Bayesian Inversion

Bangti Jin (UCL-CS, b.jin@ucl.ac.uk)

Course “Inverse Problems & Imaging”

Outline

finite-dimensional inverse problem

$$F(X) = Y,$$

- X, Y: the unknown coefficient and the noisy data
- $F : \mathbb{R}^m \mapsto \mathbb{R}^n$: forward map ... convolution, Radon transform ...
- efficient algorithms for finding a Tikhonov minimizer

$$\frac{1}{2}\|F(x) - y\|^2 + \alpha\psi(x)$$

- Question: How plausible is the Tikhonov minimizer ?

Motivation

- $\Rightarrow$ tools for assessing the reliability of the inverse solution
  - Bayesian inference is one principled framework
  - probabilistic numerics ?
- basic idea: regard the unknown X and the data Y as *random variables*, and encode the prior knowledge in a probability distribution + Bayes rule.

starting point: Bayes' formula, i.e., for two random variables X and Y the conditional probability of X given Y is given by

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)},$$

- x, y: realization of X, Y
- $p_{Y|X}(y|x)$: **likelihood function** — building block I
  information in the data y (noise statistics of y)
- $p_X(x)$: **prior distribution** — building block II
  a prior knowledge (before collecting the data)

the unnormalized posteriori $p(x, y)$ defined by

$$p(x, y) = p_{Y|X}(y|x)p_X(x),$$

and shall often write

$$p_{X|Y}(x|y) \propto p(x, y)$$

the posteriori $p_{X|Y}(x|y)$ up to a multiplicative constant

$p_{X|Y}(x|y)$ **holds the full information about the inverse problem**

$\Rightarrow$ **calibrating the uncertainties of the inverse solutions.**

likelihood function $p_{Y|X}(y|x) \Leftarrow$ the noise statistics

- all sources of errors are lumped into data noise
- a careful modeling of all errors in y is essential for extracting useful information

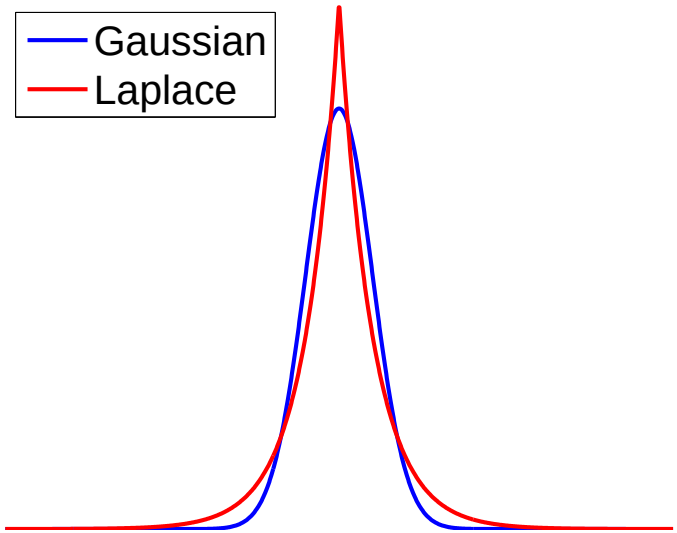
The most popular noise model is the additive Gaussian model

