Modelling Transcriptional Regulation with Gaussian processes

Neil Lawrence and Magnus Rattray
School of Computer Science
University of Manchester

7th November 2007

Outline

Online Resources

- All source code and slides are available online
  - This talk available from my home page (see talks link on side).
  - MATLAB examples in the 'oxford' toolbox (vrs 0.13).
    - `http://www.cs.man.ac.uk/~neill/oxford/`.
  - And the 'gpsim' toolbox (vrs 0.1).
    - `http://www.cs.man.ac.uk/~neill/gpsim/`.
  - MATLAB commands used for examples given in `typewriter font`.

Introduction to Gaussian Processes

- TFAs can be seen as *latent chemical species*.
- In Magnus' talk we saw how they can be modelled with Kalman filters.
- Gaussian processes (GPs) are probabilistic models for functions.
  [O'Hagan, 1978, 1992, Rasmussen and Williams, 2006]
- GPs allow inference about functions in the presence of uncertainty.

Defining a Distribution over Functions

- Gaussian Process
  - What is meant by a distribution over functions?
  - Functions are infinite dimensional objects:
    - Defining a distribution over functions seems non-sensical.

- Gaussian Distribution
  - Start with a standard Gaussian distribution.
  - Consider the distribution over a fixed number of instantiations of the function.
  - A multi-variate Gaussian distribution is defined by a mean and a covariance matrix.
  - We consider the special case where the mean is zero,

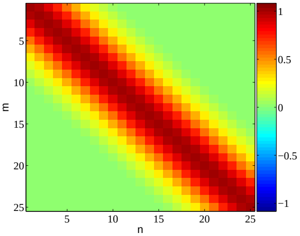
$$N(\mathbf{f}|\mathbf{0}, \mathbf{K}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\mathbf{K}|^{\frac{1}{2}}} \exp\left(-\frac{\mathbf{f}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{f}}{2}\right).$$

Covariance Functions

- RBF Kernel Function

$$k(\mathbf{x}_m, \mathbf{x}_n) = \alpha \exp\left(-\frac{\left|\left|\mathbf{x}_m - \mathbf{x}_n\right|\right|^2}{2l^2}\right)$$

- Covariance matrix is built using the *inputs* to the function $\mathbf{x}_n$.
  - For the example above it was based on Euclidean distance.
  - The covariance function is also know as a kernel.



