Modelling gene expression dynamics with Gaussian processes

Regulatory Genomics and Epigenomics
March 10^{th} 2016

Magnus Rattray
Faculty of Life Sciences
University of Manchester

Talk Outline

Introduction to Gaussian process regression

Example 1. Hierarchical models: batches and clusters

Example 2. Branching models: perturbations and bifurcations

Example 3. Differential equations: Pol-II to mRNA dynamics

Gaussian processes: flexible non-parametric models

Probability distributions over functions

$$f(t) \sim \mathcal{GP}\left(\text{mean}(t), \text{cov}(t, t')\right)$$

Covariance function $k = \text{cov}(t, t')$ defines typical properties,

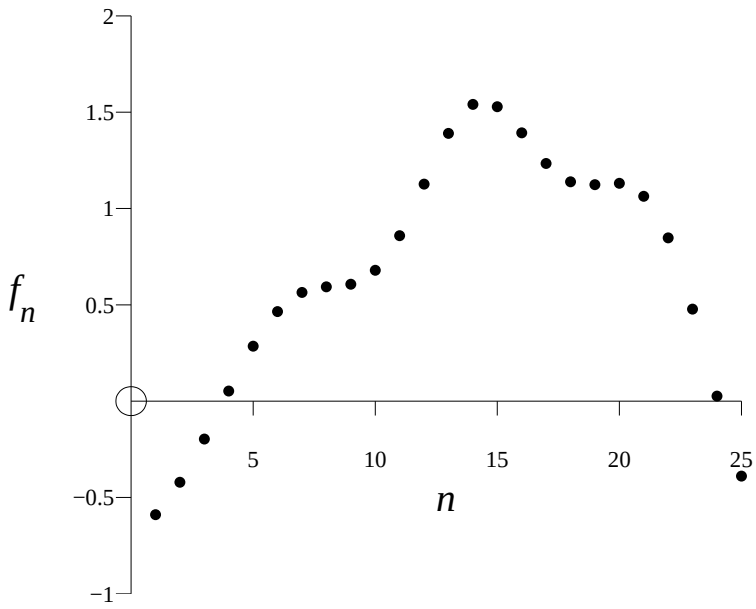
- Static . . . Dynamic
- Smooth . . . Rough
- Stationary. . . non-Stationary
- Periodic. . . Chaotic

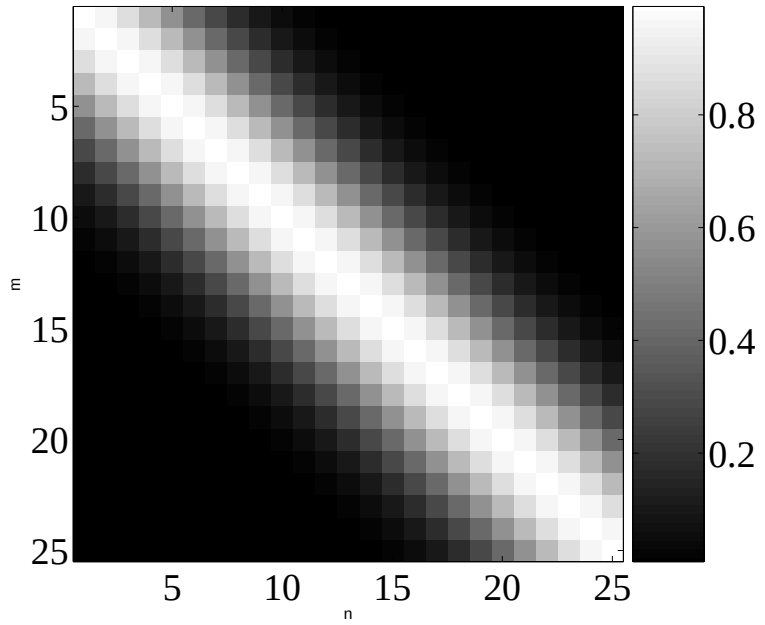
The covariance function has parameters tuning these properties

Bayesian Machine Learning perspective: Rasmussen & Williams "Gaussian Processes for Machine Learning" (MIT Press, 2006)

Gaussian processes

Samples from a 25-dimensional multivariate Gaussian distribution:



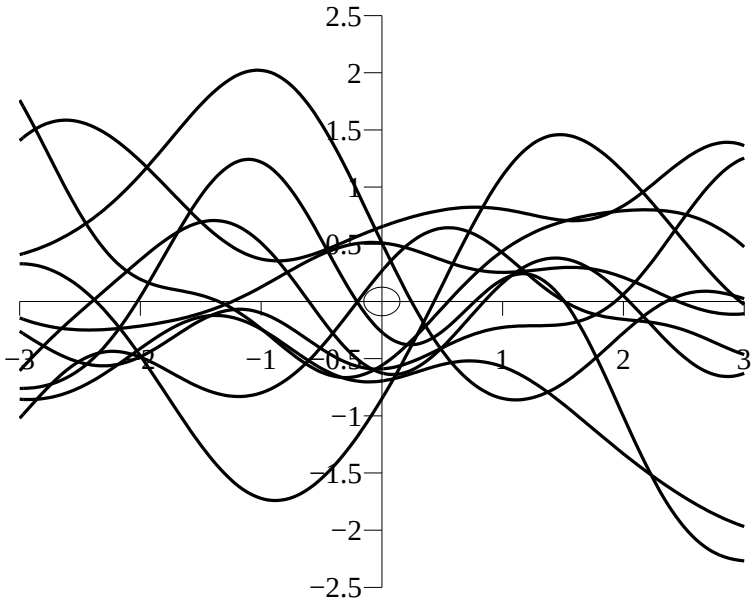


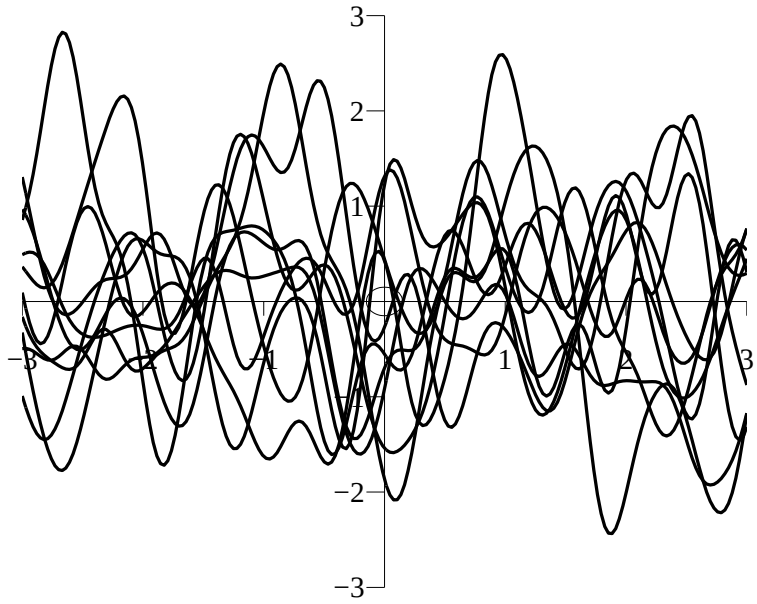
$$[f_1, f_2, \ldots, f_{25}] \sim \mathcal{N}(0, C)$$

Learning and Inference in Computational Systems Biology, MIT Press

Gaussian processes

Take dimension $\to \infty$

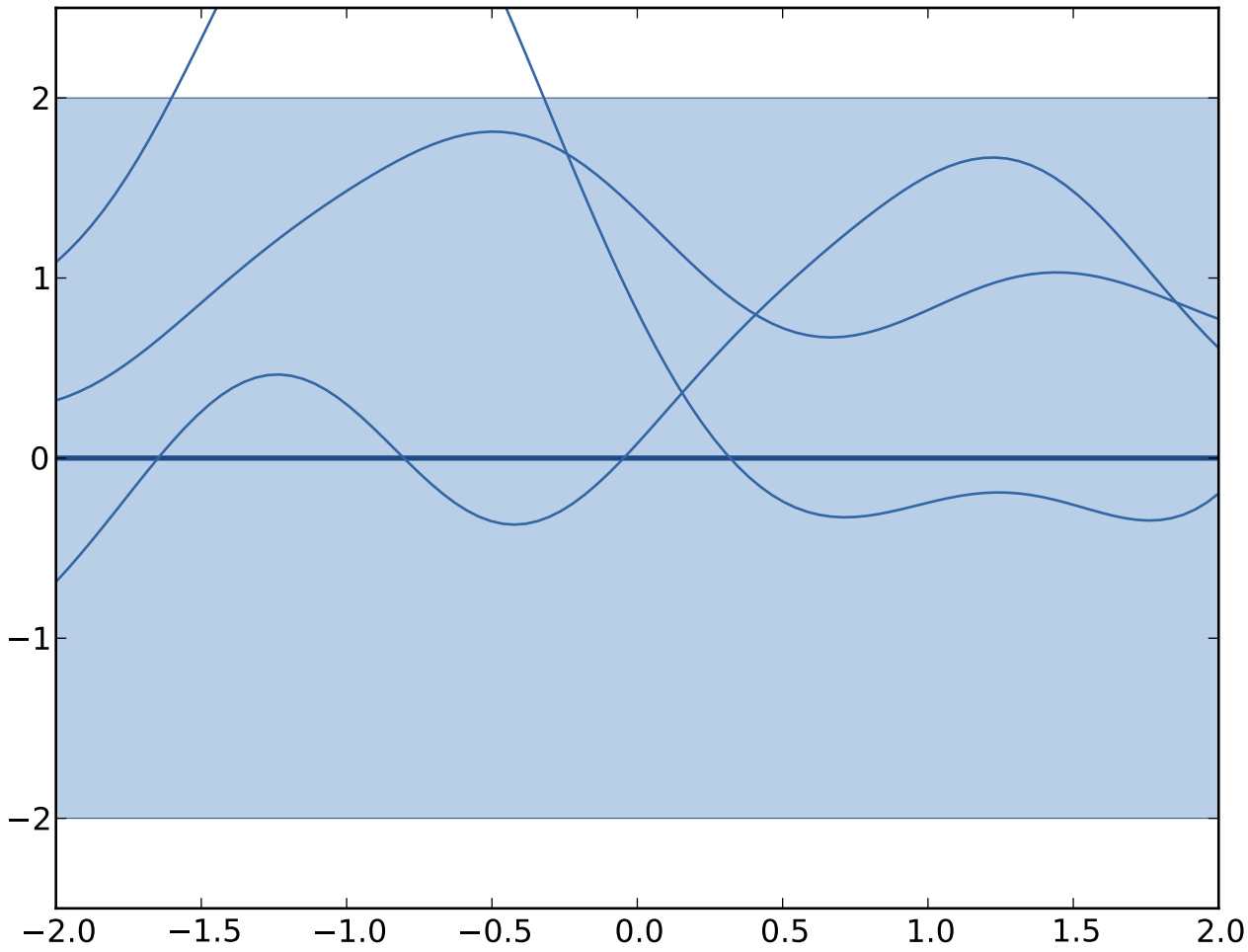




$$f \sim \mathcal{GP}(0, k) \qquad k\left(t, t'\right) = \exp\left(-\frac{(t - t')^2}{l^2}\right)$$

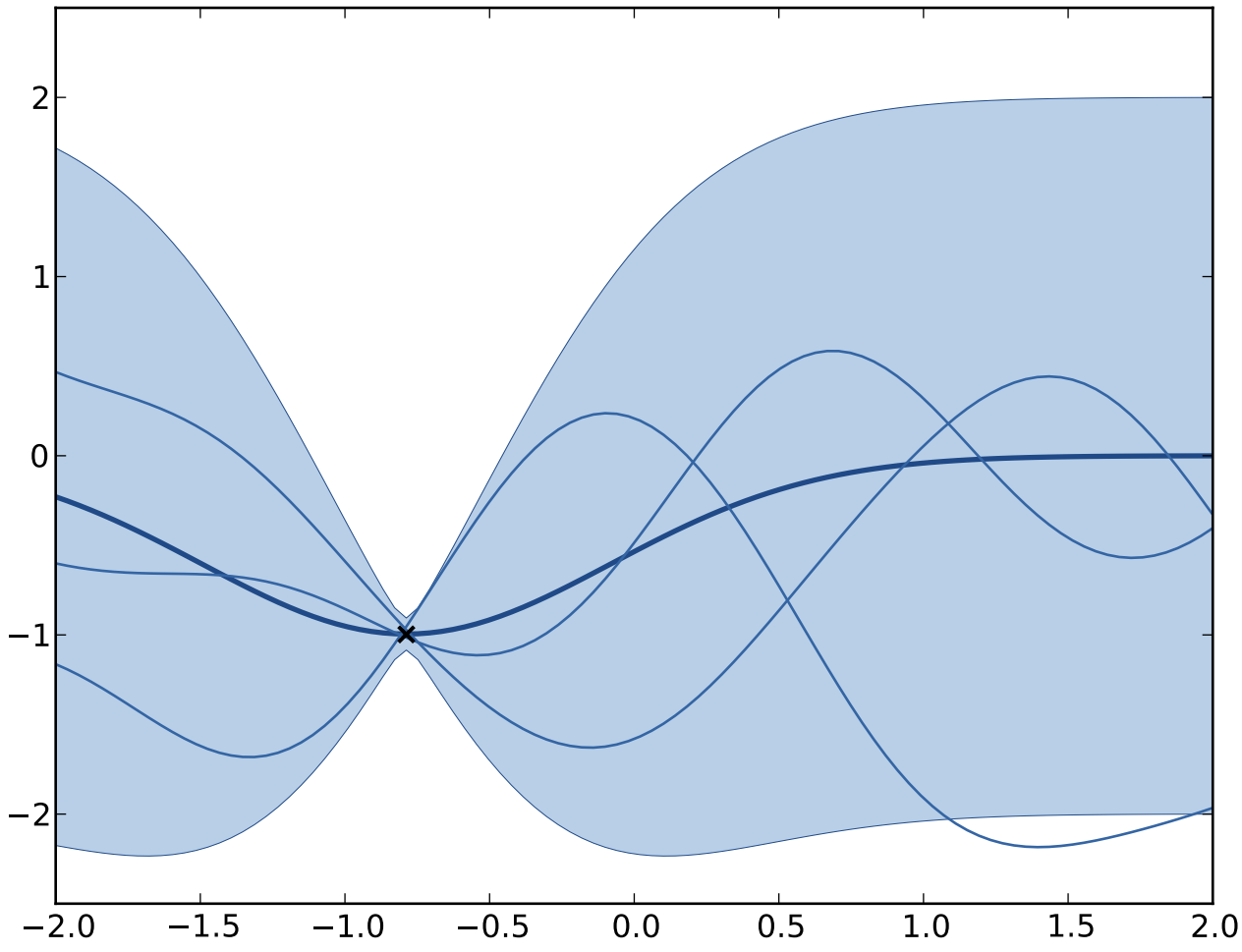
Learning and Inference in Computational Systems Biology, MIT Press

