An example of Bayesian reasoning

Consider the one-dimensional deconvolution problem with various degrees of prior information.

Model:

$$g(t) = \int_{-\infty}^{\infty} a(t-s) f(s) ds + e(t),$$

where

$$a(t)\big|_{|t|\to\infty} = 0 \quad \text{(rapidly)}.$$

The problem, even *without discretization errors* is ill-posed.

Analytic solution

Noiseless model:

$$g(t) = \int_{-\infty}^{\infty} a(t-s) f(s) ds.$$

Apply the Fourier transform,

$$\widehat{g}(k) = \int_{-\infty}^{\infty} e^{-ikt} g(t) dt.$$

The Convolution Theorem implies

$$\widehat{g}(k) = \widehat{a}(k)\widehat{f}(k),$$

so, by inverse Fourier transform,

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{itk} \frac{\widehat{g}(k)}{\widehat{a}(k)} dk.$$

Example: Gaussian kernel, noisy data:

$$a(t) = \frac{1}{\sqrt{2\pi\omega^2}} \exp\left(-\frac{1}{2\omega^2}t^2\right).$$

In the frequency domain,

$$\widehat{a}(k) = \exp\left(-\frac{1}{2}\omega^2 k^2\right).$$

If

$$\widehat{g}(k) = \widehat{a}(k)\widehat{f}(k) + \widehat{e}(k),$$

the estimate $f_{\rm est}(t)$ of f based on the Convolution Theorem is

$$f_{\rm est}(t) = f(t) + \frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{e}(k) \exp\left(\frac{1}{2}\omega^2 k^2 + ikt\right) dk,$$

which may not be even well defined unless $\widehat{e}(k)$ drops *faster than superexponentially* at high frequencies.

Discretized problem

Finite sampling, finite interval:

$$g(t_j) = \int_0^1 a(t_j - s) f(s) ds + e_j, \quad 1 \le j \le m.$$

Discrete approximation of the integral by a quadrature rule. Use the piecewise constant approximation with uniformly distributed sampling,

$$\int_0^1 a(t_j - s) f(s) ds \approx \sum_{j=1}^n \frac{1}{n} a(t_j - s_k) f(s_k) = \sum_{j=1}^n a_{jk} x_k,$$

where

$$s_k = \frac{k}{n}, \quad x_k = f(s_k), \quad a_{jk} = \frac{1}{n} a(t_j - s_k).$$

Additive Gaussian noise: likelihood

Discrete equation

$$b = Ax + e, \quad b_j = g(t_j).$$

Stochastic extension: X, E and B random variables,

$$B = AX + E.$$

Assume that E is Gaussian white noise with variance σ^2,

$$E \sim \mathcal{N}(0, \sigma^2 I), \quad \text{or} \quad \pi_{\text{noise}}(e) \propto \exp\left(-\frac{1}{2\sigma^2}\|e\|^2\right),$$

where σ^2 is assumed known. Likelihood density

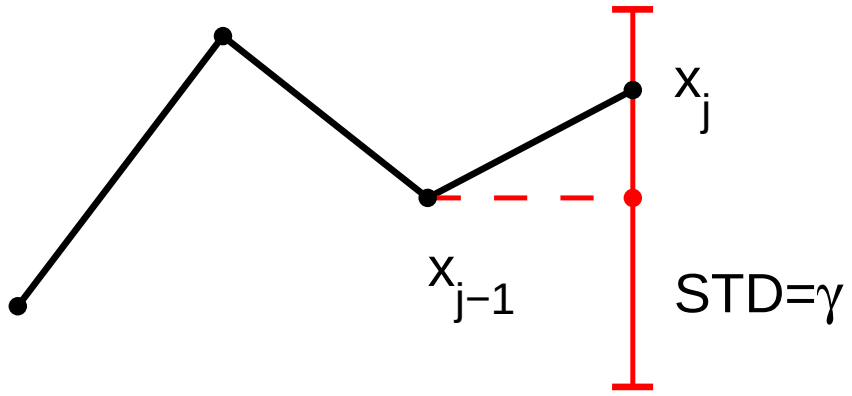
$$\pi(b \mid x) \propto \exp\left(-\frac{1}{2\sigma^2}\|b - Ax\|^2\right).$$

Prior according to the information available

Prior information: "*The signal f varies rather slowly, the slope being not more than of the order one, and vanishes at $t = 0$.*"

Write an *autoregressive Markov model* (ARMA):

$$X_j = X_{j-1} + \sqrt{\theta} W_j, \quad W_j \sim \mathcal{N}(0,1), \quad X_0 = 0.$$



How do we choose the variance θ of the innovation process?

Use the slope information: The slope is

$$S_j = n(X_j - X_{j-1}) = n\sqrt{\theta}W_j$$

and

$$\mathrm{P}\left\{|S_j| < 2n\sqrt{\theta}\right\} \approx 0.95, \quad 2n\sqrt{\theta} = \text{two standard deviations.}$$

A reasonable choice would then be

$$2n\sqrt{\theta} = 1, \quad \text{or} \quad \theta = \frac{1}{4n^2}.$$

The ARMA model in matrix form:

$$X_j - X_{j-1} = \sqrt{\theta} W_j, \quad 1 \le j \le n,$$

gives

$$LX = \sqrt{\theta} W, \quad W \sim \mathcal{N}(0, I),$$

where

$$L = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}.$$

Prior density: $(1/\sqrt{\theta})LX = W$, so

$$\pi_{\text{prior}}(x) \propto \exp(-\frac{1}{2\theta}\|Lx\|^2), \quad \theta = \frac{1}{4n^2}.$$

