EFFICIENT REMOTE SENSING IMAGE CLASSIFICATION WITH GAUSSIAN PROCESSES AND FOURIER FEATURES

Pablo Morales[1], *Adrián Pérez-Suay*[2], *Rafael Molina*[1] *and Gustau Camps-Valls*[2]

[1]Dept. of Computer Science and Artificial Intelligence, University of Granada, Spain
[2]Image Processing Laboratory (IPL), Universitat de València, Spain

ABSTRACT

This paper presents an efficient methodology for approximating kernel functions in Gaussian process classification (GPC). Two models are introduced. We first include the standard random Fourier features (RFF) approximation into GPC, which largely improves the computational efficiency and permits large scale remote sensing data classification. In addition, we develop a novel approach which avoids randomly sampling a number of Fourier frequencies, and alternatively *learns* the optimal ones using a variational Bayes approach. The performance of the proposed methods is illustrated in complex problems of cloud detection from multispectral imagery.

Index Terms— Gaussian Process Classification (GPC), random Fourier features, Variational Inference, Cloud detection, Seviri/MSG

1. INTRODUCTION

"... Nature almost surely operates by combining chance with necessity, randomness with determinism..."
–Eric Chaisson, Epic of Evolution: Seven Ages of the Cosmos

Current Earth-observation (EO) applications for image classification have to deal with an unprecedented big amount of heterogeneous and complex data sources. Spatio-temporally explicit classification methods are a requirement in a variety of Earth system data processing applications. Upcoming missions such as the super-spectral Copernicus Sentinels, EnMAP, and FLEX will soon provide unprecedented data streams. Very high resolution (VHR) sensors like Worldview-3 also pose big challenges to data processing. The challenge is not only attached to optical sensors but also to infrared sounders and radar images, which increased in spectral, spatial and temporal resolution. Besides, we should not forget the extremely large remote sensing data archives already collected by several past missions, such ENVISAT, Cosmo-SkyMED, Landsat, SPOT, or Seviri/MSG. These large-scale data problems require enhanced processing techniques that should be accurate, robust and fast. Standard classification algorithms cannot cope with this new scenario efficiently.

Research funded by the European Research Council (ERC) under the ERC-CoG-2014 SEDAL project (grant agreement 647423), the Spanish Ministry of Economy and Competitiveness (MINECO) through the projects TIN2013-43880-R, TIN2015-64210-R, and DPI2016-77869-C2-2-R, and the EUMETSAT through contract EUM/RSP/SOW/14/762293.

In the last decade kernel methods have dominated the field of remote sensing image classification [1, 2]. In particular, a kernel method called support vector machine (SVM, [3]) was gradually introduced in the field, and quickly became a standard for image classification. However, kernel machines are still not widely adopted in real practice because of the high computational cost when dealing with large scale problems. Roughly speaking, given n examples for training, kernel methods typically need to store in memory *kernel matrices* of size $n \times n$ and to process them using standard linear algebra tools (e.g. matrix inversion, factorization, eigendecomposition) that typically scale cubically, $\mathcal{O}(n^3)$. This important constraint hampers applicability to large scale EO problems.

An interesting kernel method for classification is the Gaussian Process classifier (GPC) [4]. The GPC method was originally introduced in remote sensing classification in [5], where very good capabilities for land cover classification from multispectral and hyperspectral imagery were shown. GPC has interesting theoretical properties as it approaches the classification problem with a solid Bayesian treatment. Unfortunately, like any other kernel method, the computational cost is very high since large matrices need to be inverted. Furthermore, the use of non-conjugate observation models in classification problems renders impossible the calculation of the marginal distribution needed for inference. These problems could be addressed with the introduction of inducing points and approximate inference [6], but this comes at the price of a vast number of parameters to estimate.

In this paper, we introduce an alternative pathway to perform efficient GP classification. We follow the work in [7] to approximate the kernel matrix with projections on a reduced set of random Fourier features (RFF). This makes classification possible with millions of examples, as recently shown in [8] for SVM land cover problems. Here, two contributions are given. First, we include the RFF kernel approximation into the GP formalism, yielding the so-called RFF-GPC method. In addition, we introduce a novel approach to avoid *randomly sampling* a number of Fourier frequencies. Instead, we propose *learning* the optimal ones in a variational Bayes fashion. Therefore, Fourier frequencies are no longer *fixed*, but *latent variables* to be estimated directly from data. We will call this method Variational Fourier Features (VFF) GPC.

