# Computational inverse problems

Nuutti Hyvönen, Jenni Heino,
Pauliina Hirvi

nuutti.hyvonen@aalto.fi, pauliina.hirvi@aalto.fi

Twelfth lecture

# Hypermodels

In the statistical framework, the prior densities usually depend on some parameters such as variance or mean. Typically — or at least thus far —, these parameters are assumed to be known.

Some classical regularization methods can be viewed as construction of estimators based on the posterior density (e.g., Tikhonov regularization). The regularization parameter, which corresponds to the parameter that defines the prior distribution, is not assumed to be know, but selected using, e.g., the Morozov discrepancy principle.

What happens if it is not clear how to choose these 'prior parameters' in the statistical framework?

If a parameter is not know, it can be estimated as a part of the statistical inference problem based on the data. This leads to hierarchical models that include hypermodels for the parameters defining the prior density.

Assume that the prior distribution depends on a parameter $\alpha$ which is not assumed to be known. Then we write the prior as a conditional density, that is,

$$\pi_{\mathrm{pr}}(x \,|\, \alpha).$$

Assuming we have a hyperprior for $\alpha$, i.e.,

$$\pi_{\mathrm{hyper}}(\alpha),$$

we can write the joint distribution of $x$ and $\alpha$ as

$$\pi(x, \alpha) = \pi_{\mathrm{pr}}(x \,|\, \alpha)\pi_{\mathrm{hyper}}(\alpha).$$

Assuming a likelihood model $\pi(y \,|\, x)$ for the measurement data $y$, we get the posterior density for $x$ and $\alpha$, given $y$, from the Bayes formula:

$$\pi(x, \alpha \,|\, y) \propto \pi(y \,|\, x)\pi(x, \alpha) = \pi(y \,|\, x)\pi(x \,|\, \alpha)\pi_{\mathrm{hyper}}(\alpha).$$

In general, the hyperprior density $\pi_{\text{hyper}}$ may depend on some hyperparameter $\alpha_0$. In such a case, the main reason for the use of a hyperprior model is that the construction of the posterior is assumed to be more robust with respect to fixing a value for the hyperparameter $\alpha_0$ than fixing a value for $\alpha$.

Sometimes $\alpha_0$ can also be treated as a random variable with a respective probability density. Then, we would write

$$\pi_{\text{hyper}}(\alpha \mid \alpha_0),$$

giving rise to nested hypermodels.

## Example: Hypermodel for a deconvolution problem

(Adapted from the textbook by Calvetti and Somersalo, Chapter 10)

Consider a one-dimensional deconvolution problem, the goal of which is to estimate a signal $f : [0, 1] \to \mathbb{R}$ from noisy, blurred observations modelled as

$$y_i = g(s_i) = \int_0^1 \mathcal{A}(s_i, t) f(t) dt + e(s_i), \qquad 1 \le i \le m,$$

where $\{s_i\}_{i=1}^m \subset [0, 1]$ are the uniformly distributed measurement points, the blurring kernel is defined to be

$$\mathcal{A}(s, t) = \exp\left(-\frac{1}{2\omega^2}(t - s)^2\right),$$

and the noise is Gaussian, or more precisely $e \sim \mathcal{N}(0, \sigma^2 I)$.

To begin with, we discretize the model as

$$y = Ax + e,$$

where $A \in \mathbb{R}^{m \times n}$ is obtained by approximating the integral with a suitable quadrature rule, and the vector $x$ contains the values of the unknown signal at the discretization points $\{t_j\}_{j=0}^{n}$ that we have chosen to be distributed uniformly over the interval $[0, 1]$. To be more precise,

$$x_j = f(t_j), \qquad t_j = \frac{j}{n}, \qquad 0 \leq j \leq n.$$

For simplicity we assume it is known that $f(0) = x_0 = 0$, and define the actual unknown $x$ to be

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n.$$

Assume that as prior information we know that the signal is continuous except for a possible jump discontinuity at a *known* location.

Let us start with a Gaussian first order smoothness prior,

$$\pi_{\mathrm{pr}}(x) \propto \exp\left(-\frac{1}{2\gamma^2}\|Lx\|^2\right),$$

where $L$ is a first order finite difference matrix (recall that $x_0 = 0$),

$$L = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

It is easy to see that $L$ is invertible and

$$L^{-1} = \begin{bmatrix} 1 & & & \\ 1 & 1 & & \\ \vdots & \ddots & \ddots & \\ 1 & \dots & 1 & 1 \end{bmatrix}$$

is a lower triangular matrix. Since $\frac{1}{\gamma}L$ is the whitening matrix of $X \in \mathbb{R}^n$ distributed according to $\pi_{\mathrm{pr}}(x)$ — see the eighth lecture —, it follows that

$$X \,=\, L^{-1}W, \qquad W \sim \mathcal{N}(0, \gamma^2 I).$$

Due to the particular shape of $L^{-1}$, this relation can alternatively be given as a Markov process:

$$X_j = X_{j-1} + W_j, \qquad W_j \sim \mathcal{N}(0, \gamma^2),\ j = 1, \dots, n, \quad X_0 = 0.$$

Next, we aim at fine-tuning the the above smoothness prior so that it allows a jump discontinuity over the interval $[t_{k-1}, t_k]$.

To this end, we modify the above Markov model (only) at $j = k$ by setting

$$X_k = X_{k-1} + W_k, \qquad W_k \sim \mathcal{N}\left(0, \frac{\gamma^2}{\delta^2}\right),$$

where $\delta < 1$ is a parameter controlling the variance of $W_k$, i.e., the expected size of the jump.

