Supplementary Material for "Robust and Scalable Gaussian Process Regression and Its Applications"

Yifan Lu[1], Jiayi Ma[1]*, Leyuan Fang[2], Xin Tian[1], and Junjun Jiang[3]
[1] Wuhan University, China [2] Hunan University, China [3] Harbin Institute of Technology, China
{lyf048, xin.tian}@whu.edu.cn, {jyma2010, fangleyuan}@gmail.com, jiangjunjun@hit.edu.cn

In the following supplementary material, we first present additional analysis of our model experimentally in Sect. 1. Sect. 2 provides more detailed results. Then, in Sect. 3, we visualize how the SVI in our model works with large-scale data. Following this, more experimental details and results of feature matching are provided in Sect. 4. Finally, Sect. 5 gives the detailed derivations of the formulas in our model.

1. Additional Analyses

1.1. Number of Inducing Variables

We change the number of inducing variables and measure the moments-matching divergence $\mathrm{KL}(p(\mathbf{f}_*|\mathcal{D})||q(\mathbf{f}_*))$ between the true test posterior $p(\mathbf{f}_*|\mathcal{D})$ and the approximate test posterior $q(\mathbf{f}_*)$ on Neal data with 100 inliers and 100 outliers. Fig. S1 shows the KL divergence as the number of inducing variables increases. As we can see, our model is able to match the true exact GP model when the number of inducing variables is more than 15. This is because our model akin to the work of VFE [7] approximates the posterior rigiously, while maintaining high robustness to outliers. This provides a great convenience for speeding up our model without losing regression performance.

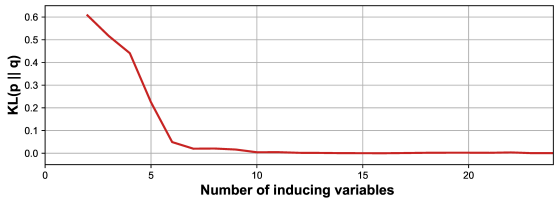

Figure S1. KL divergence between the true test posterior and the approximate test posterior of our model as the number of inducing variables increases.

1.2. Impact of Outlier Ratio

To further measure the robustness of our method to outliers, we fix the inlier number to 100 with varing outlier ratios, and measure the KL divergence mentioned above. The baseline GPR-VFE [7] is included for comparison and the results are reported in Fig. S2, where the median and standard error bars are calculated by repeating the experiment 10 times. From the results, we find that our model almost always gives a perfect match to the true posterior when the outlier ratios are below 60%. The overfitting of our model starts to emerge slightly only when the outlier ratios reach 80%. In such high outlier ratio case, numerous outliers affect the posterior variance of our model, thus judging some outliers that agree with the posterior mean as inliers. We note that when facing massive outliers (*e.g.*, 80%), the posterior mean provided by our model is still accurate, and only the posterior variance is slightly affected by the outliers.

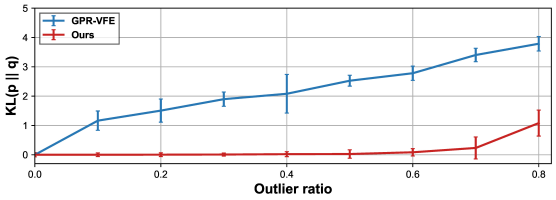

Figure S2. KL divergence as the outlier ratio grows for our model and GPR-VFE.

1.3. Non-uniform Outliers

We inject mixture Gaussian, independent Gaussian, and two types of structured outliers on Neal (Fig. 7 in the main paper), and show the mean and standard deviation in Tab. S1.

2. Detailed Results

We provide a more detailed results on Neal data with different outlier ratios in Tab. S2, where the standard deviation and KL divergence between the true test posterior and the approximate test posterior are reported. As we can see, our method always has the best results. This superiority is even more evident on KL, as our method fits the posterior mean

*Corresponding Author

Table S1. Results of injecting different non-uniform outliers. **Bold** indicates the best.

