Correlated prior and time series

Václav Šmídl

May 4, 2020

Transformation of Gaussian Distribution

Consider normal distribution

$$e \sim \mathcal{N}(0,1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}e^2\right),$$

the distribution of

$$x = f(e) = \mu + \sigma e \qquad \Leftrightarrow \qquad e = g(x) = \frac{x-\mu}{\sigma}$$

is:

$$p(x) = p_e\left(\frac{x-\mu}{\sigma}\right)|g'(x)| = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)\frac{1}{\sigma} = \mathcal{N}(\mu,\sigma)$$

Multivariate Normal

In multivariate case:

$$\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, I), \quad \Leftrightarrow \quad e_i \sim \mathcal{N}(0,1), \;\; i = 1, \ldots, n$$

$$\propto \exp\left(-\frac{1}{2}\boldsymbol{e}^T\boldsymbol{e}\right) \qquad \prod_{i=1}^{n} p(e_i) \propto \prod_{i=1}^{n} \exp(-\frac{1}{2}e_i^2)$$

either with transformation

$$\boldsymbol{x} = \mu + S\boldsymbol{e} \quad \Leftrightarrow \quad \boldsymbol{e} = S^{-1}(\boldsymbol{x} - \mu)$$

$$\boldsymbol{e}^T\boldsymbol{e} = (\boldsymbol{x} - \mu)^T S^{-T} S^{-1}(\boldsymbol{x} - \mu)$$

$$\boldsymbol{x} \sim \frac{1}{\sqrt{2\pi}^n} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \mu)^T S^{-T} S^{-1}(\boldsymbol{x} - \mu)\right) |S^{-1}|$$

with $\Sigma = SIS^T$ and $\det(\Sigma) = |S|^2$

Least squares – different way I

Consider regression problem:

$$\begin{aligned} \boldsymbol{y} &= X\boldsymbol{\theta} + \sigma\boldsymbol{e} & \boldsymbol{e} &\sim \mathcal{N}(\mathbf{0}, I), \\ p(\boldsymbol{y}|X, \boldsymbol{x}) &= \mathcal{N}(X\boldsymbol{\theta}, \sigma^2 I) & & \\ \boldsymbol{\theta} &= \alpha\boldsymbol{e}_{\boldsymbol{\theta}} & \boldsymbol{e}_{\boldsymbol{\theta}} &\sim \mathcal{N}(\mathbf{0}, I) \\ p(\boldsymbol{\theta}) &= \mathcal{N}(\mathbf{0}, I), & & \end{aligned}$$

Build joint distribution

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{\theta} \end{bmatrix} = \begin{bmatrix} \sigma I & X \\ 0 & I \end{bmatrix} \begin{bmatrix} \boldsymbol{e} \\ \boldsymbol{e}_{\boldsymbol{\theta}} \end{bmatrix}$$

$$p(\boldsymbol{y}, \boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma I & X \\ 0 & I \end{bmatrix} \begin{bmatrix} \sigma I & 0 \\ X^T & I \end{bmatrix}\right) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma^2 I + XX^T & X^T \\ X & I \end{bmatrix}\right)$$

Least squares – different way II

Joint

$$p(\boldsymbol{y}, \boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma^2 I + XX^T & X \\ X^T & I \end{bmatrix}\right)$$

with decompositon

$$\begin{aligned} p(\boldsymbol{y}) &= \mathcal{N}(\mathbf{0}, \sigma^2 I + XX^T), \\ p(\boldsymbol{\theta}|\boldsymbol{y}) &= \mathcal{N}(\hat{\boldsymbol{\theta}}, \Sigma_{\boldsymbol{\theta}}) \\ \hat{\boldsymbol{\theta}} &= X^T(\sigma^2 I + XX^T)^{-1}\boldsymbol{y}, \\ \Sigma_{\boldsymbol{\theta}} &= I - X^T(\sigma^2 I + XX^T)^{-1}X \end{aligned}$$

decomposition into:

1. generative model,
2. different form of least squares. (inverse in dimension of $\boldsymbol{y}$!)

Generative models of 2nd order polynomials

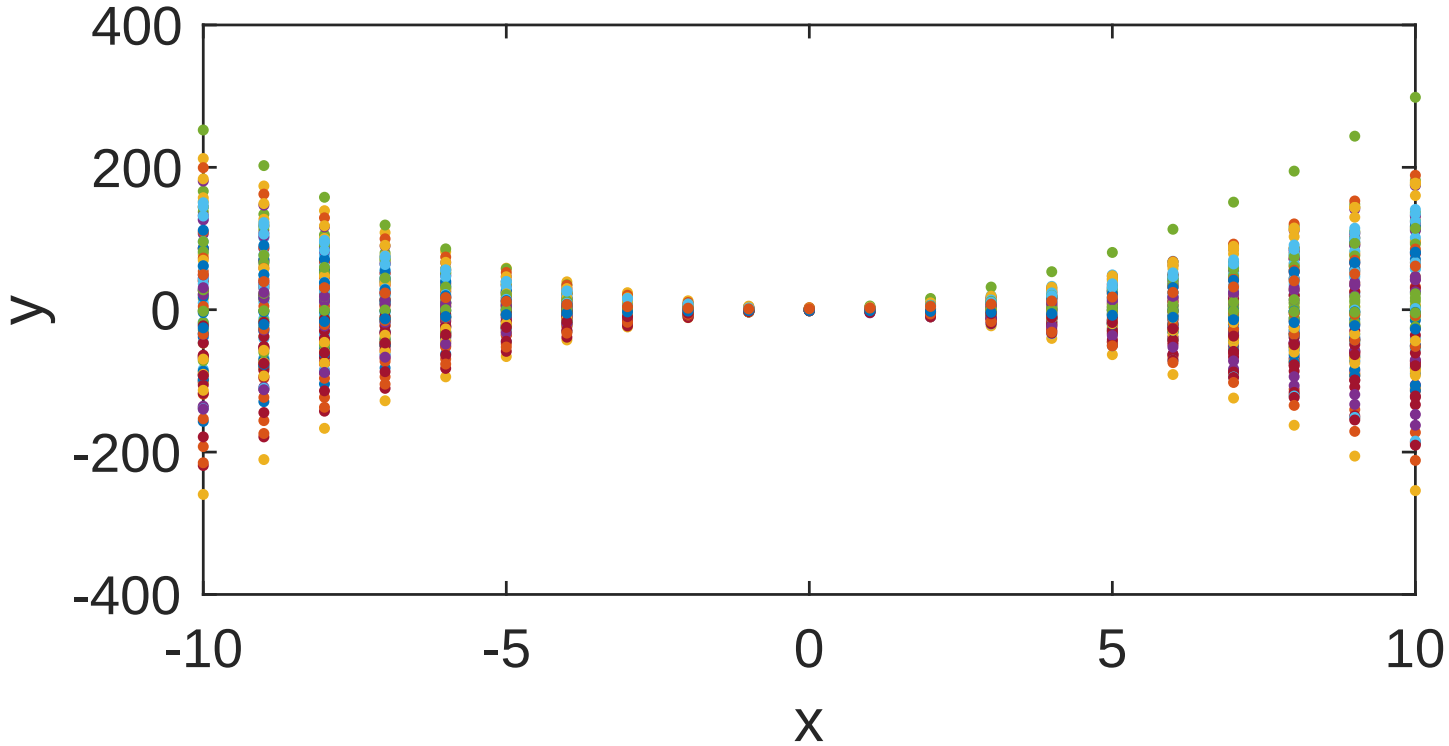
we have a least squares model:

$$p(\boldsymbol{y}|M,\boldsymbol{\theta}) = \mathcal{N}(M\boldsymbol{\theta}, \sigma^2 I), \qquad p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, I),$$

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \qquad p(\boldsymbol{y}) = \mathcal{N}(\mathbf{0}, \sigma^2 I + MM^T),$$

