Gaussian processes for switching regimes

Citation for published version:
Storkey, A 1998, Gaussian processes for switching regimes. in *ICANN 98: Proceedings of the 8th International Conference on Artificial Neural Networks, Skövde, Sweden, 2–4 September 1998.* Perspectives in Neural Computing, Springer London, pp. 621-626. https://doi.org/10.1007/978-1-4471-1599-1_94

Digital Object Identifier (DOI):
10.1007/978-1-4471-1599-1_94

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
ICANN 98

Gaussian Processes for Switching Regimes

Amos Storkey

Neural Systems Group, Imperial College, London amoss@ic.ac.uk

Abstract. It has been shown that Gaussian processes are a competitive tool for nonparametric regression and classification. Furthermore they are equivalent to neural networks in the limit of an infinite number of neurons. Here we show that the versatility of Gaussian processes at defining different textural characteristics can be used to recognise different regimes in a signal switching between different sources.

1 Introduction

The use of Gaussian processes [3] to tackle many of the standard neural network problems was reintroduced by Williams [6] and prompted by recent work showing that neural networks and Gaussian processes were closely related. Neal [2] showed that in the limit of an infinite number of neurons, the two were equivalent. It was also noted the linear models and radial basis functions were special cases of Gaussian processes [1]. Rasmussen showed that Gaussian processes were competitive on a number of benchmark problems [4]. Here we look at the problems of non stationary signals, specifically the case of switching signals. This situation has received attention in the past [5]. Here we show that Gaussian processes are a useful tool for tackling this problem.

2 Gaussian Processes for Regression

Consider a set of points $\{\mathbf{x_i}\}$, which consist of the points in input space at which we will later receive data $\{\boldsymbol{x}_i\}$ $i = 1, 2, \ldots, n$ and the points $\{\boldsymbol{x}_i\}$ $i = n+1, 2, \ldots, m$ at which we would like to make predictions. We use a superscript D (for DATA) to denote an m-vector truncated to the elements $i = 1, 2, \ldots, n$, and a superscript P (for PREDICTION) to denote an m-vector truncated to the elements $i = n + 1, \ldots, m$.

We suppose for now that there is a true unknown function $f(\boldsymbol{x})$ which generates datum f_i at point $\boldsymbol{x}_i$. This datum is corrupted by measurement noise η_i, assumed for now to be Gaussian, mean zero, variance σ^2. We define the random variable y_i by

$$y_i = \begin{cases} f_i + \eta_i & i = 1, 2, \ldots, n \\ f_i & i = n + 1, \ldots, m \end{cases}$$

So y_i combines the possible values of the data to be received (including measurement noise) with the possible values of the predictions (without measurement noise). Now $\boldsymbol{y} = (y_1, y_2, \ldots, y_m)$ contains all the values of interest, and so we wish to find some prior distribution over $\boldsymbol{y}$.

We define this distribution in two stages. First of all we have assumed that η_i is $Gauss(0, \sigma^2)$. We now assume that the prior function over $\boldsymbol{f} = (f_1, f_2, \ldots, f_m)$ can be expressed as a multivariate Gaussian

$$P(\boldsymbol{f}|H) = \frac{1}{Z'} \exp\left(-\frac{1}{2}(\boldsymbol{f}-\boldsymbol{\mu})^T C^{-1}(\boldsymbol{f}-\boldsymbol{\mu})\right)$$

where $\boldsymbol{\mu} = \boldsymbol{\mu}(H)$ is some mean vector, $C = C(H)$ is some covariance matrix and Z' is the relevant normalisation constant. H stands for any set of hyperparameters, which, for now, are assumed to be known.

Then $\boldsymbol{f}$ and $\boldsymbol{\eta}$ are independent Gaussian random variables, and so $\boldsymbol{y}$ is a sum of independent Gaussian distributed random variables, and therefore has a prior distribution of

$$P(\boldsymbol{y}|H) = \frac{1}{Z} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T Q^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)$$

where

$$Q_{ij} = \begin{cases} C_{ij} + \delta_{ij}\sigma^2 & i, j \leq n \\ C_{ij} & \text{otherwise} \end{cases}$$

For future use, we partition Q into the form

$$\begin{pmatrix} Q^{DD} & Q^{DP} \\ Q^{PD} & Q^{PP} \end{pmatrix}$$

where Q^{DD} is $n \times n$ and Q^{PP} is $(m-n) \times (m-n)$. Note that $Q^{PD} = (Q^{DP})^T$.

