Bayesian Estimation, Large Deviations, and Incomplete Data

Rudolf Kulhavý

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

Outline

- Bayesian Estimation
- Data Compression
- Approximate Posterior Distribution
- Key Properties of Approximation
- Illustrative Example
- Concluding Remarks

Bayesian Estimation

sequence of random variables

$$\boldsymbol{X} = (X_1, \ldots, X_k), \quad X_i \in \mathfrak{X}, \quad \mathfrak{X} \text{ finite}$$

independent and identically distributed

$$S_\theta(x), \quad \theta \in \mathfrak{T}, \quad \mathfrak{T} \text{ finite}, \quad S_\theta(x) > 0$$

posterior distribution of Θ

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

Some Notions

type of sequence $\boldsymbol{x} = (x_1, \ldots, x_k)$

$$R_{\boldsymbol{x}}(a) = \frac{1}{k} N_{\boldsymbol{x}}(a) = \frac{1}{k} \sum_{i=1}^{k} \mathbf{I}_{\boldsymbol{x}_k}(a)$$

Shannon entropy

$$H(R) = \sum_{a \in \mathcal{X}} R(a) \log \frac{1}{R(a)}$$

relative entropy, Kullback-Leibler distance, I-divergence

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

Distribution of Sample

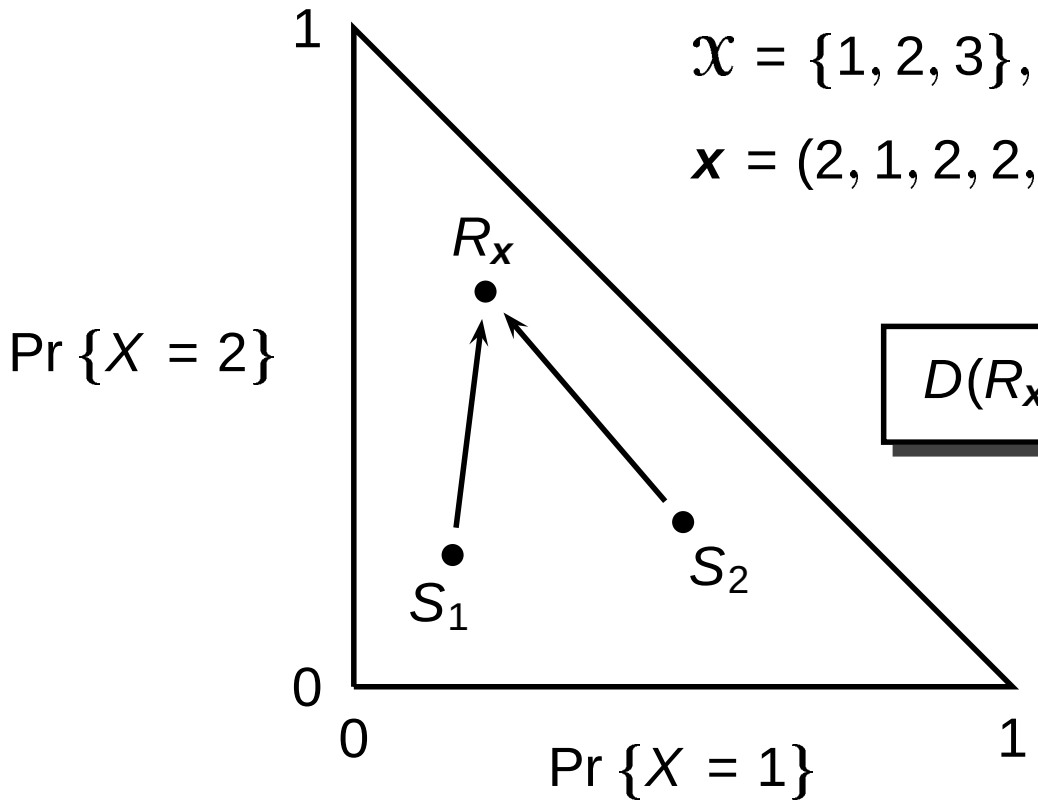
joint distribution of $\boldsymbol{X} = (X_1, \ldots, X_k)$

$$S_\theta^k(\boldsymbol{x}) = \prod_{i=1}^{k} S_\theta(x_i)$$

$$= \exp \sum_{i=1}^{k} \log S_\theta(x_i)$$

$$= \exp\{-k\, H(R_{\boldsymbol{x}})\}\ \exp\{-k\, D(R_{\boldsymbol{x}} \,\|\, S_\theta)\}$$