Consider samples:

- $\boldsymbol{\theta}^{(i)}$ from $p(\boldsymbol{\theta})$,
- $\boldsymbol{y}^{(i)}$ from $p(\boldsymbol{y}|M,\boldsymbol{\theta}^{(i)})$ to generate samples from the joint $(\boldsymbol{y}^{(i)}, \boldsymbol{x}^{(i)})$



Least squares – prediction

Provide predictions $\overline{\boldsymbol{y}}$ at given $\overline{\boldsymbol{x}}$ and known $(\boldsymbol{x},\boldsymbol{y})$

$$p(\overline{\boldsymbol{y}}|\overline{\boldsymbol{x}},\boldsymbol{x},\boldsymbol{y}) = \int p(\overline{\boldsymbol{y}}|\overline{\boldsymbol{x}},\theta)p(\theta|\boldsymbol{x},\boldsymbol{y})d\boldsymbol{\theta}$$

Alternative:

$$\begin{bmatrix}\boldsymbol{y}\\ \overline{\boldsymbol{y}}\\ \boldsymbol{\theta}\end{bmatrix} = \begin{bmatrix}\sigma I & 0 & X\\ 0 & \sigma I & \overline{X}\\ 0 & 0 & \alpha I\end{bmatrix}\begin{bmatrix}\boldsymbol{e}\\ \overline{\boldsymbol{e}}\\ \boldsymbol{e}_{\boldsymbol{\theta}}\end{bmatrix}$$

$$p(\boldsymbol{y},\overline{\boldsymbol{y}},\boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{0},\begin{bmatrix}\sigma^2 I + XX^T & X\overline{X}^T & X\\ \overline{X}X^T & \sigma^2 I + XX^T & \overline{X}\\ X^T & \overline{X}^T & I\end{bmatrix}\right)$$

$$p(\boldsymbol{y},\overline{\boldsymbol{y}}) = \mathcal{N}\left(\mathbf{0},\begin{bmatrix}\sigma^2 I + XX^T & X\overline{X}^T\\ \overline{X}X^T & \sigma^2 I + \overline{XX}^T\end{bmatrix}\right)$$

We do not need parameters to predict! (we need all data!)
All we need is kernel $K_{xx} = XX^T, K_{x\overline{x}}, K_{\overline{xx}}$

Gaussian process (non-parametric model)

Definition

[Rasmussen, 2006] A Gaussinan process is a collection of random variables, any finite number of which have a joint Gaussian distribution
Value of a process at location $\boldsymbol{x}$ is

$$\begin{aligned} f(\boldsymbol{x}) &\sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')) \\ m(\boldsymbol{x}) &= \mathrm{E}\left(f(\boldsymbol{x})\right), \\ k(\boldsymbol{x}, \boldsymbol{x}') &= \mathrm{E}\left((f(\boldsymbol{x}) - m(\boldsymbol{x}))(f(\boldsymbol{x}') - m(\boldsymbol{x}'))\right), \end{aligned}$$

where $m(\boldsymbol{x})$ si the mean and $k(\boldsymbol{x}, \boldsymbol{x}')$ is a correlation function. $\boldsymbol{x}'$ is any other point in the collection.

- in polynomial regression $k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T \boldsymbol{x}$
- "better" kernels:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{1}{2l}|\boldsymbol{x} - \boldsymbol{x}'|^2\right)$$
$$k(\boldsymbol{x}, \boldsymbol{x}') = (1 + |\boldsymbol{x} - \boldsymbol{x}'|^2)^{-\alpha}$$

with hyper-parameters l, α ...

Example

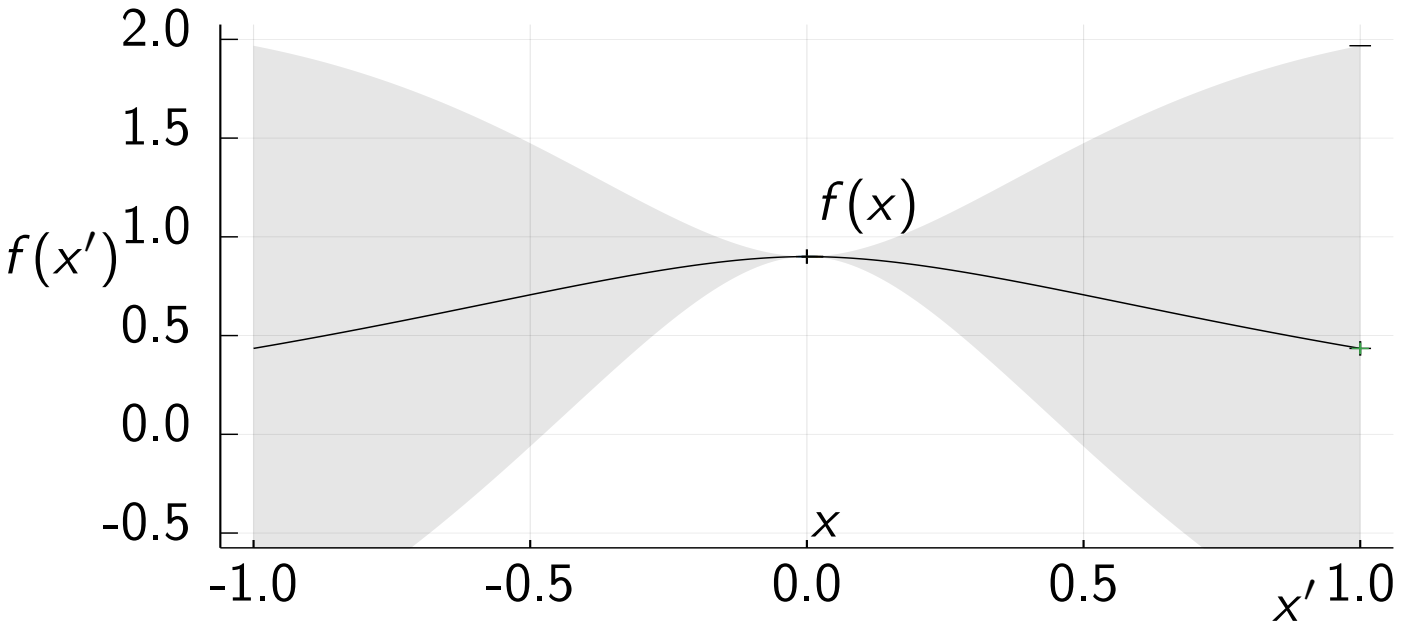
Joint distribution of two values $y = f(x) + e$

$$
\begin{aligned}
p(f(x), \overline{f(x)}) &= \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} k(x,x) & k(x,\overline{x}) \\ k(\overline{x},x) & k(\overline{x},\overline{x}) \end{bmatrix}\right) \\
&= \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & k(x,\overline{x}) \\ k(\overline{x},x) & 1 \end{bmatrix}\right) \\
p(y, \overline{y}) &= \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma+1 & k(x,\overline{x}) \\ k(\overline{x},x) & 1+\sigma \end{bmatrix}\right) \\
p(\overline{y}|y) &= \mathcal{N}\left(k(x,x')f(x), (1+\sigma - k(x,x')^2/(1+\sigma))\right), \\
p(\overline{y}) &= \mathcal{N}\left(0, 1+\sigma\right),
\end{aligned}
$$

