(Supplemental Material) Appendix:

Markov Modulated Gaussian Cox Processes for Semi-Stationary Intensity Modeling of Events Data

Minyoung Kim [1 2]

A. Continuous-Time Markov Chain

An r-state continuous-time Markov chain (CTMC) is specified by the initial state probability $\pi_i = P(X(0) = i)$ for $i = 1, \ldots, r$, and the transition rate matrix Q whose off-diagonal Q_{ij} $(i \neq j)$ defines the probability rate of state change from i to j, namely

$$Q_{ij} = \lim_{\Delta t \to 0} \frac{P(X(t + \Delta t) = j | X(t) = i)}{\Delta t}. \tag{1}$$

For convenience, we define the diagonal entries of Q such that Q has 0 row sums, i.e., $Q_{ii} := -\sum_{j \neq i} Q_{ij}$. We often denote the initial state probabilities as a row vector π (i.e., $\pi = [\pi_1, \ldots, \pi_r]$).

Then, using the differential notations and denoting the infinitesimal time duration by dt, the following holds for all $i, j = 1, \ldots, r$,

$$P(X(t + dt) = j | X(t) = i) = \mathbb{I}_{\{i=j\}} + Q_{ij} dt. \tag{2}$$

Hence the marginal state distribution conforms to the first-order linear ODE. More specifically, letting $v(t)$ be the r-dim row vector of the state distribution at time t (i.e., $[v(t)]_i = p(X(t) = i)$), we can rewrite (2) in a vector notation,

$$\dot{v}(t) = v(t) Q. \tag{3}$$

With the initial condition $v(0) = \pi$, the ODE (3) admits the closed-form solution $v(t) = \pi e^{tQ}$, where e^A denotes the matrix exponential of a square matrix A, that is, $e^A = \sum_{k=0}^{\infty} \frac{A^k}{k!}$ with $A^0 = I$, the $(r \times r)$ identity matrix. Hence the marginal distribution can be written as:

$$P(X(t) = i) = [\pi e^{tQ}]_i. \tag{4}$$

Also, from (2) and (4), it is obvious that the joint distribution of two consecutive states having different states $(j \neq i)$ becomes:

$$P(X(t + dt) = j, X(t) = i) = [\pi e^{tQ}]_i Q_{ij} dt. \tag{5}$$

B. Integral of Matrix Exponentials and Derivatives

We describe how the integral of the matrix exponential $J_C = \int_0^T e^{tC} dt$ can be computed analytically once C is diagonalized. It is required in Sec. 4.2 of the main manuscript when computing the KL divergence between the variational CTMC distribution $q(X)$ and the model's prior $P(X)$.

Suppose that C is diagonalized as $C = UDU^{-1}$ for the diagonal matrix D (diagonal entries denoted by d_i for $i = 1, \ldots, r$) and the invertible U. From the definition of the matrix exponential, we have $e^{tC} = Ue^{tD}U^{-1}$ and e^{tD} is also a diagonal matrix with diagonal entries e^{td_i} for $i = 1, \ldots, r$. Applying the integration leads to:

$$J_C = \int_0^T Ue^{tD}U^{-1} dt = U \int_0^T e^{tD} dt \, U^{-1} = UEU^{-1}, \tag{6}$$

[1]Seoul National University of Science & Technology, Korea [2]Rutgers University, Piscataway, NJ, USA. Correspondence to: Minyoung Kim <mikim21@gmail.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

where $E = \int_0^T e^{tD} dt$ is a $(r \times r)$ diagonal matrix whose (i,i) entry is $(i = 1, \ldots, r)$:

$$E_{ii} = \begin{cases} (e^{Td_i} - 1)/d_i & \text{if } d_i \neq 0 \\ T & \text{if } d_i = 0 \end{cases} . \tag{7}$$

Note that (7) involves only scalar exponentials.