Probability Simplex

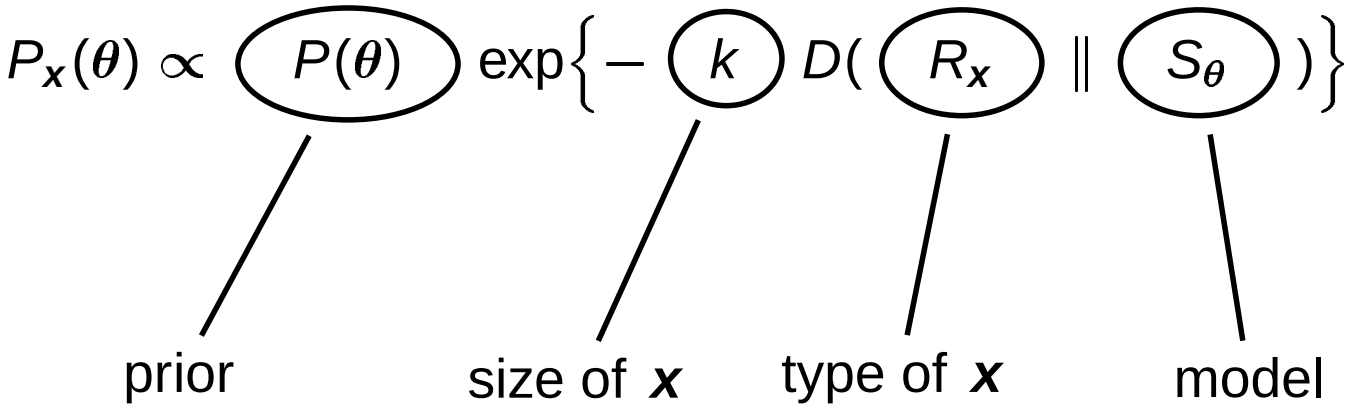


$\mathcal{X} = \{1, 2, 3\}, \ \mathcal{T} = \{1, 2\}$

$\boldsymbol{x} = (2, 1, 2, 2, 3) \Rightarrow R_{\boldsymbol{x}} = \frac{1}{5}(1, 3, 1)$

$D(R_{\boldsymbol{x}} \| S_\theta)$

Posterior Distribution



Data Compression

data statistic

$$T(\boldsymbol{X}) = \frac{1}{k} \sum_{j=1}^{k} h(X_j) = E_{R_{\boldsymbol{X}}}\, h(X), \quad h : \mathfrak{X} \mapsto \mathbb{R}^n$$

a trivial example

$$\mathfrak{X} \subset \mathbb{R},\ h_1(X) = X,\ h_2(X) = X^2$$

partial knowledge of $\boldsymbol{X}$

$$T(\boldsymbol{X}) \geq \xi \quad \Rightarrow \quad \boxed{P_\xi(\theta) \propto P(\theta)\, S_\theta^k(\xi)}$$

Distribution of Statistic

set of $\boldsymbol{x}$'s consistent with the statistic

$$\mathcal{X}_\xi = \{\boldsymbol{x} : T(\boldsymbol{x}) \geq \xi\} = \cup_{R \in \mathcal{R}_\xi} \mathcal{X}_R$$

where $\mathcal{X}_R = \{\boldsymbol{x} : R_{\boldsymbol{x}} = R\}$ and $\mathcal{R}_\xi = \{R : E_R h \geq \xi\}$

probability of all such $\boldsymbol{x}$'s

$$S_\theta^k(\xi) = S_\theta^k(\mathcal{X}_\xi) = \sum_{\boldsymbol{x} \in \mathcal{X}_\xi} S_\theta^k(\boldsymbol{x}) = \sum_{R \in \mathcal{R}_\xi} \sum_{\boldsymbol{x} \in \mathcal{X}_R} S_\theta^k(\boldsymbol{x}) =$$

$$= \sum_{R \in \mathcal{R}_\xi} |\mathcal{X}_R| \exp\{-k\, H(R)\} \exp\{-k\, D(R \| S_\theta)\}$$

Approximate Distribution of Statistic

equality to the 1st order in the exponent

$$S^k \doteq \hat{S}^k \quad \text{if} \quad \lim_{k\to\infty} \frac{1}{k} \log \frac{S^k}{\hat{S}^k} = 0$$

