WARPED GAUSSIAN PROCESS MODELLING OF TRANSCRIPTIONAL REGULATION

Ruirui Ji[1,2] and Dirk Husmeier[1]

[1]School of Mathematics & Statistics, University of Glasgow, United Kingdom
[2]School of Automation and Information Engineering, Xi'an University of Technology, China
Ruirui.Ji@glasgow.ac.uk, dirk.husmeier@glasgow.ac.uk

ABSTRACT

This article extends recent work on Gaussian process modelling of transcriptional regulation, which assumed additive Gaussian noise of constant variance, to heteroscedastic noise. Our work is based on an explicit noise model for transcriptional profiling and the concept of warped Gaussian processes.

1. INTRODUCTION

A linear model of gene expression was proposed by Barenco et al. [1]

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

where $i \in \{1, \ldots, G\}$ is a set of genes regulated by the same transcription factor 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, e.g. with microarrays or RT-PCR scans. However, TF activity may be subject to post-transcriptional regulation and hence not amenable to transcriptional profiling techniques. We therefore assume that $f(t)$ is unobservable. Equation (1) has the analytic solution

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

where transient terms have been ignored. Gao et al. [2] proposed a nonparametric Bayesian approach to inference in this model by placing a Gaussian process prior with squared exponential covariance matrix on the unknown TF activities $\mathbf{f} = (f(t_1), \ldots, f(t_T))$ at timepoints $\mathbf{t} = (t_1, \ldots, t_T)$

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K_{f,f}}) \qquad (3)$$

that is, the prior probability of the TF activities, $p(\mathbf{f})$, is a zero-mean multivariate Gaussian distribution with covariance matrix $\mathbf{K_{f,f}}$, whose elements are

$$K_{f,f}(t, t') = \exp\left(-\frac{(t-t')^2}{l^2}\right) \qquad (4)$$

where l is a scale hyperparameter. The linear form of equation (2) implies that the joint prior distribution of the expression profiles of all regulated genes

$$\mathbf{x}_i = [x_i(t_1), \ldots, x_i(t_T)]; i = 1, \ldots, G \qquad (5)$$

is described by a Gaussian process prior with a covariance matrix, $\mathbf{K}$, that depends on the scale hyperparameter l and the parameters that characterise the transcriptional regulation processes via (1):

$$p(\mathbf{x}|\boldsymbol{\theta}') = \mathcal{N}(\mathbf{0}, \mathbf{K}); \quad \mathbf{K} = \mathbf{K}(\boldsymbol{\theta}') \qquad (6)$$
$$\boldsymbol{\theta}' = (l, B_1, \ldots, B_G, S_1, \ldots, B_G, D_1, \ldots, D_G)$$

See [2, 3] for explicit expressions. To relate the unknown true gene expression profiles $\mathbf{x}_i$ to noisy measurements $\mathbf{y}_i = [y_i(t_1), \ldots, y_i(t_T)]; i = 1, \ldots, G$, the standard approach (e.g. [4], Sect. 6.4.2) assumes additive Gaussian noise of constant variance σ^2:

$$p(\mathbf{y}|\mathbf{x}, \sigma^2) = \mathcal{N}(\mathbf{y}|\mathbf{x}, \sigma^2\mathbf{I}) \qquad (7)$$

where $\mathbf{I}$ is the identity matrix. The marginalisation over $\mathbf{y}$ is analytically tractable and gives, with the definition $\boldsymbol{\theta} = (\boldsymbol{\theta}', \sigma^2)$:

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{x}, \sigma^2)p(\mathbf{x}|\boldsymbol{\theta}')d\mathbf{x} = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{C}(\boldsymbol{\theta}))$$
$$\mathbf{C}(\boldsymbol{\theta}) = \mathbf{K}(\boldsymbol{\theta}') + \sigma^2\mathbf{I} \qquad (8)$$

Inference of the parameters $\boldsymbol{\theta}$ can then be achieved in a maximum likelihood or Bayesian framework, as described in standard textbooks on Gaussian processes [4, 5].