Next we derive the gradient of J_C with respect to C, which is needed when taking derivatives of $\mathrm{KL}(q(X)||P(X))$ in Appendix C, one of the objectives in the variational inference. To do this, we use the technique in (Kalbfleisch & Lawless, 1985) where they derived the gradients of the matrix exponential functions. Therein, for a square matrix A whose elements are functions of some parameters $\boldsymbol{\eta}$, it is shown that the partial derivative of e^{tA} wrt each scalar element $\eta \in \boldsymbol{\eta}$ can be written as:

$$\frac{\partial e^{tA}}{\partial \eta} = UVU^{-1}, \tag{8}$$

where $A = UDU^{-1}$ is the diagonalization of A (at current η), V is the $(r \times r)$ matrix defined as

$$V_{ij} = \begin{cases} G_{ij} \frac{e^{td_i} - e^{td_j}}{d_i - d_j} & (i \neq j) \\ G_{ii} t e^{td_i} & (i = j) \end{cases} , \tag{9}$$

and $G = U^{-1} \left(\frac{\partial A}{\partial \eta} \right) U$.

In our case, we note that the elements of the CTMC rate matrix C have certain constraints: $C_{kl} > 0$ for all $k \neq l$, and $C_{kk} = -\sum_{l \neq k} C_{kl}$. To facilitate unconstrained gradient descent optimization, we rather employ a set of *unconstrained* parameters for C instead of directly working with it. More specifically, we define $\overline{C} = \{\overline{C}_{kl}\}_{k \neq l}$ to be the unconstrained parameters, from which C can be recovered as $C_{kl} = \exp(\overline{C}_{kl})$ for $k \neq l$, and C_{kk}'s accordingly. Then it is easy to see that $\frac{\partial C}{\partial \overline{C}_{kl}}$ for $k \neq l$, is the all-zero matrix except for two entries:

$$\left[\frac{\partial C}{\partial \overline{C}_{kl}} \right]_{kl} = C_{kl}, \quad \left[\frac{\partial C}{\partial \overline{C}_{kl}} \right]_{kk} = -C_{kl}. \tag{10}$$

This lets us express the partial derivative of J_C wrt $\eta = \overline{C}_{kl}$ $(k \neq l)$ as:

$$\frac{\partial J_C}{\partial \overline{C}_{kl}} = \int_0^T \frac{\partial e^{tC}}{\partial \overline{C}_{kl}} dt = UHU^{-1}, \tag{11}$$

where $C = UDU^{-1}$ is the diagonalization of C, and $H = \int_0^T V dt$ (with V from (9)) is a $(r \times r)$ matrix whose (i,j) entry $(i,j = 1, \ldots, r)$ is defined as

$$H_{ij} = \begin{cases} \frac{G_{ij}}{d_i - d_j} \left(\frac{e^{Td_i} - 1}{d_i} - \frac{e^{Td_j} - 1}{d_j} \right) & (i \neq j) \\ G_{ii} \frac{(Td_i - 1)e^{Td_i} + 1}{d_i^2} & (i = j) \end{cases} \tag{12}$$

Note that from (10) we have G in (9) now defined as:

$$G = C_{kl} \left[U^{-1} \right]_{:k} \left(\left[U \right]_{l:} - \left[U \right]_{k:} \right), \tag{13}$$

where $\left[A \right]_{k:}$ and $\left[A \right]_{:k}$ indicate the k-th row and column vectors of the matrix A, respectively.

C. Gradient Derivations for ELBO

As shown in (12) of the main manuscript, the ELBO objective is comprised of three terms. We derive gradients for each term in the subsequent sections. To recap, the parameters with which the objectives are differentiated are: i) $\Theta = \{\boldsymbol{\theta}_m = \{\theta_m^i\}_{i=1}^r, \boldsymbol{\theta}_k = \{\theta_k^i\}_{i=1}^r\}$ is the parameters of the r mean and covariance functions of the prior GP, ii) $\Omega = \{Q, \pi\}$ is the

CTMC parameters for the prior state trajectory distribution, and iii) $\Lambda := \{C, \alpha, \boldsymbol{\mu} := \{\mu^i\}_{i=1}^r, \boldsymbol{\Sigma} := \{\Sigma^i\}_{i=1}^r\}$ is the variational parameters for the posterior where the former two are the parameters for the CTMC $q(X)$, and the rest for $q(\mathbf{f})$.

