

John Bardsley, University of Montana

Bayesian Scientific Computing

**Accompanies Chapter 10 of the text
Bayesian Scientific Computing, Calvetti & Somersalo**

An Example

Suppose $f : [0, 1] \rightarrow \mathbb{R}$ is a continuous signal discretized as follows:

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

A good prior for $\mathbf{x} = [x_1, \dots, x_n]^T$, supposing the signal varies gradually, is

$$\pi_{\text{prior}}(\mathbf{x}) \propto \exp\left(-\frac{1}{2\gamma^2}\|\mathbf{L}\mathbf{x}\|^2\right),$$

where \mathbf{L} is the first order derivative matrix

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

This corresponds to the Markov Model

$$X_j = X_{j-1} + W_j, \quad N(0, \gamma^2).$$

This corresponds to the Markov Model

$$X_j = X_{j-1} + W_j, \quad N(0, \gamma^2).$$

Now suppose we believe that in the interval $[t_{k-1}, t_k]$ the signal could have a large jump. Then for $j = k$, we might assume

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

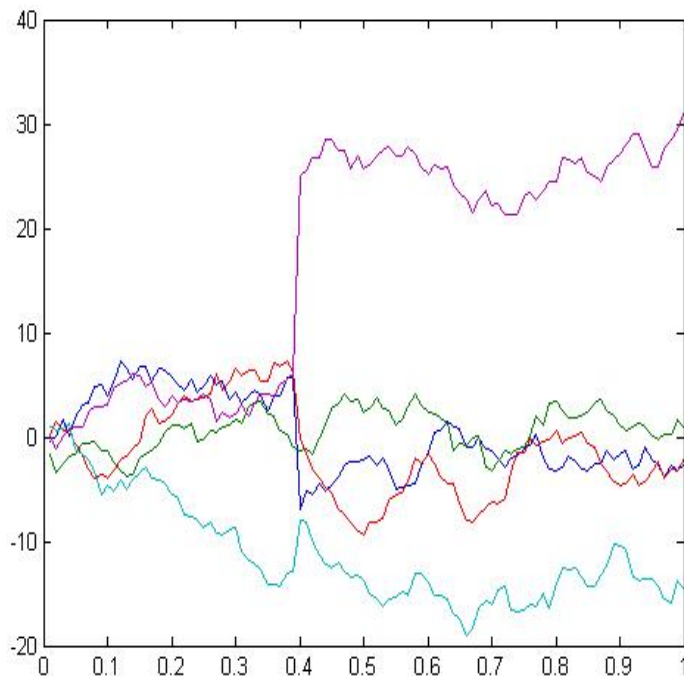
Then the prior becomes

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

where $\mathbf{D}^{1/2} \in \mathbb{R}^{n \times n}$ has the form

$$\mathbf{D}^{1/2} = \text{diag}\left(1, \dots, 1, \underbrace{\delta}_{k^{\text{th}}}, 1, \dots, 1\right).$$

Below are five draws from this prior with $\gamma = 1$ and $\delta = 0.02$ at $k=40$.



NOTE: Here we've assumed that we know: (i) the jump location, and (ii) the variance associate with the jump.

Example Continued

More common in signal processing is the knowledge of the presence of jumps, but not of their locations and variances. In this case, our Markov model takes the form

$$X_j = X_{j-1} + W_j, \quad N\left(0, \frac{1}{\theta_j}\right), \quad \theta_j > 0,$$

leading to the prior

$$\pi_{\text{prior}}(\mathbf{x}|\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\|\mathbf{D}^{1/2}\mathbf{L}\mathbf{x}\|^2\right),$$

where $\mathbf{D}^{1/2} \in \mathbb{R}^{n \times n}$ has the form

$$\mathbf{D}^{1/2} = \text{diag}(\theta_1^{1/2}, \theta_2^{1/2}, \dots, \theta_n^{1/2}).$$

NOTE: recall homework #2.

