Patch kernels for Gaussian processes in high-dimensional imaging problems

Matthew Chung Hai Lee
Imperial College London
London, SW7 2AZ
matthew.lee13@imperial.ac.uk

Hugh Salimbeni
Imperial College London
London, SW7 2AZ
h.salimbeni15@imperial.ac.uk

Marc Peter Deisenroth
Imperial College London
London, SW7 2AZ
m.deisenroth@imperial.ac.uk

Ben Glocker
Imperial College London
London, SW7 2AZ
b.glocker@imperial.ac.uk

Abstract

In real-world, high dimensional machine learning problems, such as imaging, many of the available input features can be uninformative and dominate relevant signals which hurts training of predictive models. This is a predominant problem in medical imaging where inference is made on large volumetric image data. Gaussian processes are generally well suited to medical imaging applications as they work well when few training examples are available and provide vital uncertainty information. However, kernel methods are often problematic in such settings as commonly used kernel functions rely on distance metrics measured across whole data points which can over-attribute importance to irrelevant feature dimensions. We introduce patch kernels for Gaussian processes, a novel way to exploit spatial information in data by breaking high-dimensional images down into smaller subregions and calculating covariances across these localised regions. This works well in imaging problems as it takes advantage of the spatial structure. We demonstrate our method on two datasets: a synthetic digit classification problem and age regression from magnetic resonance images of brains. Our results show patch kernels outperform standard kernels when there is structural information to be exploited in the input.

1 Introduction

Gaussian processes (GPs) are a powerful tool in machine learning that models arbitrary functions while providing predictive probability distributions. Any machine learning problem that can be formulated as inference on a function is a potential application for a GP [11]. As a prior over functions GPs have been useful in many applications, including geostatistics [8], robotics [5] and bioinformatics [6]. Having a fully Bayesian probabilistic interpretation means their uncertainty measures can also be utilised, such as in [9] where GPs are used to generate samples of segmentations and to provide uncertainty estimates.

However, GPs are ill suited to deal with problems where inputs are dominated by large quantities of irrelevant features. That is, when the signal is a function of only a subset of the input features. An example where this is a predominant problem is medical imaging, where the dimensionality (number of pixels) is high but the areas of interest can be a small proportion of the whole image. Still, having access to uncertainty information and working well with few training examples make GPs a prime candidate for employment in such problems.

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Commonly used kernel functions are typically functions of Euclidean distance [11]. In such scenarios, where the dimensionality is high and the image is dominated with uninformative pixels, these distances make little sense. Conventional kernels do not take into consideration the underlying structure of the image either, i.e. they do not utilise information such as which pixels are close together in image space.

In this paper, we introduce patch kernels to deal with imaging problems where much of the image is uninformative. The key idea is to compute covariances between image patches rather than across the entire image, this allows separate hyperparameters to be optimised for separate subregions in image space, allowing uninformative regions to be ignored and covariance to be measured more consistently. We demonstrate the effectiveness of patch kernels on a synthetic dataset and on age regression from human brain magnetic resonance images. Our approach works well with image and volumetric data as it takes into consideration the underlying 2D or 3D structure during modelling. Our method is a conceptually straightforward yet powerful solution to problems where generating signals are correlated with specific locations in images.

2 Method

We describe patch kernels on volumetric data, though it applies to 2D data in a straightforward manner. Given a volume $\mathbf{X}$, the common approach to perform either GP regression or classification is to flatten $\mathbf{X}$ into a 1D vector. However, all spatial information is lost by the flattening operation. Instead, we break the volume down into several sub volumes (3D patches), calculating covariances on each sub volume separately and combining them by summation. The idea behind this is that we assume that informative features or structures are present in localised regions, as opposed to being present across the whole image. This also holds for background noise, or uninformative regions. For example, in brain magnetic resonance images, most of the volume is background and contains no informative features, further we postulate that specific regions of the brain contribute to different regression targets in varying amounts, hence covariance should depend on such regions in varying quantities. A kernel function which is computed patch-wise then summed can be seen as a direct sum kernel and allows for separate hyper-parameters, namely length scales and variances, for each sub region to be optimised. This means regions which are more or less significant to the regression task can have their influence increased or decreased.

