ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

Gaussian process modelling of multiple short time series

Hande Topa[1] and Antti Honkela[2] *

1- Helsinki Institute for Information Technology HIIT,
Department of Information and Computer Science
Aalto University, Espoo, Finland

2- Helsinki Institute for Information Technology HIIT,
Department of Computer Science
University of Helsinki, Helsinki, Finland

Abstract. We study effective Gaussian process (GP) modelling of multiple short time series. These problems are common for example when applying GP models independently to each gene in a gene expression time series data set. Such sets typically contain very few time points and hence naive application of common GP modelling techniques can lead to severe over-fitting in a significant fraction of the fitted models, depending on the details of the data set. We propose avoiding over-fitting by constraining the GP length-scale to values that are compatible with the spacing of the time points. We demonstrate that this eliminates otherwise serious over-fitting in real experiment using GP model to rank single nucleotide polymorphisms (SNPs) based on their likelihood of being under natural selection.

1 Introduction

Gaussian processes (GPs) are widely applied non-parametric probabilistic models for continuous data [10]. Because of their non-parametric nature, they can flexibly adapt to differently sized data sets and can easily accommodate for example non-uniformly sampled data. GPs are computationally very convenient, because they permit exact marginalisation of the latent process in regression with a Gaussian likelihood. Most methods development work on GPs in machine learning has focused on developing efficient inference for large data sets. This is an important area, as naive inference algorithms suffer from cubic computational complexity with respect to the data set size, and the recently developed methods can successfully reduce this significantly.

In this paper we focus instead on the other frontier of GP applications in data sets with a very large number of small independent instances. GPs for such applications have recently gathered significant interest in computational systems biology, where they provide a very powerful model for sparsely and often irregularly sampled gene expression time series [5, 11, 6, 3, 4, 1, 13]. Reliable fitting of very large number of independent models is important in many applications of these models, such as ranking of targets of gene regulators [3].

*This work was supported by the Academy of Finland (grant 259440) and by Alfred Kordelin Foundation.

Most gene expression time series are very short with a great majority having less than 9 time points [2], so computational complexity of any GP inference method will typically not be an issue. Instead, the application of GP methods in these problems will face other problems due to lack of and sparseness of data. Depending on the specific problem, this can easily lead to either over-fitting or under-fitting. When fitting the models automatically to a large number, possibly several thousands, of instances, it is impractical to manually locate and fix these problematic fits.

In this paper we present methods for setting constraints or more restrictive priors to kernel length-scale parameters that help avoid these phenomena.

2 Gaussian Process Modelling

A GP is a stochastic process $\{f(\mathbf{t})|\mathbf{t} \in \mathcal{T}\}$ for which the marginal distribution at any finite sub-collection of points $\mathbf{t}_1, \ldots, \mathbf{t}_n$ is multivariate Gaussian [10]. The process is completely defined by the mean function $\mu(\mathbf{t})$ and the covariance function $k(\mathbf{t}, \mathbf{t}')$, that also define the mean vector and covariance matrix of the multivariate Gaussian over the sub-collection. For simplicity, we set the mean function to zero $\mu(\mathbf{t}) \equiv 0$ by subtracting the mean of data.

The most widely used covariance function for GPs in machine learning is the squared exponential covariance [10]

$$k_{\mathrm{SE}}(\mathbf{t}, \mathbf{t}') = \sigma_f^2 \exp\left(-\frac{r^2}{2\ell^2}\right), \tag{1}$$

where $r = ||\mathbf{t} - \mathbf{t}'||$. The covariance depends on two positive hyperparameters: variance σ_f^2 and length-scale ℓ. The length-scale parameter ℓ governs the range of dependencies in the process. A short length-scale corresponds to rapidly varying functions with weak long-range dependencies, while a large length-scale corresponds to slowly varying functions. Extremely small length-scale may lead to a situation where each observation is treated as essentially independent, which makes the model overfit.

2.1 GP-based approach to ranking time series

Recently, GPs have been successfully used to develop methods for ranking genomic markers according to their temporal behaviours [4, 11]. For example, one may be interested in the most active genes whose expression levels change during a time interval, or in the SNPs whose allele frequencies show changes across generations, being affected under natural selection. GPs are very useful for modelling temporal behaviours of these genomic markers efficiently in contrast to the commonly used pairwise comparisons which fail to exploit the full temporal behaviour of the markers of interest.

