On Dual Expression of Prior Information in Bayesian Parameter Estimation

R. Kulhavý and L. Tesař

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

Outline

- **Bayesian estimation**
  - *Statistical view*
  - *Information view*
- **Conjugate prior**
  - *A priori knowledge of distribution of data*
- **Minimum-distance prior**
  - *Partial knowledge of distribution of data*
- **Summary**

Likelihood Based Inference

- General regression

$$y_k, z_k = z(u^k, y^{k-1}), \quad k = m+1, \ldots, N+m$$

- Model

$$s_\theta(y|z), \quad \theta \in \Theta \subset R^m$$

- Likelihood function

$$l_N(\theta) = q_\theta(y_{m+1}^{N+m}, u_{m+1}^{N+m} \mid y^m, u^m) \propto \prod_{k=m+1}^{N+m} s_\theta(y_k | z_k)$$

- Posterior density

$$p_N(\theta) \propto p_0(\theta)\, l_N(\theta)$$

Information Based Inference

- **Empirical density**

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

- **Conditional inaccuracy**

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

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

- **Likelihood**

$$l_N(\theta) \propto \exp(-N K(r_N : s_\theta))$$

- **Posterior density**

$$p_N(\theta) \propto p_0(\theta) \exp(-N K(r_N : s_\theta))$$

General Form of Conjugate Prior

$$p_\nu(\theta) \propto \exp(-\nu K(r_\nu : s_\theta))$$

$\nu > 0$

Degree of belief in the prior

$r_\nu(y,z)$

"Prior" density of data

Conjugate Prior

$$\begin{cases} p_\nu(\theta) \propto \exp(-\nu K(r_\nu : s_\theta)) & \text{... Prior} \\ l_N(\theta) \propto \exp(-N K(r_N : s_\theta)) & \text{... Likelihood} \end{cases}$$

$$p_{\nu+N}(\theta) \propto \exp\left(-(\nu+N)\, K(r_{\nu+N} : s_\theta)\right) \quad \text{... Posterior}$$