Posterior density

From Bayes' formula, the posterior density is

$$\begin{aligned} \pi_{\text{post}}(x) &= \pi(x \mid b) \propto \pi_{\text{prior}}(x)\pi(b \mid x) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\|b - Ax\|^2 - \frac{1}{2\theta}\|Lx\|^2\right). \end{aligned}$$

This is a Gaussian density, whose maximum is the Maximum A Posteriori (MAP) estimate,

$$x_{\text{MAP}} = \text{argmin}\left(\frac{1}{2\sigma^2}\|b - Ax\|^2 + \frac{1}{2\theta}\|Lx\|^2\right),$$

or, equivalently, the least squares solution of the system

$$\begin{bmatrix} \sigma^{-1}A \\ \theta^{-1/2}L \end{bmatrix} x = \begin{bmatrix} \sigma^{-1}b \\ 0 \end{bmatrix}.$$

Posterior covariance: by completing the square,

$$\frac{1}{2\sigma^2}\|b-Ax\|^2+\frac{1}{2\theta}\|Lx\|^2 = x^{\mathrm{T}}\underbrace{\left(\frac{1}{\sigma^2}A^{\mathrm{T}}A+\frac{1}{\theta}L^{\mathrm{T}}L\right)}_{=\Gamma^{-1}}x+2x^{\mathrm{T}}\left(\frac{1}{\sigma^2}A^{\mathrm{T}}b\right)+\frac{1}{\sigma^2}\|b\|^2$$

$$= \left(x-\Gamma\frac{1}{\sigma^2}A^{\mathrm{T}}b\right)^{\mathrm{T}}\Gamma^{-1}\left(x-\Gamma\frac{1}{\sigma^2}A^{\mathrm{T}}b\right)+\text{terms independent of } x.$$

Denoting

$$x_0 = \Gamma\frac{1}{\sigma^2}A^{\mathrm{T}}b = \left(\frac{1}{\sigma^2}A^{\mathrm{T}}A+\frac{1}{\theta}L^{\mathrm{T}}L\right)^{-1}\frac{1}{\sigma^2}A^{\mathrm{T}}b.$$

The posterior density is

$$\pi_{\rm post}(x) \propto \exp\left(-\frac{1}{2}(x-x_0)^{\rm T}\Gamma^{-1}(x-x_0)\right),$$

or in other words,

$$\pi_{\rm post} \sim \mathcal{N}(x_0, \Gamma),$$

$$\Gamma = \left(\frac{1}{\sigma^2}A^{\rm T}A + \frac{1}{\theta}L^{\rm T}L\right)^{-1},$$

$$x_0 = \Gamma\frac{1}{\sigma^2}A^{\rm T}b = \left(\frac{1}{\sigma^2}A^{\rm T}A + \frac{1}{\theta}L^{\rm T}L\right)^{-1}\frac{1}{\sigma^2}A^{\rm T}b = x_{\rm MAP}.$$

Example

Gaussian convolution kernel,

$$a(t) = \frac{1}{\sqrt{2\pi\omega^2}} \exp\left(-\frac{1}{2\omega^2}t^2\right), \quad \omega = 0.05.$$

Number of discretization points = 160.

See the attached Matlab file `Deconvloution.m`:

Test 1: Input function triangular pulse,

$$f(t) = \begin{cases} \alpha t, & t \le t_0 = 0.7, \\ \beta(1-t), & t > t_0, \end{cases}$$

where

$$\alpha = 0.6, \quad \beta = \alpha \frac{t_0}{1-t_0} = 1.4 > 1.$$

Hence, the prior information about the slopes is slightly too conservative.

Posterior credibility envelope

Plotting the MAP estimate and the 95% *credibility envelope*:

If X is distributed according to the Gaussian posterior density,

$$\mathrm{var}(X_j) = \Gamma_{jj}.$$

With 95% a posteriori probability

$$(x_0)_j - 2\sqrt{\Gamma_{jj}} < X_j < (x_0)_j + 2\sqrt{\Gamma_{jj}}.$$

Try the algorithm with various values of θ.

Test 2: Input function, whose slope is *locally* larger than one:

$$f(t) = 10(t-0.5)\exp\left(-\frac{1}{2\delta^2}(t-0.5)^2\right), \quad \delta^2 = 1000.$$

The slope in the interval $[0.4, 0.6]$ is much larger than one, while outside the interval it is very small. The algorithm does not perform well.

Assume that we, for some source, *have the following prior information*:

Prior information: "*The signal is almost constant (the slope of the order 0.1, say) outside the interval* $[0.4, 0, 6]$, *while within this interval, the slope can be of order 10.*"

Non-stationary ARMA model

Write Write an *autoregressive Markov model* (ARMA):

$$X_j = X_{j-1} + \sqrt{\theta_j} W_j, \quad W_j \sim \mathcal{N}(0,1), \quad X_0 = 0,$$

i.e., the variance of the innovation process varies.