1D Example

Joint distribution of two values $y = f(x) + e$,

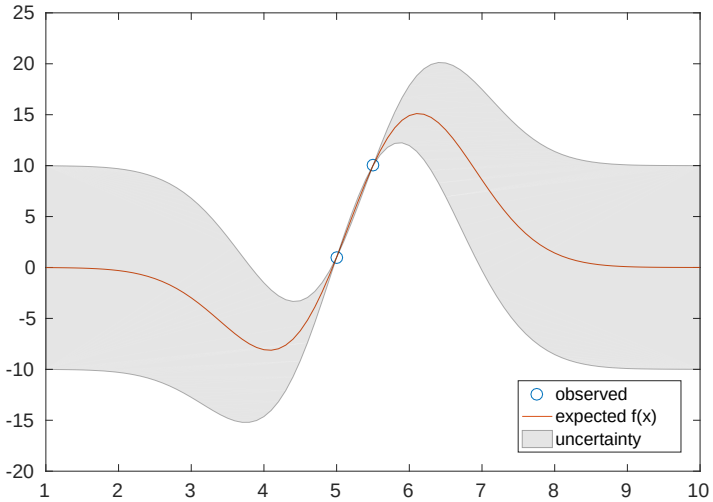
$$p(f(\overline{x})|y) = \mathcal{N}\left(k(x,\overline{x})y, (1+\sigma-k(x,\overline{x})^2)\right),$$
$$p(\overline{y}|y) = \mathcal{N}\left(k(x,\overline{x})y, (1+\sigma-k(x,\overline{x})^2/(1+\sigma))\right),$$



1D Example – adding points

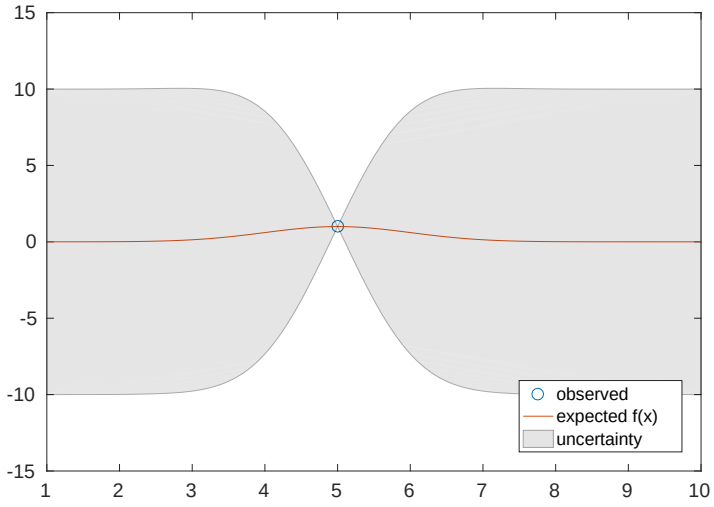
Joint distribution of two values $y = f(x) + e$

$$p(\overline{y}|y) = \mathcal{N}\left(K(X,\overline{X})y, K(\overline{X},\overline{X}) + \sigma I - K(\overline{X},X)\left(\sigma I + K(X,X)\right)^{-1} K(X,\overline{X})\right),$$