Method	Mixture Gaussian		Independent Gaussian	
	MAE	KL	MAE	KL
GPR-VFE	0.467±0.122	7.683±4.695	0.971±0.082	2.403±0.172
GPR-St	0.278±0.619	6.839±50.594	0.248±0.487	0.512±0.184
GPR-Lap	0.082±0.048	3.827±4.145	0.314±0.167	1.674±0.214
GPR-MEM	0.068±0.057	6.420±1.232	0.918±0.323	2.334±0.656
Ours	**0.027±0.022**	**0.068±0.141**	**0.026±0.009**	**0.027±0.020**
Method	Structured #1		Structured #2	
	MAE	KL	MAE	KL
GPR-VFE	1.099±0.403	457.143±416.136	1.140±0.059	3.356±0.316
GPR-St	0.176±0.445	28.174±78.706	0.479±0.647	0.543±0.216
GPR-Lap	0.164±0.170	304.562±314.627	0.392±0.167	2.114±0.564
GPR-MEM	0.098±0.080	5.458±1.546	1.172±0.072	3.245±0.335
Ours	**0.023±0.006**	**0.022±0.022**	**0.021±0.005**	**0.018±0.006**

Table S2. Detailed results on Neal data with different outlier ratios. **Bold** indicates the best.

Method	Neal					
	10%		50%		80%	
	MAE	KL	MAE	KL	MAE	KL
GPR-VFE	0.189±0.210	1.396±0.750	0.583±0.089	2.347±0.204	0.757±0.113	3.867±0.273
GPR-St	0.036±0.080	0.741±0.354	0.092±0.799	1.508±1.159	0.773±0.101	4.077±0.443
GPR-Lap	0.031±0.009	6.770±15.891	0.426±0.019	1.730±0.200	0.350±0.043	3.945±0.852
GPR-MEM	0.043±0.057	2.030±1.308	0.064±0.010	7.227±2.286	0.774±0.086	4.057±0.306
Ours	**0.012±0.008**	**0.001±0.138**	**0.026±0.007**	**0.231±0.169**	**0.032±0.012**	**0.753±0.700**

and variance very well.

3. Visual Illustration of SVI

Stochastic variational inference (SVI) can dramatically improve the speed of our model without losing much performance. Fig. S3 visualize the intermediate results of the optimization process on Neal data, where the top group contains 1667 data points with 40% outliers and the bottom contains 3334 data points with 70% outliers. In each plot of the figure showing the optimization process, we only show the currently processed mini-batch data. We also visualize the field of p, which indicates the inlier probability at any location. The deeper the blue color, the higher the probability. With SVI, we see our model gradually converges to the correct result. It enables our model embrace large-scale data as it is easier to calculate the natural gradient for mini batch than traversing the whole training data.

4. Details of Feature Matching

4.1. Datasets

YFCC100M: The Yahoo's YFCC100M dataset [6] collected 100 million photos from Internet and was later organized into 72 scenes reconstructed with the Structure from Motion software VisualSfM, providing bundle adjusted camera poses, intrinsics and triangulated point clouds. Following [9], 4 sequences (*i.e.*, Buckingham palace, Sacre coeur, Reichstag, and Notre dame front facade) are set as our test set, which contains 4000 image pairs in total. [3] is used to recover the camera poses and generate ground truth.

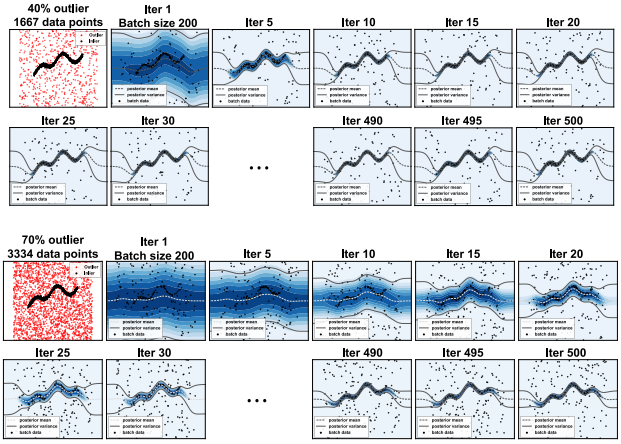

Figure S3. Illustration of SVI in our model for robust Gaussian process regression. The training data in the top group includes 1667 data points with 40% outliers and the bottom contains 3334 data points with 70% outliers. The field of p is also visualized and represented by the shades of blue.

HPatches: The HPatches benchmark contains 116 scenes with 696 unique pictures, where the first 57 scenes are taken under different illumination and the remaining 59 scenes undergo viewpoint changes. Each scene contains one reference image and five target images with ground-truth homography provided for each target image. The SIFT [4] and HardNet [5] are used to detect feature points and generate descriptors, respectively.