Gaussian processes for inference: Bayesian Regression



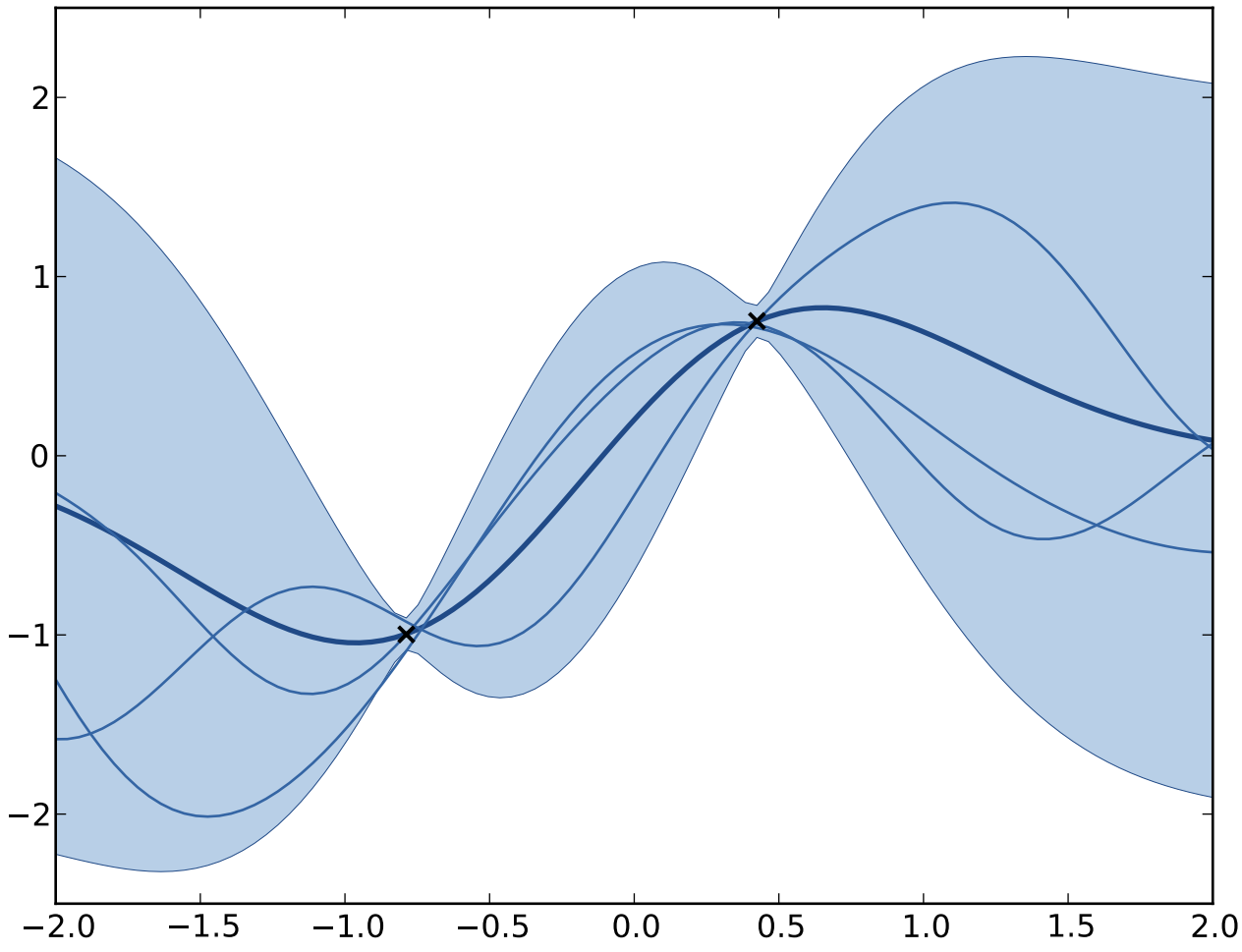
Figures from James Hensman

Regression example



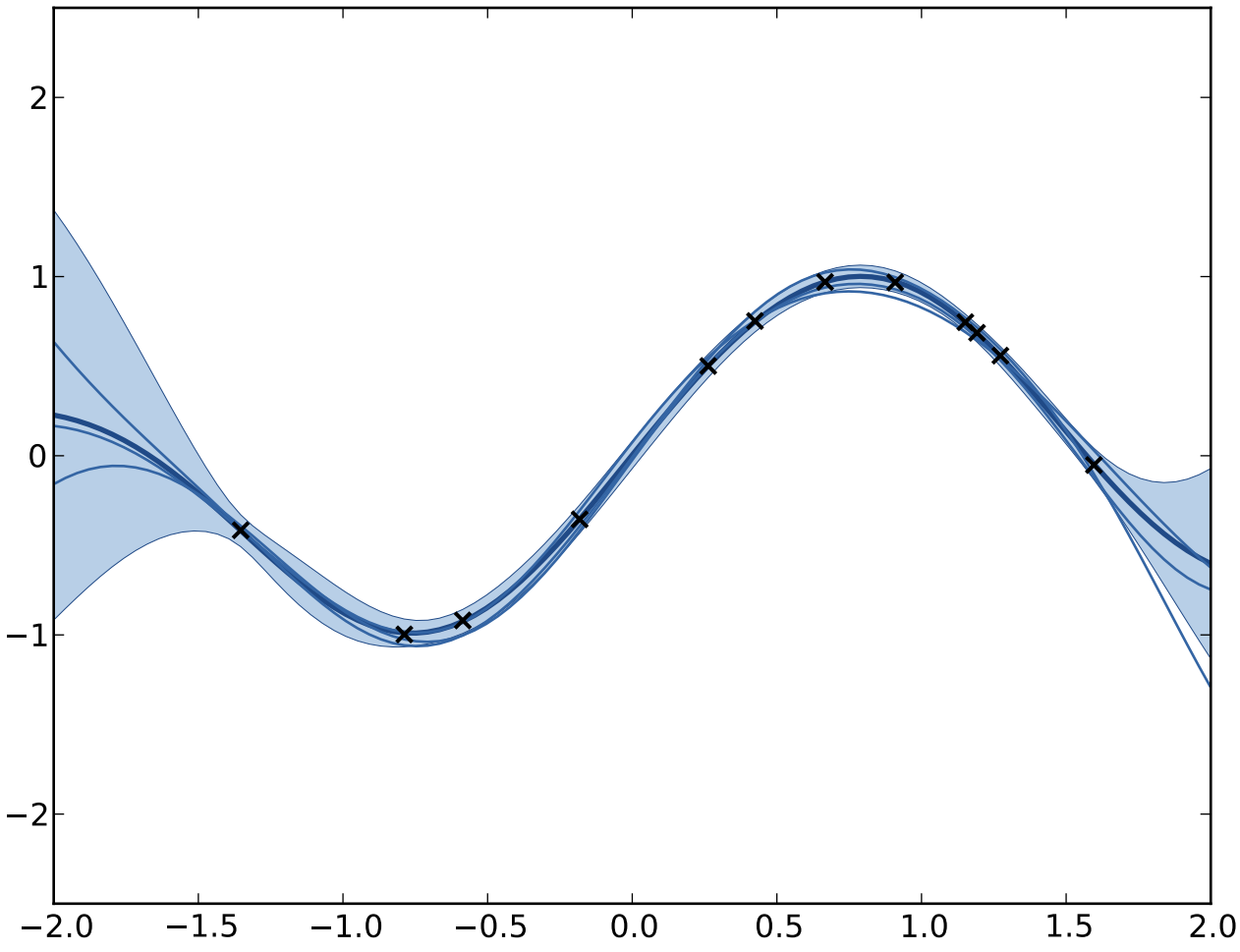
Figures from James Hensman

Regression example



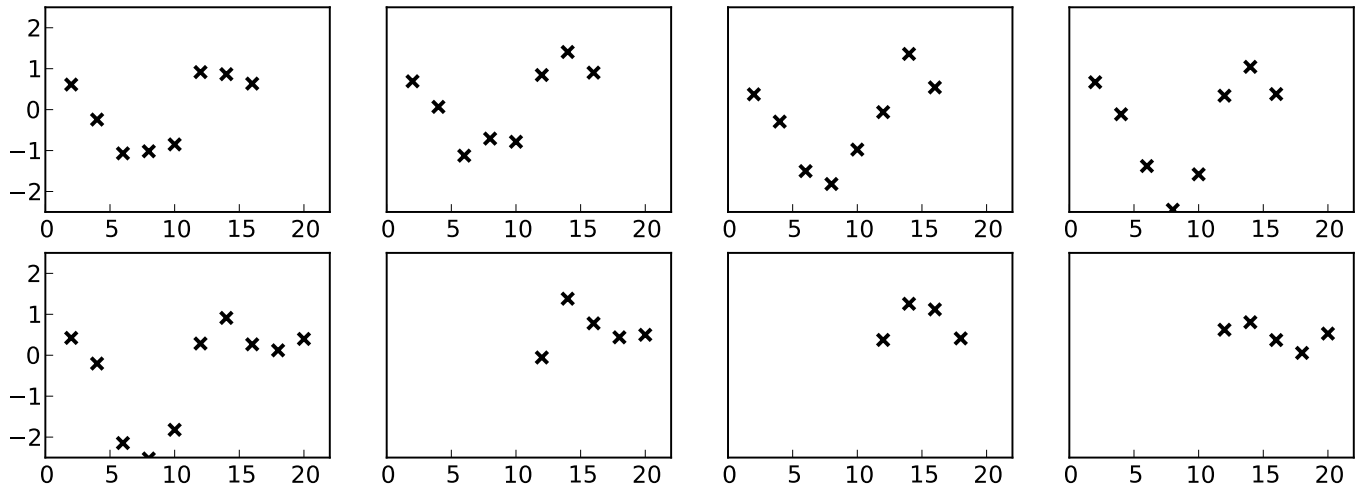
Figures from James Hensman

Regression example



Figures from James Hensman

Ex1. Hierarchical models: batches and clusters

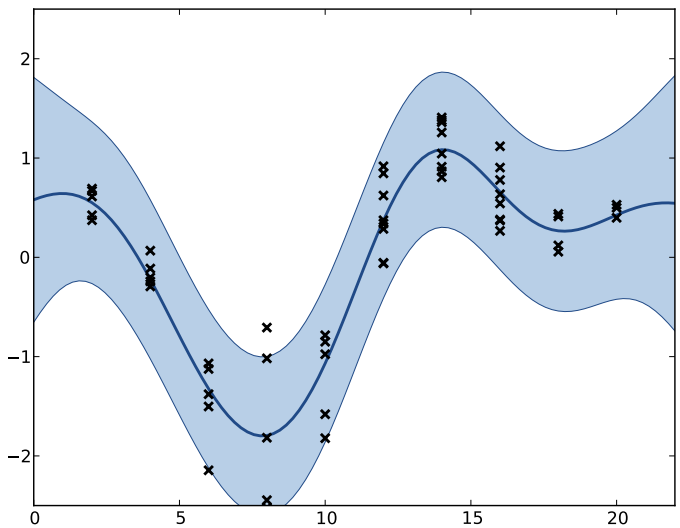


Data from Kalinka et al. "Gene expression divergence recapitulates the developmental hourglass model" *Nature* 2010

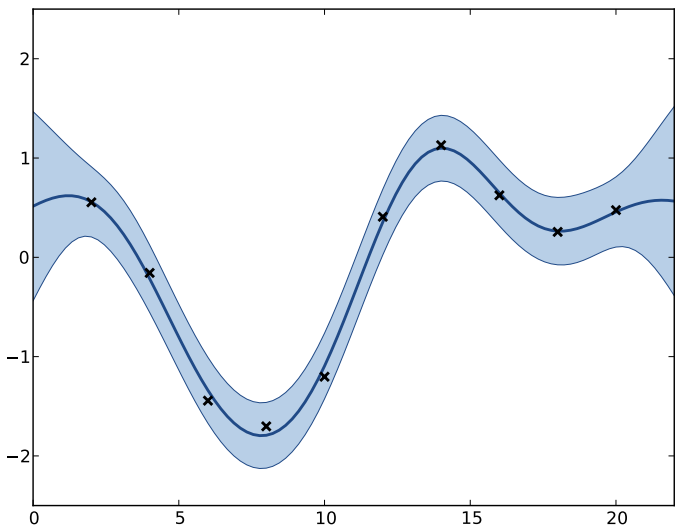
Joint work with James Hensman and Neil Lawrence

Usual processing options for time course batches

Lumped



Averaged



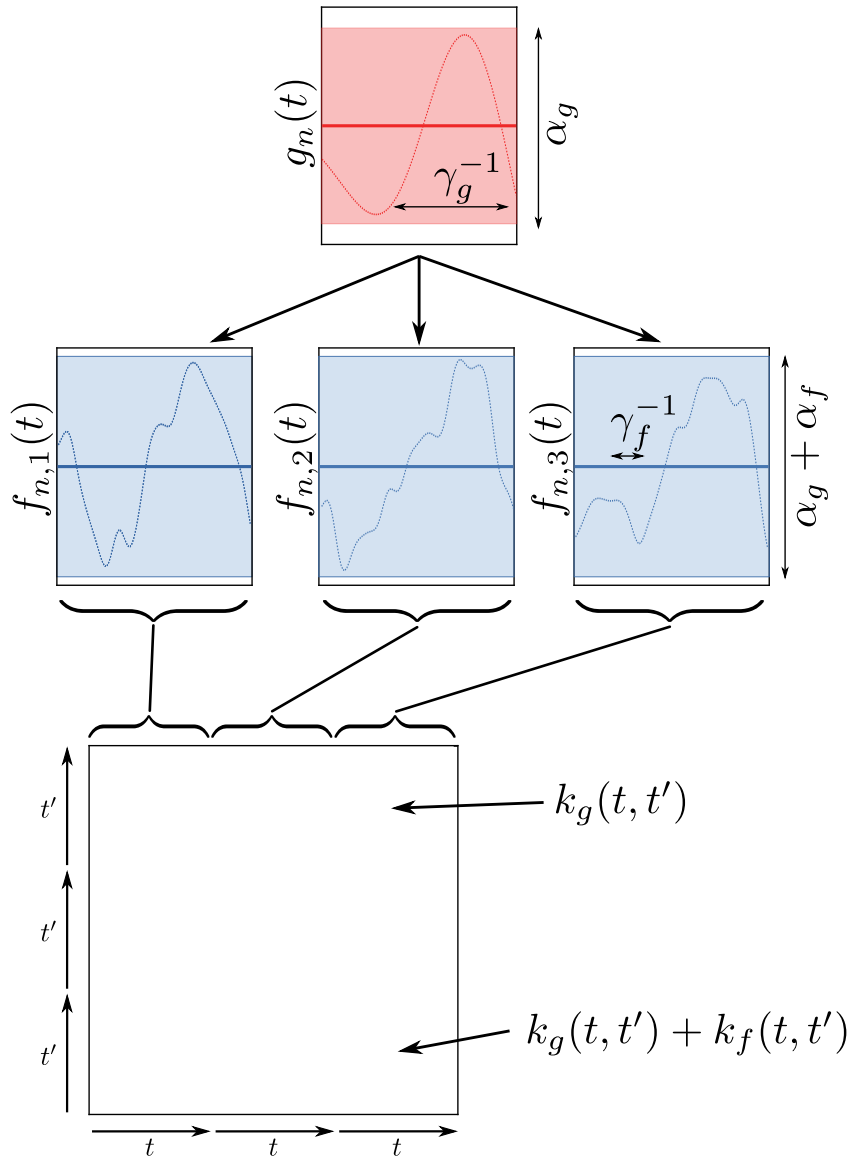
Hierarchical Gaussian process

gene:

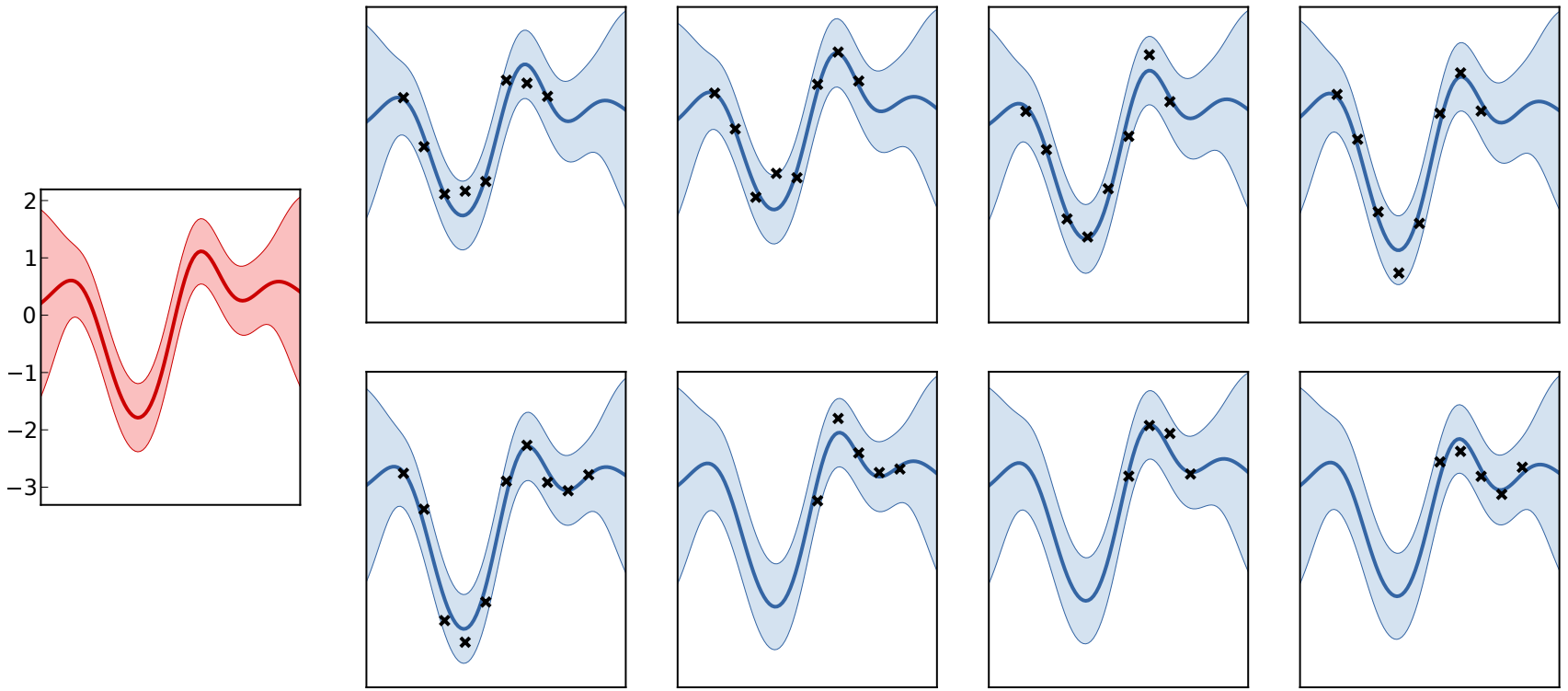
$$g(t) \sim \mathcal{GP}\left(0, k_g(t, t')\right)$$

replicate:

$$f_i(t) \sim \mathcal{GP}\left(g(t), k_f(t, t')\right)$$



Hierarchical Gaussian process



J. Hensman, N.D. Lawrence, M.Rattray "Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters" *BMC Bioinformatics* 2013