The remainder of the paper is organized as follows. Section §2 reviews the RFF approximation and introduce it into GPC, deriving RFF-GPC and the more sophisticated VFF-GPC. Section §3 summarizes the data used in this paper. Sec-

tion §4 shows the experimental results comparing the two proposed methods and standard GPC. Section §5 concludes the paper with some remarks and future outlook.

2. GAUSSIAN PROCESS CLASSIFICATION WITH FOURIER FEATURES

Gaussian processes (GPs) are Bayesian state-of-the-art tools for machine learning problems. They are well-known to the geostatistics community as kriging. While GP algorithms are typically used for regression tasks, they can be naturally extended for classification by "squashing" the real line into a $(0,1)$-interval that represents the probability of one class [4].

Namely, we are given a set of input-output data pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{0,1\}$. The GP models this relationship by means of latent variables $\{f_i = f(\mathbf{x}_i) \in \mathbb{R}\}_{i=1}^n$ that jointly follow a normal distribution $\mathcal{N}(\mathbf{0}, (k(\mathbf{x}_i, \mathbf{x}_j))_{1\leq i,j\leq n})$, with k the kernel function. The observation model for the output is given by the sigmoid function $\psi(\cdot)$ as $\mathrm{p}(y_i|f_i) = \psi(f_i) = (1+\exp(-f_i))^{-1}$.

2.1. Random Fourier Features

Standard approaches for GP classification scale as $\mathcal{O}(n^3)$. This is non-efficient and computationally prohibitive for large datasets. Following the work in [7], an elegant alternative is to approximate the kernel function with a linear one, which is computed over $\mathcal{O}(D)$ explicitly mapped random Fourier features. This substitutes $n \times n$ matrix inversions for $\mathcal{O}(D) \times \mathcal{O}(D)$ ones, which reduces the computational cost to $\mathcal{O}(nD^2)$, with D typically much lower than n.

For instance, the well-known squared exponential (SE) Gaussian kernel $k(\mathbf{x}, \mathbf{x}') = \gamma \cdot \exp(-\|\mathbf{x}-\mathbf{x}'\|^2/(2\sigma^2))$ can be approximated as $k(\mathbf{x}, \mathbf{x}') \approx \mathbf{z}(\mathbf{x})^\intercal \mathbf{z}(\mathbf{x}')$, with

$$\mathbf{z}(\mathbf{x})^\intercal = \sqrt{\gamma D^{-1}} \left(\cos\left(\sigma^{-1}\mathbf{w}_1^\intercal \mathbf{x}\right), \sin\left(\sigma^{-1}\mathbf{w}_1^\intercal \mathbf{x}\right), \dots \right.$$
$$\left. \dots, \cos\left(\sigma^{-1}\mathbf{w}_D^\intercal \mathbf{x}\right), \sin\left(\sigma^{-1}\mathbf{w}_D^\intercal \mathbf{x}\right)\right) \in \mathbb{R}^{2D}. \quad (1)$$

The precision of the approximation exponentially increases with the number D of Fourier frequencies $\mathbf{w}_i \in \mathbb{R}^d$, which are independently sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ [7]. In matrix notation, given n data points, the kernel matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$ is approximated by the explicitly projected data $\mathbf{Z} = [\mathbf{z}_1 \cdots \mathbf{z}_n]^\intercal \in \mathbb{R}^{n\times D}$ as $\mathbf{K} \approx \mathbf{Z}\mathbf{Z}^\intercal$.

To approximate the GP classifier, we will consider a Bayesian linear model which uses these new features. Within a variational inference approach, hyperparameters γ and σ in eq. (1) will be estimated by maximizing the marginal likelihood of the observed data. For VFF-GPC, the Fourier frequencies $\mathbf{w}_i$ will be also optimized.

2.2. Derivation of the proposed methods

2.2.1. Model formulation

We consider the standard logistic observation model working on Fourier features $\mathbf{z}_i$:

$$\mathrm{p}(y_i = 1|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{W}) = (1+\exp(-\boldsymbol{\beta}^\intercal \mathbf{z}_i))^{-1}, \quad (2)$$

where $\mathbf{z}_i$ is given by eq. (1), $\boldsymbol{\theta} = (\theta_1 = \sqrt{\gamma}, \theta_2 = \sigma)$, and $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_D]^\intercal \in \mathbb{R}^{D\times d}$.