CPC: The Community Photo Collection (CPC) dataset [8] includes unstructured images of landmarks collected from Flicker, where the image pairs are wide-baseline with different resolutions. 1000 image pairs are selected from the dataset for testing with ground-truth fundamental matrix available.

4.2. Evaluation Metrics

Precision is defined as the ratio of the true inliers among those preserved "inlier" by a matching algorithm. Recall is defined as the percentage of preserved true inliers among the whole inliers contained in the original putative set. And F1-score is defined as the ratio of $2 \times \text{Precision} \times \text{Recall}$ and $\text{Precision} + \text{Recall}$.

Normalized symmetric geometry distance (NSGD) is computed as the SGD (in pixels) divided by the length of image diagonals [1], where SGD compares the estimated fundamental matrix with the ground-truth fundamental matrix by iteratively generating points on the borders of the images, and then measuring the epipolar distances. The *homography error* metric defined in SuperPoint [2] compares the estimated homography with the ground-truth homography using the corners of images.

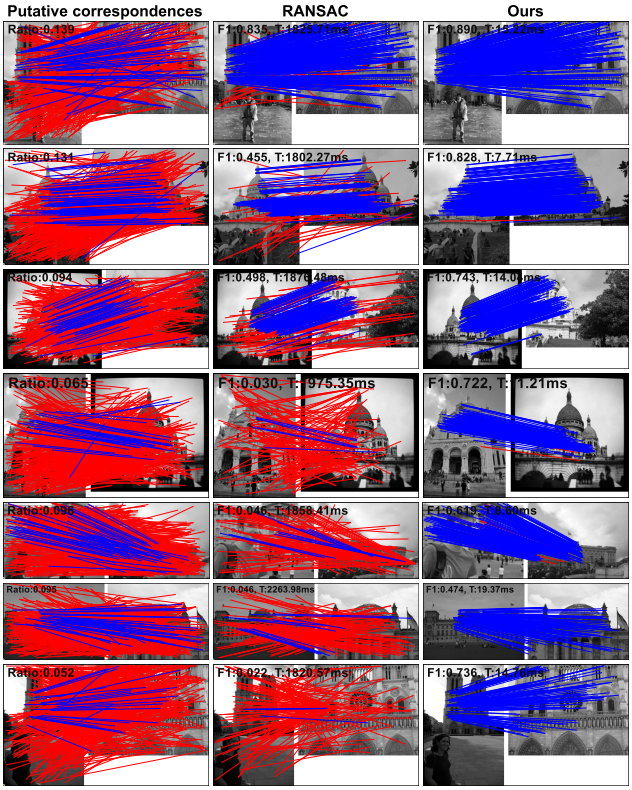

Figure S4. Additional qualitative illustration of our model on feature matching. The test scenarios come from YFCC100M. We show the true positive in blue and false positive in red. For visibility, in the image pairs, at most 500 randomly selected matches are presented. Best viewed in color.

4.3. Details of Incorporating Ratio Information

The ratio information is a by-product of nearest neighbor matching. It is the ratio of the descriptors distance of the nearest to the second nearest neighbor. If the ratio is small, it means that the correspondence is well differentiated, and it is more likely to be an inlier, and vice versa. Previous works simply use a user-defined threshold to pre-reject some outliers, called "ratio test", which significantly reduces the proportion of outliers but also removes many inliers. Here we integrate the ratio scores into our model as a prior which is expressed as $p(z_i = 1) = \gamma \cdot \max\{1, \frac{1}{r_i} - 0.5\}$, where $r_i \in [0, 1]$ denotes the ratio scores. Compared to simple ratio test, our incorporation of ratio information is more theoretically sound and does not lose inliers.

4.4. More Qualitative Illustration

Fig. S4 shows additional qualitative illustrations of our model in default setting on YFCC100M dataset with RANSAC method as a baseline competitor. The F1-score and running time are also reported for each scenario. As we can see, our model takes only a few milliseconds and has almost no mismatches. It demonstrates extremely high speed and robustness.