Hierarchical Gaussian process for clustering

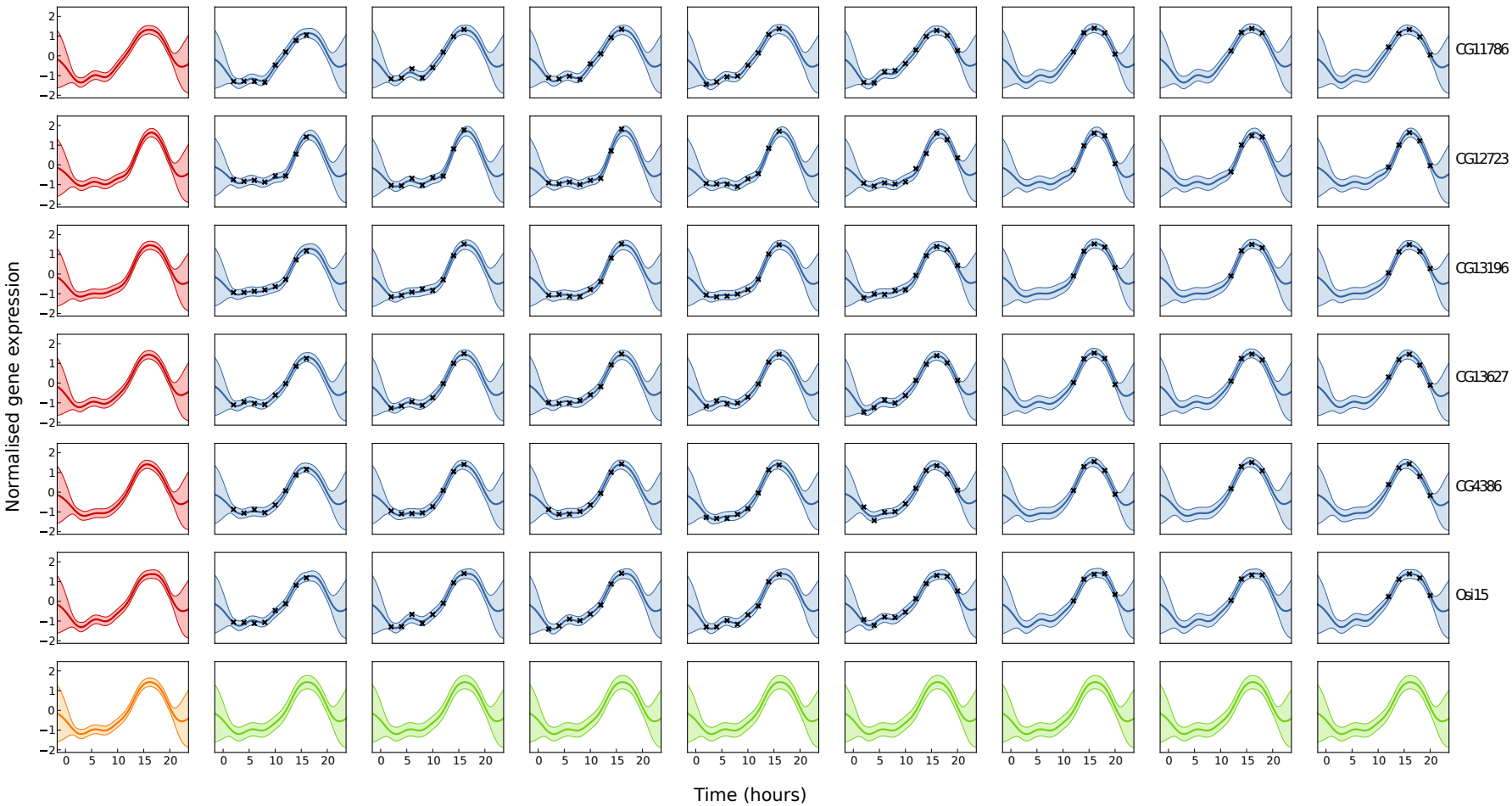
An extended hierarchy

$$h(t) \sim \mathcal{GP}\Big(0, k_h(t, t')\Big) \text{ cluster}$$

$$g_i(t) \sim \mathcal{GP}\Big(h(t), k_g(t, t')\Big) \text{ gene}$$

$$f_{ir}(t) \sim \mathcal{GP}\Big(g_i(t), k_f(t, t')\Big) \text{ replicate}$$

Hierarchical Gaussian process for clustering



Hierarchical Gaussian process for clustering

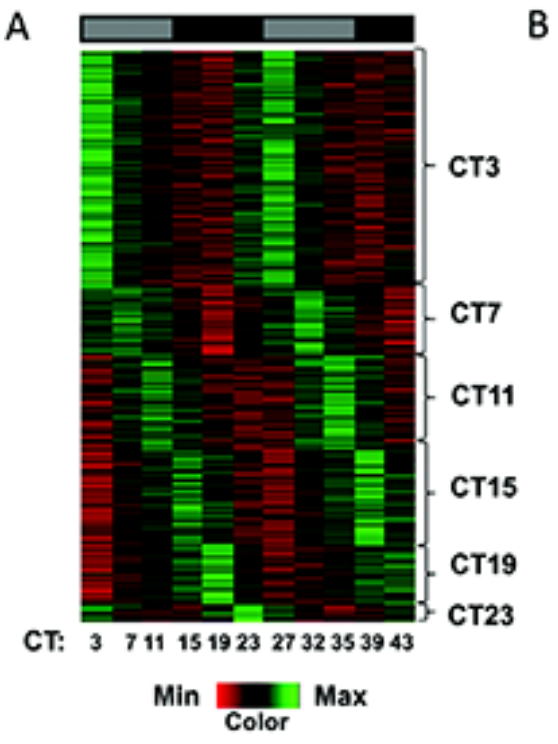
Modifying an existing algorithm to include this model of replicate and cluster structure leads to more meaningful clustering

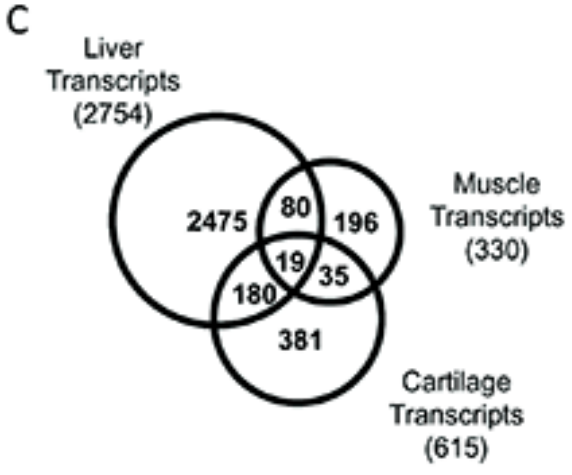
	MF	BP	CC	$\mathcal{L}$	N. clust.
agglomerative HGP	**0.46**	**0.16**	**0.50**	**7360.8**	50
agglomerative GP	0.39	0.13	0.36	6203.7	128
Mclust (concat.)	0.39	0.07	0.25	1324.0	26
Mclust (averaged)	0.40	0.08	0.24	-736.2	20

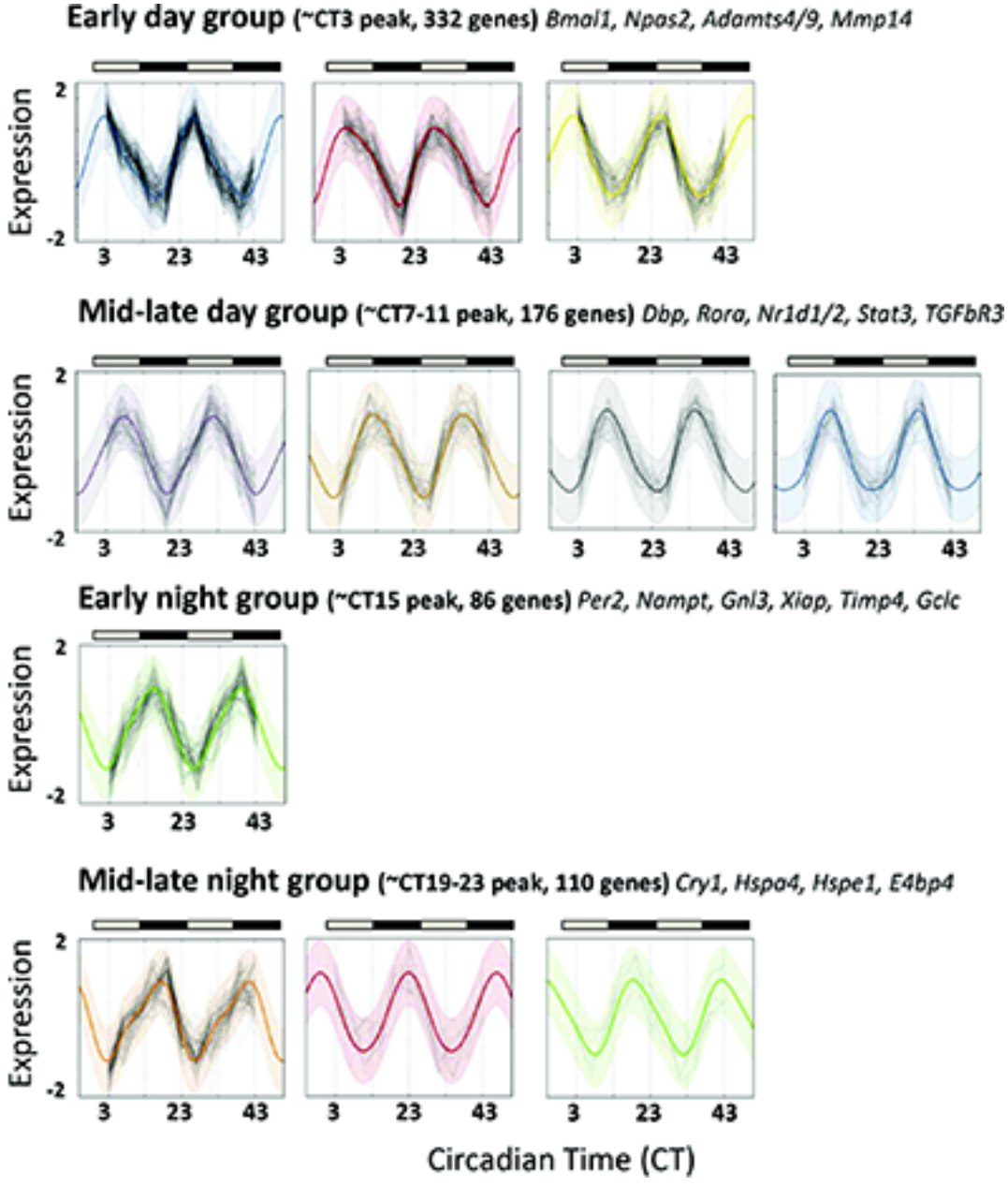
Variational Bayes algorithm is more efficient, allowing Bayesian clustering of $>$10K profiles with a Dirichlet Process prior