Some of these parameters are constrained, including the CTMC transition rate matrices C and Q as discussed in the previous section. For instance, the initial state distributions α and π have to lie in the probability simplex, while the covariance matrices Σ^i's should be positive definite. Since the unconstrained re-parametrization to reflect these constraints are relatively well known and straightforward[1], we provide derivatives with the original parameters directly, unlike C (and Q) for which we take derivatives wrt the unconstrained parameters $\overline{C}$ (and $\overline{Q}$).

C.1. Gradients of $\mathrm{KL}\big(q(\mathbf{F}_{\mathcal{Z}})||P(\mathbf{F}_{\mathcal{Z}})\big)$

This objective is the sum of the KL divergence between Gaussians, more specifically,

$$\mathrm{KL}\big(q(\mathbf{f}_{\mathcal{Z}})||P(\mathbf{f}_{\mathcal{Z}})\big) = \sum_{i=1}^{r} \frac{1}{2}\left[\log\frac{\left|K^i_{\mathcal{Z}^i,\mathcal{Z}^i}\right|}{\left|\Sigma^i\right|} - M_i + \mathrm{Tr}\Big(\big(K^i_{\mathcal{Z}^i,\mathcal{Z}^i}\big)^{-1}\Sigma^i\Big) + (\mu^i - m^i_{\mathcal{Z}^i})^\top\big(K^i_{\mathcal{Z}^i,\mathcal{Z}^i}\big)^{-1}(\mu^i - m^i_{\mathcal{Z}^i})\right] \tag{14}$$

It is obvious that the i-th ($i = 1, \ldots, r$) summand of the objective is dependent solely on $(\theta^i_m, \theta^i_k, \mu^i, \Sigma^i)$. The derivatives of each term of the i-th summand in (14) can be derived as follows. Here, we use the notation $[\cdot]_j$ for the parameters θ^i_m and θ^i_k to indicate individual *scalar* element. For simplicity, we also drop the dependency on $\mathcal{Z}^i$ in notation (e.g., $K^i_{\mathcal{Z}^i,\mathcal{Z}^i}$ and $m^i_{\mathcal{Z}^i}$ simply written as K^i and m^i, respectively).

$$\frac{\partial \log\left|K^i\right|}{\partial[\theta^i_k]_j} = \mathrm{Tr}\left(\big(K^i\big)^{-1}\frac{\partial K^i}{\partial[\theta^i_k]_j}\right) \tag{15}$$

$$\frac{\partial \log\left|\Sigma^i\right|}{\partial \Sigma^i} = \big(\Sigma^i\big)^{-1} \tag{16}$$

$$\frac{\partial \mathrm{Tr}\Big(\big(K^i\big)^{-1}\Sigma^i\Big)}{\partial[\theta^i_k]_j} = -\mathrm{Tr}\left(\big(K^i\big)^{-1}\Sigma^i\big(K^i\big)^{-1}\frac{\partial K^i}{\partial[\theta^i_k]_j}\right) \tag{17}$$

$$\frac{\partial \mathrm{Tr}\Big(\big(K^i\big)^{-1}\Sigma^i\Big)}{\partial \Sigma^i} = \big(K^i\big)^{-1} \tag{18}$$

$$\frac{\partial(\mu^i - m^i)^\top\big(K^i\big)^{-1}(\mu^i - m^i)}{\partial[\theta^i_k]_j} = -(\mu^i - m^i)^\top\big(K^i\big)^{-1}\frac{\partial K^i}{\partial[\theta^i_k]_j}\big(K^i\big)^{-1}(\mu^i - m^i) \tag{19}$$

$$\frac{\partial(\mu^i - m^i)^\top\big(K^i\big)^{-1}(\mu^i - m^i)}{\partial[\theta^i_m]_j} = -2(\mu^i - m^i)^\top\big(K^i\big)^{-1}\frac{\partial m^i}{\partial[\theta^i_m]_j} \tag{20}$$

$$\frac{\partial(\mu^i - m^i)^\top\big(K^i\big)^{-1}(\mu^i - m^i)}{\partial \mu^i} = 2\big(K^i\big)^{-1}(\mu^i - m^i) \tag{21}$$

