PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Gaussian processes with optimal kernel construction for neuro-degenerative clinical onset prediction

Liane S. Canas, Benjamin Yvernault, David M. Cash, Erika Molteni, Tom Veale, et al.

Liane S. Canas, Benjamin Yvernault, David M. Cash, Erika Molteni, Tom Veale, Tammie Benzinger, Sébastien Ourselin, Simon Mead, Marc Modat, "Gaussian processes with optimal kernel construction for neuro-degenerative clinical onset prediction," Proc. SPIE 10575, Medical Imaging 2018: Computer-Aided Diagnosis, 105750G (27 February 2018); doi: 10.1117/12.2293242

SPIE.

Event: SPIE Medical Imaging, 2018, Houston, Texas, United States

Gaussian Processes with optimal kernel construction for neuro-degenerative clinical onset prediction

Liane S Canas[a*], Benjamin Yvernault[a], David M. Cash[a,b], Erika Molteni[a], Tom Veale[c], Tammie Benzinger[d], Sébastien Ourselin[a], Simon Mead[e], and Marc Modat[a,b]

[a]Translational Imaging Group, Centre for Medical Image Computing, University College London, UK
[b]Dementia Research Centre, UCL Institute of Neurology, London, UK
[c]Institute of Neurology, University College London, UK
[d]Washington University School of Medicine, St. Louis, MO, USA
[e]MRC Prion Unit, Department of Neurodegenerative Disease, UCL Institute of Neurology, London, UK

ABSTRACT

Gaussian Processes (GP) are a powerful tool to capture the complex time-variations of a dataset. In the context of medical imaging analysis, they allow a robust modelling even in case of highly uncertain or incomplete datasets. Predictions from GP are dependent of the covariance kernel function selected to explain the data variance. To overcome this limitation, we propose a framework to identify the optimal covariance kernel function to model the data.The optimal kernel is defined as a composition of base kernel functions used to identify correlation patterns between data points. Our approach includes a modified version of the Compositional Kernel Learning (CKL) algorithm, in which we score the kernel families using a new energy function that depends both the Bayesian Information Criterion (BIC) and the explained variance score. We applied the proposed framework to model the progression of neurodegenerative diseases over time, in particular the progression of autosomal dominantly-inherited Alzheimer's disease, and use it to predict the time to clinical onset of subjects carrying genetic mutation.

Keywords: Gaussian Process, Covariance kernel functions, Neurodegenerative Diseases, Compositional kernel Learning, Disease progression model, Clinical onset prediction

1. DESCRIPTION OF PURPOSE

Neurodegenerative diseases are characterised by a progressive alteration of the physiology and morphology of the brain, which leads to an irreversible impairment of cognitive functions. These processes result in levels of brain atrophy that can be quantified using high-resolution structural MRI. Brain atrophy has proven to be an imaging biomarker that tracks well with disease progression and is present during the preysymptomatic phase of the disease. Accurate knowledge of the disease progression pattern in the stage that precedes clinical symptoms might lead to intervention at an earlier stage of the disease that would be more beneficial to the patient. One group of presymptomatic individuals of particular interest are those that carry an autosomal dominantly-inhertited mutation known to cause Alzheimer's disease (AD). While these individuals are nearly certain to develop AD, more accurately predicting their proximity to symptom onset would be beneficial to better understand the timing of different disease processes and thus plan trials accordingly. In addition, improving the accuracy of predicting onset could reduce and prolong anxiety for at-risk individuals.

The majority of models used to study AD learn the behaviour of the biomarkers over time, applying this information to perform a subjects' stratification between healthy control, mild cognitive impairment or AD. These predictive models that stage subjects are for example based on event-based models,[1] or mixed effect models.[2] These paradigms require the knowledge of the clinical onset and/or the binary discretisation of the biomarkers

* Further author information: liane.canas.15@ucl.ac.uk

Medical Imaging 2018: Computer-Aided Diagnosis, edited by Nicholas Petrick, Kensaku Mori, Proc. of SPIE Vol. 10575, 105750G · © 2018 SPIE · CCC code: 1605-7422/18/$18 · doi: 10.1117/12.2293242

changes. To overcome the discrete characterisation of the subjects, Challis *et al*[3] proposed a Bayesian framework for subjects stratification. However, the proposed method only concerns the diagnosis and stratification of the subjects, neglecting their prognosis. To tackle this limitation, Hyun and collaborators[4] suggested to use a Gaussian process (GP) to delineate the biomarkers trajectories, whilst incorporating the spatial and temporal features of longitudinal neuroimaging data for both diagnostic prognosis.