J. Hensman, M.Rattray, N.D. Lawrence "Fast non-parametric clustering of time-series data" *IEEE TPAMI* 2015

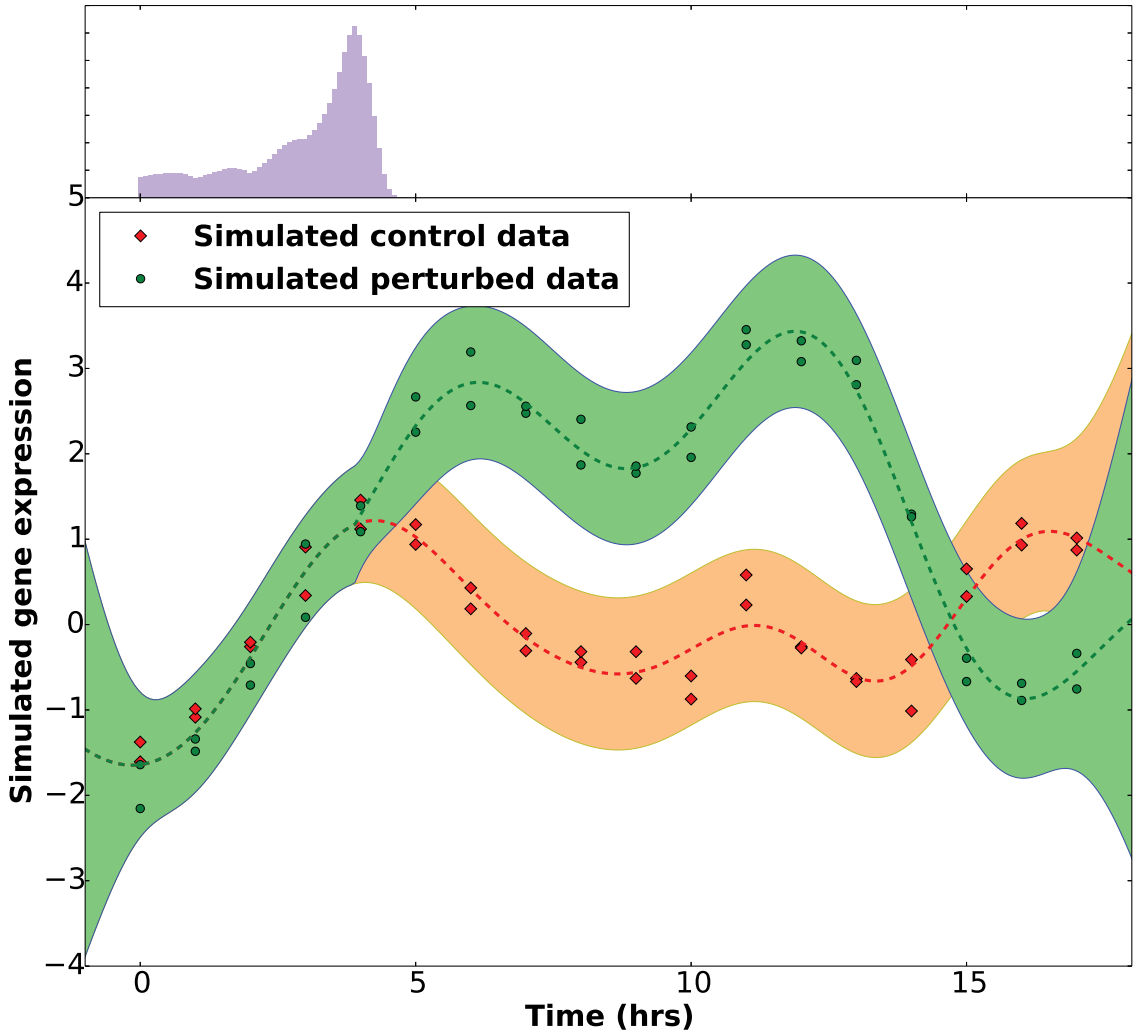
Clustering with a periodic covariance





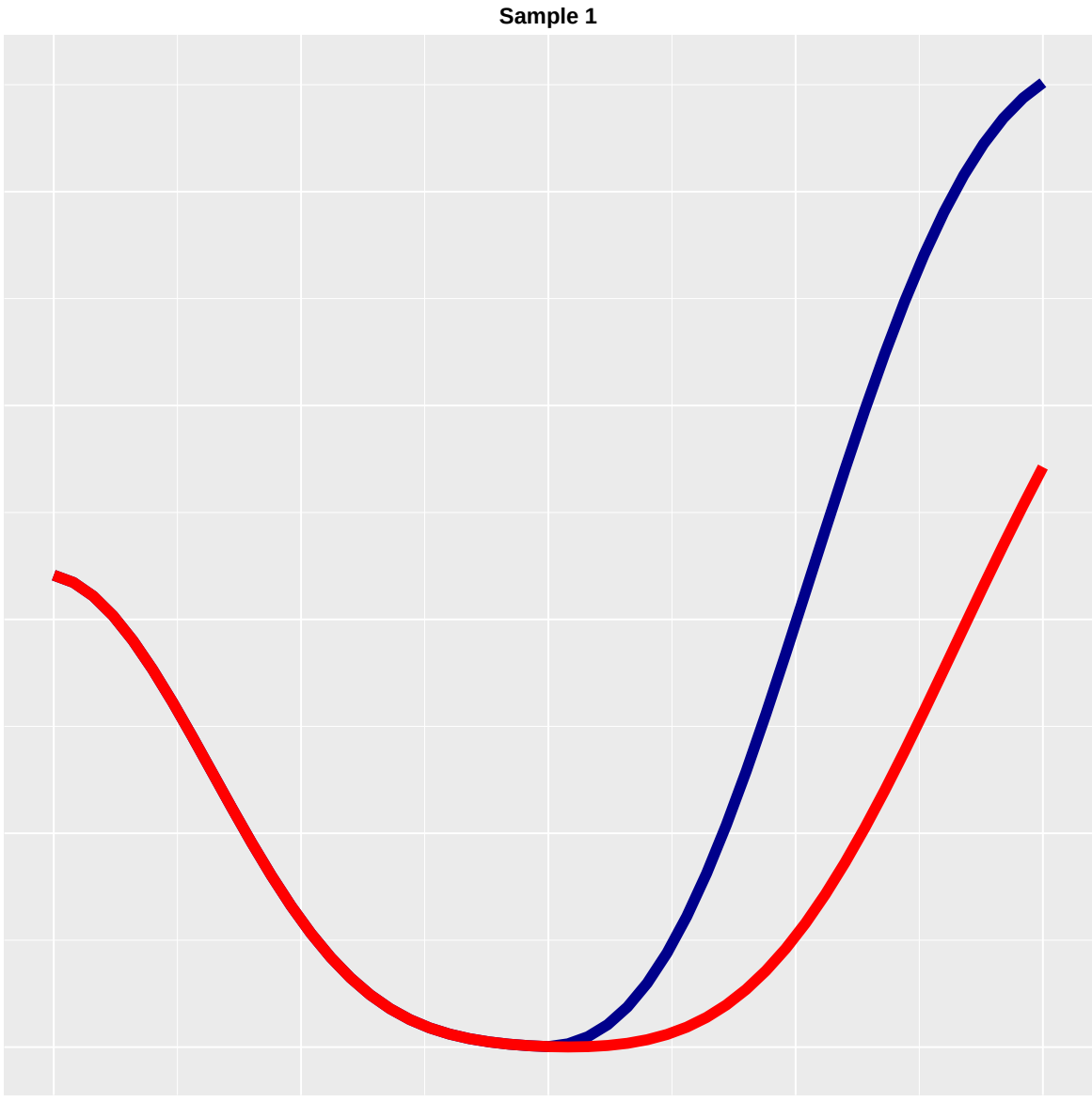


Ex2. Branching models: perturbations and bifurcations

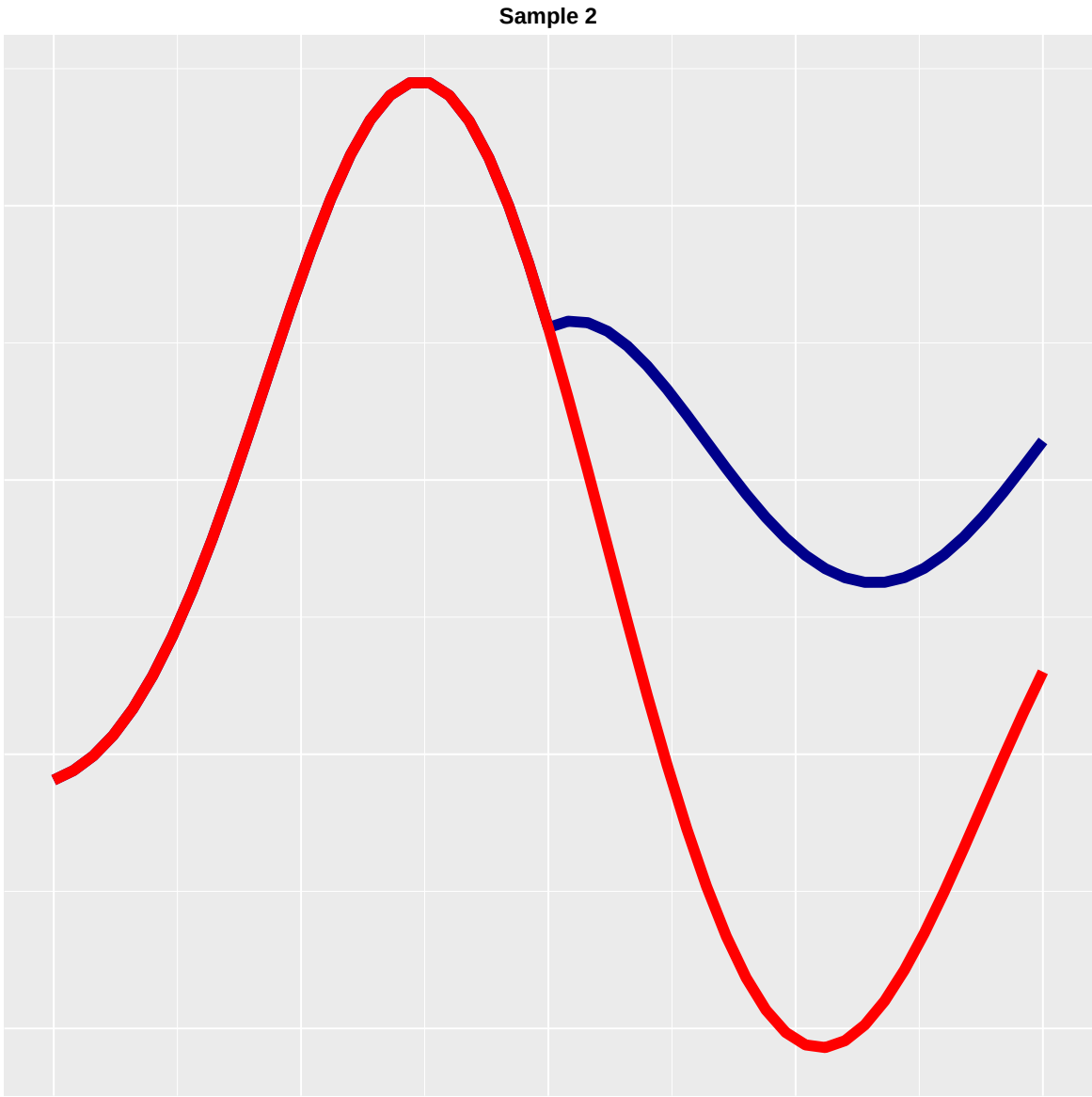


Joint work with Jing Yang, Chris Penfold and Murray Grant

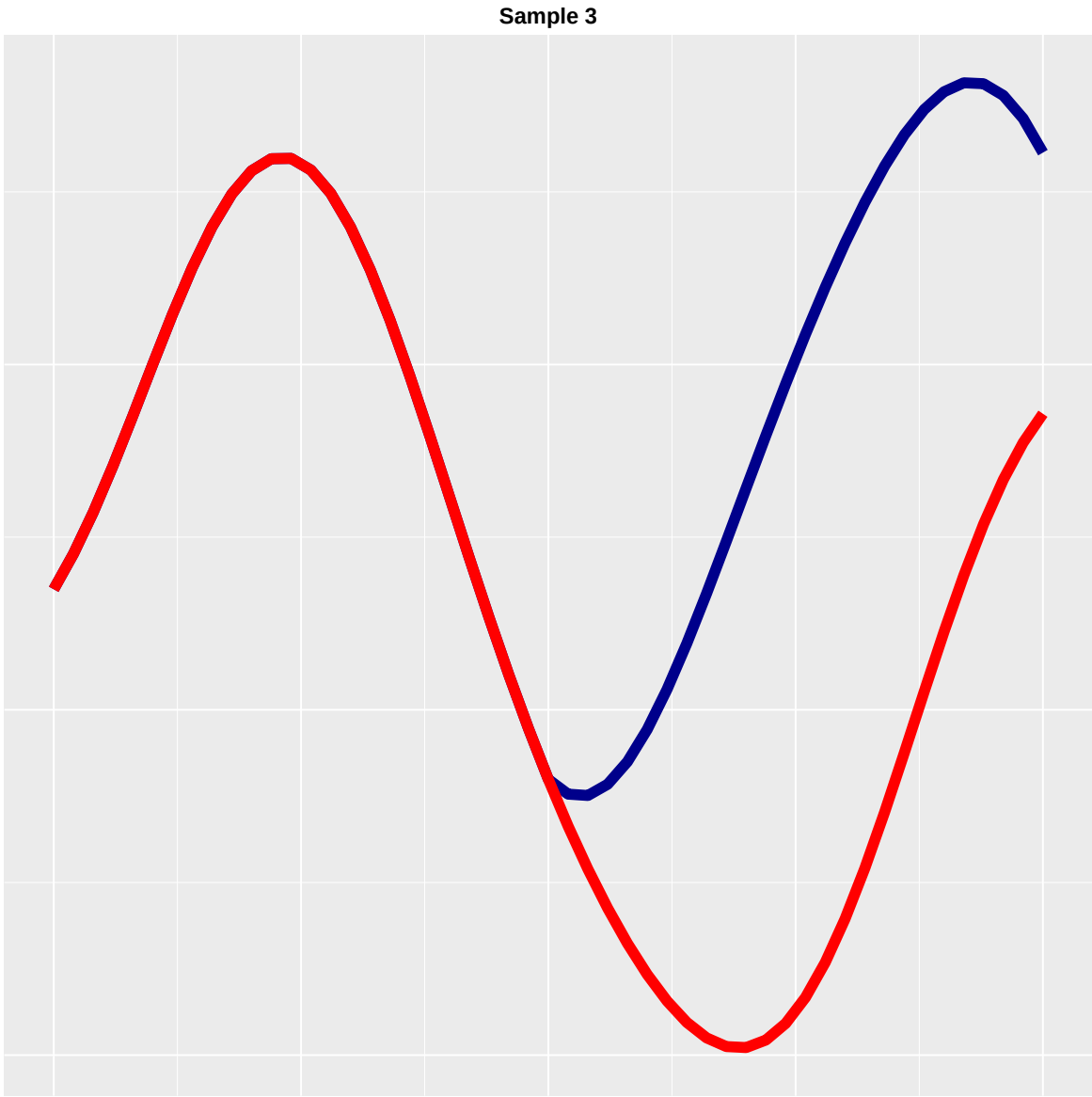
Samples from a branching model



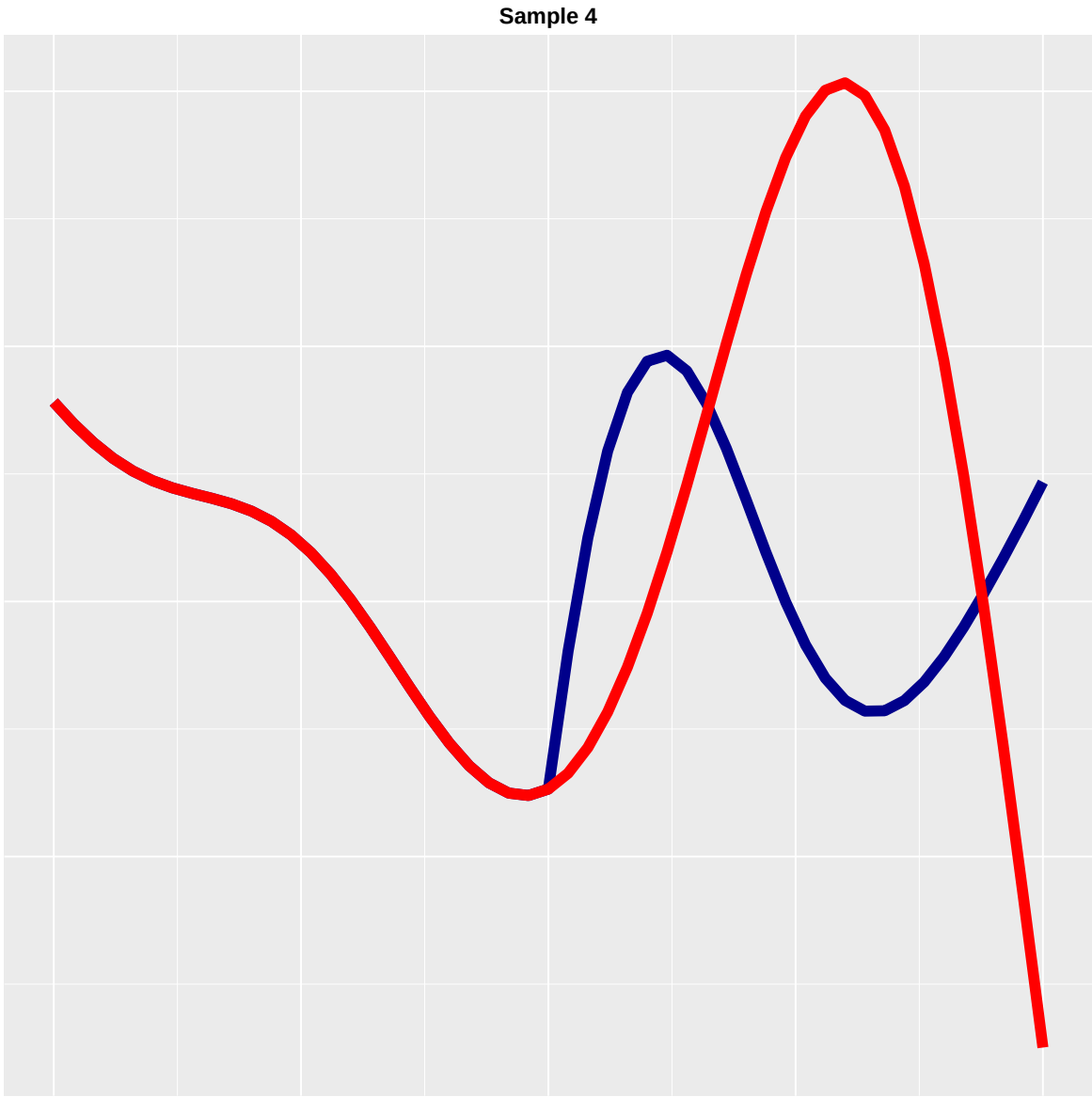
Samples from a branching model



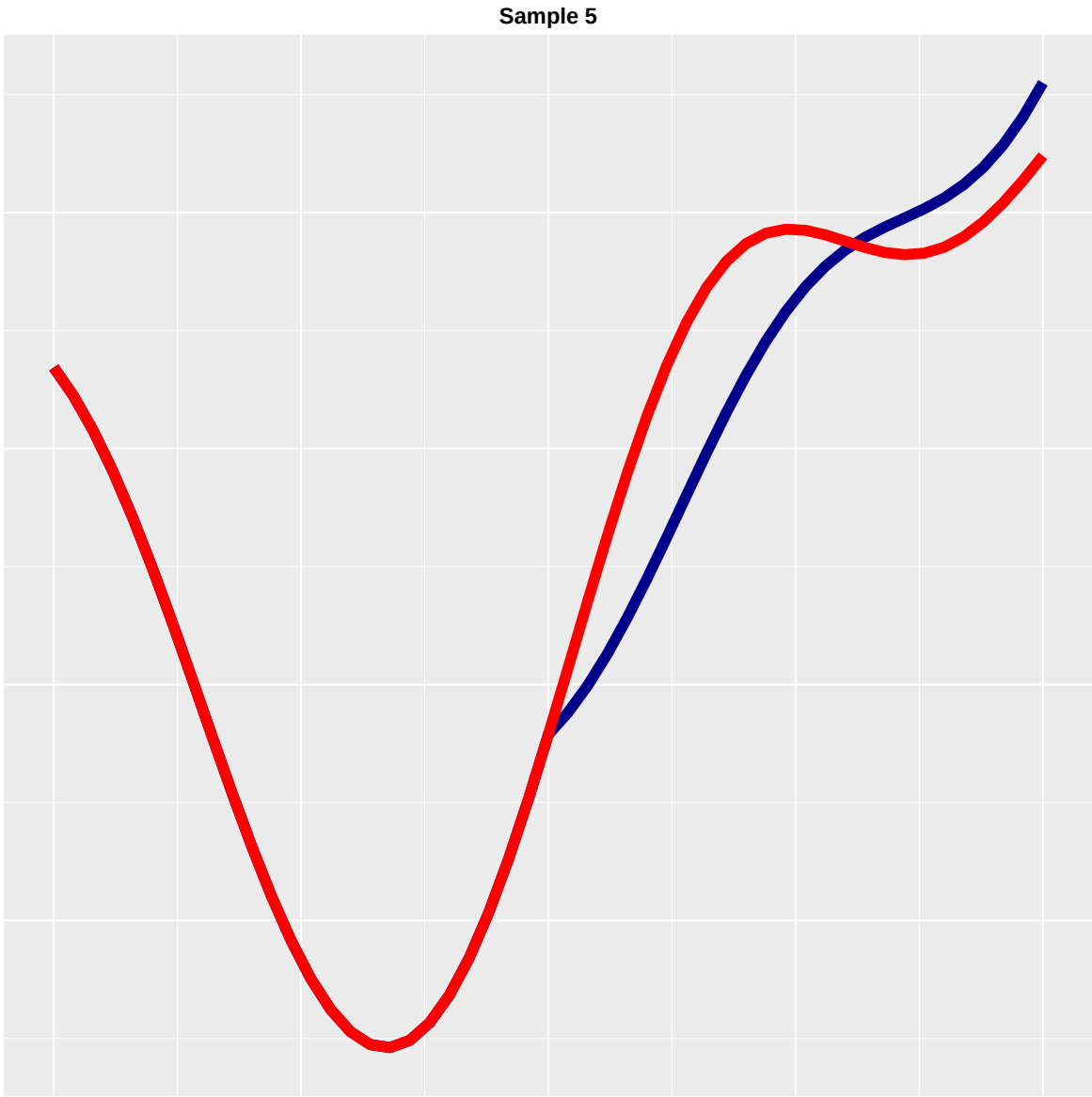
Samples from a branching model



Samples from a branching model

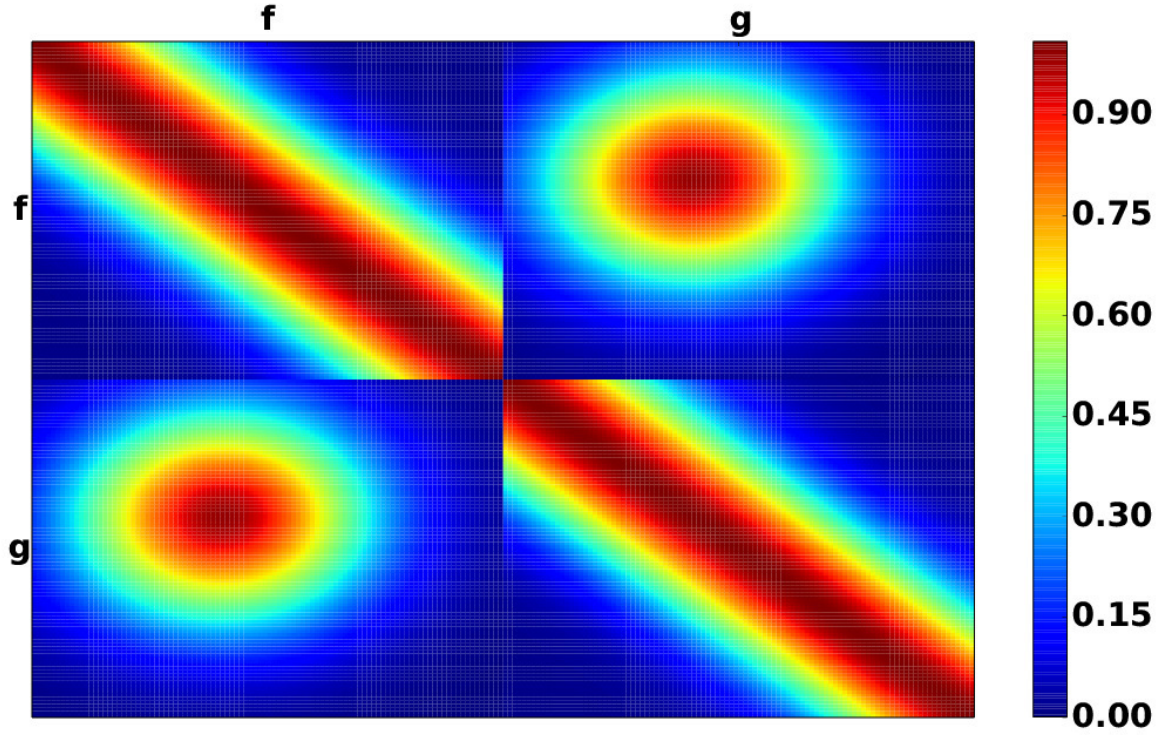


Samples from a branching model



Joint distribution to two functions crossing at t_p

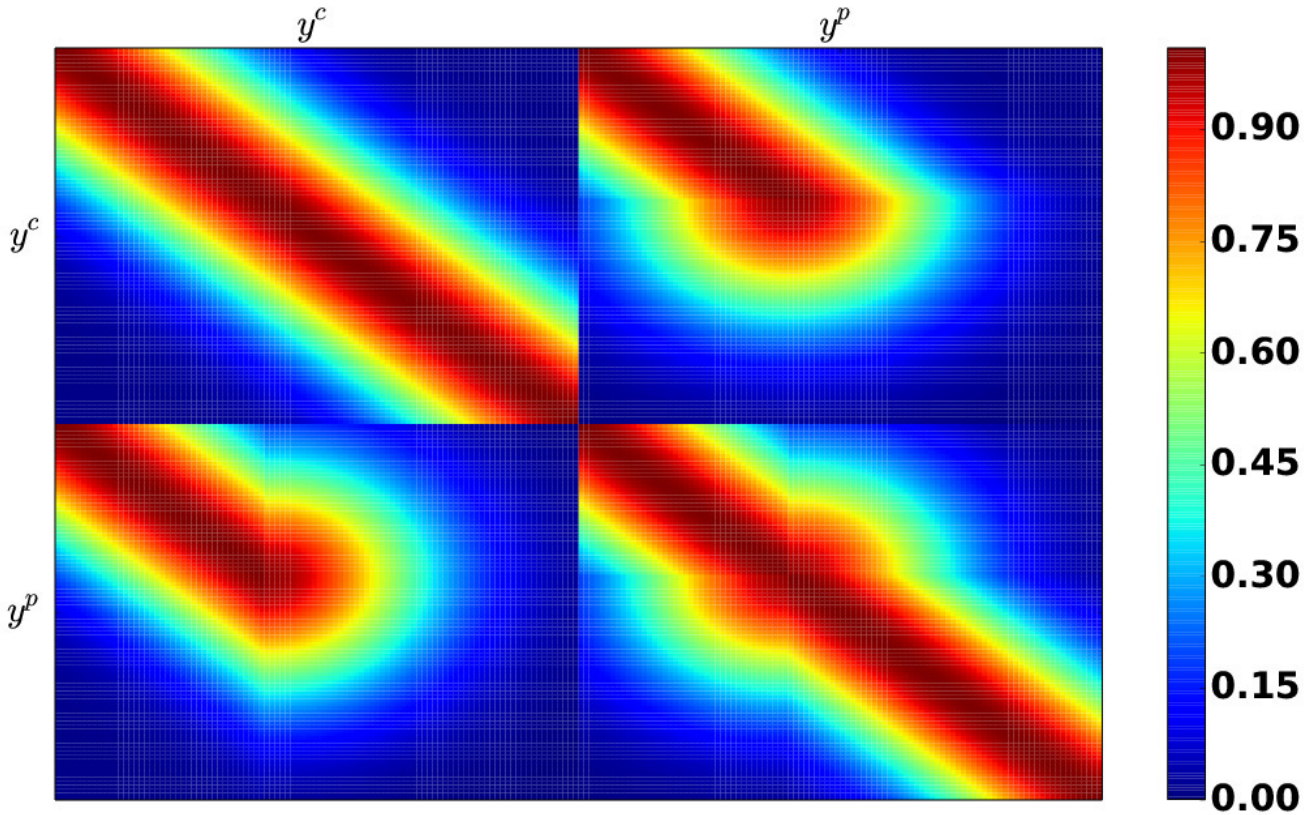
$$f \sim \mathcal{GP}(0, K)\ , \quad g \sim \mathcal{GP}(0, K)\ , \quad g(t_p) = f(t_p)$$



$$\Sigma = \begin{pmatrix} K_{ff} & K_{fg} \\ K_{gf} & K_{gg} \end{pmatrix} = \begin{pmatrix} K(\mathbf{T}, \mathbf{T}) & \frac{K(\mathbf{T}, t_p) K(t_p, \mathbf{T})}{k(t_p, t_p)} \\ \frac{K(\mathbf{T}, t_p) K(t_p, \mathbf{T})}{k(t_p, t_p)} & K(\mathbf{T}, \mathbf{T}) \end{pmatrix} \qquad (1)$$