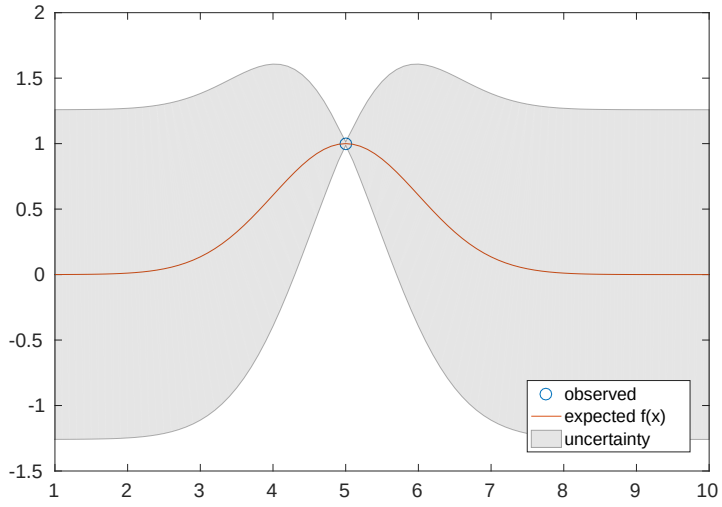


Sensitivity to kernel parameters

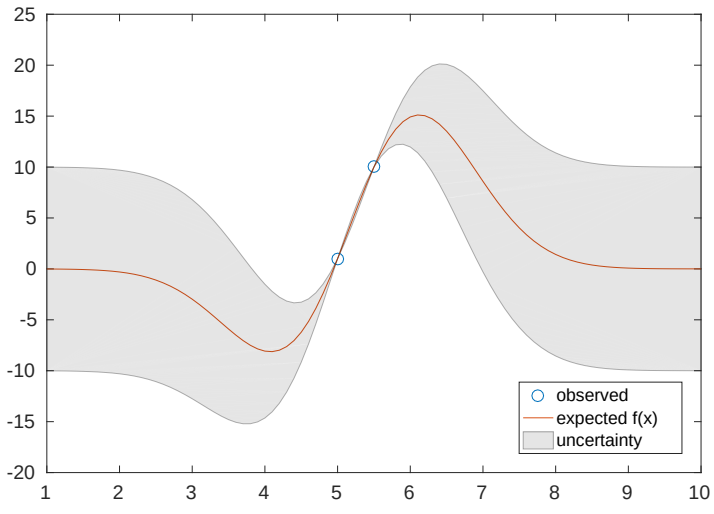
$\ell = 1, \sigma = 0.1$



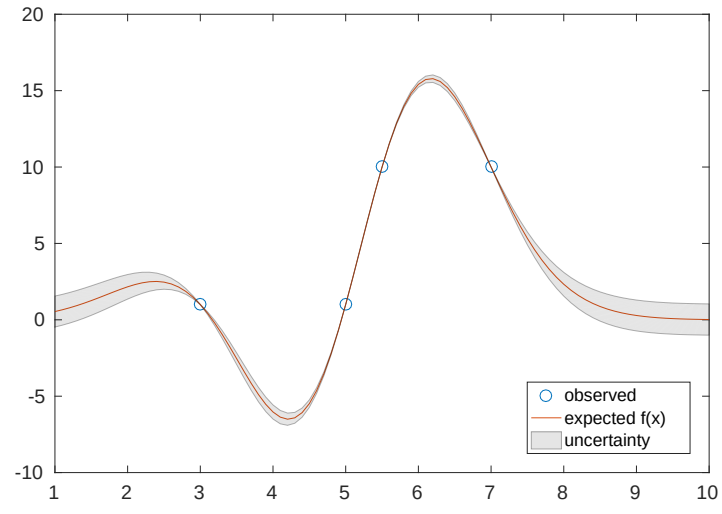
$\ell = 0.1, \sigma = 0.1$



$\ell = 1, \sigma = 0.1$

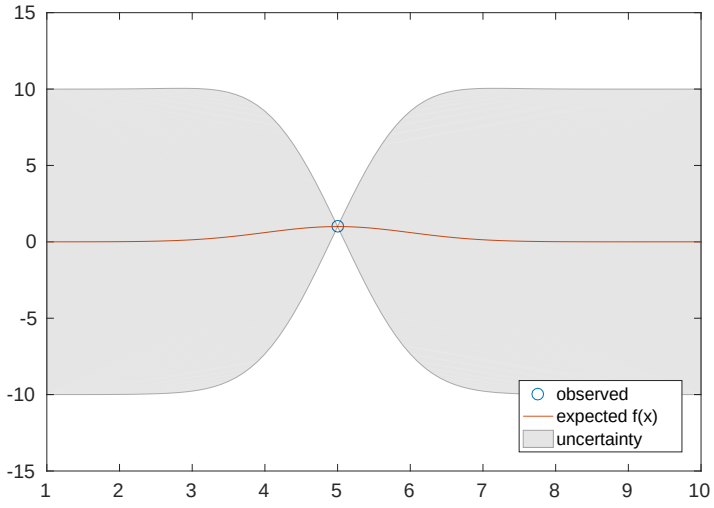


$l = 1, \sigma = 0.1$

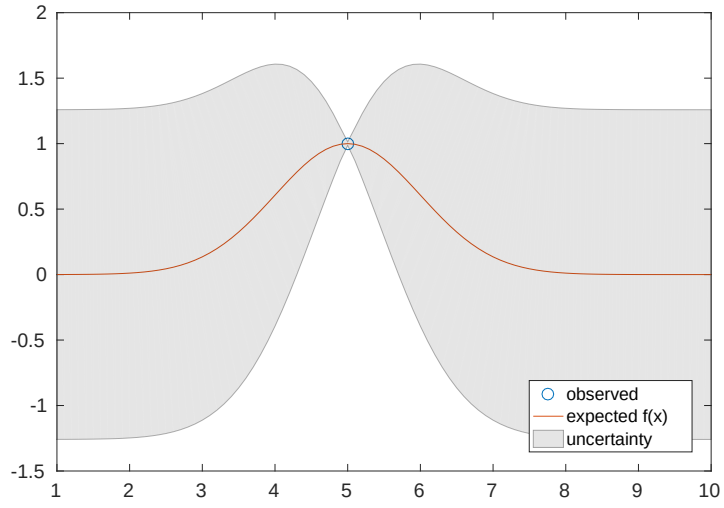


Sensitivity to kernel parameters (function smoothness)

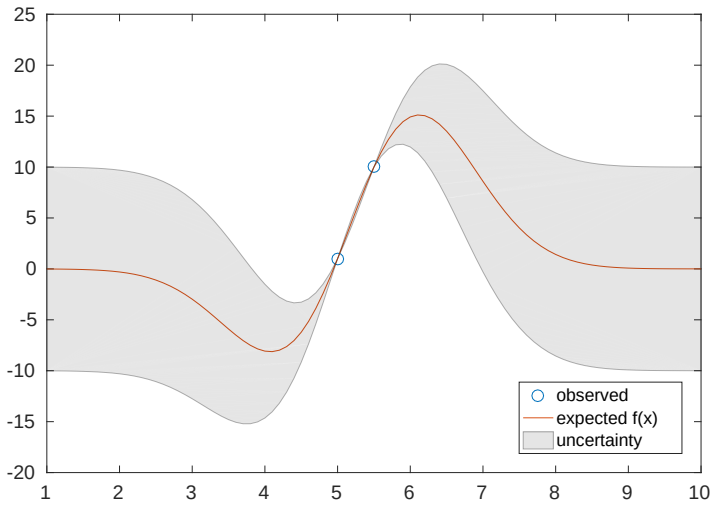
$\ell = 1, \sigma = 0.1$



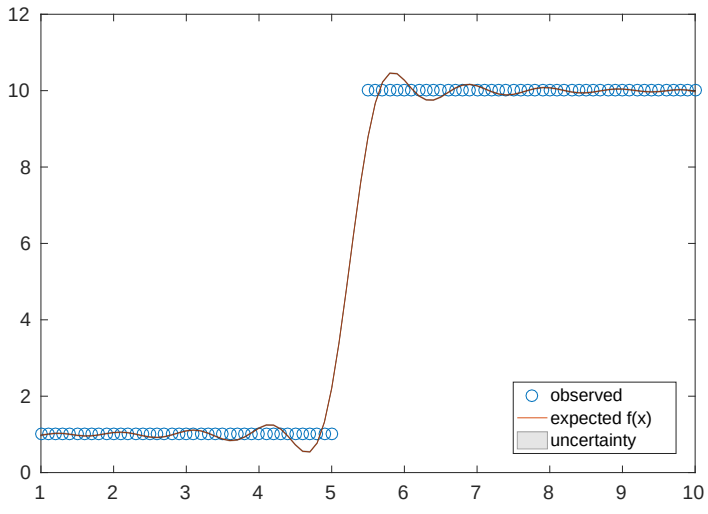
$\ell = 0.1, \sigma = 0.1$



$\ell = 1, \sigma = 0.1$



$\ell = 1, \sigma = 0.1$



Learning kernel parameters

Technically we have:

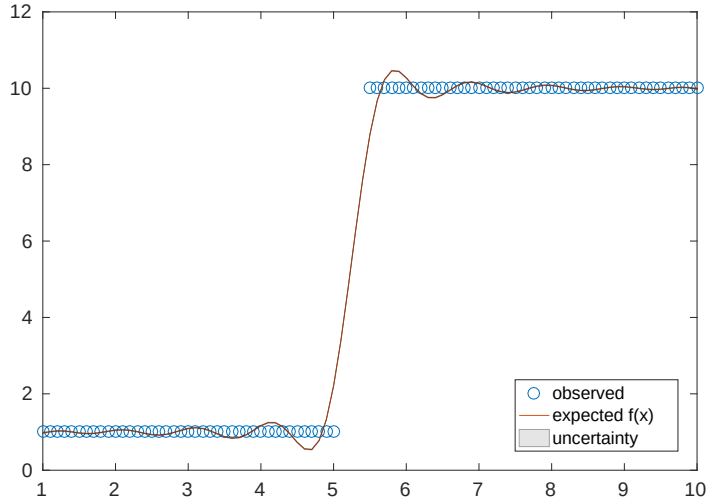
$$p(\boldsymbol{y}|\theta) = \mathcal{N}\left(\mathbf{0}, \sigma^2 I + K_\theta(X, X^T)\right)$$

we can learn

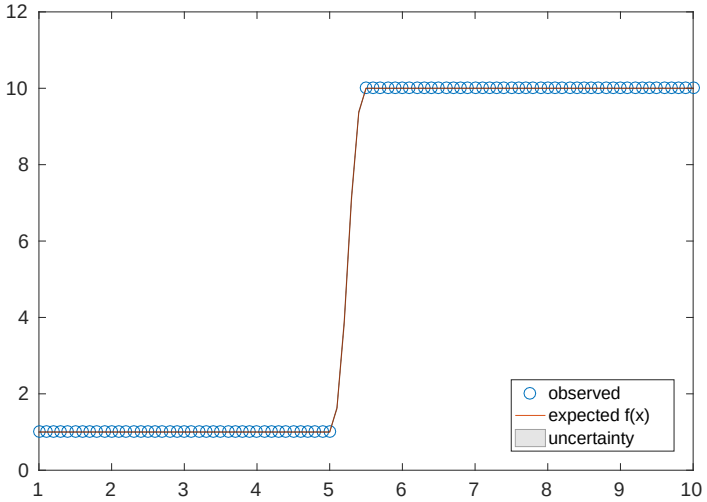
$$\theta = \arg\max_\theta p(\boldsymbol{y}|\theta)$$

which implies minimization of $|\sigma^2 I + K_\theta(X, X^T)|$. Or VB, EM, MCMC...

$\ell = 1, \sigma = 0.1$



$\hat{\ell} = 0.55, \hat{\sigma} = 1e-6$



Non-parametric models

- Data interpolators
- Nice analytical properties (Gaussian)
- Algebra on kernels: $K(X, \overline{X}) = K_1(X, \overline{X}) + K_2(X, \overline{X})$
- Extensions
  - Student-t processes, etc.
  - Non-stationary processes
  - Hierarchical processes
- Book:
  - Carl Edward Rasmussen and Christopher K.I. Williams. Gaussian processes for machine learning. Vol. 2. No. 3. Cambridge, MA: MIT press, 2006.