Kalaitzis et al. have recently proposed a GP-based approach for ranking differentially expressed gene expression time courses [4]. In this approach, genes are ranked according to their Bayes factors (BFs), where BFs are the ratio of the marginal likelihoods under "time-dependent" and "time-independent" GP models.

In the "time-independent" model, the observations ($\mathbf{D}$) are assumed to have been randomly generated around a constant mean (no temporal dependency) whereas the time dependency in the "time-dependent" model is modelled by a squared exponential covariance function (k_{SE} in Eq. 1). Assuming the noise is additive white Gaussian, the corresponding white noise covariance matrix is given by

$$\mathbf{k}_{\text{W}} = \sigma_n^2 \mathbf{I}, \tag{2}$$

where $\mathbf{I}$ is an n-by-n identity matrix and σ_n^2 is the noise variance parameter.

Therefore, we can define our hypotheses as following:

H_0 : The data has come from a "time-independent" model $\equiv \mathbf{D} \sim GP(\mathbf{0}, \mathbf{\Sigma_0})$
H_1 : The data has come from a "time-dependent" model $\equiv \mathbf{D} \sim GP(\mathbf{0}, \mathbf{\Sigma_1})$,

where $\mathbf{\Sigma_0} = \mathbf{k}_{\text{W}}$ and $\mathbf{\Sigma_1} = \mathbf{k}_{\text{SE}} + \mathbf{k}_{\text{W}}$.

Bayes factors (BF) can be computed as [4, 11]:

$$\text{BF} = \frac{p(\mathbf{D}|\hat{\theta}_{\mathbf{1}}, H_1)}{p(\mathbf{D}|\hat{\theta}_{\mathbf{0}}, H_0)}, \tag{3}$$

where $\hat{\theta}_{\mathbf{1}}$ and $\hat{\theta}_{\mathbf{0}}$ include the maximum likelihood estimates of the hyperparameters in the corresponding GP models. Note that $\mathbf{\Sigma_1}$ becomes equivalent to $\mathbf{\Sigma_0}$ when $\ell \to \infty$ and $\sigma_f^2 \to 0$. Therefore, a Bayes factor of 1 indicates that it is more likely that the time course data have been generated from a "time-independent" model, while larger BFs indicate that it is more likely that the data have been generated from a "time-dependent" model.

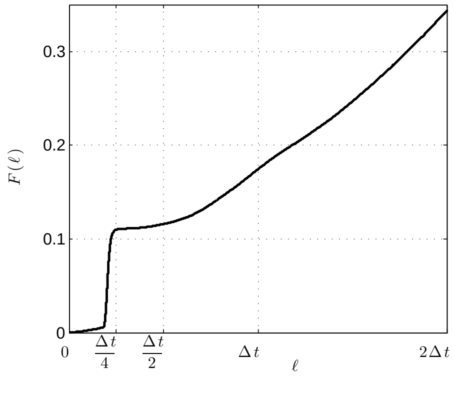


Fig. 1: Empirical cumulative distribution of length-scale estimates for the SNPs with $BF > 3$. Δt was set to 4, which is the shortest distance between the consecutive time points.

2.2 Length-scale bounds

Naive application of common GP modelling techniques can lead to severe over-fitting or under-fitting, depending on the details of the data set. We propose avoiding over-fitting by constraining the GP length-scale ℓ to values that focus most of the energy spectrum to frequencies below the Nyquist frequency corresponding to the sampling in the data set. According to the Nyquist sampling theorem, the Nyquist frequency $s_n = \frac{1}{2\Delta t}$ is the maximal frequency that can be identified in the spectral representation of the sampled signal [12, 8], where Δt is the sampling interval in the data set. This can be generalized to the nonuniformly sampled data as long as the samples satisfy the Nyquist rate on the average [7]. Therefore, we define Δt conservatively as the shortest distance between consecutive data points to obtain the least restrictive bound.

