Assessing the impact of non-additive noise on modelling transcriptional regulation with Gaussian processes.

Vinny Davies[1] and Dirk Husmeier[1]

[1] School of Mathematics and Statistics, University of Glasgow, Scotland

E-mail for correspondence: v.davies.1@research.gla.ac.uk

Abstract: In transcriptional regulation, transcription factors (TFs) are often unobservable at mRNA level or may be controlled outside of the system being modelled. Gaussian processes are a promising approach for dealing with these difficulties as a prior distribution can be defined over the latent TF activity profiles and the posterior distribution inferred from the observed expression levels of potential target genes. However previous approaches have been based on the assumption of additive Gaussian noise to maintain analytical tractability. We investigate the influence of a more realistic form of noise on a biologically accurate system based on Michaelis-Menten kinetics.

Keywords: Transcriptional regulation, Gaussian processes, additive and multiplicative noise, Michaelis-Menten kinetics

1 Introduction

A particular challenge in the quantitative modelling of transcriptional regulation is that transcription factors (TFs), the regulatory proteins at the heart of the process, are frequently subject to post-translational modification, which may affect their DNA binding capability. Consequently, gene expression levels of TFs contain only limited information about their actual activities. A promising approach to deal with these difficulties was proposed in Gao et al. (2008), inspired by the work of Barenco et al. (2006). The authors advocate the use of Gaussian processes to define prior distributions over the latent TF activity profiles. Inference is soundly based on the principles of non-parametric Bayesian statistics, consistently inferring the posterior distribution of the unknown TF activities from the observed expression levels of potential target genes, and inferring regulatory network structures after marginalizing over the unknown TF activity profiles.
The choice of a non-parametric prior distribution from the Gaussian process family is not a restrictive modelling assumption. Somewhat more restrictive is the assumption of additive Gaussian noise, which can be found in all

previous applications (Gao et al. (2008), Honkela et al. (2010), etc.). Previous work by Rocke and Durbin (2001) showed that mRNA concentrations obtained from microarray experiments are of a more complex form and the purpose of this work is to investigate what effect this deviation from additive Gaussianity has on the inference in transcriptional regulation.

2 Method

A linear model of gene expression was proposed by Barenco et al. (2006)

$$\frac{dx_i(t)}{dt} = B_i + S_i f(t) - D_i x_i(t) \tag{1}$$

where $i \in \{1, \ldots, G\}$ is a set of genes regulated by the same TF, $x_i(t)$ are the (unknown) true gene expression levels at time point t, $f(t)$ is the (unknown) TF activity, B_i is the basal transcription rate of gene i, S_i is the sensitivity to binding of TF, and D_i is a decay rate. We assume that (noisy) measurements of $x_i(t)$ can be obtained, however TF activity is unknown and therefore $f(t)$ is assumed to be unobservable.
Eq. (1) has the analytical solution:

$$x_i(t) = \frac{B_i}{D_i} + S_i \int_0^t \exp(-D_i(t-u)) f(u) du. \tag{2}$$

Gao et al. (2008) proposed a non-parametric Bayesian approach to inference in this model by placing a Gaussian process prior with a squared exponential covariance matrix on the unknown TF activities $\mathbf{f} = (f(t_1), \ldots, f(t_T))$ at timepoints $\mathbf{t} = (t_1, \ldots, t_T)$. The linear form of the model implies that the joint prior distribution of the expression profiles of all regulated genes, $\mathbf{x}_i$, is described by a Gaussian process prior with a covariance matrix, $\mathbf{K}$, that depends on the hyperparameters of the prior, θ_h, as well as the parameters that characterise the transcriptional regulation processes via eq. (2):

$$\begin{aligned} p(\mathbf{x}|\boldsymbol{\theta}') &= \mathcal{N}(\mathbf{B}./\mathbf{D}, \mathbf{K}); \quad \mathbf{K} = \mathbf{K}(\boldsymbol{\theta}') \\ \boldsymbol{\theta}' &= (\theta_h, B_1, \ldots, B_G, S_1, \ldots, B_G, D_1, \ldots, D_G) \end{aligned} \tag{3}$$

where $B./D$ is a point-wise vector division. See Davies and Husmeier (2013) for details.
To relate the unknown true gene expression profiles $\mathbf{x}_i = (x_i(t_1), \ldots, x_i(t_T))$ to noisy measurements $\mathbf{y}_i = (y_i(t_1), \ldots, y_i(t_T))$, Gao et al. (2008) assumed additive Gaussian noise of constant variance σ^2. The marginalisation over $\mathbf{y}$ is analytically tractable and gives:

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int \mathcal{N}(\mathbf{y}|\mathbf{x}, \sigma^2\mathbf{I}) \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{K}(\boldsymbol{\theta}')) d\mathbf{x} = \mathcal{N}(\mathbf{y}|\mathbf{B}./\mathbf{D}, \mathbf{K}(\boldsymbol{\theta}') + \sigma^2\mathbf{I}) \tag{4}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}', \sigma^2)$. Inference of the parameters $\boldsymbol{\theta}$ can then be achieved in a maximum likelihood or Bayesian framework; see Bishop (2006).