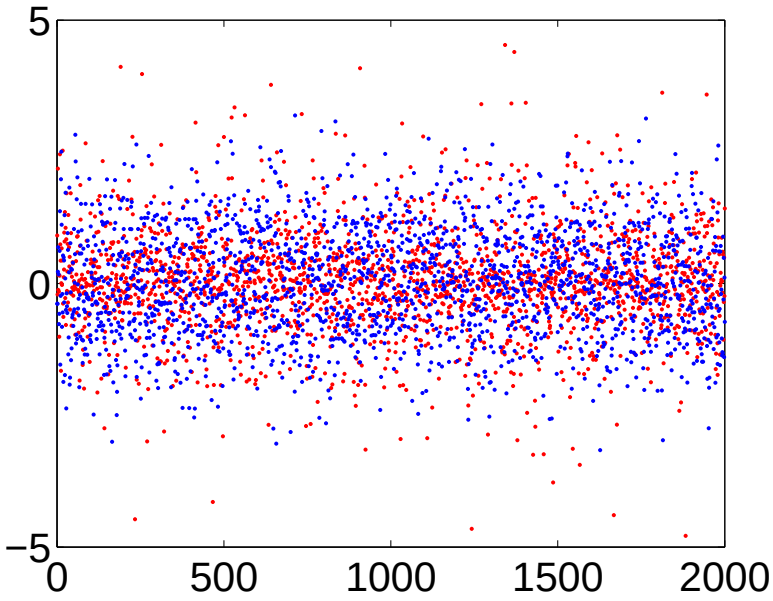
$$y = y^\dagger + \xi,$$

- $\xi \in \mathbb{R}^n$ is a realization of i.i.d. Gaussian r.v. $N(0, \sigma^2)$
- ξ is independent of the true data $y^\dagger$ (and hence x) $\Rightarrow$

$$p_{Y|X}(y|x) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\|F(x)-y\|^2}.$$

other noise models are also possible: Laplace, Poisson, Gamma ...





Gaussian distribution vs. Laplace distribution

The prior $p_X(x)$: the prior knowledge about the solution x in a probabilistic manner.

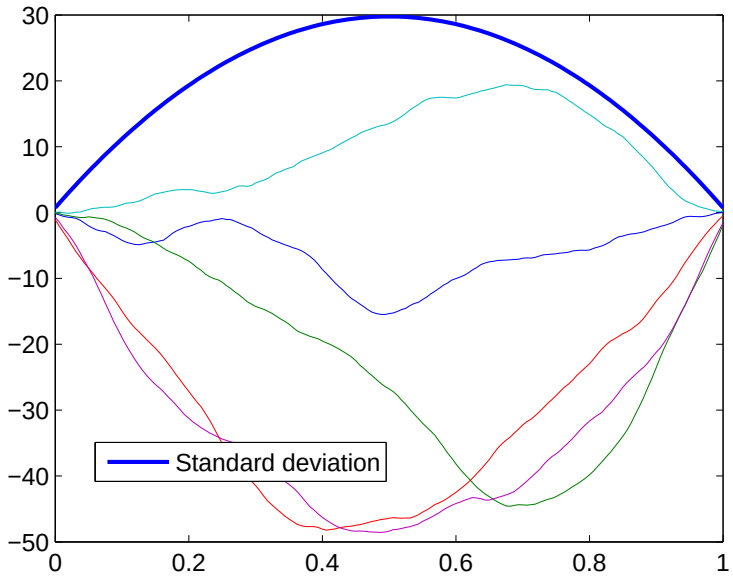
- prior knowledge: expert opinion, historical investigations, statistical studies and anatomical knowledge etc.
- the prior plays the role of regularization in a stochastic setting prior modeling stays at the heart of Bayesian modeling

but it is an art ...

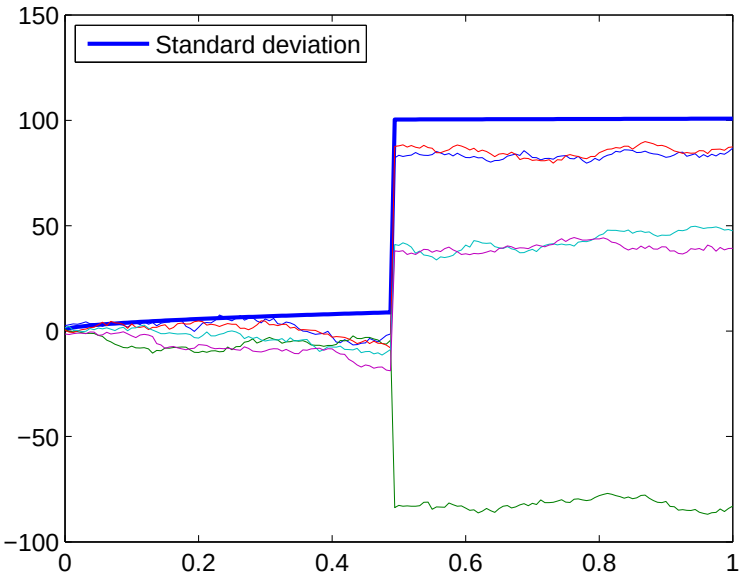
- One very versatile prior model is Markov random field