$$r_{\nu+N}(y,z) = \frac{\nu}{\nu+N}\, r_\nu(y,z) + \frac{N}{\nu+N}\, r_N(y,z)$$

Coin Tossing

Model density

Head

$$s_\theta(y=1)=\theta$$

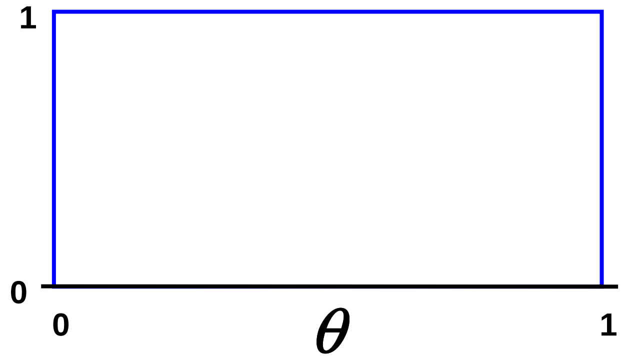
$$s_\theta(y=0)=1-\theta$$

Tail

Noninformative prior density

$$p_0(\theta)\propto 1$$

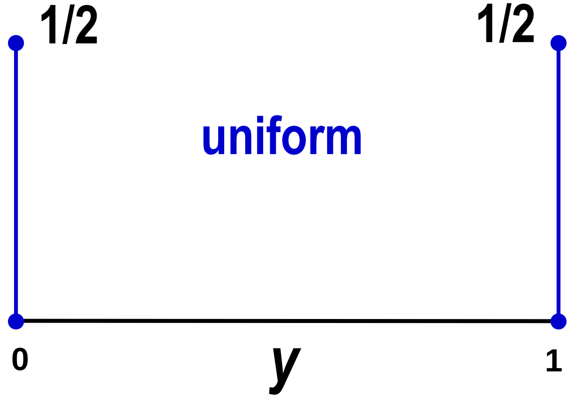
uniform



Coin Tossing

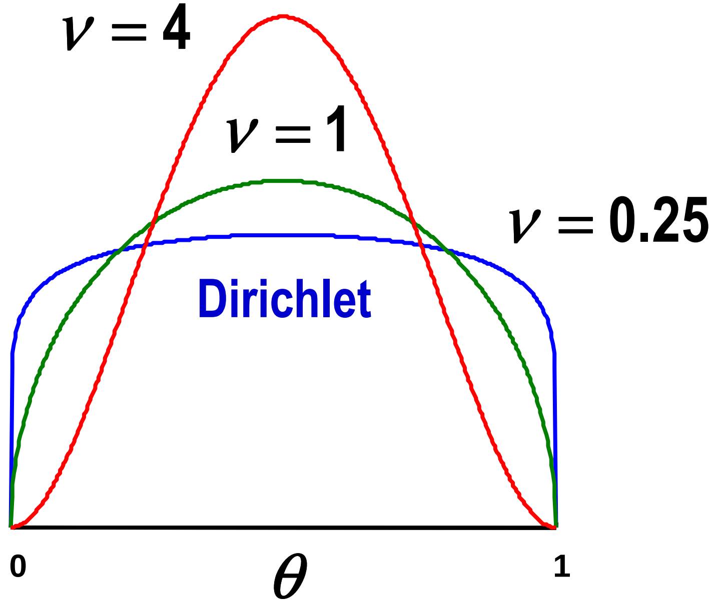
Prior density of data

$$r_\nu(y=0)=0.5$$
$$r_\nu(y=1)=0.5$$



Prior density of pars.

$$p_\nu(\theta) \propto \theta^{\nu/2}(1-\theta)^{\nu/2}$$



ARX Model

- **Model**

$$y_k = \theta' z_k + e_k, \quad e_k \sim N(0,\sigma^2), \quad \sigma^2 \text{ known}$$

- **Model density**

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

- **Log-density**

$$\log s_\theta(y \mid z) = \text{const.} - \frac{1}{2\sigma^2} \begin{bmatrix} -1 \\ \theta \end{bmatrix}' \begin{bmatrix} y \\ z \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix}' \begin{bmatrix} -1 \\ \theta \end{bmatrix}$$

ARX Model

- Conditional inaccuracy

$$K(r_\nu : s_\theta) = \text{const.} + \frac{1}{2\sigma^2}\begin{bmatrix}-1\\ \theta\end{bmatrix}' E_\nu\begin{bmatrix}y\\ z\end{bmatrix}\begin{bmatrix}y\\ z\end{bmatrix}'\begin{bmatrix}-1\\ \theta\end{bmatrix}$$

- Posterior density

$$p_\nu(\theta) \propto \exp\left(-\frac{1}{2\sigma^2}\begin{bmatrix}-1\\ \theta\end{bmatrix}' \nu \underbrace{E_\nu\begin{bmatrix}y\\ z\end{bmatrix}\begin{bmatrix}y\\ z\end{bmatrix}'}\begin{bmatrix}-1\\ \theta\end{bmatrix}\right)$$

$$\mathrm{E}_\nu\left[\begin{array}{c|c} y\,y' & y\,z' \\ \hline z\,y' & z\,z' \end{array}\right]$$

