Supplementary material for: Heteroskedastic Gaussian processes for uncertainty modeling in large-scale crowdsourced traffic data

Filipe Rodrigues, Francisco Pereira

1 Variational inference

Approximate inference in the FC-HGP is based on the variational approximation proposed in [1]. We provide here an overview of the inference algorithm.

As with standard variational approximations, we aim at finding a variational distribution $q(\mathbf{f},\mathbf{g})$ that minimizes the Kullback-Leibler (KL) divergence to the true posterior, $\mathbb{KL}(q(\mathbf{f},\mathbf{g})||p(\mathbf{f},\mathbf{g}|\mathbf{y}))$. Assuming a factorized variational distribution of the form $q(\mathbf{f},\mathbf{g}) = q(\mathbf{f})\, q(\mathbf{g})$, we can write

$$\begin{aligned}\mathbb{KL}(q(\mathbf{f})\, q(\mathbf{g})||p(\mathbf{f},\mathbf{g}|\mathbf{y})) &= \mathbb{E}_q\left[\log \frac{q(\mathbf{f})\, q(\mathbf{g})}{p(\mathbf{f},\mathbf{g}|\mathbf{y})}\right] \\ &= \mathbb{E}_q[\log q(\mathbf{f})] + \mathbb{E}_q[\log q(\mathbf{g})] \\ &\quad - \mathbb{E}_q[\log p(\mathbf{y},\mathbf{f},\mathbf{g})] + \mathbb{E}_q[\log p(\mathbf{y})]\end{aligned}$$

Defining $\mathcal{L}(q) \triangleq \mathbb{E}_q[\log p(\mathbf{y},\mathbf{f},\mathbf{g})] - \mathbb{E}_q[\log q(\mathbf{f})] - \mathbb{E}_q[\log q(\mathbf{g})]$ and re-arranging yields

$$\mathcal{L}(q) = \log p(\mathbf{y}) - \mathbb{KL}(q(\mathbf{f})\, q(\mathbf{g})||p(\mathbf{f},\mathbf{g}|\mathbf{y})).$$

Since the KL divergence is always non-negative, it becomes clear that $\mathcal{L}(q)$ lower bounds the (log) marginal likelihood of the data, i.e. $\log p(\mathbf{y}) \geqslant \mathcal{L}(q)$. Minimizing he KL divergence in then equivalent to maximizing $\mathcal{L}(q)$.

In its current form $\mathcal{L}(q)$ depends on two T-dimensional variational distributions: $q(\mathbf{f})$ and $q(\mathbf{g})$. We can obtain a simpler, tighter bound, by optimally removing the dependency on $q(\mathbf{f})$. According to the variational Bayesian theory, the optimal distribution $q^*(\mathbf{f})$ is given by [2]

$$q^*(\mathbf{f}) = \arg\max_{q(\mathbf{f})} \mathcal{L}(q) = \frac{p(\mathbf{f})}{Z(q(\mathbf{g}))} e^{\int q(\mathbf{g}) \log p(\mathbf{y}|\mathbf{f},\mathbf{g})\, d\mathbf{g}}, \tag{1}$$

where $Z(q(\mathbf{g})) = \int e^{\int q(\mathbf{g}) \log p(\mathbf{y}|\mathbf{f},\mathbf{g})\, d\mathbf{g}}\, p(\mathbf{f})\, d\mathbf{f}$ is a normalization constant needed to ensure that $q^*(\mathbf{f})$ integrates to one. Plugging $q^*(\mathbf{f})$ back into the bound $\mathcal{L}(q)$

and performing some simplifications we obtain a marginalized variational lower bound given by

$$\mathcal{L}(q) = \log Z(q(\mathbf{g})) - \mathbb{KL}(q(\mathbf{g})||p(\mathbf{g})).$$

Restricting $q(\mathbf{g})$ to be a multivariate normal distribution, such that $q(\mathbf{g}) = \mathcal{N}(\mathbf{g}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we have that

$$\begin{aligned}\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log \int e^{\int \mathcal{N}(\mathbf{g}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \log p(\mathbf{y}|\mathbf{f},\mathbf{g})\, d\mathbf{g}} \mathcal{N}(\mathbf{f}|\mathbf{0}_T, \mathbf{K}_f) d\mathbf{f} \\ - \mathbb{KL}(\mathcal{N}(\mathbf{g}|\boldsymbol{\mu}, \boldsymbol{\Sigma})||\mathcal{N}(\mathbf{g}|\mu_0 \mathbf{1}_T, \mathbf{K}_g)),\end{aligned} \tag{2}$$

where we made the dependency of the lower bound $\mathcal{L}$ on the variational parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ explicit and, similarly to $\mathbf{K}_f$, the matrix $\mathbf{K}_g$ is used to denote the covariance function $k_g(\{x_{t-1}, ..., x_{t-L}\}, \{x'_{t-1}, ..., x'_{t-L}\})$ evaluated between every pair of training inputs with the relative flow information for the noise process.