Let us walk the the above steps backwards: It is easy to see that this new Markov process can alternatively be given as

$$X \;=\; L^{-1}(D^{1/2})^{-1}W, \qquad W \sim \mathcal{N}(0, \gamma^2 I),$$

where

$$D^{1/2} = \mathrm{diag}(1, 1, \ldots, \delta, \ldots, 1, 1) \in \mathbb{R}^{n \times n}$$

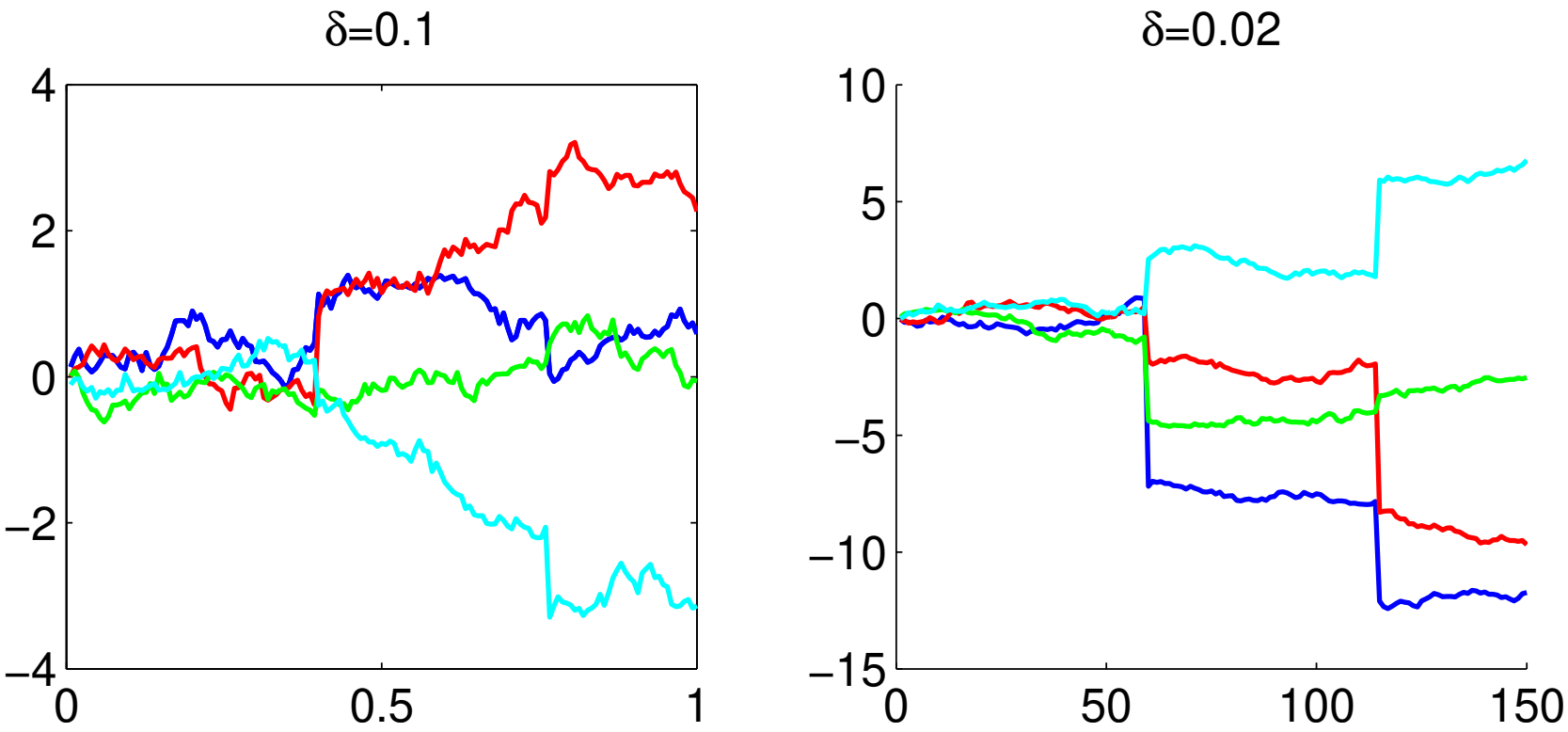
is defined so that $(D^{1/2})^{-1}$ scales the $k$th component of $W$ by $1/\delta$.

In consequence, after the above modification in the $k$th step of the Markov process *defining* $X$, the random variable $D^{1/2}LX$ is distributed according to $\mathcal{N}(0, \gamma^2 I)$, and thus we have introduced the fine-tuned 'jump prior'

$$\pi_{\mathrm{pr}}(x) \propto \exp\left(-\frac{1}{2\gamma^2}\|D^{1/2}Lx\|^2\right).$$

Let us draw samples from this kind of a prior density. We set $n = 150$ and $\gamma = 0.1$, meaning that we expect increments of the order $0.1$ at most of the subintervals. As an exception, at two known locations $t \approx 0.4$ and $t \approx 0.8$ we use $\delta < 1$ at the corresponding diagonal element of $D^{1/2}$, in anticipation of a jump of the order $\gamma/\delta = 0.1/\delta$.

Random draws from the jump discontinuity prior with two different values of $\delta$.