$$p_X(x) \propto e^{-\lambda\psi(x)},$$

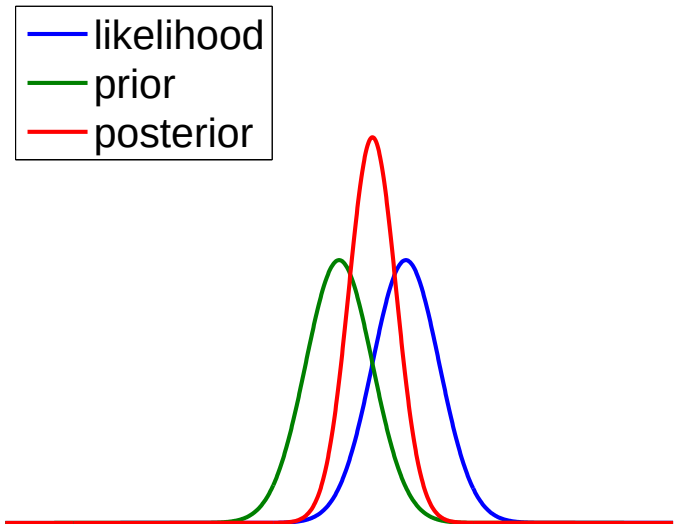
  where $\psi(x)$ is a potential function – penalty term
- $\lambda > 0$ is a scale parameter – regularization parameter



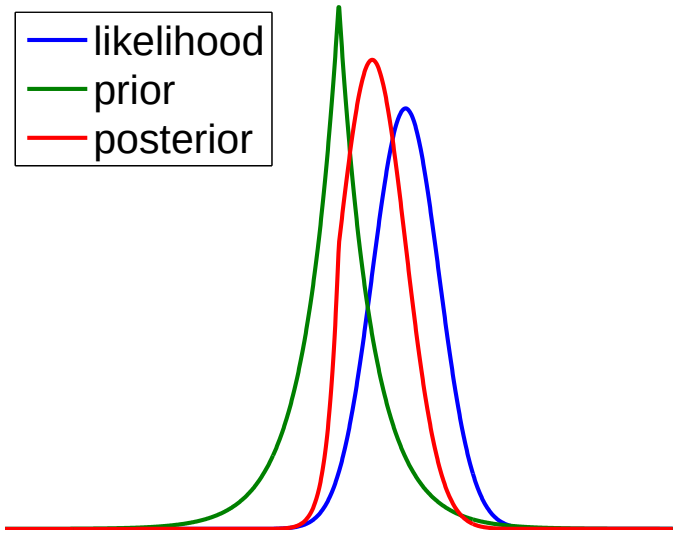
(a) smoothness



(b) total variation



Gaussian likelihood



Laplace likelihood

likelihood $p_{Y|X}(y|x)$ and the prior $p_X(x)$ may contain unknowns e.g.,

$$p_{Y|X}(y|x) = p_{Y|X,\Upsilon}(y|x,\tau) \quad \text{and} \quad p_X(x) = p_{X|\Lambda}(x|\lambda)$$

- τ,λ: precision (inverse variance) and the scale parameter
- commonly known as hyperparameters
- *Hierarchical* Bayesian provides one approach for their choices

hierarchical Bayesian modeling

- view λ and τ as random variables with their own priors
- determine them from the data y
- convenient choice: conjugate distribution

For both λ and τ, the conjugate distribution is a Gamma distribution:

$$p_\Lambda(\lambda) = G(\lambda; a_0, b_0) = \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0 - 1} e^{-b_0 \lambda},$$

$$p_\Upsilon(\tau) = G(\tau; a_1, b_1) = \frac{b_1^{a_1}}{\Gamma(a_1)} \tau^{a_1 - 1} e^{-b_1 \tau}.$$

- (a_0, b_0) and (a_1, b_1) determine the range of λ and τ
- noninformative prior is often adopted: $(a_0, b_0) \approx (1, 0)$

posterior distribution $p_{X,\Lambda,\Upsilon|Y}(x, \lambda, \tau|y)$

$$p_{X,\Lambda,\Upsilon|Y}(x, \lambda, \tau|y) \propto p_{Y|X,\Upsilon}(y|x, \tau) p_{X|\Lambda}(x|\lambda) p_\Lambda(\lambda) p_\Upsilon(\tau)$$