Covariance Samples

- `demCovFuncSample`

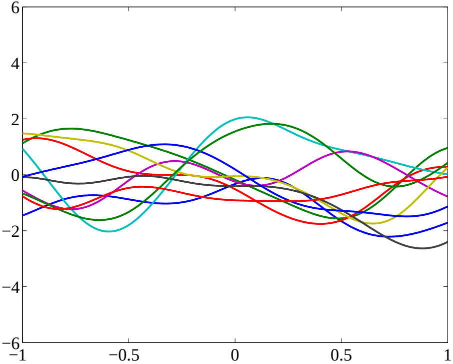


Figure: RBF kernel with $\gamma = 10$, $\alpha = 1$

- `demCovFuncSample`

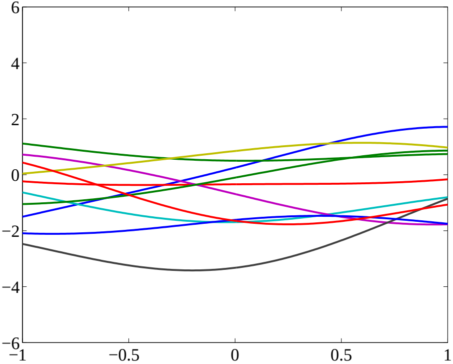


Figure: RBF kernel with $l = 1$, $\alpha = 1$

Covariance Samples

- `demCovFuncSample`

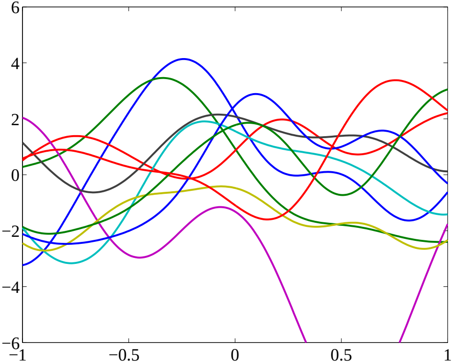


Figure: RBF kernel with $l = 0.3$, $\alpha = 4$

Covariance Samples

- `demCovFuncSample`

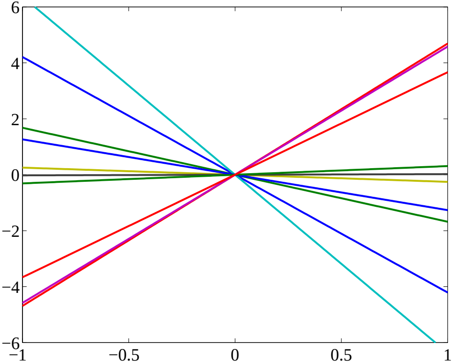


Figure: linear kernel with $\alpha = 16$

Covariance Samples

- `demCovFuncSample`

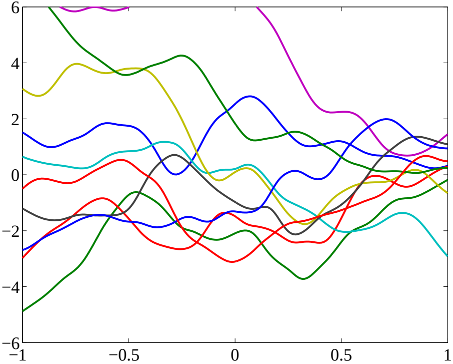


Figure: MLP kernel with $\alpha = 8$, $w = 100$ and $b = 100$

Covariance Samples

- `demCovFuncSample`

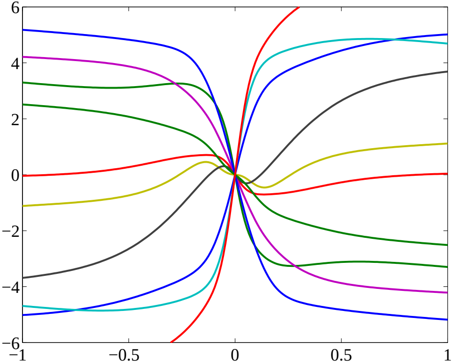


Figure: MLP kernel with $\alpha = 8$, $b = 0$ and $w = 100$

Covariance Samples

- `demCovFuncSample`

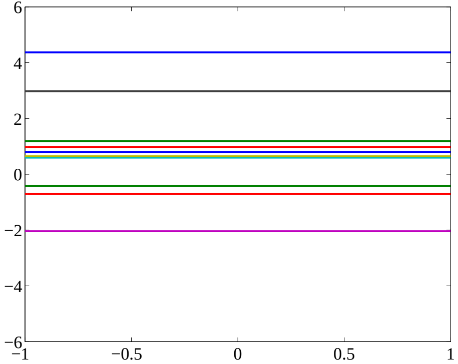


Figure: bias kernel with $\alpha = 1$ and

Covariance Samples

- `demCovFuncSample`

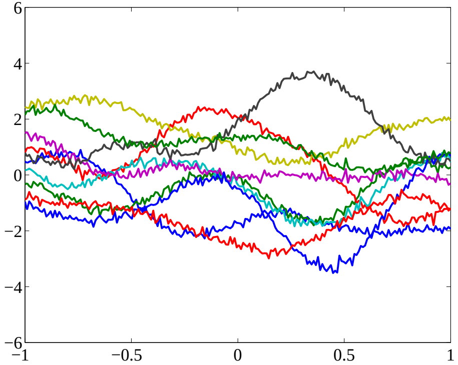


Figure: summed combination of: RBF kernel, $\alpha = 1$, $l = 0.3$; bias kernel, α =1; and white noise kernel, $\beta = 100$