How to choose the variance? Slope information:

The slope is

$$S_j = n(X_j - X_{j-1}) = n\sqrt{\theta_j} W_j$$

and

$$\mathrm{P}\left\{|S_j| < 2n\sqrt{\theta_j}\right\} \approx 0.95, \quad 2n\sqrt{\theta} = \text{two standard deviations.}$$

A reasonable choice would then be

$$2n\sqrt{\theta_j} = \begin{cases} 10, & 0.4 < t_j < 0.6 \\ 0.1 & \text{else} \end{cases}$$

Define the variance vector $\theta \in \mathbb{R}^n$,

$$\theta_j = \frac{25}{n^2}, \quad \text{if } 0.4 < t_j < 0.6 \quad \theta_j = \frac{1}{400n^2}, \quad \text{else,}$$

The ARMA model in matrix form:

$$D^{-1/2}LX = W \sim \mathcal{N}(0, I),$$

where

$$D = \text{diag}(\theta).$$

This leads to a prior model

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

Posterior density

From Bayes' formula, the posterior density is

$$\begin{aligned} \pi_{\text{post}}(x) &= \pi(x \mid b) \propto \pi_{\text{prior}}(x)\pi(b \mid x) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\|b - Ax\|^2 - \frac{1}{2}\|D^{-1/2}Lx\|^2\right). \end{aligned}$$

This is a Gaussian density, whose maximum is the Maximum A Posteriori (MAP) estimate,

$$x_{\text{MAP}} = \text{argmin}\left(\frac{1}{2\sigma^2}\|b - Ax\|^2 + \frac{1}{2}\|D^{-1/2}Lx\|^2\right),$$

or, equivalently, the least squares solution of the system

$$\begin{bmatrix} \sigma^{-1}A \\ D^{-1/2}L \end{bmatrix} x = \begin{bmatrix} \sigma^{-1}b \\ 0 \end{bmatrix}.$$

Posterior mean and covariance

Gaussian posterior density

$$\pi_{\text{post}} \sim \mathcal{N}(x_0, \Gamma),$$

where

$$\Gamma = \left(\frac{1}{\sigma^2} A^{\mathrm{T}} A + L^{\mathrm{T}} D^{-1} L \right)^{-1},$$

$$x_0 = \Gamma \frac{1}{\sigma^2} A^{\mathrm{T}} b = \left(\frac{1}{\sigma^2} A^{\mathrm{T}} A + L^{\mathrm{T}} D^{-1} L \right)^{-1} \frac{1}{\sigma^2} A^{\mathrm{T}} b = x_{\text{MAP}}.$$

Test 2 continued:

Use the non-stationary ARMA model based prior.

Try different values for the variance in the interval.

Observe the effect on the posterior covariance envelopes.

Test 3: Go back to the stationary ARMA model, and assume this time that

$$f(t) = \begin{cases} 1, & \tau_1 < t < \tau_2, \\ 0 & \text{else} \end{cases}$$

where $\tau_1 = 5/16$, $\tau_2 = 9/16$.

Observe that the prior is badly in conflict with the reality.

Notice that the posterior envelopes are very tight, but the true signal is not within.

Is it wrong to say: "*The p% credibility envelopes contain the true signal with probability p%.*"

Smoothness prior and discontinuities

Assume that the prior information is:

Prior information: "*The signal is almost constant (the slope of the order 0.1, say), but is suffers a jump of order one somewhere around* $t_50 = 6/16$ *and* $t_90 = 9/16$."

Build an ARMA model based on this information:

$$\theta_j = \frac{1}{400n^2}, \quad \text{except around } j = 50 \text{ and } j = 90,$$

and adjust the variances around the jumps to correspond to the jump amplitude, e.g.,

$$\theta_j = \frac{1}{4},$$

corresponding to standard deviation $\sqrt{\theta_j} = 0.5$.

Notice the large posterior variance around the jumps!

Prior based on incorrect information

Note: I avoid the term *incorrect prior*: prior is what we believe a priori, it is not right or wrong, but *evidence* may prove to be against it or supporting it.

In simulations, it is easy to judge a prior incorrect, because we control the "truth".

Test 4: What happens if we have incorrect information concerning the jumps?

- We believe that there is a third jump, which does not exist in the input; or
- the prior information concerning the jump locations is completely wrong.

A step up: unknown variance

Going back to the original problem, but with less prior information:

Prior information: "*The signal f varies rather slowly, and vanishes at $t = 0$, but no particular information about the slope is available.*"

As before, start with the initial ARMA model:

$$X_j = X_{j-1} + \sqrt{\theta} W_j, \quad W_j \sim \mathcal{N}(0,1), \quad X_0 = 0,$$

or in the matrix form,

$$LX = \sqrt{\theta} W, \quad W \sim \mathcal{N}(0, I).$$

As before, we write the prior, *pretending that we knew the variance*,

$$\pi_{\text{prior}}(x \mid \theta) = C \exp\left(-\frac{1}{2\theta} \|Lx\|^2 \right).$$

Normalizing constant: the integral of the density has to be one,

$$1 = C(\theta) \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2\theta}\|Lx\|^2\right) dx,$$

and with the change of variables

$$x = \sqrt{\theta} z, \quad dx = \theta^{n/2} dz,$$

we obtain

$$1 = \theta^{n/2} C(\theta) \underbrace{\int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}\|Lz\|^2\right) dz}_{\text{independent of } \theta},$$

so we deduce that

$$C(\theta) \propto \theta^{-n/2},$$

and we may write

$$\pi_{\text{prior}}(x \mid \theta) \propto \exp\left(-\frac{1}{2\theta}\|Lx\|^2 - \frac{n}{2}\log\theta\right).$$

Stochastic extension

Since θ is not known, it will be treated as a random variable Θ.

Any information concerning Θ is then coded the prior probability density called the *hyperprior*, $\pi_{\rm hyper}(\theta)$.

The inverse problem is to infer on the *pair* of unknown, (X, Θ).