Note that $\frac{\partial K^i}{\partial[\theta^i_k]_j}$ and $\frac{\partial m^i}{\partial[\theta^i_m]_j}$ indicate the derivatives of the kernel matrix K^i and the mean vector m^i, which are of the same dimensions as K^i and m^i, respectively, obtained by differentiating each element.

C.2. Gradients of $\mathrm{KL}\big(q(X)||P(X)\big)$

The objective, as written below, is dependent only on the CTMC parameters of the prior and the variational distributions on the state trajectory $X(\cdot)$.

$$\mathrm{KL}\big(q(X)||P(X)\big) = \sum_{i=1}^{r}\left\{\alpha_i\log\frac{\alpha_i}{\pi_i} + [\alpha J_C]_i\left(C_{ii} - Q_{ii} + \sum_{j\neq i} C_{ij}\log\frac{C_{ij}}{Q_{ij}}\right)\right\} \tag{22}$$

We take derivate of the objective with respect to individual scalar parameters, namely α_k (and π_k) for $k = 1, \ldots, r$, and $\overline{C}_{kl}$ (and $\overline{Q}_{kl}$) for $k, l = 1, \ldots, r$ with $k \neq l$. For notational simplicity, we denote the objective in (22) by KL_X.

[1] For the positive definite constraints, for instance, one can use the Cholesky decomposition, optionally with exponentiation of diagonal elements (Pinheiro & Bates, 1996).

$$\frac{\partial KL_X}{\partial \alpha_k} = 1 + \log \frac{\alpha_k}{\pi_k} + \sum_{i=1}^{r} [J_C]_{ki} \left(C_{ii} - Q_{ii} + \sum_{j \neq i} C_{ij} \log \frac{C_{ij}}{Q_{ij}} \right) \tag{23}$$

$$\frac{\partial KL_X}{\partial \pi_k} = -\frac{\alpha_k}{\pi_k} \tag{24}$$

$$\frac{\partial KL_X}{\partial \overline{C}_{kl}} = \sum_{i=1}^{r} \left\{ \left[\alpha \frac{\partial J_C}{\partial \overline{C}_{kl}} \right]_i \left(C_{ii} - Q_{ii} + \sum_{j \neq i} C_{ij} \log \frac{C_{ij}}{Q_{ij}} \right) + [\alpha J_C]_i \left(\frac{\partial C_{ii}}{\partial \overline{C}_{kl}} + \sum_{j \neq i} \frac{\partial C_{ij}}{\partial \overline{C}_{kl}} \left(1 + \log \frac{C_{ij}}{Q_{ij}} \right) \right) \right\} \tag{25}$$

$$\frac{\partial KL_X}{\partial \overline{Q}_{kl}} = -\sum_{i=1}^{r} [\alpha J_C]_i \left(\frac{\partial Q_{ii}}{\partial \overline{Q}_{kl}} + \sum_{j \neq i} \frac{\partial Q_{ij}}{\partial \overline{Q}_{kl}} \frac{C_{ij}}{Q_{ij}} \right) \tag{26}$$

Here, the derivatives $\frac{\partial C_{ij}}{\partial \overline{C}_{kl}}$ (similarly $\frac{\partial Q_{ij}}{\partial \overline{Q}_{kl}}$) and $\frac{\partial J_C}{\partial \overline{C}_{kl}}$ follow (10) and (11), respectively, as described in Appendix B.