Example: Gaussian noise model + Laplace prior

$$p_{Y|X,\Upsilon}(y|x,\tau) \propto \tau^{-\frac{n}{2}} e^{-\frac{\tau}{2}\|F(x)-y\|^2},$$
$$p_{X|\Lambda}(x|\lambda) \propto \lambda^m e^{-\lambda\|x\|_1}$$

fixed λ and τ + maximum a posteriori estimate $x_{\text{map}} \Rightarrow$

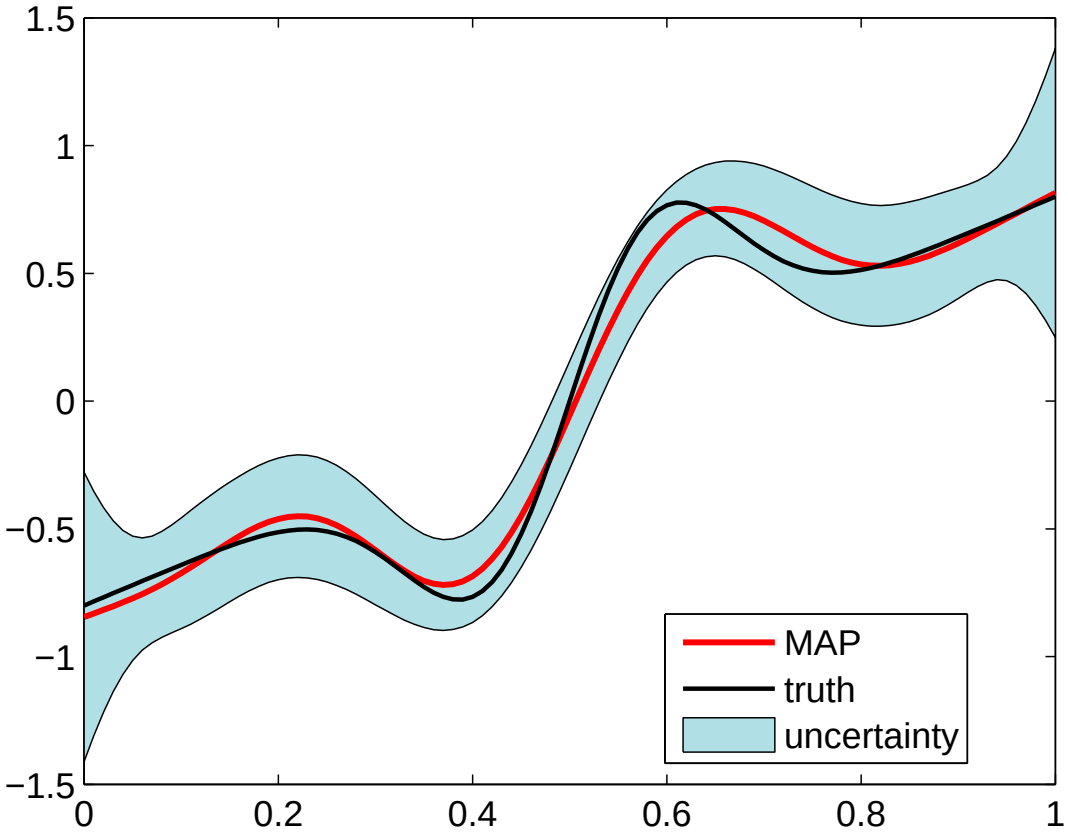
$$\begin{aligned} x_{\text{map}} &= \arg\max_x p_{X,\Lambda,\Upsilon|Y}(x,\lambda,\tau|y) \\ &= \arg\min_x \left\{ \tfrac{\tau}{2}\|F(x)-y\|^2 + \lambda\|x\|_1 \right\} \end{aligned}$$

the functional in the curly bracket is

$$\tfrac{1}{2}\|F(x)-y\|^2 + \lambda\tau^{-1}\|x\|_1$$

Tikhonov regularization + sparsity constraint, with $\alpha = \lambda\tau^{-1}$

A Tikhonov minimizer is an MAP estimate of some Bayesian formulation.