Let us assign the prior distributions $\mathrm{p}(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta}|\mathbf{0}, \mathbf{I})$ and $\mathrm{p}(\mathbf{W}) = \mathcal{N}(\mathbf{W}|\mathbf{0}, \mathbf{I})$ to the weights $\boldsymbol{\beta} \in \mathbb{R}^{2D}$ and Fourier frequencies $\mathbf{W}$, respectively. A flat improper prior distribution is assumed for $\boldsymbol{\theta}$, that is, $\mathrm{p}(\boldsymbol{\theta}) \propto$ const. Therefore, the joint p.d.f. reads

$$\mathrm{p}(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{W}) = \mathrm{p}(\mathbf{y}|\boldsymbol{\beta}, \mathbf{W}, \boldsymbol{\theta})\mathrm{p}(\boldsymbol{\beta})\mathrm{p}(\mathbf{W}). \quad (3)$$

Following the Bayesian approach, we seek to optimally estimate $\boldsymbol{\theta}$ and $\mathbf{W}$ by $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{W}}$ respectively, and then derive the posterior distribution $\mathrm{p}(\boldsymbol{\beta}|\mathbf{y}, \hat{\boldsymbol{\theta}}, \hat{\mathbf{W}})$. However, the sigmoid likelihood in eq. (2) makes these computations analytically intractable, and we will resort to the *variational* inference approximation method [9, Section 10.6].

In the case of RFF-GPC, we do not consider $\mathbf{W}$ as model's parameters in eqs. (2)–(3). Instead, they are randomly sampled and fixed from now on.

2.2.2. Variational inference

In order to estimate $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{W}}$ we have to integrate out $\boldsymbol{\beta}$ in eq. (3). This is not analytically possible because of the sigmoid in $\mathrm{p}(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{W})$. To overcome this problem, we use the variational bound [9, Section 10.6]

$$\log\left(1+e^x\right) \leq \lambda(\xi)(x^2-\xi^2) + \frac{x-\xi}{2} + \log\left(1+e^\xi\right), \quad (4)$$

where $x, \xi \in \mathbb{R}$ and $\lambda(\xi) = (1/2\xi)\left((1+e^{-\xi})^{-1} - 1/2\right)$. For a fixed x, this bound is minimum when $\xi^2 = x^2$. This produces the following lower bound for eq. (3):

$$\log \mathrm{p}(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{W}) \geq -\frac{1}{2}\boldsymbol{\beta}^\intercal\left(\mathbf{Z}^\intercal(2\boldsymbol{\Lambda})\mathbf{Z} + \mathbf{I}\right)\boldsymbol{\beta} + \mathbf{v}^\intercal\mathbf{Z}\boldsymbol{\beta}$$
$$- \frac{1}{2}\operatorname{Tr}\left(\mathbf{W}\mathbf{W}^\intercal\right) + C(\xi_i), \quad (5)$$

where $\mathbf{v} = \mathbf{y} - (1/2)\cdot\mathbf{1}_{n\times 1}$ and $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda(\xi_1), \dots, \lambda(\xi_n))$.

Using the exponential function on eq. (5), integrating out $\boldsymbol{\beta}$, and maximizing on $\boldsymbol{\theta}$ and $\mathbf{W}$ (both are in $\mathbf{Z}$), we obtain:

$$(\hat{\boldsymbol{\theta}}, \hat{W}) = \arg\max_{(\boldsymbol{\theta}, \mathbf{W})} \left(\mathbf{v}^\intercal\mathbf{Z}\left(\mathbf{Z}^\intercal(2\boldsymbol{\Lambda})\mathbf{Z} + \mathbf{I}\right)^{-1}\mathbf{Z}^\intercal\mathbf{v} - \right.$$
$$\left. - \log\left|\mathbf{Z}^\intercal(2\boldsymbol{\Lambda})\mathbf{Z} + \mathbf{I}\right| - \frac{1}{2}\operatorname{Tr}\left(\mathbf{W}\mathbf{W}^\intercal\right)\right). \quad (6)$$

