# Summary for Bayesian Smoothing Spline Methods

Ziang Zhang

## 1 Smoothing Spline Problem

Consider a data set $\{y_i, x_i, i \in [n]\}$, and a nonparametric model $y_i = g(x_i) + \epsilon_i$ where $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$ and $x_i \in [a, b]$, then the smoothing spline aims to solve the following problem:

$$\arg\min_{g \in C^2} \bigg\{ \sum_i \Big( y_i - g(x_i) \Big)^2 + \lambda \int_a^b g''(x)^2 dx \bigg\}.$$

If $\{x_i, i \in [n]\}$ is a set of distinct points, then the solution to the above problem is well known to be a cubic spline with knots being $\{x_i, i \in [n]\}$. The same minimization problem can be generalized to other models, by replacing the sum of squares with negative log likelihood function, a procedure called *penalized likelihood.*

## 2 ARIMA Approach to Smoothing Spline

Let $\mathbf{g} := (g(x_1), ..., g(x_n))^T$ and $\mathbf{y} := (y_1, ..., y_n)^T$, then the minimization problem above can be vectorized as:

$$(\mathbf{y} - \mathbf{g})^T(\mathbf{y} - \mathbf{g}) + \mathbf{g}^T K \mathbf{g},$$

where $K$ is a positive semi-definite matrix defined as $\frac{1}{\sigma_s^2} Q^T R^{-1} Q$ and $\sigma_s^2 = \frac{1}{\lambda}$. For simplicity, assume all the knots are equally spaced with unit distance, then the $(n-2) \times n$ matrix $Q$ will be the second order difference matrix defined as:

$$Q = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 \\ 0 & \vdots & & & \ddots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 1 & -2 & 1 \end{bmatrix} \tag{1}$$

The $(n-2) \times (n-2)$ matrix R is a strictly positive definite matrix that can be computed as:

$$R = \begin{bmatrix} \frac{2}{3} & \frac{1}{6} & 0 & 0 & 0 & \cdots & 0 \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 0 & \cdots & 0 \\ & & & & & \ddots & \\ 0 & 0 & 0 & \cdots & 0 & \frac{1}{6} & \frac{2}{3} \end{bmatrix} \tag{2}$$

Note that the $Q$ matrix can be viewed as a second order difference operator, that functions on the parameter vector $\mathbf{g}$ to get its second order difference vector $\boldsymbol{\gamma} := Q\boldsymbol{g}$. Hence $\boldsymbol{g}^T K \boldsymbol{g}$ can be equivalently computed as $\boldsymbol{\gamma}^T R^{-1} \boldsymbol{\gamma}$. In other words, adding the penalty term is equivalent to assigning a prior to the second order difference vector $\boldsymbol{\gamma} \sim N(0, \sigma_s^2 R)$.

Let $\boldsymbol{\gamma} := (\gamma_3, ..., \gamma_n)^T \in \mathbb{R}^{n-2}$, with $\gamma_i := g(x_i) - 2g(x_{i-1}) + g(x_{i-2})$, then we can notice that the covariance matrix $R$ of $\boldsymbol{\gamma}$ has the same structure as a MA(1) model. Let $\boldsymbol{\xi} := (\xi_2, ..., \xi_n)^T \sim N(0, \sigma_s^2 I)$, and define $\gamma_i = \theta\xi_{i-1} + \xi_i$, then we can write $\boldsymbol{\gamma} = \Theta\boldsymbol{\xi}$, where the $(n-2) \times (n-1)$ coefficient matrix $\Theta$ is defined as:

$$\Theta = \begin{bmatrix} \theta & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \theta & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \theta & 1 & 0 & \cdots & 0 \\ & & & & & \ddots & \\ 0 & 0 & 0 & \cdots & 0 & \theta & 1 \end{bmatrix} \tag{3}$$

To solve the coefficient $\theta$, we want to find the value such that $\Theta\Theta^T = R$. This gives $\theta = 2 \pm \sqrt{3}$, and we will select $\theta = 2 - \sqrt{3} < 1$ so that the process is invertible. Putting such MA(1) prior on the second order difference vector implies a ARIMA(0,2,1) prior will be assigned to the parameter vector $\boldsymbol{g}$ [Brown and De Jong, 2001].

