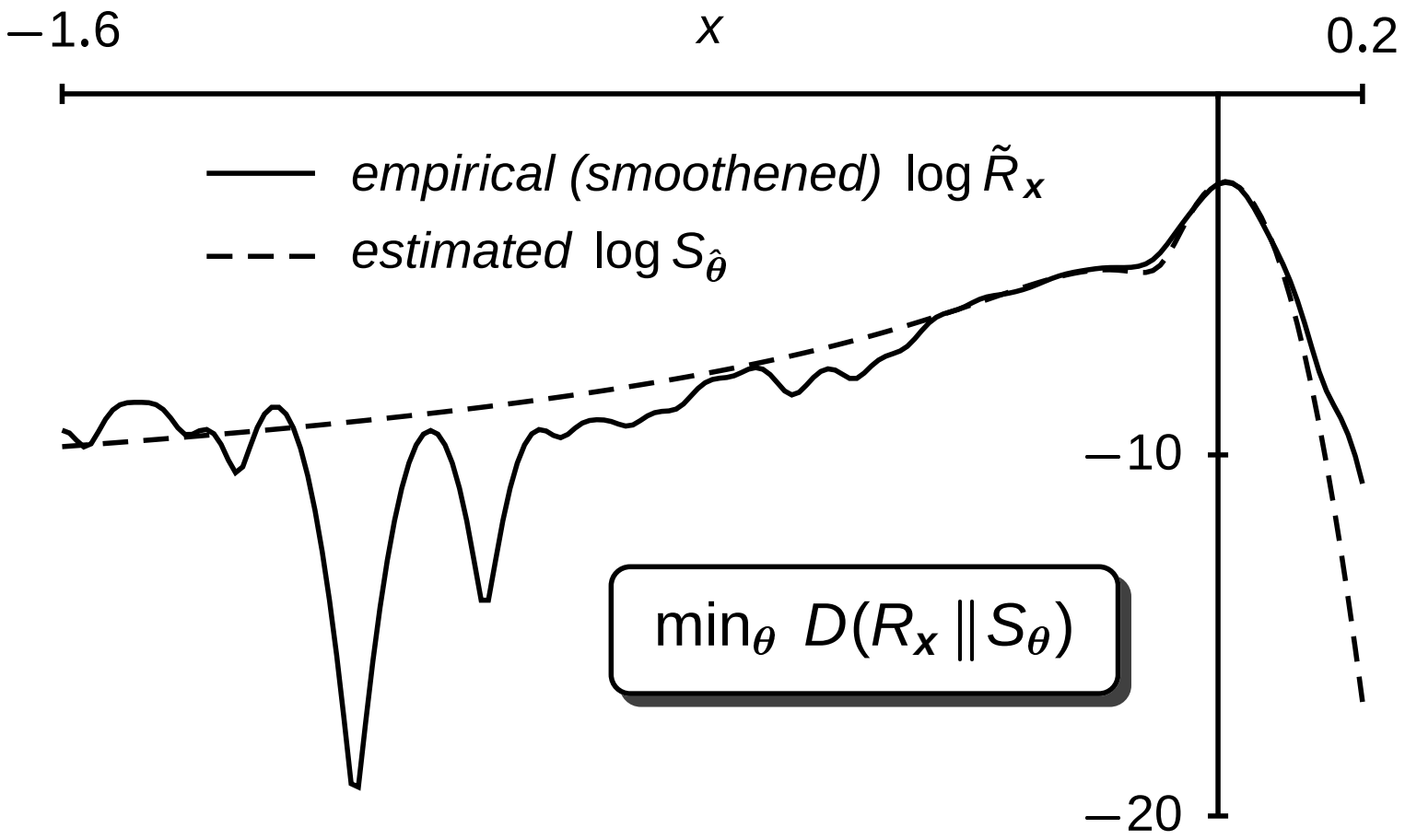
$$\theta = (\alpha, \beta, \gamma, \delta, \nu, \epsilon)$$



BALLOON: Parameter Estimation



BALLOON: Parameter Estimation



What's a proper distance of empirical and model distributions?

Ann. Math. Statist. **22** (1951) 79–86

ON INFORMATION AND SUFFICIENCY

By S. Kullback and R. A. Leibler

The George Washington University and Washington, D. C.

Measuring a "Distance" of Distributions

$$D(R \| S) = \sum_{x \in \mathcal{X}} R(x) \log \frac{R(x)}{S(x)}$$

$$D(R \| S) = \int_{\mathcal{X}} r(x) \log \frac{r(x)}{s(x)} \, \mathrm{d}x$$

Kullback-Leibler distance

relative entropy

informational divergence

discrimination information

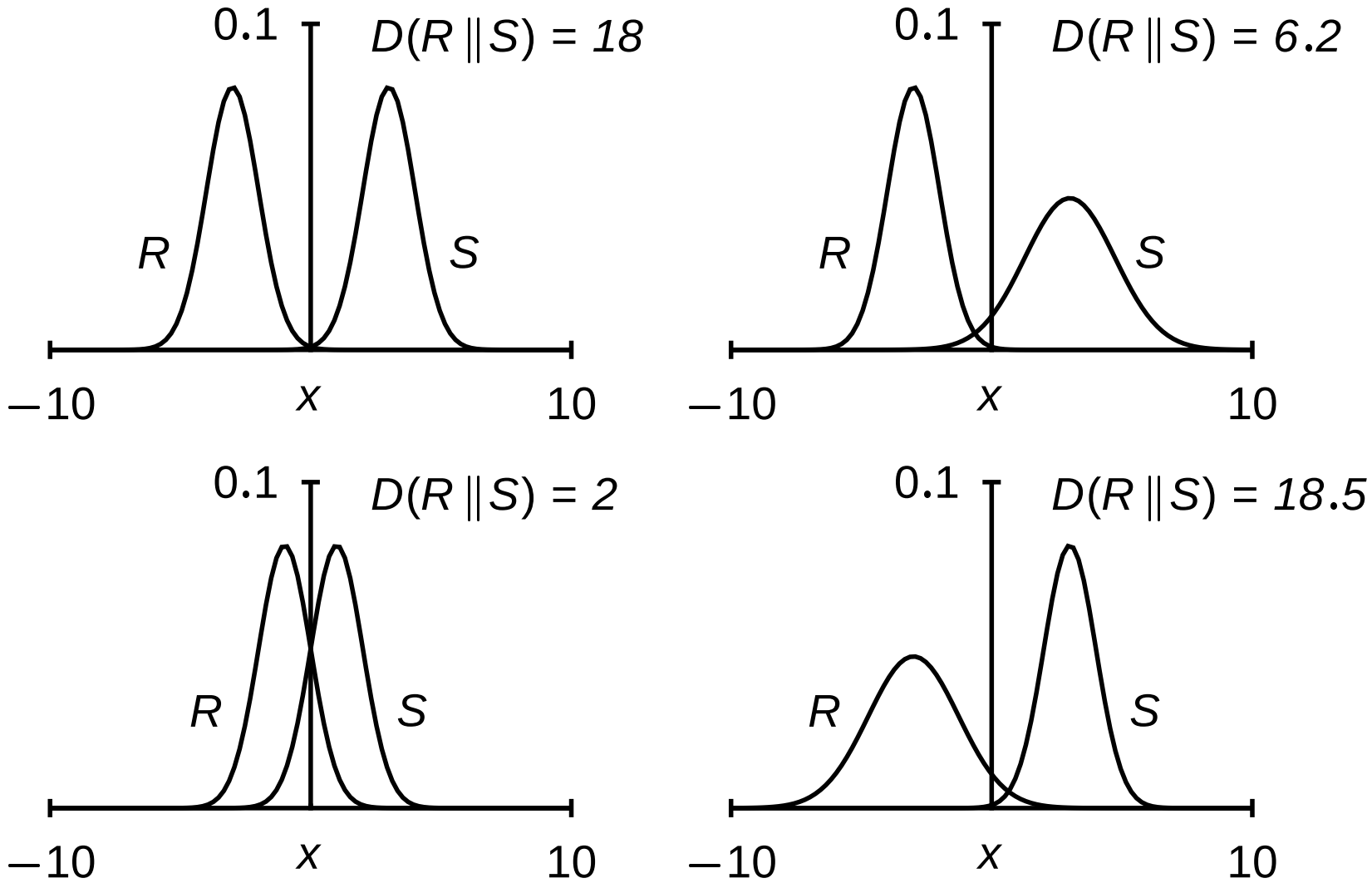
K.-L. Distance of Normal Distributions

$$R \sim \text{Normal}\,(\mu_R, {\sigma_R}^2)$$

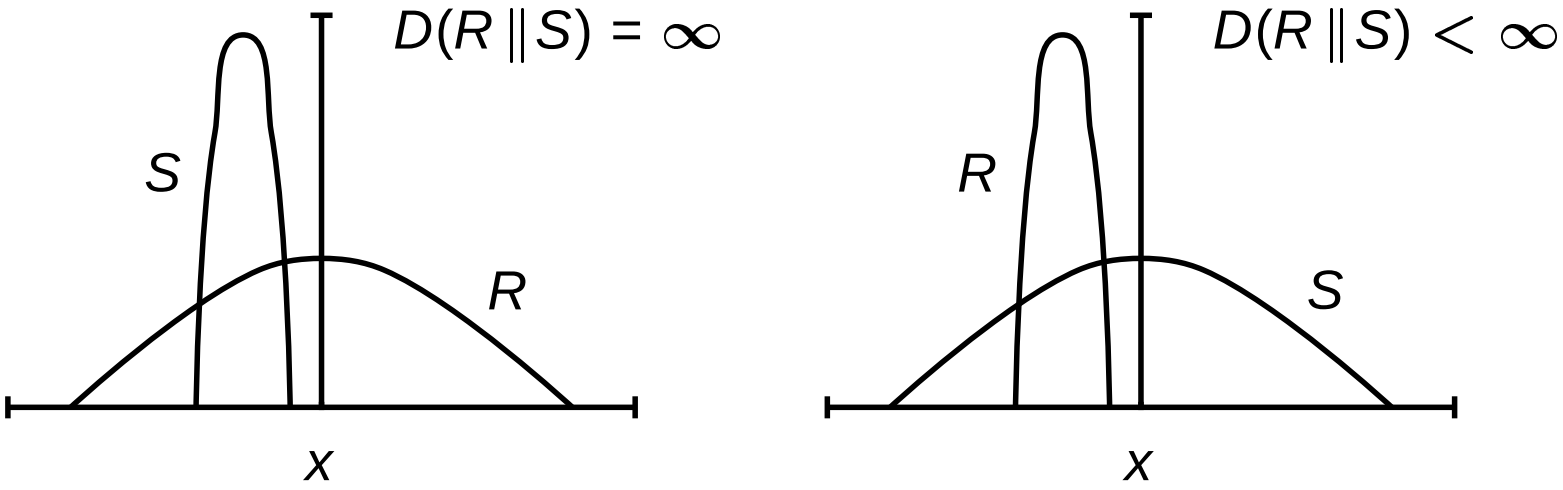
$$S \sim \text{Normal}\,(\mu_S, {\sigma_S}^2)$$

$$D(R \,\|\, S) = \frac{1}{2}\left(\frac{{\sigma_R}^2}{{\sigma_S}^2} - \log \frac{{\sigma_R}^2}{{\sigma_S}^2} - 1 \right) + \frac{1}{2} \frac{(\mu_R - \mu_S)^2}{{\sigma_S}^2}$$