As the additive noise was assumed to be Gaussian, the likelihood density corresponding to the considered measurement is

$$\pi(y \,|\, x) \propto \exp\left(-\frac{1}{2\sigma^2}\|y - Ax\|^2\right),$$

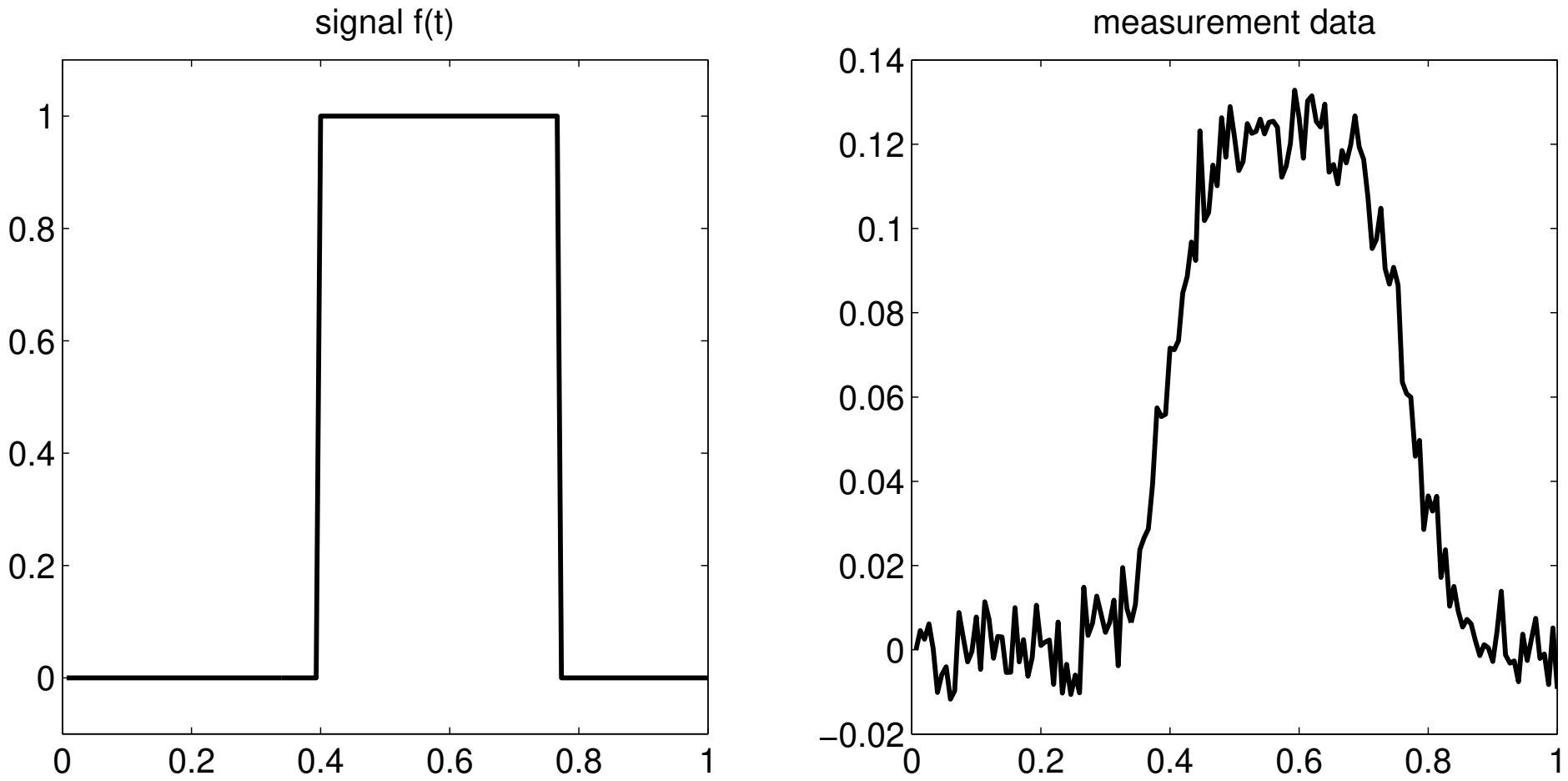
and due to the Bayes formula, the posterior density can thus be written as

$$\pi(x \,|\, y) \propto \exp\left(-\frac{1}{2\sigma^2}\|y - Ax\|^2 - \frac{1}{2\gamma^2}\|D^{1/2}Lx\|^2\right).$$

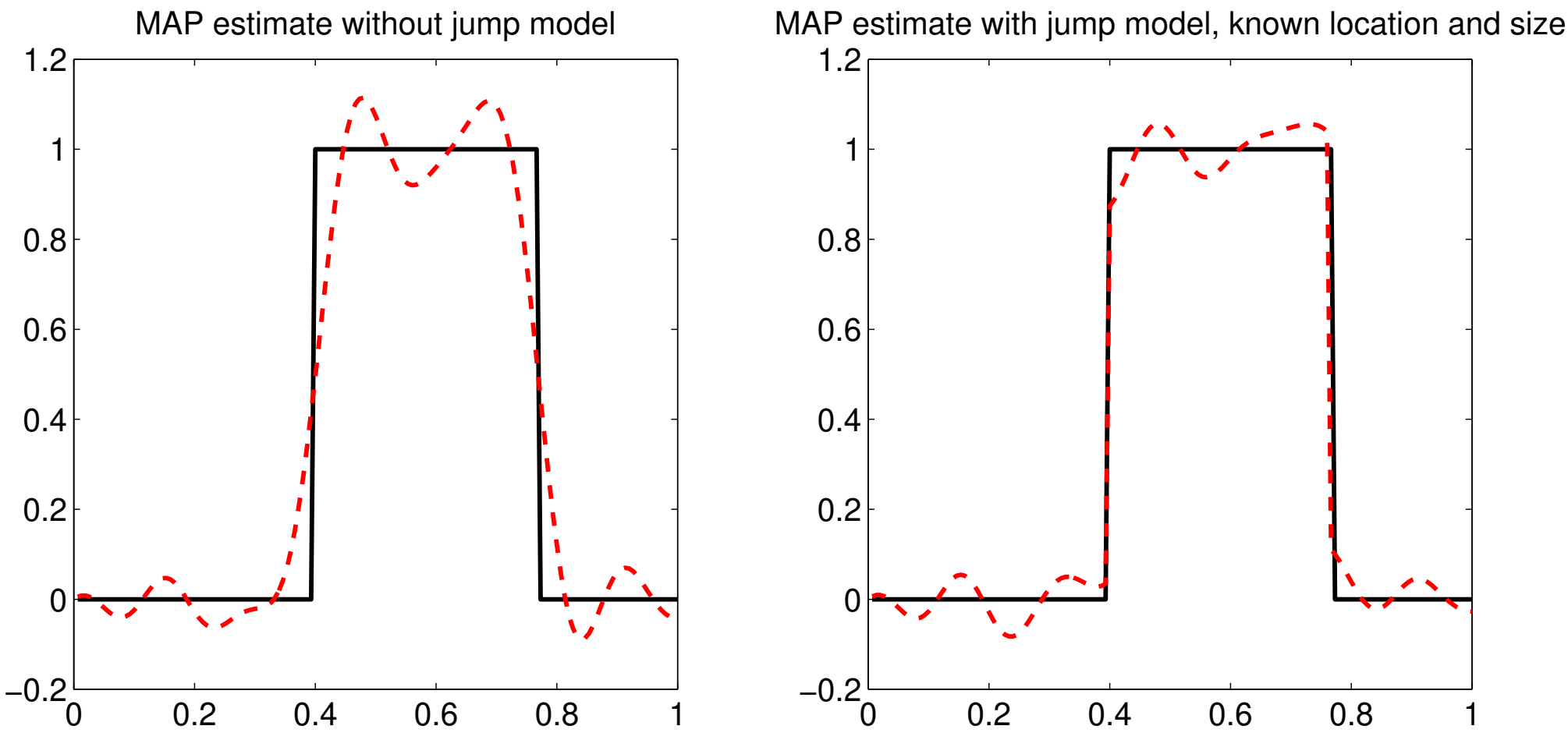
Using the results for Gaussian densities from previous lectures, the mean of the posterior, which is also the MAP and the CM estimate, can be written explicitly as