ARX Model

- ***Model Based Viewpoint:***

  Prior information about the model parameters

$$\mathrm{E}(\theta),\ \mathrm{Cov}(\theta)$$

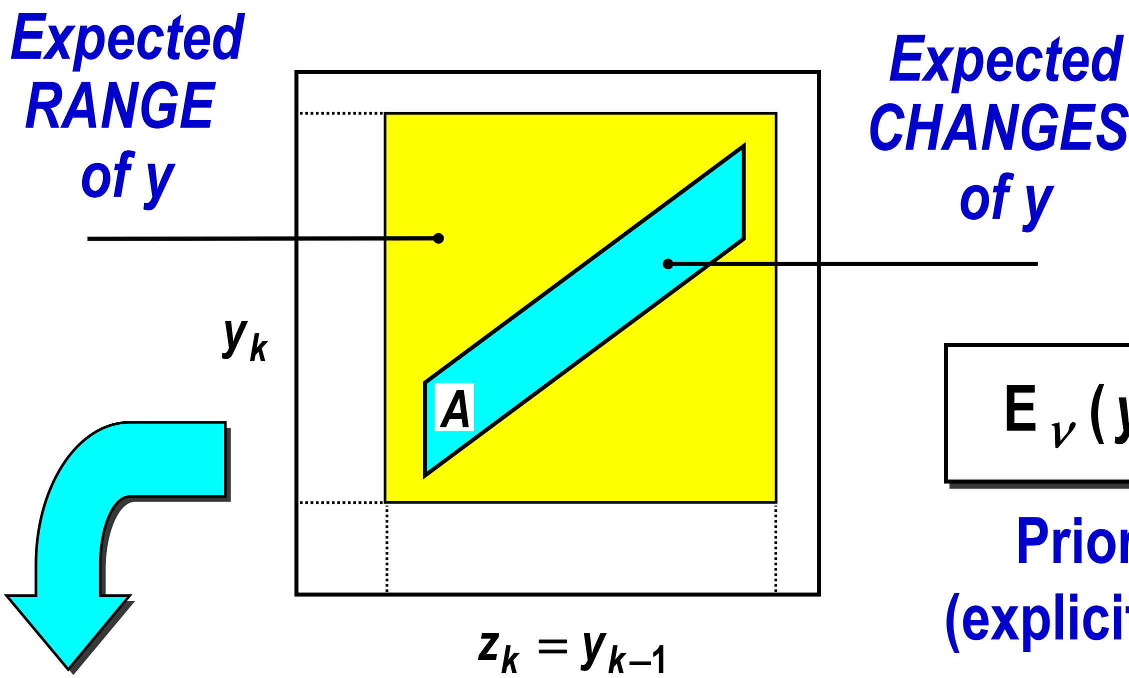
- ***Empirical Viewpoint:***

  Prior information about the measured data

$$\mathrm{E}_{\nu}(yz),\ \mathrm{E}_{\nu}(zz')$$

Which way is the 'right' one?

Prior Knowledge of Data



$$E_{\nu}(yz),\ E_{\nu}(zz')$$

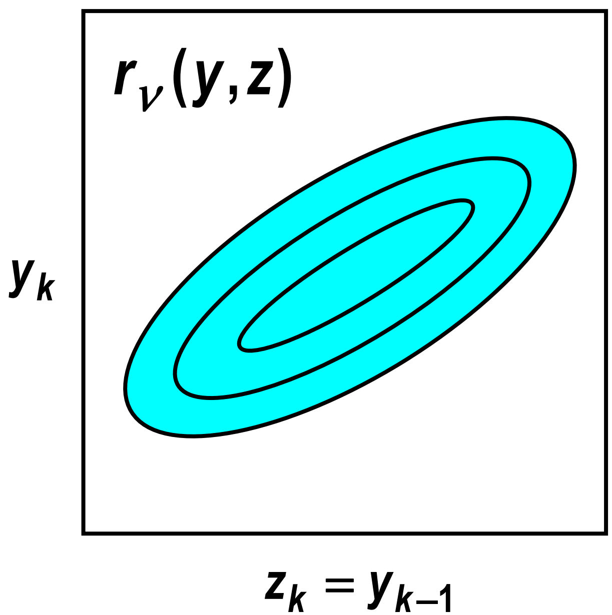
Prior expectations
(explicit or Monte Carlo)

Uniform prior

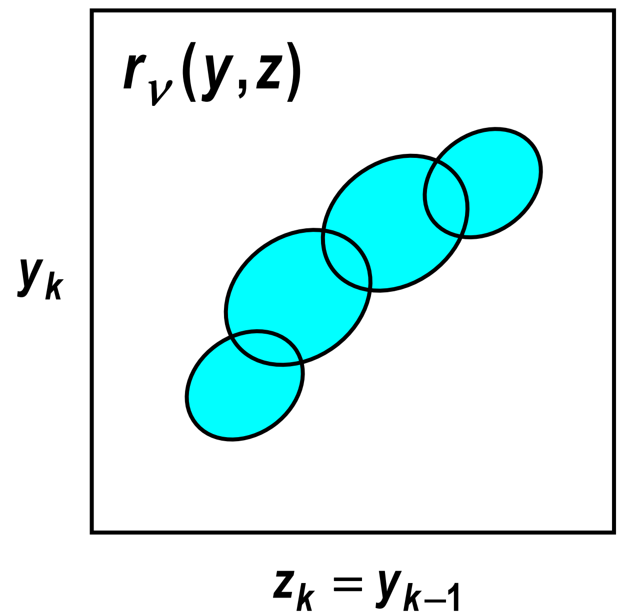
$$r_{\nu}(y,z) = \begin{cases} \text{const.} & \text{if } (y,z) \in A \\ 0 & \text{if } (y,z) \notin A \end{cases}$$