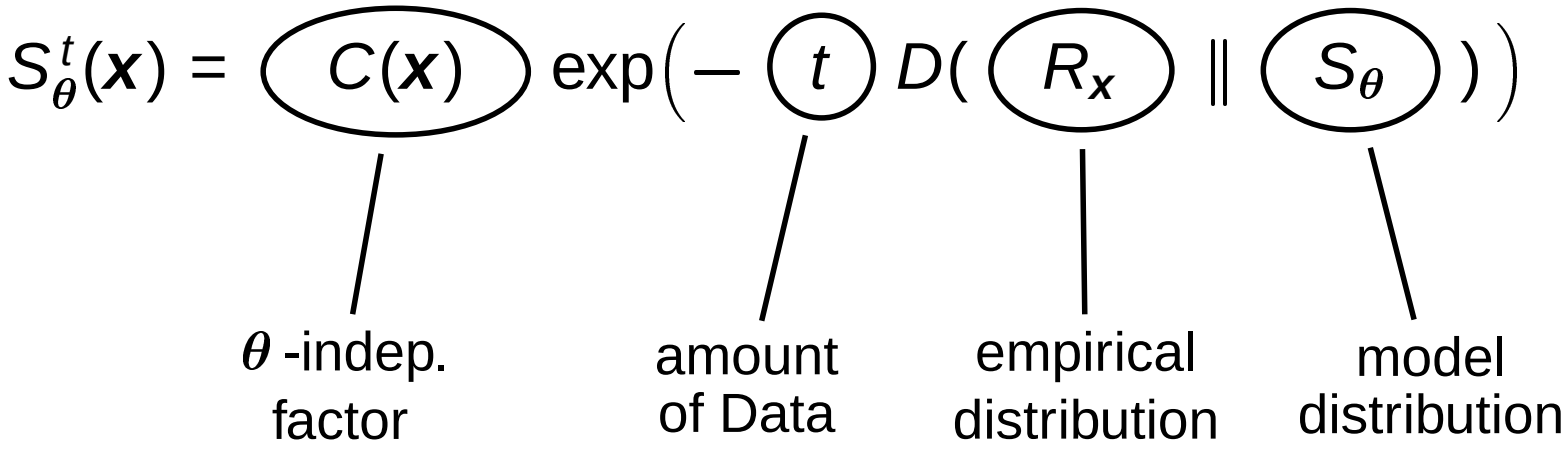
K.-L. Distance of Normal Distributions



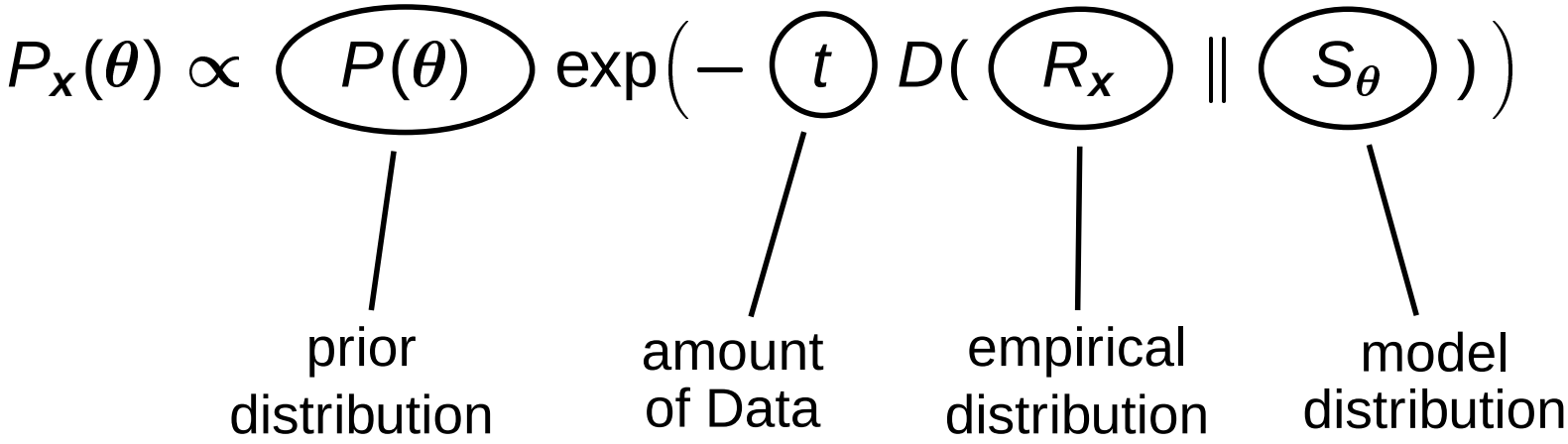
Distributions with Different Support



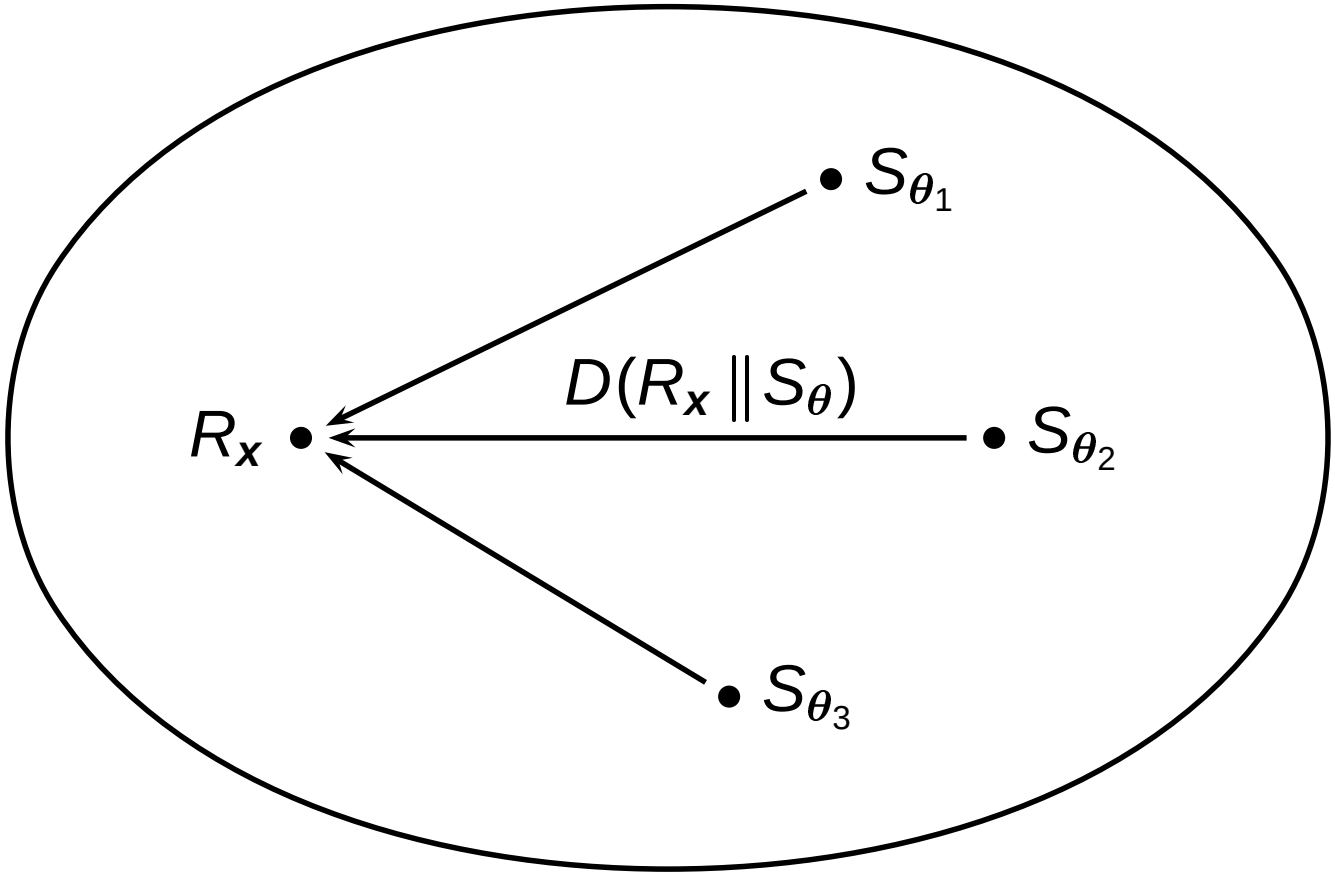
Joint Distribution of Sample



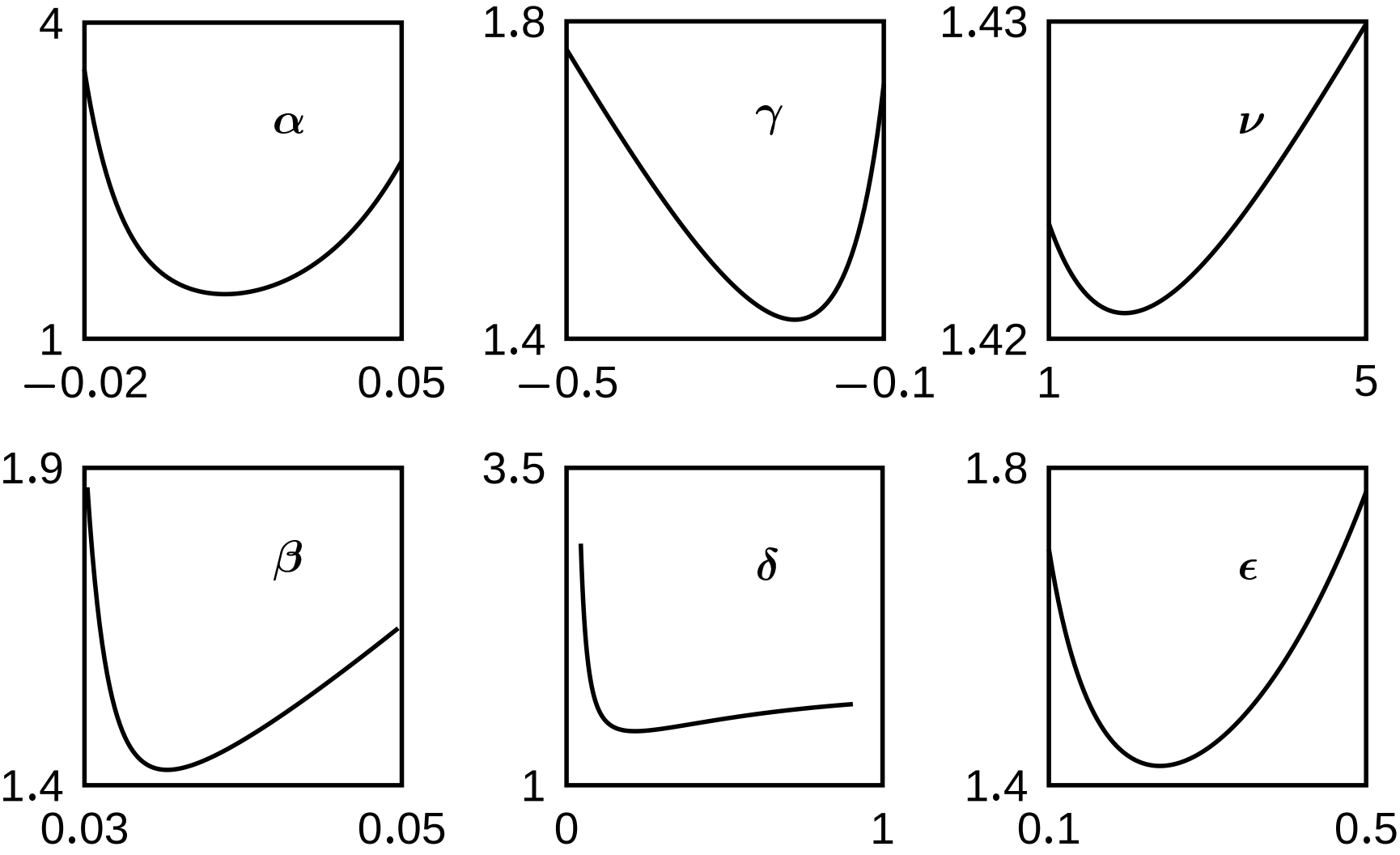
Posterior Distribution of Parameter



Minimum Distance Estimation

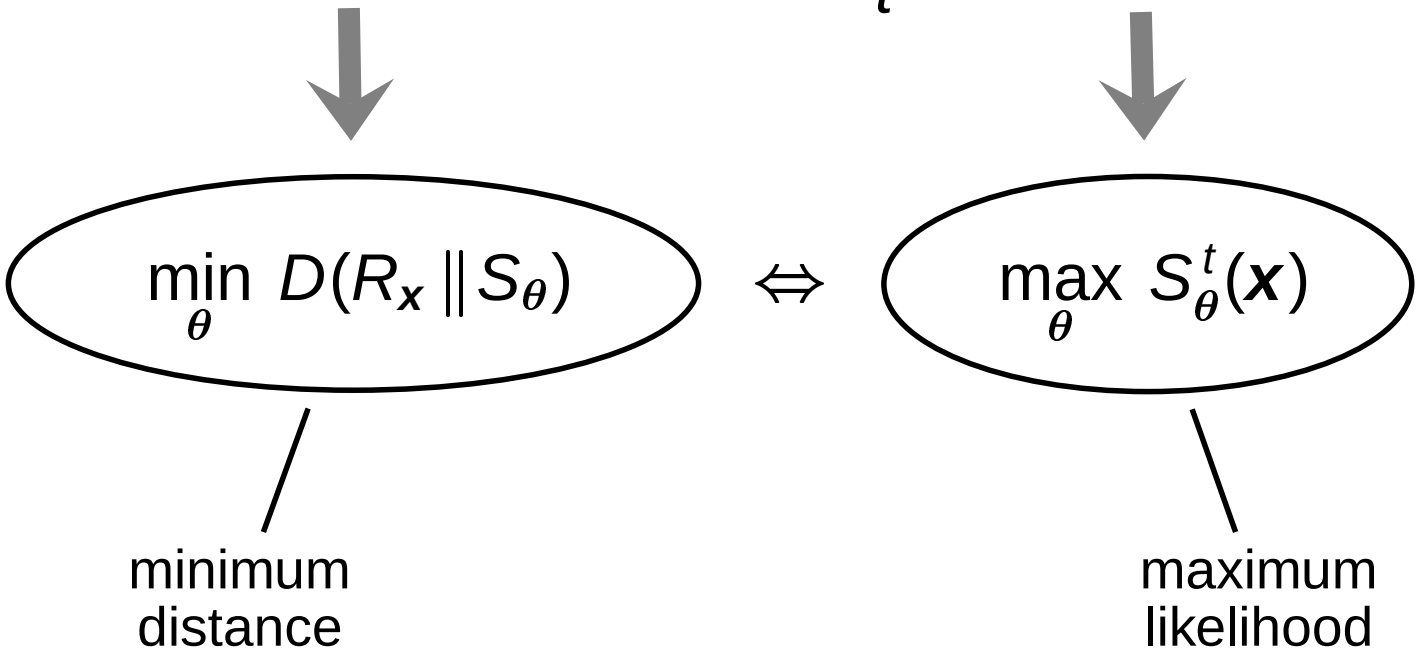


$$D\left(R_x \,\|\, \epsilon \text{ Normal}\,(\alpha, \beta^2) + (1-\epsilon) \text{ Log Student}\,(\gamma, \delta^2, \nu)\right)$$



K.-L. Distance and Likelihood

$$D(R_{\boldsymbol{x}} \,\|\, S_{\theta}) \;=\; C(\boldsymbol{x}) - \frac{1}{t} \log S_{\theta}^{t}(\boldsymbol{x})$$



Minimum Distance Estimation

"true" parametric distribution	empirical distribution
$\min_{\theta} D(S_{\theta_0} \| S_{\theta})$	$D(R_{\boldsymbol{x}} \| S_{\theta})$
$\Updownarrow$	$\Big\downarrow \; R_{\boldsymbol{x}} \to S_{\theta_0}$
$\max_{\theta} E_{S_{\theta_0}}\left(\log S_{\theta}\right)$	
$E_{S_{\theta_0}}(\cdot) \approx E_{R_{\boldsymbol{x}}}(\cdot)$	$D(S_{\theta_0} \| S_{\theta})$

What to do when the empirical distribution is only partially known?

Data Statistic

$$E_{R_{\boldsymbol{x}}}(h) = \frac{1}{t} \sum_{k=1}^{t} h(x_k) \triangleq \xi \quad \text{given} \quad h : \mathcal{X} \to \mathbf{R}^n$$