The Normalizing Constant for π_{prior}

Now that our prior depends upon auxiliary variables, the normalizing constant for π_{prior} can't be ignored (note that we will optimize over BOTH \mathbf{x} and $\boldsymbol{\theta}$). We have

$$\begin{aligned}\pi_{\text{prior}}(\mathbf{x}|\boldsymbol{\theta}) &= \sqrt{\frac{\det(\mathbf{L}^T \mathbf{D} \mathbf{L})}{(2\pi)^n}} \exp\left(-\frac{1}{2} \|\mathbf{D}^{1/2} \mathbf{L} \mathbf{x}\|^2\right) \\ &\propto \det(\mathbf{D})^{1/2} \exp\left(-\frac{1}{2} \|\mathbf{D}^{1/2} \mathbf{L} \mathbf{x}\|^2\right) \\ &= \left(\prod_{i=1}^n \theta_i^{1/2}\right) \exp\left(-\frac{1}{2} \sum_{j=1}^n \theta_j [\mathbf{L} \mathbf{x}]_j^2\right) \\ &= \exp\left(-\frac{1}{2} \sum_{j=1}^n \theta_j [\mathbf{L} \mathbf{x}]_j^2 + \frac{1}{2} \sum_{j=1}^n \ln \theta_j\right)\end{aligned}$$

Bayes' Law Again

Bayes' Law can handle the extra parameter θ , called a hyperparameter:

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

where $\pi_{\text{hyper}}(\theta)$ is known as the hyper-prior.

Bayes' Law and the Exponential Hyper-prior

Bayes' Law can handle the extra parameter θ , called a hyperparameter:

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

where $\pi_{\text{hyper}}(\boldsymbol{\theta})$ is known as the hyper-prior.

Let's make a distributional assumption on the θ_i 's:

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

where $\pi_+(\boldsymbol{\theta}) = \max(\boldsymbol{\theta}, \mathbf{0})$.

The Maximum A Posteriori Estimator

Now let's assume the linear statistical model

$$\mathbf{y} = \mathbf{A}\mathbf{x} + N(\mathbf{0}, \sigma^2\mathbf{I}).$$

Then the likelihood takes the form

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

and hence,

$$\begin{aligned} -\ln \pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) &\propto -\ln \pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) - \ln \pi_{\text{prior}}(\mathbf{x}|\boldsymbol{\theta}) - \ln \pi_{\text{hyper}}(\boldsymbol{\theta}) \\ &= \frac{1}{2\sigma^2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \frac{1}{2} \sum_{j=1}^n \theta_j [\mathbf{L}\mathbf{x}]_j^2 \\ &\quad - \frac{1}{2} \sum_{j=1}^n \ln \theta_j + \frac{\gamma}{2} \sum_{j=1}^n \theta_j \end{aligned}$$

Minimizing $-\ln \pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$

The simplest approach is two stage:

Stage 0: Initialize θ_0 , set $k = 1$.

Stage 1: Compute

$$\mathbf{x}^k = \arg \min_{\mathbf{x}} \left\{ \frac{1}{2\sigma^2} \|\mathbf{Ax} - \mathbf{b}\|^2 + \frac{1}{2} \sum_{j=1}^n \theta_{k-1,j} [\mathbf{Lx}]_j^2 \right\}$$

Stage 2: Compute

$$\boldsymbol{\theta}^k = \arg \min_{\boldsymbol{\theta}} \left\{ f(\boldsymbol{\theta}) = \frac{1}{2} \sum_{j=1}^n \left\{ \theta_j [\mathbf{Lx}_k]_j^2 - \ln \theta_j + \gamma \theta_j \right\} \right\}.$$

Then set $k = k + 1$ and return to Stage 1.

Stage 2 solution

Note that the computation of θ_k in Stage 2 is straightforward: solve

$$\nabla f(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} (\theta_1 [\mathbf{L}\mathbf{x}_k]_1^2 - \ln \theta_1 + \gamma \theta_1) \\ \vdots \\ \frac{\partial}{\partial \theta_n} (\theta_n [\mathbf{L}\mathbf{x}_k]_n^2 - \ln \theta_n + \gamma \theta_n) \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

Stage 2 solution

Note that the computation of θ_k in Stage 2 is straightforward: solve

$$\nabla f(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} (\theta_1 [\mathbf{Lx}_k]_1^2 - \ln \theta_1 + \gamma \theta_1) \\ \vdots \\ \frac{\partial}{\partial \theta_n} (\theta_n [\mathbf{Lx}_k]_n^2 - \ln \theta_n + \gamma \theta_n) \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

or

$$\theta_j^k = \frac{1}{[\mathbf{Lx}_k]_j^2 + \gamma}, \quad j = 1, \dots, n.$$

Thus we have the following clean formula for Stage 2:

$$\boldsymbol{\theta}^k = \frac{\mathbf{1}}{[\mathbf{Lx}_k]^2 + \boldsymbol{\gamma}},$$

where division and multiplication are component-wise, and $\mathbf{1}$ and $\boldsymbol{\gamma}$ are constant vectors of 1's and γ 's, respectively.