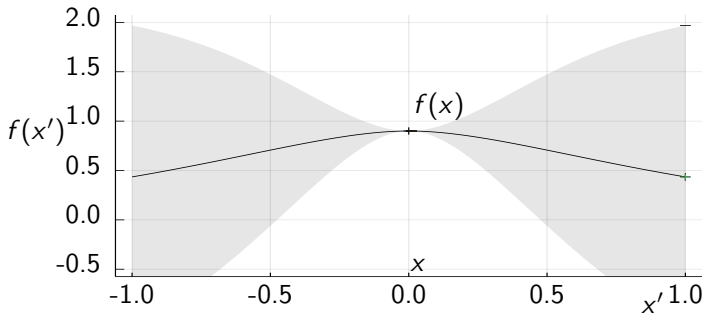
Assignment

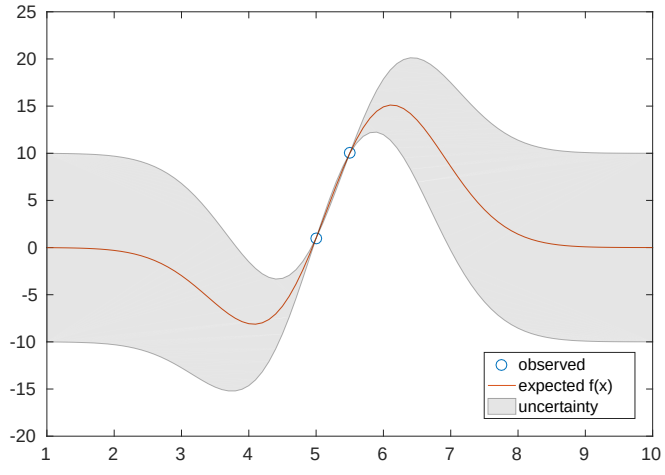
Derive a Gaussian process approximation for

- 1 point of observation
- 2 points of observation

Choose freely

- kernel,
- prediction range...
- y





Back to State-space model

Consider state space

$$x_n = ax_{n-1} + \sqrt{q}v_n,$$
$$y_n = cx_n + \sqrt{r}w_n$$

as a regression problem: $\boldsymbol{y} = [y_1, y_2 \dots y_n]$ and $\boldsymbol{x} = [x_1, x_2 \dots, x_n]$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} c & & \\ & c & \\ & & c \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix} + \sqrt{r} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \end{bmatrix} \qquad p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\mathbb{C}\boldsymbol{x}, rI)$$

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} 0 & a & 0 & \dots \\ 0 & 0 & a & \dots \\ & & & \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix} + \sqrt{q} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \end{bmatrix} \qquad p(\boldsymbol{x}) = ?$$

Correlated model

Prior on $\boldsymbol{x}$:

$$\begin{aligned} p(x_1) = \mathcal{N}(\hat{x}_1, q_1) \quad &\Leftrightarrow \quad x_1 = \hat{x}_1 + \sqrt{p_1}e_1, \qquad e_1 \sim \mathcal{N}(0,1) \\ p(x_i|x_{i-1}) = \mathcal{N}(ax_{i-1}, q) \quad &\qquad x_i = ax_{i-1} + \sqrt{q}e_i, \qquad i = 2, \ldots, n \end{aligned}$$

$$\begin{aligned} p_1 e_1 &= x_1 - \hat{x}_1 \\ q e_2 &= x_2 - a x_1 \\ q e_3 &= x_3 - a x_2 \\ &\vdots \end{aligned} \Leftrightarrow \begin{bmatrix} \sqrt{p_1} & & & \\ & \sqrt{q} & & \\ & & \sqrt{q} & \\ & & & \ddots \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & & & \\ -a & 1 & & \\ & -a & 1 & \\ & & & \ddots \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \end{bmatrix} -$$

Implies

$$\Sigma_0 \boldsymbol{e} = \mathbb{A}\boldsymbol{x} - \mu_0 \quad \Rightarrow \quad \mathbb{A}^{-1}(\Sigma_0 \boldsymbol{e} + \mu_0) = \boldsymbol{x} \quad \Rightarrow \quad p(\boldsymbol{x}) = \mathcal{N}(\mathbb{A}^{-1}\hat{\mu}_0, \mathbb{A}^{-1}\Sigma_0\mathbb{A}^{-}$$

Higher order correlation

Consider vector

$$\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$$

with prior probabilities

$$\begin{aligned} x_1 &= \hat{x}_1 + \sigma_1 e_1, & e_1 &\sim \mathcal{N}(0,1) \\ x_2 &= a x_1 + \hat{x}_2 + \sigma_2 e_1, & e_1 &\sim \mathcal{N}(0,1) \\ x_i &= a x_{i-1} + b x_{i-2} + \sigma e_i, & i &= 2, \ldots, n \end{aligned}$$

$$\begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \sigma & \\ & & & \ddots \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & & & \\ -a & 1 & & \\ -b & -a & 1 & \\ & & & \ddots \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \end{bmatrix} - \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ 0 \\ 0 \end{bmatrix}$$

Implies

$$\Sigma_0 \boldsymbol{e} = A\boldsymbol{x} - \mu_0 \quad \Rightarrow \quad A^{-1}(\Sigma_0 \boldsymbol{e} + \mu_0) = \boldsymbol{x} \quad \Rightarrow \quad \boldsymbol{x} = \mathcal{N}(A^{-1}\hat{\mu}_0, A^{-1}\Sigma_0 A^{-T})$$

Special case #1

State Space model for $a = 0, c = 1$

$$x_n = \sqrt{q} v_n,$$
$$y_n = x_n + \sqrt{r} w_n$$

We have

$$p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}, r\boldsymbol{I}), \quad p(\boldsymbol{x}) = \mathcal{N}(q\boldsymbol{I}) \quad \Leftrightarrow \quad p(y_i|x_i) = \mathcal{N}(x_i, r), \quad p(x_i) = \mathcal{N}(q)$$

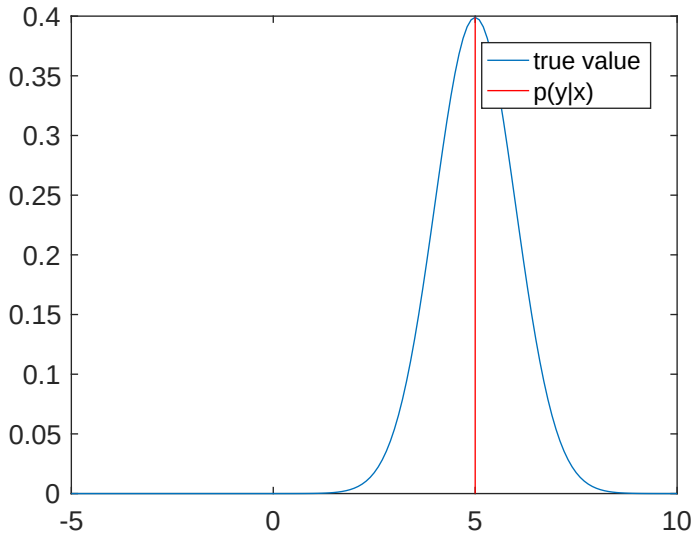
Lecture #2. (Normal observation y of constant parameter m)

Bayes rule for Normal

Likelihood of unknown x:

$$p(y|x) = \mathcal{N}(x, r)$$
$$\propto \exp\left(-\frac{1}{2}\left(\frac{x-y}{r}\right)^2\right)$$

With observation y.