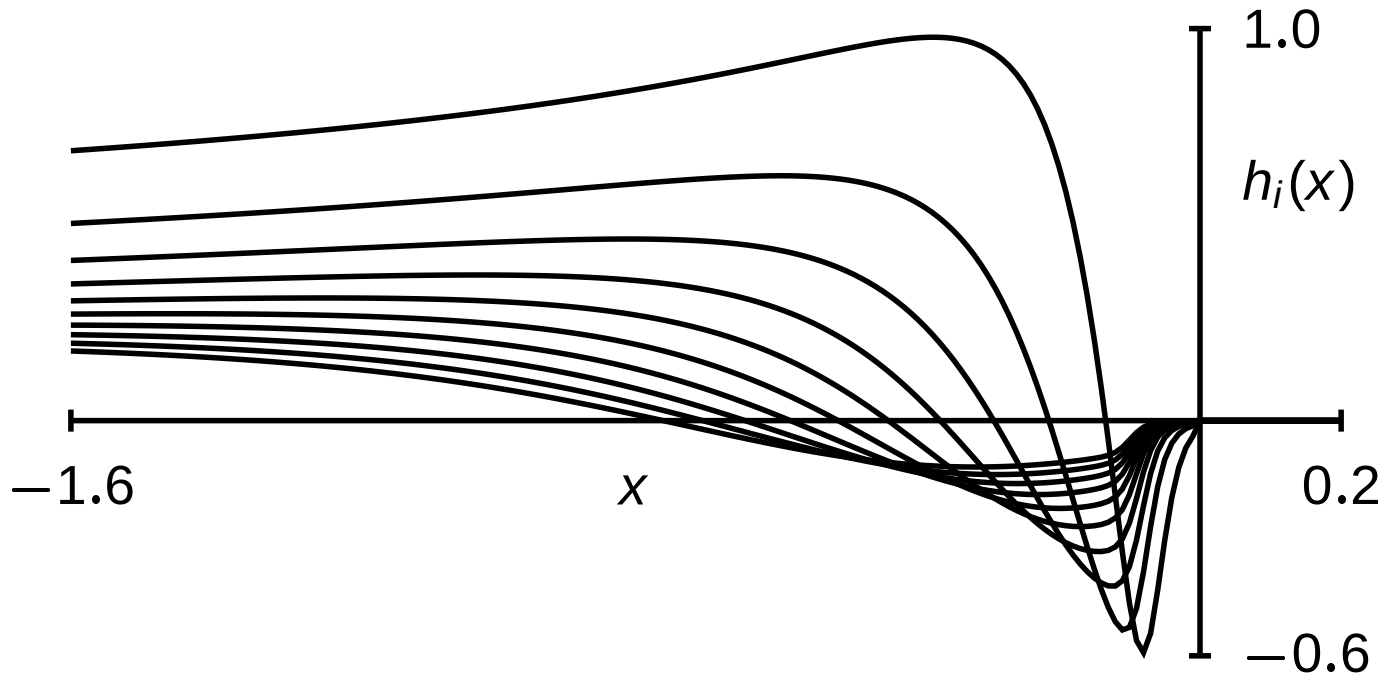
implies ambiguity

$$R_{\boldsymbol{x}} \in \mathcal{R}_{\xi} \triangleq \{R : E_R(h) = \xi\}$$

BALLOON: Choice of Statistic

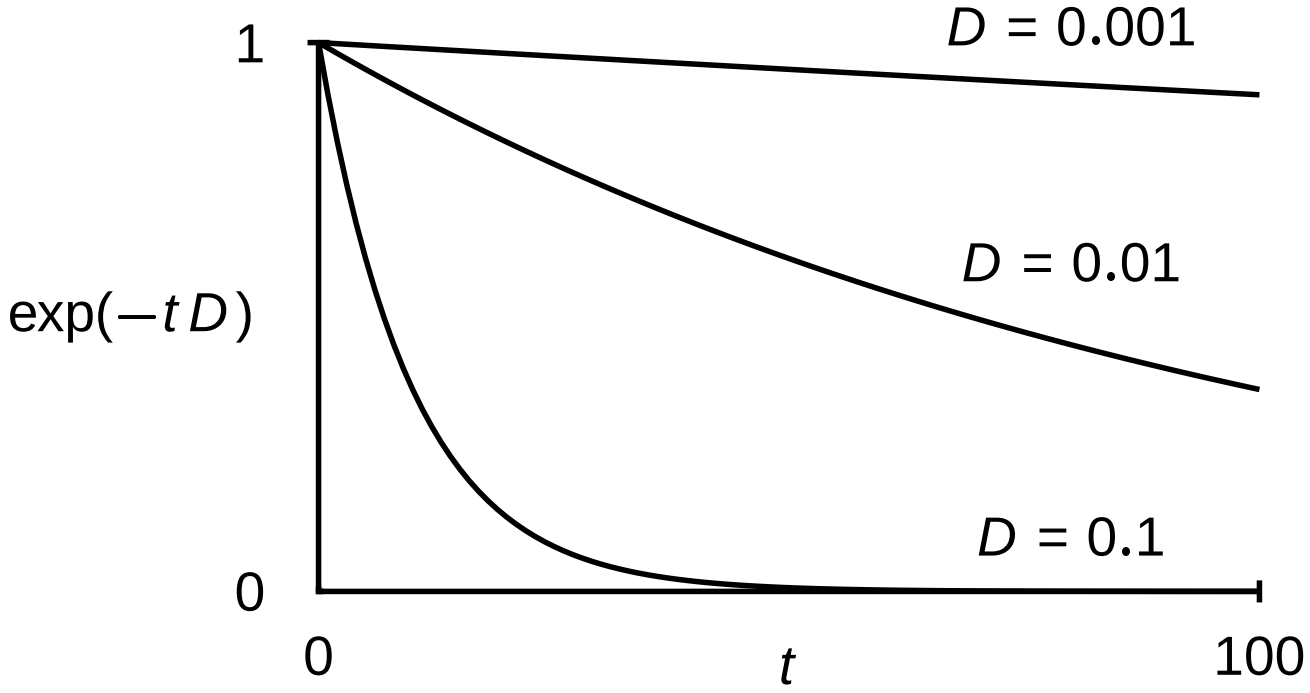
$$h_i(x) = \log S_{\gamma_i}(x) - \log S_{\gamma_{i+1}}(x), \quad i = 1, \ldots, n$$

$$E_{R_{\boldsymbol{x}}}(h_i) = \log S^t_{\gamma_i}(\boldsymbol{x}) - \log S^t_{\gamma_{i+1}}(\boldsymbol{x}), \quad i = 1, \ldots, n$$



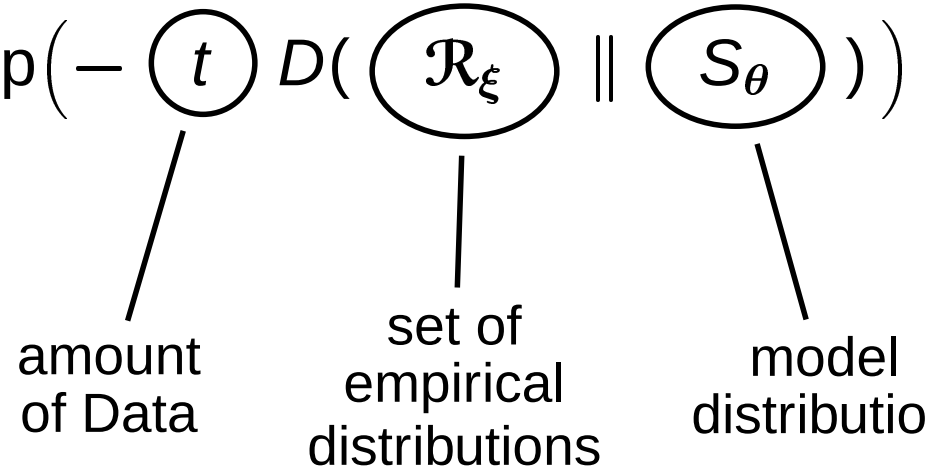
Probability of Large Deviations

$$S_\theta^t(R_{\boldsymbol{x}} \in \mathcal{R}_\xi) \;=\; \exp\Big(-\,t \min_{R \,\in\, \mathcal{R}_\xi} D(R \,\|\, S_\theta)\Big)\;\; \exp\Big(-\,t\, o(1)\Big)$$

