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, Helsinki, Finland
hande.topa@aalto.fi

[2] Helsinki Institute for Information Technology HIIT
Department of Computer Science
University of Helsinki, Helsinki, Finland
antti.honkela@hiit.fi

Abstract

We present techniques for effective Gaussian process (GP) modelling of multiple short time series. These problems are common when applying GP models independently to each gene in a gene expression time series data set. Such sets typically contain very few time points. Naive application of common GP modelling techniques can lead to severe over-fitting or under-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 focus most of the energy spectrum to frequencies below the Nyquist frequency corresponding to the sampling frequency in the data set. Under-fitting can be avoided by more informative priors on observation noise. Combining these methods allows applying GP methods reliably automatically to large numbers of independent instances of short time series. This is illustrated with experiments with both synthetic data and real gene expression data.

1 Introduction

Gaussian processes (GPs) are a widely applied non-parametric probabilistic model for continuous data (Rasmussen and Williams, 2006). 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 (see, e.g. Quiñonero Candela and Rasmussen, 2005; Snelson and Ghahramani, 2006; Rasmussen and Williams, 2006; Titsias, 2009). 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 (Lawrence et al., 2007; Gao et al., 2008; Kirk and Stumpf, 2009; Stegle et al., 2010; Liu et al., 2010; Honkela et al., 2010; Kalaitzis and Lawrence, 2011; Cooke et al., 2011; Titsias et al., 2012). Reliable fitting of very large number of

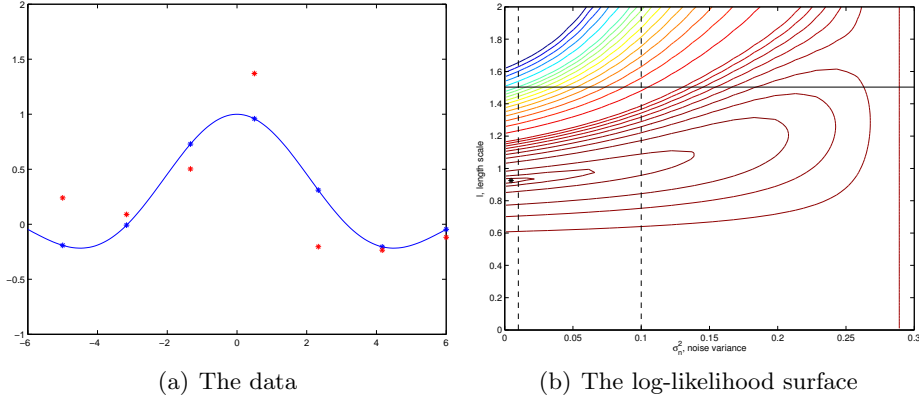|	
(a) The data	(b) The log-likelihood surface

Figure 1: (a) The underlying function (blue line) for the synthetic example and an example data set instance (red dots). (b) Contour plot of the log-likelihood surface for different ℓ and σ_n^2 values corresponding to the data set in (a) showing the maximum likelihood solution (black dot) with a very small $\sigma_n^2 \approx 0.005$ and $\ell \approx 0.9$, when no bounds are introduced.

independent models is important in many applications of these models, such as ranking of targets of gene regulators (Honkela et al., 2010).

Most gene expression time series are very short with a great majority having less than 9 time points (Ernst et al., 2005), 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 some model parameters that help avoid these phenomena. Earlier heuristic variants of the length-scale bound have been applied in some previous works (Honkela et al., 2011; Titsias et al., 2012) without detailed justification, but here we present a new rigorous derivation for the bound.

2   GP modelling methods for small data sets

A GP is a stochastic process $\{f(\mathbf{x})|\mathbf{x} \in \mathcal{X}\}$ for which the marginal distribution at any finite sub-collection of points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is multivariate Gaussian (Rasmussen and Williams, 2006). The process is completely defined by the mean function $\mu(\mathbf{x})$ and the covariance function $k(\mathbf{x}, \mathbf{x}')$, that also define the mean vector and covariance matrix of the multivariate Gaussian over the sub-collection. For simplicity, we assume the mean function is identically zero $\mu(\mathbf{x}) \equiv 0$.