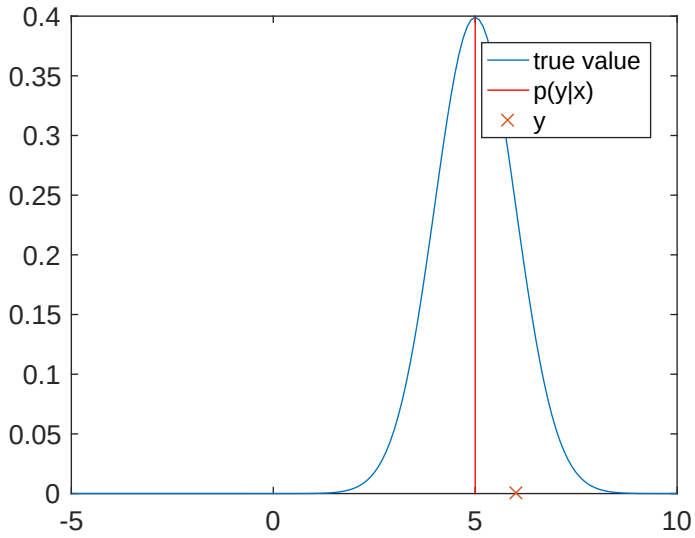
$$p(x|y,\phi) = \mathcal{N}\left(\left(\frac{r+q}{qr}\right)^{-1}\frac{0r+yq}{qr}, \left(\frac{r+q}{qr}\right)^{-1}\right) = \mathcal{N}\left(\left(\frac{qy}{r+q}\right), \frac{qr}{q+r}\right)$$

Bayes rule for Normal

Likelihood of unknown x:

$$p(y|x) = \mathcal{N}(x, r)$$
$$\propto \exp\left(-\frac{1}{2}\left(\frac{x-y}{r}\right)^2\right)$$

With observation y.



$$p(x|y,\phi) = \mathcal{N}\left(\left(\frac{r+q}{qr}\right)^{-1}\frac{0r+yq}{qr}, \left(\frac{r+q}{qr}\right)^{-1}\right) = \mathcal{N}\left(\left(\frac{qy}{r+q}\right), \frac{qr}{q+r}\right)$$

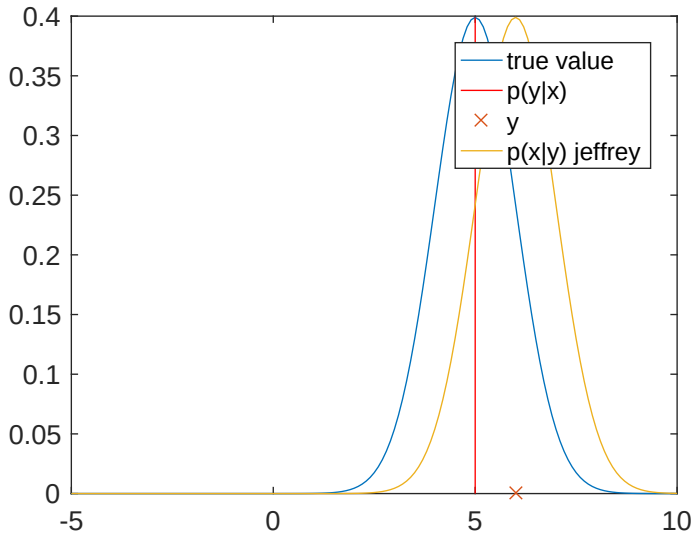
Bayes rule for Normal

Likelihood of unknown x:

$$p(y|x) = \mathcal{N}(x, r)$$
$$\propto \exp\left(-\frac{1}{2}\left(\frac{x-y}{r}\right)^2\right)$$

With observation y.
Posterior with Jeffrey's prior

$$p(x|y) \propto p(x|y)$$
$$= \mathcal{N}(y, r)$$

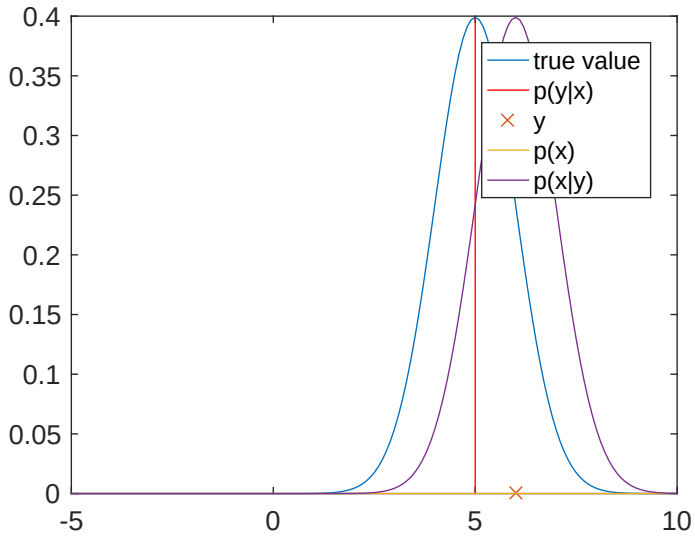


$$p(x|y,\phi) = \mathcal{N}\left(\left(\frac{r+q}{qr}\right)^{-1}\frac{0r+yq}{qr}, \left(\frac{r+q}{qr}\right)^{-1}\right) = \mathcal{N}\left(\left(\frac{qy}{r+q}\right), \frac{qr}{q+r}\right)$$

Bayes rule for Normal

Likelihood of unknown x:

$$p(y|x) = \mathcal{N}(x, r)$$
$$\propto \exp\left(-\frac{1}{2}\left(\frac{x-y}{r}\right)^2\right)$$



With observation y.
Posterior with Jeffrey's prior

$$p(x|y) \propto p(x|y)$$
$$= \mathcal{N}(y, r)$$

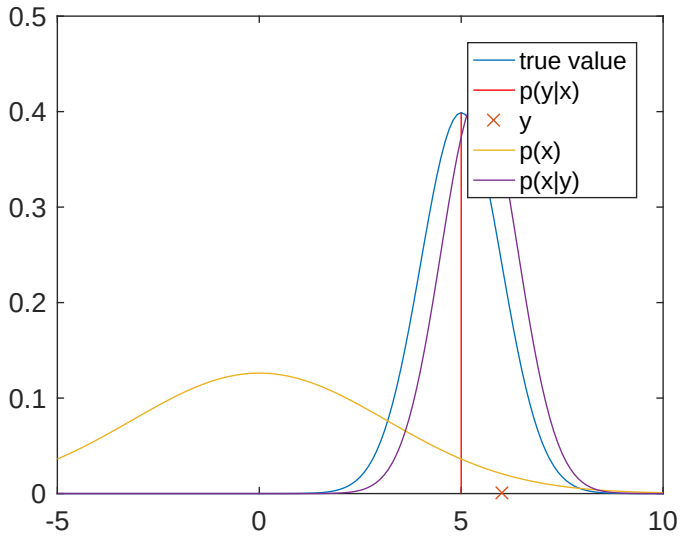
with normal prior $p(x) = \mathcal{N}(0, q)$

$$p(x|y,\phi) = \mathcal{N}\left(\left(\frac{r+q}{qr}\right)^{-1}\frac{0r+yq}{qr}, \left(\frac{r+q}{qr}\right)^{-1}\right) = \mathcal{N}\left(\left(\frac{qy}{r+q}\right), \frac{qr}{q+r}\right)$$

Bayes rule for Normal

Likelihood of unknown x:

$$p(y|x) = \mathcal{N}(x, r)$$
$$\propto \exp\left(-\frac{1}{2}\left(\frac{x-y}{r}\right)^2\right)$$



With observation y.
Posterior with Jeffrey's prior

$$p(x|y) \propto p(x|y)$$
$$= \mathcal{N}(y, r)$$

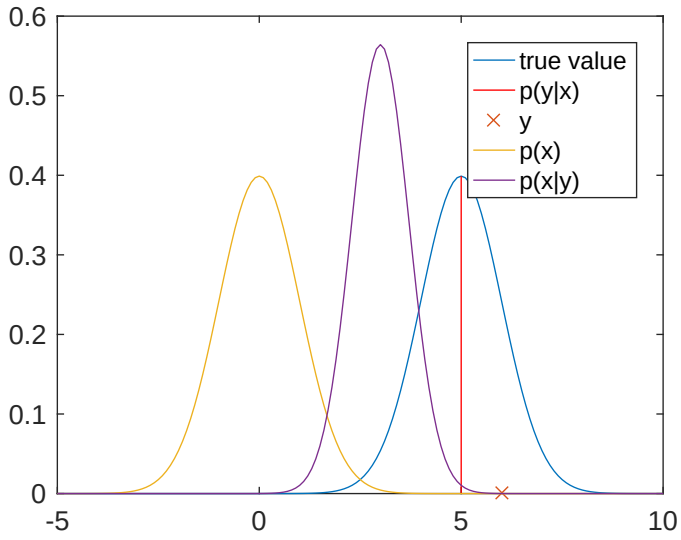
with normal prior $p(x) = \mathcal{N}(0, q)$

$$p(x|y,\phi) = \mathcal{N}\left(\left(\frac{r+q}{qr}\right)^{-1}\frac{0r+yq}{qr}, \left(\frac{r+q}{qr}\right)^{-1}\right) = \mathcal{N}\left(\left(\frac{qy}{r+q}\right), \frac{qr}{q+r}\right)$$