C.3. Gradients of $\mathbb{E}_{q(X,\mathbf{f})}\big[\log P(\mathcal{D}|X,\mathbf{f})\big]$

Recall from Sec. 4.3 in the main article that this objective can be approximated as:

$$\mathbb{E}_{q(X,\mathbf{f})}\big[\log P(\mathcal{D}|X,\mathbf{f})\big] = \sum_{i=1}^{r} \big(\text{ELL}_i - \text{ENO}_i\big) \approx \sum_{i=1}^{r} \big(\widehat{\text{ELL}}_i - \widehat{\text{ENO}}_i\big). \tag{27}$$

Here $\widehat{\text{ELL}}_i$ is the Monte-Carlo estimate of $\text{ELL}_i = \sum_{n=1}^{N} [\alpha e^{t_n C}]_i \mathbb{E}_{q(f^i(t_n))}\big[\log (f^i(t_n))^2\big]$ for $i = 1, \ldots, r$, that is,

$$\widehat{\text{ELL}}_i = \sum_{n=1}^{N} \frac{[\alpha e^{t_n C}]_i}{S} \sum_{s=1}^{S} \log \big(f_n^{i\,(s)}\big)^2 \tag{28}$$

where $\epsilon_{in}^{(s)}$ $(s = 1, \ldots, S)$ are iid random samples from standard normal for each (i, n), and $f_n^{i\,(s)} = \tilde{\mu}_i(t_n) + (\tilde{\sigma}_i^2(t_n))^{1/2} \epsilon_{in}^{(s)}$ with

$$\tilde{\mu}_i(t) = m^i(t) + K^i_{t,\mathcal{Z}^i} (K^i_{\mathcal{Z}^i,\mathcal{Z}^i})^{-1} (\mu^i - m^i_{\mathcal{Z}^i}), \tag{29}$$

$$\tilde{\sigma}_i^2(t) = K^i_{t,t} - K^i_{t,\mathcal{Z}^i} (K^i_{\mathcal{Z}^i,\mathcal{Z}^i})^{-1} K^i_{\mathcal{Z}^i,t} + K^i_{t,\mathcal{Z}^i} (K^i_{\mathcal{Z}^i,\mathcal{Z}^i})^{-1} \Sigma^i (K^i_{\mathcal{Z}^i,\mathcal{Z}^i})^{-1} K^i_{\mathcal{Z}^i,t}. \tag{30}$$

Also, $\widehat{\text{ENO}}_i$ is the grid-based numerical integration of $\text{ENO}_i = \int_0^T [\alpha e^{tC}]_i \mathbb{E}_{q(f^i(t))}\big[(f^i(t))^2\big] dt$. Specifically, with the G uniform grid points $\{\tilde{t}_g\}_{g=1}^G$ over $[0, T]$, we have (dropping the dependency on $\mathcal{Z}^i$ in notation for simplicity):

$$\begin{aligned} \widehat{\text{ENO}}_i = {} & (\mu^i - m^i)^\top \big(K^i\big)^{-1} \Psi_0^i \big(K^i\big)^{-1} (\mu^i - m^i) + 2\Psi_1^i \big(K^i\big)^{-1} (\mu^i - m^i) - \text{Tr}\Big(\big(K^i\big)^{-1} \Psi_0^i\Big) + \\ & \text{Tr}\Big(\big(K^i\big)^{-1} \Sigma^i \big(K^i\big)^{-1} \Psi_0^i\Big) + \Psi_2^i + \Psi_3^i, \end{aligned} \tag{31}$$

where the Ψ's are defined as follows (by denoting $w_g^i = [\alpha e^{\tilde{t}_g C}]_i \Delta t$ with $\Delta t = \tilde{t}_{g+1} - \tilde{t}_g$),

$$\Psi_0^i = \sum_{g=1}^{G} w_g^i K^i_{\mathcal{Z}^i,\tilde{t}_g} K^i_{\tilde{t}_g,\mathcal{Z}^i}, \;\; \Psi_1^i = \sum_{g=1}^{G} w_g^i m^i(\tilde{t}_g) K^i_{\tilde{t}_g,\mathcal{Z}^i}, \;\; \Psi_2^i = \sum_{g=1}^{G} w_g^i (m^i(\tilde{t}_g))^2, \;\; \Psi_3^i = \sum_{g=1}^{G} w_g^i K^i_{\tilde{t}_g,\tilde{t}_g} \tag{32}$$