1.5
1
0.5
0
−0.5
−1
−1.5
0
0.2
0.4
0.6
0.8
1
MAP
truth
uncertainty

unknown parameters λ and $\tau \Rightarrow$ hierarchical model
conjugate prior on λ and $\tau \Rightarrow$ posterior distribution

$$p_{X,\Lambda,\Upsilon|Y}(x,\lambda,\tau|y) \propto \tau^{\frac{n}{2}+a_1-1} e^{-\frac{\tau}{2}\|F(x)-y\|^2} \cdot \lambda^{m+a_0-1} e^{-\lambda\|x\|_1} \cdot e^{-b_1\tau} \cdot e^{b_0\lambda}.$$

ways of exploring the posterior $p_{X,\Lambda,\Upsilon|Y}(x,\lambda,\tau|y)$

- joint MAP estimate $(x,\lambda,\tau)_{\text{map}}$, i.e.,

$$(x,\lambda,\tau)_{\text{map}} = \arg\min_{x,\lambda,\tau} J(x,\lambda,\tau),$$

where

$$J(x,\lambda,\tau) = \tfrac{\tau}{2}\|F(x)-y\|^2 + \lambda\|x\|_1 - \tilde{a}_0 \ln\lambda + b_0\lambda - \tilde{a}_1 \ln\tau + b_1\tau.$$

augmented Tikhonov regularization for sparsity constraint

augmented Tikhonov regularization

$$J(x, \lambda, \tau) = \tfrac{\tau}{2}\|F(x) - y\|^2 + \lambda\|x\|_1 - \tilde{a}_0 \ln \lambda + b_0\lambda - \tilde{a}_1 \ln \tau + b_1\tau.$$

- the first two terms recover Tikhonov regularization
- the rest automatically determines the regularization parameter.
- It remains a point estimate and ignores the statistical fluctuations
  $\Rightarrow$ full Bayesian treatment

distinct features

- $p_{X,\Lambda,\Upsilon|Y}(x,\lambda,\tau|y)$ is a probability distribution, and is an ensemble of solutions consistent with y (to various extent)

$$\mu = \int x p_{X|Y}(x|y)dx,$$

$$C = \int (x-\mu)(x-\mu)^{\mathrm{t}} p_{X|Y}(x|y)dx.$$

- the crucial role of proper statistical modeling in designing useful regularization formulations for practical problems.
- hierarchical modeling provides a flexible regularization, *partially* resolving the issue of choosing a regularization parameter.

posteriori $p(x)$ lives in a high-dimensional space $\Rightarrow$ noninformative
$\Rightarrow$ compute *summarizing* statistics, e.g., mean μ and covariance C

$$\mu = \int x p(x) dx \quad \text{and} \quad C = \int (x-\mu)(x-\mu)^{\mathrm{t}} p(x) dx.$$

very high-dimensional integrals, and quadrature rules are inefficient
Ex: $m = 100$, 2 points/dir $\Rightarrow 2^{100} \approx 1.27 \times 10^{30}$ points
more efficient approach

- Monte Carlo methods, especially Markov chain Monte Carlo

Monte Carlo simulation

- draw a large set of i.i.d. samples $\{x^{(i)}\}_{i=1}^{N}$ from the target distribution $p(x)$
- approximate the expectation $E_p[f]$ of any function $f : \mathbb{R}^m \to \mathbb{R}$ by the sample mean $E_N[f]$

$$E_N[f] \equiv \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)}) \to E_p[f] = \int f(x)p(x)dx \quad \text{as } N \to \infty.$$

- the Monte Carlo integration error $e_N[f]$ by

$$e_N[f] = E_p[f] - E_N[f] \approx \mathrm{Var}_p[f]^{\frac{1}{2}} N^{-1/2}\nu,$$

$\nu \sim N(0,1)$

  - the error $e_N[f]$ is $O(N^{-1/2})$
  - with a constant $\sim$ the variance of the integrand f
  - the estimate is independent of the dimensionality m

Generating a large set of i.i.d. samples from an implicit and high-dimensional joint distribution is highly nontrivial.

- nonlinear inverse problems and nongaussian models

Markov chain Monte Carlo: general-purposed approach for exploring posteriori $p(x)$

- basic idea: given $p(x)$, construct an aperiodic and irreducible Markov chain such that its stationary distribution is $p(x)$.
- By running the chain for sufficiently long, simulated values from the chain are dependent samples from $p(x)$, and used for computing summarizing statistics.
- Metropolis: simulating energy levels of atoms in a crystalline structure (1950s)
- Hastings: statistical problems (1970s)
- in inverse problems: 1990s ...

