1st Symposium on Advances in Approximate Bayesian Inference, 2018 1–9

Scalable GAM using sparse variational Gaussian Processes

Vincent Adam VINCENT.ADAM@PROWLER.IO
and **Nicolas Durrande** NICOLAS@PROWLER.IO
and **ST John** ST@PROWLER.IO
PROWLER.io, 72 Hills Road, Cambridge, CB2 1LA, United Kingdom

Abstract

Generalized additive models (GAMs) are a widely used class of models of interest to statisticians as they provide a flexible way to design interpretable models of data beyond linear models. We here propose a scalable and well-calibrated Bayesian treatment of GAMs using Gaussian processes and leveraging recent advances in variational inference. We use sparse Gaussian processes to represent each component and exploit the additive structure of the model to efficiently represent a Gaussian *a posteriori* coupling between the components. Although our scheme is motivated by GAMs, it can be used for arbitrary predictors, beyond additive ones.

Keywords: GAM, Gaussian Process, Variational Inference

1. Introduction

Generalized additive models (GAMs) are a class of interpretable regression models with non-linear yet additive predictors (Hastie and Tibshirani, 1990). Their Bayesian treatment requires the specification of priors over functions. Here, we use Gaussian processes (GPs) (Rasmussen and Williams, 2006) and propose an approximate inference algorithm that is both well-calibrated and scalable with both the number of data points and additive components. We extend the variational pseudo-point GP approximation (Titsias, 2009; Bauer et al., 2016) to posterior dependencies across GPs. This approximation provides state-of-the art performance for GP regression and provides approximations to the posterior distributions in the form of a GP. This approach has been successfully extended to the multiple GP setting using a factorized (mean-field) approximation of the posterior across GPs (Saul et al., 2016; Adam et al., 2016). However, it suffers from the known variance underestimation of mean-field approximations and therefore can lead to poor predictions or can bias learning (Turner and Sahani, 2011). Adam (2017) introduced additional structure to the posterior distribution by allowing some coupling across GPs inducing variables but this was at the cost of scalability.

2. Background

2.1. Regression with multiple GPs

We consider additive models with factorizing likelihood $p(Y \mid X) = \prod_{n=1}^{N} p(y_n \mid \sum_c f_c(x_n))$, where $f_1, \ldots, f_C$ are functions from $\mathcal{X}_c \to \mathbb{R}$. The specific form of the likelihood is arbitrary. We denote $\mathcal{F} = \{f_1, ..., f_C\}$ such that $p(\mathcal{F})$ constitutes the joint distribution over

the processes. $\mathcal{F}(x) = [f_1(x), ..., f_C(x)]$ is the vector of function evaluations at x. We are interested in computing the joint posterior $p(\mathcal{F} \mid X, Y)$.

2.2. Variational Inference

The classical variational lower bound (or ELBO) to the marginal likelihood is given by

$$\log p(Y \mid X) \geq \mathbb{E}_{q(\mathcal{F})}[\log p(Y \mid \mathcal{F}, X)] - \mathrm{KL}[q(\mathcal{F}) \,\|\, p(\mathcal{F})] = \mathcal{L}(q). \tag{1}$$

This is the optimization objective in the Variational Free Energy (VFE) approximation. We choose $q(\mathcal{F})$ to be a multivariate Normal distribution with mean $\mu_{\mathcal{F}}$ and covariance $\Sigma_{\mathcal{F}}$, which is not an approximation in the conjugate likelihood setting. This leads to

$$\mathcal{L}(q) = \mathbb{E}_{q(\mathcal{F})}[\log p(Y \mid \mathcal{F})] + \frac{1}{2}\operatorname{tr}[K_{\mathcal{F}}^{-1}\Sigma_{\mathcal{F}}] + \frac{1}{2}\mu_{\mathcal{F}}^{\intercal}K_{\mathcal{F}}^{-1}\mu_{\mathcal{F}} - \frac{1}{2}\log|\Sigma_{\mathcal{F}}|. \tag{2}$$

The expectation term in equation (2) is intractable in most cases and needs to be approximated. See Hensman et al. (2015) for deterministic approximations and Salimbeni and Deisenroth (2017) for stochastic approximations.

