

---

# Large-Scale Cox Process Inference using Variational Fourier Features

---

## Supplementary Material

### Contents

<b>A Example Rate Functions</b>	<b>2</b>
<b>B Derivation of Objective</b>	<b>2</b>
<b>C Derivation of <math>\Psi</math> Matrix for Fourier Features</b>	<b>3</b>
C.1 Notation . . . . .	3
C.2 One-Dimensional Kernel . . . . .	3
C.3 Sum Kernel . . . . .	6
C.4 Product Kernel . . . . .	6
<b>D Supplementary Figures</b>	<b>7</b>

## A. Example Rate Functions

The three example functions in Figure 2 are

$$\lambda_1(s) = 2 \exp(-s/15) + \exp(-((s - 25)/10)^2), \quad (1)$$

$$\lambda_2(s) = 5 \sin(s^2) + 6, \quad (2)$$

$$\lambda_3(s) = \text{piecewise linear}, \quad (3)$$

where  $\lambda_3$  goes through the following points:  $(0, 2), (25, 3), (50, 1), (75, 2.5), (100, 3)$ . The domains are  $\mathcal{T}_1 = [0, 50]$ ,  $\mathcal{T}_2 = [0, 5]$ , and  $\mathcal{T}_3 = [0, 100]$ . The average number of events per draw are  $\bar{\lambda}_1 = 46.92$ ,  $\bar{\lambda}_2 = 33.49$ , and  $\bar{\lambda}_3 = 224.37$ .

## B. Derivation of Objective

We arrive at our objective by considering the KL divergence between the true posterior  $p(f(\cdot) | \mathcal{D})$  and our sparse approximation  $q(f(\cdot))$ :

$$p(f^* | \mathcal{D}) = \iint p(f^* | f, \Theta) p(f, \Theta | \mathcal{D}) d\Theta df \quad (4)$$

$$q(f^*) = \iint p(f^* | u, \Theta) q(u, \Theta) d\Theta df \quad (5)$$

We want to minimize their KL divergence  $\mathcal{K}$ ; writing out the full probability distribution of everything:

$$\mathcal{K} = \text{KL}[q(f^*, f, u, \Theta) || p(f^*, f, u, \Theta | \mathcal{D})] \quad (6)$$

$$= -\mathbb{E}_{q(f^*, f, u, \Theta)} \left[ \log \frac{p(f^* | u, f, \Theta) p(u | f, \Theta) p(f, \Theta | \mathcal{D})}{p(f^* | u, f, \Theta) p(f | u, \Theta) q(u, \Theta)} \right] \quad (7)$$

$$\stackrel{(a)}{=} -\mathbb{E}_{q(f, u, \Theta)} \left[ \log \frac{p(u | f, \Theta) p(\mathcal{D} | f, \Theta) p(f | \Theta) p(\Theta) / p(\mathcal{D})}{p(f | u, \Theta) q(u, \Theta)} \right] \quad (8)$$

$$= -\mathbb{E}_{q(f, u, \Theta)} \left[ \log \frac{p(u | f, \Theta) p(f | \Theta) p(\mathcal{D} | f, \Theta) p(\Theta) / p(\mathcal{D})}{p(f | u, \Theta) q(u, \Theta)} \right] \quad (9)$$

$$\stackrel{(b)}{=} -\mathbb{E}_{q(f, u, \Theta)} \left[ \log \frac{p(f | u, \Theta) p(u | \Theta) p(\mathcal{D} | f, \Theta) p(\Theta) / p(\mathcal{D})}{p(f | u, \Theta) q(u, \Theta)} \right] \quad (10)$$

$$= -\mathbb{E}_{q(f, u, \Theta)} \left[ \log \frac{p(u | \Theta) p(\mathcal{D} | f, \Theta) p(\Theta) / p(\mathcal{D})}{q(u, \Theta)} \right] \quad (11)$$

$$= -\mathbb{E}_{q(f, u, \Theta)} \left[ \log \frac{p(u | \Theta) p(\mathcal{D} | f, \Theta) p(\Theta)}{q(u, \Theta)} \right] + \log p(\mathcal{D}) \quad (12)$$

where we made use of (a)

$$p(f, \Theta | \mathcal{D}) = p(f, \Theta, \mathcal{D}) / p(\mathcal{D}) = p(\mathcal{D} | f, \Theta) p(f | \Theta) p(\Theta) / p(\mathcal{D}) \quad (13)$$

and (b)