The Gamma hyper-prior

See `OneDBlurExp.m` for the implementation of the above approach with the 1D image deblurring example.

The exponential hyper-prior is unsatisfying because $0 < \theta_j \ll 1$'s corresponds to large jumps. Such values should come from the tails of the distribution. Let's try something different: let

$$\pi_{\text{prior}}(\mathbf{x}|\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\|\mathbf{D}^{-1/2}\mathbf{L}\mathbf{x}\|^2\right),$$

where \mathbf{D} and \mathbf{L} are as above, and the hyper-prior given by the Gamma distribution (mean $\alpha\theta_0$, variance $k\theta^2$)

$$\pi_{\text{hyper}}(\boldsymbol{\theta}|\alpha, \theta_0) \propto \prod_{j=1}^n \theta_j^{\alpha-1} \exp\left(-\frac{\theta_j}{\theta_0}\right)$$

The

$$-\ln \pi(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y}) \propto \frac{1}{2\sigma^2} \|\mathbf{Ax} - \mathbf{b}\|^2 + \frac{1}{2} \sum_{j=1}^n \theta_j^{-1} [\mathbf{Lx}]_j^2 \\ - \left(\alpha - \frac{3}{2} \right) \sum_{j=1}^n \ln \theta_j + \sum_{j=1}^n \frac{\theta_j}{\theta_0}$$

In the above two-stage algorithm, the first stage remains the same (note, however that now we are using θ_j^{-1}), while the second stage changes. In particular, the quadratic theorem gives **Stage 2** update

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_0 \left(\frac{\alpha - 1.5}{2} + \sqrt{\frac{(\alpha - 1.5)^2}{4} + \frac{(\mathbf{Lx}_k)^2}{2\boldsymbol{\theta}_0}} \right)$$

NOTE: Have a look at `OneDBLurGamma.m`.

The inverse-Gamma hyperprior

PROBLEM: Implement the above approach using the inverse-gamma hyperprior: (mean $\frac{\theta_0}{\alpha-1}$, variance $\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$)

$$\pi_{\text{hyper}}(\boldsymbol{\theta}|\alpha, \theta_0) \propto \prod_{j=1}^n \theta_j^{-\alpha-1} \exp\left(-\frac{\theta_0}{\theta_j}\right)$$

Once again, only stage 2 will change. Compute

$$\begin{aligned} -\ln \pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) &\propto \frac{1}{2\sigma^2} \|\mathbf{Ax} - \mathbf{b}\|^2 + \frac{1}{2} \sum_{j=1}^n \theta_j^{-1} [\mathbf{Lx}]_j^2 \\ &\quad + \frac{1}{2} \sum_{j=1}^n \ln \theta_j - \ln \pi_{\text{hyper}}(\boldsymbol{\theta}|\alpha, \theta_0), \end{aligned}$$

and then, for **Stage 2**,

$$\boldsymbol{\theta}_k = \text{NEW EXPRESSION.}$$

The inverse-Gaussian hyperprior

PROBLEM: Implement the above approach using the inverse-Gaussian hyperprior: (mean μ , variance μ^3/λ)

$$\pi_{\text{hyper}}(\boldsymbol{\theta}|\mu, \lambda) \propto \prod_{j=1}^n \theta_j^{-3/2} \exp\left(-\frac{\lambda(\theta_j - \mu)^2}{2\mu^2\theta_j}\right)$$

Once again, only stage 2 will change. Compute

$$\begin{aligned} -\ln \pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) &\propto \frac{1}{2\sigma^2} \|\mathbf{Ax} - \mathbf{b}\|^2 + \frac{1}{2} \sum_{j=1}^n \theta_j^{-1} [\mathbf{Lx}]_j^2 \\ &\quad + \frac{1}{2} \sum_{j=1}^n \ln \theta_j - \ln \pi_{\text{hyper}}(\boldsymbol{\theta}|\mu, \lambda), \end{aligned}$$

and then, for **Stage 2**,

$$\boldsymbol{\theta}_k = \text{NEW EXPRESSION.}$$

Sampling from the posterior

Above are examples in which:

- we can't sample from $\pi(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y})$;
- we can sample from $\pi(\mathbf{x} | \boldsymbol{\theta}, \mathbf{y})$ and $\pi(\boldsymbol{\theta} | \mathbf{x}, \mathbf{y})$.

Thus Gibbs sampling is a natural choice.

