

Supplementary Material for: Robust Multi-class Gaussian Process Classification

Daniel Hernández-Lobato
 ICTEAM - Machine Learning Group
 Université catholique de Louvain
 Place Sainte Barbe, 2
 Louvain-La-Neuve, 1348, Belgium
 danielhernandezlobato@gmail.com

José Miguel Hernández-Lobato
 Department of Engineering
 University of Cambridge
 Trumpington Street, Cambridge
 CB2 1PZ, United Kingdom
 jmh233@eng.cam.ac.uk

Pierre Dupont
 ICTEAM - Machine Learning Group
 Université catholique de Louvain
 Place Sainte Barbe, 2
 Louvain-La-Neuve, 1348, Belgium
 pierre.dupont@uclouvain.be

1 EP updates for RMGPC

Before describing the EP updates for the proposed model, we consider the following re-parameterization of \mathcal{Q} and the approximate terms $\tilde{\psi}_{ik}$, with $i = 1, \dots, n$ and $k \neq y_i$, and $\tilde{\psi}_i$, with $i = 1, \dots, n$:

$$p_i = \sigma(q_i), \quad \tilde{p}_{ik} = \sigma(\tilde{q}_{ik}), \quad \tilde{p}_i = \sigma(\tilde{q}_i), \quad (1)$$

where q_i , \tilde{q}_{ik} and \tilde{q}_i are new parameters taking values in \mathbb{R} and $\sigma(\cdot)$ is the logistic function

$$\sigma(x) = \frac{1}{1 + \exp(-x)}. \quad (2)$$

The logistic function is used to simplify the updates and to improve the numerical stability of the algorithm, especially when the posterior probability of $z_i = 1$ is very close to the extreme values 0 or 1. For the sake of clarity, we present the update operations that do not consider any damping, *i.e.*, $\epsilon = 1$. Incorporating the damping effect in the EP updates is straight-forward and is hence omitted.

1.1 Updates corresponding to the Likelihood

In this section we describe in detail the EP updates for the approximate terms $\tilde{\psi}_{ik}$, with $i = 1, \dots, n$ and $k \neq y_i$, corresponding to the likelihood. Recall that these factors are updated in parallel. Specifically, for each $\tilde{\psi}_{ik}$ we compute $\mathcal{Q}^{\setminus \tilde{\psi}_{ik}} \propto \mathcal{Q} / \tilde{\psi}_{ik}$ as follows:

$$\begin{aligned} (\Sigma_k^{\setminus \tilde{\psi}_{ik}})_{ii} &= ((\Sigma_k)_{ii}^{-1} - \tilde{\nu}_{ik}^{-1})^{-1}, & (\Sigma_{y_i}^{\setminus \tilde{\psi}_{ik}})_{ii} &= ((\Sigma_{y_i})_{ii}^{-1} - (\tilde{\nu}_{ik}^{y_i})^{-1})^{-1}, \\ \mu_{ki}^{\setminus \tilde{\psi}_{ik}} &= (\Sigma_k^{\setminus \tilde{\psi}_{ik}})_{ii} ((\Sigma_k)_{ii}^{-1} \mu_{ki} - \tilde{\nu}_{ik}^{-1} \tilde{\mu}_{ik}), & \mu_{y_i i}^{\setminus \tilde{\psi}_{ik}} &= (\Sigma_{y_i}^{\setminus \tilde{\psi}_{ik}})_{ii} ((\Sigma_{y_i})_{ii}^{-1} \mu_{y_i i} - (\tilde{\nu}_{ik}^{y_i})^{-1} \tilde{\mu}_{ik}^{y_i}), \\ q_i^{\setminus \tilde{\psi}_{ik}} &= q_i - \tilde{q}_{ik}. \end{aligned} \quad (3)$$

The other parameters of $\mathcal{Q}^{\tilde{\psi}_{ik}}$ are not required since the exact factor ψ_{ik} does not depend on them. Given each $\mathcal{Q}^{\tilde{\psi}_{ik}}$, we compute the updated approximate factor $\tilde{\psi}_{ik}$ as follows:

$$\begin{aligned}\tilde{\nu}_{ik} &= \frac{(\Sigma_k^{\tilde{\psi}_{ik}})_{ii} + (\Sigma_{y_i}^{\tilde{\psi}_{ik}})_{ii}}{\alpha_{ik}(m_{y_i} - m_k)} - (\Sigma_k^{\tilde{\psi}_{ik}})_{ii}, & \tilde{\nu}_{ik}^{y_i} &= \frac{(\Sigma_k^{\tilde{\psi}_{ik}})_{ii} + (\Sigma_{y_i}^{\tilde{\psi}_{ik}})_{ii}}{\alpha_{ik}(m_{y_i} - m_k)} - (\Sigma_{y_i}^{\tilde{\psi}_{ik}})_{ii}, \\ \tilde{\mu}_{ik} &= m_k - \alpha_{ik}\tilde{\nu}_{ik}, & \tilde{\mu}_{ik}^{y_i} &= m_{y_i} + \alpha_{ik}\tilde{\nu}_{ik}^{y_i}, \\ \tilde{q}_{ik} &= -\log(l)/(l-1) - \log\Phi(u_{ik}).\end{aligned}\quad (4)$$

where $\Phi(\cdot)$ is the cumulative probability function of a standard Gaussian distribution and

$$\begin{aligned}u_{ik} &= \frac{\mu_{y_i i}^{\tilde{\psi}_{ik}} - \mu_{k i}^{\tilde{\psi}_{ik}}}{\sqrt{(\Sigma_{y_i}^{\tilde{\psi}_{ik}})_{ii} + (\Sigma_k^{\tilde{\psi}_{ik}})_{ii}}}, & Z_{ik} &= \sigma(-q_i^{\tilde{\psi}_{ik}})\Phi(u_{ik}) + \sigma(q_i^{\tilde{\psi}_{ik}})\left(l^{-\frac{1}{l-1}}\right), \\ \alpha_{ik} &= \sigma(-q_i^{\tilde{\psi}_{ik}})\frac{\mathcal{N}(u_{ik}|0, 1)}{Z_{ik}}\frac{1}{\sqrt{(\Sigma_{y_i}^{\tilde{\psi}_{ik}})_{ii} + (\Sigma_k^{\tilde{\psi}_{ik}})_{ii}}}, & m_{y_i} &= \mu_{y_i i}^{\tilde{\psi}_{ik}} + \alpha_{ik}(\Sigma_{y_i}^{\tilde{\psi}_{ik}})_{ii} \\ m_k &= \mu_{k i}^{\tilde{\psi}_{ik}} - \alpha_{ik}(\Sigma_k^{\tilde{\psi}_{ik}})_{ii}.\end{aligned}\quad (5)$$

The update of the parameter \tilde{s}_{ik} is addressed later on, since it is only required to compute $Z \approx \mathcal{P}(\mathbf{y}|\mathbf{X})$. Once each $\tilde{\psi}_{ik}$ has been updated, we recompute \mathcal{Q} as the normalized product of all the approximate terms. This gives

$$q_i = \sum_{k=1}^l \tilde{q}_{ik} + \tilde{q}_i, \quad \Sigma_k = (\mathbf{K}_k^{-1} + \mathbf{\Lambda}^k)^{-1}, \quad \boldsymbol{\mu}_k = \Sigma_k \mathbf{v}^k, \quad (6)$$

where $\mathbf{\Lambda}^k$ and \mathbf{v}^k are respectively defined as in (17) and (18), in the main manuscript. The parameters a and b of \mathcal{Q} do not change in these updates.

Once EP has converged, we compute \tilde{s}_{ik} to guarantee that $\tilde{\psi}_{ik}\mathcal{Q}^{\tilde{\psi}_{ik}}$ and $\psi_{ik}\mathcal{Q}^{\tilde{\psi}_{ik}}$ integrate up to the same value. This gives

$$\begin{aligned}\tilde{s}_{ik} &= \left(\Phi(u_{ik}) + l^{-\frac{1}{l-1}}\right) \sqrt{1 + (\tilde{\nu}_{ik}^{y_i})^{-1}(\Sigma_{y_i}^{\tilde{\psi}_{ik}})_{ii}} \sqrt{1 + \tilde{\nu}_{ik}^{-1}(\Sigma_k^{\tilde{\psi}_{ik}})_{ii}} \\ &\quad \exp\left\{\frac{\alpha_{ik}(\Sigma_{y_i}^{\tilde{\psi}_{ik}})_{ii}(1 + (\Sigma_k^{\tilde{\psi}_{ik}})_{ii})}{2(m_{y_i} - m_k)}\right\} \exp\left\{\frac{\alpha_{ik}(\Sigma_k^{\tilde{\psi}_{ik}})_{ii}(1 + (\Sigma_{y_i}^{\tilde{\psi}_{ik}})_{ii})}{2(m_{y_i} - m_k)}\right\}.\end{aligned}\quad (7)$$