# 3 SDE approach for smoothing spline

The ARIMA approach in the previous section provides a Bayesian interpretation of the smoothing spline problem. When $x_i$ is discretely observed, this ARIMA approach amounts to assign an MA(1) prior to the second order differences (second derivatives), and diffuse priors to the intercept and slope. Although the same approach can be used to deal with unequally spaced $x_i$ by cutting $\{x_i, i \in [n]\}$ into finer equally spaced resolution, the ARIMA prior does not generalize to continuously observed $\{x_i, i \in [n]\}$. For example, if in the refined resolution, the space between $x_i$ and $x_{i+1}$ is $d$, the prior distribution of $g(x_i + \frac{d}{2})$ will not be clearly defined.

An alternative method is to assign a prior on the whole unknown function $g(.)$, instead of its evaluation vector $\boldsymbol{g} := (g(x_1), .., g(x_n))^T$. This can be done through the use of SDE based prior on the function space. Let $W(t)$ denote Wiener's process (Brownian motion), a SDE based prior is assigned to $g(t)$ in the following way:

$$\frac{d^2 g(t)}{dt^2} = \sigma_s \frac{dW(t)}{dt}.$$

The derivative of $W(t)$ does not exist in ordinary definition, but can be defined as a generalized function, the *white noise* process. Such SDE will not be proper without extra condition on the intercept $g(0)$ and the slope $g'(0)$. When $g(0)$ and $g'(0)$ are fixed to be zero, this SDE is equivalent to use a *second folded Wiener's process* on $g(t)$. If $g(0)$ and $g'(0)$ are given diffuse Gaussian priors, the posterior mean of $\boldsymbol{g}$ will be the minimizer of the smoothing spline problem [Wahba, 1978].

## 3.1 Galerkin Approximation to SDE

The direct use of such SDE prior on $g(x)$ can yield posteriors for the parameter vector $\boldsymbol{g}$, but the posterior of $g(s)$ for $s \notin \{x_i, i \in [n]\}$ cannot be directly obtained from $P(\boldsymbol{g}|y)$. To reduce the problem of inference for infinite dimension parameter $g(.)$ to finite dimensional inference, it is convenient to consider the use of numerical discretization to the SDE.

Let $\{\phi_i(x), i \in [n]\}$ be the set of test functions, and $\{\varphi_i(x), i \in [n]\}$ be the set of basis functions. A weak solution $\tilde{g}(x)$ to the SDE is defined as

$$\tilde{g}(x) = \sum_{i=1}^{n} w_i \varphi_i(x),$$

such that

$$\langle \frac{d^2 \tilde{g}(t)}{dt^2}, \phi_i(t) \rangle \overset{d}{=} \langle \sigma_s \frac{dW(t)}{dt}, \phi_i(t) \rangle,$$

for any test function $\phi_i$. The coefficient vector $\boldsymbol{w} := (w_1, ..., w_n)^T$ will become the parameters to infer.

If one considers to use the same set of functions as both the test functions and the basis functions, then the approximation is called a *Galerkin Approximation.* In Lindgren and Rue [2008], the authors consider the use of linear B splines as both the test functions and basis functions. Then, the inner products $\langle \frac{d^2 \tilde{g}(t)}{dt^2}, \phi_i(t) \rangle_{i=1}^n$ can be vectorized as:

$$\langle \frac{d^2 \tilde{g}(t)}{dt^2}, \phi_i(t) \rangle_{i=1}^n = H \boldsymbol{w},$$

when knots have unit spacing, the $n \times n$ matrix $H$ is computed as:

$$H = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 \\ 0 & \vdots & & & \ddots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \tag{4}$$