3. Optimal Gaussian Posterior in variational inference

Following Opper and Archambeau (2009), we derive the expression for the optimal $\Sigma_{\mathcal{F}}$ by noting that at the optimum, $\nabla_{\Sigma_{\mathcal{F}}}\mathcal{L}(q) = 0$. This implies that

$$\Sigma_{\mathcal{F}}^{-1} = K_{\mathcal{F}}^{-1} + \nabla_{\Sigma_{\mathcal{F}}}\left[\mathbb{E}_{q(\mathcal{F})}[\log p(Y \mid \mathcal{F})]\right]. \tag{3}$$

3.1. Optimality in the additive case

In the additive case considered here, the gradient term in (3) is low rank and can be parameterized by a vector $\lambda \in \mathbb{R}^N$ as follows, with $\Lambda = \mathrm{diag}(\lambda)$:

$$\Sigma_{\mathcal{F}}^{-1} = K_{\mathcal{F}}^{-1} + (1 \otimes \Lambda)(1 \otimes \Lambda)^{\intercal}. \tag{4}$$

This parameterization requires $2N$ values, equal to that of the classical single GP regression setting described in Opper and Archambeau (2009). It also inherits the non-convexity of this objective as highlighted by Khan et al. (2012).

3.2. Optimality in the sparse additive case

Following Adam et al. (2016) we introduce for each GP indexed by c some 'inducing points' $Z_c = [z_c^{(1)}, ..., z_c^{(m)}] \in \mathcal{X}_c^m$. The vector of associated function evaluations is given by $\mathbf{U}_c = [u_c^{(1)}, ..., u_c^{(m)}] = [f_c(z_c^{(1)}), ..., f_c(z_c^{(m)})]$. We also define the stacked vector $\mathbf{U} = [\mathbf{U}_1, ..., \mathbf{U}_c]$.

Following Adam (2017), we parameterize $q(\mathcal{F}) = q(\mathbf{U})\prod_c p(f_{c\neg\mathbf{U}_c} \mid \mathbf{U}_c)$. This choice leads to a simplification of the lower bound (2) as

$$\mathcal{L}(q) = \mathbb{E}_{q(\mathcal{F})}[\log p(Y \mid \mathcal{F}, X)] - \mathrm{KL}[q(\mathbf{U}) \,\|\, p(\mathbf{U})]. \tag{5}$$

Saul et al. (2016) considered the mean field case $q(\mathbf{U}) = \prod_c q(\mathbf{U}_c)$ with each factor parameterized as a multivariate normal distribution $\mathcal{N}(\mu_{\mathbf{U}_c}, \Sigma_{\mathbf{U}_c})$. This approach does not

capture posterior dependencies across GPs. Adam (2017) parameterized $q(\mathbf{U})$ as a multivariate normal distribution $\mathcal{N}(\mu_{\mathbf{U}}, \Sigma_{\mathbf{U}})$ to include cross-GP coupling through the inducing variables $\mathbf{U}$. We extend this last approach but ask what the optimal $q(\mathbf{U})$ should be. It turns out to be (see Appendix A):

$$\Sigma_{\mathbf{U},\mathbf{U}}^{-1} = K_{\mathbf{U},\mathbf{U}}^{-1} + K_{\mathbf{U},\mathbf{U}}^{-1}\big(\textstyle\sum_c K_{\mathbf{U},f_c}\big)\Lambda\big(\textstyle\sum_{c'} K_{f_{c'},\mathbf{U}}\big)K_{\mathbf{U},\mathbf{U}}^{-1}. \tag{6}$$

This form has again $2N$ parameters which becomes an over-parameterization as soon as $N > M^2C^2/2$. Since we are interested in scalability, it is not of practical interest.