The joint prior density is

$$\pi_{\rm prior}(x, \theta) = \pi_{\rm prior}(x \mid \theta)\pi_{\rm hyper}(\theta).$$

The posterior density of (X, Θ) is, by Bayes' formula,

$$\pi_{\rm post}(x, \theta) \propto \pi_{\rm prior}(x, \theta)\pi(b \mid x) = \pi_{\rm prior}(x \mid \theta)\pi_{\rm hyper}(\theta)\pi(b \mid x).$$

Here, we assume that the only prior information concerning Θ is its positivity,

$$\pi_{\text{hyper}}(\theta)\pi_+(\theta) = \begin{cases} 1, & \theta > 0 \\ 0, & \theta \le 0 \end{cases}$$

Note: this is, in fact an improper density, since i is not integrable. In practice, we assume an upper bound that we hope will never play a role.

The posterior density is now

$$\pi_{\text{post}}(x,\theta) \propto \pi_+(\theta)\exp\left(-\frac{1}{2\sigma^2}\|b - Ax\|^2 - \frac{1}{2\theta}\|Lx\|^2 - \frac{n}{2}\log\theta\right).$$

Iterative algorithm for the MAP estimate

Sequential optimization:

1. Initialize $\theta = \theta_0 > 0$, $k = 1$.
2. Update x,
$$x^k = \text{argmax}\big\{\pi_{\text{post}}(x \mid \theta_{k-1})\big\}.$$
3. Update θ,
$$\theta^k = \text{argmax}\big\{\pi_{\text{post}}(\theta \mid x_k)\big\}.$$
4. Increase k by one and repeat from 2. until convergence.

Notice: $\pi_{\text{post}}(x \mid \theta)$ means that we simply fix θ.

Updating steps in practice:

- Updating x: Since $\theta = \theta_{k-1}$ is fixed, we simply have

$$x_k = \text{argmin}\left\{\frac{1}{2\sigma^2}\|b - Ax\|^2 + \frac{1}{2\theta_{k-1}}\|Lx\|^2\right\},$$

  which is the Good Old Least Squares Solution (GOLSQ).

- Updating θ: The likelihood does not depend on θ, and so we have

$$\theta_k = \text{argmin}\left\{\frac{1}{2\theta}\|Lx_k\|^2 + \frac{n}{2}\log\theta\right\},$$

  which is the zero of the derivative of the objective function,

$$-\frac{1}{2\theta^2}\|Lx_k\|^2 + \frac{n}{2\theta} = 0,$$

  or

$$\theta_k = \frac{\|Lx_k\|^2}{n}.$$

Undetermined variances

Qualitative description of problem: "Given a noisy indirect observation, recover a signal or an image that **varies slowly** except for **unknown number of jumps of unknown size and location.**"

Information concerning the jumps:

- The jumps should be sudden, suggesting that the variances should be mutually independent.
- There is no obvious preference of one location over others, therefore the components should be identically distributed.
- Only a few variances can be significantly large, while most of them should be small, suggesting a hyperprior that allows rare outliers.

Candidate probability densities:

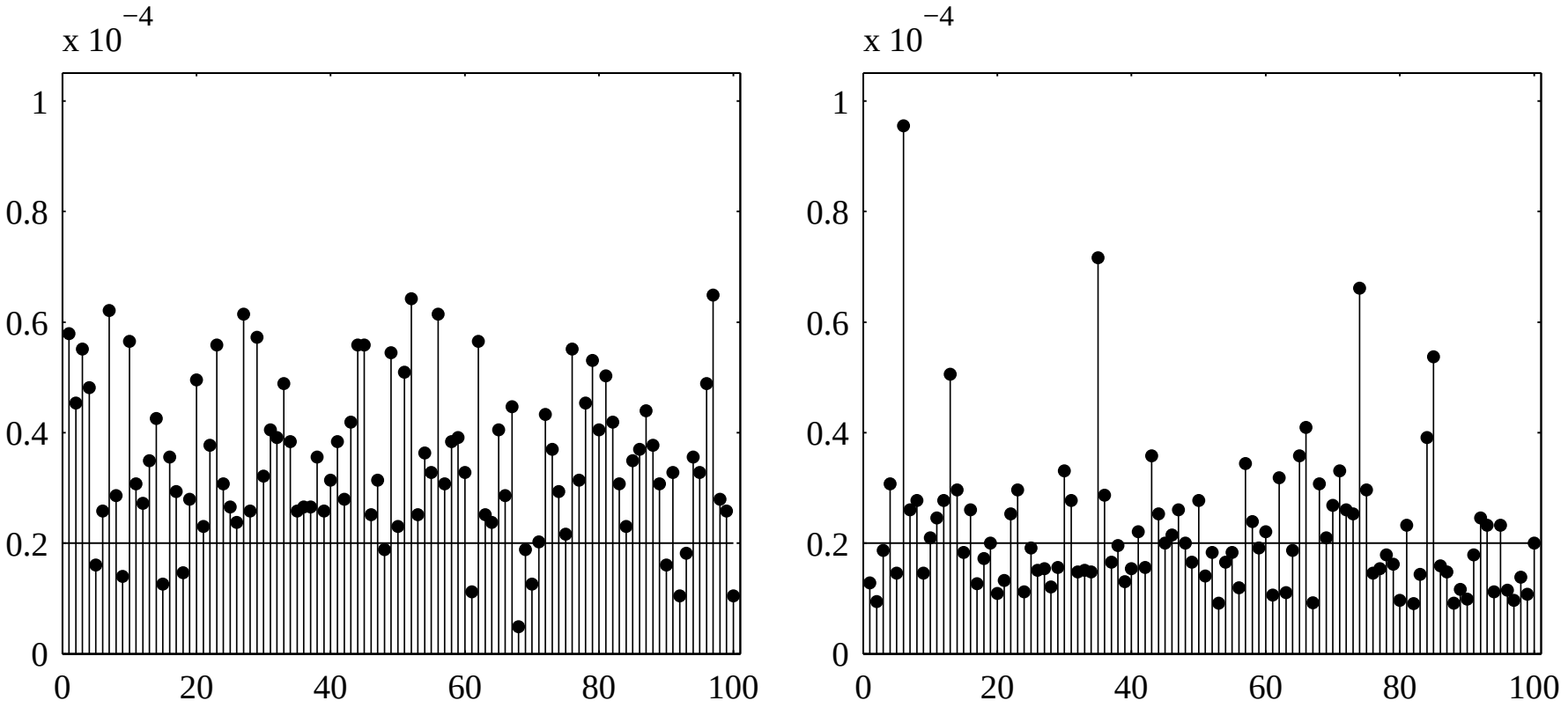
Gamma distribution,

$$\theta_j \sim \mathrm{Gamma}(\alpha, \theta_0), \quad \pi_{\mathrm{hyper}}(\theta) \propto \prod_{j=1}^{n} \theta_j^{\alpha-1} \exp\left(-\frac{\theta_j}{\theta_0}\right), \tag{1}$$

inverse Gamma distribution,

$$\theta_j \sim \mathrm{InvGamma}(\alpha, \theta_0), \quad \pi_{\mathrm{hyper}}(\theta) \propto \prod_{j=1}^{n} \theta_j^{-\alpha-1} \exp\left(-\frac{\theta_0}{\theta_j}\right). \tag{2}$$

Random Draws



Estimating the MAP

Outline of the algorithm is as follows:

1. Initialize $\theta = \theta^0$, $k = 1$.
2. Update the estimate of the increments: $x^k = \text{argmax}\, \pi(x, \theta^{k-1} \mid b)$.
3. Update the estimate of the variances θ: $\theta^k = \text{argmax}\, \pi(x^k, \theta \mid b)$.
4. Increase k by one and repeat from 2. until convergence.

In practice:

$$F(x, \theta \mid b) = -\log\big(\pi(x, \theta \mid b)\big),$$

which becomes, when the hyperprior is the gamma distribution,

$$F(x, \theta \mid b) \simeq \frac{1}{2\sigma^2}\|Az - b\|^2 + \frac{1}{2}\|D^{-1/2}Lx\|^2 + \frac{1}{\theta_0}\sum_{j=1}^{n}\theta_j - \left(\alpha - \frac{3}{2}\right)\sum_{j=1}^{n}\log\theta_j,$$

and, when the hyperprior is the inverse gamma distribution,

$$F(x, \theta \mid b) \simeq \frac{1}{2\sigma^2}\|Ax - b\|^2 + \frac{1}{2}\|D^{-1/2}Lx\|^2 + \theta_0\sum_{j=1}^{n}\frac{1}{\theta_j} + \left(\alpha + \frac{3}{2}\right)\sum_{j=1}^{n}\log\theta_j,$$

where $D = D_\theta$.

1. Updating x:

$$x^k = \text{argmin}\left(\frac{1}{2\sigma^2}\|Ax-b\|^2 + \frac{1}{2}\|D^{-1/2}Lx\|^2\right), \quad D = D_{\theta^{k-1}},$$

that is, x^k is the least squares solution of the system

$$\begin{bmatrix} (1/\sigma)A \\ D^{-1/2}L \end{bmatrix} x = \begin{bmatrix} (1/\sigma)b \\ 0 \end{bmatrix}.$$