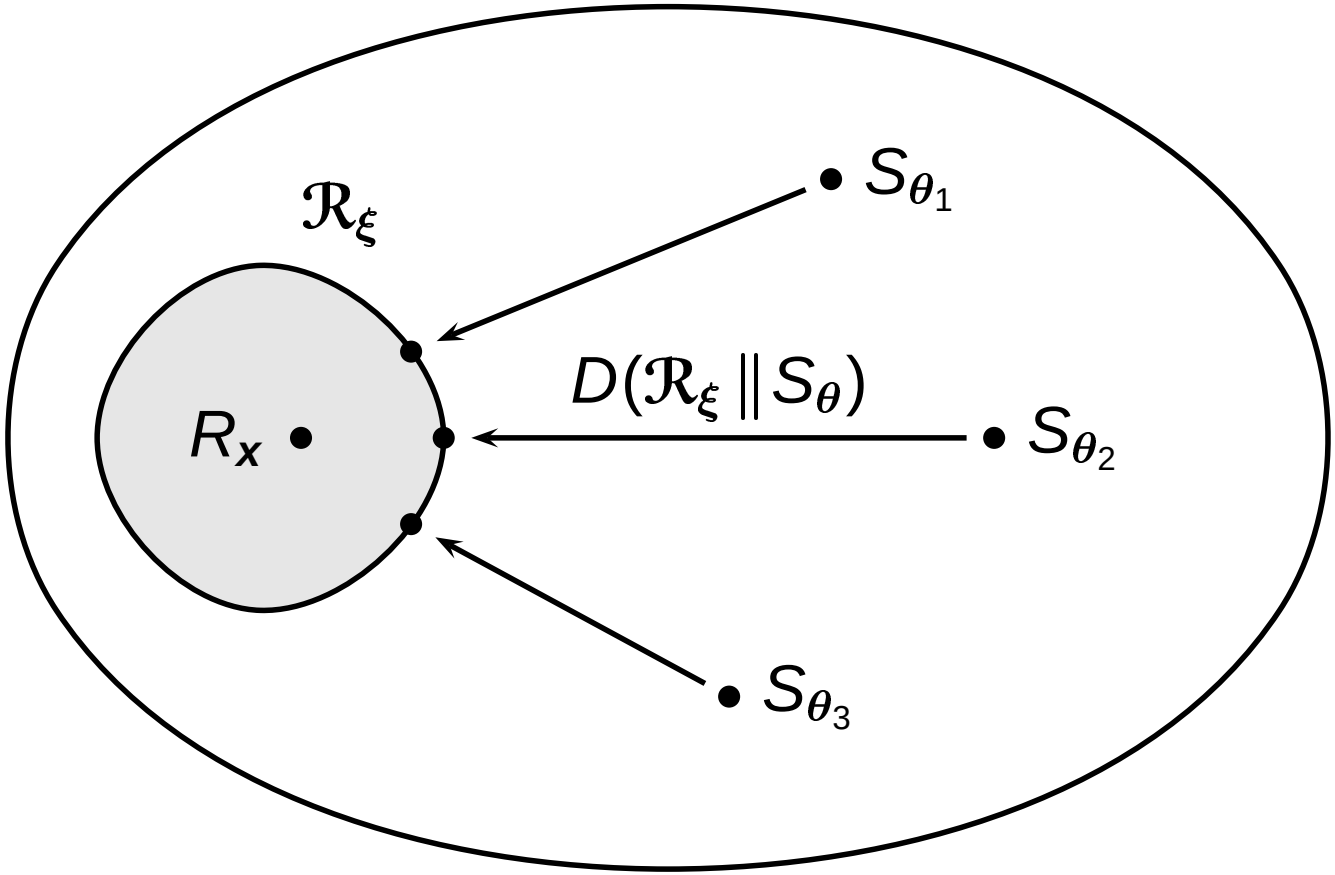


Joint Distribution of a Sample Set

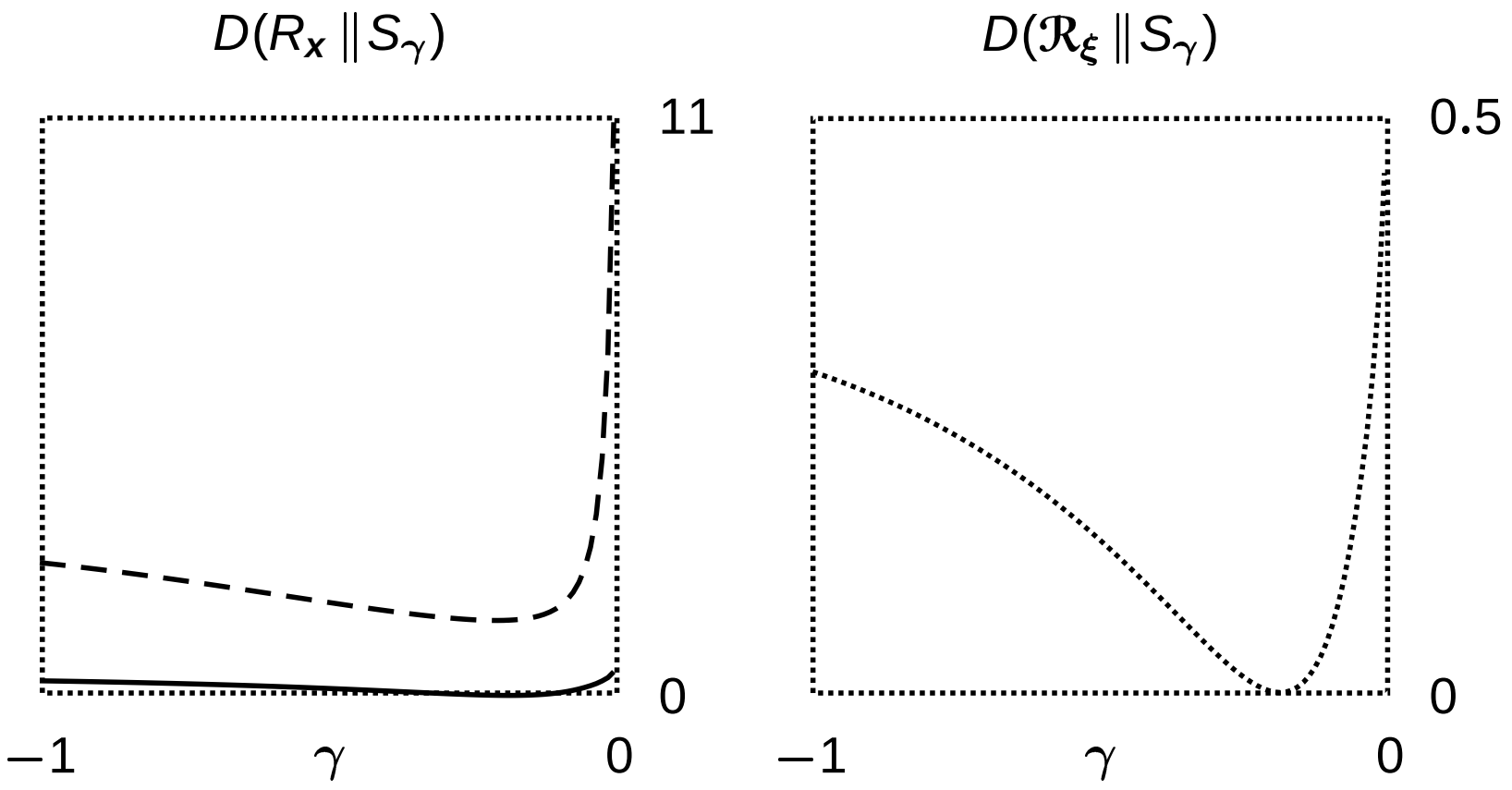
$$S_\theta^t(R_{\boldsymbol{x}} \in \mathcal{R}_\xi) \doteq \exp\Big(-\,t\; D(\,\mathcal{R}_\xi \,\|\, S_\theta\,)\Big)$$



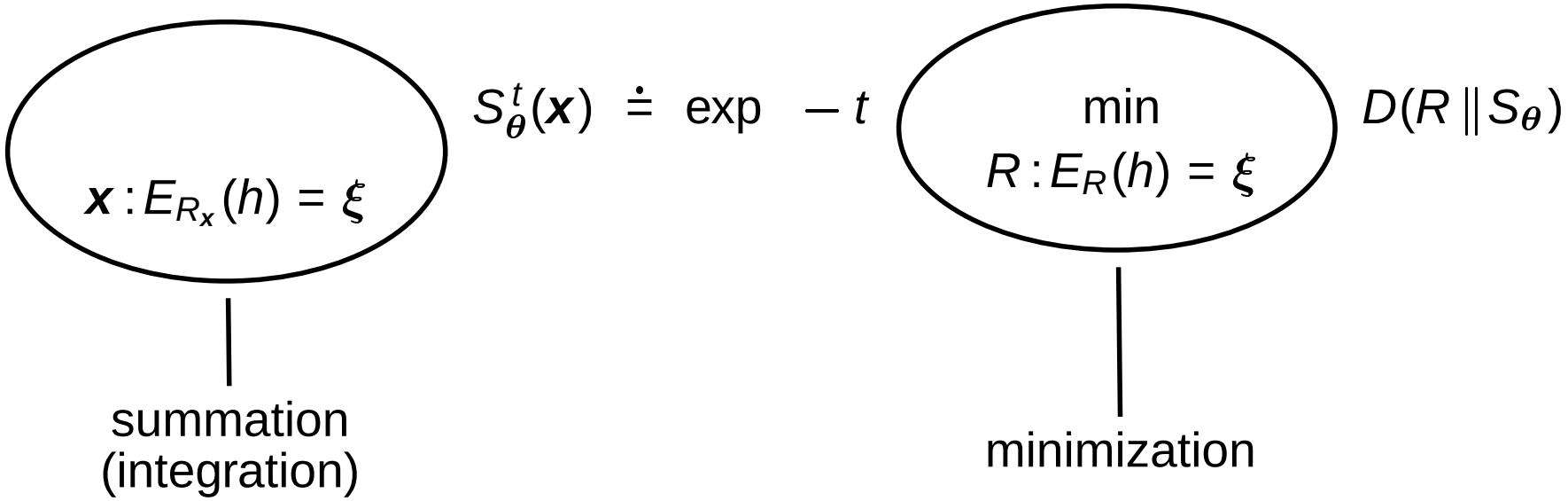
Minimum Distance Estimation



BALLOON: Minimum K.-L. Distance



Why the Large Deviation Approximation?



a significant drop in complexity!

Large Deviation Approximation

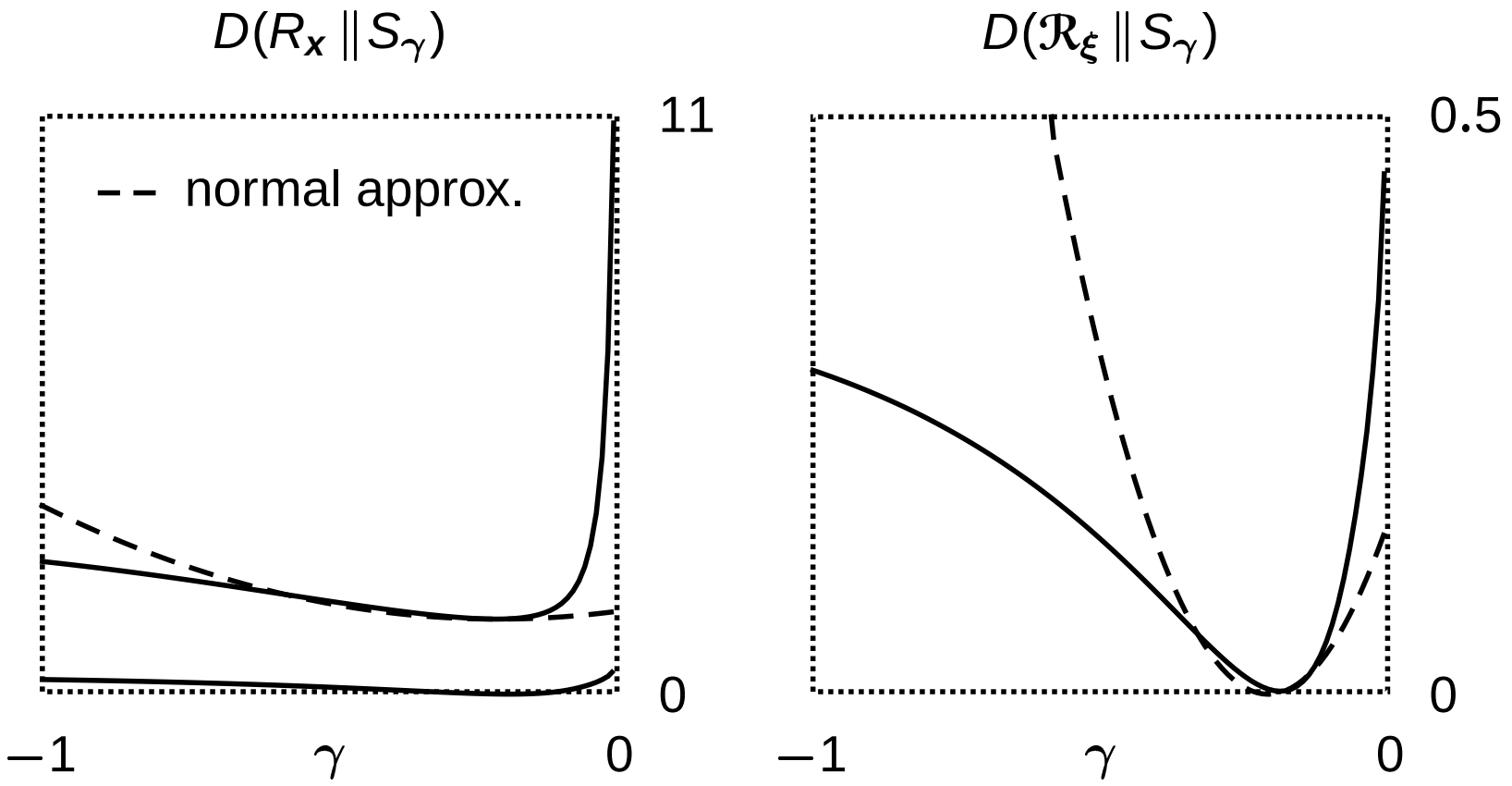
$$S_{\theta}^{t}(R_{\boldsymbol{x}} \in \mathcal{R}_{\xi}) = \exp\Big(-t \min_{R \in \mathcal{R}_{\xi}} D(R \,\|\, S_{\theta})\Big) \exp\Big(-t \; o(1)\Big)$$

Normal Approximation

$$S_{\theta}^{t}(\boldsymbol{x}) = \exp\Big(-t\, D(R_{\boldsymbol{x}} \,\|\, S_{\hat{\theta}})\Big) \exp\Big(-t \; \frac{1}{2} \; I(\hat{\theta})\,(\theta - \hat{\theta})^{2}\Big)$$

$$\cdot \exp\Big(-t \; O(|\theta - \hat{\theta}|^{3})\Big)$$

BALLOON: Minimum K.-L. Distance




How to deal with
dependent Data?