However Rocke and Durbin (2001) showed that the noise in transcriptional profiling with microarrays has the following more general form:

$$y_i(t) = c + x_i(t)\exp(\epsilon_\mu) + \epsilon_t \quad \text{where} \quad \epsilon_j \sim \mathcal{N}(0, \sigma_j^2) \tag{5}$$

where c is mean background noise, and σ_μ^2 and σ_t^2 are unknown variance parameters. Replacing $\mathcal{N}(\mathbf{y}|\mathbf{x}, \sigma^2\mathbf{I})$ in eq. (4) by the noise in eq. (5) does not give a closed-form solution, and this has therefore been ignored in previous work. The objective of the present study is to quantify the effect the deviation from additive Gaussianity has on the inference of the transcriptional regulation.

3 Data

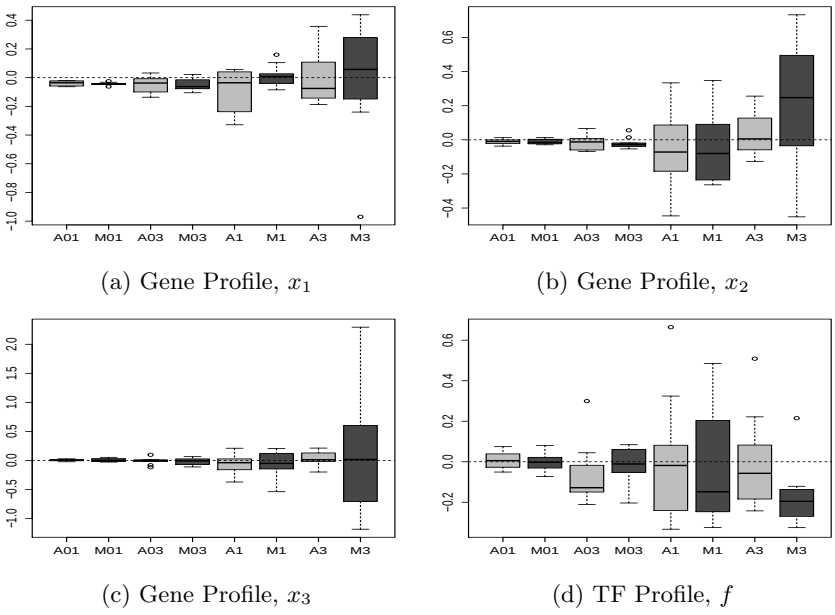
We combined a simple regulatory network for three genes with a protein signalling pathway from Vyshemirsky and Girolami (2008); see Davies and Husmeier (2013) for details. The active form of the TF is unobservable due to post-translational modification, and the processes leading to the formation of active TF is controlled outside of the subsystem being modelled. The transcriptional profiles of the downstream genes were generated by solving eq. (2) with the different kinetic parameters. 18 values from these expression profiles were then subjected to either additive Gaussian noise, or the more complex noise of eq. (5). For the non-Gaussian noise the standard deviations were chosen on a roughly log scale such that $\sigma_\mu, \sigma_t = (0.01, 0.03, 0.1, 0.3)$, with equivalent values chosen for the additive noise model to allow for a fair comparison. This was repeated 10 times for each standard deviation size and noise model.

4 Results

For a relatively small data set, our results, given in Figure 1, have shown that the deviation from additive Gaussian noise has little negative effect when $\sigma_\mu, \sigma_t = (0.01, 0.03)$. For larger standard deviations the results show a consistent deterioration in the case of non-Gaussian noise, although this cannot be easily quantified until $\sigma_\mu, \sigma_t = (0.3)$. For this level of variance, Figure 1, as well as similar results for the kinetic parameter estimates, show a roughly four fold increase in the median error.

5 Conclusion

Our work has considered the implications of having non-Gaussian noise when using Gaussian processes for modelling transcriptional regulation. This noise model violates some of the modelling assumptions and causes a deterioration in the ability of the model to perform parameter inference. We have shown that the effect of this noise is not as significant as first assumed and the negative effect only becomes apparent for larger variances.



(a) Gene Profile, x_1

(b) Gene Profile, x_2

(c) Gene Profile, x_3

(d) TF Profile, f

FIGURE 1: Box plots of the error of gene and TF profile predictions. Box plots for the additive Gaussian, 'A', and non-Gaussian, 'M', noise are given in light and dark grey respectively. The standard deviations used for σ_μ and σ_t are given under each box plot and represent the values (0.01,0.03,0.1,0.3)

References

Barenco, M. et al. (2006). Ranked prediction of p53 targets using hidden variable dynamic modelling. *Genome Biology*, **7(3)**, R25.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Davies, V. and Husmeier, D. (2013). Modelling transcriptional regulation with Gaussian processes. Technical Report. University of Glasgow *www.maths.gla.ac.uk/∼dhusmeier/MyPapers/bookChapterVinny.pdf*

Gao, P. et al. (2008). Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics*, **24**, 70 – 75.

Honkela, A. et al. (2010). Model-based method for transcription factor target identification with limited data. *PNAS*, **107(17)**, 7793 – 7798.

Rocke, D.M. and Durbin, B. (2001). A model for measurement error of gene expression analysis. *J. Comput. Biol.*, **8**, 557 – 569

Vyshemirsky, V. and Girolami, M.A. (2008). Bayesian ranking of biochemical system models. *Bioinformatics*, **24(6)**, 833 – 839.