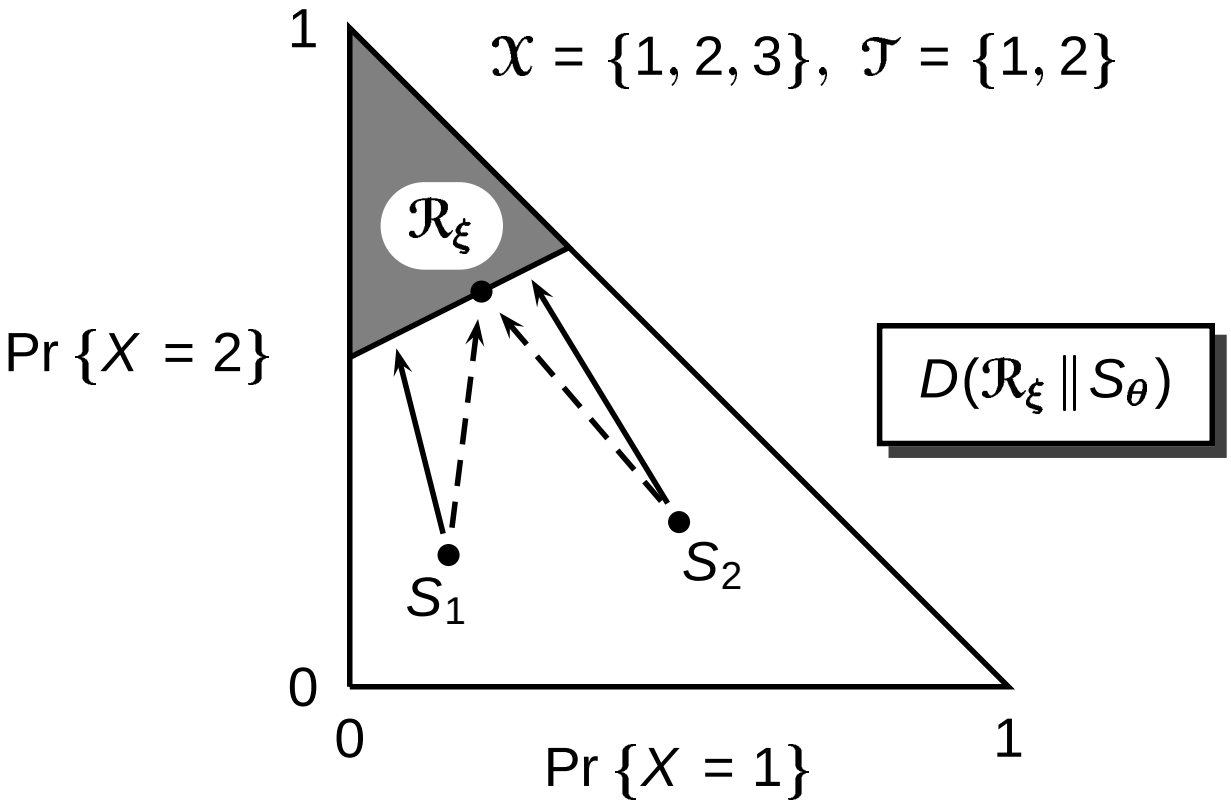
approximate expression for $S_\theta^k(\xi)$

$$S_\theta^k(\xi) \doteq \exp\left\{-k\, D(\mathcal{R}_\xi \,\|\, S_\theta)\right\},$$

$$D(\mathcal{R}_\xi \,\|\, S_\theta) = \min_{R\in\mathcal{R}_\xi} D(R \,\|\, S_\theta)$$

large deviation theorem (Sanov, 1957)

Probability Simplex



Main Result:

Approximate Posterior Distribution

$$\hat{P}_{\boldsymbol{x}}(\theta) \propto P(\theta) \exp\left\{ - k \, D(\mathcal{R}_{\xi} \,\|\, S_{\theta}) \right\}$$

prior

size of $\boldsymbol{x}$

types of $\boldsymbol{x}$'s consistent with $T(\boldsymbol{x}) \geq \xi$

model

Key Properties of Approximation

meaningful bounds

$$0 \leq D(\mathcal{R}_\xi \| S_\theta) \leq D(R_{\boldsymbol{x}} \| S_\theta)$$

monotonicity

$$\mathcal{R}_\xi \subset \mathcal{R}'_\xi \Rightarrow D(\mathcal{R}'_\xi \| S_\theta) \leq D(\mathcal{R}_\xi \| S_\theta)$$

extreme cases

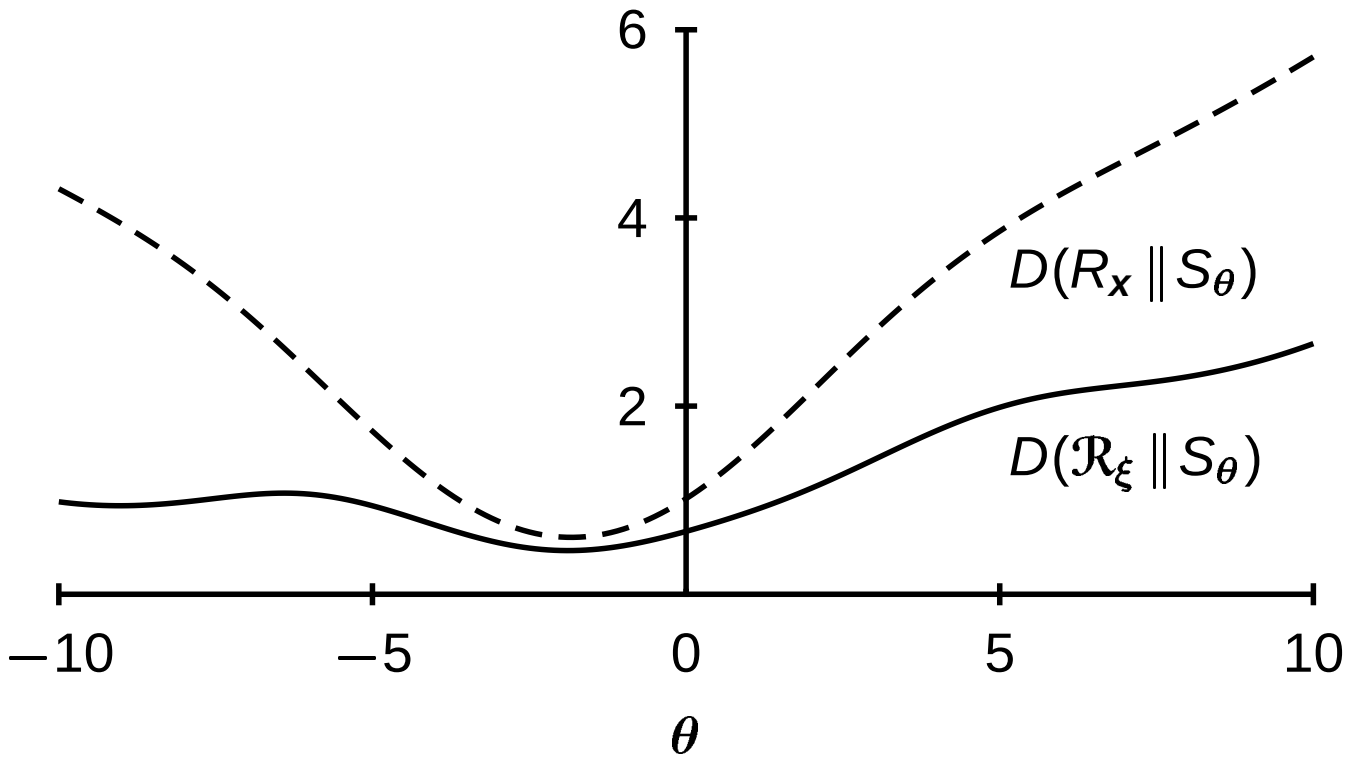
$$\mathcal{R}_\xi = \{R_{\boldsymbol{x}}\} \Rightarrow D(\mathcal{R}_\xi \| S_\theta) = D(R_{\boldsymbol{x}} \| S_\theta) \Rightarrow \hat{P}_\xi(\theta) = P_{\boldsymbol{x}}(\theta)$$

$$\mathcal{R}_\xi = \mathcal{R} \Rightarrow D(\mathcal{R}_\xi \| S_\theta) = 0 \Rightarrow \hat{P}_\xi(\theta) = P(\theta)$$