5. Derivation Details

5.1. Preliminaries of Gaussian Integrals

Two Gaussians

$$\int_{\mathbb{R}^k} \mathcal{N}(\mathbf{x}|\mathbf{a}, \mathbf{A})\mathcal{N}(\mathbf{a}|\mathbf{b}, \mathbf{B})d\mathbf{a} = \mathcal{N}(\mathbf{x}|\mathbf{b}, \mathbf{A} + \mathbf{B}) \quad \text{(S1)}$$

Quadratic forms

$$\int_{\mathbb{R}^k} (\mathbf{x} - \mathbf{c})^\top \mathbf{A}(\mathbf{x} - \mathbf{c})\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})d\mathbf{x} = (\boldsymbol{\mu} - \mathbf{c})^\top \mathbf{A}(\boldsymbol{\mu} - \mathbf{c}) + \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}) \quad \text{(S2)}$$

5.2. Derivation on q_1

Given q_2, q_3, and q_4, and denote $p_i = \mathbb{E}[z_i]$ and $\mathbf{P} = \mathrm{diag}(p_1, p_2, \cdots, p_n)$. Eq. (11) turns to

$$\begin{aligned}
\mathcal{L} &= \int q_1 q_{2-4} \log\left(p(\mathbf{f}_m)p(\mathbf{y}, \mathbf{Z}|\mathbf{f})\right) d\mathbf{f}d\mathbf{f}_m d\boldsymbol{\theta} \\
&\quad - \int q_1 \log \phi(\mathbf{f}_m) d\mathbf{f}d\mathbf{f}_m + \text{const.} \\
&= \int p(\mathbf{f}|\mathbf{f}_m)\phi(\mathbf{f}_m)\mathbb{E}_{q_{2-4}}[\log\left(p(\mathbf{f}_m)p(\mathbf{y}, \mathbf{Z}|\mathbf{f})\right)]d\mathbf{f}d\mathbf{f}_m \\
&\quad - \int p(\mathbf{f}|\mathbf{f}_m)\phi(\mathbf{f}_m) \log \phi(\mathbf{f}_m) d\mathbf{f}d\mathbf{f}_m + \text{const.}
\end{aligned} \quad \text{(S3)}$$

Focus on the term $\mathbb{E}_{q_{2-4}}[\log\left(p(\mathbf{f}_m)p(\mathbf{y}, \mathbf{Z}|\mathbf{f})\right)]$, we obtain

$$\begin{aligned}
&\mathbb{E}_{q_{2-4}}[\log\left(p(\mathbf{f}_m)p(\mathbf{y}, \mathbf{Z}|\mathbf{f})\right)] = \log p(\mathbf{f}_m|\mathbf{X}_m, \boldsymbol{\varphi}) \\
&+ \sum_{i=1}^{n} \left((1 - p_i) \log \frac{\langle 1 - \gamma \rangle}{a} + p_i \log\left(\langle \gamma \rangle \mathcal{N}(y_i|f_i, \sigma^2)\right) \right) \\
&= \log \mathcal{N}(\mathbf{f}_m|\mathbf{0}, \mathbf{K}_{mm}) + \sum_{i=1}^{n} p_i \log \mathcal{N}(y_i|f_i, \sigma^2) + \text{const.},
\end{aligned}$$

where $\langle \gamma \rangle = \exp(\mathbb{E}[\log \gamma])$ and $\langle 1 - \gamma \rangle = \exp(\mathbb{E}[\log(1 - \gamma)])$. Next, we integrate over $\mathbf{f}$, Eq. (S3) becomes

$$\begin{aligned}
\mathcal{L} &= \int \phi(\mathbf{f}_m) \int p(\mathbf{f}|\mathbf{f}_m) \left(\sum_{i=1}^{n} p_i \log \mathcal{N}(y_i|f_i, \sigma^2) \right) d\mathbf{f}d\mathbf{f}_m \\
&\quad + \int \phi(\mathbf{f}_m) \log \mathcal{N}(\mathbf{f}_m|\mathbf{0}, \mathbf{K}_{mm}) d\mathbf{f}_m \\
&\quad - \int \phi(\mathbf{f}_m) \log \phi(\mathbf{f}_m) d\mathbf{f}_m + \text{const.},
\end{aligned} \quad \text{(S4)}$$

