Gaussian Process Regression Bootstrapping

Paul Kirk

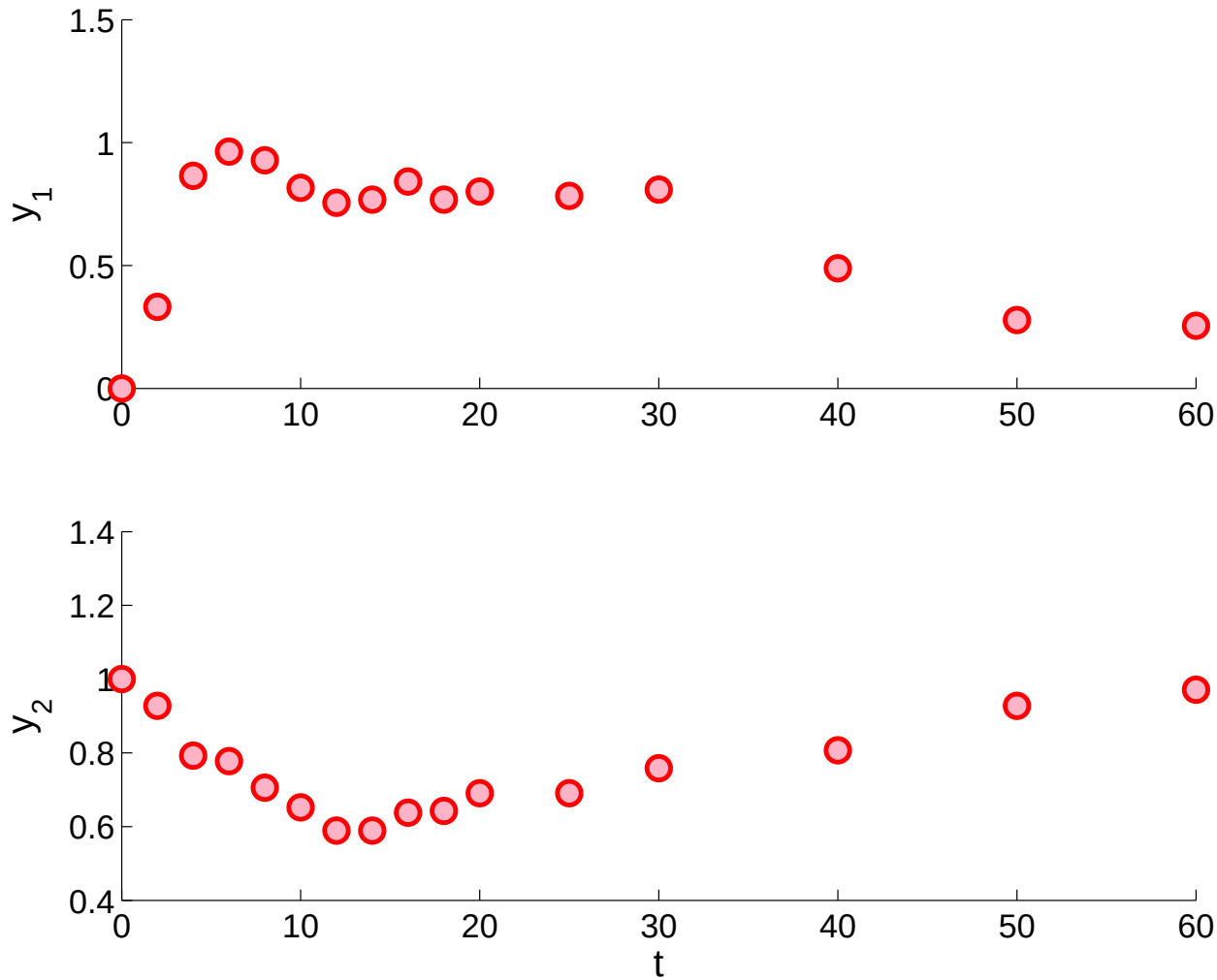
Centre for Bioinformatics & Institute of Mathematical Sciences
& Centre for Integrative Systems Biology at Imperial College London

1st April 2009

Motivating Example

We have a parametric model, $\mathcal{M}(t;\theta)$, which predicts the values of responses y_1 and y_2 at time t.