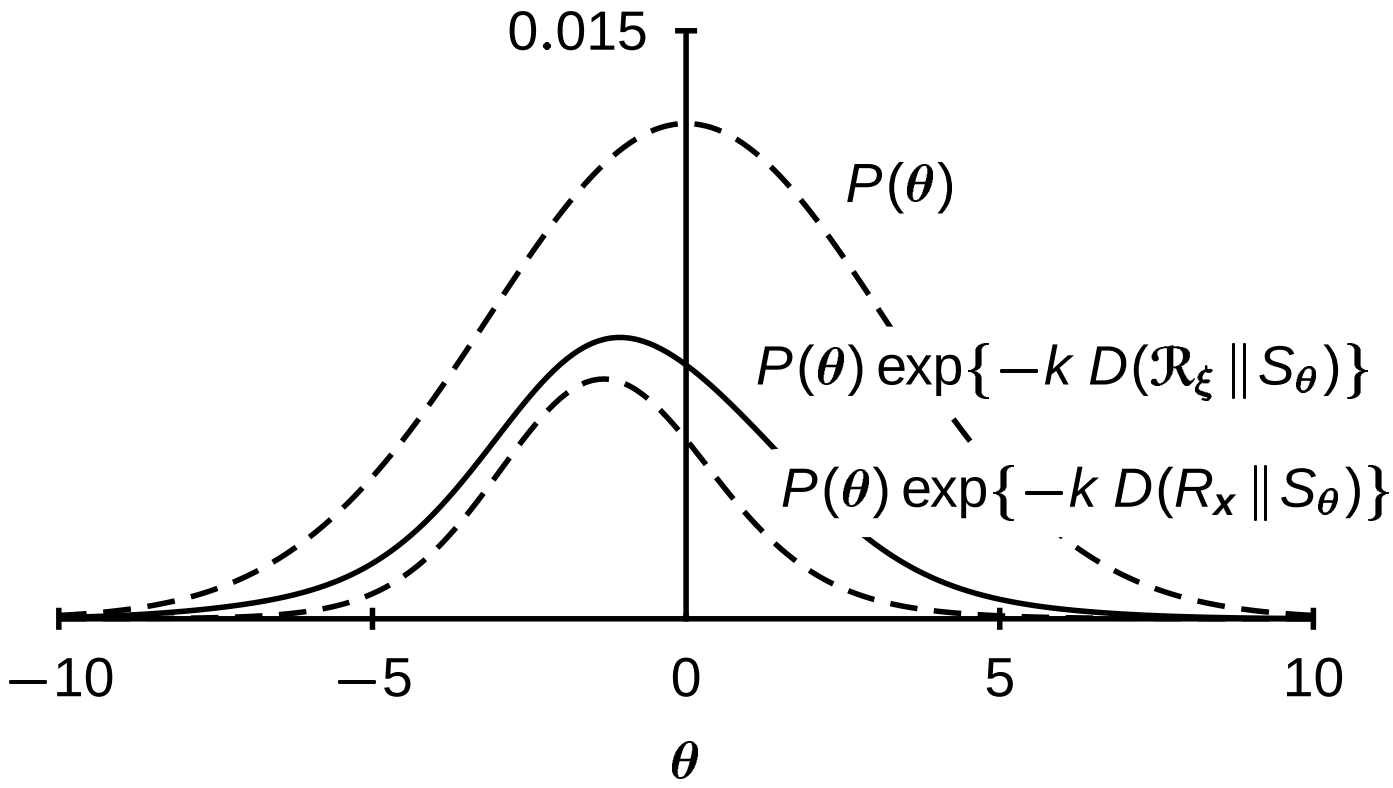
asymptotic consistency

$$R_{\boldsymbol{x}} \to S_{\theta_0} \Rightarrow D(\mathcal{R}_\xi \| S_{\theta_0}) \to 0$$

Example: Relative Entropy



Example: Unnormalized Posterior



Concluding Remarks

- a way of coping with incomplete data

  $D(\mathcal{R}_\xi \,\|\, S_\theta) = \min_{R \in \mathcal{R}_\xi} D(R \,\|\, S_\theta)$

- general approach

  i.i.d. data, Markov chains, contin. θ, contin. x

- curse of dimensionality

  computational complexity of $D(\mathcal{R}_\xi \,\|\, S_\theta)$

- shift of paradigm

  from $P_{\boldsymbol{x}}(\theta)$ *to* $D(\mathcal{R}_\xi \,\|\, S_\theta)$