Example 2: SUNSPOTS

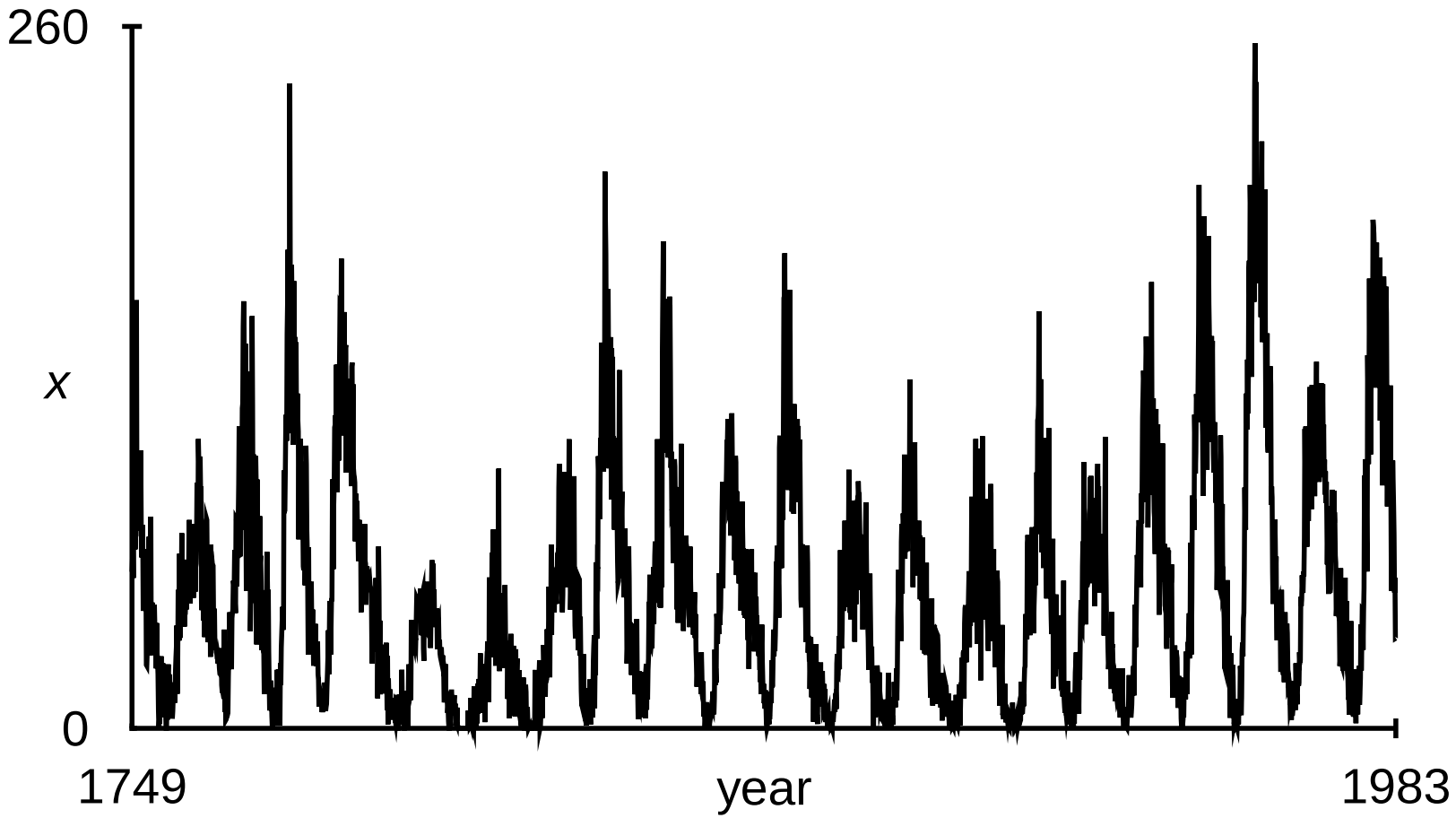
Zurich monthly sunspot numbers in 1749–1983

Source:

`http://lib.stat.cmu.edu/Datasets/Andrews/T11.1`

Andrews and Herzberg, DATA

SUNSPOTS: Time Series



Joint Distribution of Sample

model distribution $S_\theta(x_k \mid x_{k-1}, \ldots, x_{k-m})$

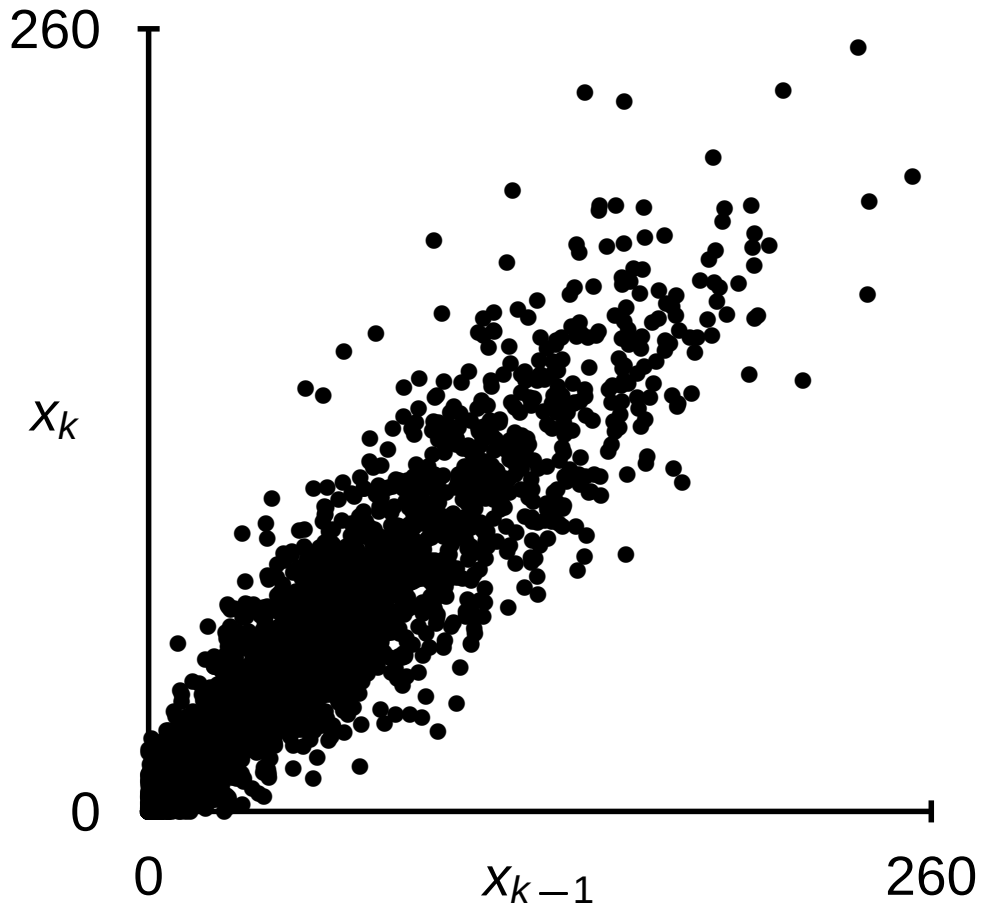
empirical distribution $R_{\boldsymbol{x}}(x_k, x_{k-1}, \ldots, x_{k-m})$

"weighting" distribution $W(x_{k-1}, \ldots, x_{k-m})$

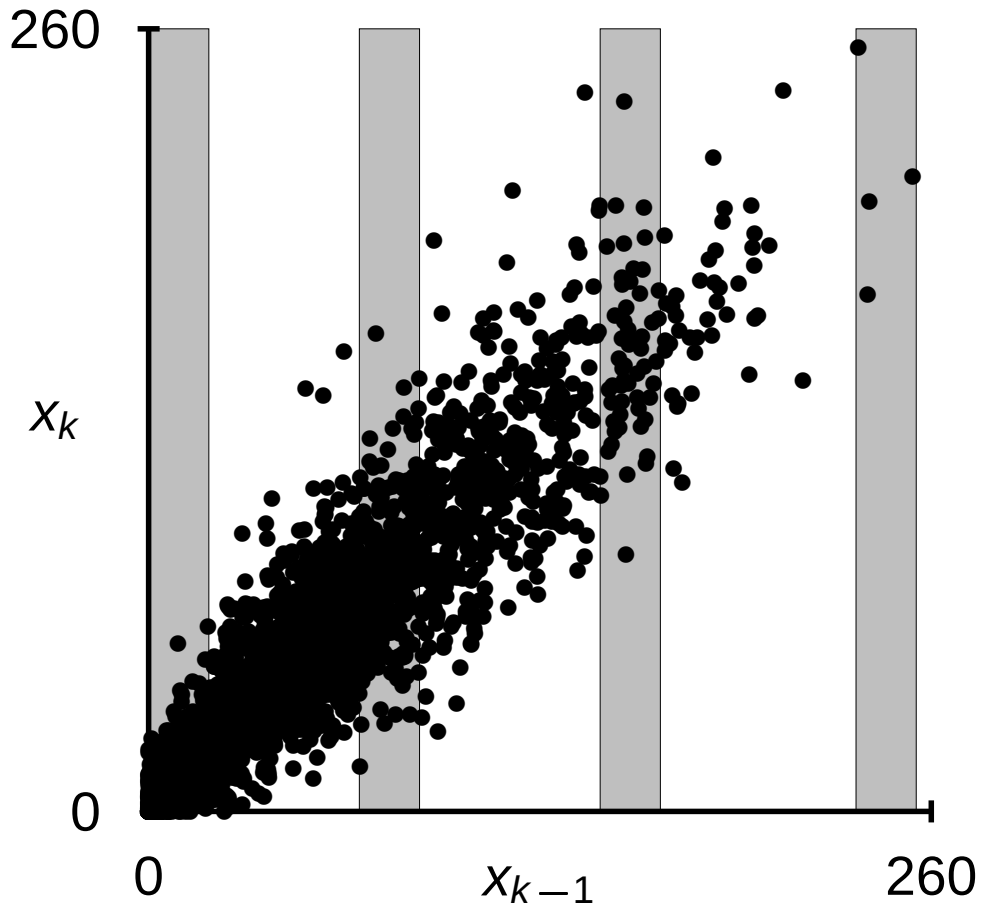
$$S_\theta^{t-m}(x_t, \ldots, x_{m+1} \mid x_m, \ldots, x_1) =$$

$$= C(\boldsymbol{x}) \exp\left(- (t-m) \min_W D(R_{\boldsymbol{x}} \| S_\theta W) \right)$$