1.2 Updates corresponding to the Prior for \mathbf{z} given ρ

In this section we describe in detail the EP updates for the approximate terms $\tilde{\psi}_i$, with $i = 1, \dots, n$ corresponding to the prior for \mathbf{z} given ρ . For each $\tilde{\psi}_{ik}$ we compute $\mathcal{Q}^{\tilde{\psi}_i} \propto \mathcal{Q}/\tilde{\psi}_i$ as follows:

$$q_i^{\tilde{\psi}_i} = q_i - \tilde{q}_i, \quad a^{\tilde{\psi}_i} = a - \tilde{a}_i + 1, \quad b^{\tilde{\psi}_i} = b - \tilde{b}_i + 1. \quad (8)$$

The other parameters of $\mathcal{Q}^{\tilde{\psi}_i}$ are not required since the exact factor ψ_i does not depend on them. Given $\mathcal{Q}^{\tilde{\psi}_i}$, we compute the updated approximate factor $\tilde{\psi}_i$ as follows:

$$\tilde{q}_i = \log a^{\tilde{\psi}_i} - \log b^{\tilde{\psi}_i}, \quad \tilde{a}_i = a^* - a^{\tilde{\psi}_i} + 1, \quad \tilde{b}_i = b^* - b^{\tilde{\psi}_i} + 1, \quad (9)$$

where

$$a^* = e_1 \frac{e_1 - e_2}{e_2 - e_1^2}, \quad b^* = (1 - e_1) \frac{e_1 - e_2}{e_2 - e_1^2}, \quad (10)$$

and

$$\begin{aligned}
e_1 &= \frac{1}{Z_i} \left(\sigma(-q_i^{\tilde{\psi}_i})(1 - \bar{\rho}^{\tilde{\psi}_i}) \frac{a^{\tilde{\psi}_i}}{a^{\tilde{\psi}_i} + b^{\tilde{\psi}_i} + 1} + \sigma(q_i^{\tilde{\psi}_i})\bar{\rho}^{\tilde{\psi}_i} \frac{a^{\tilde{\psi}_i} + 1}{a^{\tilde{\psi}_i} + b^{\tilde{\psi}_i} + 1} \right), \\
e_2 &= \frac{1}{Z_i} \left(\sigma(-q_i^{\tilde{\psi}_i})(1 - \bar{\rho}^{\tilde{\psi}_i}) \frac{a^{\tilde{\psi}_i}(a^{\tilde{\psi}_i} + 1)}{(a^{\tilde{\psi}_i} + b^{\tilde{\psi}_i} + 1)(a^{\tilde{\psi}_i} + b^{\tilde{\psi}_i} + 2)} + \right. \\
&\quad \left. + \sigma(q_i^{\tilde{\psi}_i})\bar{\rho}^{\tilde{\psi}_i} \frac{(a^{\tilde{\psi}_i} + 1)(a^{\tilde{\psi}_i} + 2)}{(a^{\tilde{\psi}_i} + b^{\tilde{\psi}_i} + 1)(a^{\tilde{\psi}_i} + b^{\tilde{\psi}_i} + 2)} \right), \\
Z_i &= \sigma(-q_i^{\tilde{\psi}_i})(1 - \bar{\rho}^{\tilde{\psi}_i}) + \sigma(q_i^{\tilde{\psi}_i})\bar{\rho}^{\tilde{\psi}_i}, \\
\bar{\rho}^{\tilde{\psi}_i} &= \frac{a^{\tilde{\psi}_i}}{a^{\tilde{\psi}_i} + b^{\tilde{\psi}_i}}. \tag{11}
\end{aligned}$$

As suggested in [1] (see Section 3.3.3), in Eq. (9) we have matched the first and the second moments of ρ instead of the sufficient statistics. The motivation is that this provides a closed form expression for the update of $\tilde{\psi}_i$. The update of the parameter \tilde{s}_i is addressed later on, since it is only required to compute $Z \approx \mathcal{P}(\mathbf{y}|\mathbf{X})$.