where

$$\begin{aligned}
&\int p(\mathbf{f}|\mathbf{f}_m)\left(\sum_{i=1}^{n} p_i \log \mathcal{N}(y_i|f_i,\sigma^2)\right) d\mathbf{f} \\
\sim &\int \mathcal{N}(\mathbf{f}|\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{f}_m, \mathbf{K}_{\mathbf{f}|\mathbf{f}_m})\left(-\frac{1}{2}(\mathbf{f}-\mathbf{y})^\top \sigma^{-2}\mathbf{P}(\mathbf{f}-\mathbf{y})\right) \\
\sim &-\frac{1}{2}\left(\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{f}_m-\mathbf{y}\right)^\top \sigma^{-2}\mathbf{P}\left(\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{f}_m-\mathbf{y}\right) \\
&-\frac{1}{2\sigma^2}\mathrm{tr}(\mathbf{P}\mathbf{K}_{\mathbf{f}|\mathbf{f}_m}) \\
\sim &\log \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{f}|\mathbf{f}_m}, \sigma^2\mathbf{P}^{-1}) - \frac{1}{2\sigma^2}\mathrm{tr}(\mathbf{P}\mathbf{K}_{\mathbf{f}|\mathbf{f}_m}) := \log Q(\mathbf{f}_m, \mathbf{y}).
\end{aligned}$$

Note that we use $\sim$ to represent equality plus an irrelevant constant term. Eq. (S4) then turns to

$$\mathcal{L} = \int \phi(\mathbf{f}_m) \log\left(\frac{p(\mathbf{f}_m)Q(\mathbf{f}_m,\mathbf{y})}{\phi(\mathbf{f}_m)}\right) d\mathbf{f}_m + \text{const.},$$

which corresponds to the Eq. (12) in the main paper. Thus, the optimal $\hat{\phi}(\mathbf{f}_m)$ is given as follows

$$\begin{aligned}
&\log\hat{\phi}(\mathbf{f}_m) \propto \log\left(p(\mathbf{f}_m)Q(\mathbf{f}_m,\mathbf{y})\right) \\
&\sim \log\mathcal{N}(\mathbf{f}_m|\mathbf{0},\mathbf{K}_{mm}) + \log\mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{f}|\mathbf{f}_m},\sigma^2\mathbf{P}^{-1}) \\
&\sim -\frac{1}{2}\left\{\mathbf{f}_m^\top\mathbf{K}_{mm}^{-1}\mathbf{f}_m + (\mathbf{y}-\boldsymbol{\mu}_{\mathbf{f}|\mathbf{f}_m})^\top\sigma^{-2}\mathbf{P}(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{f}|\mathbf{f}_m})\right\} \\
&\sim -\frac{1}{2}\left(\mathbf{f}_m-\sigma^{-2}\mathbf{A}_m\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}\mathbf{P}\mathbf{y}\right)^\top\mathbf{A}_m^{-1} \\
&\quad\left(\mathbf{f}_m-\sigma^{-2}\mathbf{A}_m\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}\mathbf{P}\mathbf{y}\right) \\
&\sim \log\mathcal{N}(\mathbf{f}_m|\sigma^{-2}\mathbf{A}_m\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}\mathbf{P}\mathbf{y},\mathbf{A}_m),
\end{aligned}$$

where $\mathbf{A}_m = \left(\mathbf{K}_{mm}^{-1}+\sigma^{-2}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}\mathbf{P}\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\right)^{-1}$ and can be simplified to $\mathbf{K}_{mm}\boldsymbol{\Sigma}\mathbf{K}_{mm}$, and $\boldsymbol{\Sigma} = (\mathbf{K}_{mm}+\sigma^{-2}\mathbf{K}_{mn}\mathbf{P}\mathbf{K}_{nm})^{-1}$. Thus, the optimal $\hat{\phi}(\mathbf{f}_m)$ is a multivariate Gaussian distribution

$$\hat{\phi}(\mathbf{f}_m) \sim \mathcal{N}(\mathbf{f}_m|\boldsymbol{\mu}_m,\mathbf{A}_m),$$

where $\boldsymbol{\mu}_m = \sigma^{-2}\mathbf{K}_{mm}\boldsymbol{\Sigma}\mathbf{K}_{mn}\mathbf{P}\mathbf{y}$, which corresponds to the Eq. (14) in the main paper.