Other Possible Priors

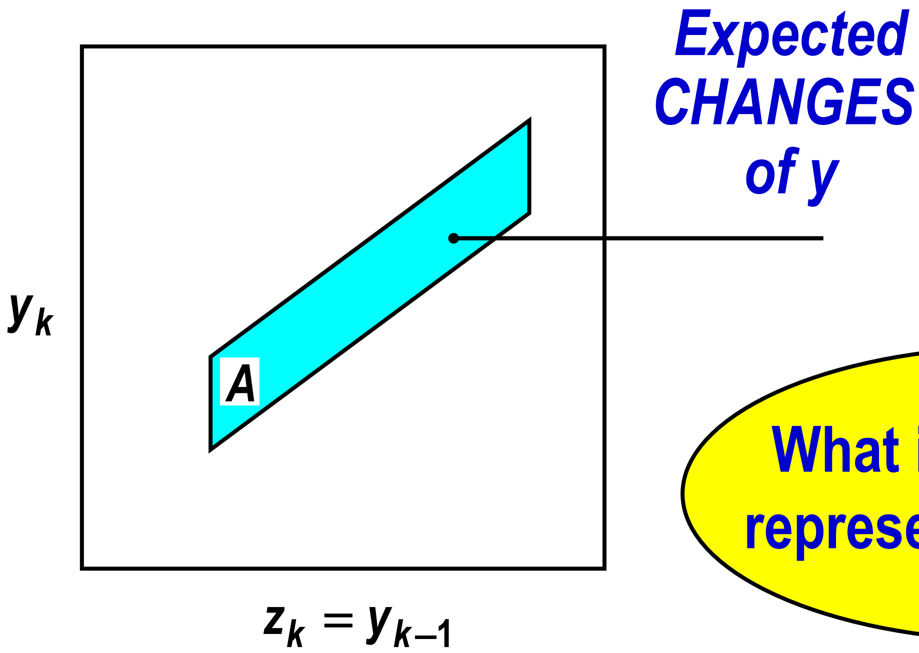
Joint Gaussian density



Mixture of Gaussian densities



Partial Knowledge of Data



What is the optimum representative of R_ν?

Prior belongs to the set of all *r*'s on *A*

$$\boldsymbol{R}_\nu = \{ r \, : \, r(y,z) = 0 \text{ for } (y,z) \notin A \}$$

Minimum "Distance" Solution

SET $\leftrightarrow$ POINT

$$\hat{p}_{\nu}(\theta) \propto \exp(-\nu D(\boldsymbol{R}_{\nu} : s_{\theta}))$$

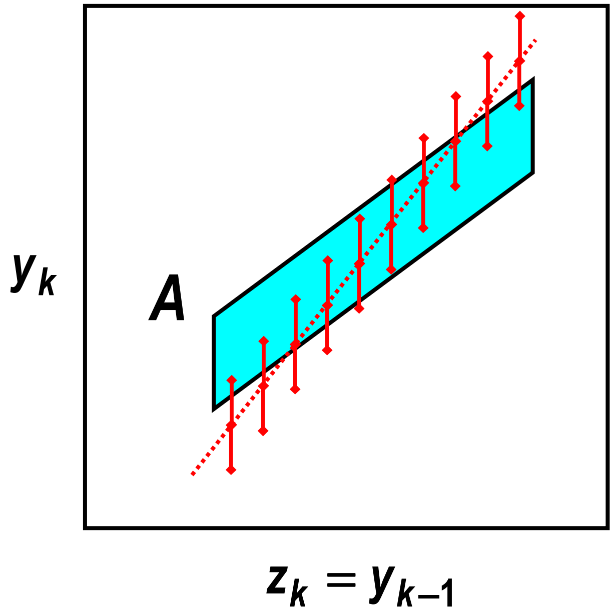
Set membership uncertainty

$$D(\boldsymbol{R}_{\nu} \,\|\, s_{\theta}) = \inf_{r \in \boldsymbol{R}_{\nu}} D(r \,\|\, s_{\theta})$$

$$D(r \,\|\, s_{\theta}) = \iint r(y,z) \log \frac{r(y,z)}{s_{\theta}(y \mid z)} \, \mathrm{d}y \, \mathrm{d}z$$

Relative entropy

Use of Partial Knowledge of Data



Prior belongs to the set of all r's on A

$$\boldsymbol{R}_\nu = \{ r : r(y,z) = 0 \text{ for } (y,z) \notin A \}$$

$$\tilde{s}_\theta(y,z) = \frac{s_\theta(y \mid z)}{\iint_A s_\theta(y \mid z)\, dy\, dz}$$