The most widely used covariance function for GPs in machine learning is the squared exponential covariance (Rasmussen and Williams, 2006)

$$k_{\mathrm{SE}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{r^2}{2\ell^2}\right),\tag{1}$$

where $r = ||\mathbf{x} - \mathbf{x}'||$. The covariance depends on two positive hyperparameters: variance σ_f^2 and length-scale ℓ. The squared exponential covariance is infinitely smooth, which is often too strong
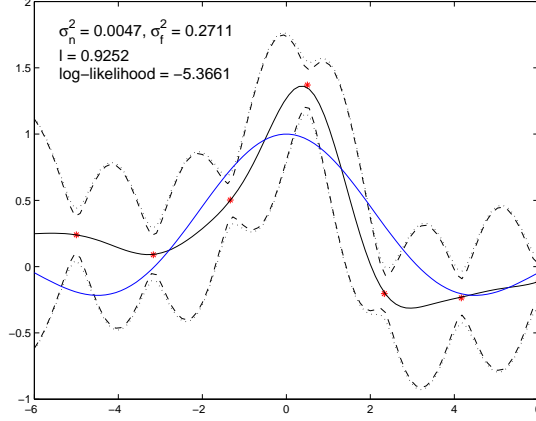
Figure 2: A GP model fitted to 7 data points generated from $f(x) = \frac{\sin(x)}{x}$ showing over-fitting to the data. The plot shows the posterior mean of the GP (black solid line) together with two standard deviation posterior credible regions both for the squared exponential covariance (dashed line) and squared exponential plus noise covariance (dotted line). The two posterior credible regions are very close to each other, indicating a very small estimated observation noise level.

an assumption. A simple generalisation is given by the Matérn class covariance functions

$$k_{\mathrm{Matern}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}r}{\ell} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu}r}{\ell} \right), \tag{2}$$

with additional positive hyperparameter ν, where K_{ν} is a modified Bessel function (Rasmussen and Williams, 2006).

Both these covariance functions have a length-scale parameter ℓ that 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 over-fit.

2.1 Illustrative Example

We illustrate fitting GPs to small data sets with synthetic data generated from a $\mathrm{sinc}(x)$ function, that is, $f(x) = \frac{\sin(x)}{x}$, by uniformly sampling 7 data points in the interval [-5,6] with a noise term which is normally distributed with mean 0 and variance 0.09. An example of such a data set is shown in Fig. 1(a).

Fitting a GP model with squared exponential covariance to the data set shown in Fig. 1(a) and selecting the maximum likelihood (ML) solution for the squared exponential covariance variance σ_f^2 leads to a likelihood surface for length scale ℓ and observation noise variance σ_n^2 shown in Fig. 1(b). The ML estimate for the noise is clearly much smaller than the generative value indicating severe over-fitting to the data. This corresponds to a fairly small value for the length-scale compared to the sampling rate in the data. The GP model corresponding to the ML fit is shown in Fig. 2.

2.2 Length-scale bounds

As seen above, 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 $f_n = \frac{1}{2\Delta t}$ is the maximal frequency that can be identified in the spectral representation of the sampled signal (Tick and Shaman, 1966; Oppenheim and Schafer, 1975). Here Δt is the sampling interval in the data set. In case of non-uniformly sampled data, 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

$$S_{\text{SE}}(s) = (2\pi\ell^2)^{D/2} \exp(-2\pi^2\ell^2 s^2), \tag{3}$$

where D is the number of dimensions and s denotes the frequency (Rasmussen and Williams, 2006). For $D = 1$, the corresponding lower bound for the length-scale can be found by solving the inequality

$$\int_{-\frac{1}{2\Delta t}}^{\frac{1}{2\Delta t}} S_{\text{SE}}(s) \, \mathrm{d}s = \text{erf}\left(\frac{\pi\ell}{\sqrt{2}\Delta t}\right) \geq \alpha, \tag{4}$$

where α denotes the fraction of the system's energy on the frequencies that are below the Nyquist frequency.

Solving for ℓ and setting α to 0.99, we can obtain the lower bound for the length scale parameter that would constrain at least 99% of the process's energy on the frequencies which are below the Nyquist frequency:

$$\ell \geq a_\ell(\alpha) = \frac{\sqrt{2}\,\text{erfinv}(\alpha)}{\pi}\Delta t \approx 0.8199 \times \Delta t, \tag{5}$$

where $\text{erfinv}(x)$ denotes the inverse of the error function.