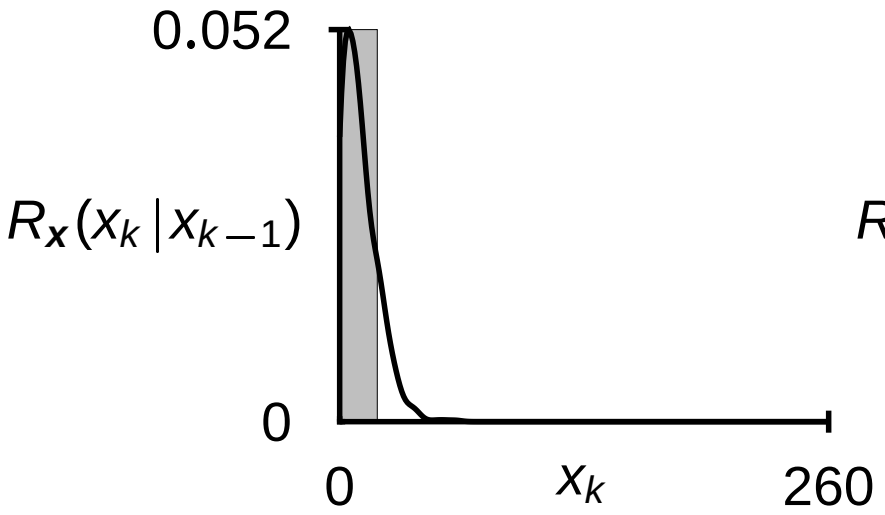
SUNSPOTS: Empirical Distribution

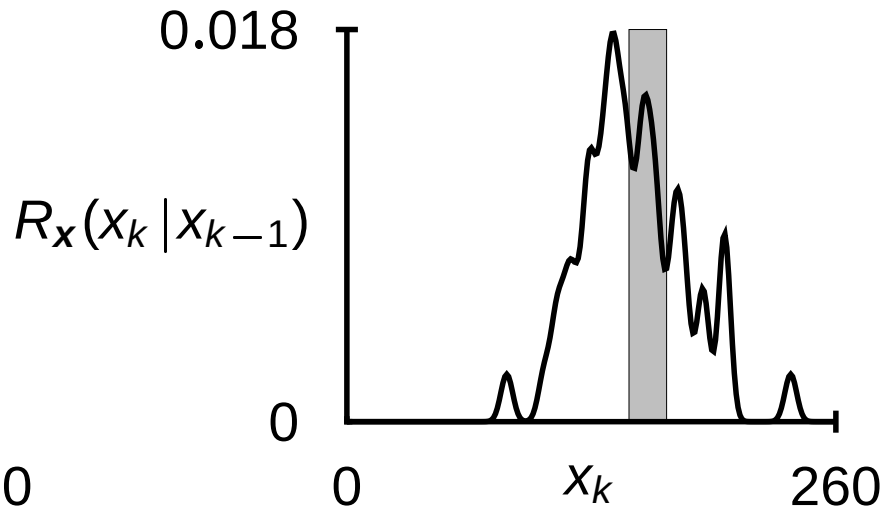


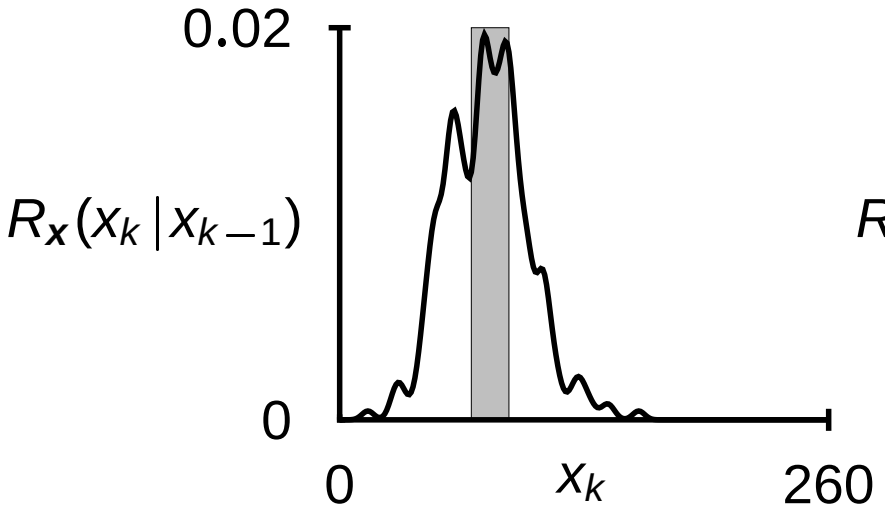
SUNSPOTS: Empirical Distribution

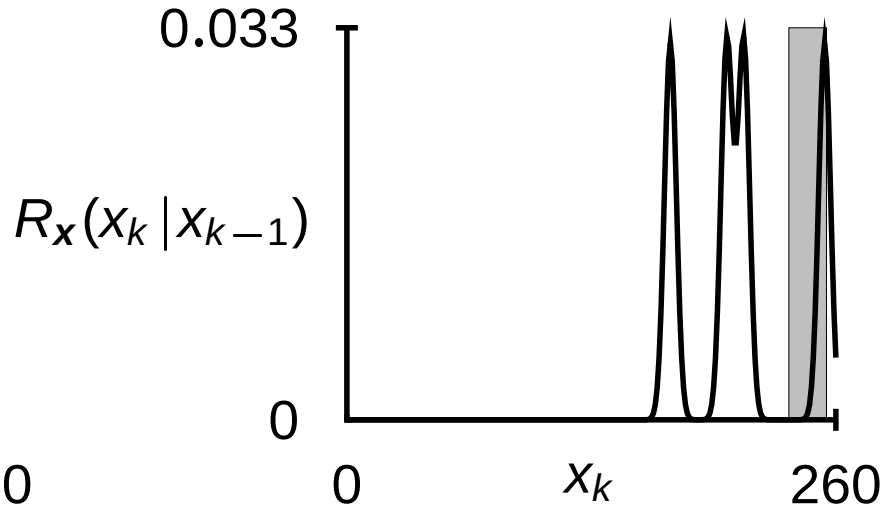


SUNSPOTS: Conditional Empirical Distrib's









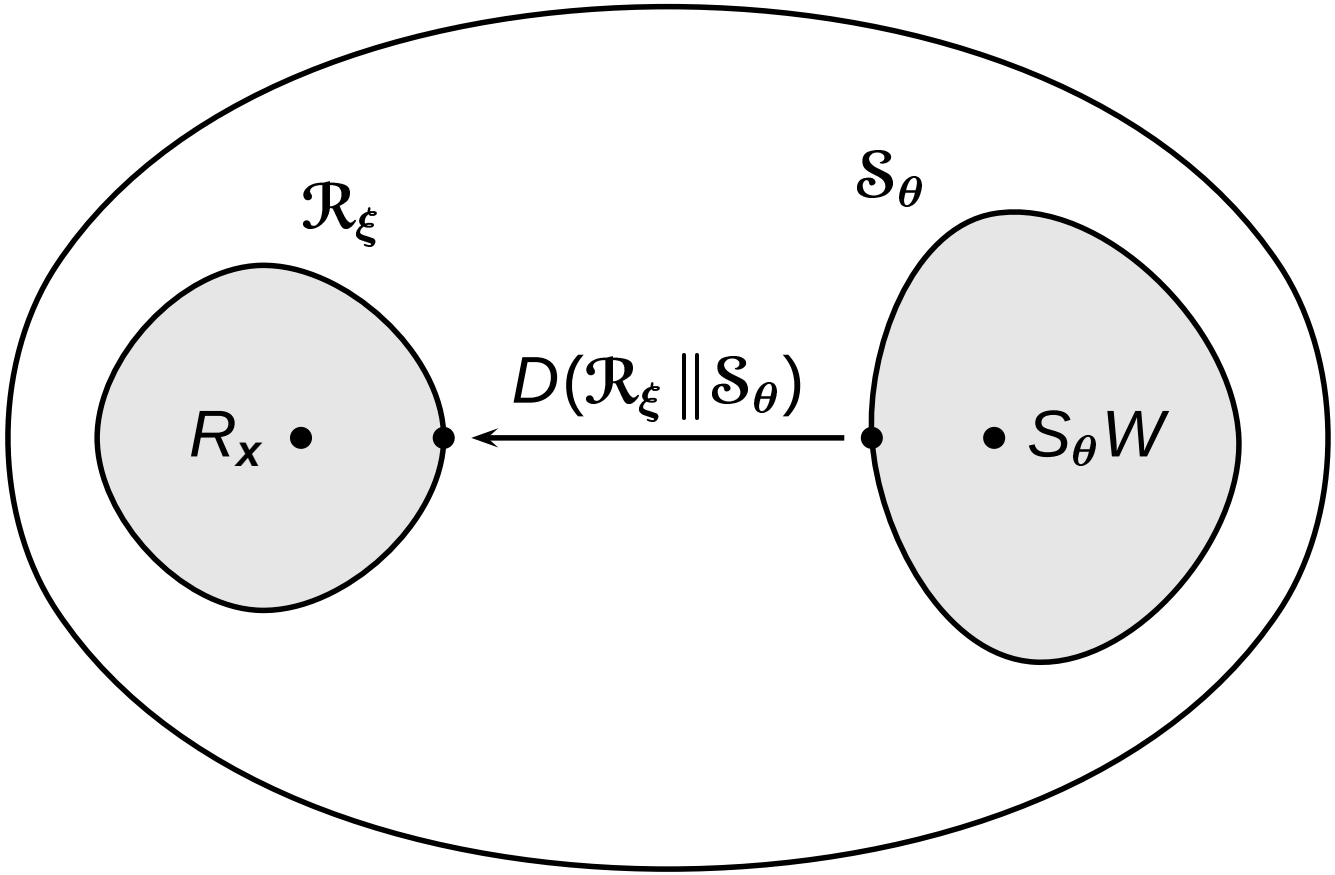
Probability of Large Deviations

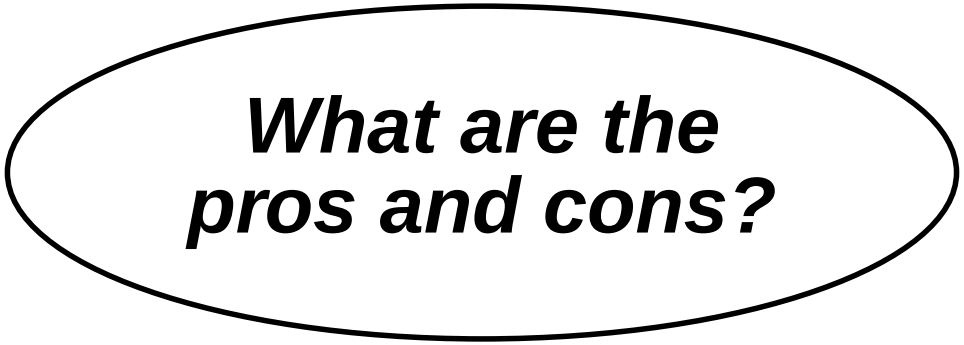
$$\text{given } E_{R_{\boldsymbol{x}}}(h) = \frac{1}{t} \sum_{k=m+1}^{t} h(x_k, x_{k-1}, \ldots, x_{k-m}) \triangleq \boldsymbol{\xi}$$

$$\text{where } h : \mathcal{X}^{m+1} \to \mathbf{R}^n$$

$$S_{\boldsymbol{\theta}}^{t}(R_{\boldsymbol{x}} \in \mathcal{R}_{\boldsymbol{\xi}}) \doteq \exp\Big(-t \min_{R \in \mathcal{R}_{\boldsymbol{\xi}}} \min_{W} D(R \,\|\, S_{\boldsymbol{\theta}} W)\Big)$$

Minimum Distance Estimation




What are the
pros and cons?