For the gradients of $\widehat{\text{ELL}}_i$ in (28), letting $w_n^i = [\alpha e^{t_n C}]_i$, we see that w_n^i is dependent only on (α, C), while the latter summation term (over s) is a function of the GP mean, covariance, and the variational parameters. Hence the gradients of $\widehat{\text{ELL}}_i$ wrt (α, C) can be derived as:

$$\frac{\partial \widehat{\text{ELL}}_i}{\partial \alpha_k} = \sum_{n=1}^{N} \frac{[e^{t_n C}]_{ki}}{S} \sum_{s=1}^{S} \log \big(f_n^{i\,(s)}\big)^2 \tag{33}$$

$$\frac{\partial \widehat{\text{ELL}}_i}{\partial \overline{C}_{kl}} = \sum_{n=1}^{N} \frac{1}{S} \left[\alpha \frac{\partial e^{t_n C}}{\partial \overline{C}_{kl}} \right]_i \sum_{s=1}^{S} \log \big(f_n^{i\,(s)}\big)^2 \tag{34}$$

where $\frac{\partial e^{t_n C}}{\partial \overline{C}_{kl}}$ can be computed from (8) with G from (13). The gradients of $\widehat{\text{ELL}}_i$ wrt $\eta \in \{\theta_m^i, \theta_k^i, \mu^i, \Sigma^i\}$ are as follows:

$$\frac{\partial \widehat{\text{ELL}}_i}{\partial \eta} = \sum_{n=1}^{N} \frac{w_n^i}{S} \sum_{s=1}^{S} \left\{ \frac{\partial \tilde{\mu}_i(t_n)}{\partial \eta} \frac{2}{f_n^{i\,(s)}} + \frac{\partial \tilde{\sigma}_i^2(t_n)}{\partial \eta} \frac{\epsilon_{in}^{(s)}}{f_n^{i\,(s)} (\tilde{\sigma}_i^2(t_n))^{1/2}} \right\} \tag{35}$$

where the partial derivatives of $\tilde{\mu}_i(t_n)$ and $\tilde{\sigma}_i^2(t_n)$ wrt specific parameters can be obtained as:

$$\begin{aligned}
\frac{\partial \tilde{\mu}_i(t_n)}{\partial [\theta_k^i]_j} &= \frac{\partial K_n^i}{\partial [\theta_k^i]_j} \big(K^i\big)^{-1}(\mu^i - m^i) - K_n^i \big(K^i\big)^{-1} \frac{\partial K^i}{\partial [\theta_k^i]_j} \big(K^i\big)^{-1}(\mu^i - m^i) && (36)\\
\frac{\partial \tilde{\mu}_i(t_n)}{\partial [\theta_m^i]_j} &= \frac{\partial m_n^i}{\partial [\theta_m^i]_j} - K_n^i \big(K^i\big)^{-1} \frac{\partial m^i}{\partial [\theta_m^i]_j} && (37)\\
\frac{\partial \tilde{\mu}_i(t_n)}{\partial \mu^i} &= \big(K^i\big)^{-1} \big(K_n^i\big)^\top && (38)\\
\frac{\partial \tilde{\sigma}_i^2(t_n)}{\partial [\theta_k^i]_j} &= \frac{\partial K_{nn}^i}{\partial [\theta_k^i]_j} - 2 \frac{\partial K_n^i}{\partial [\theta_k^i]_j} \big(K^i\big)^{-1} \big(K_n^i\big)^\top + K_n^i \big(K^i\big)^{-1} \frac{\partial K^i}{\partial [\theta_k^i]_j} \big(K^i\big)^{-1} \big(K_n^i\big)^\top + \\
&\qquad 2 \left(\frac{\partial K_n^i}{\partial [\theta_k^i]_j} - K_n^i \big(K^i\big)^{-1} \frac{\partial K^i}{\partial [\theta_k^i]_j} \right) \big(K^i\big)^{-1} \Sigma^i \big(K^i\big)^{-1} \big(K_n^i\big)^\top && (39)\\
\frac{\partial \tilde{\sigma}_i^2(t_n)}{\partial \Sigma^i} &= \big(K^i\big)^{-1} \big(K_n^i\big)^\top K_n^i \big(K^i\big)^{-1} && (40)
\end{aligned}$$