Once $\tilde{\psi}_i$ has been updated, we update \mathcal{Q} . In particular, we set

$$a = a^{\tilde{\psi}_i} + \tilde{a}_i - 1, \quad b = b^{\tilde{\psi}_i} + \tilde{b}_i - 1, \quad q_i = q_i^{\tilde{\psi}_i} + \tilde{q}_i. \tag{12}$$

Once EP has converged, we compute \tilde{s}_i to guarantee that $\tilde{\psi}_i \mathcal{Q}^{\tilde{\psi}_i}$ and $\psi_i \mathcal{Q}^{\tilde{\psi}_i}$ integrate up to the same value. This gives

$$\tilde{s}_i = \frac{B(a^{\tilde{\psi}_i}, b^{\tilde{\psi}_i})}{B(a, b)}. \tag{13}$$

2 Extra Experiments

In this section we further evaluate RMGPC, SMGPC and HTPC under different noise conditions. For this purpose we consider a linear synthetic classification problem characterized by the following labeling rule:

$$y_i = \arg \max_k f_k(\mathbf{x}_i), \tag{14}$$

where k goes from 1 to 5, *i.e.*, the problem has 5 different classes. Each latent function f_k is defined as follows:

$$f_k(\mathbf{x}) = \mathbf{w}_k^T \mathbf{x}, \tag{15}$$

where \mathbf{w}_k is a hyper-plane generated from a factorizing Gaussian distribution with zero mean and unit variance in each component. The data instances \mathbf{x} of this problem are also generated from the same distribution. The dimension d of each \mathbf{x} and, hence, of each \mathbf{w}_k , is set equal to 5.

Using the labeling rule (14) and the specifications described before, we randomly generate 100 training and test sets containing 100 and 1000 instances, respectively. A different set of latent functions f_1, \dots, f_5 is considered in each realization. This process is repeated three times, and each time we contaminate the 100 training sets with a different type of noise. Specifically, we consider three different noise scenarios:

1. **Noise near the decision boundaries:** For each instance, we contaminate each latent function f_k with some additive Gaussian noise ϵ_k with zero mean and standard deviation equal to 0.5. Under this noise scenario SMGPC is expected to be optimal since it assumes latent Gaussian noise around each latent function f_k .

2. **Arbitrary noise in the labels of the data:** The class label of 20% of the training instances are randomly chosen from the 5 different potential class labels of the problem. Under this scenario RMGPC is expected to perform significantly better than SMGPC and HTPC, which cannot consider noise in the labels of the data independently of their distance to the decision boundaries.
3. **Mix of both types of noise:** We simultaneously consider the two types of noise described before. This means that in the data there are noisy instances close to the decision boundaries and noisy instances that are independent of their distance to the decision boundaries.

In the two first scenarios we have fixed the level of noise so that the best performing model has similar error rates.

Next, we evaluate the prediction performance of RMGPC, SMGPC and HTPC using each one of the training and test sets generated. However, since the classification problem we are analyzing is linear, instead of using the Gaussian covariance function described in Eq. (22) of the main manuscript, we employ in RMGPC, SMGPC and HTPC a linear covariance function defined as:

$$c(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j. \quad (16)$$

The remaining parameters of RMGPC, SMGPC and HTPC are fixed as described in Section 4.1 of the main manuscript. Since the class distributions of the problem considered are balanced, we do not report for each method the average balanced class rate on the test set, but the prediction error. Specifically, the results for each different noise scenario are displayed for RMGPC, SMGPC and HTPC in Table 1. When the performance of a method is significantly different from the performance of RMGPC, as estimated by a Wilcoxon rank test (p-value < 1%), the corresponding error is marked with the symbol \triangleleft . The table shows that RMGPC performs best in the second and third noise scenarios while it performs only worse in the first.

Table 1: Average test error in % of each method on the synthetic problem for each noise scenario.

| Scenario | RMGPC | SMGPC | HTPC |
|----------|----------|--------------------------|--------------------------|
| 1 | 15.2±2.8 | 13.4±2.4 \triangleleft | 14.5±3.1 \triangleleft |
| 2 | 13.6±3.1 | 22.1±4.1 \triangleleft | 21.5±4.4 \triangleleft |
| 3 | 19.4±3.7 | 24.6±4.4 \triangleleft | 24.2±5.0 \triangleleft |

The figures reported in Table 1 indicate that RMGPC provides much better results when data instances whose labels strongly disagree with the assumed labeling rule (outliers) are present in the data. When such instances are not present in the data RMGPC is found to perform only slightly worse, although the differences are statistically significant.

References

- [1] Thomas Minka. *A Family of Algorithms for approximate Bayesian Inference*. PhD thesis, Massachusetts Institute of Technology, 2001.