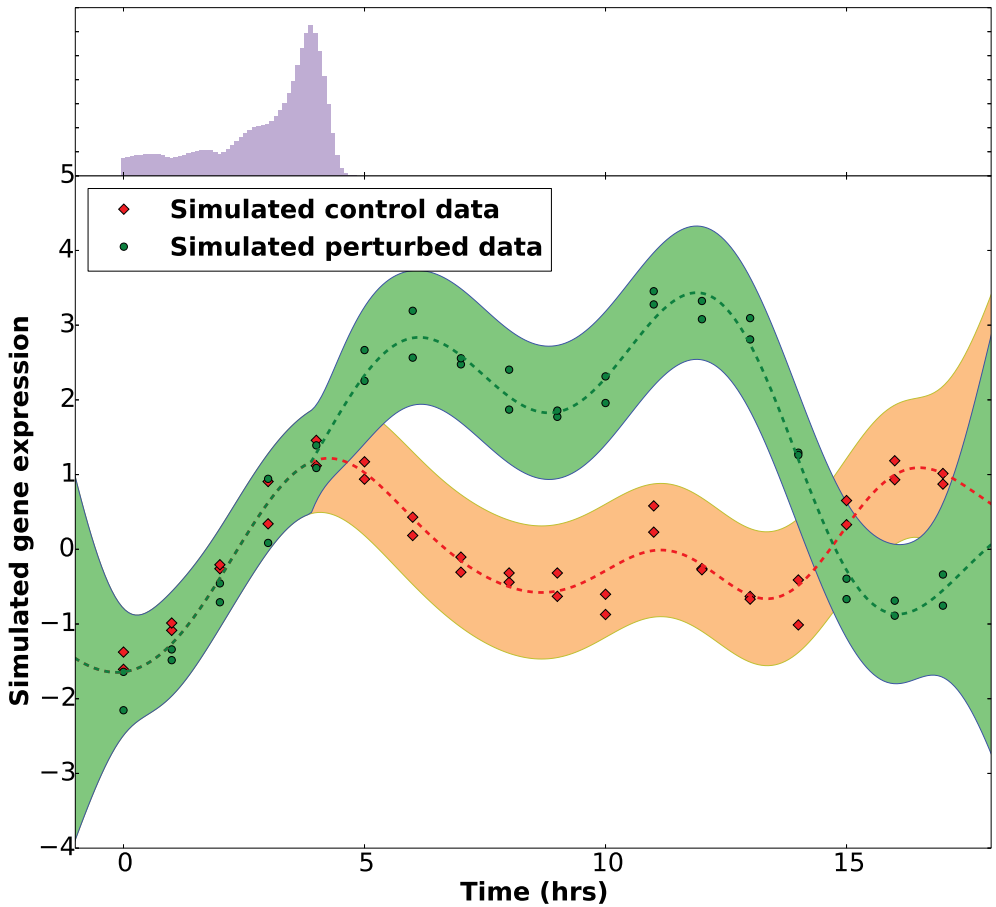
Joint distribution of two datasets diverging at t_p

$$
\begin{aligned}
y^c(t_n) &\sim \mathcal{N}(f(t_n), \sigma^2) \\
y^p(t_n) &\sim \mathcal{N}(f(t_n), \sigma^2) \quad \text{for } t_n \leq t_p \\
y^p(t_n) &\sim \mathcal{N}(g(t_n), \sigma^2) \quad \text{for } t_n > t_p
\end{aligned}
$$



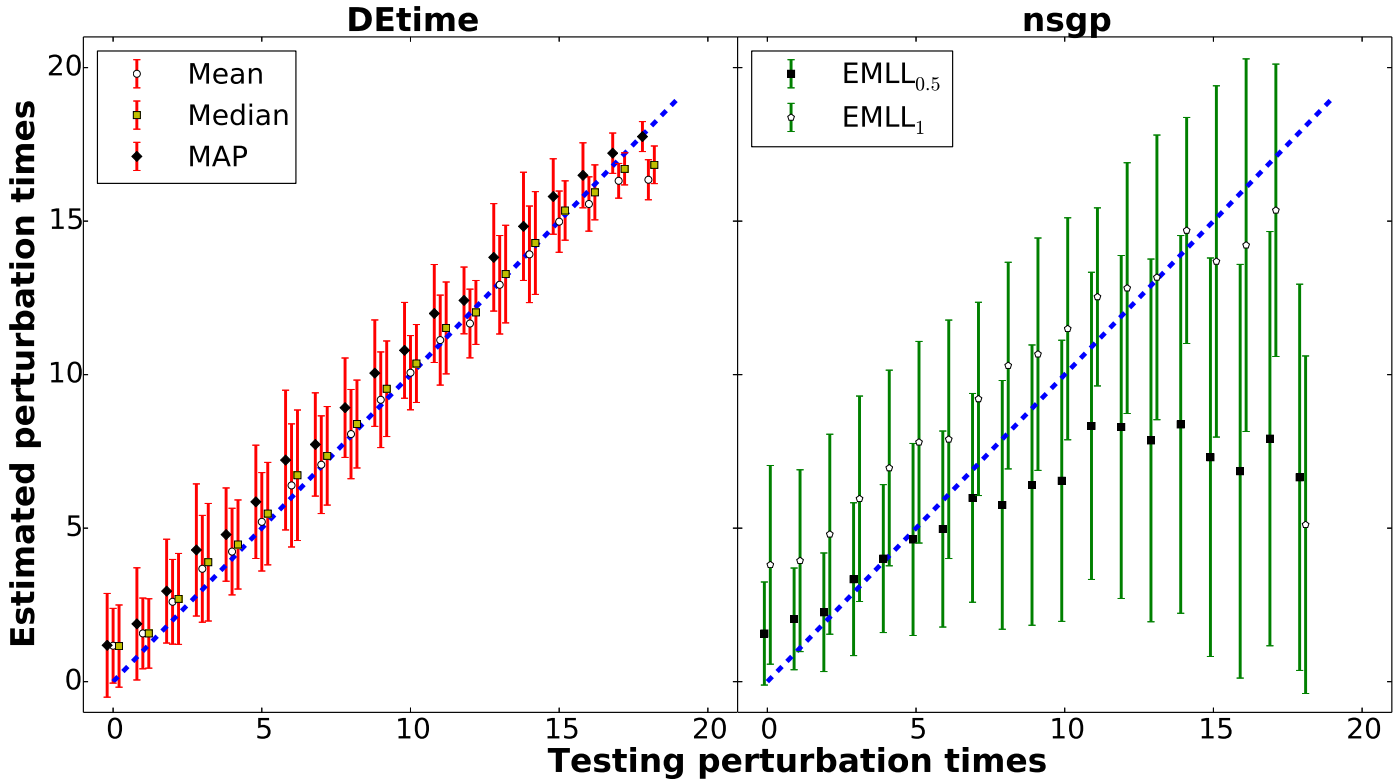
Posterior probability of the perturbation time t_p

$$p(t_p|y^c(\mathbf{T}), y^p(\mathbf{T})) \simeq \frac{p(y^c(\mathbf{T}), y^p(\mathbf{T})|t_p)}{\sum_{t=t_{\min}}^{t=t_{\max}} p(y^c(\mathbf{T}), y^p(\mathbf{T})|t)}$$



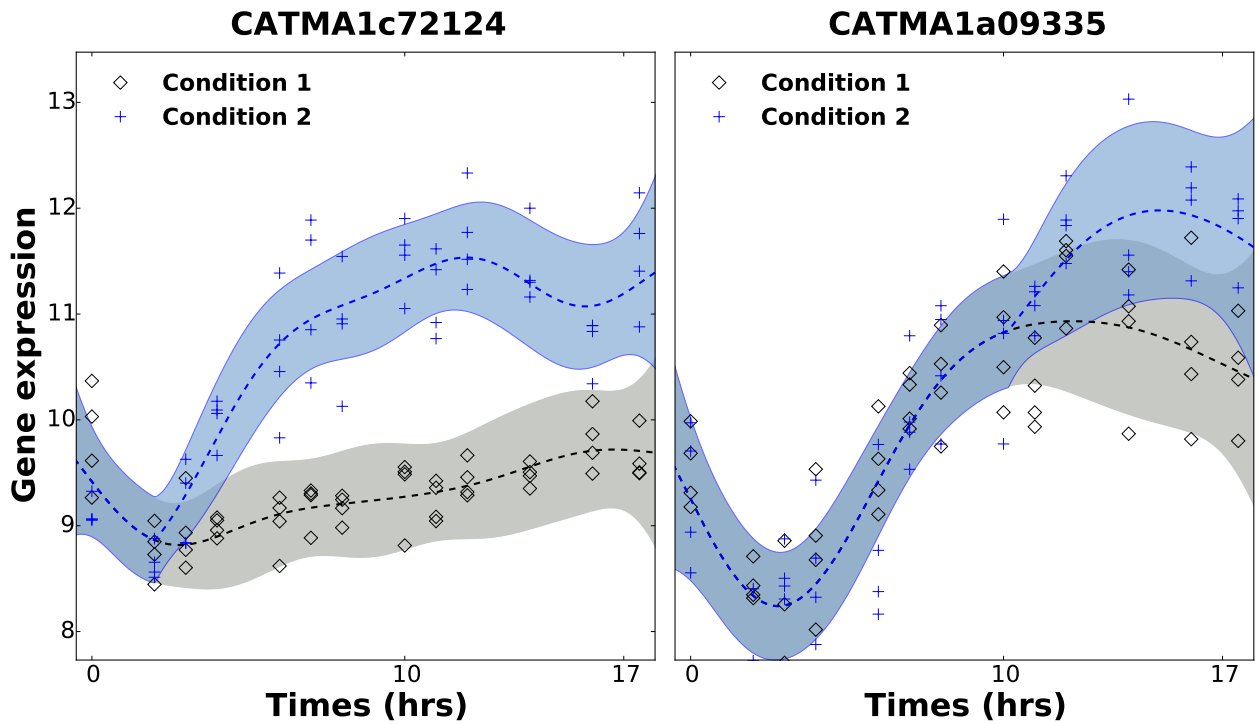
Comparison to DE thresholding approach

Alternative: use first point where some DE threshold is passed
We tested several DE metrics in the nsgp package



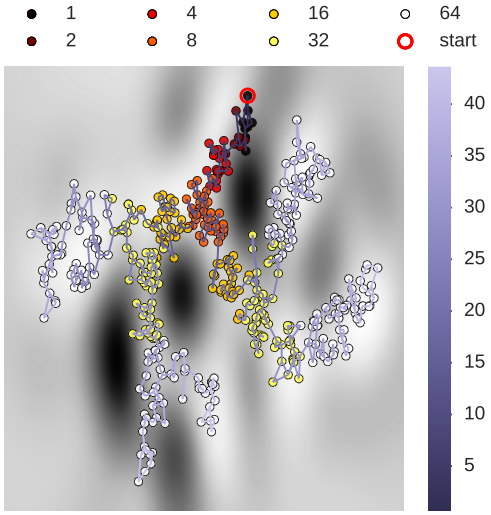
Investigating a plant's response to bacterial challenge

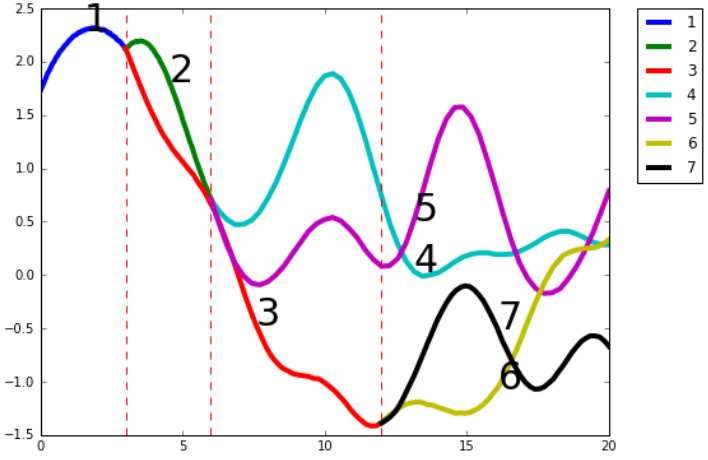
Infection with virulent *Pseudomonas syringage* pv. tomato DC3000 vs. disarmed strain DC3000*hrpA*

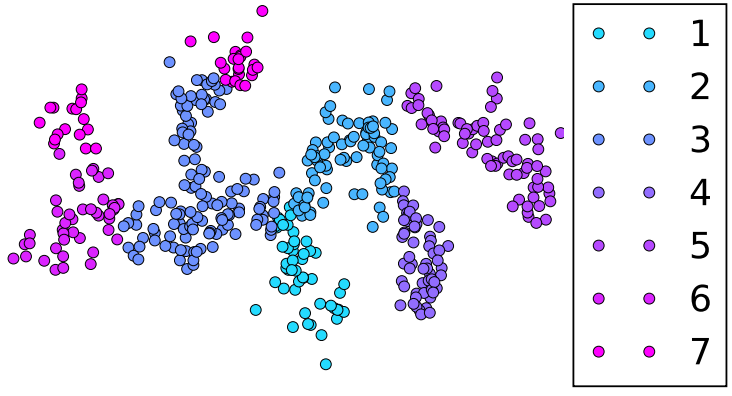


Yang *et al.* "Inferring the perturbation time from biological time course data" *Bioinformatics* (accepted) preprint: arxiv 1602.01743

Current work: modelling branching in single-cell data

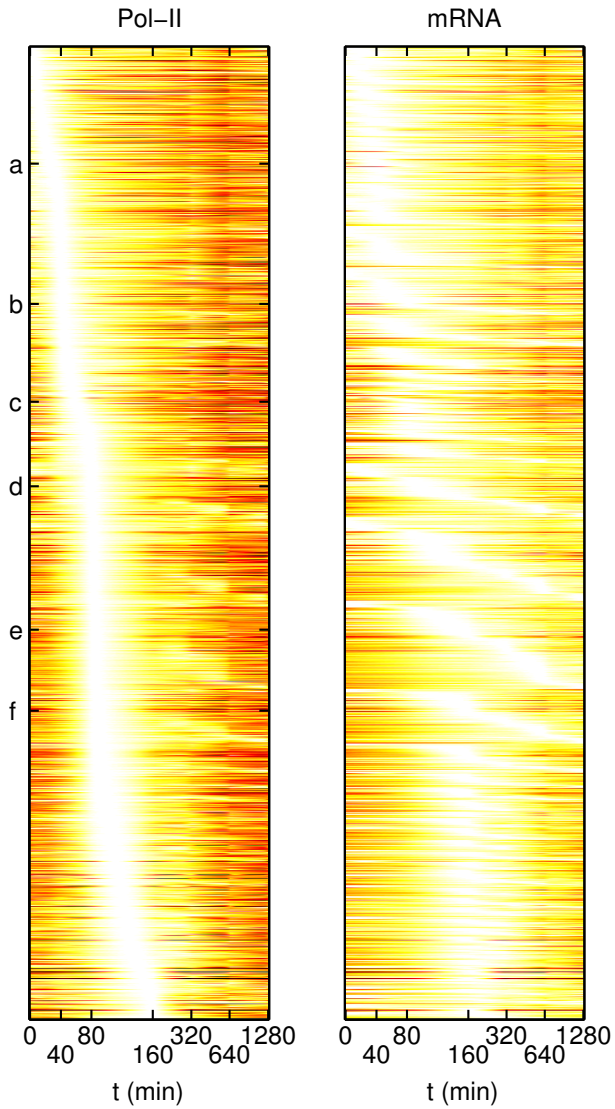


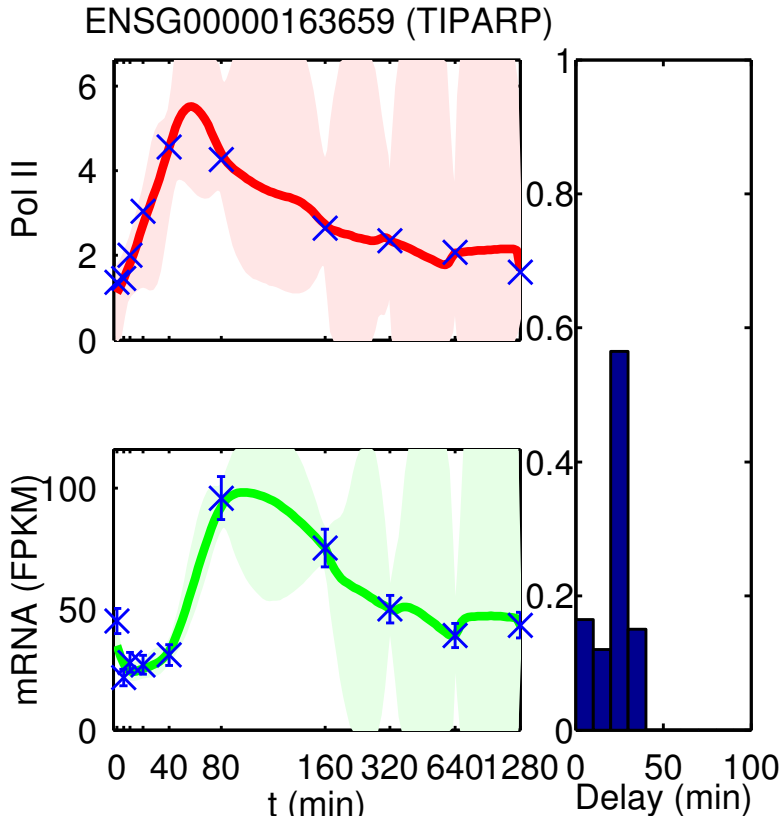




with A. Boukouvalas, M. Zweissele, J. Hensman and N. Lawrence

Ex 3. Linking Pol-II activity to mRNA profiles





Joint work with Antti Honkela, Jaakko Peltonen, Neil Lawrence

Linking Pol-II activity to mRNA profiles

$$\frac{\mathrm{d}m(t)}{\mathrm{d}t} = \beta p(t-\Delta) - \alpha m(t)$$

- $m(t)$ is mRNA concentration (RNA-Seq data)
- $p(t)$ is mRNA production rate (3' pol-II ChIP-Seq data)
- α is degradation rate (mRNA half-life $t_{1/2} = 2/\alpha$)
- Δ is processing delay