$$p(u | f, \Theta) p(f | \Theta) = p(u, f | \Theta) = p(f | u, \Theta) p(u | \Theta) \quad (14)$$

$$\mathcal{K} = -\mathbb{E}_{q(f, u, \Theta)} \log \frac{p(u | \Theta) p(\Theta)}{q(u, \Theta)} - \mathbb{E}_{q(f, u, \Theta)} \log p(\mathcal{D} | f, \Theta) + \log p(\mathcal{D}) \quad (15)$$

$$= -\mathbb{E}_{q(u, \Theta)} \log \frac{p(u, \Theta)}{q(u, \Theta)} - \mathbb{E}_{q(f, \Theta)} \log p(\mathcal{D} | f, \Theta) + \log p(\mathcal{D}) \quad (16)$$

$$= \text{KL}[q(u, \Theta) || p(u, \Theta)] - \mathcal{L}_D + \log p(\mathcal{D}) \quad (17)$$

$$= \log p(\mathcal{D}) - \mathcal{L} \quad (18)$$

So with respect to a variational distribution  $q(u)$ , maximizing the ELBO  $\mathcal{L}$  is equivalent to minimizing the KL divergence  $\mathcal{K}$ .

## C. Derivation of $\Psi$ Matrix for Fourier Features

We want to calculate the matrix

$$\Psi = \int_{\mathcal{T}} \mathbf{k}_u(\mathbf{x})^\top \mathbf{k}_u(\mathbf{x}) d\mathbf{x}, \quad (19)$$

where  $\mathbf{k}_u(\cdot) = \phi(\cdot)$ . We first calculate the elements of  $\Psi$  for a one-dimensional kernel,  $\Psi_{ij} = \int \phi_i(x)\phi_j(x) dx$ .

### C.1. Notation

We use the following short-hand notation:

$$\cos_{\omega_m} = \cos(\omega_m(x - a)), \quad \text{where } \omega_m = \frac{2\pi m}{b - a} \quad (20)$$

where  $m$  is an integer and  $\omega_m$  is the corresponding natural frequency on the interval  $[a, b]$ , and equivalently  $\cos_{\omega_n}$ ,  $\sin_{\omega_m}$ , and  $\sin_{\omega_n}$ . The Fourier features can then be written as

$$\phi_i(x) = \begin{cases} \cos_{\omega_i} & \text{for } 0 \leq i \leq M \\ \sin_{\omega_{i-M}} & \text{for } M < i \leq 2M \end{cases} \quad (21)$$

### C.2. One-Dimensional Kernel

For Fourier features with  $M$  frequencies, the first element in the feature vector is the constant 1, then there are  $M$  cosine functions, then  $M$  sine functions. This leads to six types of integrals:  $1 \times 1$ ,  $1 \times \cos$ ,  $1 \times \sin$ ,  $\cos \times \cos$ ,  $\cos \times \sin$ ,  $\sin \times \sin$ , and we have to distinguish between features with the same frequency ( $m = n$ ) or different frequencies ( $m \neq n$ ).

For a domain  $\mathcal{T} = [c, d]$  in one dimension, we need to evaluate the integrals

$$\Psi_{i,j} = \int_c^d \phi_i(x)\phi_j(x) dx. \quad (22)$$

This can be split into the following cases:

$\phi_0\phi_0 = 1 \times 1 :$

$$\Psi_{0,0} = \int_c^d 1 dx = d - c \quad (23)$$

$\phi_0\phi_i[1 \leq i \leq M] = 1 \times \cos_{\omega_m}, \quad m \geq 1 :$

$$\Psi_{0,i} = \int_c^d \cos_{\omega_m} dx = \frac{b - a}{2\pi m} \sin_{\omega_m} |_c^d \quad (24)$$

$$= \frac{b - a}{2\pi m} (\sin(\omega_m(d - a)) - \sin(\omega_m(c - a))) \quad (25)$$

$\phi_0\phi_i[M < i \leq 2M] = 1 \times \sin_{\omega_m}, \quad m \geq 1 :$

$$\Psi_{0,i} = \int_c^d \sin_{\omega_m} dx = -\frac{b - a}{2\pi m} \cos_{\omega_m} |_c^d \quad (26)$$

$$= -\frac{b - a}{2\pi m} (\cos(\omega_m(d - a)) - \cos(\omega_m(c - a))) \quad (27)$$

