# Detecting Premature Ventricular Contractions in ECG Signals with Gaussian Processes

F Melgani[1], Y Bazi[2]

[1]Dept of Information Engineering and Computer Science, Univ of Trento, Italy
[2]College of Engineering, Al-Jouf Univ, Saudi Arabia

**Abstract**

*The aim of this work is twofold. First, we propose to investigate the capabilities of a new Bayesian approach for detecting premature ventricular contractions (PVCs), namely the Gaussian process (GP) approach. Second, we report an experimental comparison of different kinds of ECG signal representations, which are the standard temporal signal morphology, the discrete wavelet transform domain, the S-transform characteristics and the high-order statistics. In general, the obtained classification results show that the GP detector can compete seriously with state-of-the-art methods since it allows to yield better overall accuracy as well as better sensitivity. In addition, among the different kinds of features explored, those based on high-order statistics appear to be the best compromise between accuracy and computational time for PVC detection.*

## 1. Introduction

The detection and classification of ECG arrhythmias such as premature ventricular contraction (PVC) is essential for the treatments of patients with heart disease. For such purpose, simple as well as sophisticated algorithms exploiting different classification strategies with different features representations of the ECG signals have been proposed [1]-[3].

In this paper, we propose to investigate the capabilities of a new Bayesian approach for detecting premature ventricular contractions (PVCs), namely the Gaussian process (GP) approach [4]-[6]. In addition, we report an experimental comparison of different kinds of ECG signal representations, which are the standard temporal signal morphology, the discrete wavelet transform domain [1], the S-transform [7], which is an extension to the ideas of wavelet transform, and the high-order statistics [3]. The main idea of GPs is to assume that the probability of belonging to a class label for an input beat is monotonically related to the value of some latent function at that beat. Such monotonic relationship is defined according to a so-called squashing function (e.g., the logistic and the probit functions). A Gaussian process prior characterized by a zero mean and a covariance matrix embedding a set of hyperparameters is placed on this latent function. The inference is made by integrating over the latent function values through an analytical approximation based on the Laplace technique. In the prediction phase, the predictive mean and variance for the approximate Gaussian posterior over the latent variable of the considered beat are first computed. Then, the approximate predictive distribution for the beat label is derived either analytically or by approximation depending on the adopted squashing function.

## 2. Gaussian process classification

Let us consider a supervised binary classification problem. Let us consider a training set $D$=($\mathbf{X}$,$\mathbf{y}$) consisting of a matrix of training beats $\mathbf{X} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad ... \quad \mathbf{x}_N]^T$ where $N$ is the number of beats and $\mathbf{y} = [y_1 \quad y_2 \quad ... \quad y_N]^T$ is the corresponding target vector. To each vector $\mathbf{x}_i \in \Re^d$ ($i$ = 1, 2, …, $N$), we associate a target (label) $y_i \in \{-1, +1\}$. Given this training set $D$, we aim to predict the label of a new test beat $\mathbf{x}_*$ by computing the output probability $p(y_* \mid D, \mathbf{x}_*)$ .

In GPC, the probability of belonging to a class label $y_i$=+1 for an input sample $\mathbf{x_i}$ is monotonically related to the value of some latent function $f_i$. Such monotonic relationship is defined according to a squashing function; which can take several forms (e.g., logistic and probit functions):

$$p(y_i = +1 \mid f_i) = \begin{cases} \dfrac{1}{1+\exp(-y_i f_i)} & \text{logistic} \\ \Phi(y_i f_i) & \text{probit} \end{cases} \quad (1)$$

where $\Phi$ is the Gaussian cumulative distribution function given by: $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) dx$

A Gaussian process prior (GP) characterized by a zero mean and a covariance matrix embedding a set of hyperparameters □ is placed on this latent function. The

prediction of the output probability for the test beat $\mathbf{x}_*$ is obtained by integrating over the latent function $f_*$ as follows:

$$p(y_* = +1 \mid D, \mathbf{x}_*, \boldsymbol{\Theta}) = \int p(y_* \mid f_*, \boldsymbol{\Theta}) p(f_* \mid D, \mathbf{x}_*, \boldsymbol{\Theta}) df_* \quad (2)$$