For the 1-dimensional Matérn covariance function the corresponding spectral density is

$$S_{\text{Matern}}(s) = \frac{2\sqrt{\pi}\gamma(\nu + \frac{1}{2})(2\nu)^\nu}{\gamma(\nu)\ell^{2\nu}}\left(\frac{2\nu}{\ell^2} + 4\pi^2 s^2\right)^{-(\nu+\frac{1}{2})}, \tag{6}$$

from which we can derive the fraction of energy in the frequency interval $[-1/2\Delta t, 1/2\Delta t]$ as

$$\int_{-\frac{1}{2\Delta t}}^{\frac{1}{2\Delta t}} S_{\text{Matern}}(s) \, \mathrm{d}s = \frac{4\ell\sqrt{2\pi}\gamma(\nu + 1/2)}{\Delta t\sqrt{v}\gamma(\nu)} \, _2F_1\left(\frac{1}{2}, \nu + \frac{1}{2}, \frac{3}{2}, -\frac{\ell^2\pi^2}{2\nu(\Delta t)^2}\right), \tag{7}$$

where $_2F_1(a, b, c, x)$ is the hypergeometric function (Abramowitz and Stegun, 1965). Given a fixed value of ν, appropriate lower bound a_ℓ for ℓ can be derived from Eq. (7) using numerical optimisation.

For Bayesian parameter estimation, a uniform prior over the length scale over the interval $[a_\ell, t_n - t_1]$ seems like a reasonable objective prior.

2.3 Variance prior

Bounding the length-scale as above can avoid most obvious over-fitting, but naive estimation of observation noise will still frequently lead to mis-estimation of the noise and hence effectively over-fitting or under-fitting the data. The easiest way to avoid this is to use informative priors on the noise that focus the distribution away from implausibly small and large values.
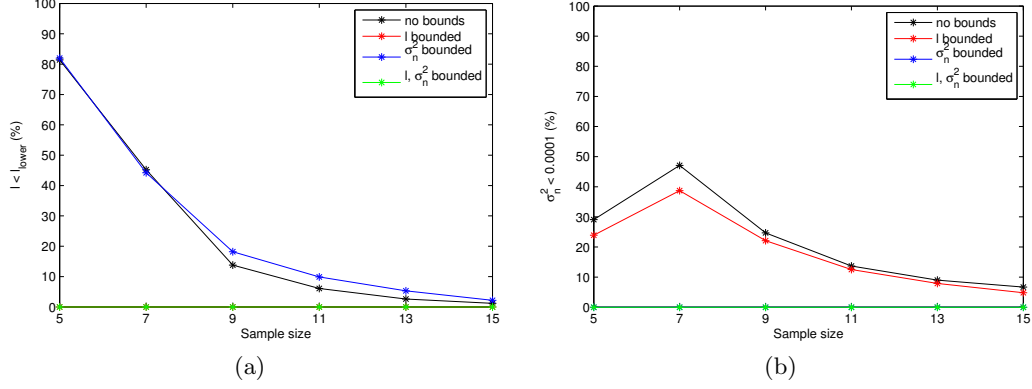
Figure 3: Fraction of over-fitted models in different setups with synthetic data. Setting the length-scale bound (a) automatically eliminates all problems with length scales and a noise bound (b) with the noise values, and the corresponding curves are overlapping at constant 0.
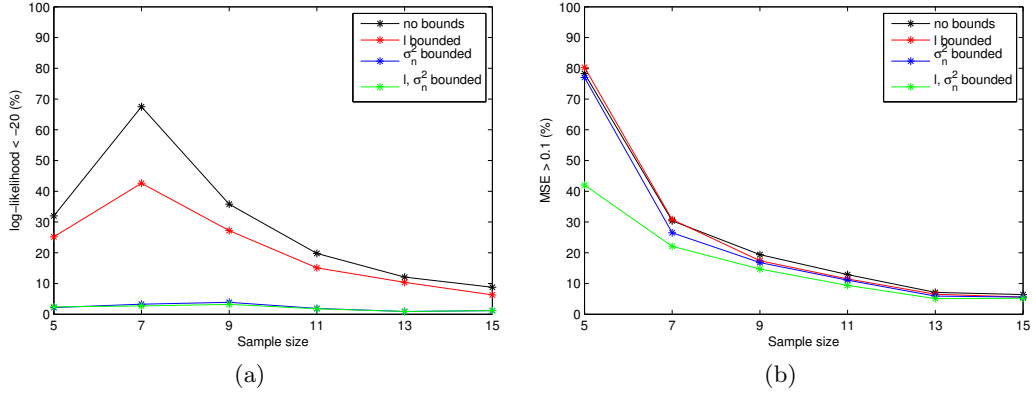

Figure 4: Fraction of models with low predictive log-likelihood values (a) and with high MSE values (b) in different setups with synthetic data.