$\phi_i\phi_i[M < i \leq 2M] = \sin_{\omega_m} \sin_{\omega_m}, \quad m = n :$

$$\Psi_{i,i} = \int_c^d \sin_{\omega_m} \sin_{\omega_m} dx = \int_c^d \sin_{\omega_m}^2 dx \quad (28)$$

$$= \int_c^d \frac{1}{2}(1 - \cos_{\omega_{2m}}) \quad (29)$$

$$= \frac{1}{2}(d - c) - \frac{b - a}{2 \times 2\pi \times 2m} \sin_{\omega_{2m}} |_c^d \quad (30)$$

$$= \frac{1}{2}(d - c) - \frac{b - a}{8\pi m} \sin_{\omega_{2m}} |_c^d \quad (31)$$

$$\phi_i \phi_j [M < i, j \leq 2M] = \sin_{\omega_m} \sin_{\omega_n}, \quad m \neq n :$$

$$\Psi_{i,j} = \int_c^d \underbrace{\sin_{\omega_m}}_u \underbrace{\sin_{\omega_n}}_{v'} dx \quad (32)$$

$$|u'| = \frac{2\pi m}{b - a} \cos_{\omega_m}, \quad v = -\frac{b - a}{2\pi n} \cos_{\omega_n} \quad (33)$$

$$= -\frac{b - a}{2\pi n} \sin_{\omega_m} \cos_{\omega_n} |_c^d + \frac{m}{n} \int_c^d \underbrace{\cos_{\omega_m}}_u \underbrace{\cos_{\omega_n}}_{v'} dx \quad (34)$$

$$|u'| = \frac{2\pi m}{b - a} (-\sin_{\omega_m}), \quad v = -\frac{b - a}{2\pi n} \sin_{\omega_n} \quad (35)$$

$$= -\frac{b - a}{2\pi n} \sin_{\omega_m} \cos_{\omega_n} |_c^d + \frac{m}{n} \left[ \frac{b - a}{2\pi n} \cos_{\omega_m} \sin_{\omega_n} |_c^d + \frac{m}{n} \int_c^d \sin_{\omega_m} \sin_{\omega_n} dx \right] \quad (36)$$

$$(1 - \frac{m^2}{n^2}) \int \dots = -\frac{b - a}{2\pi n} \sin_{\omega_m} \cos_{\omega_n} |_c^d + \frac{b - a}{2\pi n} \frac{m}{n} \cos_{\omega_m} \sin_{\omega_n} |_c^d \quad (37)$$

$$\int \dots = \frac{n^2}{n^2 - m^2} \frac{b - a}{2\pi} \left( -\frac{1}{n} \sin_{\omega_m} \cos_{\omega_n} |_c^d + \frac{m}{n^2} \cos_{\omega_m} \sin_{\omega_n} |_c^d \right) \quad (38)$$

$$= \frac{1}{n^2 - m^2} \frac{b - a}{2\pi} (m \cos_{\omega_m} \sin_{\omega_n} |_c^d - n \sin_{\omega_m} \cos_{\omega_n} |_c^d) \quad (39)$$

$$\phi_i \phi_j [1 \leq j \leq M, M < i \leq 2M] = \sin_{\omega_m} \cos_{\omega_n}, \quad m \neq n :$$

$$\Psi_{i,j} = \int_c^d \underbrace{\sin_{\omega_m}}_u \underbrace{\cos_{\omega_n}}_{v'} dx \quad (40)$$

$$|u'| = \frac{2\pi m}{b - a} \cos_{\omega_m}, \quad v = \frac{b - a}{2\pi n} \sin_{\omega_n} \quad (41)$$

$$= \frac{b - a}{2\pi n} \sin_{\omega_m} \sin_{\omega_n} |_c^d - \frac{m}{n} \int_c^d \cos_{\omega_m} \sin_{\omega_n} dx \quad (42)$$

$$\int_c^d \underbrace{\cos_{\omega_m}}_u \underbrace{\sin_{\omega_n}}_{v'} \quad (43)$$

$$|u'| = -\frac{2\pi m}{b - a} \sin_{\omega_m}, \quad v = -\frac{b - a}{2\pi n} \cos_{\omega_n} \quad (44)$$

$$= -\frac{b - a}{2\pi n} \cos_{\omega_m} \cos_{\omega_n} |_c^d - \frac{m}{n} \int_c^d \sin_{\omega_m} \cos_{\omega_n} dx \quad (45)$$