The second inner products $\langle \frac{dW(t)}{dt}, \phi_i(t) \rangle_{i=1}^n$ will have Gaussian distribution with zero mean vector and $n \times n$ covariance matrix $B_{ij} = [\langle \phi_i, \phi_j \rangle]$:

$$B = \begin{bmatrix} \frac{1}{3} & \frac{1}{6} & 0 & 0 & 0 & \cdots & 0 & 0 \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 0 & 0 & \cdots & 0 & 0 \\ 0 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 0 & \cdots & 0 & 0 \\ 0 & 0 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & \cdots & 0 & 0 \\ & & & & \ddots & & & \\ 0 & 0 & 0 & \cdots & \frac{1}{6} & \frac{2}{3} & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & \frac{1}{6} & \frac{2}{3} & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & \frac{1}{6} & \frac{1}{3} \end{bmatrix} \tag{5}$$

Therefore, combine the two vectorized equations, we know that $\boldsymbol{w}$ will prior distribution being Gaussian with zero mean and precision matrix $\frac{1}{\sigma_s^2} H^T B^{-1} H$. As noted in Lindgren and Rue [2008], this Galerkin *approximation* is actually exact, except for the boundaries. This can also be found from the fact that $H$ matrix after removing its first and last row is the $Q$ matrix from Brown and De Jong [2001], and the $B$ matrix after removing its first and last rows and columns is the $R$ matrix from Brown and De Jong [2001].

### 3.1.1 Diagonal Approximation

The use of Galerkin approximation as in Lindgren and Rue [2008] will result in $\boldsymbol{w}$ having precision matrix $H^T B^{-1} H$. Since $B$ matrix is tridiagonal, this precision matrix will be dense, and hence not applicable in INLA-typed inference method [Rue et al., 2009]. To handle this dense precision matrix problem, Lindgren and Rue [2008] proposes the use of a diagonal approximation $A$, to the covariance matrix $B$. This amounts to use uncorrelated noises to approximate the noises $\langle \frac{dW(t)}{dt}, \phi_i(t) \rangle_{i=1}^n$ which approximately have a MA(1) covariance structure. The resulting (approximate) Galerkin approximation is called the second order random walk prior (RW2).

Specifically, the covariance matrix $B$ is approximated by the diagonal matrix A defined as following(assuming unit spacing between knots):

$$A = \begin{bmatrix} \frac{1}{2} & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ & & & & \ddots & & & \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \frac{1}{2} \end{bmatrix} \tag{6}$$

### 3.1.2 Write ARIMA Model as A Weak Solution

When knots are equally spaced, the ARIMA(0,2,1) model (therefore, the exact solution) can be written as a particular weak solution to the above SDE, with slightly different basis and test functions. The RW2 model uses the same set of linear spline functions to be both the basis functions and test functions. This type of discretization using same set of functions to be both basis function and test function is called *Galerkin Approximation*, and the precision matrix of $\boldsymbol{w}$ (hence $\boldsymbol{g}$) will be $H^T B^{-1} H$. On the other hand, if the set of basis functions is not the same as the set of test functions, then the discretization method to the SDE is called *Petrov Galerkin* approximation.

If we keep the same set of linear spline functions to be the $n$ basis functions, but remove $\phi_1$ and $\phi_n$ from the set of test functions, then the same discretization procedure will yield $Q^T R^{-1} Q$ as the precision matrix, the same precision matrix of the ARIMA model. In other words, this *Petrov Galerkin* procedure will have a weak solution that is also the exact solution to the SDE, even at the boundaries.

An example illustrating the above procedure with five equally spaced knots is shown in Figure 1, where the linear spline functions to be removed from the set of test functions are in red.