Bayes rule for Normal

Likelihood of unknown x:

$$p(y|x) = \mathcal{N}(x, r)$$
$$\propto \exp\left(-\frac{1}{2}\left(\frac{x-y}{r}\right)^2\right)$$



With observation y.
Posterior with Jeffrey's prior

$$p(x|y) \propto p(x|y)$$
$$= \mathcal{N}(y, r)$$

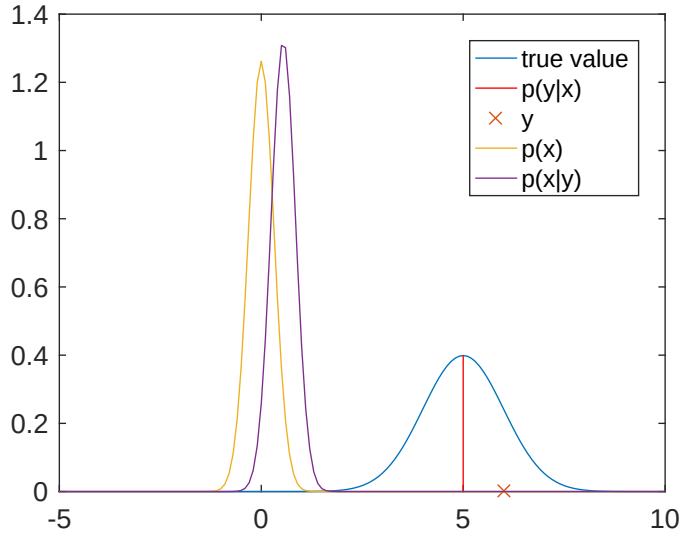
with normal prior $p(x) = \mathcal{N}(0, q)$

$$p(x|y,\phi) = \mathcal{N}\left(\left(\frac{r+q}{qr}\right)^{-1}\frac{0r+yq}{qr}, \left(\frac{r+q}{qr}\right)^{-1}\right) = \mathcal{N}\left(\left(\frac{qy}{r+q}\right), \frac{qr}{q+r}\right)$$

Bayes rule for Normal

Likelihood of unknown x:

$$p(y|x) = \mathcal{N}(x, r)$$
$$\propto \exp\left(-\frac{1}{2}\left(\frac{x-y}{r}\right)^2\right)$$



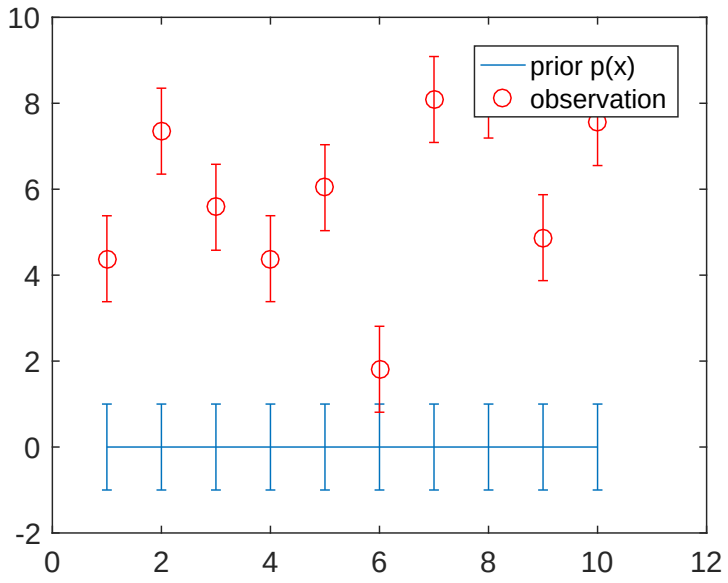
With observation y.
Posterior with Jeffrey's prior

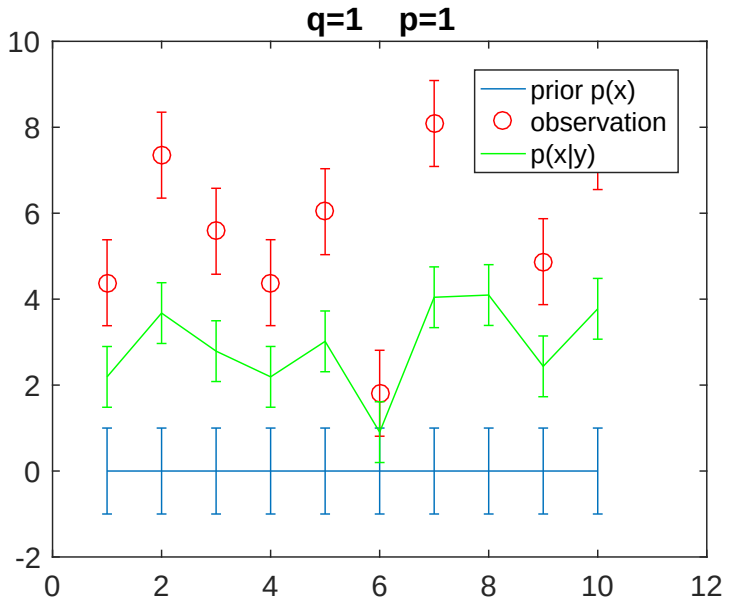
$$p(x|y) \propto p(x|y)$$
$$= \mathcal{N}(y, r)$$

with normal prior $p(x) = \mathcal{N}(0, q)$

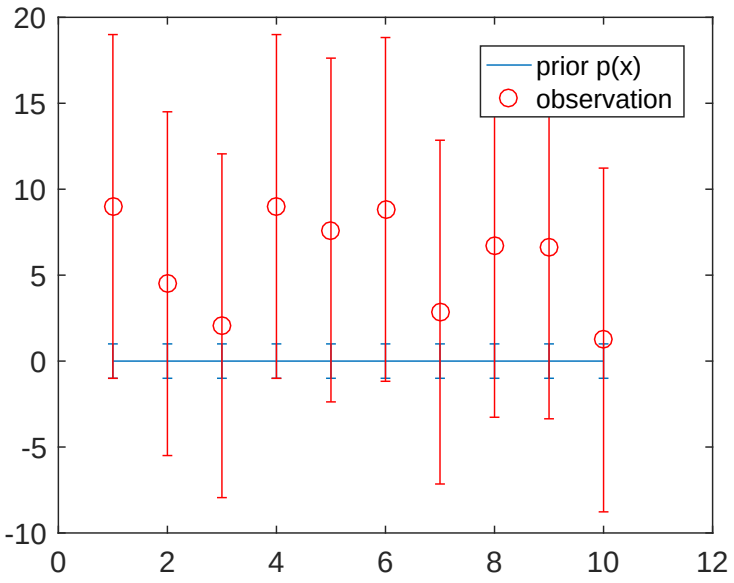
$$p(x|y,\phi) = \mathcal{N}\left(\left(\frac{r+q}{qr}\right)^{-1}\frac{0r+yq}{qr}, \left(\frac{r+q}{qr}\right)^{-1}\right) = \mathcal{N}\left(\left(\frac{qy}{r+q}\right), \frac{qr}{q+r}\right)$$