$$\Psi_{i,j} = \frac{b - a}{2\pi n} \sin_{\omega_m} \sin_{\omega_n} |_c^d - \frac{m}{n} \left[ -\frac{b - a}{2\pi n} \cos_{\omega_m} \cos_{\omega_n} |_c^d - \frac{m}{n} \int_c^d \sin_{\omega_m} \cos_{\omega_n} dx \right] \quad (46)$$

$$= \frac{b - a}{2\pi n} \sin_{\omega_m} \sin_{\omega_n} |_c^d + \frac{m}{n} \frac{b - a}{2\pi n} \cos_{\omega_m} \cos_{\omega_n} |_c^d + \left( \frac{m}{n} \right)^2 \int_c^d \sin_{\omega_m} \cos_{\omega_n} dx \quad (47)$$

$$(1 - (\frac{m}{n})^2) \int \dots = \frac{b - a}{2\pi n} (\sin_{\omega_m} \sin_{\omega_n} + \frac{m}{n} \cos_{\omega_m} \cos_{\omega_n}) |_c^d \quad (48)$$

$$\int \dots = \frac{1}{1 - \frac{m^2}{n^2}} \frac{b - a}{2\pi n} (\sin_{\omega_m} \sin_{\omega_n} |_c^d + \frac{m}{n} \cos_{\omega_m} \cos_{\omega_n} |_c^d) \quad (49)$$

$$= \frac{1}{n^2 - m^2} \frac{b-a}{2\pi} (n \sin_{\omega_m} \sin_{\omega_n} |_c^d + m \cos_{\omega_m} \cos_{\omega_n} |_c^d) \quad (50)$$

$$\phi_i \phi_j [1 \leq j \leq M, i = j+M] = \sin_{\omega_m} \cos_{\omega_m}, \quad m = n :$$

$$\Psi_{i,i+M} = \int_c^d \sin_{\omega_m} \cos_{\omega_m} dx = \frac{1}{2} \int_c^d \sin_{\omega_{2m}} dx \quad (51)$$

$$= \frac{1}{2} \left( -\frac{b-a}{2\pi \times 2m} \cos_{\omega_{2m}} |_c^d \right) \quad (52)$$

$$= -\frac{b-a}{8\pi m} \cos_{\omega_{2m}} |_c^d \quad (53)$$

$$\phi_i \phi_i [1 \leq i \leq M] = \cos_{\omega_m} \cos_{\omega_m}, \quad m = n :$$

$$\Psi_{i,i} = \int_c^d \cos_{\omega_m} \cos_{\omega_m} dx = \int_c^d \cos_{\omega_m}^2 dx = \int_c^d (1 - \sin_{\omega_m}^2) dx \quad (54)$$

$$= \int_c^d \frac{1}{2} (1 + \cos_{\omega_{2m}}) \quad (55)$$

$$= \frac{1}{2}(d-c) + \frac{b-a}{2 \times 2\pi \times 2m} \sin_{\omega_{2m}} |_c^d \quad (56)$$

$$= \frac{1}{2}(d-c) + \frac{b-a}{8\pi m} \sin_{\omega_{2m}} |_c^d \quad (57)$$

$$\phi_i \phi_j [1 \leq i, j \leq M] = \cos_{\omega_m} \cos_{\omega_n}, \quad m \neq n :$$

$$\Psi_{i,j} = \int_c^d \underbrace{\cos_{\omega_m}}_u \underbrace{\cos_{\omega_n}}_{v'} dx \quad (58)$$

$$|u'| = \frac{2\pi m}{b-a} (-\sin_{\omega_m}), \quad v = \frac{b-a}{2\pi n} \sin_{\omega_n} \quad (59)$$

$$= \frac{b-a}{2\pi n} \cos_{\omega_m} \sin_{\omega_n} |_c^d + \frac{m}{n} \int_c^d \sin_{\omega_m} \sin_{\omega_n} dx \quad (60)$$

$$= \frac{b-a}{2\pi} \left( \frac{1}{n} \cos_{\omega_m} \sin_{\omega_n} |_c^d + \frac{m}{n} \frac{1}{n^2 - m^2} m \cos_{\omega_m} \sin_{\omega_n} |_c^d - \frac{m}{n} \frac{1}{n^2 - m^2} n \sin_{\omega_m} \cos_{\omega_n} |_c^d \right) \quad (61)$$

$$= \frac{b-a}{2\pi} \left( \cos_{\omega_m} \sin_{\omega_n} |_c^d \times \underbrace{\left( \frac{1}{n} + \frac{m}{n} \frac{1}{n^2 - m^2} \right)}_{(n^2 - m^2 + m)/n} - \dots \right) \quad (62)$$