2. METHODOLOGICAL INNOVATION

The assumption of additive Gaussian noise is not biologically realistic. Gao et al. [2] formally introduced a case-dependent variance, σ_t^2, but such an over-flexible model is not amenable to statistical inference. Our approach is based on Durbin et al. [6], who proposed a general noise model for transcriptional profiling with microarrays:

$$y_i(t) = c + x_i(t)\exp(\mu_t) + \varepsilon_t$$
$$\mu_t \sim \mathcal{N}(0, \sigma_\mu^2); \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \qquad (9)$$

where $y_i(t)$ is the measured, and $x_i(t)$ the unknown true expression level of gene i at time point t, c is mean background noise, and σ_μ^2 and σ_ε^2 are unknown variance parameters. Note that this form of noise reduces to additive Gaussian noise in the limit of low expression levels,

$x_i(t) << 1$, and to log normally distributed noise in the limit of high expression levels, $x_i(t) >> 1$. Inserting (9) into (7) does not give a closed-form solution, and renders inference with Gaussian processes intractable. To proceed, we apply a result found in [7]. Using the delta method of classical statistical inference, the authors derived a variance-stabilising transformation for measured gene expression levels y of the form:

$$h(y) = U\gamma \operatorname{arsinh}(\alpha + \beta y) = \gamma \log\left(y + \sqrt{y^2 + 1}\right) \tag{10}$$

Following [8] we define the warping function

$$\left(z_i[t_1], \ldots, z_i[t_T]\right) = \mathbf{z}_i = \mathbf{h}(\mathbf{y}_i; \boldsymbol{\Psi})$$
$$= \left(h(y_i[t_1]; \boldsymbol{\Psi}), \ldots, h(y_i[t_T]; \boldsymbol{\Psi})\right) \tag{11}$$

where $\boldsymbol{\Psi} = (\alpha, \beta, \gamma)$, and we model $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_G)$ with a Gaussian process of (8):

$$p(\mathbf{z}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{C}(\boldsymbol{\theta})) \tag{12}$$

Using the standard variable transformation rule for probability densities, (12) implies the following distribution for the measured gene expression levels $\mathbf{y}$:

$$p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\Psi}) = \mathcal{N}\left(\mathbf{h}(\mathbf{y})|\mathbf{0}, \mathbf{C}(\boldsymbol{\theta})\right) \left(\frac{\partial \mathbf{h}(\mathbf{y}; \boldsymbol{\Psi})}{\partial \mathbf{y}}\right)^{-1} \tag{13}$$
$$= \mathcal{N}\left(\mathbf{h}(\mathbf{y})|\mathbf{0}, \mathbf{C}(\boldsymbol{\theta})\right) \prod_{i=1}^{G} \prod_{k=1}^{T} \left(\frac{\partial h_i(y_i(t_k); \boldsymbol{\Psi})}{\partial y_i(t_k)}\right)^{-1}$$

Inference is achieved by taking derivatives of the log likelihood $\log p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\Psi})$ with respect to both $\boldsymbol{\theta}$ and $\boldsymbol{\Psi}$, and applying a scaled conjugate gradient search for the maximum likelihood parameters. In this way, both the parameters of the covariance matrix, $\boldsymbol{\theta}$, and those of the nonlinear transformation, $\boldsymbol{\Psi}$, are learnt simultaneously under the same probabilistic framework. As we demonstrate in our simulation study, this can be expected to achieve better results than applying the transformation (10) in a separate data preprocessing step. The distribution of $z_i(t^*)$ at a new time point t^* has a Gaussian distribution

$$p(z_i(t^*)|\mathbf{x}_1, \ldots, \mathbf{x}_G, \boldsymbol{\theta}) = \mathcal{N}\left(z_i(t^*)|\hat{\mathbf{z}}(\boldsymbol{\theta}), \hat{\sigma}^2(\boldsymbol{\theta})\right) \tag{14}$$

where $\hat{\mathbf{z}}(\boldsymbol{\theta})$ and $\hat{\sigma}^2(\boldsymbol{\theta})$ are obtained by standard transformations of multivariate Gaussian distributions; see [4, 5] for explicit expressions. The distribution in the original data space is obtained by passing this Gaussian distribution through the non-linear warping function (10):

$$p(y_i(t^*)|\mathbf{x}_1, \ldots, \mathbf{x}_G, \boldsymbol{\theta}) \tag{15}$$
$$= \frac{dh[y_i(t^*)]}{dy_i(t^*)} \mathcal{N}\left(h[y_i(t^*), \boldsymbol{\Psi}]|\hat{z}(\boldsymbol{\theta}), \hat{\sigma}^2(\boldsymbol{\theta})\right)$$

Note that as opposed to (14), the distribution in (15) is *not* Gaussian. When we require a point prediction, rather than the whole distribution, it is convenient (for analytical tractability) to take the median

$$\operatorname{median}[y_i(t^*)] = h^{-1}(\hat{z}(\boldsymbol{\theta}); \boldsymbol{\Psi}) \tag{16}$$

We note that the proposed approach is a modification of the one proposed [8], with a warping function that is motivated by the transcriptional noise model (9).