where we have several short-cut notations: $K_n^i = K_{t_n, \mathcal{Z}^i}^i$, $K_{nn}^i = K_{t_n, t_n}^i$, $m_n^i = m^i(t_n)$, and as usual $K^i = K_{\mathcal{Z}^i, \mathcal{Z}^i}^i$, $m^i = m_{\mathcal{Z}^i}^i$.

For the gradients of $\widehat{\text{ENO}}_i$, we first compute the derivatives of the Ψ terms in (32). Note that (α, C) affects the Ψ terms only through $w_g^i = [\alpha e^{\tilde{t}_g C}]_i \Delta t$, and it is sufficient to derive:

$$\frac{\partial w_g^i}{\partial \alpha_k} = [e^{\tilde{t}_g C}]_{ki} \Delta t, \quad \frac{\partial w_g^i}{\partial \overline{C}_{kl}} = \left[\alpha \frac{\partial e^{\tilde{t}_g C}}{\partial \overline{C}_{kl}} \right]_i \Delta t. \tag{41}$$

The gradients of the Ψ terms wrt $(\boldsymbol{\theta}_m, \boldsymbol{\theta}_k)$ can be evaluated as follows, where for notational simplicity, the GP mean vectors and kernel matrices on the grid points are denoted as: $K_g^i = K_{\tilde{t}_g, \mathcal{Z}^i}^i$, $K_{gg}^i = K_{\tilde{t}_g, \tilde{t}_g}^i$, and $m_g^i = m^i(\tilde{t}_g)$.

$$\begin{aligned}
\frac{\partial \Psi_0^i}{\partial [\theta_k^i]_j} &= \sum_{g=1}^{G} w_g^i \left\{ \left(\frac{\partial K_g^i}{\partial [\theta_k^i]_j} \right)^\top K_g^i + \big(K_g^i\big)^\top \frac{\partial K_g^i}{\partial [\theta_k^i]_j} \right\} && (42)\\
\frac{\partial \Psi_1^i}{\partial [\theta_k^i]_j} &= \sum_{g=1}^{G} w_g^i m_g^i \frac{\partial K_g^i}{\partial [\theta_k^i]_j}, \qquad \frac{\partial \Psi_1^i}{\partial [\theta_m^i]_j} = \sum_{g=1}^{G} w_g^i \frac{\partial m_g^i}{\partial [\theta_m^i]_j} K_g^i && (43)\\
\frac{\partial \Psi_2^i}{\partial [\theta_m^i]_j} &= 2 \sum_{g=1}^{G} w_g^i m_g^i \frac{\partial m_g^i}{\partial [\theta_m^i]_j}, \qquad \frac{\partial \Psi_3^i}{\partial [\theta_k^i]_j} = \sum_{g=1}^{G} w_g^i \frac{\partial K_{gg}^i}{\partial [\theta_k^i]_j} && (44)
\end{aligned}$$