$$x_{\mathrm{CM}} = x_{\mathrm{MAP}} = \left(\frac{\sigma^2}{\gamma^2}L^{\mathrm{T}}(D^{1/2})^{\mathrm{T}}D^{1/2}L + A^{\mathrm{T}}A\right)^{-1} A^{\mathrm{T}}y.$$

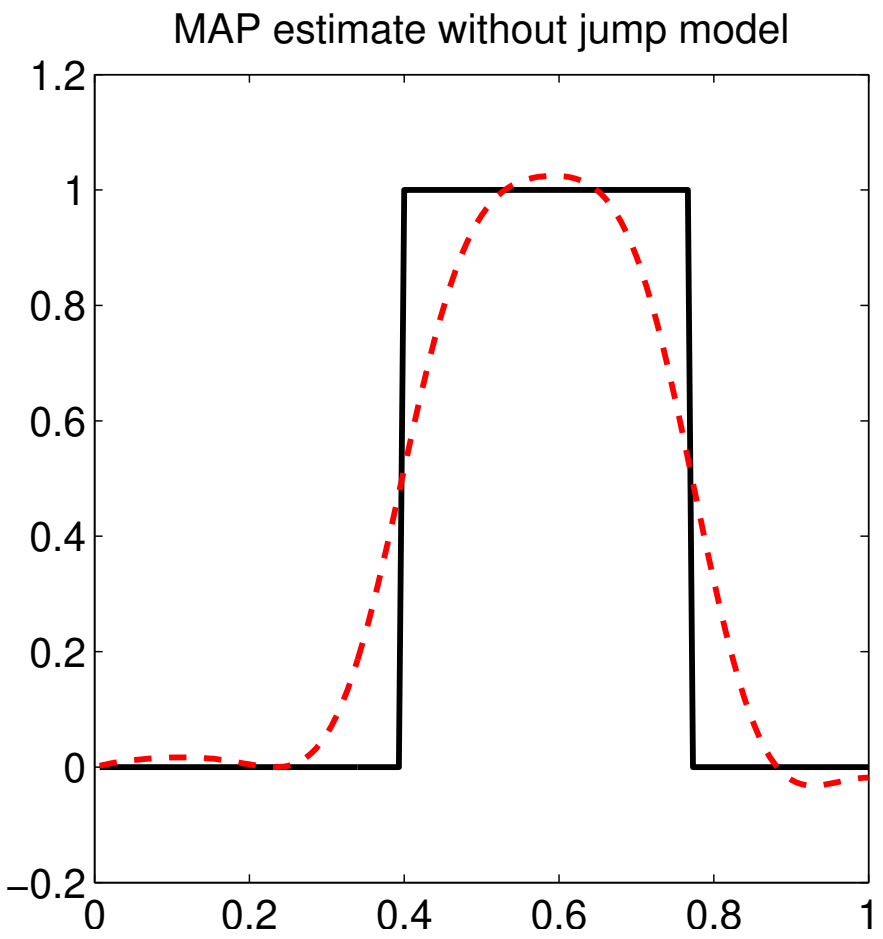
The original signal $f(t)$ and the measurement data ($\omega \approx 0.05$):