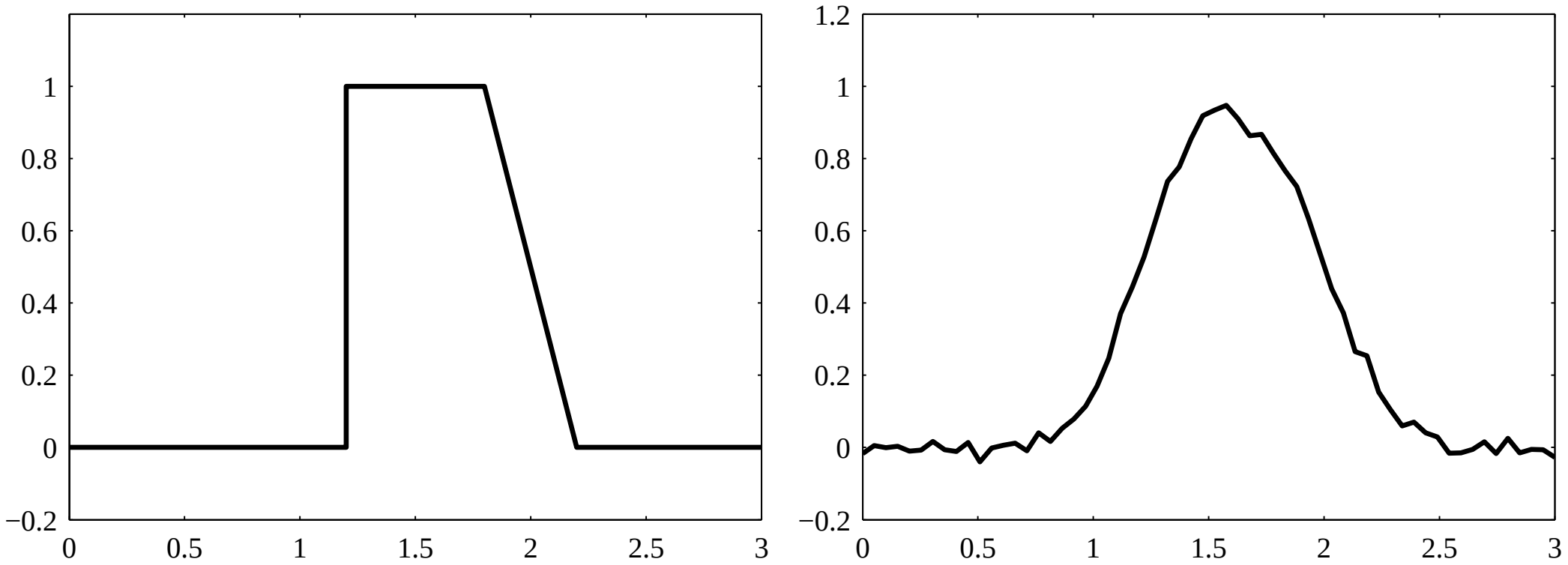
2. Updating θ: θ_j^k satisfies

$$\frac{\partial}{\partial\theta_j}F(x^k,\theta) = -\frac{1}{2}\left(\frac{z_j^k}{\theta_j}\right)^2 + \frac{1}{\theta_0} - \left(\alpha-\frac{3}{2}\right)\frac{1}{\theta_j} = 0,$$

$z^k = Lx^k$, which has an explicit solution,

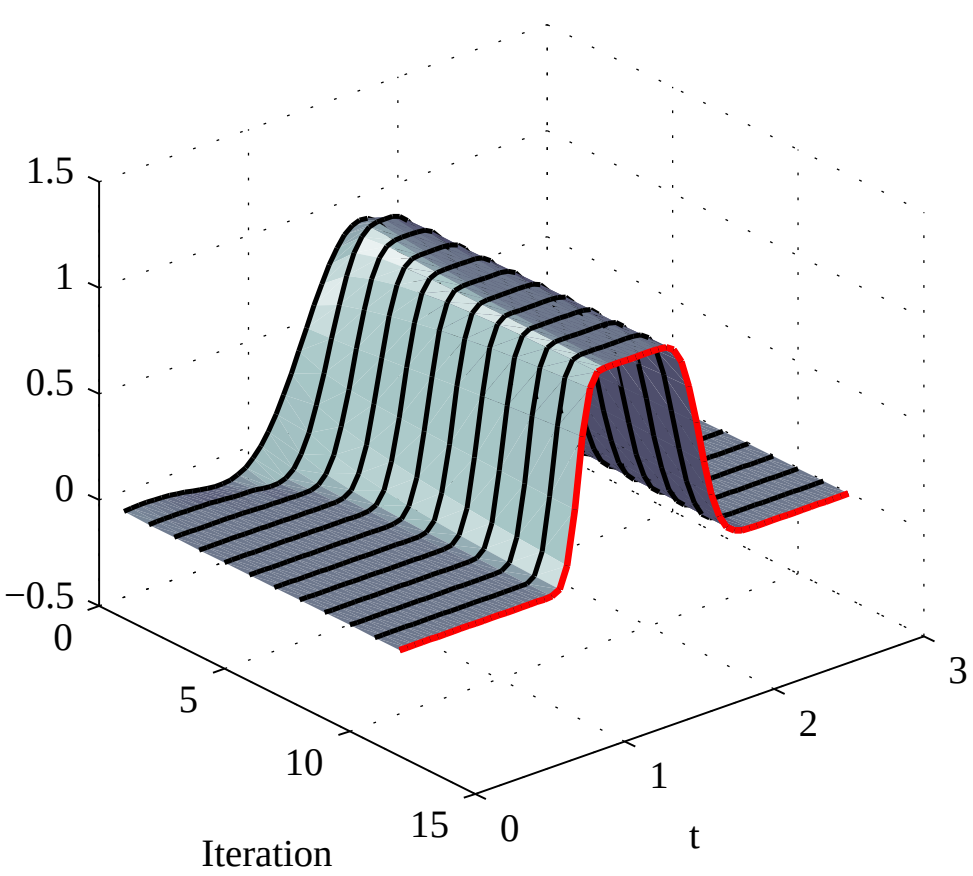
$$\theta_j^k = \theta_0\left(\eta + \sqrt{\frac{(z_j^k)^2}{2\theta_0}+\eta^2}\right), \quad \eta = \frac{1}{2}\left(\alpha-\frac{3}{2}\right).$$