4. A new parameterization for $q(\mathbf{U})$

The second term of the sum in (6) can be expressed as $BB^\intercal$ with B of size $MC \times N$. Keeping this structure arising from the additivity of the model, we propose the parameterization

$$\Sigma_{\mathbf{U},\mathbf{U}}^{-1} = K_{\mathbf{U},\mathbf{U}}^{-1} + BB^\intercal, \tag{7}$$

with B of smaller size $MC \times M$. This parameterization preserves the *structure* of the optimal covariance. It requires storing M^2C values, which is less than a direct representation of a Cholesky factor of $\Sigma_{\mathbf{U},\mathbf{U}}^{-1}$ that would require M^2C^2 parameters.

5. Summary of complexities

Time and space complexity of the sparse variational algorithms are summarized in Table 1.

Model	**Storage**	$\boldsymbol{KL(q\|p)}$	$\boldsymbol{E_q(\rho)\log p(Y\|\rho)}$
Mean field Saul et al. (2016)	$O(CM^2)$	$O(CM^3)$	$O(CM^3 + NCM^2)$
Coupled (covariance) Adam (2017)	$O(C^2M^2)$	$O(C^3M^3)$	$O(C^3M^3 + NC^2M^2)$
Coupled (precision)	$O(CM^2)$	$O(CM^3)$	$O(CM^3 + NC^2M^2)$

Table 1: Complexity of sparse variational additive models

6. Related work

Variational inference for the multi-GP setting has so far only used the mean-field (MF) approximation as described in Saul et al. (2016). When posterior dependencies are a quantity of interest, a natural approach is to increase the complexity of the variational posterior to capture these dependencies. This often results in a prohibitive increase in the complexity of the inference. Different solutions have been proposed to tackle this problem. A first approach in Giordano et al. (2015) consists of a two-step scheme where MF inference is *assumed* to provide accurate posterior mean estimates. A perturbation analysis is then performed around the MF posterior means to provide second order (covariance) estimates. A second approach consists in 'relaxing' the MF approximation by extending the variational posterior $q(\mathcal{F})$ with additional multiplicative terms capturing dependencies while keeping the computational complexity of the resulting inference scheme low (Tran et al., 2015; Hoffman and Blei, 2015). Our approach fits in this second family of extensions of the MF parameterization. It is tailored to the VFE approximation to GP models and leverages its sparsity to provide a fast and scalable inference algorithm.

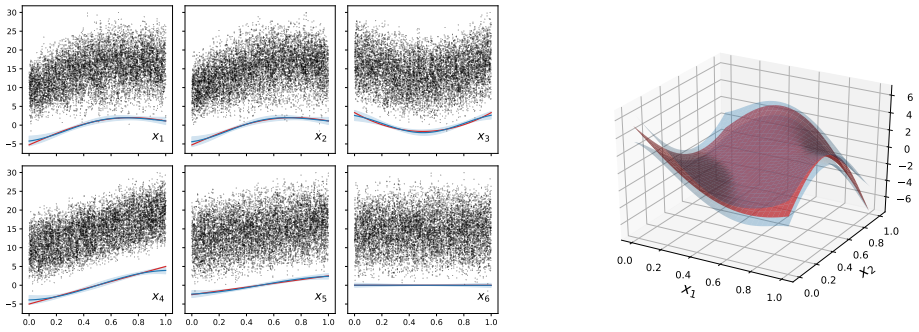

Figure 1: Left: projected observations (black points), estimated (blue) and analytically computed (red) univariate effects for each input dimension. Right: estimated (blue) and analytical (red) interaction between the variables x_1 and x_2.

7. Illustration

We consider a simple regression task consisting of approximating the following function: $f(x) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5$ with $x \in [0,1]^6$ (note that the last variable has no effect), given 5000 observation points uniformly distributed in the input space and a Gaussian observation noise with unit variance.