## 4 Practical Comparison between the three approaches

In this section, we will implement semi-parametric smoothing with each of these three approaches: RW2 with Diagonal Approximation (approach 1), RW2 without Diagonal Approximation (approach 2) and the exact ARIMA(0,2,1) method (approach 3). Through simulations under different settings, we aim to compare both

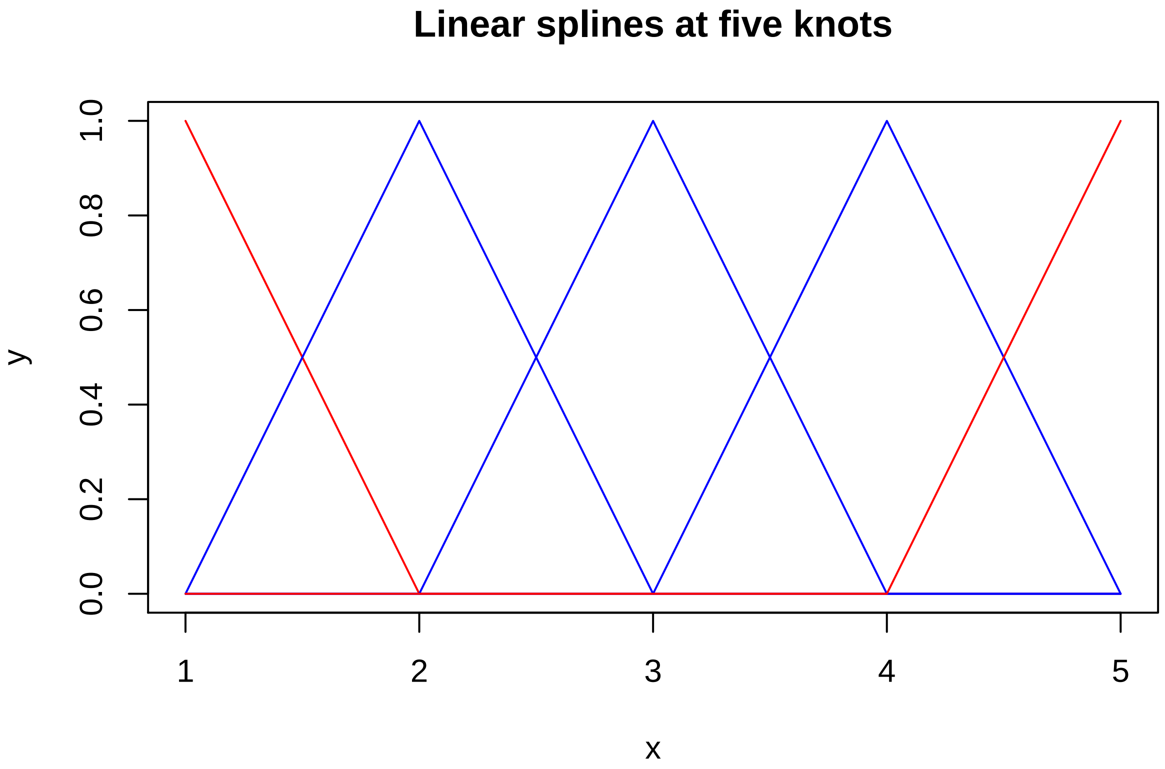

Figure 1: The set of linear splines which are used as basis/test functions in RW2. The two functions that will be removed from the test functions to construct ARIMA are colored in red.

the accuracy and the computational efficiency of these approaches.

## 4.1 Scenario 1: Ordinary Gaussian regression

In this example, we consider $n = 500$ observations generated from the following model:

$$\begin{aligned} y_i &= g(x_i) + \epsilon_i \\ \epsilon_i &\sim N(0, \sigma_\epsilon = 1) \\ g(x) &= 5\sin(0.5x) \end{aligned} \tag{7}$$

The covariate values (knots) $\{x_i, i \in [n]\}$ are equally spaced from 0.1 to 50.