The first term in (2) can be computed by noticing that

$$\int \mathcal{N}(\mathbf{g}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \log p(\mathbf{y}|\mathbf{f}, \mathbf{g})\, d\mathbf{g} = \log \mathcal{N}(\mathbf{y}|\mathbf{f}, \mathbf{R}) - \frac{1}{4}\text{tr}(\boldsymbol{\Sigma}), \tag{3}$$

where $\mathbf{R}$ is a diagonal matrix with elements $[\mathbf{R}]_{ii} = e^{[\boldsymbol{\mu}]_i - [\boldsymbol{\Sigma}]_{ii}/2}$ and $\text{tr}(\boldsymbol{\Sigma})$ denotes the trace of $\boldsymbol{\Sigma}$. Making use of the fact that

$$\int \mathcal{N}(\mathbf{y}|\mathbf{f}, \mathbf{R}) \mathcal{N}(\mathbf{f}|\mathbf{0}_T, \mathbf{K}_f)\, d\mathbf{f} = \mathcal{N}(\mathbf{y}|\mathbf{0}_T, \mathbf{K}_f + \mathbf{R}),$$

we obtain an analytical expression for the marginalized variational bound, given by

$$\begin{aligned}\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log \mathcal{N}(\mathbf{y}|\mathbf{0}_T, \mathbf{K}_f + \mathbf{R}) - \frac{1}{4}\text{tr}(\boldsymbol{\Sigma}) \\ - \mathbb{KL}(\mathcal{N}(\mathbf{g}|\boldsymbol{\mu}, \boldsymbol{\Sigma})||\mathcal{N}(\mathbf{g}|\mu_0 \mathbf{1}_T, \mathbf{K}_g)),\end{aligned} \tag{4}$$

where the KL divergence between two multivariate normal distributions is given by [3]

$$\begin{aligned}\mathbb{KL}(\mathcal{N}(\mathbf{g}|\boldsymbol{\mu}, \boldsymbol{\Sigma})||\mathcal{N}(\mathbf{g}|\mu_0 \mathbf{1}_T, \mathbf{K}_g)) = \frac{1}{2} \log \frac{|\mathbf{K}_g|}{|\boldsymbol{\Sigma}^{-1}|} \\ + \frac{1}{2}\text{tr}(\mathbf{K}_g^{-1}\boldsymbol{\Sigma}) + \frac{1}{2}(\mu_0 \mathbf{1}_T - \boldsymbol{\mu})^{\text{T}} \mathbf{K}_g^{-1} (\mu_0 \mathbf{1}_T - \boldsymbol{\mu}).\end{aligned}$$

By finding the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ that maximize the bound in (4), we are simultaneously finding the variational distribution $\mathcal{N}(\mathbf{g}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ that is closest to the true posterior. Since the optimization of $\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is non-linear, a conjugate gradients procedure is used. It is important to note that the bound in (4) can also be used to optimize the hyper-parameters of covariance functions k_f and

k_g, thereby implementing type-II maximum likelihood for model selection. In practice, in order to simplify this optimization problem and reduce the computational complexity, we follow the reparametrization procedure proposed in [1] based on [4], which defines

$$\boldsymbol{\mu} = \mathbf{K}_g\left(\boldsymbol{\Lambda} - \frac{1}{2}\mathbf{I}_T\right)\mathbf{1}_T + \mu_0\mathbf{1}_T$$
$$\boldsymbol{\Sigma} = (\mathbf{K}_g^{-1} + \boldsymbol{\Lambda})^{-1}, \tag{5}$$

for some diagonal matrix $\boldsymbol{\Lambda}$, thereby effectively reducing the number of parameters from $T + T(T+1)/2$ to T.