We choose a kernel dedicated to sensitivity analysis and tailored to the structure of the function at stake [Durrande et al. (2013)]. Given univariate squared exponential kernels $g_1, \dots, g_8$ we define the kernel as $k(x,y) = \sigma_0 + \sum_{i=1}^{6} s_i(x_i, y_i) + s_7(x_1, y_1)s_8(x_2, y_2)$ with

$$s_i(x_i, y_i) = g_i(x_i, y_i) - \frac{\int_0^1 g_i(x_i, s)\,\mathrm{d}s \int_0^1 g_i(y_i, s)\,\mathrm{d}s}{\iint_0^1 g_i(s_i, t)\,\mathrm{d}s\,\mathrm{d}t}. \tag{8}$$

Since the number of observations is relatively large and the kernel has an additive structure (it is the sum of 8 kernels), we choose the *sparse additive model* described above. We choose 16 regularly spaced one-dimensional inducing points for each kernel $s_1, \dots, s_6$ and 16 points distributed as a 4×4 grid for the bi-dimensional kernel $s_7 s_8$. The final model is obtained by maximizing the ELBO with respect to the variational parameters and the hyper-parameters of the g_i. Given the structure of the model and the fact that inducing inputs are dedicated to model components, it is then possible to decompose the model predictions and to represent separately all the components of the ANOVA representation of the test function. Figure 1 shows that the model accurately approximates the test function and that the proposed framework is helpful to reveal its inner structure.

8. Conclusion

We presented a method that provides a fast, scalable and well-calibrated Bayesian treatment of GAMs. Although motivated by GAMs, our structured variational distribution may be used in models where the predictor is non-additive but where the posterior is well-approximated by a unimodal distribution.

References

Vincent Adam. Structured variational inference for coupled Gaussian processes. *arXiv preprint arXiv:1711.01131*, 2017.

Vincent Adam, James Hensman, and Maneesh Sahani. Scalable transformed additive signal decomposition by non-conjugate Gaussian process inference. In *Machine Learning for Signal Processing (MLSP), 2016 IEEE 26th International Workshop on*, pages 1–6. IEEE, 2016.

Matthias Bauer, Mark van der Wilk, and Carl Edward Rasmussen. Understanding probabilistic sparse Gaussian process approximations. In *Advances in Neural Information Processing Systems*, pages 1533–1541, 2016.

Nicolas Durrande, David Ginsbourger, Olivier Roustant, and Laurent Carraro. Anova kernels and rkhs of zero mean functions for model-based sensitivity analysis. *Journal of Multivariate Analysis*, 115:57–67, 2013.

Ryan J Giordano, Tamara Broderick, and Michael I Jordan. Linear response methods for accurate covariance estimates from mean field variational bayes. In *Advances in Neural Information Processing Systems*, pages 1441–1449, 2015.

Trevor Hastie and Robert Tibshirani. *Generalized additive models*. Wiley Online Library, 1990.

James Hensman, Alexander G de G Matthews, and Zoubin Ghahramani. Scalable variational Gaussian process classification. 2015.

Matthew Hoffman and David Blei. Stochastic structured variational inference. In *Artificial Intelligence and Statistics*, pages 361–369, 2015.

Emtiyaz Khan, Shakir Mohamed, and Kevin P Murphy. Fast bayesian inference for non-conjugate Gaussian process regression. In *Advances in Neural Information Processing Systems*, pages 3140–3148, 2012.

Manfred Opper and Cédric Archambeau. The variational Gaussian approximation revisited. *Neural computation*, 21(3):786–792, 2009.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, 2006.

Hugh Salimbeni and Marc Deisenroth. Doubly stochastic variational inference for deep Gaussian processes. *arXiv preprint arXiv:1705.08933*, 2017.

Alan D Saul, James Hensman, Aki Vehtari, and Neil D Lawrence. Chained Gaussian processes. In *Artificial Intelligence and Statistics*, pages 1431–1440, 2016.

Michalis K Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pages 567–574, 2009.

Dustin Tran, David Blei, and Edo M Airoldi. Copula variational inference. In *Advances in Neural Information Processing Systems*, pages 3564–3572, 2015.