Suppose we have now received data at points $\boldsymbol{x}^D$ given by $\boldsymbol{y}^D = \boldsymbol{y}^*$. Then we obtain the posterior distribution

$$P(\boldsymbol{y}^P|\boldsymbol{y}^D = \boldsymbol{y}^*, H) = \frac{P(\boldsymbol{y}, \boldsymbol{y}^D = \boldsymbol{y}^*|H)}{P(\boldsymbol{y}^D = \boldsymbol{y}^*|H)}$$

$$= \frac{Z^D}{Z} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T Q^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) - \frac{1}{2}(\boldsymbol{y}^D - \boldsymbol{\mu}^D)^T (Q^{DD})^{-1}(\boldsymbol{y}^D - \boldsymbol{\mu}^D)\right)$$

This simplifies to the posterior distribution we want

$$P(\boldsymbol{y}^P|\boldsymbol{y}^D = \boldsymbol{y}^*, H) = \frac{1}{Z^P} \exp\left(-\frac{1}{2}(\boldsymbol{y}^P - \hat{\boldsymbol{y}})^T S^{-1}(\boldsymbol{y}^P - \hat{\boldsymbol{y}})\right)$$

where $S = (Q^{PP} - Q^{PD}(Q^{DD})^{-1}Q^{DP})$ and $\hat{\boldsymbol{y}} = Q^{PD}(Q^{DD})^{-1}\boldsymbol{y}^D + \boldsymbol{\mu}$ by the partitioned inverse equations. Note that we only need to invert matrices Q^{DD} and S, which are $n \times n$ and $(m-n) \times (m-n)$ respectively. There is no need to invert any $m \times m$ matrices such as Q. Here the formulation in [6, 1, 4] is extended to the multivariate predictor case.

This Gaussian process approach has a number of advantages. These are

- The posterior distribution can be calculated analytically.
- The prior form is very flexible: many different forms of covariance matrix can be used, each giving a different type of textural structure to the signal.

- Prior knowledge about functional forms can meaningfully be represented by a Gaussian process: the hyperparameters relate directly to length scales.

There is a computational disadvantage to this method: It involves calculating the inverse of an $n \times n$ matrix, involving $o(n^3)$ computations. Hence the computational power needed increases significantly with the size of the dataset, making it less suitable for cases where many data are available.

3 Types of Covariance Functions

We have said nothing yet of the form of the covariance function C. In fact for the Gaussian distributions above to be meaningful for all points in input space, the distributions need to satisfy the Chapman-Kolmogorov equations. This is done if the covariance function is that of a Gaussian process. For this to be the case, the covariance function must be positive semidefinite symmetric, and C_{ij} must depend on variables $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, and no other $\boldsymbol{x}_k$. Furthermore the mean $\mu_i = \mu(\boldsymbol{x}_i)$.

Given a set of scaling hyperparameters θ_1, θ_2, r_l, a common choice for C is

$$C(\boldsymbol{x}_i, \boldsymbol{x}_j; H) = \theta_1 \exp\left(-\frac{1}{2}\sum_{l=1}^{L}\frac{(x_i^{(l)} - x_j^{(l)})^2}{r_l^2}\right) + \theta_2$$

where L is the dimension of the input space, and l counts through each dimension. This corresponds to saying that the closer points are in input space, the more correlated their function values will be, and that the function is smooth.

4 Determining Different Signal Regimes: Gaussian Process Mixtures

Very often the data under study has not been generated from a stationary process. A common example of this is where a number of different signal sources are present, and the observable signal is created by switching between these different regimes.

Here this situation is modelled with a mixture of Gaussian processes. Latent variables represent which of the current regimes generated a sample datum. Then different hyperparameters or covariance structures can be used to represent the characteristics of the different regimes.

The great benefit of Gaussian processes is that the covariance matrix structure can represent many different signal structures and textures, from smooth curves to random fractal textures.

As it is not known which regime is generating the signal at any point, and the structure of the signals is unknown, these variables/parameters are given prior distributions which should be integrated over.