The Metropolis-Hastings algorithm

1: Initialize $x^{(0)}$ and set N;
2: **for** $i = 0 : N$ **do**
3: sample $u \sim U(0,1)$;
4: sample $x^{(*)} \sim q(x^{(i)}, x^{(*)})$
5: **if** $u < \alpha(x^{(i)}, x^{(*)})$ **then**
6: $x^{(i+1)} = x^{(*)}$;
7: **else**
8: $x^{(i+1)} = x^{(i)}$;
9: **end if**
10: **end for**

- the uniform distribution $U(0,1)$
- $p(x)$: the target distribution
- $q(x, x')$ is an easy-to-sample proposal distribution

Having generated a new state x' from $q(x, x')$, accept it as the new state of the chain with probability $\alpha(x, x')$ given by

$$\alpha(x, x') = \min\left\{1, \frac{p(x')q(x', x)}{p(x)q(x, x')}\right\}.$$

However, if we reject x', then the chain remains in the current state x.

- $p(x)$ enters only through α via the ratio $p(x')/p(x)$
  $\Rightarrow$ require $p(x)$ only up to a multipl. constant
- if q is symmetric, i.e., $q(x, x') = q(x', x)$, $\alpha(x, x')$ reduces to

$$\alpha(x, x') = \min\left\{1, \frac{p(x')}{p(x)}\right\}.$$

The Metropolis-Hastings algorithm guarantees that the Markov chain converges to $p(x)$ for any reasonable $q(x, x')$. There are many possible choices for $q(x, x')$, the defining ingredient of the algorithm.

random walker sampler

- If $q(x, x') = f(x' - x)$ for p.d.f. f, then $x^{(*)} = x^{(i)} + \xi$, $\xi \sim f$
- Markov chain is driven by a random walk
- f: uniform, multivariate normal or t-distribution
- With i.i.d. Gaussian distribution $N(0, \sigma^2)$,

$$x_j^{(*)} = x_j^{(i)} + \xi, \quad \text{with } \xi \sim N(\xi; 0, \sigma^2)$$

σ^2 controls the size of the random walks, and should be carefully tuned to improve the MCMC convergence.

Heuristically, the optimal acceptance ratio should be around 0.25 for some model problems.

- the first samples are poor approximations as samples from $p(x)$
- discards these initial samples (burning-in period)
- assess the convergence of the MCMC chains

- If the state space is high dim., it is difficult to update the entire vector x in one single step since $\alpha(x, x')$ is often very small.
- to update a part of the components of x each time and to implement an updating cycle inside each step
- The extreme case is the Gibbs sampler Geman-Geman, 1984

which updates a single component each time.
to update x_i of x, proposal $q(x, x')$: the full conditional

$$q(x, x') = \begin{cases} p(x_i'|x_{-i}) & x'_{-i} = x_{-i}, \\ 0 & \text{otherwise}, \end{cases}$$

where x_{-i} denotes $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m)^t$
these proposals are automatically accepted

example: Gibbs sampler for Gaussian noise + smoothness prior $p(\lambda) \propto \lambda^{a_0-1} e^{-b_0\lambda}$ on the scale parameter λ, i.e., posteriori

$$p(x,\lambda) \propto e^{-\frac{\tau}{2}\|Ax-y\|^2} \cdot \lambda^{\frac{m}{2}} e^{-\frac{\lambda}{2}x^t W x} \lambda^{a_0-1} e^{-b_0\lambda},$$

where the matrix W encodes the local interaction structure

full conditional $p(x_i|x_{-i},\lambda)$

$$p(x_i|x_{-i},\lambda) \sim N(\mu_i,\sigma_i^2), \quad \mu_i = \frac{b_i}{2a_i}, \quad \sigma_i = \frac{1}{\sqrt{a_i}},$$

with a_i and b_i given by

$$a_i = \tau \sum_{j=1}^{n} A_{ji}^2 + \lambda W_{ii} \quad \text{and} \quad b_i = 2\tau \sum_{j=1}^{n} \mu_j A_{ji} - \lambda \mu_p,$$

and $\mu_j = y_j - \sum_{k\neq i} A_{jk}x_k$ and $\mu_p = \sum_{j\neq i} W_{ji}x_j + \sum_{k\neq i} W_{ik}x_k$. Lastly, we deduce the full conditional for λ:

$$p(\lambda|x) \sim G\left(\lambda; \tfrac{m}{2} + a_0, \tfrac{1}{2}x^{\mathrm{t}}Wx + \beta_0\right).$$