For gene expression data, one can for example use posterior variances of the inferred expression levels from pre-processing both for microarrays (Liu et al., 2005; Du et al., 2008) and RNA-sequencing (Turro et al., 2011; Glaus et al., 2012). This kind of approach has been applied e.g. in (Lawrence et al., 2007; Gao et al., 2008; Honkela et al., 2010; Titsias et al., 2012).

As an alternative, Cooke et al. (2011) use an empirical variance estimate from multiple replicates as the approximate lower bound and total variance as the approximate upper bound for a semi-empirical prior on the variance.

3 Experiments

We present experimental results highlighting the over-fitting caused by bad length-scale estimates on synthetic data and real gene expression data.

(a) No bounds

(b) Bounded ℓ

(c) Bounded σ_n^2
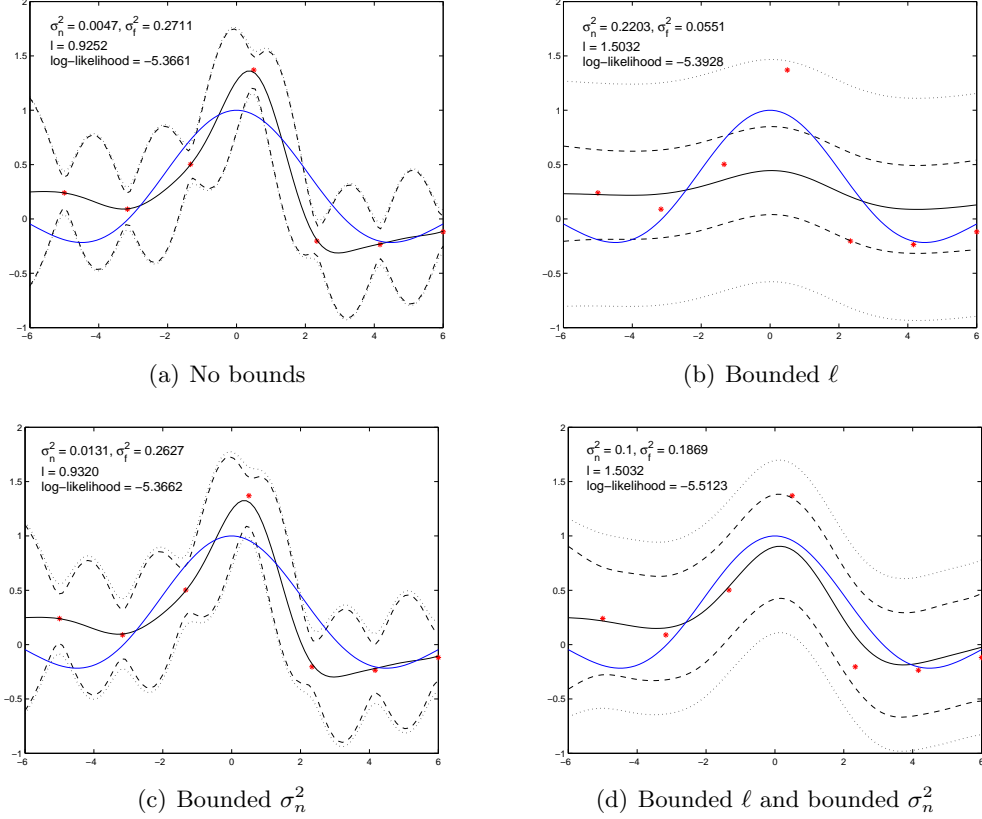
(d) Bounded ℓ and bounded σ_n^2

Figure 5: Examples of GP model fits for a synthetic data instance. Using no bounds in (a) leads to severe over-fitting which is compensated by under-fitting by introducing the ℓ bound ($[a_\ell, \infty) = [1.5032, \infty)$, from $\Delta t \approx 1.8334$) in (b), while introducing only the σ_n^2 bound ($[0.01, 0.1]$) does not help avoid over-fitting in (c). Introducing the ℓ and σ_n^2 bounds together in (d) leads to a reasonably good fit to the original function.