3. SIMULATION

We tested the performance of the proposed scheme on data simulated in a similar manner as described in [9]. We assume that the unobservable TF activity has the form

$$f(t) = \sum_{j=1}^{4} a_j \exp\left(-\frac{(t - \mu_j)^2}{\sigma^2}\right) \tag{17}$$

with $\sigma = 1.5, a_1 = a_2 = 1.5, a_3 = a_4 = 0.5, \mu_1 = 4, \mu_2 = 6, \mu_3 = 8.5$ and $\mu_4 = 10.5$. We generated three gene expression profiles $\mathbf{x} = [x_i(t_1), \ldots, x_i(t_T)]$, $i = 1, 2, 3$, from (1–2) over 100 time points regularly spaced from 0 and 18, using the settings as: $B_1 = 0.01, B_2 = 7.5 \times 10^{-2}, B_3 = 2.5 \times 10^{-3}, S_1 = 1.0, S_2 = 0.4, S_3 = 0.4, D_1 = 1.0, D_2 = 0.05, D_3 = 0.001$. We then sampled six data points from each target gene, which provided the training data. Unlike [9], we did not simply add iid Gaussian noise, but simulated noisy measurements from (9), with parameter settings $c = 0, \sigma_\mu^2 = 0.01, \sigma_\varepsilon^2 = 0.01$. We compared five approaches: *GP:* standard Gaussian processes, as in [2]; *asinhGP:* standard Gaussian processes after pre-processing the data according to the transformation (10), estimating the parameters $\boldsymbol{\Psi}$ as described in [7]; *GPtanh:* warped Gaussian processes, with the mixture-of-tanh warping function proposed in [8]; *GPlog*: warped Gaussian processes with a logarithm warping function; and *GPasinh:* warped Gaussian processes with the warping function proposed in the present paper. For all applications, we optimized the parameters $\boldsymbol{\theta}$ and (if applicable) $\boldsymbol{\Psi}$ in a maximum likelihood sense using scaled conjugate gradients, and computed the posterior median according to (16). Table 1 shows the mean absolute prediction error for the estimated chemical kinetic parameters defined in (1), the gene expression profiles and TF activities at the fixed time points. Figures 1-5 show a comparison between the five methods in terms of the error distribution for various quantities, obtained from 20 datasets.

Table 1. The errors of estimated parameters, inferred gene expression profiles and transcriptional factor activity on simulated dataset (mean and median for 20 repeats).

model	GP	asinhGP	GPtanh	GPlog	GPasinh
B_{mean}	0.597	0.591	0.459	0.413	**0.181**
D_{mean}	0.319	0.356	0.193	0.367	**0.182**
S_{mean}	0.377	**0.265**	0.298	0.479	0.292
$Gene_{mean}$	0.515	0.589	0.533	0.433	**0.078**
TF_{mean}	0.430	0.497	0.232	0.281	**0.168**
B_{median}	0.486	0.442	0.384	0.353	**0.165**
D_{median}	0.269	0.251	**0.181**	0.217	0.197
S_{median}	0.410	**0.178**	0.266	0.395	0.300
$Gene_{median}$	0.252	0.439	0.159	0.144	**0.069**
TF_{median}	0.371	0.507	0.199	0.190	**0.162**

Our findings suggest that the proposed warped Gaussian process tends to achieve the lowest prediction error
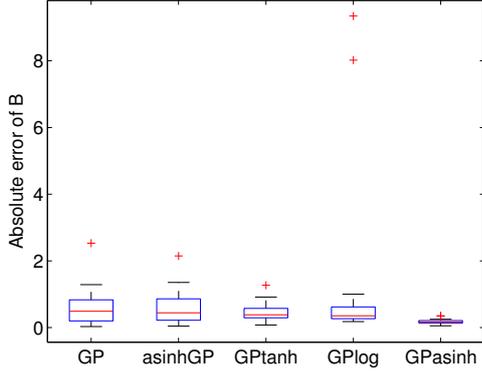
Figure 1. Boxplot of the median, the 25th and 75th percentiles, and the outliers of the absolute error of the basal transcription rate, B_i in (1), over 20 datasets.
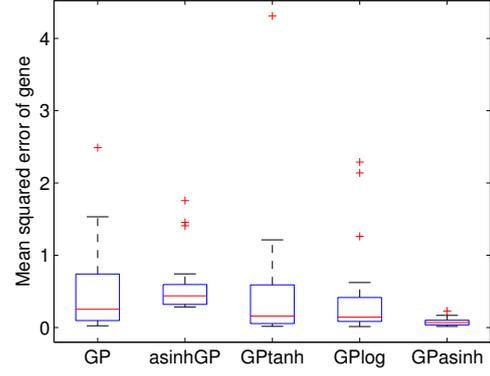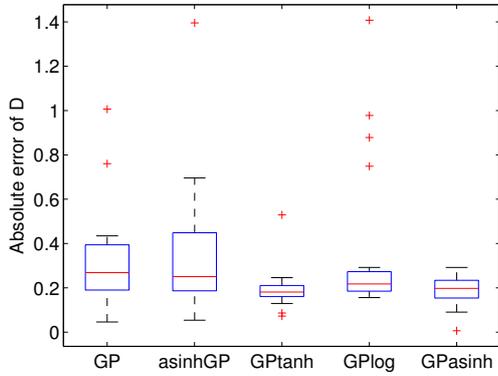