Gaussian Process Regression

- `demRegression`

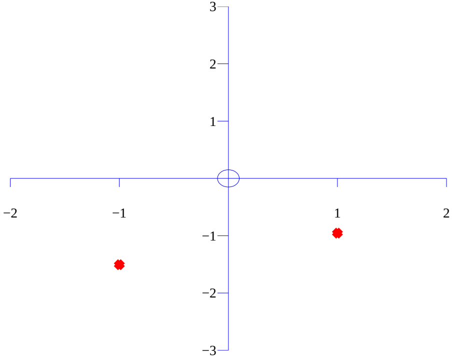


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

- `demRegression`

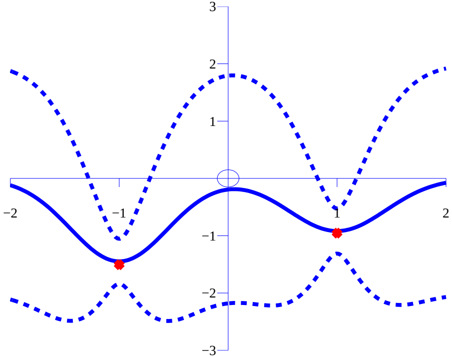


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

- `demRegression`

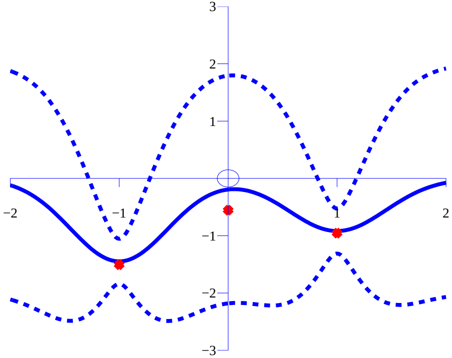


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

- `demRegression`

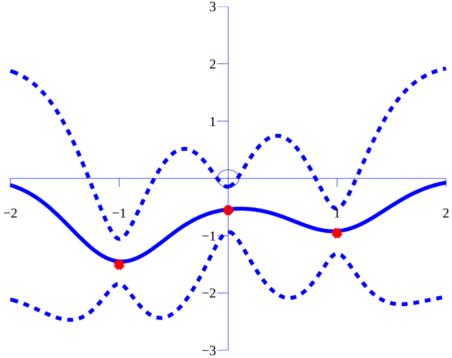


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

- `demRegression`

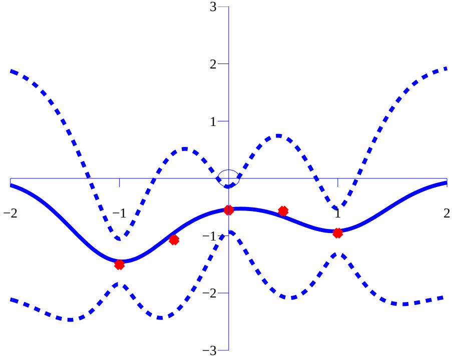


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

- `demRegression`

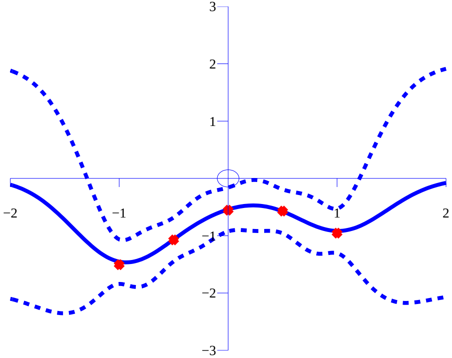


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

- `demRegression`

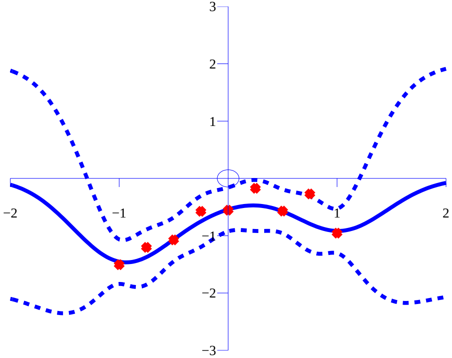


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

- `demRegression`

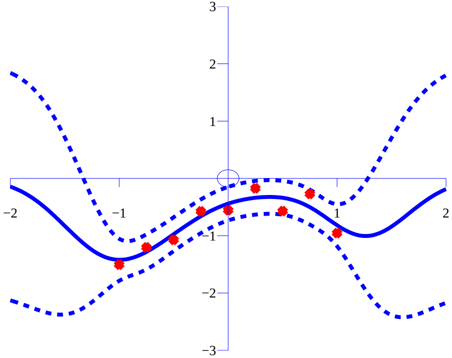


Figure: Examples include WiFi localization, C14 callibration curve.