Next, we tackle gradients of each term in (31). The first term, $\text{TERM}_1 = (\mu^i - m^i)^\top \big(K^i\big)^{-1} \Psi_0^i \big(K^i\big)^{-1} (\mu^i - m^i)$ has:

$$\begin{aligned}
\frac{\partial \text{TERM}_1}{\partial [\theta_k^i]_j} &= (\mu^i - m^i)^\top \big(K^i\big)^{-1} \frac{\partial \Psi_0^i}{\partial [\theta_k^i]_j} \big(K^i\big)^{-1} (\mu^i - m^i) \; - \\
&\qquad 2 (\mu^i - m^i)^\top \big(K^i\big)^{-1} \frac{\partial K^i}{\partial [\theta_k^i]_j} \big(K^i\big)^{-1} \Psi_0^i \big(K^i\big)^{-1} (\mu^i - m^i) && (45)\\
\frac{\partial \text{TERM}_1}{\partial [\theta_m^i]_j} &= -2 \left(\frac{\partial m^i}{\partial [\theta_m^i]_j} \right)^\top \big(K^i\big)^{-1} \Psi_0^i \big(K^i\big)^{-1} (\mu^i - m^i) && (46)\\
\frac{\partial \text{TERM}_1}{\partial \mu^i} &= 2 \big(K^i\big)^{-1} \Psi_0^i \big(K^i\big)^{-1} (\mu^i - m^i) && (47)
\end{aligned}$$

The second term, $\text{TERM}_2 = \Psi_1^i \big(K^i\big)^{-1}(\mu^i - m^i)$ admits:

$$\begin{aligned}
\frac{\partial \text{TERM}_2}{\partial [\theta_k^i]_j} &= \frac{\partial \Psi_1^i}{\partial [\theta_k^i]_j}\big(K^i\big)^{-1}(\mu^i - m^i) - \Psi_1^i\big(K^i\big)^{-1}\frac{\partial K^i}{\partial [\theta_k^i]_j}\big(K^i\big)^{-1}(\mu^i - m^i) && (48)\\
\frac{\partial \text{TERM}_2}{\partial [\theta_m^i]_j} &= \frac{\partial \Psi_1^i}{\partial [\theta_m^i]_j}\big(K^i\big)^{-1}(\mu^i - m^i) - \Psi_1^i\big(K^i\big)^{-1}\frac{\partial m^i}{\partial [\theta_m^i]_j} && (49)\\
\frac{\partial \text{TERM}_2}{\partial \mu^i} &= \big(K^i\big)^{-1}\big(\Psi_1^i\big)^\top && (50)
\end{aligned}$$

The third term, $\text{TERM}_3 = \text{Tr}\Big(\big(K^i\big)^{-1}\Psi_0^i\Big)$ has:

$$\frac{\partial \text{TERM}_3}{\partial [\theta_k^i]_j} = \text{Tr}\left(\big(K^i\big)^{-1}\frac{\partial \Psi_0^i}{\partial [\theta_k^i]_j}\right) - \text{Tr}\left(\big(K^i\big)^{-1}\frac{\partial K^i}{\partial [\theta_k^i]_j}\big(K^i\big)^{-1}\Psi_0^i\right) \tag{51}$$

The fourth term, $\text{TERM}_4 = \text{Tr}\Big(\big(K^i\big)^{-1}\Sigma^i\big(K^i\big)^{-1}\Psi_0^i\Big)$ admits:

$$\begin{aligned}
\frac{\partial \text{TERM}_4}{\partial [\theta_k^i]_j} &= \text{Tr}\left(\big(K^i\big)^{-1}\Sigma^i\big(K^i\big)^{-1}\frac{\partial \Psi_0^i}{\partial [\theta_k^i]_j}\right) - 2\text{Tr}\left(\big(K^i\big)^{-1}\frac{\partial K^i}{\partial [\theta_k^i]_j}\big(K^i\big)^{-1}\Sigma^i\big(K^i\big)^{-1}\Psi_0^i\right) && (52)\\
\frac{\partial \text{TERM}_4}{\partial \Sigma^i} &= \big(K^i\big)^{-1}\Psi_0^i\big(K^i\big)^{-1} && (53)
\end{aligned}$$

The remaining Ψ_2^i and Ψ_3^i have been done already.

References

Kalbfleisch, J. D. and Lawless, J. F. The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, 80:863–871, 1985.

Pinheiro, J. C. and Bates, D. M. Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, 6(3):289–296, 1996.