In case of the squared exponential covariance function, the spectral density is given by

$$S_{\mathrm{SE}}(s) = (2\pi\ell^2)^{D/2} \exp(-2\pi^2\ell^2 s^2), \tag{4}$$

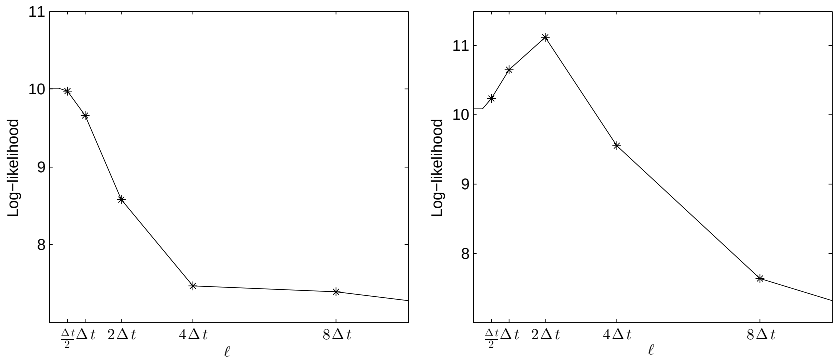
where D is the number of dimensions and s denotes the frequency [10]. For $D = 1$, the fraction α of the system's energy on the frequencies that are below the Nyquist frequency is:

$$\alpha = \int_{-\frac{1}{2\Delta t}}^{\frac{1}{2\Delta t}} S_{\mathrm{SE}}(s)\,\mathrm{d}s = \mathrm{erf}\left(\frac{\pi\ell}{\sqrt{2}\Delta t}\right). \tag{5}$$

In the next section we investigate the effects of the overfitted models on GP-based ranking methods in a real data set and we introduce a lower bound for the length-scales to overcome the emerged problems.

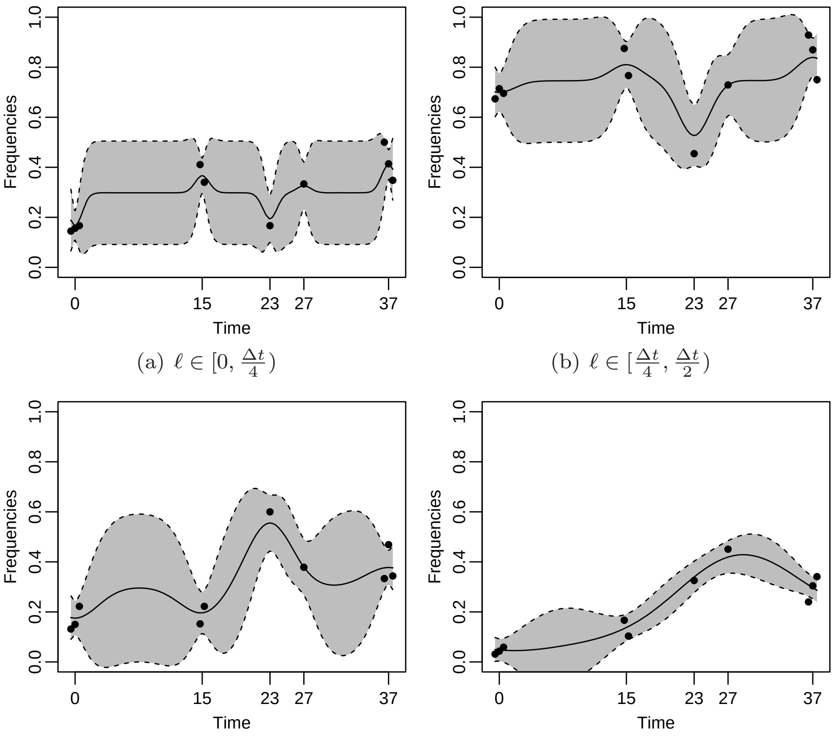
3 Results and discussion

We applied naive GP regression on a real data set which was obtained by evolve and re-sequencing methods on *Drosophila melanogaster* populations under a fluctuating temperature regime [9]. The data set contains the allele frequencies of the SNPs at the following generations: three replicates at base generation, two replicates at generation 15, single replicates at generation 23 and 27, three replicates at generation 37. In our analysis, we included only the bi-allelic SNPs, which were in total 1,257,117. Aiming to identify the SNPs which were affected under natural selection, we followed Kalaitzis et al.'s method and modelled the allele frequencies under the "time-dependent" and "time-independent" GP models by using the "gptk" R package [4]. Then, we ranked the SNPs according to their Bayes factors. We observed that approximately 10% of the highly-ranked SNPs suffered from over-fitting, i.e., they had very small length-scale estimates along with large Bayes factors (> 3), and thus they dominated the ranking deceptively. Fig. 1 shows the empirical cumulative distribution of the length-scale estimates for the SNPs which had Bayes factors larger than 3. The GPs under



(a) Overfitted example (b) Non-overfitted example

Fig. 2: *Log-likelihoods vs. length-scales (ℓ) for (a) overfitted example and (b) non-overfitted example.* The signal variance (σ_f^2) and the noise variance (σ_n^2) were optimised at each value of ℓ. Overfitted examples resulted from the fact that the maximum log-likelihood was found at a very small length-scale estimate.