After obtaining the optimal $\hat{\phi}(\mathbf{f}_m)$, we can recover $q(\mathbf{f})$ by marginalizing out $\mathbf{f}_m$ using Eq. (S2):

$$\begin{aligned}
q(\mathbf{f}) &\sim \int p(\mathbf{f}|\mathbf{f}_m)\phi(\mathbf{f}_m)d\mathbf{f}_m \\
&\sim \int \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}_{\mathbf{f}|\mathbf{f}_m},\mathbf{K}_{\mathbf{f}|\mathbf{f}_m})\mathcal{N}(\mathbf{f}_m|\boldsymbol{\mu}_m,\mathbf{A}_m)d\mathbf{f}_m \\
&\sim \int \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}_{\mathbf{f}|\mathbf{f}_m},\mathbf{K}_{\mathbf{f}|\mathbf{f}_m})\mathcal{N}(\boldsymbol{\mu}_{\mathbf{f}|\mathbf{f}_m}|\boldsymbol{\mu}_{\mathbf{f}},\mathbf{K}_{nm}\boldsymbol{\Sigma}\mathbf{K}_{mn})d\boldsymbol{\mu}_{\mathbf{f}|\mathbf{f}_m} \\
&= \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}_{\mathbf{f}},\mathbf{A}),
\end{aligned}$$

where $\boldsymbol{\mu}_{\mathbf{f}} = \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\boldsymbol{\mu}_m$ and $\mathbf{A} = \mathbf{K}_{\mathbf{f}|\mathbf{f}_m} + \mathbf{K}_{nm}\boldsymbol{\Sigma}\mathbf{K}_{mn}$.

5.3. Derivation on Remark 1

With the optimal $\hat{\phi}(\mathbf{f}_m)$, we expand $\mathbb{E}[\log\mathcal{N}(\mathbf{y}|\mathbf{f},\sigma^2\mathbf{I})]$ and as follows

$$\begin{aligned}
&\mathbb{E}[\log\mathcal{N}(\mathbf{y}|\mathbf{f},\sigma^2\mathbf{I})] \\
&= \int p(\mathbf{f}|\mathbf{f}_m)\hat{\phi}(\mathbf{f}_m)\log\mathcal{N}(\mathbf{y}|\mathbf{f},\sigma^2\mathbf{I})d\mathbf{f}d\mathbf{f}_m \\
&= \int \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}_{\mathbf{f}},\mathbf{A})\log\mathcal{N}(\mathbf{y}|\mathbf{f},\sigma^2\mathbf{I})d\mathbf{f} \\
&= -\frac{1}{2}\log 2\pi - \frac{1}{2}\log|\sigma^2\mathbf{I}| \\
&\quad -\frac{1}{2\sigma^2}\int\mathcal{N}(\mathbf{f}|\boldsymbol{\mu}_{\mathbf{f}},\mathbf{A})(\mathbf{y}-\mathbf{f})^\top(\mathbf{y}-\mathbf{f})d\mathbf{f}.
\end{aligned}$$

Utilize Eq. (S2), it becomes

$$\begin{aligned}
&\mathbb{E}[\log\mathcal{N}(\mathbf{y}|\mathbf{f},\sigma^2\mathbf{I})] \\
&= -\frac{1}{2}\log 2\pi - \frac{1}{2}\log|\sigma^2\mathbf{I}| \\
&\quad -\frac{1}{2\sigma^2}(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{f}})^\top(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{f}}) - \frac{1}{2\sigma^2}\mathrm{tr}(\mathbf{A}) \\
&= \log\mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{f}},\sigma^2\mathbf{I}) - \frac{1}{2\sigma^2}\mathrm{tr}(\mathbf{A}) \\
&= \sum_{i=1}^{n}\left(\log\mathcal{N}(y_i|\boldsymbol{\mu}_{\mathbf{f}i},\sigma^2) - \frac{1}{2\sigma^2}\mathbf{A}_{ii}\right) \\
&= \sum_{i=1}^{n}\mathbb{E}[\log\mathcal{N}(y_i|f_i,\sigma^2)].
\end{aligned}$$

Therefore, we obtain

$$\langle\mathcal{N}_i\rangle = \mathcal{N}(y_i|\boldsymbol{\mu}_{\mathbf{f}i},\sigma^2)\exp(-\frac{1}{2\sigma^2}\mathbf{A}_{ii}),$$

which recovers the Eq. (15) in the main paper.