$$D(\boldsymbol{R}_\nu \,\|\, s_\theta) = D(\tilde{s}_\theta \,\|\, s_\theta)$$

$$= -\log \iint_A s_\theta(y \mid z)\, dy\, dz$$

Special Case: Linear Constraints

- Prior statistic

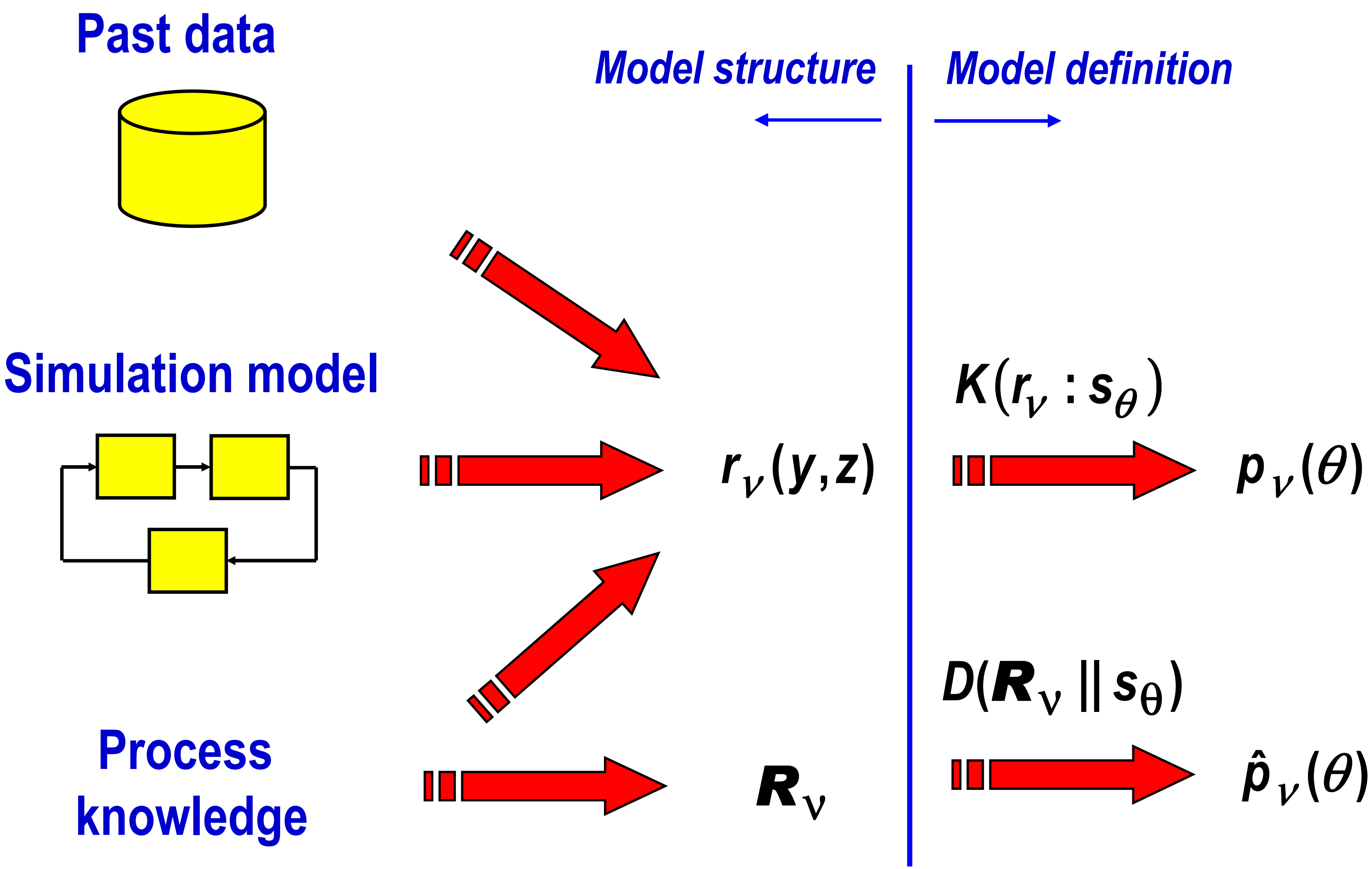
$$\boldsymbol{R}_\nu = \left\{ \iint r(y,z)\, h(y,z)\, \mathrm{d}y\, \mathrm{d}z = \bar{h}_\nu \right\}$$

- Convex optimization problem

$$D(\boldsymbol{R}_\nu \,\|\, s_\theta) = \min_{\lambda \in R^n} \left[\, \psi(\theta,\lambda) - \lambda' \bar{h}_\nu \,\right]$$

$$\psi(\theta,\lambda) = \log \iint s_\theta(y|z)\, \exp(\lambda' h(y,z))\, \mathrm{d}y\, \mathrm{d}z$$

Summary



Past data
Model structure
Model definition
Simulation model
$K(r_\nu : s_\theta)$
$r_\nu(y,z)$
$p_\nu(\theta)$
$D(\boldsymbol{R}_\nu \| s_\theta)$
Process
knowledge
$\boldsymbol{R}_\nu$
$\hat{p}_\nu(\theta)$