The second part of the integral (2) represents the distribution of the latent variable corresponding to the test beat $\mathbf{x}_*$. It is obtained by further integrating over $\mathbf{f} = [f_1 \quad f_2 \quad \ldots \quad f_N]$:

$$p(f_* \mid D, \mathbf{x}_*, \boldsymbol{\Theta}) = \int p(f_* \mid \mathbf{X}, \mathbf{x}_*, \mathbf{f}, \boldsymbol{\Theta}) p(\mathbf{f} \mid D, \boldsymbol{\Theta}) d\mathbf{f} \quad (3)$$

where $p(\mathbf{f}|D,\Box)$ is the posterior over the latent variables:

$$\begin{aligned} p(\mathbf{f} \mid D, \boldsymbol{\Theta}) &= p(\mathbf{y} \mid \mathbf{f}, \boldsymbol{\Theta}) p(\mathbf{f} \mid \mathbf{X}, \boldsymbol{\Theta}) / p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\Theta}) \\ &= \left( \prod_{i=1}^{N} p(y_i \mid f_i, \boldsymbol{\Theta}) \right) p(\mathbf{f} \mid \mathbf{X}, \boldsymbol{\Theta}) / p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\Theta}) \end{aligned} \quad (4)$$

$p(\mathbf{y}|\mathbf{f},\Box)$ is the probability of each observed class label given the latent function value. It can be one of the forms adopted in (1). $p(\mathbf{y}|\mathbf{X},\Box)$ is the marginal likelihood and $p(\mathbf{f}|\mathbf{X},\Box)$ is the GP prior over the latent functions:

$$p(\mathbf{f} \mid \mathbf{X}, \boldsymbol{\Theta}) = \frac{1}{(2\pi)^{N/2} |\mathbf{K}|} \exp\left\{ -\frac{1}{2} \mathbf{f} \mathbf{K}^{-1} \mathbf{f} \right\} \quad (5)$$

where each term of the covariance function $\mathbf{K}$ is a function of $\mathbf{x}_i$ and $\mathbf{x}_j$. A popular covariance function is the squared exponential (or Gaussian RBF), i.e.

$$k(x_i^{(m)}, x_j^{(m)}) = \sigma^2 \exp\left( -\frac{\sum_{m=1}^{d} (x_i^{(m)} - x_j^{(m)})^2}{2l^2} \right) \quad (6)$$

where $\sigma$ is the variance and $l$ is the length scale; they form the hyperparameter vector $\Box$, i.e. $\Box=[l\ \sigma]$.

Since the integrals in equations (2) and (3) are not analytically tractable due to the nonlinearity in the likelihood terms, analytical approximation or Monte Carlo methods have been adopted. In next section, we describe the well known analytical approximation based on the Laplace algorithm.

## 3. Laplace algorithm

The Laplace approximation uses a Gaussian approximation $q(\mathbf{f}|D,\Box)$ to the non-Gaussian posterior in the integral (3). This approximation is based on the second order Taylor expansion of $\log p(\mathbf{f}|D,\Box)$ around the maximum of the posterior:

$$p(\mathbf{f} \mid D, \boldsymbol{\Theta}) \cong q(\mathbf{f} \mid D, \boldsymbol{\Theta}) = N(\mathbf{f} \mid \hat{\mathbf{f}}, \mathbf{A}^{-1}) \quad (7)$$

Where $\hat{\mathbf{f}}$ and $\mathbf{A}$ are the mean and the covariance matrix, respectively, and they are given by:

$$\hat{\mathbf{f}} = \operatorname{argmax}_{\mathbf{f}} \ p(\mathbf{f} \mid D, \boldsymbol{\Theta}) \quad (8)$$

$$\mathbf{A} = -\nabla\nabla \log p(\mathbf{f} \mid D, \boldsymbol{\Theta}) \big|_{\mathbf{f}=\hat{\mathbf{f}}} \quad (9)$$

The covariance matrix represents the Hessian of the negative log posterior at the maximum point. In order to compute $\hat{\mathbf{f}}$ and $\mathbf{A}$ we can use the posterior $p(\mathbf{f}|D,\Box)$ formulated in (4). By taking the logarithm of this posterior and introducing the expression (5) for GP priors, we obtain the following expression:

$$\begin{aligned} \boldsymbol{\Psi}(\mathbf{f}) = \log p(\mathbf{y} \mid \mathbf{f}, \boldsymbol{\Theta}) - \frac{1}{2} \mathbf{f}^T \mathbf{K}^{-1} \mathbf{f} - \frac{1}{2} \log |\mathbf{K}| \\ - \frac{N}{2} \log 2\pi - \log p(y \mid \mathbf{X}, \boldsymbol{\Theta}) \end{aligned} \quad (10)$$

Differentiating equation (10) with respect to $\mathbf{f}$ we obtain:

$$\begin{cases} \nabla \boldsymbol{\Psi}(\mathbf{f}) = \nabla \log p(\mathbf{y} \mid \mathbf{f}, \boldsymbol{\Theta}) - \mathbf{K}^{-1} \mathbf{f} \\ \nabla\nabla \boldsymbol{\Psi}(\mathbf{f}) = \nabla\nabla \log p(\mathbf{y} \mid \mathbf{f}, \boldsymbol{\Theta}) - \mathbf{K}^{-1} \end{cases} \quad (11)$$