Let s_k denote the regime which generated datum k. For now let us assume there are two possible regimes. Then we can form the Gaussian process prior

$$P(\boldsymbol{y}|H) = \sum_{\boldsymbol{s} \in B(m)} p_{\boldsymbol{s}} \varPhi_{\boldsymbol{s}}(\boldsymbol{y}; H)$$

where $B(m)$ is the set of binary vectors of length m. $\varPhi$ is a Gaussian kernel of the form

$$\varPhi_{\boldsymbol{s}} = \frac{1}{Z_{\boldsymbol{s}}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu}(H))^T Q_{\boldsymbol{s}}(H)^{-1}(\boldsymbol{y} - \boldsymbol{\mu}(H))\right)$$

with $\boldsymbol{\mu}(H) = (\mu_{s_1}, \mu_{s_2}, \ldots, \mu_{s_m})^T$ where $\mu_1, \mu_2 \in H$. The covariance function $Q_{\boldsymbol{s}}$ is given by

$$(Q_{\boldsymbol{s}})_{ij} = \begin{cases} \theta_{1s} \exp\left(-\frac{1}{2}\sum_{l=1}^{L} \frac{(x_i^{(l)} - x_j^{(l)})^2}{r_{ls}^2}\right) + \theta_{2s} + \delta_{ij}\sigma^2 & \text{for } s = s_i = s_j \\ 0 & \text{otherwise} \end{cases}$$

which says that within a given regime we have the usual smooth functions, but there is no correlation between the points in different regimes.

Now let us assume that the switching regime is a Poisson process, rate λ. Therefore the probability of $y(t_k) = y(t_{k-1})$ is given by the probability of an even number of switches between the two time points:

$$\cosh(\lambda(t_k - t_{k-1})) \exp(-\lambda(t_k - t_{k-1}))$$

Hence $P(y(t_k) \neq y(t_{k-1})) = \sinh(\lambda(t_k - t_{k-1})) \exp(-\lambda(t_k - t_{k-1}))$ where λ is a hyperparameter. Other switching priors could equally well be used, and this formalism could easily be extended to multiple regimes.

All the priors are now defined, and the problem can be passed through the usual Gaussian process machinery. The first level of inference involves a tractable Gaussian marginalisation. The second level of inference involves an intractable integration over the hyperparameters and latent variables.

5 Integrating-Out or Maximisation?

Calculating the inverse covariance for a Gaussian process is computationally intensive. Therefore integrating out hyperparameters using Monte-Carlo Markov chain approaches can be very slow for large data sizes. The approach we take here is to sample from the posterior over the latent variables and the use a maximum posterior value for the hyperparameters. This is a form of GEM algorithm, where a sample distribution over the latent indicator variables is used instead of the true distribution.

The great benefit of this approach is that the latent variables can be Gibbs sampled, and each Gibbs sample step involves changing only one row/column of the covariance matrix. Hence the partitioned inverse equations can be used to calculate the inverse of the matrix in $o(n^2)$ flops, reducing the computational load significantly. Furthermore, because the covariance matrix has a block structure, the cost of inverting matrices is reduced. The steps of the algorithm are:

- Choose suitable values of the hyperparameters, H.
- Choose suitable starting values for the latent variables s.
- Gibbs sample the latent variables to get an E-step expression for $P(s|data)$:

  $$P(s|data) \leftarrow P(s|H, data) = \frac{1}{Z}\int dsP(data|H, s)P(s) \simeq \sum_k \delta(s - s_k)$$

  for a sample $\{s_k\}$
- Move towards the maximum posterior value of the hyperparameters, assuming that this distribution is the true distribution for s. This involves maximising

  $$\int dP(s|data) \ln P(H|s, data) \simeq \frac{1}{G}\sum_{g=1}^{G} \ln P(H|\boldsymbol{s}_g, data)$$

  where G is the chosen sample size (the GM step).
- Repeat the steps until suitably near convergence.

6 Example

These methods were tested on a number of toy problems. We introduce one of them. Here a signal is generated from two smooth functions of different regularity and size. In this example, the signal in figure 1 was used. It was generated by the function illustrated, made up of two separate sin waves. The only prior information given was that mentioned above. Hence no knowledge was presumed about the functional form, or periodicity of the data.