GP are a sensible way to represent the data (i.e, the features considered such as brain regions volumes) in a non-parametric way with a mean and a covariance kernel function. The covariance kernel function detects the pattern of the features that most precisely explain the response variable (i.e., diagnosis and/or proximity to clinical onset). Consequently, the performance of the GP is highly dependent of the kernel function used in a specific context.[5,6] In order to improve the performance of kernel-based prediction models, Duvenaud *et al*[7] have proposed a structure discovery algorithm through a compositional kernel search. Their study has successfully shown that, in supervised prediction tasks, the automatically learning of kernels outperform both variety of kernel classes commonly used and kernel combination methods. As a brute force scheme is impractical due to the highly dimensional problem, their algorithm searches over a space of based kernels and operations using a greedy search approach: at each stage it chooses the highest scoring kernel and expands it by applying operations, such as addition or multiplication, with other basis kernel functions. However, this method does not account for the replacement of the kernel selected in a previous level; furthermore, the method is design to preferentially select of the best model based on the Bayesian Information Criterion[8] (BIC), which it only takes into account the balance between the marginal likelihood of the predictions estimated based on the training set, and the complexity of the model.

In our approach, the optimal kernel search is based on depth-first search in a pre-pruned tree, applying a greedy search to select the highest scored kernel in each layer of the tree. Our approach also considers possible that a given branch of our search tree could be better that the one selected in the previous layer. We also introduce a new energy function to evaluate the kernels function performance and a cost-function to select the optimal kernel. The energy function introduced takes into account the model predictions for the validation set and also considers the balance between the model performance and its complexity by including a term in the energy function in which the BIC is considered.

2. METHODS

We implemented a non-parametric kernel-based model $\mathcal{M} : y = f(X)+\varepsilon,\ f \sim \mathcal{GP}(\mu_f; K),\ \varepsilon \sim \mathcal{N}(\mu_\varepsilon; \sigma)$ to infer the time to clinical onset of Alzheimer's disease, y, given a set of biomarkers $\mathbf{X} \in \mathcal{X}$ features space, composed by the volume of brain regions extracted from MRI. The function $f(x)$ describes the variance of the features that explains the response variable, by implementing a GP with μ_f equals to 0 and covariance kernel function K, used to determine the pattern of the inductive generalization of the features considered.

Our algorithm, represented in figure 1, requires only a set of covariance kernel functions $b \in \mathcal{B}$, defined in this paper as basis kernel functions, the design matrix of the features $\mathbf{X}$, a *c-by-N* matrix of c brain regions, extracted from N subjects, and the response variable vector $\mathbf{y}$, a *1-by-N* vector with the time to onset corresponding to each of the observations in the matrix $\mathbf{X}$. We considered as basis covariance function k with hyperparameters θ: the linear kernel ($k_{LIN}, \theta_{LIN} \in \{\sigma_b, \sigma_v, w\}$, equation 1), periodic kernel ($k_{PER}, \theta_{PER} \in \{\sigma, p, w\}$, equation 2), squared exponential ($k_{SE}, \theta_{SE} \in \{\sigma, w\}$, equation 3), matern function ($k_{MA}, \theta_{MA} \in \{\sigma, r, w\}$, equation 4) and linear logistic ($k_{LINLOG}, \theta_{LINLOG} \in \{W, a, b\}$, equation 5).

$$k_{LIN}(x, x') = \sigma_b^2 + \sigma_v^2(x - w)(x' - w) \tag{1}$$

$$k_{PER}(x, x') = \sigma^2 \exp\left(-\frac{2\sin^2(\frac{\pi(x-x')}{p})}{w^2}\right) \tag{2}$$