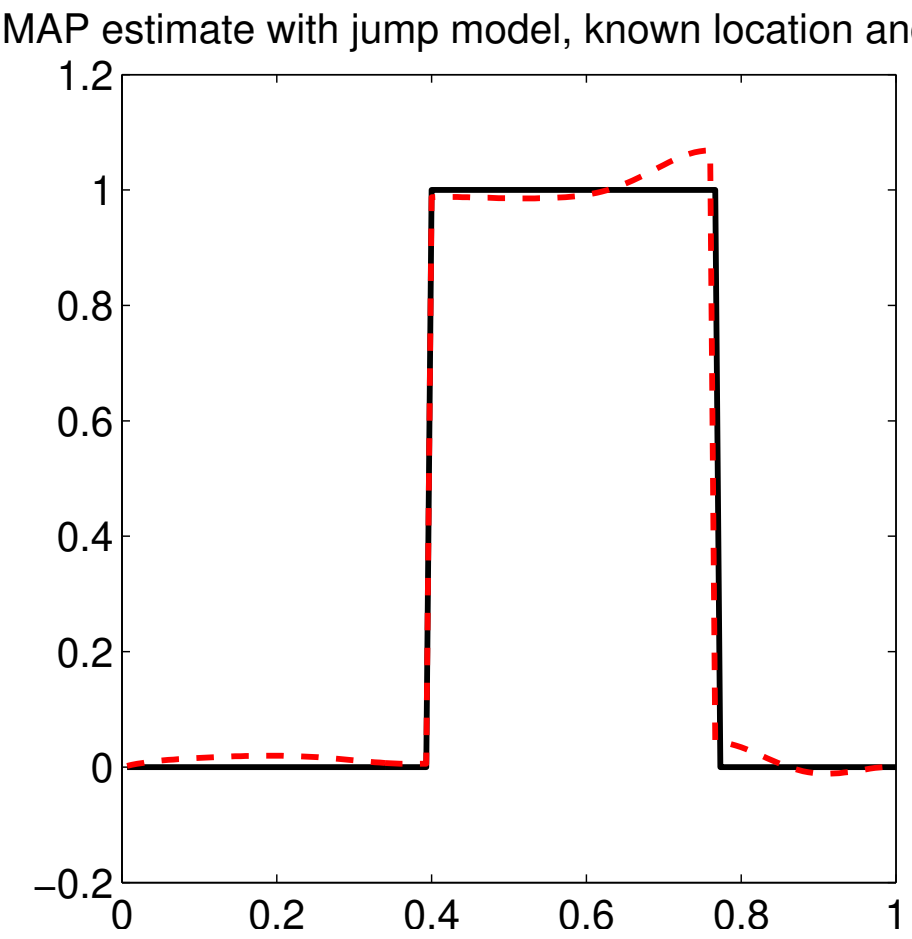


Posterior estimates for $f$ without the discontinuity model (i.e., with the mere first order smoothness prior) and with the discontinuity model with known locations and jump sizes ($\gamma = 0.1$):



Next we choose $\gamma = 0.01$ that corresponds to increments of the order of $0.01$ at each subinterval, and scale $\delta$ accordingly so that it is in accordance with jump sizes of the order $1$.



Assume next that the locations and expected sizes of the jumps are not known, but we expect *a slowly varying signal that could have a few jumps at unknown locations*.

We modify the Markov model to allow different increments at different positions:

$$X_j = X_{j-1} + W_j, \qquad W_j \sim \mathcal{N}\left(0, \frac{1}{\theta_j}\right), \quad \theta_j > 0, \quad j = 1, \ldots, n.$$

The corresponding prior model can be obtained in the same way as above:

$$\pi_{\mathrm{pr}}(x) \propto \exp\left(-\frac{1}{2}\|D^{1/2}Lx\|^2\right),$$

where this time around

$$D^{1/2} = \mathrm{diag}(\theta_1^{1/2}, \theta_2^{1/2}, \ldots, \theta_n^{1/2}).$$

If we knew the vector $\theta = [\theta_1, \ldots, \theta_n]^{\mathrm{T}}$, we could proceed as previously.

If $\theta \in \mathbb{R}^n$ is not know, it can be considered as a random variable and its estimation can be included as a part of the inference problem. To this end, we need to write the conditional density

$$\pi_{\mathrm{pr}}(x \,|\, \theta).$$

In this case, the normalizing constant of the density $\pi_{\mathrm{pr}}(x \,|\, \theta)$ is no longer a constant, but depends on the random variable $\theta$ and thus *cannot* be ignored.

Recall the probability density of a $n$-variate Gaussian distribution:

$$\pi(z) \,=\, \left(\frac{1}{(2\pi)^n \det(\Gamma)}\right)^{1/2} \exp\left(-\frac{1}{2} z^{\mathrm{T}} \Gamma^{-1} z\right),$$

where the mean is assumed to be zero.

In our case, $\Gamma = (L^{\mathrm{T}} D L)^{-1}$, where $D = \mathrm{diag}(\theta) \in \mathbb{R}^{n \times n}$. Recall that the determinant of a triangular matrix is the product of its diagonal elements, meaning that $\det(L) = \det(L^{\mathrm{T}}) = 1$. Moreover, the determinant of an inverse matrix is the inverse of the determinant of the original matrix. Hence, it holds that

$$\det(\Gamma)^{-1} \;=\; \det(L^{\mathrm{T}} D L) = \det(L^{\mathrm{T}}) \det(D) \det(L) = \prod_{j=1}^{n} \theta_j,$$

and the properly normalized density becomes

$$\begin{aligned} \pi_{\mathrm{pr}}(x \,|\, \theta) \;&=\; \left( \frac{\prod_{j=1}^{n} \theta_j}{(2\pi)^n} \right)^{1/2} \exp\left( -\frac{1}{2} \| D^{1/2} L x \|^2 \right) \\ &=\; \frac{1}{(2\pi)^{n/2}} \exp\left( -\frac{1}{2} \| D^{1/2} L x \|^2 + \frac{1}{2} \sum_{j=1}^{n} \log \theta_j \right) \Bigg). \end{aligned}$$

Next we need to choose a hyperprior density for $\theta$. Qualitatively, we should allow some components of $\theta$ to deviate strongly from the 'average'.

We decide to use an $\ell_1$-type impulse prior with a positivity constraint:

$$\pi_{\text{hyper}}(\theta) \propto \pi_+(\theta) \exp\left(-\frac{\gamma}{2}\sum_{j=1}^{n}\theta_j\right)$$

where $\pi_+(\theta)$ is one if all components of $\theta$ are positive, and zero otherwise, and $\gamma > 0$ is a hyperparameter.