3.1 Synthetic data

We generated synthetic data using a procedure similar to the one described in Sec. 2.1. We created 1000 independent instances of the data set with different noise realisations. We repeated this sampling $n = 5, 7, \ldots, 15$ equally spaced points on the interval $[-5, 6]$.

For each of these data sets, we fitted the model in four different scenarios:

1. Unconstrained ℓ parameter estimation and noise σ_n^2 estimation;

2. Unconstrained σ_n^2 estimation and bounded ℓ in the range $[a_\ell, \infty)$;

3. Bounded σ_n^2 in the range $[0.01, 0.1]$ and unconstrained ℓ; and

4. Bounded σ_n^2 in the range $[0.01, 0.1]$ and bounded ℓ in the range $[a_\ell, \infty)$.

For each n and each scenario, we recorded the number of fits with $\ell < a_\ell$ and $\sigma_n^2 < 0.0001$, both of which are indications of over-fitting. The results are shown in Fig. 3. For very small data sets, practically all instances lead to a short length scale, and a significant fraction has a very small estimated noise variance. For larger data sets the fractions drop, but remain well above zero.

Alternatively, to compare how the fitted GP models coincide with the true underlying function, we calculated the mean squared errors (MSE) by using 10 test points, equally spaced in the range [-6,5]. MSE values can be found simply by calculating the mean of the squared differences between the true underlying function values and the predicted function values at the test points. Additionally, we also computed the predictive log-likelihood values for the true function values at the test points. Small MSE values, and large log-likelihood values can be considered as the indicators of a good fit, whereas high MSE values and low log-likelihood values indicate the opposite.

The frequencies of the instances with very low predictive log-likelihood values (smaller than -20) and with very high MSE values (larger than 0.1) can be seen in Fig. 4(a) and Fig. 4(b), respectively. It is clear that the frequencies are very high if no bounds are set to the parameters. Once the parameter bounds are introduced, the frequencies start to decrease, with the largest decrease occuring in the instances with small sample sizes. Furthermore, for each instance, we recorded in which setup the fitted GP model leads to the smallest MSE and the largest predictive log-likelihood values. In Table 1 and Table 2, the fractions of the largest predictive log-likelihood values and the smallest MSE values in four different setups are presented, supporting the fact that using the length scale and noise variance bounds together improves the model fit drastically especially in the instances with small sample sizes.

Model fits for an example realisation, with sample size 7, suffering from over-fitting with un-bounded estimation are shown in Fig. 5. The over-fitting in the unbounded estimate in Fig. 5(a) is clearly remedied by introducing the length-scale bound in Fig. 5(b), but this makes the model under-fit. Also constraining the noise variance to sensible values in Fig. 5(d) leads to a reasonably good fit considering the amount of data.

Table 1: Fractions (%) of the largest log-likelihoods in different setups with synthetic data (n denotes the sample size)

	$n=5$	$n=7$	$n=9$	$n=11$	$n=13$	$n=15$
NO BOUNDS	2.2	3.0	9.6	10.6	11.9	11.6
ℓ BOUNDED	1.0	8.5	11.6	14.9	13.7	13.0
σ_n^2 BOUNDED	28.3	34.2	44.1	37.6	39.7	34.6
ℓ AND σ_n^2 BOUNDED	68.5	54.3	34.7	36.9	34.7	40.8

Table 2: Fractions (%) of the smallest MSEs in different settings with synthetic data (n denotes the sample size)

	$n=5$	$n=7$	$n=9$	$n=11$	$n=13$	$n=15$
NO BOUNDS	8.1	9.4	13.7	19.1	22.3	23.6
ℓ BOUNDED	5.7	14.3	20.0	22.6	25.0	26.0
σ_n^2 BOUNDED	27.0	36.2	38.8	28.9	26.5	21.5
ℓ AND σ_n^2 BOUNDED	59.2	40.1	27.5	29.4	26.2	28.9

3.2   Gene expression data

We apply the model to fruit fly *Drosophila melanogaster* developmental gene expression time series from (Tomancak et al., 2002) using the differential-equation-based gene regulation model from (Lawrence et al., 2007; Gao et al., 2008; Honkela et al., 2010).