5.4. Derivation on q_2

Suppose q_1, q_3, and q_4 are given and note $\hat{n} = \mathrm{tr}(\mathbf{P})$. Using Eq. (10) and focusing on the terms involving γ, we obtain

$$\begin{aligned}
\log\hat{q}_2(\gamma) &\sim \log p(\gamma) + \sum_{i=1}^{n}\left((1-p_i)\log(1-\gamma) + p_i\log\gamma\right) \\
&\sim \log\gamma^{B_a+\hat{n}-1} + \log(1-\gamma)^{B_b+n-\hat{n}-1}.
\end{aligned}$$

Taking the exponential of both sides and normalizing, the $\hat{q}_2(\gamma)$ follows a Beta distribution:

$$\hat{q}_2(\gamma) = \mathrm{Beta}(\gamma|B_a+\hat{n}, B_b+n-\hat{n}).$$

5.5. Derivation on q_3

Given q_1, q_2, and q_4, according to Eq. (10), we obtain

$$\log\hat{q}_3(\mathbf{Z}) = \sum_{i=1}^{n}\left((1-z_i)\log\frac{\langle 1-\gamma\rangle}{a} + z_i\log\left(\langle\gamma\rangle\langle\mathcal{N}_i\rangle\right)\right)$$

We see that $\hat{q}_3(\mathbf{Z})$ can be further factorized into $\hat{q}_3(\mathbf{Z}) = \prod_{i=1}^{n} \hat{q}_3^{[i]}(z_i)$. Considering $\hat{q}_3^{[i]}(z_i)$ and taking the exponential of both sides, we have

$$\hat{q}_3^{[i]}(z_i) \propto \{\langle 1-\gamma \rangle / a\}^{1-z_i} \{\langle \gamma \rangle \langle \mathcal{N}_i \rangle\}^{z_i}. \quad \text{(S5)}$$

As $z_i \in \{0, 1\}$ is a binary indicator variable, the normalization constant of $\hat{q}_3^{[i]}(z_i)$ is simply obtained by the summation for z_i

$$\langle 1-\gamma \rangle / a + \langle \gamma \rangle \langle \mathcal{N}_i \rangle.$$

Thus, after normalization, $\hat{q}_3^{[i]}(z_i)$ has the following form

$$\hat{q}_3^{[i]}(z_i) = (1-p_i)^{1-z_i} p_i^{z_i},$$

where

$$p_i = \frac{\langle \gamma \rangle \langle \mathcal{N}_i \rangle}{\langle 1-\gamma \rangle / a + \langle \gamma \rangle \langle \mathcal{N}_i \rangle}.$$

Note that the normalization constant is divided into each term of Eq. (S5). This is because z_i is a binary variable and only one term has a non-zero value.

5.6. Derivation on q_4

Given q_1, q_2, and q_3, since $q_4(\sigma^2, \boldsymbol{\varphi}, \mathbf{X}_m)$ obeys the Dirac delta distribution, we directly maximize the lower bound Eq. (11) with respect to $(\sigma^2, \boldsymbol{\varphi}, \mathbf{X}_m)$.

First we consider σ^2. The lower bound Eq. (11) becomes

$$\begin{aligned}
\mathcal{L} &\sim \int p(\mathbf{f}|\mathbf{f}_m)\phi(\mathbf{f}_m) \left(\sum_{i=1}^{n} p_i \log \mathcal{N}(y_i|f_i, \sigma^2) \right) d\mathbf{f} d\mathbf{f}_m \\
&\sim \int \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}_\mathbf{f}, \mathbf{A}) \left(\sum_{i=1}^{n} p_i \log \mathcal{N}(y_i|f_i, \sigma^2) \right) d\mathbf{f} \\
&\sim -\frac{1}{2\sigma^2} \int \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}_\mathbf{f}, \mathbf{A})(\mathbf{y}-\mathbf{f})^\top \mathbf{P}(\mathbf{y}-\mathbf{f}) d\mathbf{f} \\
&\quad - \frac{1}{2} \log \sigma^2 \sum_{i=1}^{n} p_i
\end{aligned}$$