### 4.1.1 If true variance is known to be 1:

First, we consider the case where $\sigma_\epsilon$ is known to be 1. In this scenario, we only need to specify a prior for the smoothing parameter $\sigma_s$. For all the methods, we use a PC prior [Simpson et al., 2017] such that $P(\sigma_s > 2) = 0.5$. The smoothing results are summarized in the plot below, with red lines denoted approach 1, blue lines denoted approach 2 and green lines denoted approach 3. Solid line denotes posterior mean, and dashed lines denote the 95 percent interval.

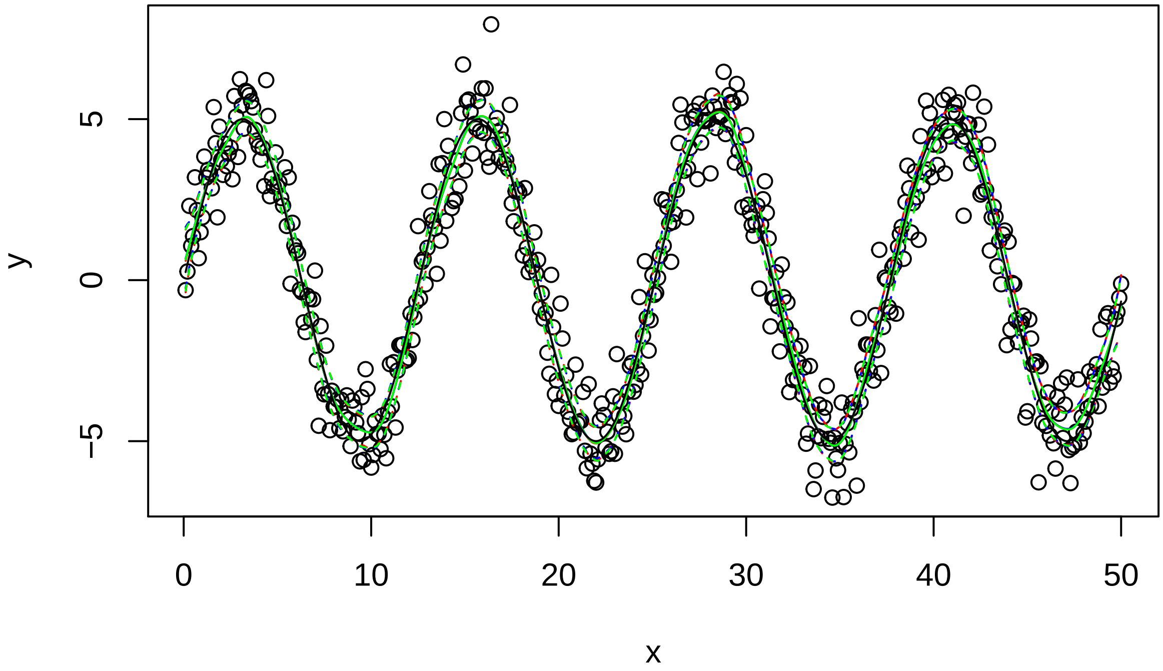

Figure 2: Smoothing result for Guassian response with known variance

The Mean Square Errors (MSE) computed from the three approaches are respectively 0.0342, 0.0336 and 0.0333. Also, based on the Figure 2 above, we can see the three approaches are almost indistinguishable in terms of smoothing result. However, approach 1 only takes 0.169 seconds to run. On the other hand, approach 2 takes 7.123 seconds and approach 3 takes 7.203 seconds. So the computational efficiency are quite different among the three approaches.

### 4.1.2 If true variance is a unknown parameter:

In this case, we no longer assume that the true variance parameter $\sigma_\epsilon$ is known. Instead, we also assign a PC prior on $\sigma_\epsilon$ such that $P(\sigma_\epsilon > 2) = P(\sigma_s > 2) = 0.5$. The results are summarized at Figure 3 below.