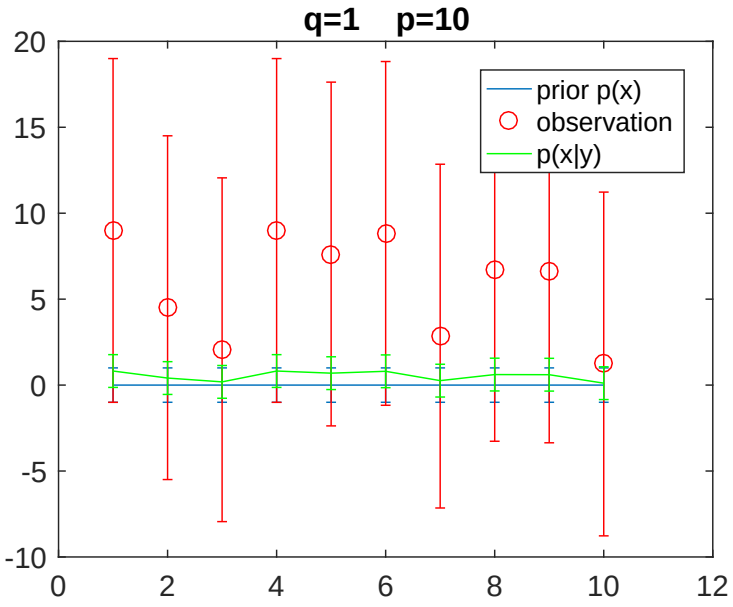
Independent Kalman



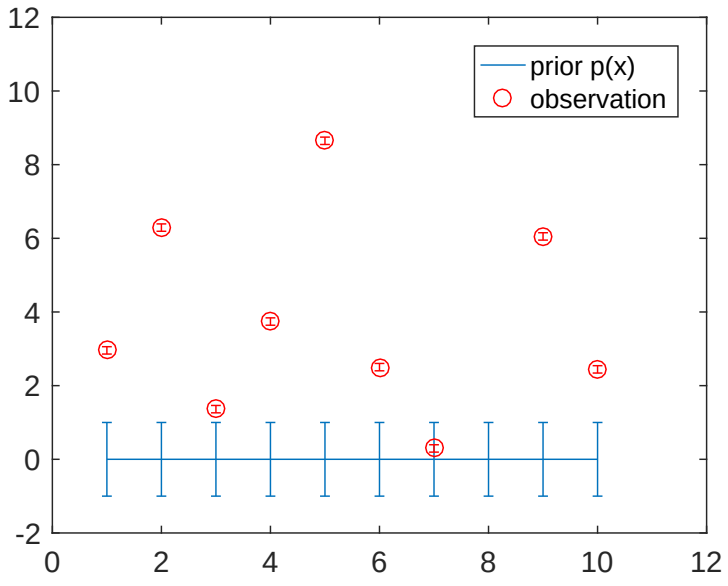


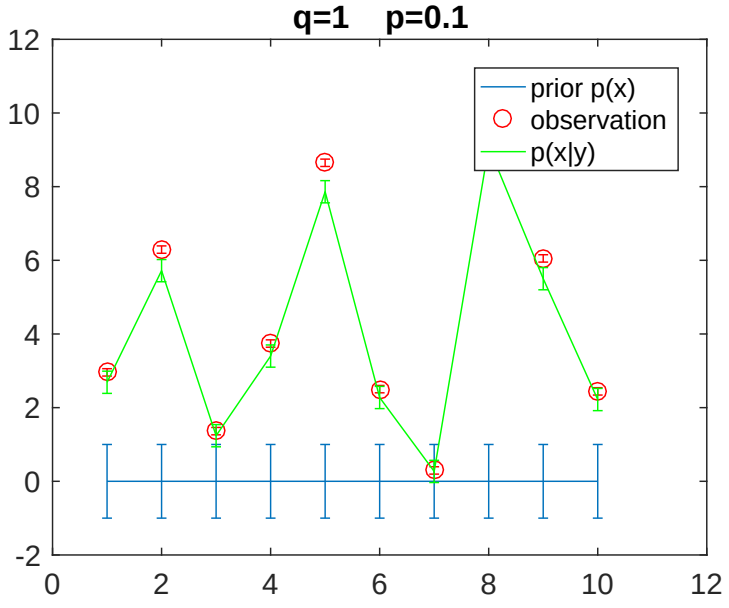
Independent Kalman





Independent Kalman





Special case #2

State Space model for $a = 1, q = 0; c = 1$

$$x_n = x_{n-1}$$
$$y_n = x_n + \sqrt{r} w_n$$

We have

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{y}) &\propto p(y_1|x_1)p(y_2|x_2)\dots p(x_1)p(x_2)\dots \\ &\propto p(y_1|x_1)p(y_2|x_2)\dots p(x_1)\delta(x_2 - x_1)\dots \\ &\propto p(y_1|x_1)p(y_2|x_1)\dots p(x_1) \end{aligned}$$

Multiple observation of a single unknown x_1.

Random walk prior

independent prior

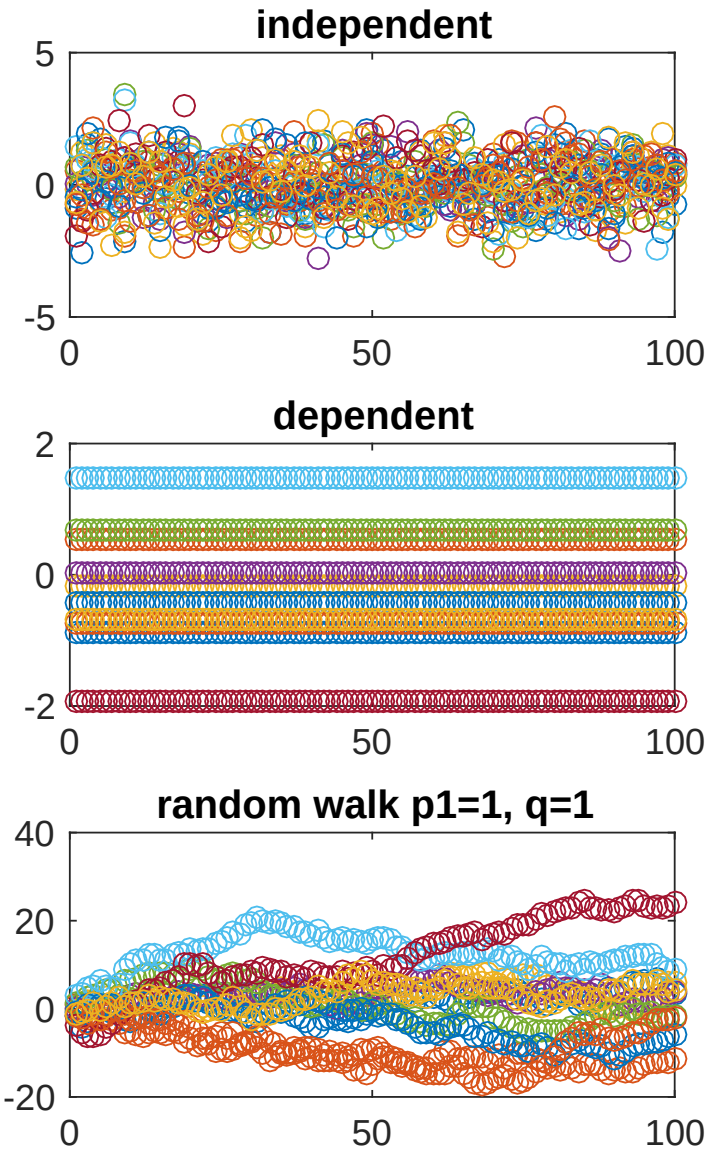
$$p(x_i) = \mathcal{N}(0, q)$$

dependent prior

$$p(x_i) = \delta(x_{i-1})$$

Random walk

$$p(x_i) = x_{i-1} + \mathcal{N}(0, q)$$



Inference with random walk