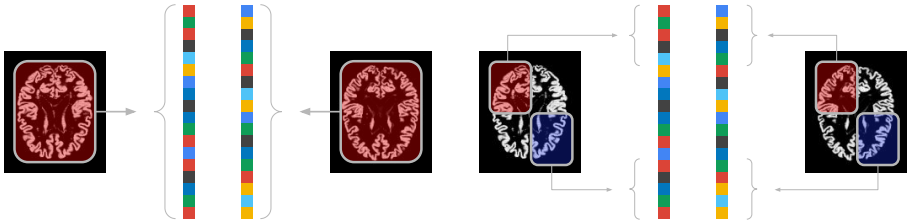

Figure 1: Illustration highlighting the difference between whole image comparison on the left and patch wise comparison on the right, where colour overlays in the image signify different hyper-parameters for the different kernels

Once any valid kernel $k(\mathbf{x}, \mathbf{x}')$ is chosen, and the volume $\mathbf{X}$ has been broken down into P patches $\{\mathbf{x}_1, \ldots, \mathbf{x}_P\}$, the covariance between two volumes $\mathbf{X}$ and $\mathbf{X}'$ is defined as $\sum_P k(\mathbf{x}_p, \mathbf{x}'_p)$. We define patches by specifying patch size and strides, this flexible formulation allows for overlapping patches, reducing the restriction on where structures of interest may lie in a volume. Since this is just a formulation of a kernel, standard inference and optimisation techniques apply without need for modification.

3 Experiments

We experiment on two datasets, a synthetic dataset of MNIST [10] digits embedded into square background images, which consist of Gaussian noise, normalised to be in $[0, 1]$, where each digit is embedded so that it is always in the centre, and noise is generated separately for each embedding. We take one in every 10 images from the MNIST dataset for practical considerations in order to

increase the number of experiments that can be run, leaving 5500 training and 1000 test points. The dataset is turned into a binary classification problem by separating odd and even numbers into two classes. We use the scalable variational GP [7] for inference, using 200 inducing points in each experiment, initialising inducing points by taking 200 random samples from the training data. The

Figure 2: Sample images of MNIST digits embedded into background images with percentage of uninformative noise of 0%, 18.4%, 32.2%, 42.7%, 51% and 88.9% from left to right.

second dataset we demonstrate results on is the IXI dataset [2] with 570 brain MRIs and use SPM12 [3] to pre-process the data. All T1-weighted images had grey matter probability maps extracted using SPM12 default settings and were registered to Montreal Neurological Institute (MNI) space. The task is to regress patient age from these probability maps. Age regression has been shown to provide a useful neuroimaging technique to study brain ageing in the context of brain disease [4]. For our experiments the images are also down-sampled at a $(3, 3, 3)$ rate, by taking the sum of values within a block. This down-sampled volumes are of size $41 \times 49 \times 41$ reducing memory load, so more experimental results could be demonstrated. We validate results by using 2 fold cross validation.

We utilise GPflow [1] for all experiments. To benchmark our method, we vectorise images and volumes and use the squared exponential (SE) kernel, a standard choice of kernel. We use this same choice of kernel in our patch kernels. When measure the percentage of the image which is uninformative we consider the 28×28 MNIST image as informative, and the background image in which is in embedded as uninformative. Figure 3 shows the effect of uninformative background noise in the image on a GP using a the standard SE kernel. The figure illustrates how quickly performance degrades as we increase the amount of uninformative features in the data. Increasing the percentage of uninformative information past 60% causes the standard kernel to dramatically drop in accuracy. However, the patch kernel consistently performs well, this is because patch kernels allows for different length scales and variances for different regions. This means contributions to covariance from uninformative regions can be reduced and regions where generating signals are can be increased. As previously mentioned, this can be considered as a regularised form of ARD.

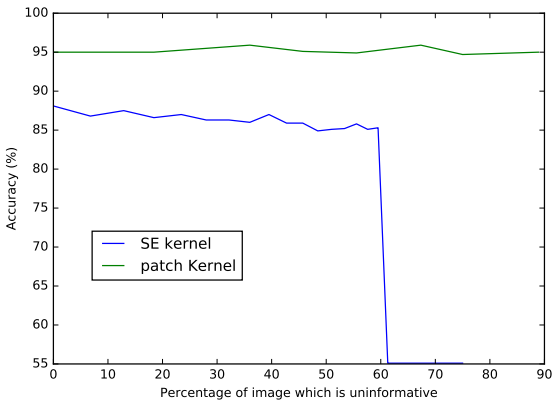

Figure 3: Gaussian process classification accuracy vs. amount of corruption in the image. Where patch sizes and strides of 7×7 were used for benchmarking purposes.