$$k_{SE}(x, x') = \sigma^2 \exp\left(-\frac{(x - x')}{2w^2}\right) \tag{3}$$

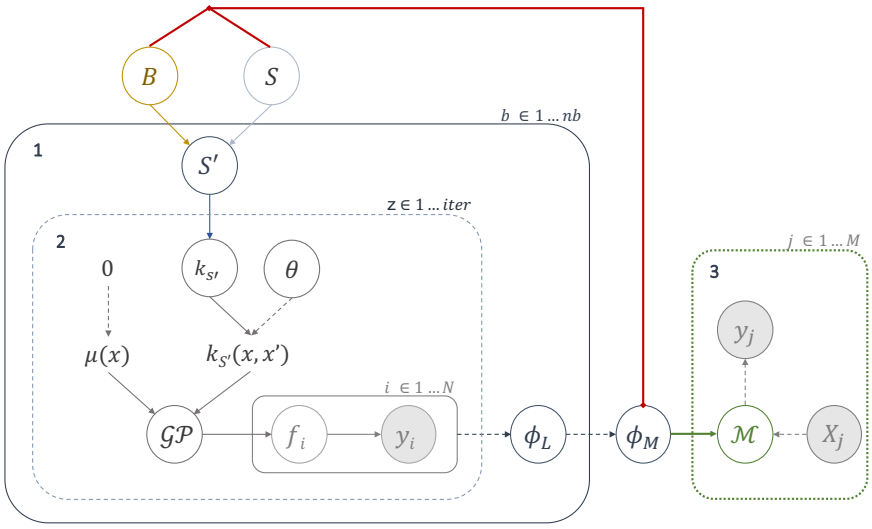


Figure 1. Graphical representation of the framework. $\mathcal{B}$ is the set of basis kernels. $\mathcal{S}$ is the kernel function initialised at each layer, whilst $\mathcal{S}'$ represents the kernel function after the expansion. f denotes the latent function, whereas y is the vector of values predicted using the estimated latent function. The subscripts i denote sampled latent values for each point to a maximum of N data points. $\mathcal{M}$ is the model selected and used to estimate y_j predictions of j observation given a set of features $\mathbf{X}_j$. The process 1 is repeated until the cost function ϕ_M converges - red line. The white circles represent functions or set of functions and the dark circles represent values of variables.

$$k_{MA}(x,x') = \sigma^2 \exp\left(1 + \frac{(x-x')\sqrt{r}}{w}\right)\left(-\frac{(x-x')\sqrt{r}}{w}\right), r \in \{3,5\} \tag{4}$$

$$k_{LINLOG}(x,x') = W \text{logit}^{-1}(ax+b) - 0.5 \tag{5}$$

The model initialisation in the first layer $\mathcal{S}$ is $\varnothing$ followed by the evaluation of the algorithm performance when each one of the basis kernel functions is used. $\mathcal{S}'$ is the result of the expansion of $\mathcal{S}$ by operations with $\mathcal{B}$. This procedure (figure 1 - process 1) is performed until all basis kernel functions are tested.

The best kernel in each layer is selected by maximisation of the energy function, equation 7, where α and β are constants, $\hat{y}_{i,b}$ is the model estimations and $y_{i,b}$ is the vector of observed values correspondent to the response variable. The first term of the equation 7 evaluates the accuracy of the prediction of the model in the testing set, whereas the second term constrains the complexity of the model based on the marginal likelihood of the predictions of the training set. Regarding the parameters of ϕ_L, it is required to compute the predictions of the model based on the kernel in analysis. Further, the estimation of $\hat{y}_{i,b}$ requires to find the best hyperparameters of each kernel. The hyperparameters θ of the kernel functions are estimated via the maximisation of the marginal likelihood of the model, $p(y|X,\theta)$, as described in equation 6; i.e., the marginalisation over the kernel parameters is performed by maximum *a posteriori* algorithm (MAP), and that the hyperparameters θ are estimated by bootstrapping.

$$\{\hat{\theta},\hat{\sigma}\} = \text{argmax}_{\sigma,\theta} p(\theta,\sigma|\mathcal{M}) = \text{argmin}_{\sigma,\theta}[-\text{log}p(\mathcal{M}|\theta,\sigma) + \text{log}p(\theta,\sigma)] \tag{6}$$

$$\text{argmax}_{b\in B}(\phi_L(b)) = \alpha\left[1 - \sum_{i=1}^{N}\left(\frac{\text{var}(\hat{y}_{i,b} - y_{i,b})}{\text{var}(y_{i,b})}\right)\right] + \beta\left[1 - \exp\left(\frac{BIC_b}{\max(BIC)}\right)\right] \tag{7}$$

In each of the kernel functions computed, we also consider a separate length scale $\boldsymbol{w}$ for each predictor p, $\mathbf{w} \in \{w_1 ... w_p\}$, by implementing the automatic relevance determination (ARD) method. This method allows

to establish different weights - relevance - for the features considered, as well as it allows to include in the same model biomarkers with different scales. By including the ARD approach we automatically extract relevant features among redundant predictors.