Table 3: Proportion of over-fitted single-target cascaded differential equation models according to different criteria in different experimental setups. '.' indicates a situation that is not possible because of the bounds.

	BIN		MEF2		TIN		TWI	
	$\ell < a_\ell$	$\sigma_n^2 < 0.01$	$\ell < a_\ell$	$\sigma_n^2 < 0.01$	$\ell < a_\ell$	$\sigma_n^2 < 0.01$	$\ell < a_\ell$	$\sigma_n^2 < 0.01$
No bounds	0.0%	0.0%	0.0%	0.3%	13.9%	2.9%	0.0%	1.6%
ℓ bounded	.	0.0%	.	0.3%	.	4.1%	.	1.6%
σ_n^2 fixed	1.8%	.	1.9%	.	5.3%	.	7.1%	.
ℓ bounded, σ_n^2 fixed	.	.	.	.	.	.	.	.

The experiments were run using a modified version of the *tigre* Bioconductor package (Honkela et al., 2011). For each model, we tested 4 different scenarios:

1. Unconstrained ℓ parameter estimation and noise σ_n^2 estimation;

2. Unconstrained σ_n^2 estimation and bounded ℓ;

3. Fixed σ_n^2 from pre-processing and unconstrained ℓ; and

4. Fixed σ_n^2 and bounded ℓ.

For the last two, the noise variances were obtained from data pre-processing as in (Honkela et al., 2010). The models were run for 6795 genes passing the significant activity filter described in (Honkela et al., 2010).

Single-target cascaded differential equation models As the first example, we tested single-target cascaded differential equation models linking observed regulator TF mRNA to candidate target mRNA. We ran the model for 4 TFs: BIN, MEF2, TIN and TWI. Fractions of genes with $\ell < a_\ell$ or $\sigma_n^2 < 0.01$ for each TF for each setting are listed in Table 3.

The results show a moderate number of genes that exhibit symptoms of over-fitting. The numbers vary significantly for different TFs, as the strength of the driving signal is different. The expression input for TIN is especially weak and sharply peaked, which leads to relatively larger number of over-fitted models without the precautions proposed in this paper. An example of such a model is illustrated in Fig. 6.

Multiple-target models As the second example we consider a slightly more difficult task of fitting a single-layer differential equation model with no TF mRNA input (Gao et al., 2008). We fit the model to each gene in turn with two fixed known targets, MEF2 and TIN. These are both known targets of TWI (Zinzen et al., 2009), so these models could plausibly be used to discover further targets of TWI. Each model was fitted independently without using any information from the previous models.

Fractions of genes with different symptoms of over-fitting are listed in Table 4. The numbers are in most cases slightly higher than in Table 3, demonstrating that this is a more challenging task.

4   Discussion

We have presented methods for improving large-scale learning of GP models for very many instances of small data sets. We presented a novel rigorous derivation for a bound for sensible length scales
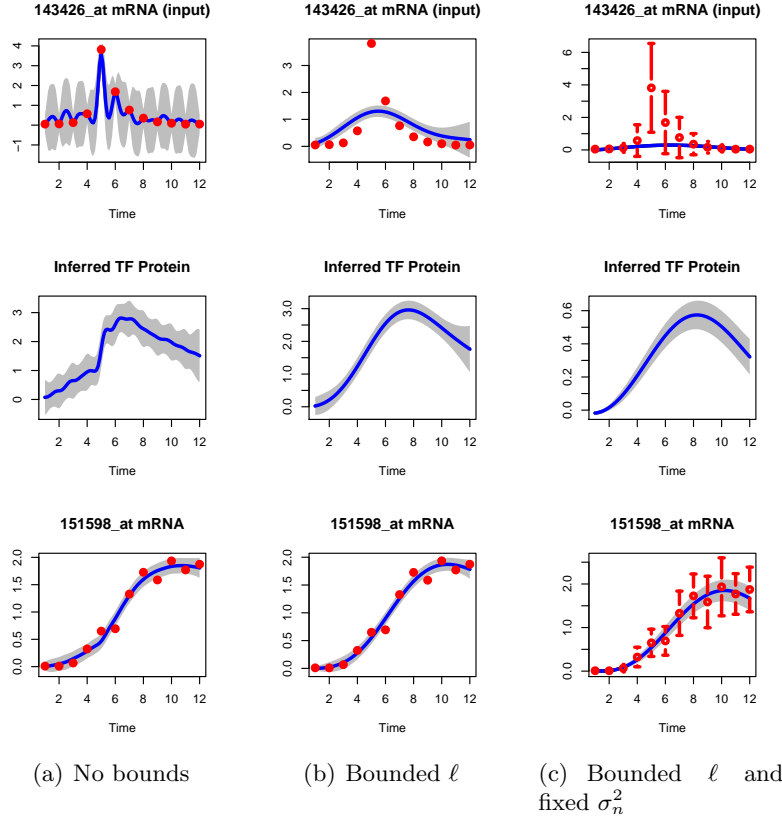
Figure 6: Example of a single-target cascaded ODE gene regulation model. Using no bounds in (a) leads to severe over-fitting which is mostly corrected by ℓ bound in (b). More accurate noise model in (c) further improves the accuracy slightly.