Hyper-models for variance estimation

Let's return to the nonlinear least squares problem

$$y_i = \mathbf{A}(\mathbf{x})_i + N(0, \sigma^2)$$

which has likelihood function

$$-\ln \pi(\mathbf{y}|\mathbf{x}, \sigma^2) = \frac{1}{2\sigma^2} \sum_{i=1}^n (\mathbf{A}(\mathbf{x})_i - y_i)^2$$

If we assume that σ^2 is unknown, Bayes' rule takes the form

$$\pi(\mathbf{x}, \sigma^2|\mathbf{y}) \propto \pi(\mathbf{y}|\mathbf{x}, \sigma^2)\pi_{\text{prior}}(\mathbf{x}|\sigma^2)\pi_{\text{hyper}}(\sigma^2).$$

Assuming

$$\pi_{\text{prior}}(\mathbf{x}|\sigma^2) \sim U(\mathbf{l}, \mathbf{u})$$

where \mathbf{l} and \mathbf{u} are lower and upper-bound vectors, and $\sigma^2 \sim \text{Gamma}(\alpha, \theta_0)$ so that

$$\pi_{\text{hyper}}(\sigma^2) \propto (\sigma^2)^{\alpha-1} \exp(-\sigma^2/\theta_0),$$

where the mean $\alpha\theta_0 = \text{mse}$, we can sample from the posterior in a Gibbsian fashion: given $(\sigma_k^2, \mathbf{x}_k)$

1. sample $\sigma_{k+1}^2 \sim \pi(\sigma^2|\mathbf{x}_k, \mathbf{y})$ using **conjugate prior** or inverse-cdf methods;
2. sample $\mathbf{x}_{k+1} \sim \pi(\mathbf{x}_k|\sigma_k^2, \mathbf{y})$ using Metropolis-Hastings.

Conjugate Prior Method: Note that if

$$z_i \sim N(\mu, \sigma^2), \quad i = 1, \dots, n,$$

with σ^2 unknown, and the prior

$$\pi_{\text{prior}}(\sigma^2) \sim \text{Inv - Gamma}(\alpha, \theta_0)$$

is assumed, then

$$\begin{aligned} \pi(\sigma^2 | \mathbf{z}) &\propto \pi(\mathbf{z} | \sigma^2) \pi_{\text{prior}}(\sigma^2) \\ &\sim \text{Inv - Gamma} \left(\alpha + \frac{n}{2}, \theta_0 + \frac{1}{2} \sum_{i=1}^n (z_i - \mu)^2 \right) \end{aligned}$$

So in the previous slide, step 1 of Gibbs has the form

1. sample $\sigma_{k+1}^2 \sim \text{Inv - Gamma} \left(\alpha + \frac{n}{2}, \theta_0 + \frac{1}{2} \sum_{i=1}^n (\mathbf{A}(\mathbf{x})_i - y_i)^2 \right)$

NOTE: See `NonlinExample1MCMCVarSampConj.m` and accompanying hand-written notes.

MATLAB: `1/gamrnd(a, 1/t0)` gives a sample from $\text{Inv - Gamma}(a, t_0)$.

Inverse-CDF Method: for step 1 in Gibbs above:

(a) First, evaluate $\pi(\sigma^2 | \mathbf{x}_k, \mathbf{y})$, which is proportional to

$$\exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (\mathbf{A}(\mathbf{x})_i - y_i)^2 + (\alpha - 1) \ln \sigma^2 - \sigma^2 / \theta_0\right),$$

on an appropriate grid (can be hard);

(b) next, create the empirical cumulative distribution vector/function Φ ;

(c) then draw $t \sim U(0, 1)$, and choose σ_k^2 so that $\Phi(\sigma_k^2) = t$ (approximately).

NOTE: See `NonlinExample1MCMCVarSamp.m`.

Sampling Images

Now, let's return to the image processing example

$$\mathbf{b} \sim \mathbf{A}\mathbf{x} + N(\mathbf{0}, \sigma^2\mathbf{I})$$

$$\mathbf{x} \sim N(\mathbf{0}, \delta^{-1}\mathbf{L}^T\mathbf{L})$$

$$\sigma^2 \sim \text{Gamma}(a, t_0) \propto (\sigma^2)^{a-1} \exp(-\sigma^2/t_0)$$

$$\delta \sim \text{Rayleigh}(a_0) \propto (\delta/a_0^2) \exp(-\delta/2a_0^2)$$

Sampling Images

Now, let's return to the image processing example

$$\mathbf{b} \sim \mathbf{Ax} + N(\mathbf{0}, \sigma^2 \mathbf{I})$$

$$\mathbf{x} \sim N(\mathbf{0}, \delta^{-1} \mathbf{L}^T \mathbf{L})$$

$$\sigma^2 \sim \text{Gamma}(a, t_0) \propto (\sigma^2)^{a-1} \exp(-\sigma^2/t_0)$$