Figure 2. Boxplot of the median, the 25th and 75th percentiles, and the outliers of the absolute error of the decay rate, D_i in (1), over 20 datasets.

Figure 3. Boxplot of the median, the 25th and 75th percentiles, and the outliers of the absolute error of the sensitivity, S_i in (1), over 20 datasets.

and outperforms the competing approaches for the majority of the evaluation criteria. For certain reconstruction errors (reconstruction of the decay D_i, Figure 2, and

Figure 4. Boxplot of the median, the 25th and 75th percentiles, and the outliers of the mean squared error for all genes expression profiles, over 20 datasets
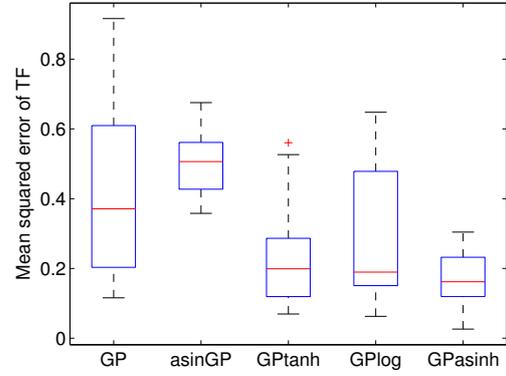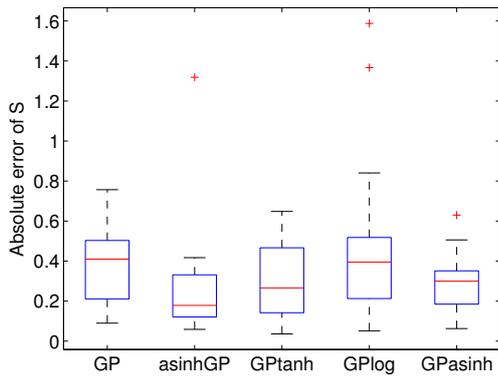

Figure 5. The boxplot of the median, the 25th and 75th percentiles, and the outliers of the mean squared error for the inferred transcription factor activity, over 20 datasets.

the TF activity, Figure 5) the transformation based on the mixture-of-tanh warping function, proposed in [8], is on a par with our method - but not consistently. It is particularly striking that nonlinearly transforming the data in a pre-processing step, as in [7], using the same warping function (10) as for our warped GP, achieves comparable result only for estimating the kinetic parameters S_i (Figure 3), but is otherwise outperformed by our method. This confirms the conjecture raised earlier that systematically inferring the parameters of the warping transformation simultaneously with the hyperparameters of the GP achieves better results than following [7] and applying the warping transformation in a separate data preprocessing step.

4. REAL-DATA APPLICATION

We have applied the warped Gaussian processes to the transcriptional profiles from Barenco's study [1], which reflect the expression levels of five target genes, DDB2, BIK, TNFRSF20b, p21 and hPA26, under the influence of regulation by a known transcription factor, P53. The

training data encompass five gene expression levels at 7 time points (0,2,4,6,8,10,12 hours). We compared our estimates of the kinetic parameters, defined in (1), with those referenced in Barenco's work. The results are shown in Table 2 and Figures 6-7 and suggest that the proposed warped Gaussian process with the arsinh transfer function of (10) infers kinetic parameters that, overall, are in good agreement with the parameters found in [1]. In terms of constructing the gene expression profiles, the warped GPs with both warping function, the arsinh function of (10) and the mixture-of-tanh functions from [8], are on a par, both outperforming the competing schemes. In terms of agreement of the inferred kinetic parameters with those from [1], our new warped GP with the arsinh function outperforms all other methods.