Table 4: Proportion of over-fitted multiple-target non-cascaded differential equation models according to different criteria in different experimental setups.

	$\ell < a_\ell$	$\sigma_n^2 < 0.01$
No bounds	7.3%	8.3%
ℓ bounded	.	6.7%
σ_n^2 fixed	6.7%	.
ℓ bounded, σ_n^2 fixed	.	.

for squared exponential and Matérn covariance functions. The bound is based on constraining the energy spectrum of the GP covariance to frequencies that can plausibly be reconstructed from the data. This can be intuitively justified by the fact that the data was collected at such sampling rate in the first place, which encodes a prior assumption about the time scale of interest. This bound clearly helps avoid many cases of obvious over-fitting, as illustrated with both synthetic and real data experiments.

The usual underlying reason for the small length-scale fits is often that a smooth model may not be a very good fit to the specific instance of data. The GP model may attempt to compensate for this by using more functional degrees of freedom than seems plausible and essentially modelling each single observation independently. For highly constrained models such as the linear ODE

9

models in Sec. 3.2 the results are usually not very severe, but for more flexible models such as ones incorporating gene-dependent delays the over-fitting can be even more severe problem.

One may wonder whether the observed small length-scales are caused by using point estimates for the parameters. This does not appear to be the case, as illustrated by the likelihood surface in Fig. 1(b) which shows a very smooth maximum in small length-scale regime. This is further supported by the fact that some form of length-scale bounds was needed by Titsias et al. (2012) for a method completely based on MCMC.

We believe the presented length-scale bound is an important ingredient for large-scale learning of multiple independent GP models for short time series that are very common in biology. Carefully justified priors will be especially important in the future when moving beyond the simple linear ODE models of Honkela et al. (2010) to more realistic and flexible models. Using such models without suitable constraints or suitably constrained priors can easily lead to unexpected over-fitting failure modes.

References

M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables.* Courier Dover, New York, NY, U.S.A., 1965. ISBN 0-486-61272-4.

E. J. Cooke, R. S. Savage, P. D. W. Kirk, R. Darkins, and D. L. Wild. Bayesian hierarchical clustering for microarray time series data with replicates and outlier measurements. *BMC Bioinformatics*, 12:399, 2011. doi: 10.1186/1471-2105-12-399. URL http://dx.doi.org/10.1186/1471-2105-12-399.

P. Du, W. A. Kibbe, and S. M. Lin. lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, 24(13):1547–1548, Jul 2008. doi: 10.1093/bioinformatics/btn224. URL http://dx.doi.org/10.1093/bioinformatics/btn224.

J. Ernst, G. J. Nau, and Z. Bar-Joseph. Clustering short time series gene expression data. *Bioinformatics*, 21 Suppl 1:i159–i168, Jun 2005. doi: 10.1093/bioinformatics/bti1022. URL http://dx.doi.org/10.1093/bioinformatics/bti1022.

P. Gao, A. Honkela, M. Rattray, and N. D. Lawrence. Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics*, 24(16):i70–i75, Aug 2008. doi: 10.1093/bioinformatics/btn278. URL http://dx.doi.org/10.1093/bioinformatics/btn278.

P. Glaus, A. Honkela, and M. Rattray. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*, 28(13):1721–1728, Jul 2012. doi: 10.1093/bioinformatics/bts260. URL http://dx.doi.org/10.1093/bioinformatics/bts260.