At the maximum of $\boldsymbol{\Psi}(\mathbf{f})$ we have:

$$\hat{\mathbf{f}} = \mathbf{K}\left( \nabla \log p(\mathbf{y} \mid \hat{\mathbf{f}}, \boldsymbol{\Theta}) \right) \quad (12)$$

and the covariance matrix is approximated by the curvature at the mode of the negative inverse Hessian:

$$\mathbf{A} = -(\nabla\nabla \boldsymbol{\Psi}(\hat{\mathbf{f}}))^{-1} = (\mathbf{K}^{-1} + \mathbf{W})^{-1} \quad (13)$$

where:

$$\mathbf{W} = -\nabla\nabla \log(p(\mathbf{y} \mid \hat{\mathbf{f}}, \boldsymbol{\Theta})) \quad (14)$$

Since (12) is nonlinear, the computation of $\hat{\mathbf{f}}$ is achieved by numerical methods such as the Newton method. After this computation, the Laplace approximation to the posterior is completely defined by:

$$q(\mathbf{f} \mid D, \Theta) = N(\hat{\mathbf{f}}, (\mathbf{K}^{-1} + W)^{-1}) \tag{15}$$

The prediction of the test beat $\mathbf{x}_*$ is evaluated by replacing the computed Gaussian approximation into the equation (2):

$$q(y_* = +1 \mid D, \mathbf{x}_*, \Theta) = \int p(y_* \mid f_*, \Theta) q(f_* \mid D, \mathbf{x}_*, \Theta) df_* \tag{16}$$

where $q(f_* \mid D, \mathbf{x}_*, \Theta)$ is a Gaussian with mean and variance given as follows:

$$\begin{cases} \boldsymbol{\mu}_* = \mathbf{k}(\mathbf{x}_*)^T \mathbf{K}^{-1} \hat{\mathbf{f}} \\ \boldsymbol{\sigma}_*^2 = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}(\mathbf{x}_*)^T \left(\mathbf{K} + \mathbf{W}^{-1}\right)^{-1} \mathbf{k}(\mathbf{x}_*) \end{cases} \tag{17}$$

and $\mathbf{k}(\mathbf{x}_*) = [k(\mathbf{x}_1, \mathbf{x}_*) \quad k(\mathbf{x}_2, \mathbf{x}_*) \quad \ldots \quad k(\mathbf{x}_N, \mathbf{x}_*)]^T$ is a vector of prior covariances between $\mathbf{x}_*$ and the training input matrix $\mathbf{X}$. It is worth noting that if the probit form is adopted for the squashing function, then the prediction (14) can be evaluated analytically:

$$q(y_* = +1 \mid D, \mathbf{x}_*, \Theta) = \Phi\left(\frac{\boldsymbol{\mu}_*}{\sqrt{1 + \boldsymbol{\sigma}_*^2}}\right) \tag{18}$$

The general form of the Laplace algorithm can be summarized as follows:

***Training phase:***

*Step* 1: Given the training set $D$=($\mathbf{X}$,$\mathbf{y}$) and the hyperparameter vector □, compute the covariance matrix $\mathbf{K}$.

*Step* 2: Compute $\hat{\mathbf{f}}$ from (12) using the iterative Newton method.

*Step* 3: Compute the Hessian matrix A related to the negative log posterior at the maximum point $\hat{\mathbf{f}}$ from (13).

***Test phase:***

*Step* 4: Given a test beat $x_*$, compute $q(y_* = +1 \mid D, \mathbf{x}_*, \Theta)$ according to (18), and if it is greater or equal to 0.5 assign the label '+1' to $x_*$, otherwise choose label '-1'.

# 4. Experimental results

The experiments were conducted on the basis of ECG data from the MIT-BIH arrhythmia database [8]. The beats refer to the recordings of 45 patients. These recordings were subdivided into two groups, one of 18 and the other of 27 recordings. While the first group was used both for training and testing purposes, the second one was exploited just for testing the detection system on completely unseen recordings. For feeding the classification process, we adopted in this study different representation of the ECG signals, which are the standard temporal signal morphology, the discrete wavelet transform domain, the S-transform characteristics and the high-order statistics. In addition, for each representation, we considered also three temporal features that are the QRS complex duration, the RR interval (i.e., time span between two consecutive R points representing the distance between the QRS peaks of the present and previous beats), and the RR interval averaged over the ten last beats [2].