The posterior distribution can then be written as

$$\pi(x,\theta \,|\, y) \propto \pi(y \,|\, x)\pi(x,\theta) = \pi(y \,|\, x)\pi(x \,|\, \theta)\pi_{\text{hyper}}(\theta)$$

$$\propto \exp\left( -\frac{1}{2\sigma^2}\|y - Ax\|^2 - \frac{1}{2}\|D^{1/2}Lx\|^2 - \frac{\gamma}{2}\sum_{j=1}^{n}\theta_j + \frac{1}{2}\sum_{j=1}^{n}\log\theta_j \right)$$

if all components of $\theta$ are positive, and $\pi(x,\theta \,|\, y) = 0$ otherwise. It is straightforward to see that the corresponding MAP estimate is the minimizer of the functional

$$F(x,\theta) = \left\| \begin{bmatrix} \frac{1}{\sigma}A \\ D^{1/2}L \end{bmatrix} x - \begin{bmatrix} \frac{1}{\sigma}y \\ 0 \end{bmatrix} \right\|^2 + \gamma\sum_{j=1}^{n}\theta_j - \sum_{j=1}^{n}\log\theta_j.$$

over $(x,\theta) \in \mathbb{R}^n \times \mathbb{R}^n_+$.

We apply a two stage minimization algorithm:

Choose some initial guesses for $x$ and $\theta$. Then, repeat the following two steps until convergence is achieved:

1. Keep $\theta$ fixed and update $x$ to be the least squares solution of

$$\begin{bmatrix} \frac{1}{\sigma}A \\ D^{1/2}L \end{bmatrix} x = \begin{bmatrix} \frac{1}{\sigma}y \\ 0 \end{bmatrix},$$

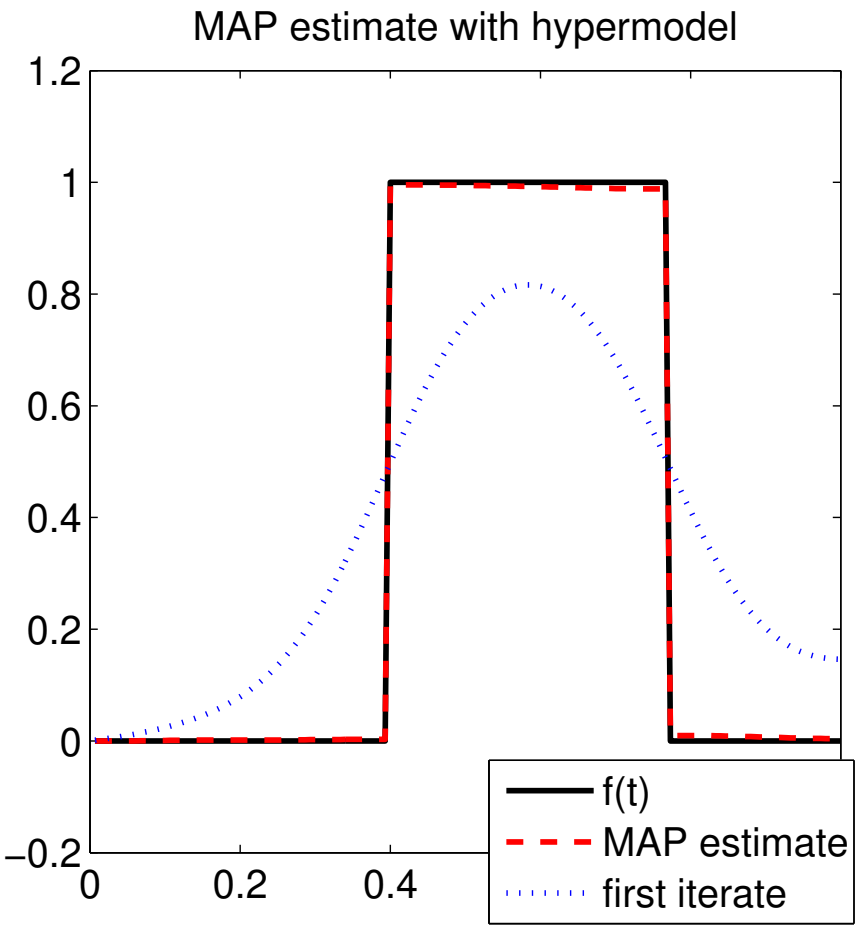
   where $D = \mathrm{diag}(\theta)$.