A. Honkela, C. Girardot, E. H. Gustafson, Y.-H. Liu, E. E. M. Furlong, N. D. Lawrence, and M. Rattray. Model-based method for transcription factor target identification with limited data. *Proc Natl Acad Sci U S A*, 107(17):7793–7798, Apr 2010. doi: 10.1073/pnas.0914285107. URL http://dx.doi.org/10.1073/pnas.0914285107.

A. Honkela, P. Gao, J. Ropponen, M. Rattray, and N. D. Lawrence. tigre: Transcription factor inference through Gaussian process reconstruction of expression for Bioconductor. *Bioinformatics*, 27(7):1026–1027, Apr 2011. doi: 10.1093/bioinformatics/btr057. URL http://dx.doi.org/10.1093/bioinformatics/btr057.

A. A. Kalaitzis and N. D. Lawrence. A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. *BMC Bioinformatics*, 12:180, 2011. doi: 10.1186/1471-2105-12-180. URL http://dx.doi.org/10.1186/1471-2105-12-180.

P. D. W. Kirk and M. P. H. Stumpf. Gaussian process regression bootstrapping: exploring the effects of uncertainty in time course data. *Bioinformatics*, 25(10):1300–1306, May 2009. doi: 10.1093/bioinformatics/btp139. URL `http://dx.doi.org/10.1093/bioinformatics/btp139`.

N. D. Lawrence, G. Sanguinetti, and M. Rattray. Modelling transcriptional regulation using Gaussian processes. In B. Schölkopf, J. C. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems*, volume 19, pages 785–792. MIT Press, Cambridge, MA, 2007.

Q. Liu, K. K. Lin, B. Andersen, P. Smyth, and A. Ihler. Estimating replicate time shifts using Gaussian process regression. *Bioinformatics*, 26(6):770–776, Mar 2010. doi: 10.1093/bioinformatics/btq022. URL `http://dx.doi.org/10.1093/bioinformatics/btq022`.

X. Liu, M. Milo, N. D. Lawrence, and M. Rattray. A tractable probabilistic model for Affymetrix probe-level analysis across multiple chips. *Bioinformatics*, 21(18):3637–3644, Sep 2005. doi: 10.1093/bioinformatics/bti583. URL `http://dx.doi.org/10.1093/bioinformatics/bti583`.

A. V. Oppenheim and R. W. Schafer. *Digital Signal Processing*. Prentice-Hall, 1975.

J. Quiñonero Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *J. Mach. Learn. Res.*, 6:1939–1959, Dec. 2005. ISSN 1532-4435.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006. ISBN 0-262-18253-X.

E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18, Cambridge, MA, 2006. MIT Press.

O. Stegle, K. J. Denby, E. J. Cooke, D. L. Wild, Z. Ghahramani, and K. M. Borgwardt. 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. doi: 10.1089/cmb.2009.0175. URL `http://dx.doi.org/10.1089/cmb.2009.0175`.

L. J. Tick and P. Shaman. Sampling rates and appearance of stationary gaussian processes. *Technometrics*, 8:91–106, 1966.

M. K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelfth International Workshop on Artificial Intelligence and Statistics*, volume 5, pages 567–574, Clearwater Beach, FL, 16-18 April 2009. JMLR W&CP 5.

M. K. Titsias, A. Honkela, N. D. Lawrence, and M. Rattray. Identifying targets of multiple co-regulating transcription factors from expression time-series by Bayesian model comparison. *BMC Syst Biol*, 6(1):53, May 2012. doi: 10.1186/1752-0509-6-53. URL `http://dx.doi.org/10.1186/1752-0509-6-53`.

P. Tomancak, A. Beaton, R. Weiszmann, E. Kwan, S. Shu, S. E. Lewis, S. Richards, M. Ashburner, V. Hartenstein, S. E. Celniker, and G. M. Rubin. Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biol*, 3(12):RESEARCH0088, 2002.

E. Turro, S.-Y. Su, A. Gonalves, L. J. M. Coin, S. Richardson, and A. Lewin. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol*, 12(2):R13, 2011. doi: 10.1186/gb-2011-12-2-r13. URL `http://dx.doi.org/10.1186/gb-2011-12-2-r13`.

R. P. Zinzen, C. Girardot, J. Gagneur, M. Braun, and E. E. M. Furlong. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, 462(7269):65–70, Nov 2009. doi: 10.1038/nature08531. URL `http://dx.doi.org/10.1038/nature08531`.