$$= \frac{b-a}{2\pi(n^2 - m^2)} \left( (n^2 - m^2) \frac{1}{n} \cos_{\omega_m} \sin_{\omega_n} + \frac{m}{n} m \cos_{\omega_m} \sin_{\omega_n} - \frac{m}{n} n \sin_{\omega_m} \cos_{\omega_n} \right) |_c^d \quad (63)$$

$$= \frac{b-a}{2\pi(n^2 - m^2)} \left( \cos_{\omega_m} \sin_{\omega_n} \left[ \frac{n^2 - m^2}{n} + \frac{m^2}{n} \right] - m \sin_{\omega_m} \cos_{\omega_n} \right) |_c^d \quad (64)$$

$$= \frac{b-a}{2\pi(n^2 - m^2)} \left( n \cos_{\omega_m} \sin_{\omega_n} - m \sin_{\omega_m} \cos_{\omega_n} \right) |_c^d \quad (65)$$

$$(66)$$

### C.3. Sum Kernel

For a multi-dimensional sum kernel, the resulting  $\Psi$  matrix has block structure. The diagonal blocks  $\Psi_{(i,i)}$  are equivalent to the one-dimensional case  $\Psi^{(i)}$ , except as we integrate over all dimensions, we get a factor  $(d_j - c_j)$  for each dimension  $j \neq i$ . The off-diagonal blocks  $\Psi_{(i,j)}$ ,  $i \neq j$ , correspond to the integrals

$$\int_{c_i}^{d_i} \phi_m(x_i) dx_i \int_{c_j}^{d_j} \phi_n(x_j) dx_j$$

which are the outer product of the first rows of the corresponding one-dimensional  $\Psi$  matrices, except again we get a factor  $(d_k - c_k)$  for each dimension  $k \notin \{i, j\}$ . The calculation can be simplified by constructing diagonal blocks  $\Psi^{(i)} / (d_i - c_i)$  and off-diagonal blocks

$$\frac{\Psi_{1,:}^{(i)} \otimes \Psi_{1,:}^{(j)}}{(d_i - c_i)(d_j - c_j)}$$

and finally scaling the overall matrix by the volume  $V = \prod_i (d_i - c_i)$ .

### C.4. Product Kernel

For a product kernel in  $D$  dimensions, the different dimensions do not interact with each other, and the full  $\Psi$  matrix is given by the Kronecker product of the one-dimensional matrices:

$$\Psi = \bigotimes_{d=1}^D \Psi_d. \quad (67)$$

## D. Supplementary Figures

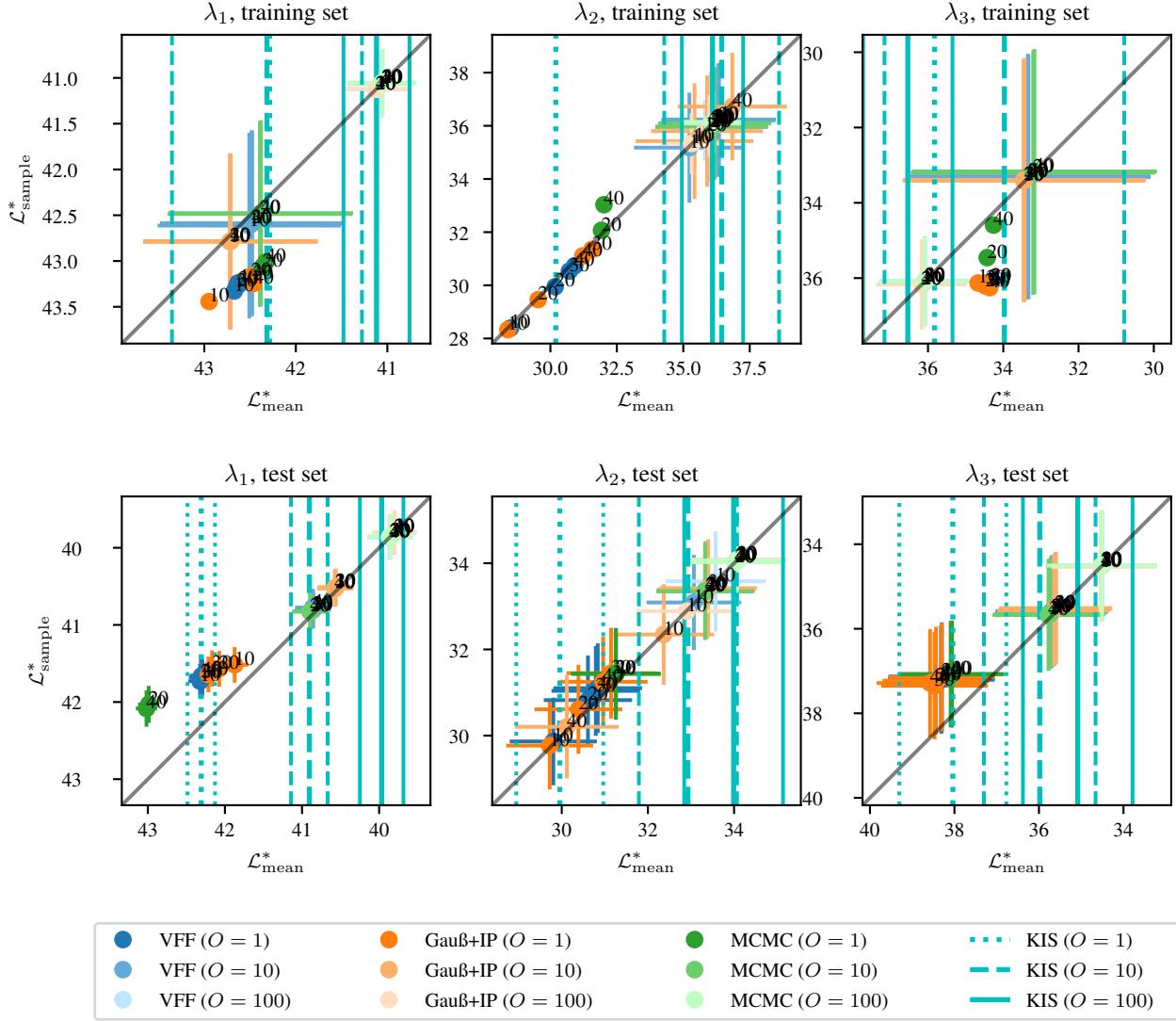


Figure S1. Test set likelihoods for the 1D examples  $\lambda_1$ ,  $\lambda_2$ , and  $\lambda_3$ , comparing  $L^*_{\text{sample}}$  and  $L^*_{\text{mean}}$ . We show results for VFF, Gauß+IP, MCMC, and KIS, for different training set sizes and for different numbers of features (frequencies or points, numbers next to each dot). The test set contains 100 observations; the error bars show the error of the mean across observations. For KIS, the confidence interval is denoted by the thinner lines. This shows that using the mean rate instead of the full samples is generally a good approximation. Exceptions to this are when the uncertainty in the posterior is large; this is more relevant for small numbers of observations in the training set.