$$\delta \sim \text{Rayleigh}(a_0) \propto (\delta/a_0^2) \exp(-\delta/2a_0^2)$$

$$\begin{aligned} \pi(\mathbf{x}, \delta, \sigma^2 | \mathbf{b}) &\propto \pi(\mathbf{b} | \mathbf{x}, \delta, \sigma^2) \pi(\mathbf{x} | \delta, \sigma^2) \pi(\delta, \sigma^2) \\ &= \pi(\mathbf{b} | \mathbf{x}, \delta, \sigma^2) \pi(\mathbf{x} | \delta) \pi(\delta) \pi(\sigma^2) \\ &= \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{Ax} - \mathbf{b}\|^2 - \frac{\delta}{2} \|\mathbf{Lx}\|\right) \\ &\quad \times \exp\left((a-1) \ln \sigma^2 - \sigma^2/t_0\right) \\ &\quad \times \exp\left(-\frac{1}{2} (\delta/a_0)^2 + \frac{n+2}{2} \ln \delta\right) \end{aligned}$$

NOTE: $\delta^{n/2}$ comes from the normalization term for the prior.

Gibbs Sampling

Given \mathbf{x}_{k-1} , σ_{k-1}^2 , and δ_{k-1} ,

1. sample $\mathbf{x}_k \sim \pi(\mathbf{x}|\delta_{k-1}, \sigma_{k-1}^2, \mathbf{b})$ where

$$\pi(\mathbf{x}|\delta_{k-1}, \sigma_{k-1}^2, \mathbf{b}) \propto \exp\left(-\frac{1}{2\sigma_{k-1}^2}\|\mathbf{Ax} - \mathbf{b}\|^2 - \frac{\delta_{k-1}}{2}\|\mathbf{Lx}\|^2\right);$$

Gibbs Sampling

Given \mathbf{x}_{k-1} , σ_{k-1}^2 , and δ_{k-1} ,

1. sample $\mathbf{x}_k \sim \pi(\mathbf{x}|\delta_{k-1}, \sigma_{k-1}^2, \mathbf{b})$ where

$$\pi(\mathbf{x}|\delta_{k-1}, \sigma_{k-1}^2, \mathbf{b}) \propto \exp\left(-\frac{1}{2\sigma_{k-1}^2}\|\mathbf{Ax} - \mathbf{b}\|^2 - \frac{\delta_{k-1}}{2}\|\mathbf{Lx}\|^2\right);$$

2. sample $(\delta_k, \sigma_k^2) \sim \pi(\delta, \sigma^2|\mathbf{x}_k, \mathbf{b})$, by first sampling σ_k^2 from

$$\pi(\sigma^2|\mathbf{x}_k, \mathbf{b}) \propto \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{Ax} - \mathbf{b}\|^2 + (a-1)\ln\sigma^2 - \sigma^2/t_0\right)$$

Gibbs Sampling

Given \mathbf{x}_{k-1} , σ_{k-1}^2 , and δ_{k-1} ,

1. sample $\mathbf{x}_k \sim \pi(\mathbf{x}|\delta_{k-1}, \sigma_{k-1}^2, \mathbf{b})$ where

$$\pi(\mathbf{x}|\delta_{k-1}, \sigma_{k-1}^2, \mathbf{b}) \propto \exp\left(-\frac{1}{2\sigma_{k-1}^2}\|\mathbf{Ax} - \mathbf{b}\|^2 - \frac{\delta_{k-1}}{2}\|\mathbf{Lx}\|^2\right);$$

2. sample $(\delta_k, \sigma_k^2) \sim \pi(\delta, \sigma^2|\mathbf{x}_k, \mathbf{b})$, by first sampling σ_k^2 from

$$\pi(\sigma^2|\mathbf{x}_k, \mathbf{b}) \propto \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{Ax} - \mathbf{b}\|^2 + (a-1)\ln\sigma^2 - \sigma^2/t_0\right)$$

and then sampling δ_k from

$$\pi(\delta|\mathbf{x}_k, \mathbf{b}) \propto \exp\left(-\frac{\delta}{2}\|\mathbf{Lx}_k\|^2 - \frac{1}{2}(\delta/a_0)^2 + \frac{n+2}{2}\ln\delta\right)$$

NOTE: See `OneDBLurSamp.m`. Implement the conjugate prior for sampling from σ^2 (see `OneDBLurSampConj.m`).