The following layers of the tree correspond to the composition of a kernel function based on a set of operations between the basis kernel functions $b \in \mathcal{B}$ and the covariance function $\mathcal{S}$ obtained in the previous layer. We consider as possible operations the addition, product and replacement of basis kernel, replacement of the branch previously chosen. Considering the pre-pruning behaviour of our algorithm, we do not need to define the maximum search depth. Rather, we defined a cost-function:

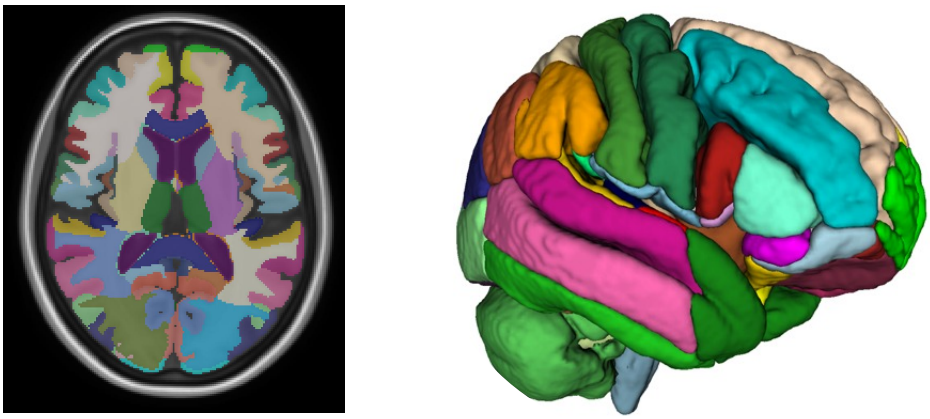
$$\phi_M : (|\phi_L(l-1) - \phi_L(l)| \leq 0.01 \cup \phi_L(l) \geq 0.9)$$

which defines the stopping criterion of the kernel search, where l is the number of layers.

3. RESULTS

3.1 Imaging Data and Pre-processing

We evaluate the effectiveness of our approach by predicting the years to clinical onset of Inherited Alzheimer's disease patients, using clinical and imaging data of the Dominantly Inherited Alzheimer Network (DIAN) study.[9] Our sample is composed of 320 controls, 240 symptomatic subjects and 70 subjects that converted to symptomatic. The subjects' T1-weighted images were processed using the Geodesical Information Flows[10] algorithm, and the volumes of brain region were used as features, as illustrated by Figure 2. We regress the impact of confounding effects in the features, such as age and the total intracranial volume. The feature selection was performed using an elastic net regression.



A: Extraction of structural features from T1w. **B:** 3D representation of the T1w parcellation .

Figure 2. Imaging biomarkers used as features in the model: structural features extracted from the T1w images.

3.2 Prediction Model

We divided our sample in three sub-samples: training set that corresponds to 70% of the asymptomatic and converted subjects, the validation set comprises 20% of the subjects belonging to the aforementioned groups and testing set, which correspond to the remaining 10% of our sample. The optimal kernel is optimized using the training and testing sets. To our knowledge, currently there is no other method to predict the time to clinical onset for AD using GP. Bearing this in mind, the validation of our approach aims to justify that the optimal kernel selection is an effective way to detect the pattern of the features considered, without a selection of the function that better explains their evolution over time.

When compared with kernel functions defined *a priori* to explain time-series, our approach selected a kernel function with equivalent coefficient of determination, but lower BIC (table 1). This fact suggests that our approach is able to find a function that explains conveniently the variance of the features, with lower level of

Table 1. Evaluation of the kernels used to predict the time to clinical onset, given a set of structural features. We compared the results obtained with different functions commonly used to explain time-series datasets, a variation of the Compositional Kernel Learning,[7] and our approach.