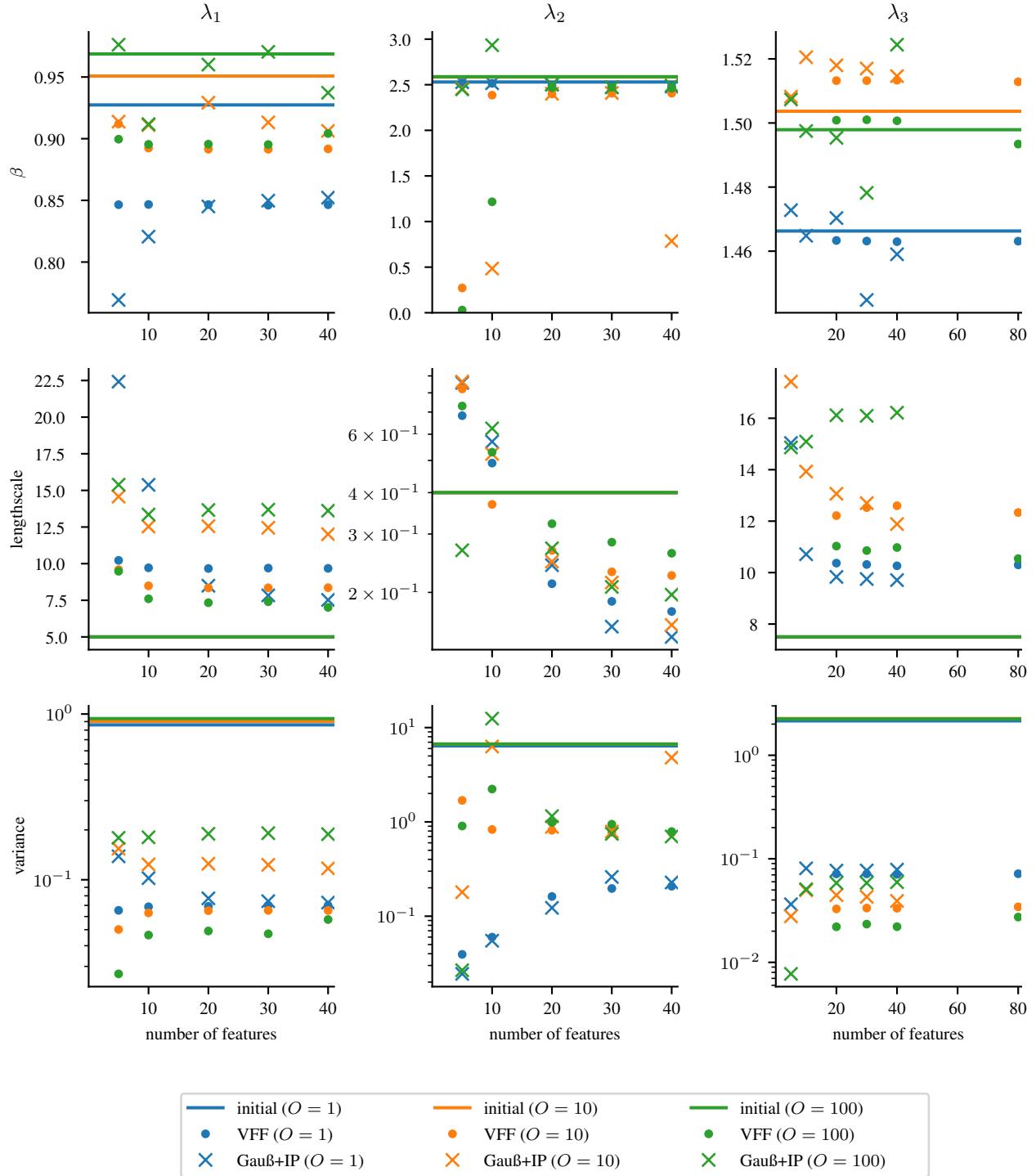
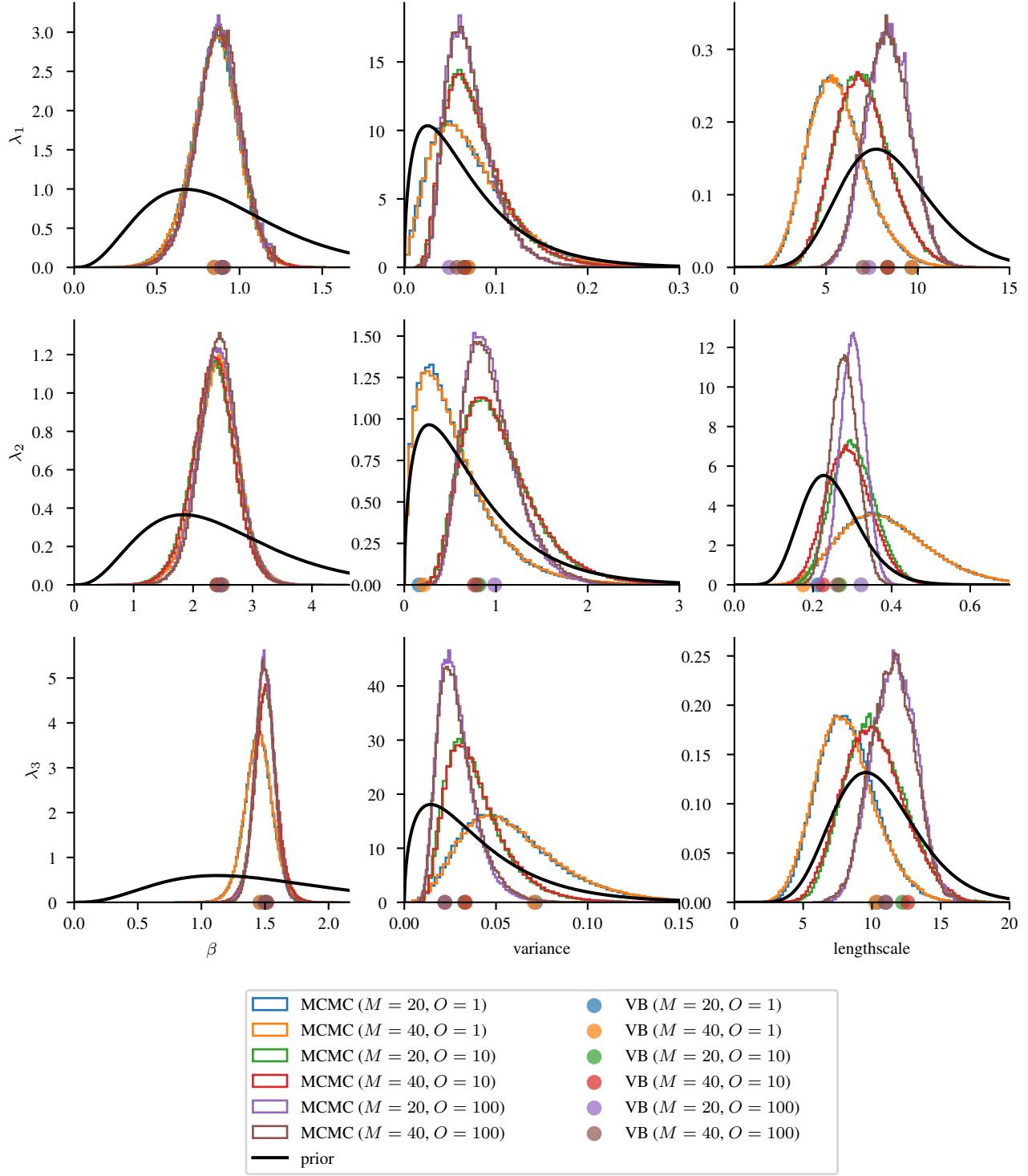