Table 2. The errors of estimated parameters and inferred gene expression profiles obtained from the real transcription profiles described in [1].

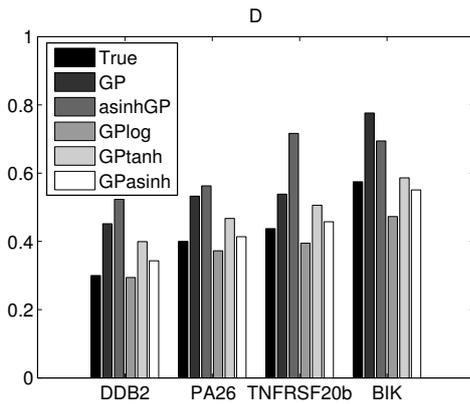| model	GP	asinhGP	GPtanh	GPlog	GPasinh
D	0.117	0.157	0.049	0.036	**0.02**
S	0.467	0.336	0.234	0.251	**0.178**
$Gene$	0.182	0.214	**0.158**	0.164	**0.158**

Figure 6. Comparison of the true and inferred decay rates, D_i in (1), as obtained with different methods from the transcription profiles in [1].

5. CONCLUSIONS

Gaussian processes have been proposed as a promising tool for modelling transcriptional regulation. However, the widely applied constant variance additive noise model (e.g. [4], Sect. 6.4.2) is oversimplistic and does not adequately reflect the intrinsic heteroscedastic nature of the noise. While warping the Gaussian process to correct for this effect tends to achieve a more reliable parameter and gene expression profile reconstruction, the empirical warping function proposed in [8] does not take into account the specific nature of the noise inherent in transcriptional regulation. In the present article we have proposed a warping function based on an explicit noise model for transcriptional regulation [6], which has been widely applied as a
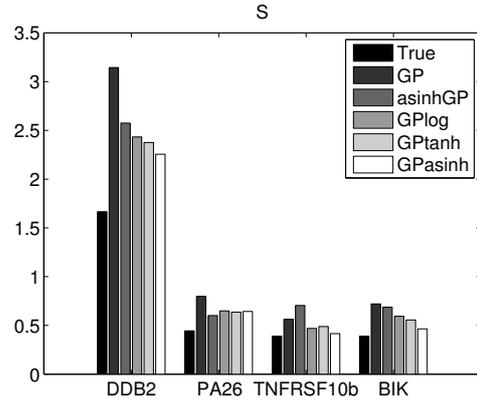

Figure 7. Comparison of the true and inferred sensitivities, S_i in (1), as obtained with different methods from the transcription profiles in [1].

variance stabilising transformation for transcriptional data [7]. We have shown that integrating this approach into the Gaussian process inference scheme achieves better results than transforming the data in a separate preprocessing step, and that our novel scheme outperforms warped Gaussian processes based on other warping functions.

6. REFERENCES

[1] M. Barenco et al., "Ranked prediction of p53 targets using hiddenvariable dynamic modeling," *Genome Biology*, vol. 7, no. 3, pp. R25, 2006.

[2] P. Gao, A. Honkela, M. Rattray, and N. D. Lawrence, "Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities," *Bioinformatics*, vol. 24, pp. i70–i75, 2008.

[3] A. Honkela et al., "Model-based method for transcription factor target identification with limited data," *PNAS*, vol. 107, no. 17, pp. 7793–7798, 2010.

[4] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, Singapore, 2006.

[5] C. E. Rasmussen and C. K. I Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[6] B.P Durbin, J.S. Hardin, D.M. Hawkins, and D.M. Rocke, "A variance-stabilizing transformation for gene-expression microarray data," *Bioinformatics*, vol. 18, no. suppl1, pp. S105–S110, 2002.

[7] W. Huber, A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron, "Variance stabilization applied to microarray data calibration and to the quantification of differential expression," *Bioinformatics*, vol. 18, no. suppl1, pp. S96–104, 2002.

[8] E. Snelsen, C.E. Rasmussen, and Z. Ghahramani, "Warped Gaussian processes," *Neural Information Processing Systems (NIPS)*, vol. 16, 2004.

[9] N.D. Lawrence, M. Rattray, P. Gao, and M. Titsias, *Gaussian Processes for Missing Species in Biochemical Systems.* In N.D. Lawrence, M. Girolami, M. Rattray, and G. Sanguinetti (eds.), *Learning and Inference in Computational Biology*, MIT Press, 2010.