Application

- A missing chemical species (e.g. transcription factor).
- Aim: infer its value with Gaussian processes.
- Differential Equation model
  - Simple linear model differential equation model recently used by Barenco et al. [2006].
  - We repeat their experiments with Gaussian processes.

Simple Linear Model

- Linear model of regulation

$$\frac{dy_i(t)}{dt} = B_i + S_i f(t) - D_i y_i(t)$$

- where:

$y_i(t)$	—	expression of the ith gene at time t.
$f(t)$	—	concentration of the transcription factor at time t.
D_i	—	gene's decay rate.
B_i	—	basal transcription rate.
S_i	—	sensitivity to the transcription factor.

Equation Solution

- Solve via Laplace Transforms
  - Solution to the equation:

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

If $f(t)$ is a zero mean Gaussian process then $y_i(t)$ is also a Gaussian process with mean $\frac{B_i}{D_i}$.

Two Properties of GPs

The integral of a GP is also a GP,

$$f(t) \sim N(\mathbf{0}, \mathbf{K}_{ff})$$

and

$$g(t) = \int_0^t f(u)\, du$$

then

$$g(t) \sim N(\mathbf{0}, \mathbf{K}_{gg}),$$

where

$$k_{gg}(t, t') = \int_0^t \int_0^{t'} k_{ff}(u, u')\, du du'$$

Two Properties of GPs

Product with deterministic function, if we have

$$f(t) \sim N(\mathbf{0}, \mathbf{K}_{ff}),$$

and

$$g(t) = f(t)\, h(t)$$

where $h(t)$ is a deterministic function then,

$$g(t) \sim N(\mathbf{0}, \mathbf{K}_{gg}),$$

where

$$k_{gg}(t, t') = h(t)\, k_{ff}(t, t')\, h(t')$$

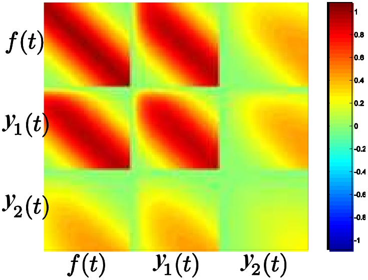
Covariance for Transcription Model

- RBF Kernel function for $f(t)$

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

- Joint distribution for $y_1(t)$, $y_2(t)$ and $f(t)$.
  - Here:

D_1	S_1	D_2	S_2
5	5	0.5	0.5



Joint Sampling of $y(t)$ and $f(t)$ from Covariance

- `gpsimTest`

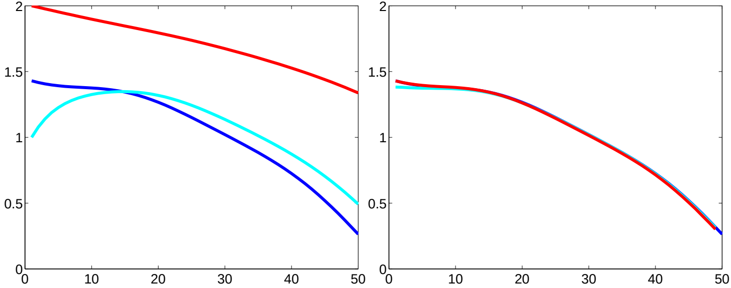


Figure: *Left*: joint samples from the transcription covariance, *blue*: $f(t)$, *cyan*: $y_1(t)$ and *red*: $y_2(t)$. *Right*: numerical solution for $f(t)$ of the differential equation from $y_1(t)$ and $y_2(t)$ (blue and cyan). True $f(t)$ included for comparison.