Lastly, we can obtain an analytical expression for $q^*(\mathbf{f})$ by making use of (3) in (1) to give

$$\begin{aligned} q^*(\mathbf{f}) &\propto \mathcal{N}(\mathbf{y}|\mathbf{f},\mathbf{R})\,\mathcal{N}(\mathbf{f}|\mathbf{0}_T,\mathbf{K}_f) \\ &= \mathcal{N}(\mathbf{f}|\mathbf{K}_f\boldsymbol{\alpha}, \mathbf{K}_f - \mathbf{K}_f(\mathbf{K}_f+\mathbf{R})^{-1}\mathbf{K}_f), \end{aligned} \tag{6}$$

where we defined $\boldsymbol{\alpha} \triangleq (\mathbf{K}_f + \mathbf{R})^{-1}\mathbf{y}$.

To make predictions for a new unobserved time t_*, we begin by making use of (**??**) and (6) to compute the posterior distribution of f_*

$$\begin{aligned} q(f_*) &= \int p(f_*|\mathbf{f})\, q^*(\mathbf{f})\, d\mathbf{f} \\ &= \int \mathcal{N}(f_*|\mathbf{k}_{f*}^{\mathrm{T}}\mathbf{K}_f^{-1}\mathbf{f},\, k_{f**} - \mathbf{k}_{f*}^{\mathrm{T}}\mathbf{K}_f^{-1}\mathbf{k}_{f*}) \\ &\quad \mathcal{N}(\mathbf{f}|\mathbf{K}_f\boldsymbol{\alpha}, \mathbf{K}_f - \mathbf{K}_f(\mathbf{K}_f+\mathbf{R})^{-1}\mathbf{K}_f)\, d\mathbf{f} \\ &= \mathcal{N}(f_*|a_*, b_*), \end{aligned}$$

where $a_* \triangleq \mathbf{k}_{f*}^{\mathrm{T}}\boldsymbol{\alpha}$ and $b_* \triangleq k_{f**} - \mathbf{k}_{f*}^{\mathrm{T}}(\mathbf{K}_f+\mathbf{R})^{-1}\mathbf{k}_{f*}$. Similarly, for the posterior of g_*, following the reparametrization in (5), we have that

$$\begin{aligned} q(g_*) &= \int p(g_*|\mathbf{g})\, q(\mathbf{g})\, d\mathbf{g} \\ &= \int \mathcal{N}(g_*|\mathbf{k}_{g*}^{\mathrm{T}}\mathbf{K}_g^{-1}\mathbf{g},\, k_{g**} - \mathbf{k}_{g*}^{\mathrm{T}}\mathbf{K}_g^{-1}\mathbf{k}_{g*}) \\ &\quad \mathcal{N}(\mathbf{g}|\mathbf{K}_g(\boldsymbol{\Lambda} - \tfrac{1}{2}\mathbf{I}_T)\mathbf{1}_T + \mu_0\mathbf{1}_T, (\mathbf{K}_g^{-1}+\boldsymbol{\Lambda})^{-1})\, d\mathbf{g} \\ &= \mathcal{N}(g_*|c_*, d_*), \end{aligned}$$

with $c_* \triangleq \mathbf{k}_{g*}^{\mathrm{T}}(\boldsymbol{\Lambda} - \frac{1}{2}\mathbf{I}_T)\mathbf{1}_T + \mu_0$ and $d_* \triangleq k_{g**} - \mathbf{k}_{g*}^{\mathrm{T}}(\mathbf{K}_g + \boldsymbol{\Lambda}^{-1})^{-1}\mathbf{k}_{g*}$, and where we made use of the Woodbury matrix identity. Finally, the predictive distribution for an unobserved time t is given by

$$\begin{aligned} q(y_*) &= \int \mathcal{N}(y_*|f_*, e^{g_*})\, q(f_*)\, q(g_*)\, df_*\, dg_* \\ &= \int \mathcal{N}(y_*|a_*, b_* + e^{g_*})\,\mathcal{N}(g_*|c_*, d_*)\, dg_*. \end{aligned} \tag{7}$$

Although this distribution is not Gaussian, we can obtain analytical expressions for its mean and variance, which are given by $\mathbb{E}_q[y_*] = a_*$ and $\mathbb{V}_q[y_*] = b_* + e^{c_* + d_*/2}$, respectively.

References

References

[1] M. K. Titsias, M. Lázaro-Gredilla, Variational heteroscedastic gaussian process regression, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 841–848.

[2] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[3] C. E. Rasmussen, C. Williams, Gaussian processes for machine learning, The MIT Press, 2005.

[4] M. Opper, C. Archambeau, The variational gaussian approximation revisited, Neural computation 21 (3) (2009) 786–792.