Approach	*R squared*	*RMSE*	*BIC*
GP - LIN	0.401	5.61	-12.5
GP - (LIN+SE)	0.392	5.64	53.2
GP - SE (ARD)	0.423	5.48	48.1
GP - Kernel optimisation	0.407	5.60	-2.67
Our approach	0.432	5.52	12.3

complexity considering the likelihood of the prediction attending to the BIC achieved. The algorithm selected a linear combination of LINLOG functions, which supports the biological assumptions regarding the features used. The results also support the assumption that this approach may be extended to other diseases, without modelling explicitly the pattern of features used to characterised them. We also evaluate the performance of the model for the prediction of the time to clinical onset of the subjects in validation set. Note however that the prediction of the time to onset is highly dependent of the data used to train the model. In this study, a large number of subjects considered have not converted to symptomatic to the date; whereby, the time to onset used as response variable is not the real age of clinical onset, but an estimation based on the age of clinical onset of the their family. In the future, we aim to validate our approach in a dataset that accounts for the uncertainty of the labels used as response variable.

4. DISCUSSION

We proposed a non-parametric Bayesian approach to predict the time to clinical onset of Alzheimer's disease. Prediction is based on a probabilistic staging of the subjects based on the biomarkers evolution over time modelled by GP. Our framework does not require any explicitly modelling of the pattern of the biomarkers over time. Towards the goal of automating the design of kernel function that better models the biomarkers considered, we introduced a space of composite kernels defined as sums, products and/or replacement of a small number of based kernels. We considered five different families of kernels that yield decompositions of the biomarkers samples into interpretable subset of samples over time.

The results have shown that the learned structure is capable of accurate extrapolation in a complex time-series, such as the evolution of brain volumes over time, and it is competitive with the other methods tested given this prediction task. The main limitation of the proposed method is the complexity of the kernel structure obtained when in presence of a high level of noise in the data. Furthermore, the evaluation of the structures found requires the estimation of the full model; thus, the metrics taken into account in table 1 to evaluate the performance of the model include the errors associated to the hyperparameters estimation and inference.

We aim to extend and further validate the proposed framework with a wider set of features extracted from other imaging modalities, as well as for other neurodegenerative diseases. We aim also to find additional support to the assumption that our model is as general as possible and it is able to characterise any type of features despite their nature. By assuming that the inter-modality relationship can be modelled as a multi-task paradigm: contribution of independent functions that explain the biomarkers progression - we will implement an Additive Gaussian process to predict the evolution of symptoms.

Moreover, the model presented in this study may also be extended to consider longitudinal data and to model individual brain changes; therefore, the model represents a promising instrument for subjects' diagnosis and prognosis.

Acknowledgements. This work is supported by the EPSRC-funded UCL Centre for Doctoral Training in Medical Imaging (EP/L016478/1), the Department of Healths NIHR-funded Biomedical Research Centre at University College London Hospitals, the Wolfson Foundation (PR/ylr/18575), MRC (UK) at University College London Hospitals NHS Foundation Trust (540649), and Alzheimer's Society UK (AS-PG-15-025).

REFERENCES

[1] Fonteijn, H. M. et al., "An event-based model for disease progression and its application in familial Alzheimer's disease and huntington's disease," *NeuroImage* **60**(3), 1880–1889 (2012).

[2] Ge, T. et al., "A kernel machine method for detecting effects of interaction between multidimensional variable sets: An imaging genetics application," *NeuroImage* **109**, 505–514 (2015).

[3] Challis, E. et al., "Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI," *NeuroImage* **112**, 232–243 (2015).

[4] Hyun, J. W. et al., "STGP: Spatio-temporal Gaussian process models for longitudinal neuroimaging data," *NeuroImage* **134**, 550–562 (2016).

[5] Rasmussen, C. and Williams, C., [*Gaussian processes for machine learning.*], vol. 14 (2004).

[6] Hwang, Y. et al., "Automatic construction of nonparametric relational regression models for multiple time series," *The 33rd International Conference on Machine Learning (ICML 2016)* **48** (2016).

[7] Duvenaud, D. et al., "Structure Discovery in Nonparametric Regression through Compositional Kernel Search," *Advances in Neural Information Processing Systems* **28** (2013).

[8] Schwarz, G., "Estimating the dimension of a model," *The Annals of Statistics* **6**(2), 461–464 (1978).

[9] Morris, J. C. et al., "Developing an international network for Alzheimer research: The Dominantly Inherited Alzheimer Network.," *Clinical investigation* **2**, 975–984 (oct 2012).

[10] Cardoso, M. J. et al., "Geodesic information flows: Spatially-variant graphs and their application to segmentation and fusion," *IEEE Transactions on Medical Imaging* **34**(9), 1976–1988 (2015).