Richard E Turner and Maneesh Sahani. Two problems with variational expectation maximisation for time-series models. *Bayesian Time series models*, pages 115–138, 2011.

Appendix A. Optimal covariance in the additive case

We first define $V(Y, \mu_{\mathcal{F}}, \Sigma_{\mathcal{F}}) = \mathbb{E}_{q(\mathcal{F})} \log p(Y \mid \mathcal{F})$. From Opper and Archambeau (2009), we know that the optimal variational precision is structured as

$$\Sigma_{\mathcal{F}}^{-1} = K_{\mathcal{F}}^{-1} + \nabla_{\Sigma_{\mathcal{F}}} V(Y, \mu_{\mathcal{F}}, \Sigma_{\mathcal{F}}).$$

For factorizing likelihood and additive predictors, and defining $\rho(\cdot) = \sum_c f_c(\cdot)$, we have

$$\begin{aligned} V(Y, \mu_{\mathcal{F}}, \Sigma_{\mathcal{F}}) &= \textstyle\sum_n \mathbb{E}_{q(\rho(x_n))}[\log p(y_n \mid \rho(x_n))] \\ &= \textstyle\sum_n v\big(y_n, \mu_{\rho(x_n)}, \sigma^2_{\rho(x_n)}\big) \end{aligned}$$

where $q(\rho(x_n))$ has variance $\sigma^2_{\rho(x_n)} = 1^\intercal \Sigma_{\mathcal{F}(x_n)} 1$.
The gradient term in the optimal precision thus can be written as

$$\begin{aligned} \nabla_{\Sigma_{\mathcal{F}}} V(Y, \mu_{\mathcal{F}}, \Sigma_{\mathcal{F}}) &= \textstyle\sum_n v'\big(y_n, \mu_{\rho(x_n)}, \sigma^2_{\rho(x_n)}\big) \nabla_{\Sigma_{\mathcal{F}}} \sigma^2_{\rho(x_n)} \\ &= \textstyle\sum_n \lambda_n^2 \sum_{cc'} e_{cn,c'n}, \end{aligned}$$

with $e_{i,j}$ the indicator matrix of size $NC \times NC$ with 1 at location (i, j). Noting $\Lambda = diag(\lambda)$, this can be rewritten in matrix form:

$$\nabla_{\Sigma_{\mathcal{F}}} V(Y, \mu_{\mathcal{F}}, \Sigma_{\mathcal{F}}) = (1 \otimes \Lambda)(1 \otimes \Lambda)^\intercal.$$

Appendix B. ELBO evaluation: additive case

We parameterize the approximate posterior as

$$q(\mathcal{F}) = \mathcal{N}(K_{\mathcal{F}}\, \alpha, \Sigma_{\mathcal{F}})$$

with

$$\Sigma_{\mathcal{F}} = \big(K_{\mathcal{F}}^{-1} + (1 \otimes \Lambda)(1 \otimes \Lambda)^\intercal\big)^{-1},$$

and optimize

$$\mathcal{L}(q) = \mathbb{E}_{q(\sum_c f_c)}[\log p(y \mid \textstyle\sum_c f_c)] - \mathrm{KL}[q(\mathcal{F}) \,\|\, p(\mathcal{F})],$$

where

$$\mathrm{KL}[q(\mathcal{F}) \,\|\, p(\mathcal{F})] = -\frac{1}{2} \log |K_{\mathcal{F}}^{-1} \Sigma_{\mathcal{F}}| + \alpha^\intercal K_{\mathcal{F}} \alpha + \mathrm{tr}[K_{\mathcal{F}}^{-1} \Sigma_{\mathcal{F}}] - NC.$$