Joint Sampling of $y(t)$ and $f(t)$ from Covariance

- `gpsimTest`

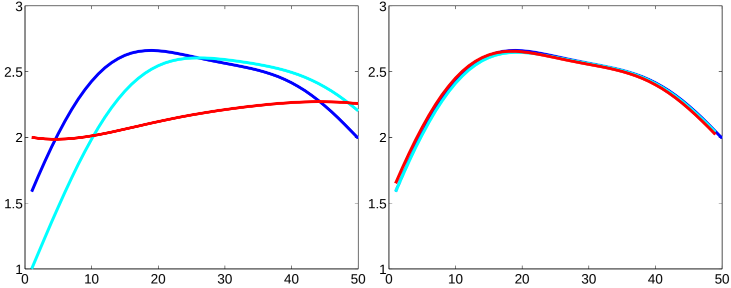


Figure: *Left*: joint samples from the transcription covariance, *blue*: $f(t)$, *cyan*: $y_1(t)$ and *red*: $y_2(t)$. *Right*: numerical solution for $f(t)$ of the differential equation from $y_1(t)$ and $y_2(t)$ (blue and cyan). True $f(t)$ included for comparison.

Joint Sampling of $y(t)$ and $f(t)$ from Covariance

- `gpsimTest`

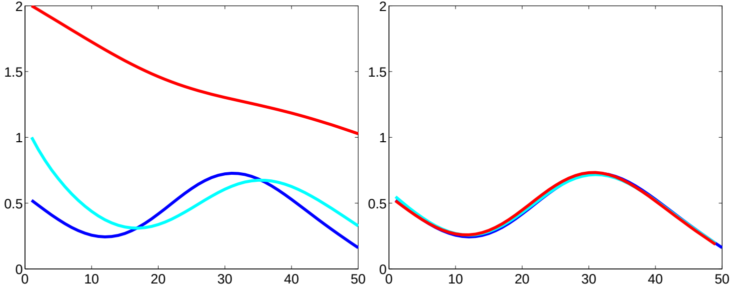


Figure: *Left*: joint samples from the transcription covariance, *blue*: $f(t)$, *cyan*: $y_1(t)$ and *red*: $y_2(t)$. *Right*: numerical solution for $f(t)$ of the differential equation from $y_1(t)$ and $y_2(t)$ (blue and cyan). True $f(t)$ included for comparison.

Results — Drosophila

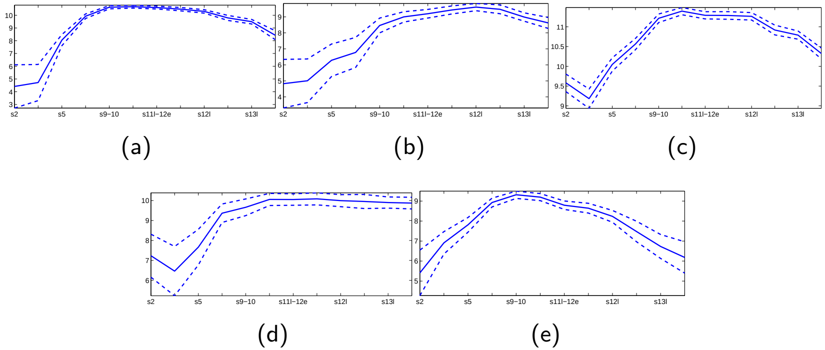
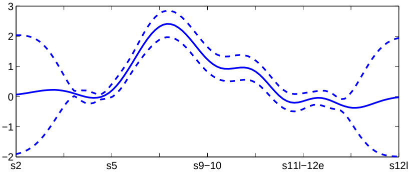
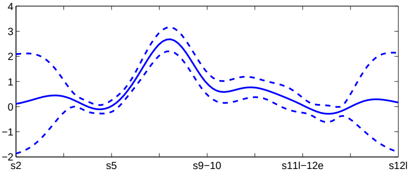


Figure: mRNA expression levels for target genes. (a) pannier, (b) hibris, (c) CG12744, (d) CG10516 (e) CG31368 .