Linking Pol-II activity to mRNA profiles

$$\frac{\mathrm{d}m(t)}{\mathrm{d}t} = \beta p(t-\Delta) - \alpha m(t)$$

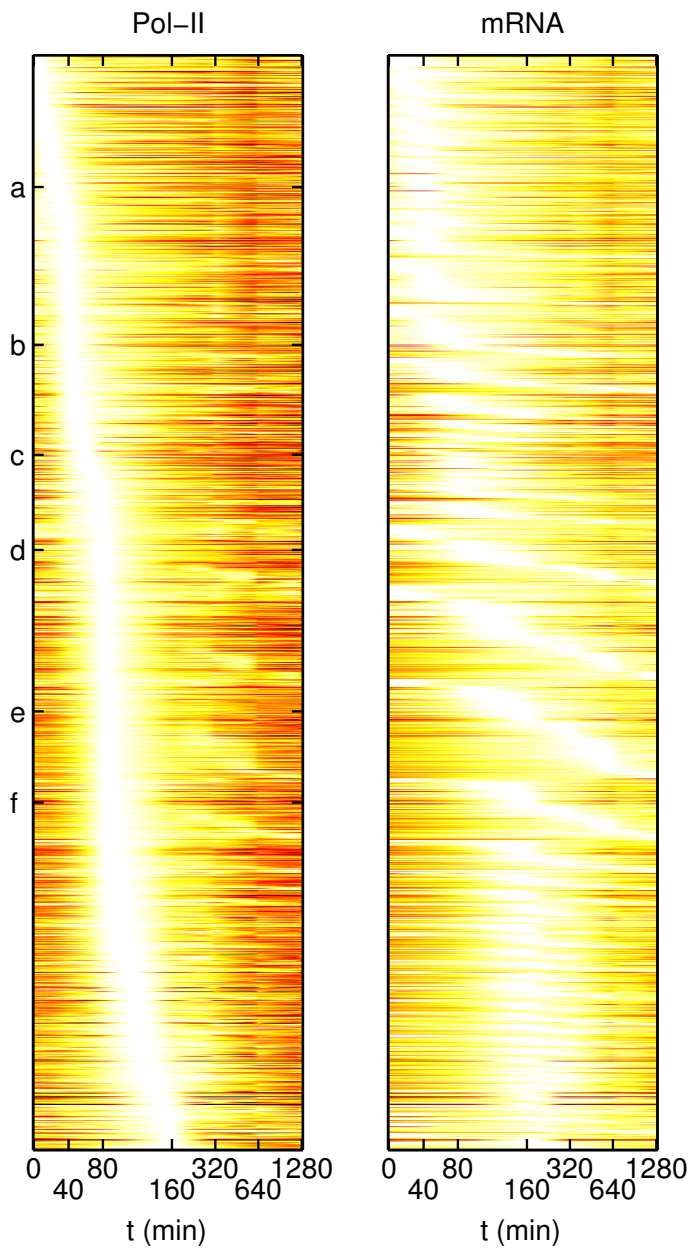
- $m(t)$ is mRNA concentration (RNA-Seq data)
- $p(t)$ is mRNA production rate (3' pol-II ChIP-Seq data)
- α is degradation rate (mRNA half-life $t_{1/2} = 2/\alpha$)
- Δ is processing delay

We model $p(t) \sim \mathcal{GP}(0, k_p)$ as a Gaussian process (GP)

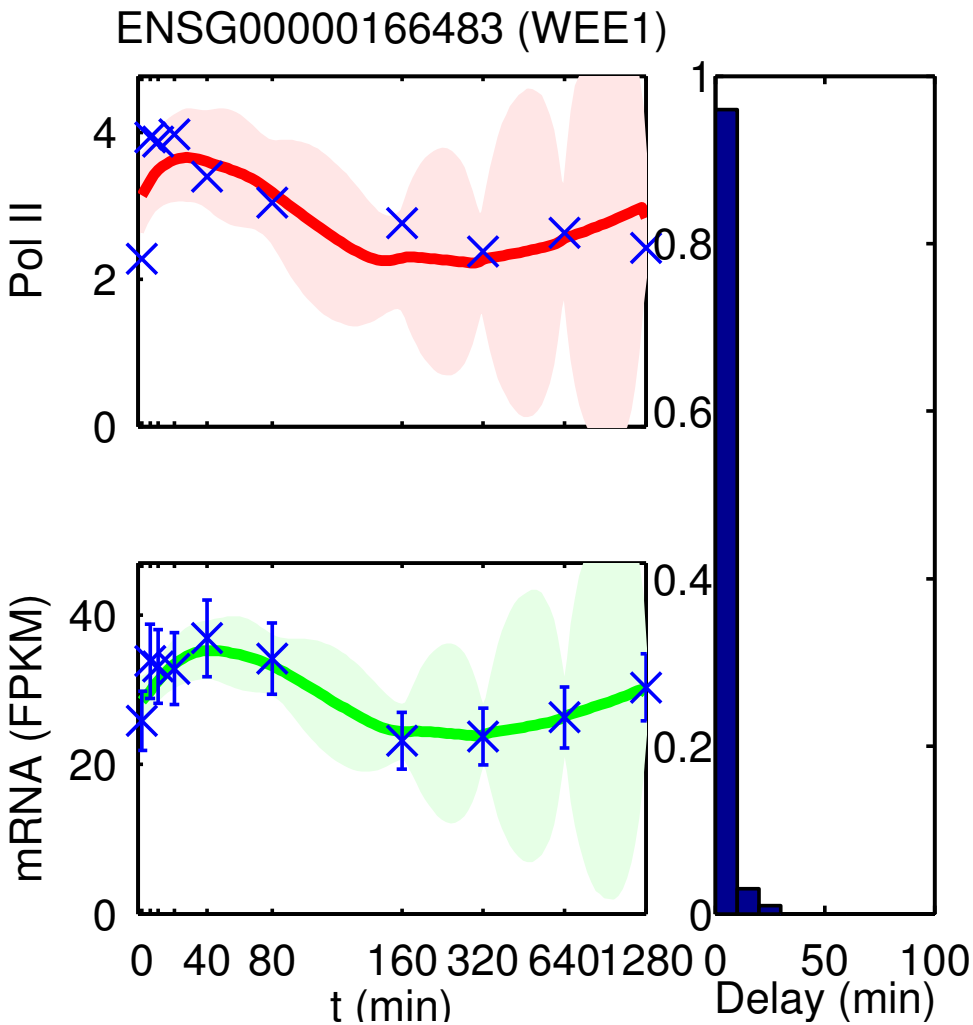
Likelihood can be worked out exactly

Bayesian MCMC used to estimate parameters α, β, Δ and GP covariance (2) and noise variance (1) parameters