According to Figure 3, the three approaches are still very similar in terms of smoothing results when the true variance is unknown. The MSE computed from the three approaches are respectively 0.0337, 0.0348 and 0.0340. The runtimes are 0.427 seconds, 40.343 seconds and 39.485 seconds.

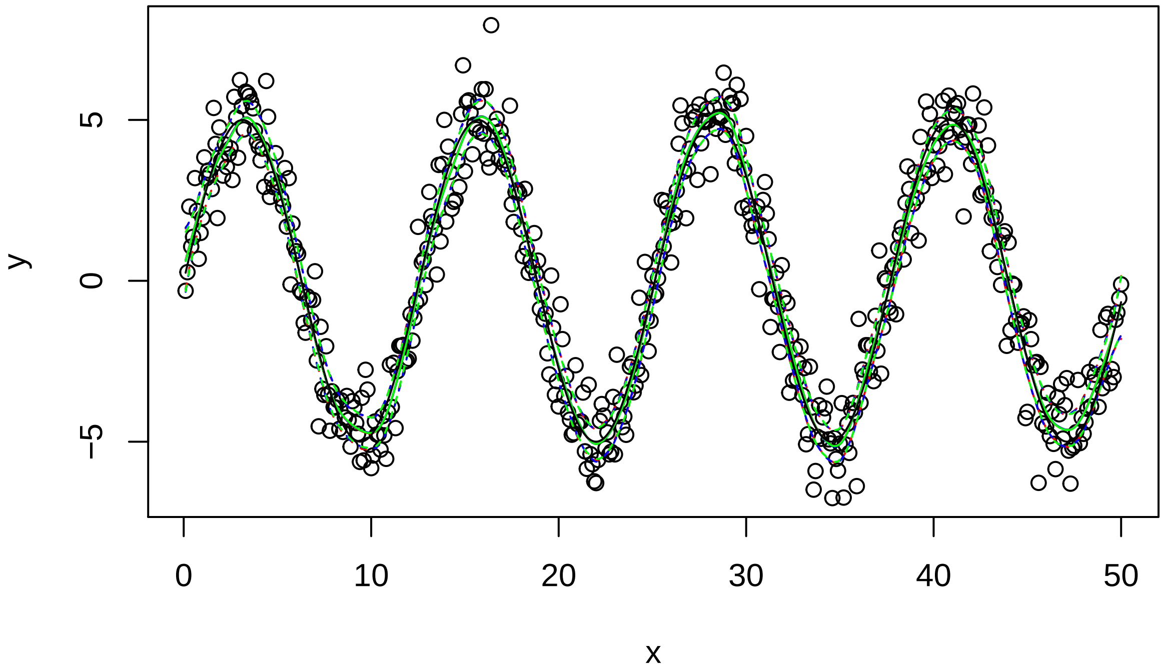

Figure 3: Smoothing result for Guassian response with unknown variance

### 4.2 Scenario 2: Poisson regression

In this example, we consider $n = 500$ data generated from the following Poisson regression model:

$$\begin{aligned} \log(\mu_i) &= g(x_i) \\ g(x) &= \log(5\sin(0.5x) + 6) \end{aligned} \tag{8}$$

The knots placement and prior settings are the same as in the previous example. Results are summarized in Figure 4 with the same colors assignment as before, and the black line denotes the true risk function.

Based on Figure 4, the smoothing results from the three approaches are still very similar. The MSE are respectively 0.0129, 0.0130 and 0.0127 for the three approaches, and the runtimes are respectively 0.220 seconds, 13.467 seconds and 12.651 seconds.