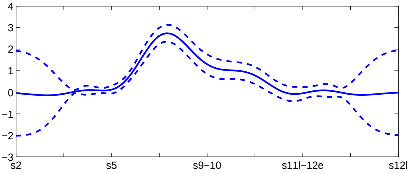
Results — Drosophila



(a)



(b)



(c)

Figure: Inferred Transcription Factor Activities

Results — Transcription Rates

- Estimation of Equation Parameters `demBarenco1`

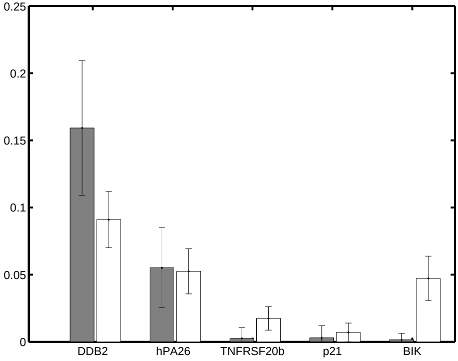


Figure: Basal transcription rates. Our results (black) compared with Barenco et al. [2006] (white).

Results — Transcription Rates

- Estimation of Equation Parameters `demBarenco1`

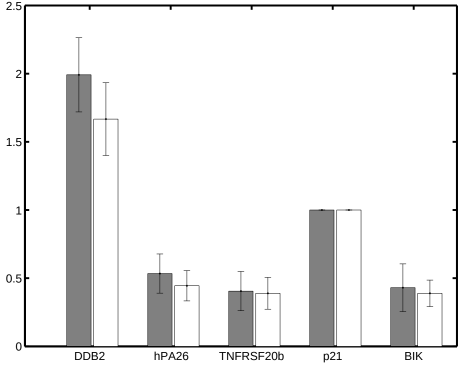


Figure: Sensitivities. Our results (black) compared with Barenco et al. [2006] (white).

Results — Transcription Rates

- Estimation of Equation Parameters `demBarenco1`

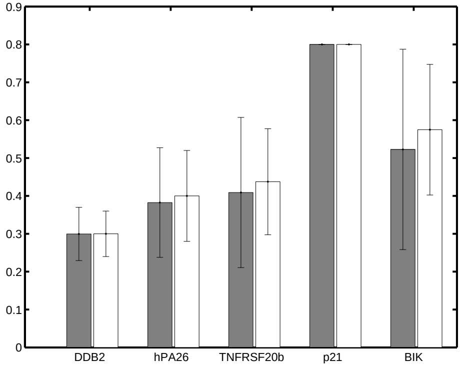
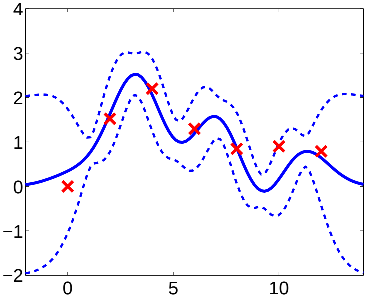


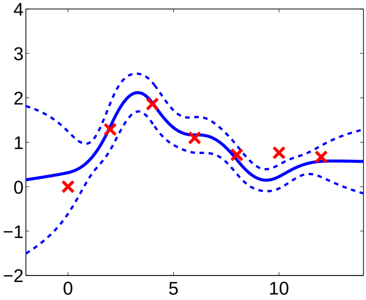
Figure: Decays. Our results (black) compared with Barenco et al. [2006] (white).

Results — Protein Concentration

- Prediction with error bars of protein concentration: $p(\mathbf{f}|\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5)$



(a)



(b)

Figure: (a) RBF covariance function (b) MLP covariance function. Also included are results from Barenco et al. [2006] as crosses.

Results — Log Space

- GP predictions in log space.