Now, fixing $\boldsymbol{\theta}$ an $\mathbf{W}$, we again resort to the variational approximation given by eq. (5) to approximate the posterior $\mathrm{p}(\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\theta}, \mathbf{W})$, which yields

$$\mathrm{p}(\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\theta}, \mathbf{W}) \approx \mathrm{q}(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta}|\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}),$$
$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \left(\mathbf{Z}^\intercal(2\boldsymbol{\Lambda})\mathbf{Z} + \mathbf{I}\right)^{-1}, \quad \boldsymbol{\mu}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\mathbf{Z}^\intercal\mathbf{v}. \quad (7)$$

Finally, the bound in eq. (5) is optimal when

$$\xi_i^2 = \langle(\boldsymbol{\beta}^\intercal\mathbf{z}_i)^2\rangle_{\mathrm{q}(\boldsymbol{\beta}|\mathbf{y}, \hat{\boldsymbol{\theta}}, \hat{\mathbf{W}})} \quad \forall i = 1, \dots, n, \quad (8)$$

which produces

$$\xi_i = \sqrt{\left(\mathbf{z}_i^\intercal \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \mathbf{Z}^\intercal \mathbf{v}\right)^2 + \mathbf{z}_i^\intercal \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \mathbf{z}_i} \quad \forall i = 1, \ldots, n. \tag{9}$$

2.2.3. *Algorithm implementation*

The proposed RFF-GPC and VFF-GPC algorithms run iteratively to compute estimations $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{W}}$ and the posterior $\mathrm{q}(\boldsymbol{\beta})$:

1. Initialize $\mathrm{q}(\boldsymbol{\beta})$ with the prior normal distribution $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \mathbf{0}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \mathbf{I}$; $\boldsymbol{\theta}$ with sensible values for the available data (i.e. using greedy approximations); and $\mathbf{W}$ with a random sample from $\mathcal{N}(\mathbf{0}, \mathbf{I})$.
2. Update the values of ξ_i's with equation (9).
3. Update $\boldsymbol{\theta}$ (and $\mathbf{W}$ for VFF-GPC) with eq. (6).
4. Update $q(\boldsymbol{\beta})$ with equation (7).
5. Go to step 2 and iterate until convergence.

For a new test example $\mathbf{x}_* \in \mathbb{R}^d$, the probability of class 1 is:

$$\mathrm{p}(y_* = 1) \approx \int \mathrm{p}(y_* = 1 | \boldsymbol{\beta}, \hat{\boldsymbol{\theta}}, \hat{\mathbf{W}}) \mathrm{q}(\boldsymbol{\beta}) \mathrm{d}\boldsymbol{\beta} \approx$$
$$\approx \psi\left(\mathbf{z}_*^\intercal \boldsymbol{\mu}_{\boldsymbol{\beta}} \cdot \left(1 + (\pi/8) \mathbf{z}_*^\intercal \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \mathbf{z}_*\right)^{-1/2}\right).$$

3. DATA COLLECTION AND PREPROCESING

We focus on the problem of cloud identification over landmarks on Seviri Meteosat Second Generation (MSG) data. This satellite mission constitutes a fundamental tool for weather forecasting, providing images of the full Earth disc every 15 minutes. Cloud contamination detection over landmarks is a essential step in the MSG processing chain, as undetected clouds are one of the most significant sources of error in landmark matching (see Fig. 1).

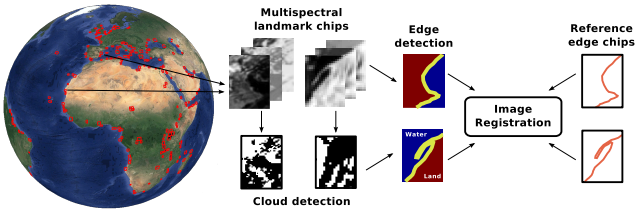

Fig. 1: Landmarks are essential in image registration and geometric quality assessment. Any misclassification of a landmark due to cloud contamination degrades the correlation matching, which is a cornerstone for the image navigation and registration (INR) algorithms.

The dataset used in the experiments was provided by EUMETSAT, and contains MSG/SEVIRI Level 1.5 acquisitions for 200 landmarks of variable size for a whole year (2010), which are mainly located over the coastline, islands, or inland waters. We selected all multispectral images from a particular landmark location, which involves 3,679,200 chip images. In addition, Level 2 cloud products were provided for each landmark observation, so the Level 2 cloud mask [10] is used as the best available 'ground truth' to validate the results. We framed the problem as a pixel-wise classification one.