We show the viability of patch kernels on images of size 84×84, where noise accounts for 88.89% of the image and GPs with standard kernels cannot find a non-trivial solution. To obtain an upper bound for the performance of patch kernels we can use the optimal patch size and step size for this setting. Using a 28×28 patches and strides of 28 in both height and width, we know that we should be able to recover our benchmark case. Not only is the patch kernel able to obtain good accuracy with these settings, but we can further improve accuracy by using smaller patches, which is encouraging.

Allocating different covariance contributions from different areas of the informative image seems to help make better predictions. We also find experimentally that when patch sizes become too large the accuracy drops: Despite the signal being present in the patch, noise still dominates. Unlike with

Table 1: Accuracy of patch kernel GPs with different patch and stride values

Patch size	Step size	Accuracy
28	28	88.2 %
14	14	**95.2 %**
56	28	55.1 %

synthetic data, in real-world applications we do not know a priori what the optimal patch and stride sizes should be. Figure 4 shows results from a grid search over the space of parameters, where colour values correspond to the mean absolute error obtained for age regression on the IXI dataset for the different parameter settings. Table 2 gives a more detailed breakdown of results shown in the top left of the figure, where the best results are found. We show clearly that patch kernels outperform the standard SE kernel.

Table 2: Mean absolute error on IXI dataset

	MAE
Standard SE kernel	5.547
Patch Kernel (patch size 2, stride 2)	5.415
Patch Kernel (patch size 4, stride 2)	5.027
Patch Kernel (patch size 4, stride 4)	**4.939**
Patch Kernel (patch size 8, stride 2)	5.119
Patch Kernel (patch size 8, stride 4)	5.084
Patch Kernel (patch size 8, stride 8)	5.385
Patch Kernel (patch size 12, stride 4)	5.198

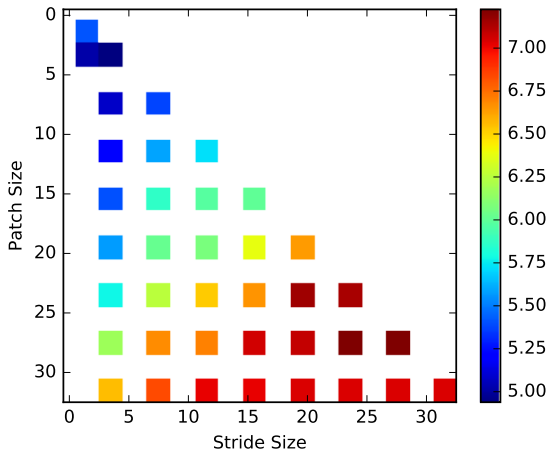

Figure 4: Mean absolute error for GPs using patch kernels, with different patch and stride values

4 Discussion

We propose the use of patch kernels for image to number regression and classification, postulating that informative image features are found in localised regions. Showing experimentally that when generating signals are localised patch kernels outperform standard kernels and that taking advantage of spatial information in image tasks is very useful. Our work opens the door to more research in the field of applying GPs to imaging tasks, such as image-based diagnosis, that would benefit from a fully Bayesian model with uncertainty estimates, an important property in clinical applications.

References

[1] Gpflow. https://github.com/GPflow/GPflow.

[2] Ixi dataset. http://brain-development.org/ixi-dataset/.

[3] Statistical parametric mapping. http://www.fil.ion.ucl.ac.uk/spm/software/spm12/.

[4] J. H. Cole, R. Leech, and D. J. Sharp. Prediction of brain age suggests accelerated atrophy after traumatic brain injury. *Annals of Neurology*, 2015.

[5] M. P. Deisenroth, D. Fox, and C. E. Rasmussen. Gaussian processes for data-efficient learning in robotics and control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):408–423, 2015.

[6] P. Gao, A. Honkela, M. Rattray, and N. D. Lawrence. Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics*, 24(16):i70–i75, 2008.

[7] J. Hensman, A. G. d. G. Matthews, and Z. Ghahramani. Scalable variational gaussian process classification. In *Proceedings of AISTATS*, 2015.

[8] D. Krige. A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of Chemical, Metallurgical, and Mining Society of South Africa*, 1951.

[9] M. Lê, J. Unkelbach, N. Ayache, and H. Delingette. GPSSI: Gaussian process for sampling segmentations of images. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9351:38–46, 2015.

[10] Y. LeCun and C. Cortes. MNIST handwritten digit database. 2010.

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