Taking derivative with respect to σ^2 and setting it to zero, we obtain a closed-form expression (*i.e.*, Eq. (20) in the main paper) using Eq. (S2) :

$$\begin{aligned}
\hat{\sigma}^2 =& \frac{1}{\hat{n}} \int \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}_\mathbf{f}, \mathbf{A})(\mathbf{y}-\mathbf{f})^\top \mathbf{P}(\mathbf{y}-\mathbf{f}) d\mathbf{f} \\
=& \frac{1}{\hat{n}} (\mathbf{y}-\boldsymbol{\mu}_\mathbf{f})^\top \mathbf{P}(\mathbf{y}-\boldsymbol{\mu}_\mathbf{f}) + \frac{1}{\hat{n}} \mathrm{tr}(\mathbf{P}\mathbf{A}).
\end{aligned}$$

Next, we focus on the remaining hyperparameters $(\boldsymbol{\varphi}, \mathbf{X}_m)$. Using the reverse Jensen's inequality and Eq. (S1), we recover the Eq. (21) in the main paper:

$$\begin{aligned}
\mathcal{L} &\sim \int \phi(\mathbf{f}_m) \log \left(\frac{p(\mathbf{f}_m) Q(\mathbf{f}_m, \mathbf{y})}{\phi(\mathbf{f}_m)} \right) d\mathbf{f}_m \\
&\geq \log \int \cancel{\phi(\mathbf{f}_m)} \frac{p(\mathbf{f}_m) Q(\mathbf{f}_m, \mathbf{y})}{\cancel{\phi(\mathbf{f}_m)}} d\mathbf{f}_m \\
&= \log \int \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{f}|\mathbf{f}_m}, \sigma^2 \mathbf{P}^{-1}) \mathcal{N}(\mathbf{f}_m|\mathbf{0}, \mathbf{K}_{mm}) d\mathbf{f}_m \\
&\quad - \frac{1}{2\sigma^2} \mathrm{tr}(\mathbf{P}\mathbf{K}_{\mathbf{f}|\mathbf{f}_m}) \\
&\sim \log \int \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{f}|\mathbf{f}_m}, \sigma^2 \mathbf{P}^{-1}) \\
&\quad \mathcal{N}(\boldsymbol{\mu}_{\mathbf{f}|\mathbf{f}_m}|\mathbf{0}, \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}) d\boldsymbol{\mu}_{\mathbf{f}|\mathbf{f}_m} - \frac{1}{2\sigma^2} \mathrm{tr}(\mathbf{P}\mathbf{K}_{\mathbf{f}|\mathbf{f}_m}) \\
&= \log \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}_\mathbf{y}) - \frac{1}{2\sigma^2} \mathrm{tr}(\mathbf{P}\mathbf{K}_{\mathbf{f}|\mathbf{f}_m}) := \mathcal{L}_2.
\end{aligned}$$

The partial derivatives of $\mathcal{L}_2$ with respect to $(\boldsymbol{\varphi}, \mathbf{X}_m)$ is

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}_2 = \frac{1}{2} \mathrm{tr}\Big((\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - \mathbf{K}_\mathbf{y}^{-1}) \frac{\partial \mathbf{K}_\mathbf{y}}{\partial \boldsymbol{\theta}} \Big) - \frac{1}{2\sigma^2} \mathrm{tr}(\mathbf{P} \frac{\partial \mathbf{K}_{\mathbf{f}|\mathbf{f}_m}}{\partial \boldsymbol{\theta}}),$$

where $\boldsymbol{\alpha} = \mathbf{K}_\mathbf{y}^{-1}\mathbf{y}$.

References

[1] Jia-Wang Bian, Yu-Huan Wu, Ji Zhao, Yun Liu, Le Zhang, Ming-Ming Cheng, and Ian Reid. An evaluation of feature matchers for fundamental matrix estimation. In *Proceedings of the British Machine Vision Conference*, 2019. 2

[2] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 224–236, 2018. 2

[3] Jared Heinly, Johannes L Schonberger, Enrique Dunn, and Jan-Michael Frahm. Reconstructing the world* in six days*(as captured by the yahoo 100 million image dataset). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3287–3295, 2015. 2

[4] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 2

[5] Anastasya Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Advances in Neural Information Processing Systems*, pages 4829–4840, 2017. 2

[6] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 2

[7] Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Proceedings of the Artificial Intelligence and Statistics*, pages 567–574, 2009. 1

[8] Kyle Wilson and Noah Snavely. Robust global translations with 1dsfm. In *Proceedings of the European Conference on Computer Vision*, pages 61–75, 2014. 2

[9] Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, and Hongen Liao. Learning two-view correspondences and geometry using order-aware network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5845–5854, 2019. 2