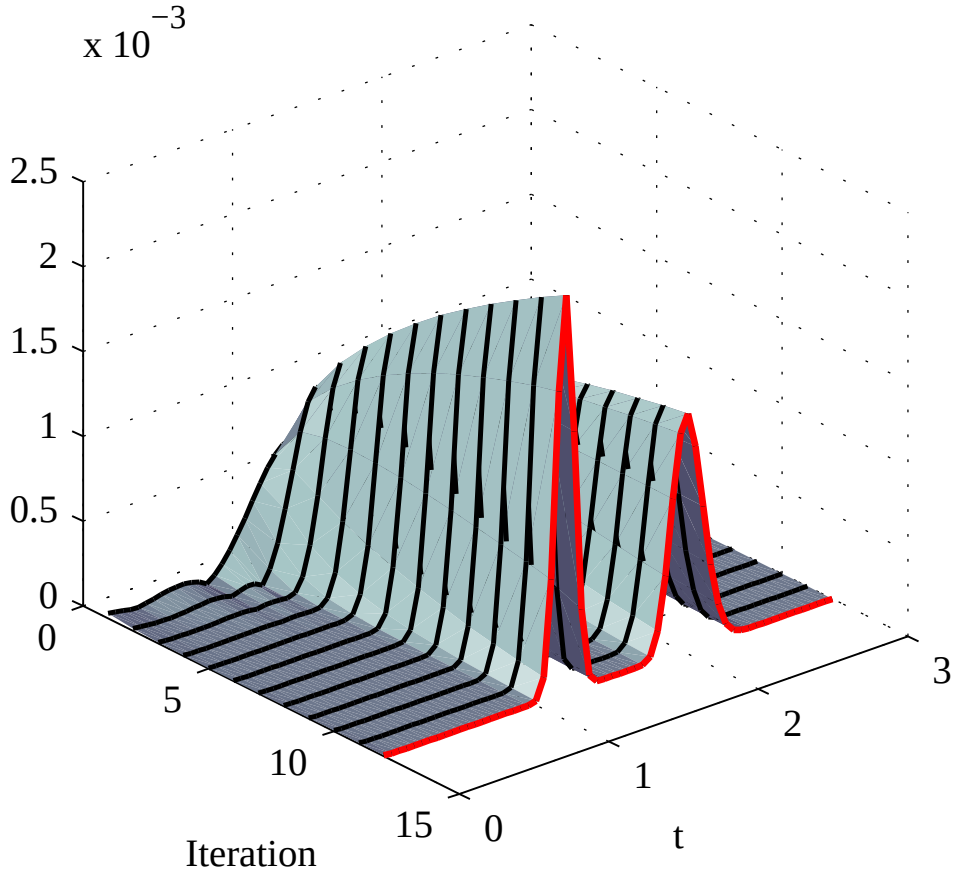
Computed example: Signal and data



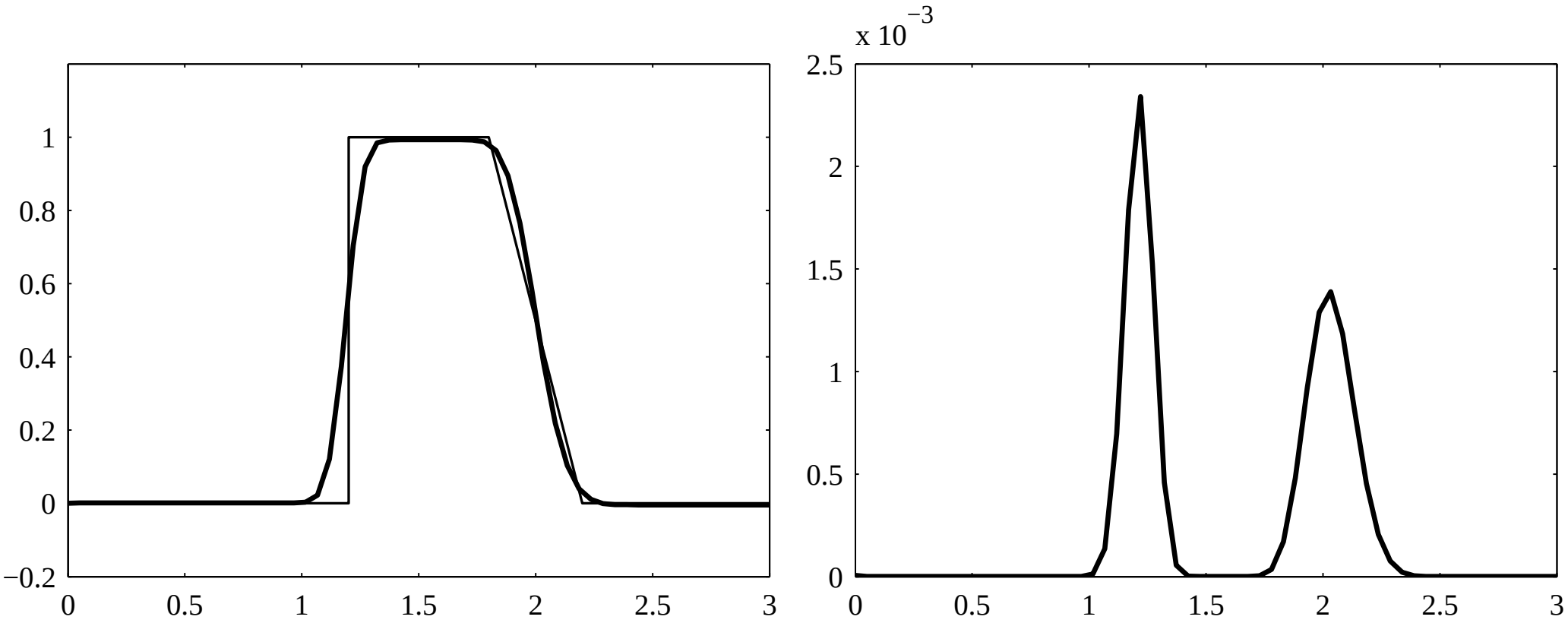
Gaussian blur, noise level 2%

MAP estimate, Gamma hyperprior