When the Gaussian process was tested on this problem it was generally able to distinguish the different regimes. The graphs in figure 1 give the predictive mean, and error bars for the two signals. The true signals are given as solid lines.

The methods were also tested on other similar problems, and problems where the signal mean differed between the signals. The model distinguished between the different regimes. Problems are sometimes encountered when the Poisson prior is such that switching is infrequent. This means that local maxima in the posterior of the latent variables are surrounded by regions of very low probability, and so the Gibbs sampler can get stuck, and not properly sample the whole data space. Occasionally resetting the Gibbs sampler with different starting positions appears to help solve this problem.

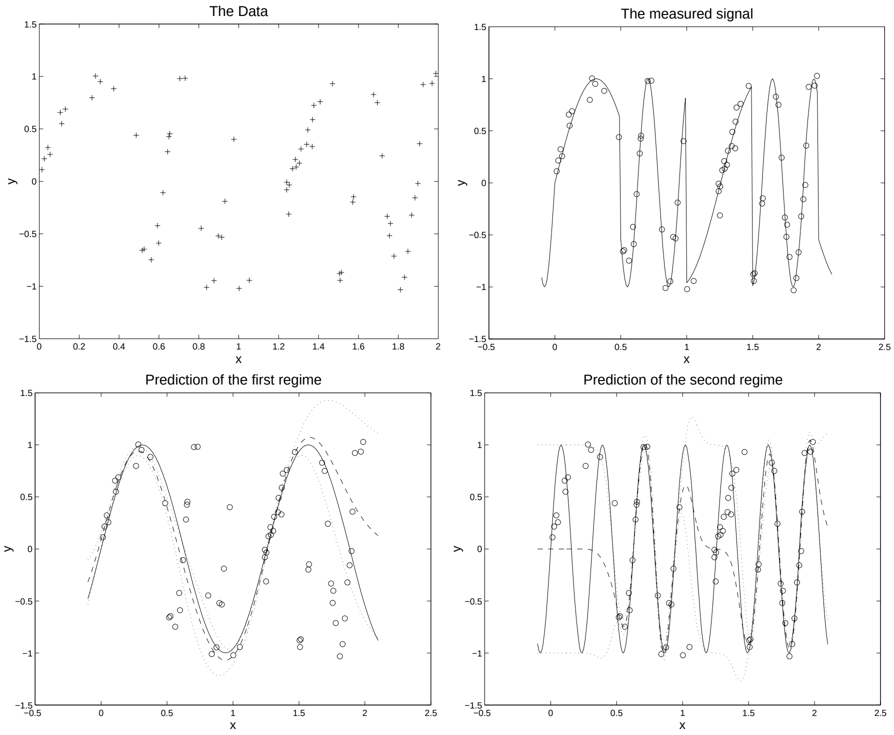

Fig. 1. A test problem. Predictors (dashed) and error bars (dotted) for the two generating functions (solid)

7 Conclusions

Gaussian processes can represent many types of functions because of the versatility of the covariance structure. This enables regimes with different second order statistical properties to be recognised, while at the same time allowing prior information about the signal form to be properly represented. These methods could be extended to higher dimensions, for example to recognise different textures in two dimensional data.

References

1. M. N. Gibbs and D. J. C. MacKay. Efficient implementation of Gaussian processes. Preprint, 1997.
2. R. Neal. *Lecture Notes in Statistics 118: Bayesian Learning for Neural Networks.* Springer Verlag, 1996.
3. A. O'Hagan. On curve fitting and optimal design for regression. *Journal of the Royal Statistical Society B*, 40:1–42, 1978.

4. C. E. Rasmussen. *Evaluation of Gaussian Processes and other Methods for Non-Linear Regression*. PhD thesis, University of Toronto., 1996.
5. A. S. Weigend, M. Mangeas, and A. N. Srivastava. Nonlinear gated experts for time series: Discovering regimes and avoiding overfitting. *International Journal of Neural Systems*, 6:373–399, 1995.
6. C. Williams. Regression with Gaussian processes. In S. W. Ellacott, J. C. Mason, and I. J.Anderson, editors, *Mathematics of Neural Networks: Models, Algorithms and Applications*. Kluwer, 1995. Published 1997.

This article was processed using the LaTeX macro package with LLNCS style