Based on previous studies [10, 11] and in order to simplify the classification task, the different illumination conditions (and hence difficulty) over the landmarks are studied by splitting the day in four ranges (sub-problems). These depend on the solar zenith angle (SZA) value, which is directly related with the light condition: *high* (SZA$<$SZA$_{\mathrm{median}}$), *mid* (SZA$_{\mathrm{median}}$ $<$SZA$<$ 80°), *low* (80°$<$SZA$<$90°), and *night* (SZA$>$90°). A number of features were extracted from the images, involving band ratios, spatial, contextual and morphological features, as well as discriminative cloud detection scores (details will be given at the time of the conference). All in all, data dimension involves $d = 16$ features for the three daylight problems, and $d = 6$ for the nighttime one.

4. EXPERIMENTAL RESULTS

For the experimental setting, we sampled the classes evenly to avoid skewed results and thus generated balanced sets of diferent sizes. The cardinality of the training sets are increasingly selected as $n \in \{200, 300, 500, 1000, 2000\}$ for GPC (recall its $\mathcal{O}(n^3)$ cost) and $n \in \{500, 1000, 2000, 4000\}$ for the proposed RFF-GPC and VFF-GPC. This allows us to study the trade-off between accuracy and computational cost of the models.We set the number of random Fourier features to $D = 100$ for both RFF-GPC and VFF-GPC. The behavior of all the methods is measured over an (also balanced) test set of size $n_{\mathrm{test}} = 15000$.

In Figure 2 we show the overall accuracy (OA, percentage of correctly predicted test data) and CPU-time for the two proposed methods and the standard GPC. RFF-GPC and VFF-GPC obtain similar (or slightly better) accuracies than GPC at the overlapping experiments ($n = 500, 1000, 2000$). However, they are faster, specially RFF-GPC, which is around ten times more efficient. This improvement factor can even reach a value of 100 (see $n = 2000$ for the *mid* dataset).

In the case of VFF-GPC, CPU-time decreases when moving from $n = 500$ to $n = 1000$ for all the data sets. This may be due to the large number of hyperparameters $(\boldsymbol{\theta}, \mathbf{W})$ to be estimated, which stands at $2 + dD = 1602$ for daylight data sets and 602 for nighttime one. Therefore, although each iteration is faster for $n = 500$, the shortage of data leads to a much greater number of iterations until convergence. There are also other experimental reasons which suggest that VFF-GPC is specially conceived for larger data sets. For example, it starts to significantly outperform RFF-GPC for $n = 4000$, and it also performs better than GPC as n increases.

RFF-GPC and VFF-GPC are not only competitive fast approximations for GPC, but they are also expected to scale well to big data problems (prohibitive for the $\mathcal{O}(n^3)$ cost of GPC). This can be observed from the CPU-time evolution, and it is specially clear for RFF-GPC. As argued before, VFF-GPC would become more efficient in a large scale framework due to the amount of hyperparameters to be estimated.

In order to better illustrate the remarkable enhancement, we show the OA vs CPU-time trade-off for data set *low* in Figure 3. Left and top positions correspond, respectively, to faster and more accurate results. Interestingly, GPC dominates the right-bottom area, which indicates that it is clearly