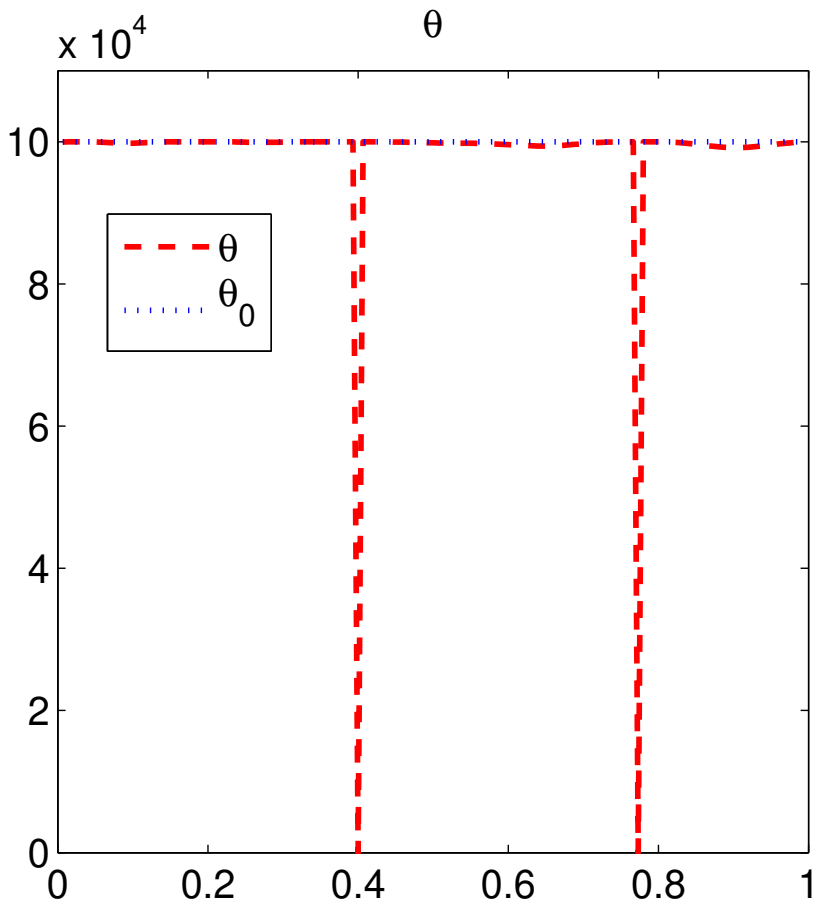
2. Fix $x$ and update $\theta$ by minimizing $F(x, \cdot)$ with respect to the second variable. An easy calculation shows that this minimizer can be given componentwise as

$$\theta_j = \frac{1}{w_j^2 + \gamma}, \qquad j = 1, \ldots, n,$$

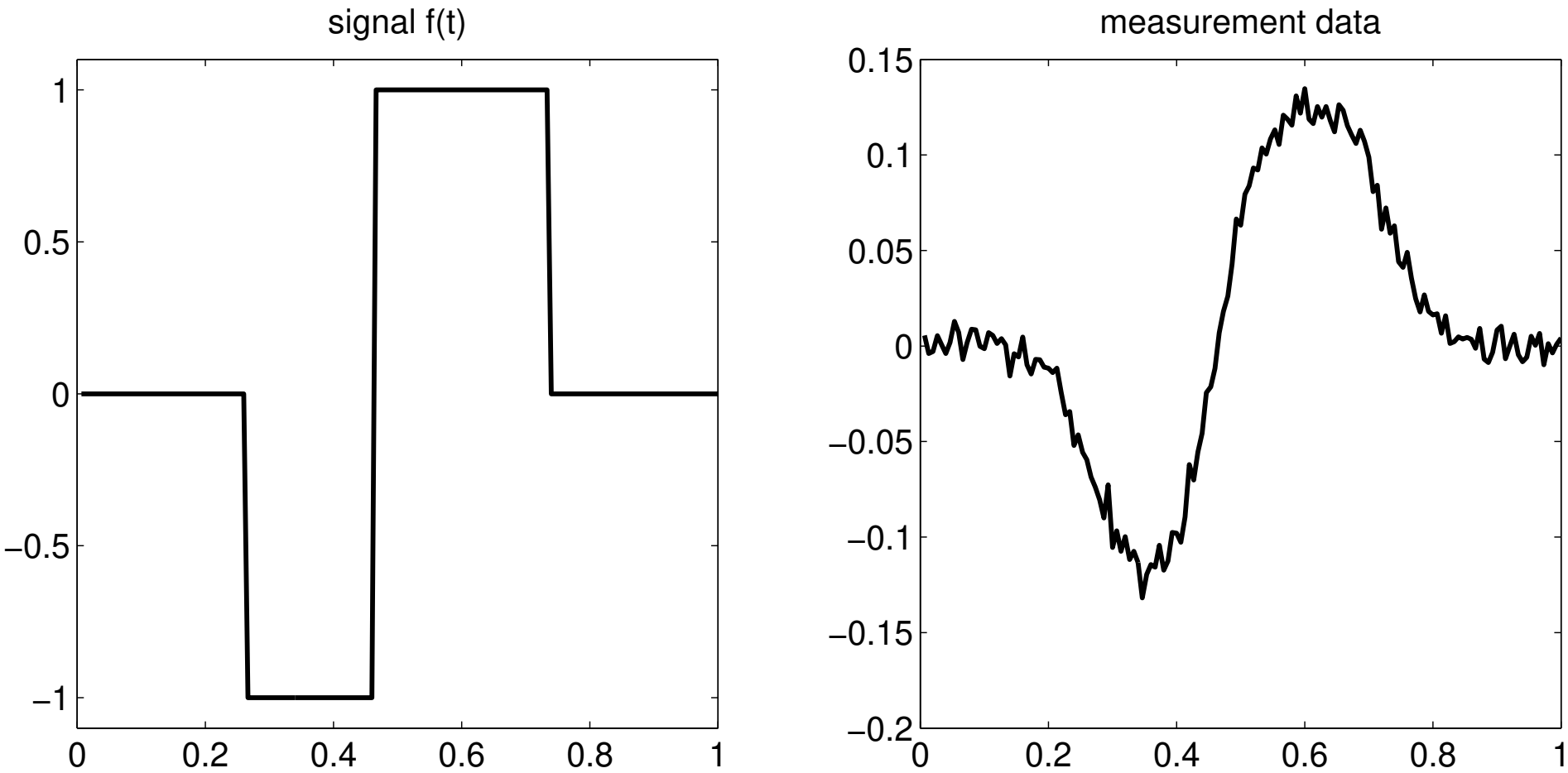
   where $w = Lx \in \mathbb{R}^n$ is the vector of increments corresponding to $x$.

MAP estimates for $x$ and $\theta$ provided by the above alternating algorithm with $\gamma = 10^{-5}$ and the initial guesses $x_0 = 0$ and $\theta_{0,j} = 1/\gamma$, $j = 1, \ldots, n$. The data is the same as depicted on page 448.