Table 1. Classification results reported in terms of overall accuracy (OA), sensitivity (Se) and specificity (Sp) achieved on both 18 and 27 records.

| | 18 Records | | | 27 Records | | |
|---|---|---|---|---|---|---|
| Feature Typology | OA (%) | Se (%) | Sp (%) | OA (%) | Se (%) | Sp (%) |
| Morph. | 94.8 | 95.3 | 94.7 | 94.1 | 77.6 | 94.9 |
| W2 | 94.8 | 95.3 | 94.7 | 94.1 | 77.6 | 94.9 |
| W3 | 94.8 | 95.2 | 94.7 | 94.0 | 77.4 | 94.8 |
| W4 | 94.8 | 95.3 | 94.7 | 94.0 | 77.4 | 94.8 |
| S-transf. | **97.1** | **97.6** | **97.0** | 93.6 | 82.3 | 94.1 |
| HOS2 | 96.4 | 97.2 | 96.3 | **96.9** | **84.7** | **97.5** |
| HOS3 | 95.7 | 96.0 | 95.6 | 90.9 | 88.5 | 91.0 |
| HOS4 | 95.2 | 95.4 | 95.2 | 88.8 | 90.0 | 88.8 |

Table 2. Computational times required by each feature typology during the training and test phases (the test time refers to a single beat).

| Feature Typology | Training Time [s] | Test Time [ms] |
|---|---|---|
| Morph. | 416 | 6 |
| W2 | 429 | 9 |
| W3 | 546 | 9 |
| W4 | 592 | 11 |
| S-transf. | 320 | 1130 |
| HOS2 | 774 | 43 |
| HOS3 | 566 | 45 |
| HOS4 | 1103 | 141 |

In order to train the GP classifier and to assess its accuracy, we selected randomly from the 18 records 600 beats for the training set (i.e., 300 samples for both PVC and non-PVC classes, respectively). The classification

performance was evaluated in terms of three standard measures which are: 1) the overall accuracy (OA); 2) the sensitivity (Se); and 3) the specificity (Sp). Concerning the GP classifier, we adopted in the experiments the squared exponential covariance function characterized by the hyperparameter vector □=[$l$ σ]. During the training phase, the determination of this optimal hyperparameter vector is made according to the Bayesian learning procedure described in [7]. Table 1 reports the classification results obtained for the different feature typologies. As can be seen, the best accuracy achieved on the 18 records was obtained for the S-transform since the OA, Se, and Sp were equal to 97.1%, 97.6%, and 97.0%, respectively. Concerning the unseen 27 records, the best accuracy was obtained for the HOS features (i.e., cumulants of the second order) and the OA, Se and Se were equal to 96.9%, 84.7%, 97.5%, respectively. It is worth noting that for all 45 records, the cumulants of the second order showed the best classification accuracy as the OA, Se, and Sp were equal to 96.7%, 90.9%, 96.9%, respectively.

From these results it appears clearly that, among the different kinds of features explored, those based on cumulants of the second order appear to be the best compromise between accuracy and computational time (see Table 2).

## 5. Conclusion

The obtained classification results show that: 1) the GP detector can compete seriously with state-of-the-art methods [1] since it allows to yield better overall accuracy as well as better sensitivity (96.7% and 90.9% against 95.2% and 82.9%, respectively); 2) since the GP detector does not exhibit a particular sensitivity to the curse of dimensionality, all extracted features are exploited and no prior (tricky and time-consuming) feature reduction step is required; 3) the GP detector maintains a high generalization capability when passing from recordings seen during training to completely unseen recordings; 4) among the different kinds of features explored, those based on high-order statistics appear to be the best compromise between accuracy and computational time for PVC detection.

## References


[1] Inan OT, Giovangrandi L, Kovacs J TA. Robust neural network based classification of premature ventricular contractions using wavelet transform and timing interval features 2006; IEEE Trans. Biomedical Engineering; 53:2507-2515.

[2] De Chazal F, Reilly RB. A patient adapting heart beat classifier using ECG morphology and heartbeat interval features 2006; IEEE Trans. Biomedical Engineering; 53:2535-2543.

[3] Osowski S, Linh TH, and Markiewicz T. Support vector machine-based expert system for reliable heart beat recognition 2004; IEEE Trans. Biomedical Engineering; 51:582-589.

[4] Williams CKI, Barber D. Bayesian Classification with Gaussian Processes 1998; IEEE Trans. Pattern Analysis and Machine Intelligence; 20:1342-1351.

[5] Minka TP. A family of algorithm for approximate Bayesian inference 2001; Ph.D. thesis Massachusetts Institute of Technology.

[6] Rasmussen C, Williams CKI. Gaussian process for machine learning 2006; The MIT press.

[7] Stockwell RG, Mansinha L, Lowe RP. Localization of the complex spectrum: The S– transform 1996; IEEE Trans. Signal Processing; 44:998–1001.

[8] Mark R, Moody G. MIT-BIH Arrhythmia Database 1997 [Online]. Available: http://ecg.mit.edu/dbinfo.html.



Address for correspondence
Melgani Farid
Dept. of Information Engineering and Computer Science, Univ. of Trento, Via Sommarive, 14, I-38050 Trento, Italy.
E-mail: melgani@disi.unitn.it