B.1. Computing marginals of $\Sigma_{\mathcal{F}^{(n)}}$

$$\begin{aligned}\Sigma_{\mathcal{F}} &= \left(K_{\mathcal{F}}^{-1} + (1\otimes\Lambda)(1\otimes\Lambda)^\intercal\right)^{-1} \\ &= K_{\mathcal{F}} - K_{\mathcal{F}}(1\otimes\Lambda)\big(I + (1\otimes\Lambda)^\intercal K_{\mathcal{F}}(1\otimes\Lambda)\big)^{-1}(1\otimes\Lambda)^\intercal K_{\mathcal{F}} \\ &= K_{\mathcal{F}} - K_{\mathcal{F}}(1\otimes\Lambda)A^{-1}(1\otimes\Lambda)^\intercal K_{\mathcal{F}},\end{aligned}$$

with

$$\begin{aligned}A &= I + (1\otimes\Lambda)^\intercal K_{\mathcal{F}}(1\otimes\Lambda) \\ &= I + \textstyle\sum_c \Lambda^\intercal K_c \Lambda.\end{aligned}$$

where $K_c = K_{f_c,f_c}$.
To evaluate the ELBO we need, for each data point (x_n, y_n), the marginal $q(\sum_c f_c^n)$. This corresponds to the diagonal elements of

$$\begin{aligned}\Sigma_{\text{sum}} =& (1\otimes I)^\intercal \Sigma_{\mathcal{F}}(1\otimes I)^\intercal \\ =& (1\otimes I)^\intercal [K_{\mathcal{F}} - K_{\mathcal{F}}(1\otimes\Lambda)A^{-1}(1\otimes\Lambda)^\intercal K_{\mathcal{F}}](1\otimes I)^\intercal \\ =& \textstyle\sum_c K_c - \sum_{c,c'} K_c \Lambda A^{-1} \Lambda K_{c'} \\ =& \textstyle\sum_c K_c - \sum_{c,c'} (L_A^{-T}\Lambda K_{c'})^T (L_A^{-T}\Lambda K_{c'})\end{aligned}$$

B.2. Computing the KL

$$\begin{aligned}|K_{\mathcal{F}}^{-1}\Sigma_{\mathcal{F}}| &= |K_{\mathcal{F}}^{-1}|/|\Sigma_{\mathcal{F}}^{-1}| \\ &= |K_{\mathcal{F}}^{-1}|/|K_{\mathcal{F}}^{-1} + (1\otimes\Lambda)(1\otimes\Lambda)^\intercal| \\ &= |K_{\mathcal{F}}^{-1}|/[|I + (1\otimes\Lambda)^\intercal K_{\mathcal{F}}(1\otimes\Lambda)||I||K_{\mathcal{F}}^{-1}|] \\ &= 1/|A|\end{aligned}$$

$$\begin{aligned}\operatorname{tr}(K_{\mathcal{F}}^{-1}\Sigma_{\mathcal{F}}) &= \operatorname{tr}(K_{\mathcal{F}}^{-1}(K_{\mathcal{F}} - K_{\mathcal{F}}(1\otimes\Lambda)A^{-1}(1\otimes\Lambda)^\intercal K_{\mathcal{F}})) \\ &= \operatorname{tr}(I - (1\otimes\Lambda)A^{-1}(1\otimes\Lambda)^\intercal K_{\mathcal{F}}) \\ &= NC - \operatorname{tr}(A^{-1}(1\otimes\Lambda)^\intercal K_{\mathcal{F}}) \\ &= NC - \textstyle\sum_c \operatorname{tr}(\Lambda A^{-1}\Lambda^\intercal K_c) \\ &= NC - \operatorname{tr}(\Lambda A^{-1}\Lambda^\intercal \textstyle\sum_c K_c)\end{aligned}$$

In the end,

$$\mathrm{KL}[q(\mathcal{F}) \,\|\, p(\mathcal{F})] = \frac{1}{2}\log|A| + \alpha^\intercal K_{\mathcal{F}}\alpha - \operatorname{tr}(\Lambda A^{-1}\Lambda^\intercal \textstyle\sum_c K_c).$$