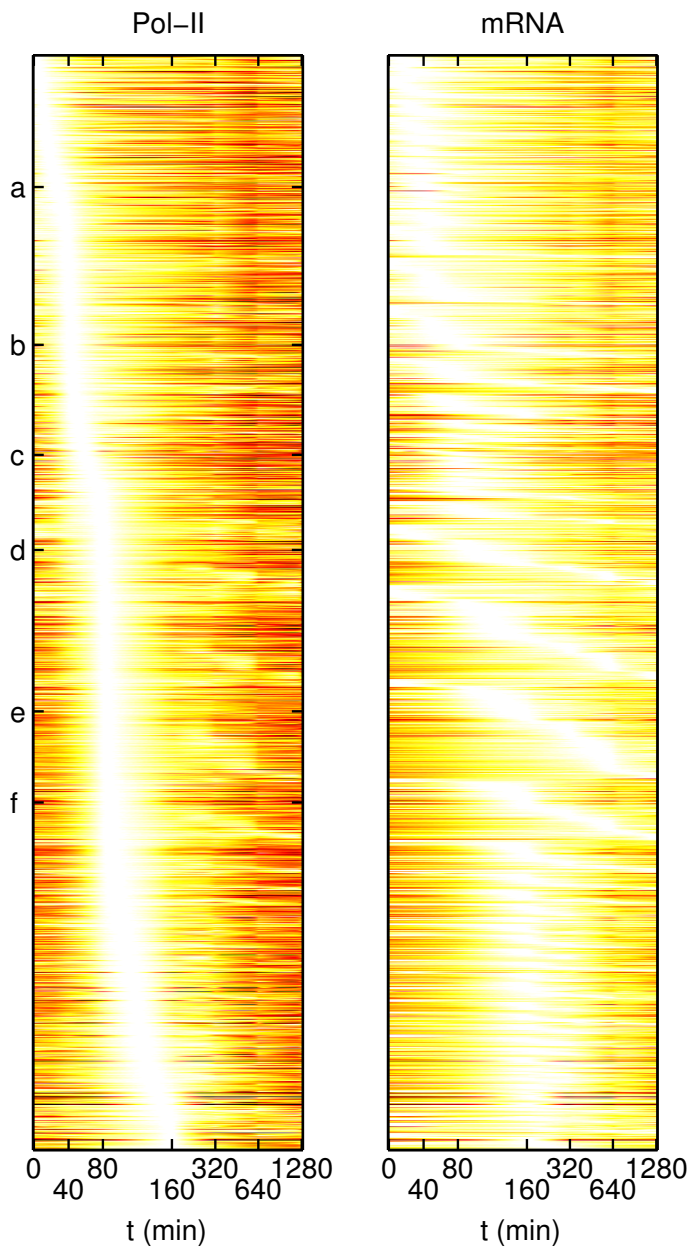
Example fits



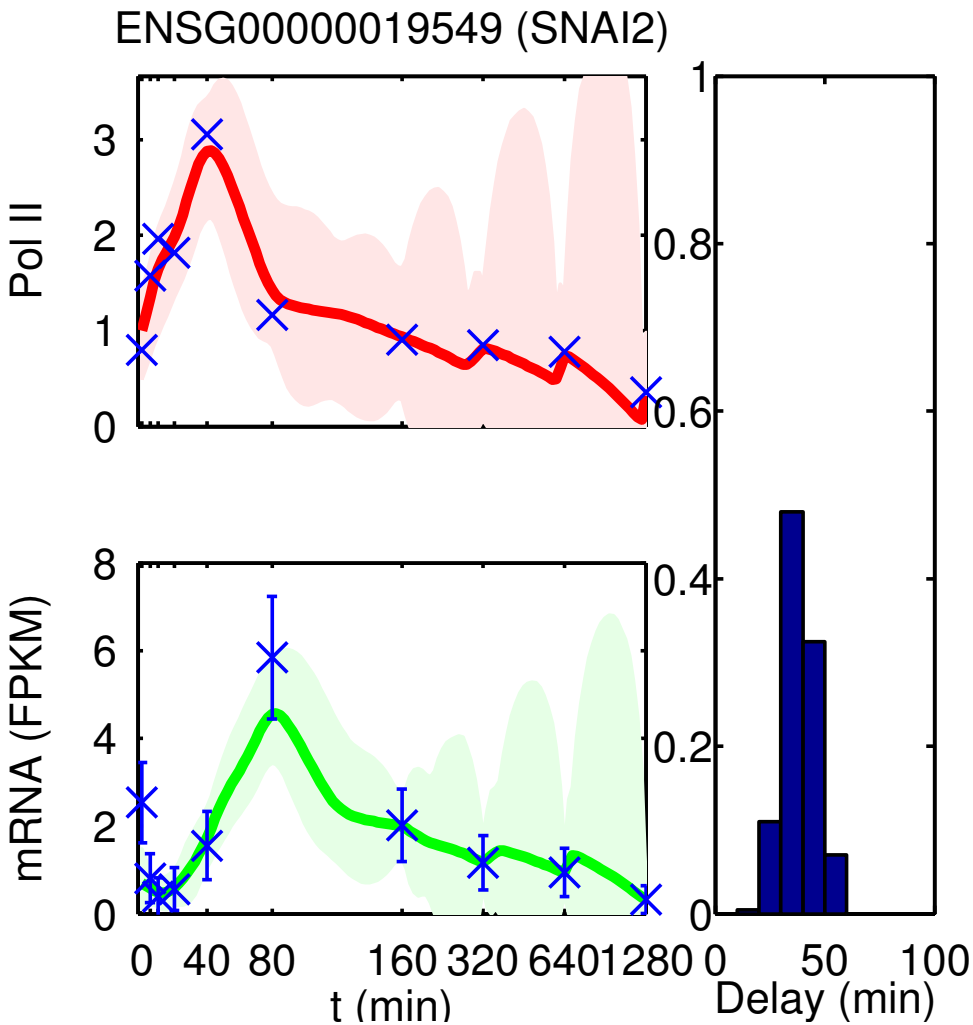
a: Early pol-II, no production delay



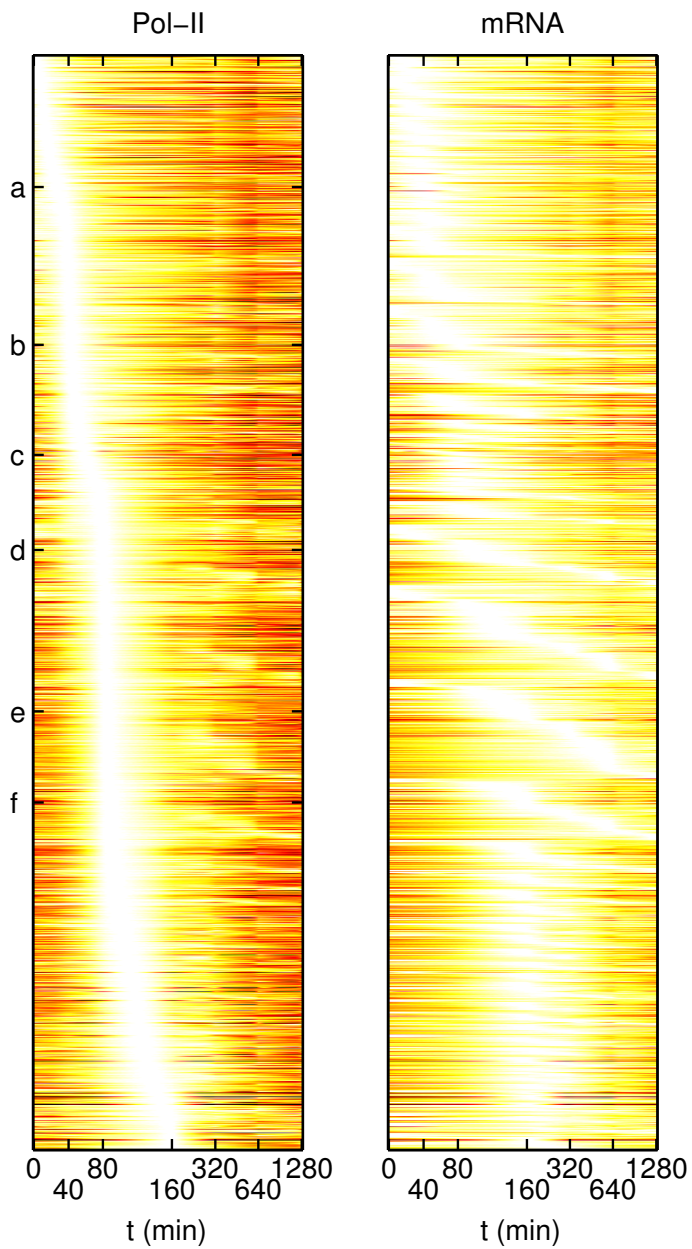
Example fits



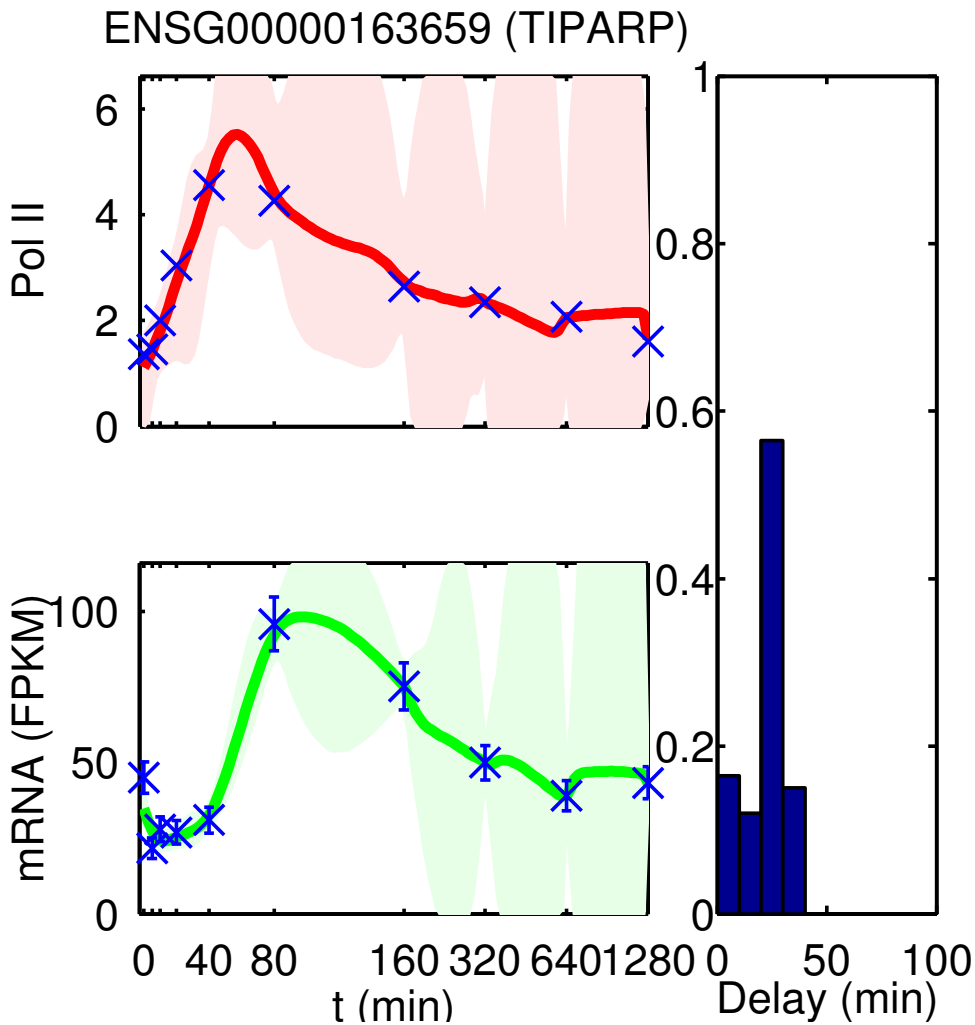
b: Early pol-II, delayed production



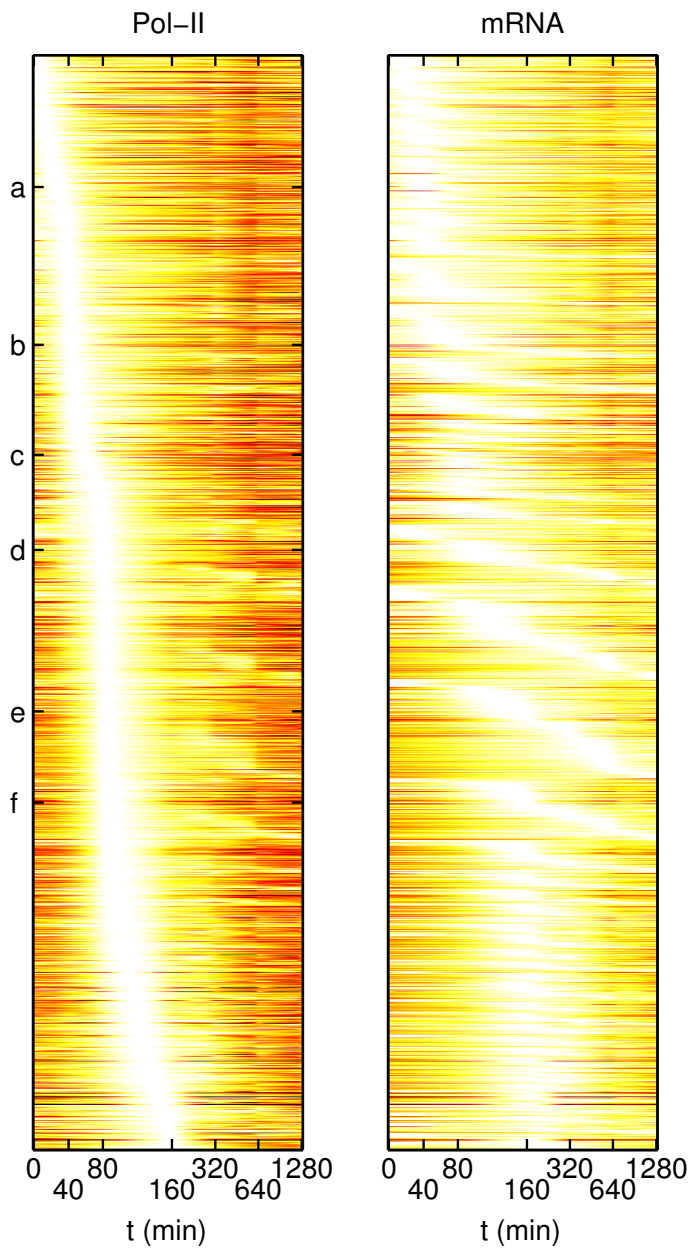
Example fits



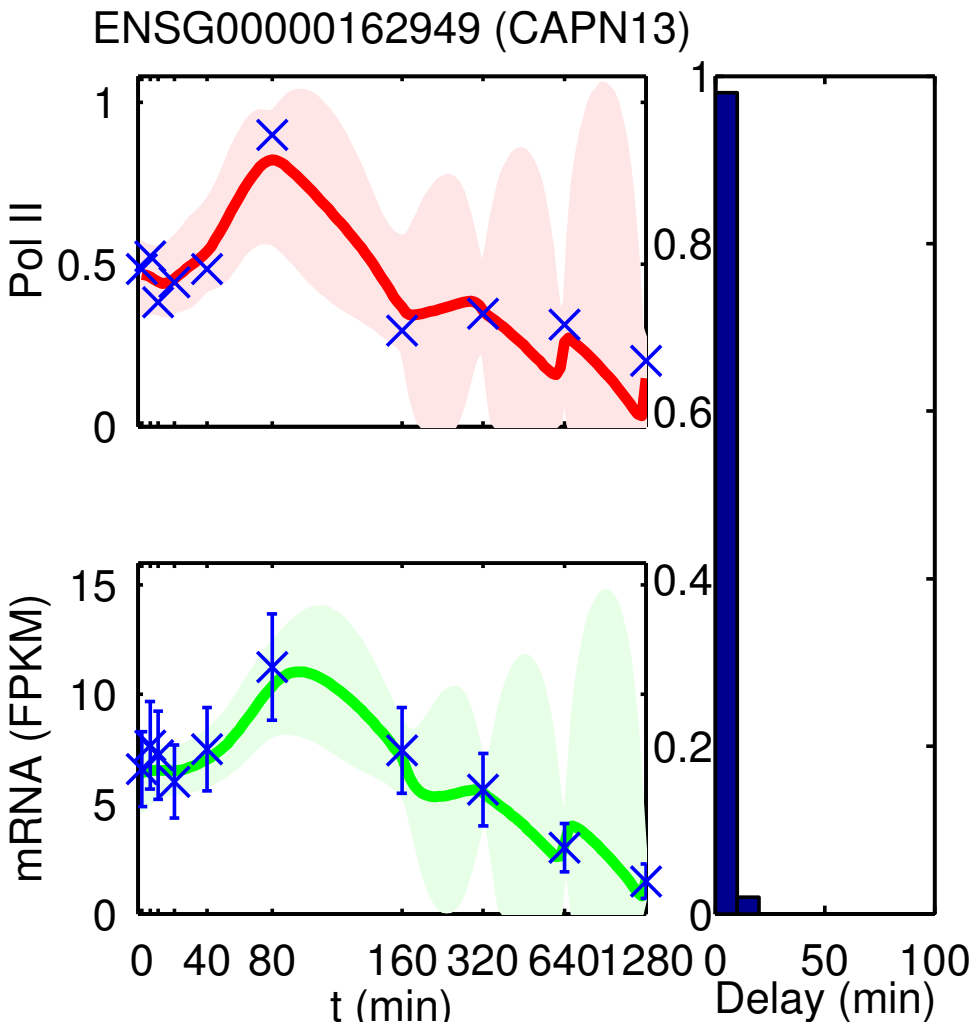
c: Later pol-II, delayed production



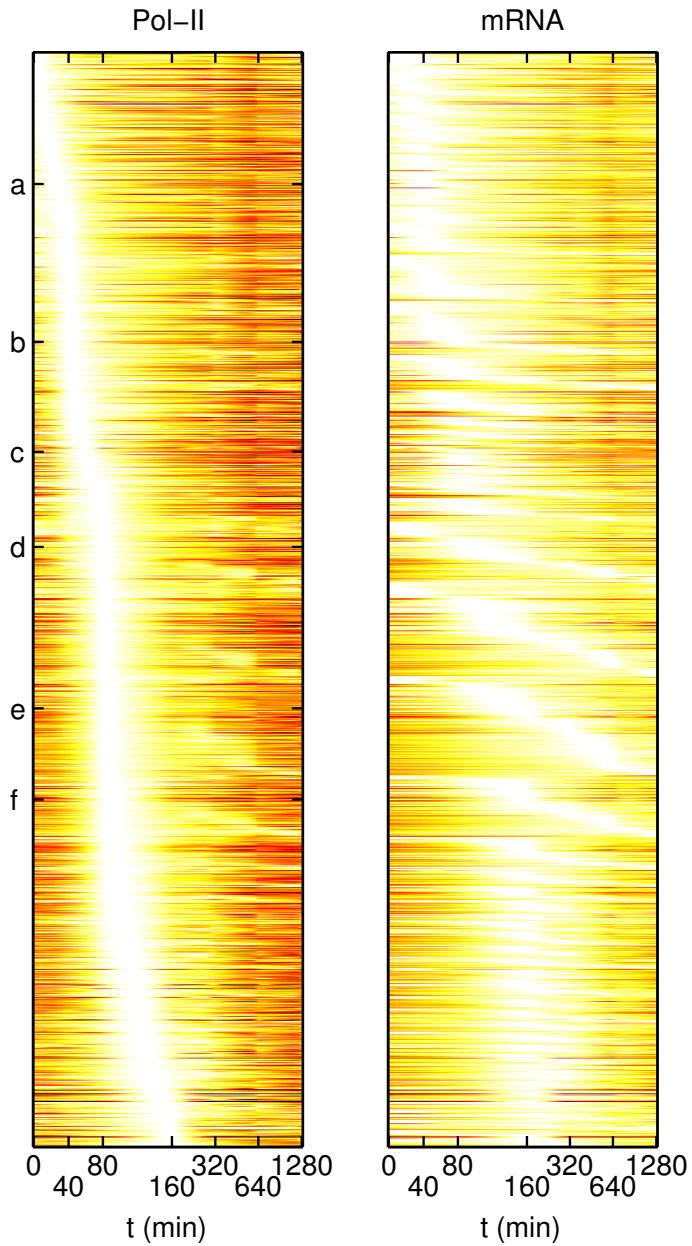
Example fits



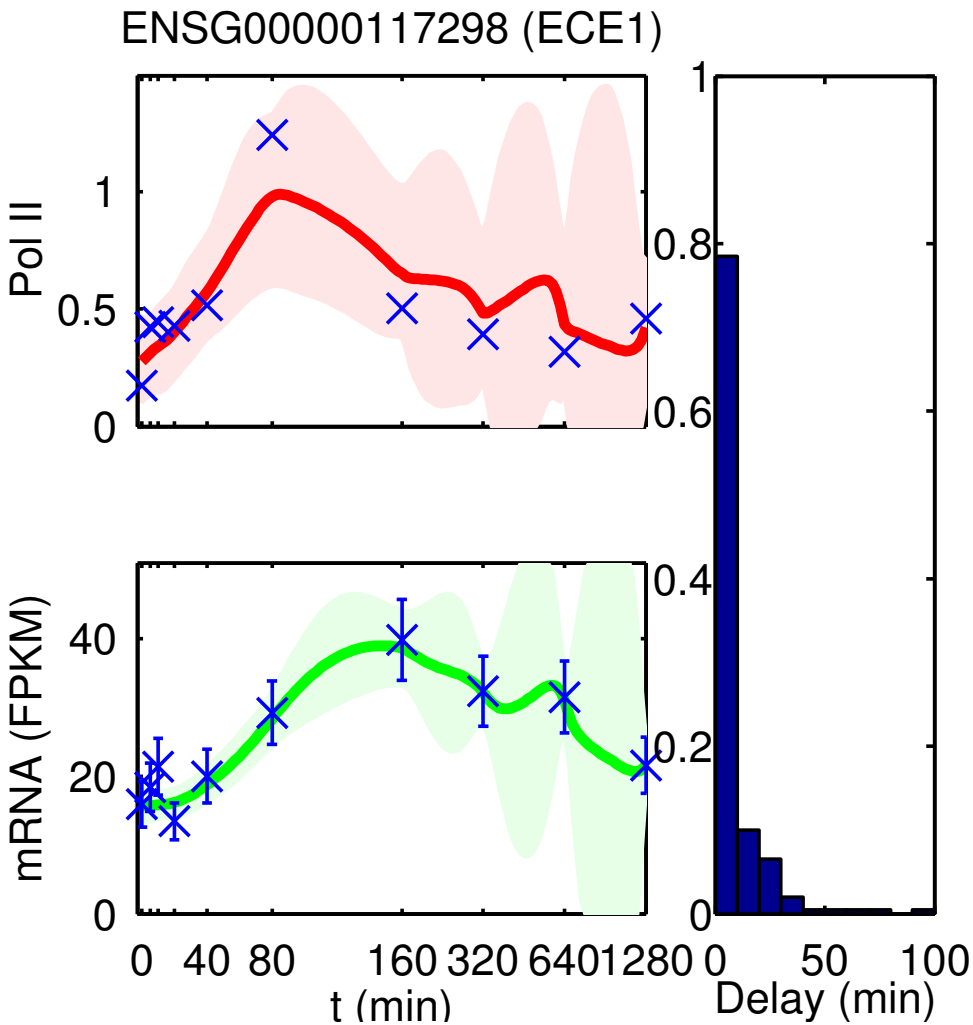
d: Late pol-II, no delay



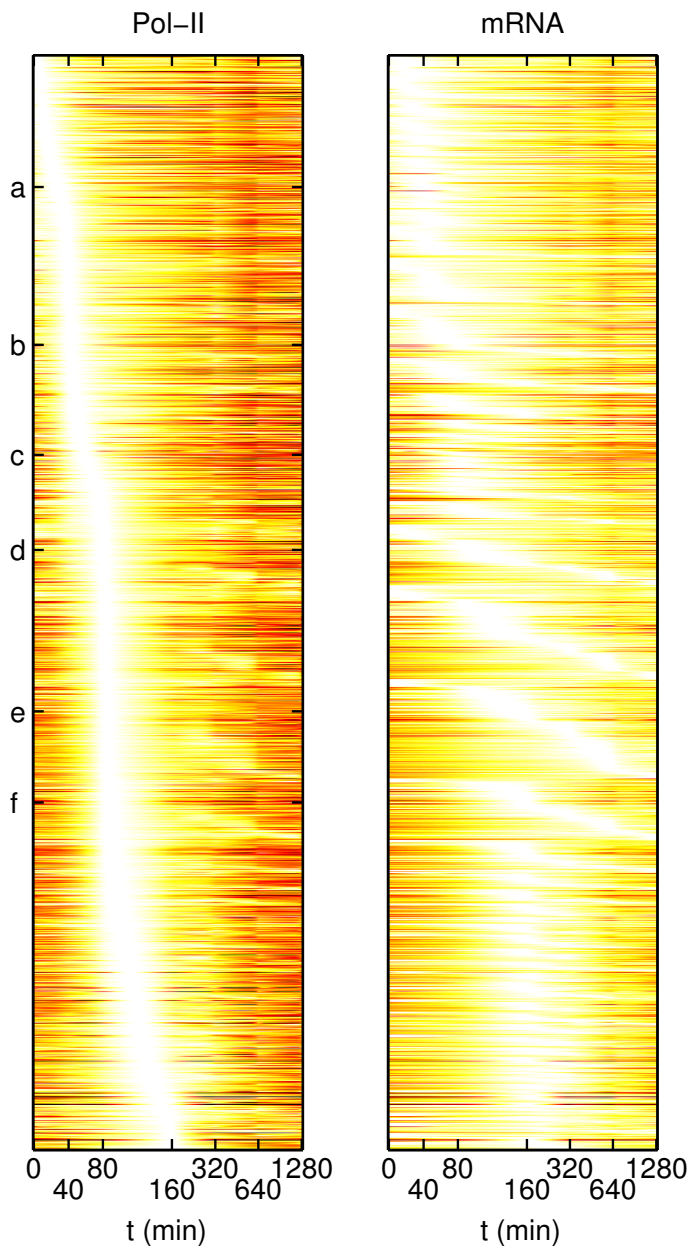
Example fits



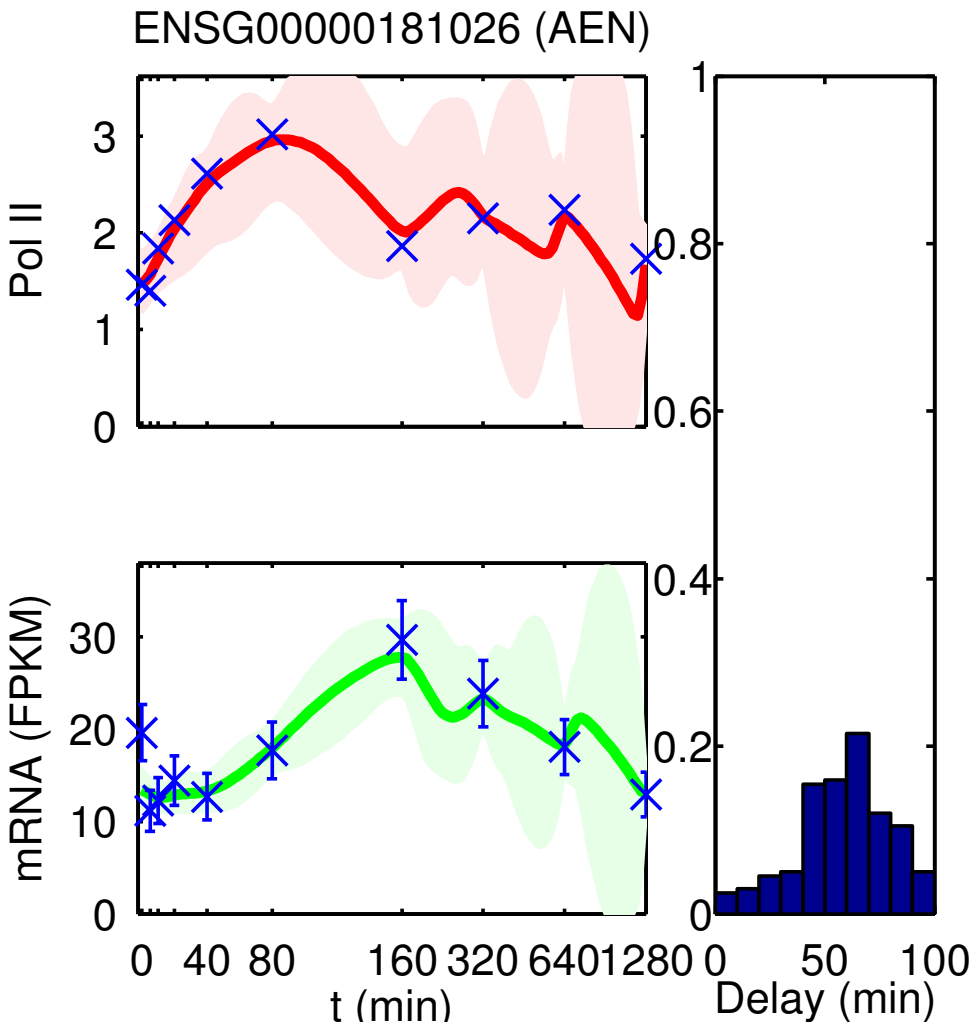
e: Late pol-II, no delay



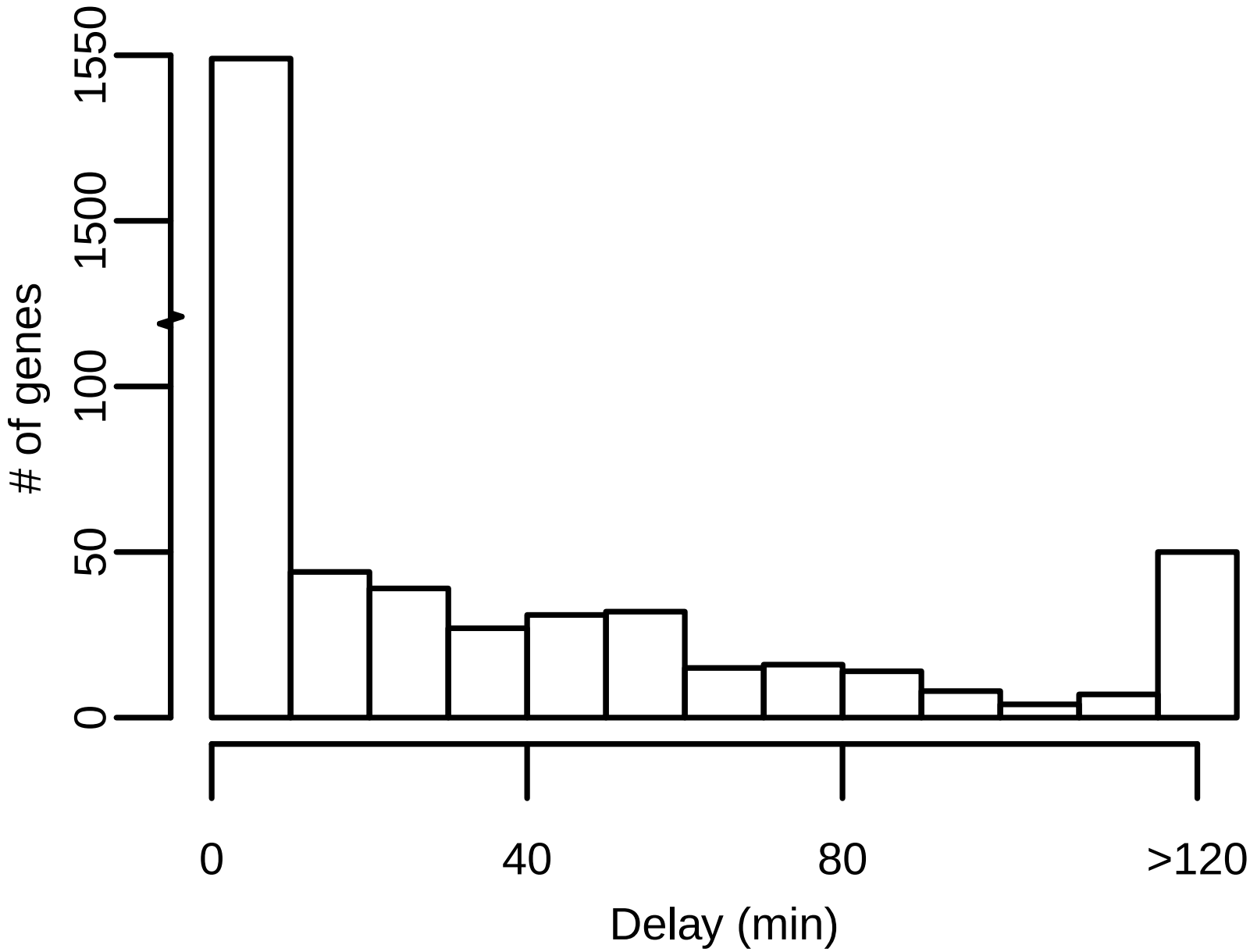
Example fits



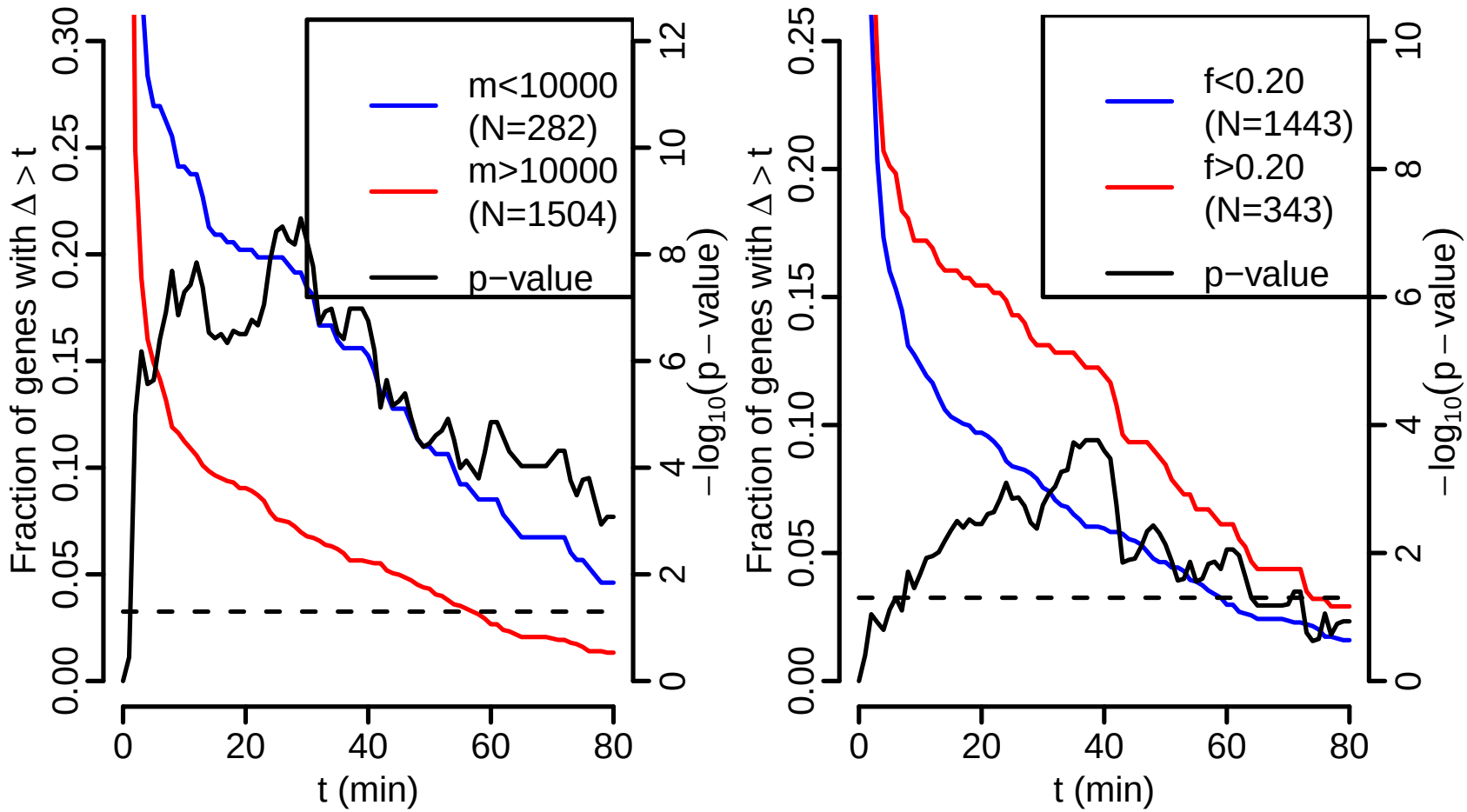
f: Late pol-II, delayed production



Large processing delays observed in 11% of genes

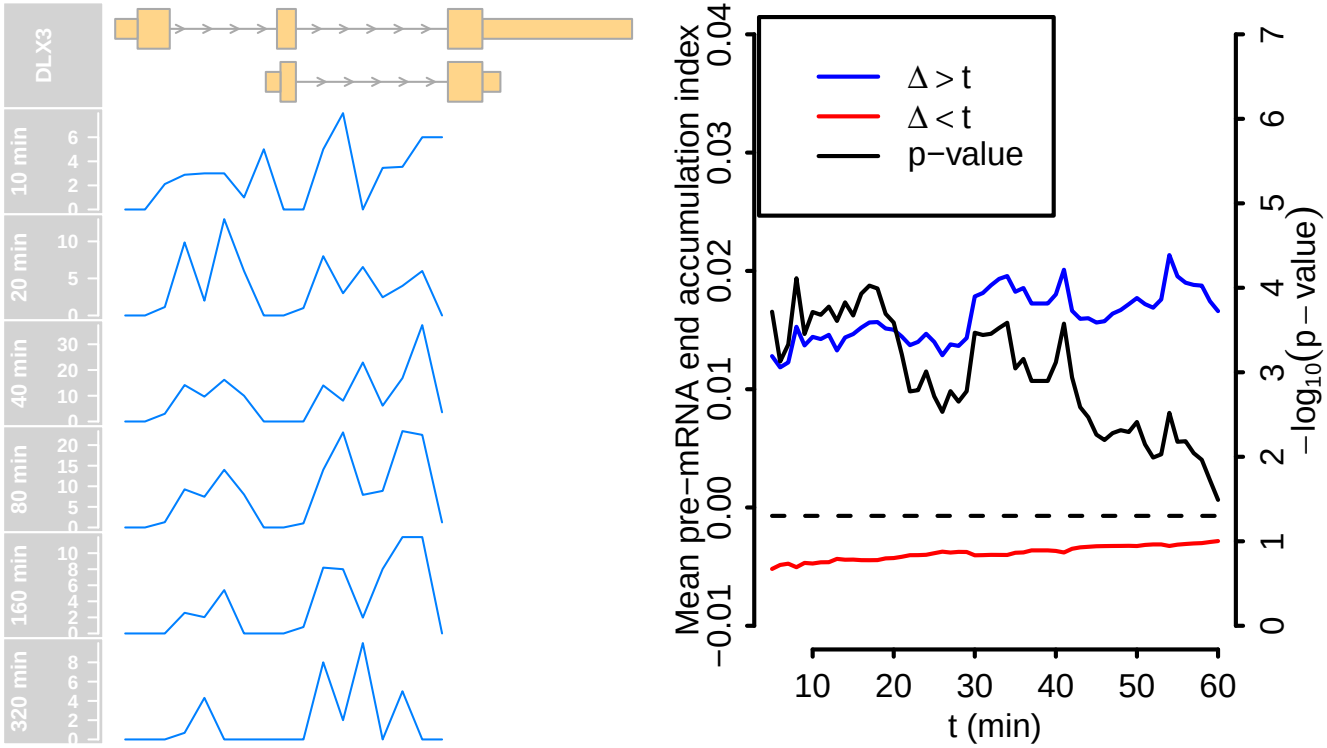


Delay linked with gene length and intron structure



Δ: delay m: gene length f: final intron length / gene length

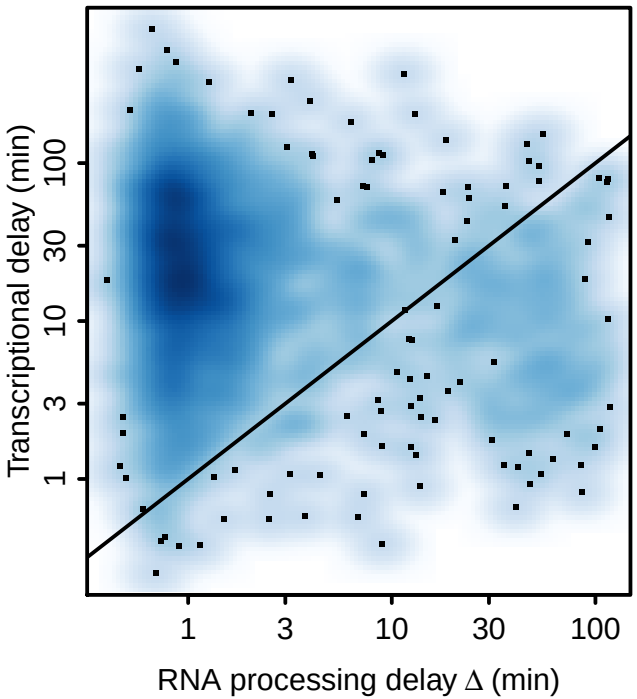
Delayed genes show evidence of late-intron retention



Left: density of RNA-Seq reads uniquely mapping to the introns in the DLX3 gene

Right: Differences in the mean pre-mRNA accumulation index in long delay genes (blue) and short delay genes (red)

Comparison of processing and transcription times