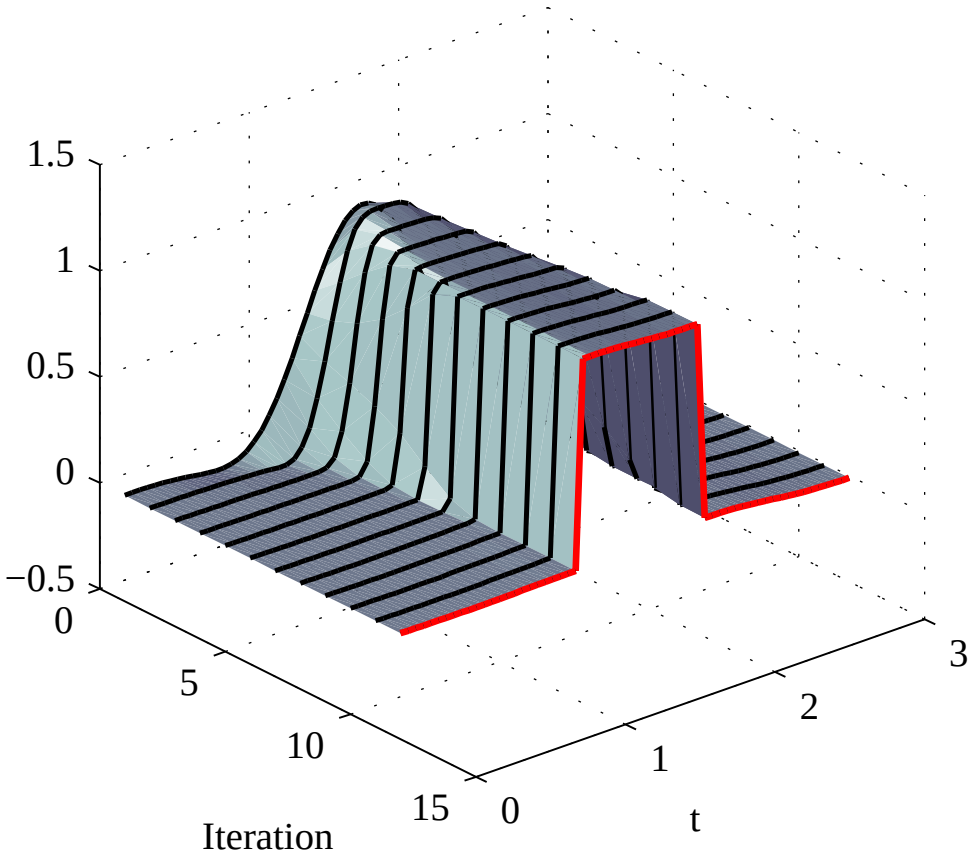


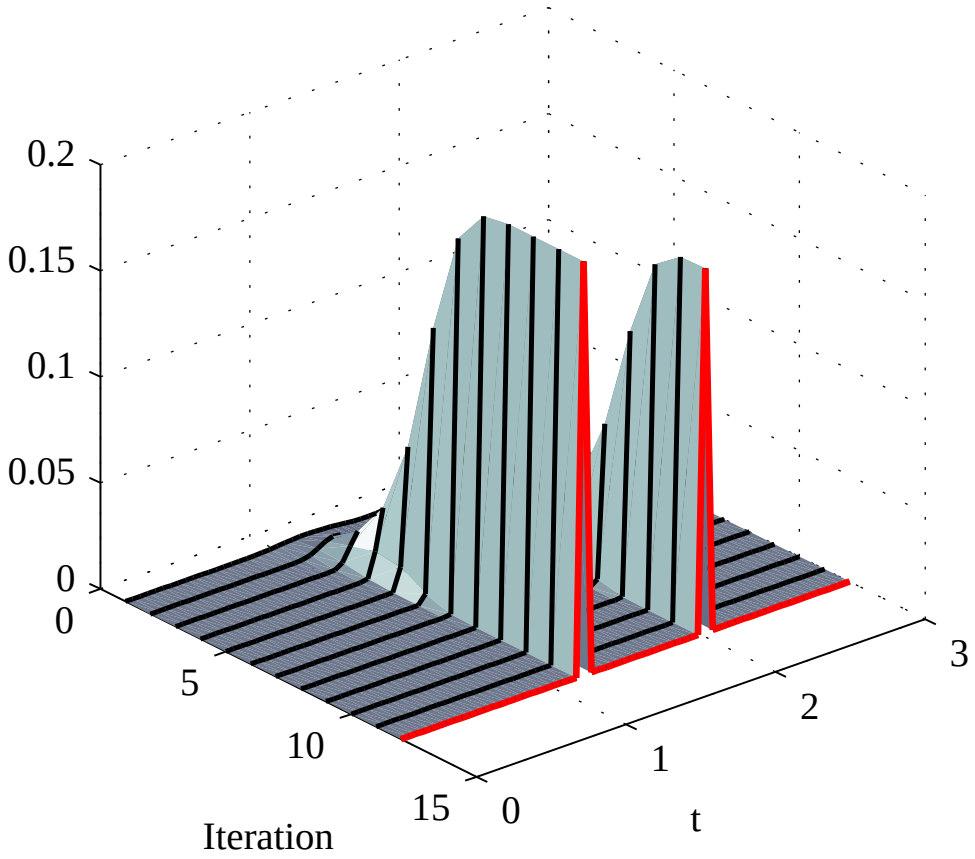


MAP estimate, Gamma hyperprior



MAP estimate, Inverse Gamma hyperprior





MAP estimate, Inverse Gamma hyperprior