B.3. Summary

$$
\begin{aligned}
A &= I + \textstyle\sum_c \Lambda^\intercal K_c \Lambda \\
\mathrm{KL}[q(\mathcal{F}) \,\|\, p(\mathcal{F})] &= \frac{1}{2}\log|A| + \alpha^\intercal K_{\mathcal{F}}\alpha - \mathrm{tr}(\Lambda A^{-1}\Lambda^\intercal \textstyle\sum_c K_c) \\
\mu_{\mathrm{sum}} &= \textstyle\sum_c K_c \alpha_c \\
\Sigma_{\mathrm{sum}} &= \textstyle\sum_c \mathrm{diag}(K_c) - \sum_{c,c'} \mathrm{diag}(K_c \Lambda A^{-1} \Lambda K_{c'})
\end{aligned}
$$

Appendix C. ELBO evaluation: sparse additive case

We parameterize an approximate posterior over the inducing values as

$$q(\mathbf{U}) = \mathcal{N}(K_{\mathbf{U},\mathbf{U}}\alpha, \Sigma_{\mathbf{U},\mathbf{U}})$$

with

$$\Sigma_{\mathbf{U},\mathbf{U}} = (K_{\mathbf{U},\mathbf{U}}^{-1} + BB^\intercal)^{-1},$$

where $B = [B_1, \ldots, B_C] \in \mathbb{R}^{MC \times M}$. We optimize

$$\mathcal{L}(q) = \mathbb{E}_{q(\sum_c f_c)}[\log p(Y \mid \textstyle\sum_c f_c) - \mathrm{KL}[q(\mathbf{U}) \,\|\, p(\mathbf{U})]$$

with

$$\mathrm{KL}[q(\mathbf{U}) \,\|\, p(\mathbf{U})] = -\frac{1}{2}\log|K_{\mathbf{U},\mathbf{U}}^{-1}\Sigma_{\mathbf{U},\mathbf{U}}| + \alpha^\intercal K_{\mathbf{U},\mathbf{U}}\alpha + \mathrm{tr}(K_{\mathbf{U},\mathbf{U}}^{-1}\Sigma_{\mathbf{U},\mathbf{U}}) - MC.$$

C.1. Computing marginals of $\Sigma_{\mathcal{F}}$

We have

$$
\begin{aligned}
\Sigma_{\mathbf{U},\mathbf{U}} &= (K_{\mathbf{U},\mathbf{U}}^{-1} + BB^\intercal)^{-1} \\
&= K_{\mathbf{U},\mathbf{U}} - K_{\mathbf{U},\mathbf{U}}B(I + B^\intercal K_{\mathbf{U},\mathbf{U}}B)^{-1}B^\intercal K_{\mathbf{U},\mathbf{U}} \\
&= K_{\mathbf{U},\mathbf{U}} - K_{\mathbf{U},\mathbf{U}}BA^{-1}B^\intercal K_{\mathbf{U},\mathbf{U}}
\end{aligned}
$$

where $A = I + B^\intercal K_{\mathcal{F}} B$, so

$$
\begin{aligned}
\mu_F &= K_{\mathcal{F}\mathbf{U}}K_{\mathbf{U},\mathbf{U}}^{-1}K_{\mathbf{U},\mathbf{U}}\alpha \\
&= K_{\mathcal{F}\mathbf{U}}\alpha
\end{aligned}
$$

and

$$
\begin{aligned}
\Sigma_{\mathcal{F}} &= K_{\mathcal{F}} - K_{\mathcal{F}\mathbf{U}}(K_{\mathbf{U},\mathbf{U}}^{-1} - K_{\mathbf{U},\mathbf{U}}^{-1}\Sigma_{\mathbf{U},\mathbf{U}}K_{\mathbf{U},\mathbf{U}}^{-1})K_{\mathbf{U}\mathcal{F}} \\
&= K_{\mathcal{F}} - K_{\mathcal{F}\mathbf{U}}(K_{\mathbf{U},\mathbf{U}}^{-1} - K_{\mathbf{U},\mathbf{U}}^{-1}[K_{\mathbf{U},\mathbf{U}} - K_{\mathbf{U},\mathbf{U}}BA^{-1}B^\intercal K_{\mathbf{U},\mathbf{U}}]K_{\mathbf{U},\mathbf{U}}^{-1})K_{\mathbf{U}\mathcal{F}} \\
&= K_{\mathcal{F}} - K_{\mathcal{F}\mathbf{U}}(BA^{-1}B^\intercal)K_{\mathbf{U}\mathcal{F}}.
\end{aligned}
$$