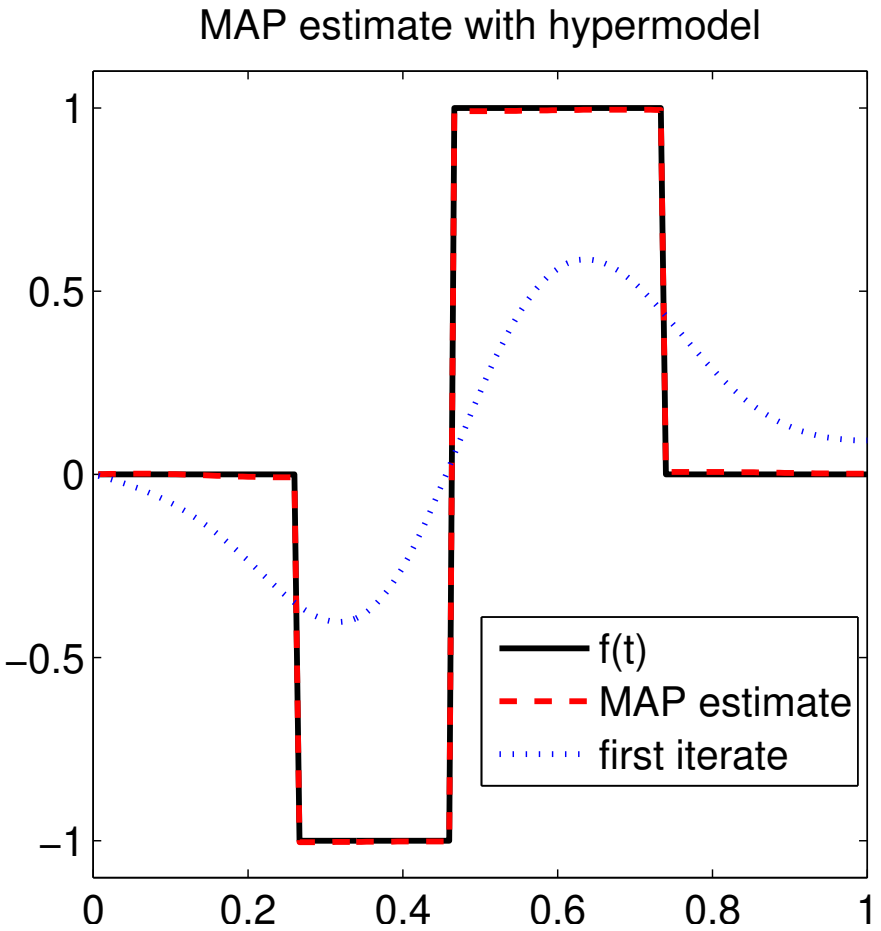


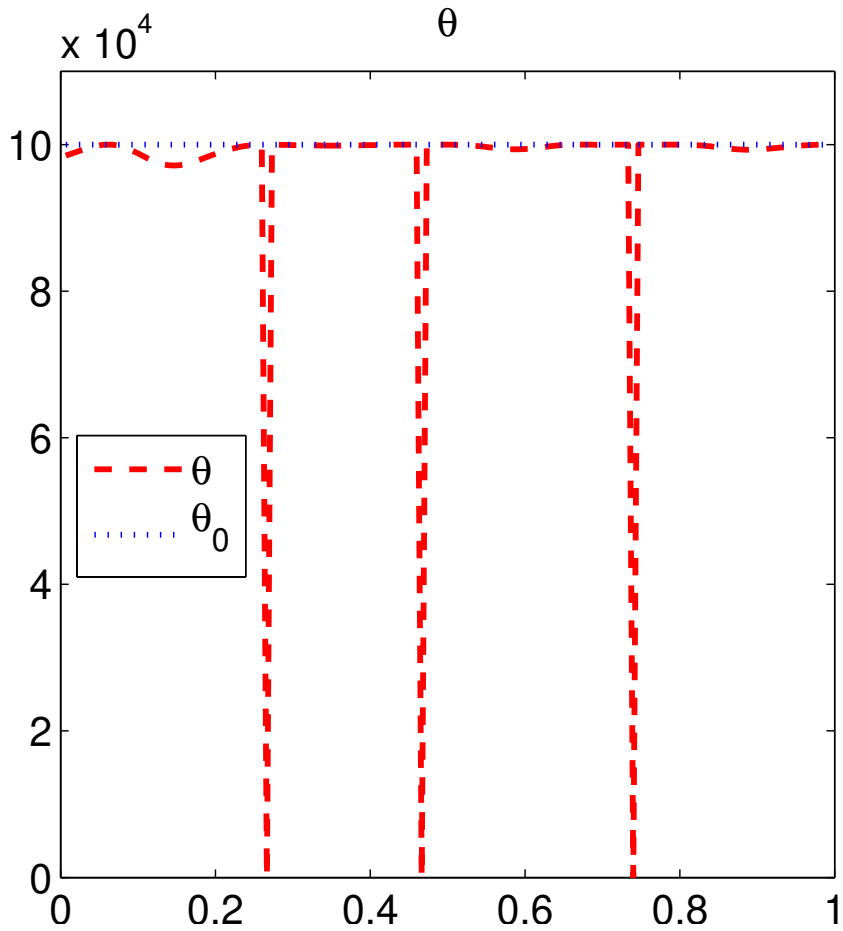



Another example: The original signal $f(t)$ and the measurement data.



MAP estimates for $x$ and $\theta$ provided by the above alternating algorithm with $\gamma = 10^{-5}$ and the initial guesses $x_0 = 0$ and $\theta_{0,j} = 1/\gamma$, $j = 1, \ldots, n$.



# The End.