## 5 Conclusion

In summary, the cubic smoothing spline problem can be interpreted as a Bayesian inference problem. The penalty term on the second derivative can be understood as an improper prior on the function space. When covariate is discretely observed with equal spacing, the improper prior corresponds to an ARIMA(0,2,1) model on the evaluation values [Brown and De Jong, 2001]. When covariate is continuously observed with

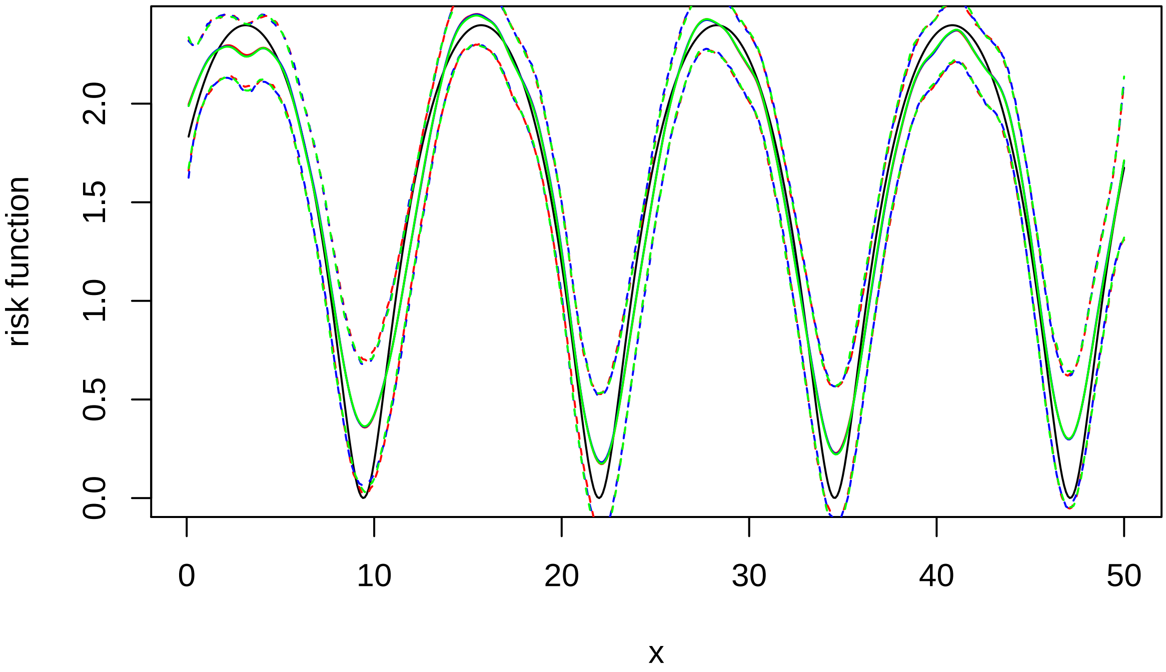

Figure 4: Smoothing result for Poisson response

potentially unequal spacing, the improper prior corresponds to the second folded Wiener's process with *diffuse* initial conditions, the so called SDE-based prior.

For the SDE-based prior, it is convenient to use discretization method in terms of basis functions and test functions, such as Galerkin Approximation method. To simplify the computations for the dense precision matrix of the Galerkin Approximation, the RW2 method [Lindgren and Rue, 2008] is developed by approximating the tri-diagonal covariance matrix with its diagonal approximation.

Based on the simulation studies we have performed in this summary, these three methods have very similar performance in practice. However, after applying the diagonal approximation to the covariance matrix, the RW2 method has much better computational efficiency compared to the original Galerkin approximation method, or the exact method with ARIMA(0,2,1) model.

# References

Patrick E Brown and Piet De Jong. Nonparametric smoothing using state space techniques. *Canadian Journal of Statistics*, 29(1):37–50, 2001.

Finn Lindgren and Håvard Rue. On the second-order random walk model for irregular locations. *Scandinavian journal of statistics*, 35(4):691–700, 2008.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

Daniel Simpson, Håvard Rue, Andrea Riebler, Thiago G Martins, and Sigrunn H Sørbye. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, 32(1): 1–28, 2017.

Grace Wahba. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(3):364–372, 1978.