Therefore

$$\begin{aligned}\Sigma_{sum} &= \sum_c K_{f_c} - \sum_{c,c'} K_{f_c\mathbf{U}}(BA^{-1}B^\intercal)K_{\mathbf{U}f_{c'}}. \\ &= \sum_c K_{f_c} - \sum_{c,c'} K_{f_c\mathbf{U}_c}(B_cA^{-1}B_{c'}^\intercal)K_{\mathbf{U}_{c'}f_{c'}}. \\ &= \sum_c K_{f_c} - \sum_{c,c'}(L_A^{-1}B_c^\intercal K_{\mathbf{U}_c f_{c'}})^T(L_A^{-1}B_{c'}^\intercal K_{\mathbf{U}_{c'}f_{c'}}).\end{aligned}$$

The Cholesky decomposition of $A = L_A L_A^T$ is of cost $O(M^3)$. Solving operations $L^{-1}B_c^T$ for each additive term cost a total of $O(CM^3)$. Predicting the marginal predictor variances then costs an extra $O(NC^2M^2)$. In total, the computational cost of posterior predictions are $O(CM^3 + NC^2M^2)$

C.2. Computing the KL

As in the additive case, we have

$$|K_{\mathbf{U},\mathbf{U}}^{-1}\Sigma_{\mathbf{U},\mathbf{U}}| = 1/|A|,$$

and

$$\begin{aligned}\operatorname{tr}(K_{\mathbf{U},\mathbf{U}}^{-1}\Sigma_{\mathbf{U},\mathbf{U}}) &= \operatorname{tr}(K_{\mathbf{U},\mathbf{U}}^{-1}(K_{\mathbf{U},\mathbf{U}} - K_{\mathbf{U},\mathbf{U}}BA^{-1}B^\intercal K_{\mathbf{U},\mathbf{U}})) \\ &= \operatorname{tr}(I - BA^{-1}B^\intercal K_{\mathbf{U},\mathbf{U}}) \\ &= MC - \textstyle\sum_c \operatorname{tr}(B_cA^{-1}B_c^\intercal K_{\mathbf{U}_c,\mathbf{U}_c})\end{aligned}$$

In the end,

$$\mathrm{KL}[q(\mathbf{U}) \,\|\, p(\mathbf{U})] = \frac{1}{2}\log|A| + \alpha^\intercal K_{\mathbf{U},\mathbf{U}}\alpha - \textstyle\sum_c \operatorname{tr}(B_cA^{-1}B_c^\intercal K_{\mathbf{U}_c,\mathbf{U}_c}).$$

C.3. Summary

$$\begin{aligned}A &= I + B^\intercal K_{\mathcal{F}} B \\ \mathrm{KL}[q(\mathbf{U}) \,\|\, p(\mathbf{U})] &= +\frac{1}{2}\log|A| + \alpha^\intercal K_{\mathbf{U},\mathbf{U}}\alpha - \textstyle\sum_c \operatorname{tr}(B_cA^{-1}B_c^\intercal K_{\mathbf{U}_c,\mathbf{U}_c}) \\ \mu_{\text{sum}} &= \textstyle\sum_c K_{f_c,\mathbf{U}_c}\alpha_c \\ \Sigma_{\text{sum}} &= \textstyle\sum_c \operatorname{diag}(K_{f_c,f_c}) - \sum_{c,c'} \operatorname{diag}(K_{f_c,\mathbf{U}_c}B_cA^{-1}B_{c'}^\intercal K_{\mathbf{U}_{c'},f_{c'}})\end{aligned}$$