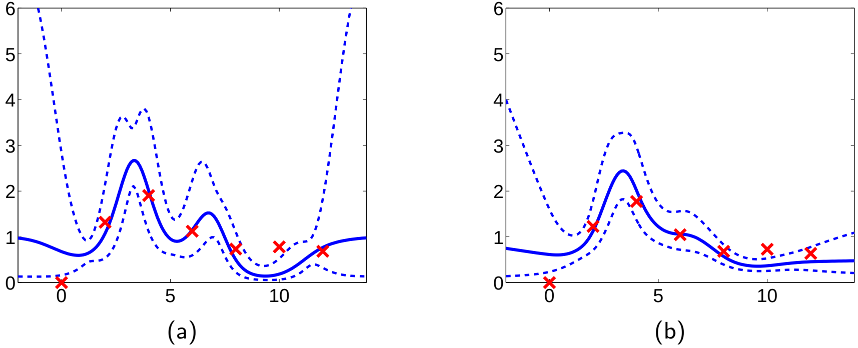


Figure: (a) RBF covariance function (b) MLP covariance function. Also included are results from Barenco et al. [2006] as crosses.

Results — $\log(1 + \exp(x))$ Constrained

- GP predictions in log space.

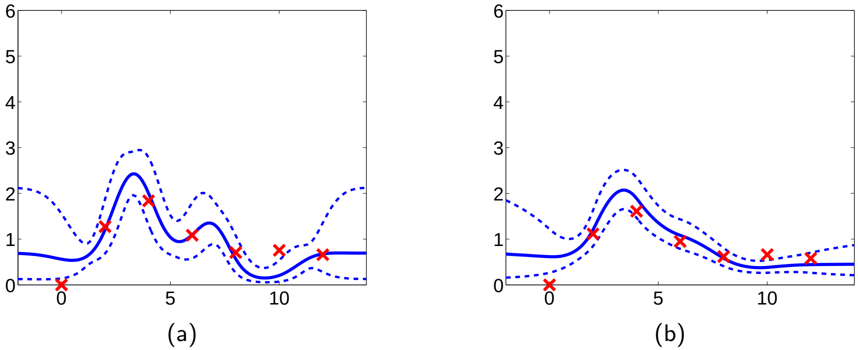


Figure: (a) RBF covariance function (b) MLP covariance function. Also included are results from Barenco et al. [2006] as crosses.

Results — Sigmoid

- GP predictions in log space.

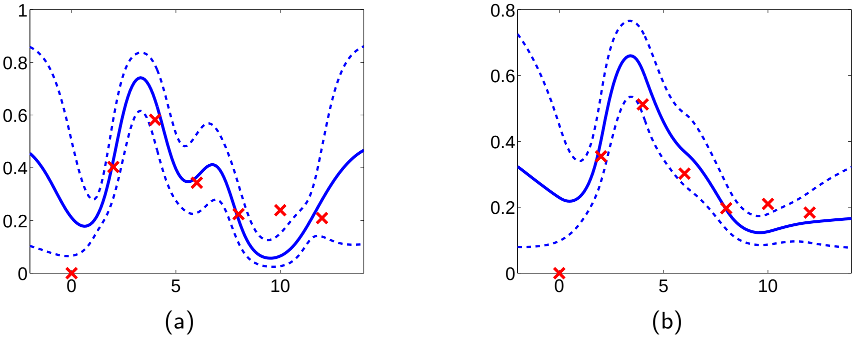


Figure: (a) RBF covariance function (b) MLP covariance function. Also included are results from Barenco et al. [2006] as crosses.

Transcription Model Summary

- Progress so far and Future work
  - Elegant solution of a problem with indirect observations.
  - Already extended to non-linear response equations (using Laplace approximation).
  - Extending to systems with *multiple transcription factors* (Pei Gao).
  - Validating with Markov chain Monte-Carlo (Michalis Titsias).
  - Sensitivities which change over time (Antti Honkela)

References

M. Barenco, D. Tomescu, D. Brewer, R. Callard, J. Stark, and M. Hubank. Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biology*, 7(3):R25, 2006.

A. O'Hagan. Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society, B*, 40:1–42, 1978.

A. O'Hagan. Some Bayesian numerical analysis. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 345–363, Valencia, 1992. Oxford University Press.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006. ISBN 026218253X.