Figure S2. Optimized hyperparameters in variational inference for VFF and Gauß+IP for the 1D examples  $\lambda_1$ ,  $\lambda_2$ , and  $\lambda_3$ , for different numbers of observations  $O$  in the training set. Horizontal lines denote the initial values. This shows how the point estimates converge with an increasing number of features. VFF tends to converge faster. Note that an insufficient number of inducing features is generally associated with a too large lengthscale. The constant offset  $\beta$  can be estimated well using our heuristics, whereas it is more difficult to estimate the variance *a priori*.



*Figure S3.* Histograms of the hyperparameter posterior distributions obtained through MCMC, for the same observations as in Figure S2. Dots denote corresponding VB point estimate. In most cases,  $M = 20$  and  $M = 40$  frequencies cannot be distinguished. The exception is  $\lambda_2$  with  $O = 100$  observations, where a larger number of frequencies allows us to resolve the oscillations with a slightly smaller lengthscale. For the comparatively smooth  $\lambda_1$  and  $\lambda_3$ , the lengthscale increases with the number of observations, as the data becomes more even and less affected by the shot noise. For  $\lambda_2$ , with less observations the inference smoothes over the oscillations. In all cases, larger numbers of observations result in tighter and more peaked distributions.

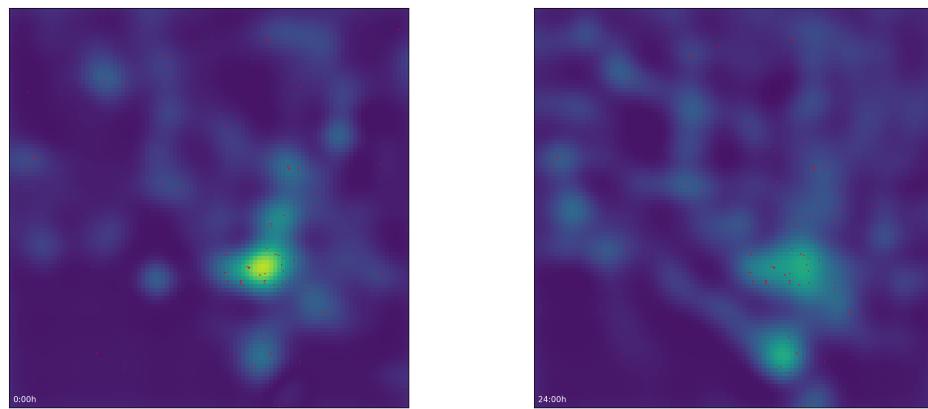


Figure S4. Inferred rate for midnight 00:00 (left) and 24:00 (right) for a spatiotemporal model including day-of-time but with a non-periodic kernel for the time dimension. With a periodic kernel, the rates for both cases would be equal.