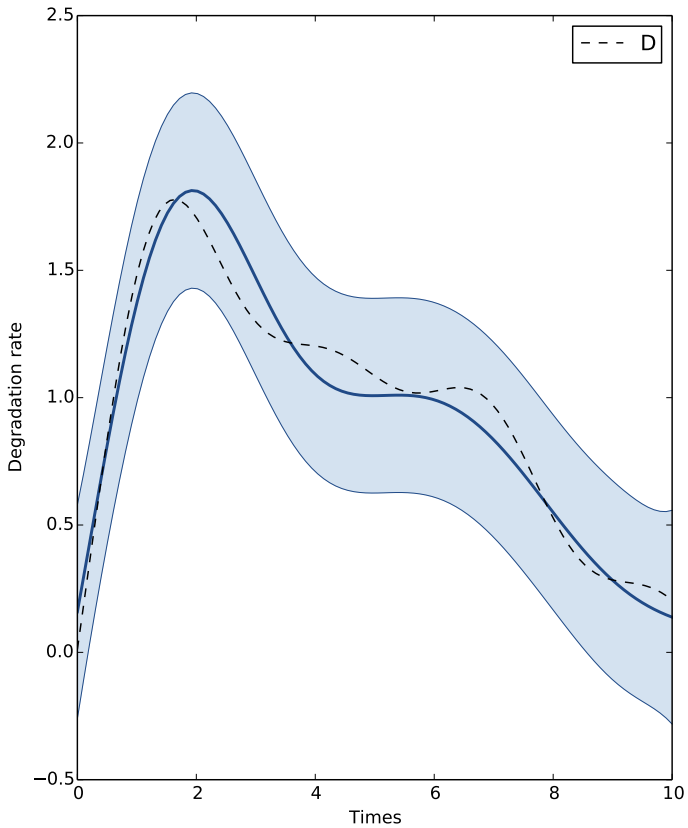
Transcription time = length/velocity
estimate from Danko *Mol Cell* 2013

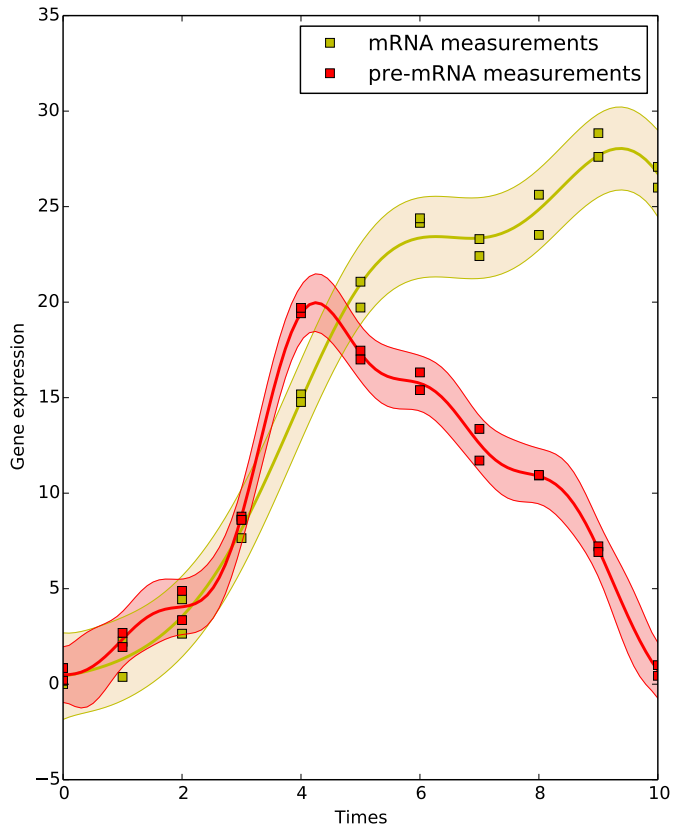
Transcription time > processing delay
in 87% of genes

Honkela *et al.* "Genome-wide modeling of transcription kinetics reveals patterns of RNA production delays" *PNAS* 2015

Extension - inferring time-varying degradation rates

$$\frac{\mathrm{d}m(t)}{\mathrm{d}t} = \beta p(t) - \alpha(t)m(t)$$





Conclusion

- Gaussian process models provide a good mix of flexibility and tractability in temporal and spatial modelling:
  - Hierarchical modelling, e.g. replicates, clusters, species
  - Periodic models without strong sinuisoidal assumptions
  - Tractable under branching and bifurcations
  - Tractable under linear operations on functions
- Easy addition and multiplication of kernels
- Good approximate inference algorithms for non-Gaussian data
- Codebase is growing making methods increasingly flexible and computationally efficient (e.g. GPy, GPflow and some R)

Funding: BBSRC (EraSysBio+ SYNERGY), EU FP7 (RADIANT)

Collaborators: Neil Lawrence, James Hensman, Antti Honkela, Jaakko Peltonen, Jing Yang, Alexis Boukouvalas, Max Zweissele

Our existential crisis

Commentary

Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era

Douglas B. Kell[1]* and Stephen G. Oliver[2]