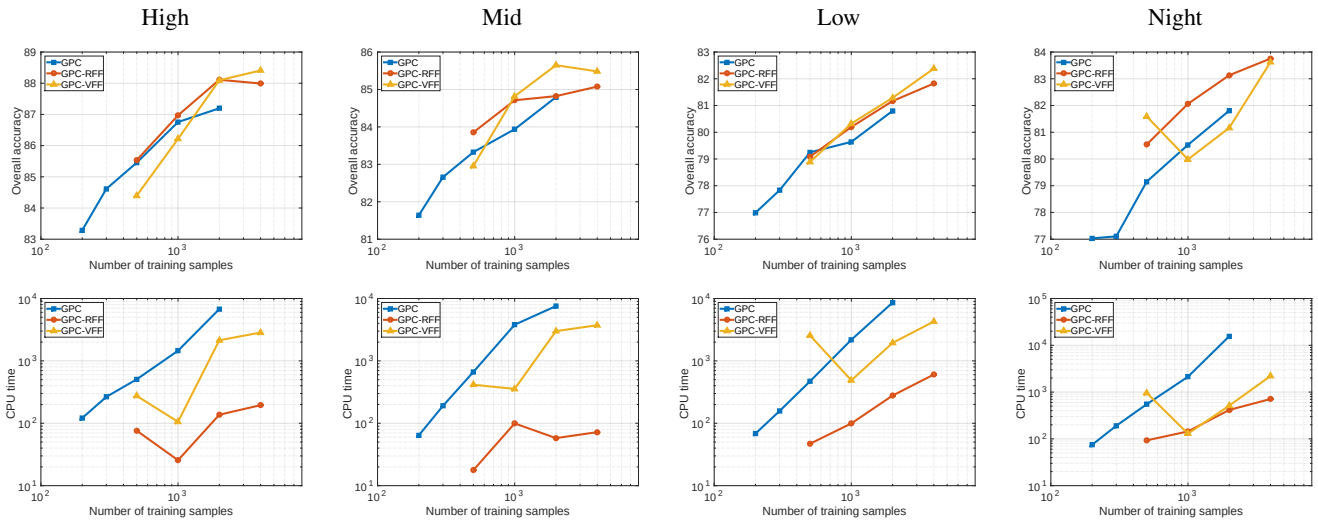

Fig. 2: Overall accuracy (upper row) and CPU execution time (lower row), both against the number of training instances.

outperformed by the proposed methods. RFF-GPC seems to obtain the best trade-off, whereas VFF-GPC can reach slightly higher OA at the cost of a (slightly) higher computational efficiency.

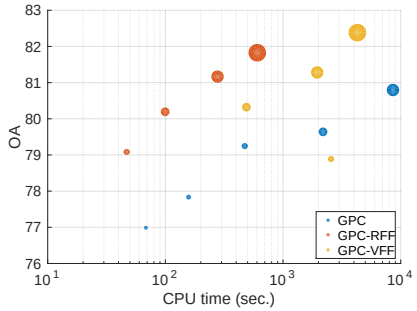

Fig. 3: OA vs CPU-time trade-off for the *low* data set. Point size is proportional to the number of samples used for training.

5. CONCLUSIONS AND FUTURE WORK

In this work we presented two efficient approximations to GPC. The first one, RFF-GPC, is performed through random Fourier features. The second, VFF-GPC, goes one step beyond by optimizing over the Fourier frequencies within a variational Bayes approach. Compared to the original GPC, the experimental results show a high competitiveness in accuracy, a remarkable decrease in computational cost, and an excellent trade-off between them. These results encourage us to expand the experimentation to larger scale problems, trying to exploit the potential of VFF-GPC when dealing with big data. Other prior distributions and inference methods will also be explored in the future.

6. REFERENCES

[1] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geosc. Rem. Sens.*, vol. 43, no. 6, pp. 1351–1362, Jun 2005.

[2] G. Camps-Valls and L. Bruzzone, Eds., *Kernel methods for Remote Sensing Data Analysis*, Wiley & Sons, UK, Dec 2009.

[3] G. Camps-Valls, L. Gómez-Chova, J. Calpe, E. Soria, J. D. Martín, L. Alonso, and J. Moreno, "Robust support vector method for hyperspectral data classification and knowledge discovery," *IEEE Trans. Geosc. Rem. Sens.*, vol. 42, no. 7, pp. 1530–1542, Jul 2004.

[4] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, New York, 2006.

[5] Y. Bazi and F. Melgani, "Gaussian process approach to remote sensing image classification," *IEEE Trans. Geosc. Rem. Sens.*, vol. 48, no. 1, pp. 186–197, Jan 2010.

[6] Andreas Damianou, *Deep Gaussian Processes and Variational Propagation of Uncertainty*, Ph.D. thesis, University of Sheffield, 2015.

[7] Ali Rahimi and Benjamin Recht, "Random features for large-scale kernel machines," in *Advances in neural information processing systems*, 2007, pp. 1177–1184.

[8] A. Pérez-Suay, L. Gómez-Chova, J. Amorós, and G. Camps-Valls, "Randomized kernels for large scale earth observation applications," *Remote Sensing of Environment*, 2017.

[9] Christopher M Bishop, "Pattern recognition," *Machine Learning*, vol. 128, 2006.

[10] M. Derrien and H. Le Gléau, "MSG/SEVIRI cloud mask and type from SAFNWC," *International Journal of Remote Sensing*, vol. 26, no. 21, pp. 4707–4732, 2005.

[11] J. Hocking, P. N. Francis, and R. Saunders, "Cloud detection in meteosat second generation imagery at the met office," Tech. Rep. 540, Universitat de València, 2010.