K.-L. Distance

- ❏ minimum distance
  - ⇨ optimization

☛ *computationally easier*

Probability

- ❏ marginal probability
  - ⇨ summation (integration)

K.-L. Distance

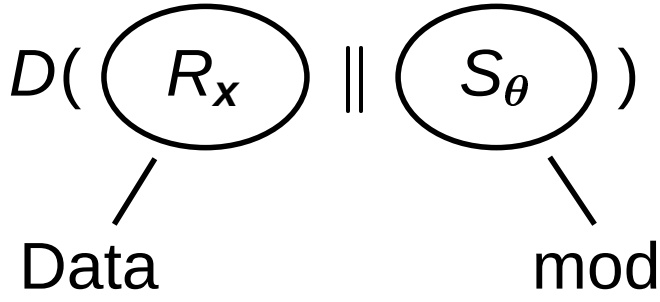
- ❏ approximation theory

Probability

- ❏ probability theory & statistics

- ☛ *conceptually simpler*
- ☛ *ready for further approximation*

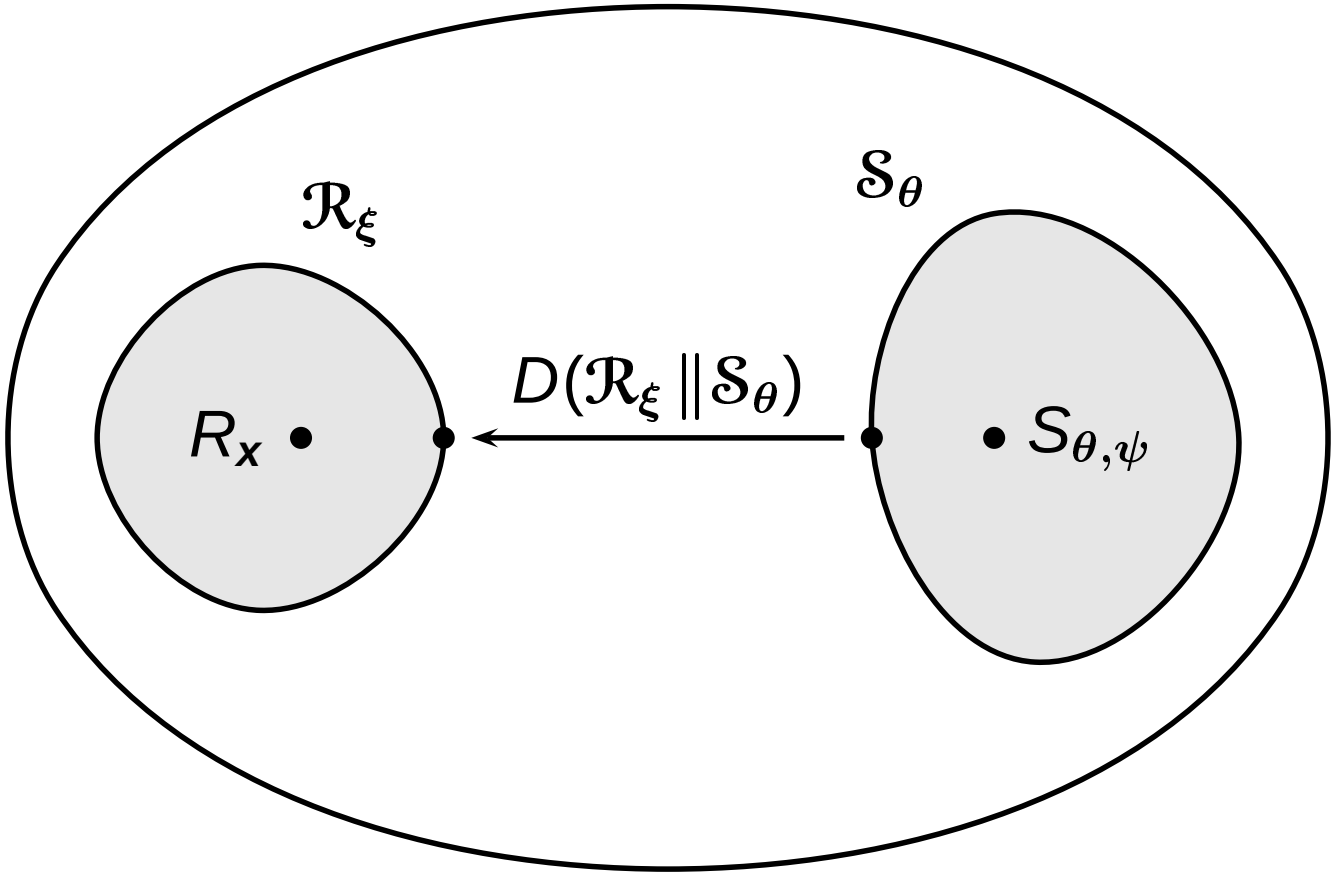
Symmetric Role of Data & Model



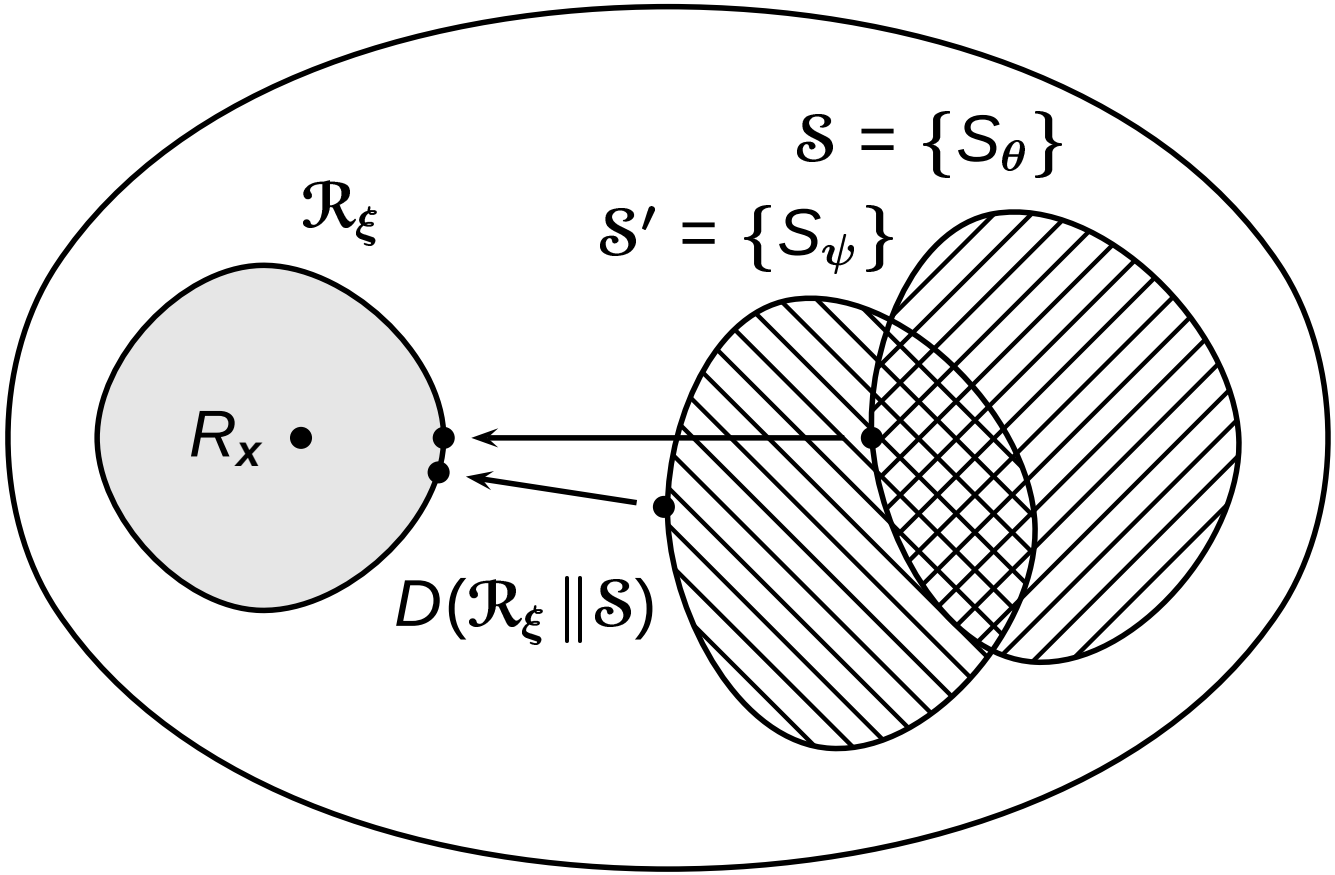
admits straightforward extensions

$$D\left(\ \{R : E_R(h) = \xi\}\ \|\ \{S_{\theta,\psi} : \psi \in \mathcal{N}\}\ \right)$$

Robust Identification



Adaptive Identification



K.-L. Distance

- ❏ *absolute* measure of goodness-of-fit

$$D(R_{\boldsymbol{x}} \,\|\, S_{\theta}) \Leftrightarrow S_{\theta}^{t}(\boldsymbol{x})$$

$$D(\mathcal{R}_{\boldsymbol{\xi}} \,\|\, S_{\theta}) \Leftrightarrow S_{\theta}^{t}(\mathcal{R}_{\boldsymbol{\xi}})$$

Posterior Prob.

- ❏ *relative measure* of goodness-of-fit

$$P_{\boldsymbol{x}}(S_{\theta}) \propto P(\theta)\, S_{\theta}^{t}(\boldsymbol{x})$$

$$P_{\boldsymbol{\xi}}(S_{\theta}) \propto P(\theta)\, S_{\theta}^{t}(\mathcal{R}_{\boldsymbol{\xi}})$$

A Big Open Issue

approximation of Kullback-Leibler distance

$$D(\mathcal{R}_\xi \,\|\, S_\theta) \doteq ?$$

- $D(\mathcal{R}_\xi \,\|\, S_\theta) \doteq D(\mathcal{R}_\xi \,\|\, \hat{S}_\theta)$, $\{\hat{S}_\theta\}$ exponential
- $D(\mathcal{R}_\xi \,\|\, S_\theta) \doteq \| E_{R_x}(h) - E_{S_\theta}(h) \|_W^2$