(a) $\ell \in [0, \frac{\Delta t}{4})$ (b) $\ell \in [\frac{\Delta t}{4}, \frac{\Delta t}{2})$

(c) $\ell \in [\frac{\Delta t}{2}, \Delta t)$ (d) $\ell \in [\Delta t, 2\Delta t)$

Fig. 3: *Example GP fits in different groups of length-scale (ℓ) estimates.* The problem of over-fitting wears off as $\ell > \Delta t$. Confidence regions are shown for ± 2 standard deviation. Replicates at the same time points are shifted by 0.5 for better visualisation.

the “time-dependent” model tend to overfit if the maximum likelihood is found at a relatively small length-scale estimate (see Fig. 2). Therefore, a practical solution to prevent such cases from dominating the ranking would be to set a lower bound (ℓ_{min}) on the length-scale estimates. Visual inspection of the GP maximum likelihood fits shown in Fig. 3 suggests that setting ℓ_{min} to $\approx \Delta t$ would be an appropriate choice, which also focuses most of the energy spectrum to frequencies below the Nyquist rate (note that as $\frac{\ell}{\Delta t} \geq 1$, $\alpha \to 1$ in Eq. 5).

For Bayesian parameter estimation, a uniform prior over the length-scale over the interval $[\ell_{min}, t_n - t_1]$ seems like a reasonable objective prior. The usage of such priors would help eliminate the false positives by moving down the inflated Bayes factors caused by the overfitted models in the ranked list.

The derivation presented here was for squared exponential covariance, but similar bound can be derived for Matérn covariance functions too.

References

[1] Cooke, E.J., Savage, R.S., Kirk, P.D.W., Darkins, R., Wild, D.L.: Bayesian hierarchical clustering for microarray time series data with replicates and outlier measurements. BMC Bioinformatics 12, 399 (2011)

[2] Ernst, J., Nau, G.J., Bar-Joseph, Z.: Clustering short time series gene expression data. Bioinformatics 21 Suppl 1, i159–i168 (Jun 2005)

[3] Honkela, A., Girardot, C., Gustafson, E.H., Liu, Y.H., Furlong, E.E.M., Lawrence, N.D., Rattray, M.: Model-based method for transcription factor target identification with limited data. Proc Natl Acad Sci U S A 107(17), 7793–7798 (Apr 2010)

[4] Kalaitzis, A.A., Lawrence, N.D.: A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. BMC Bioinformatics 12, 180 (2011)

[5] Kirk, P.D.W., Stumpf, M.P.H.: Gaussian process regression bootstrapping: exploring the effects of uncertainty in time course data. Bioinformatics 25(10), 1300–1306 (May 2009)

[6] Liu, Q., Lin, K.K., Andersen, B., Smyth, P., Ihler, A.: Estimating replicate time shifts using Gaussian process regression. Bioinformatics 26(6), 770–776 (Mar 2010)

[7] Marvasti, F.: Nonuniform Sampling Theory and Practice. Kluwer Academic/Plenum Publishers, New York (2001)

[8] Oppenheim, A.V., Schafer, R.W.: Digital Signal Processing. Prentice-Hall (1975)

[9] Orozco-Ter Wengel, P., Kapun, M., Nolte, V., Kofler, R., Flatt, T., Schlötterer, C.: Adaptation of Drosophila to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. Molecular Ecology 21, 4931–4941 (2012)

[10] Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA (2006)

[11] Stegle, O., Denby, K.J., Cooke, E.J., Wild, D.L., Ghahramani, Z., Borgwardt, K.M.: A robust Bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. J Comput Biol 17(3), 355–367 (Mar 2010)

[12] Tick, L.J., Shaman, P.: Sampling rates and appearance of stationary Gaussian processes. Technometrics 8, 91–106 (1966)

[13] Titsias, M.K., Honkela, A., Lawrence, N.D., Rattray, M.: Identifying targets of multiple co-regulating transcription factors from expression time-series by Bayesian model comparison. BMC Syst Biol 6(1), 53 (May 2012)