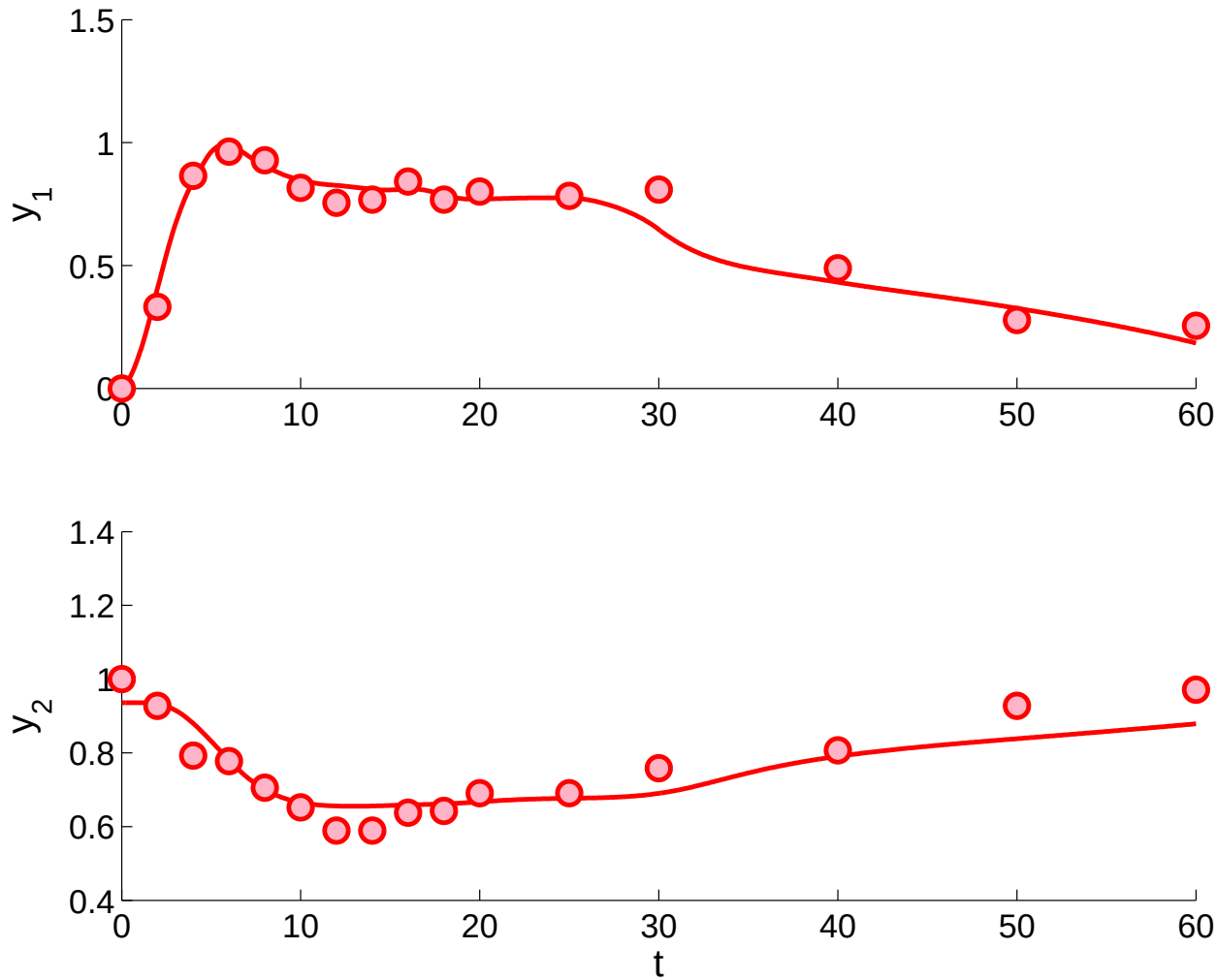
We observe the following data ...



Motivating Example

We have a parametric model, $\mathcal{M}(t;\theta)$, which predicts the values of responses y_1 and y_2 at time t.

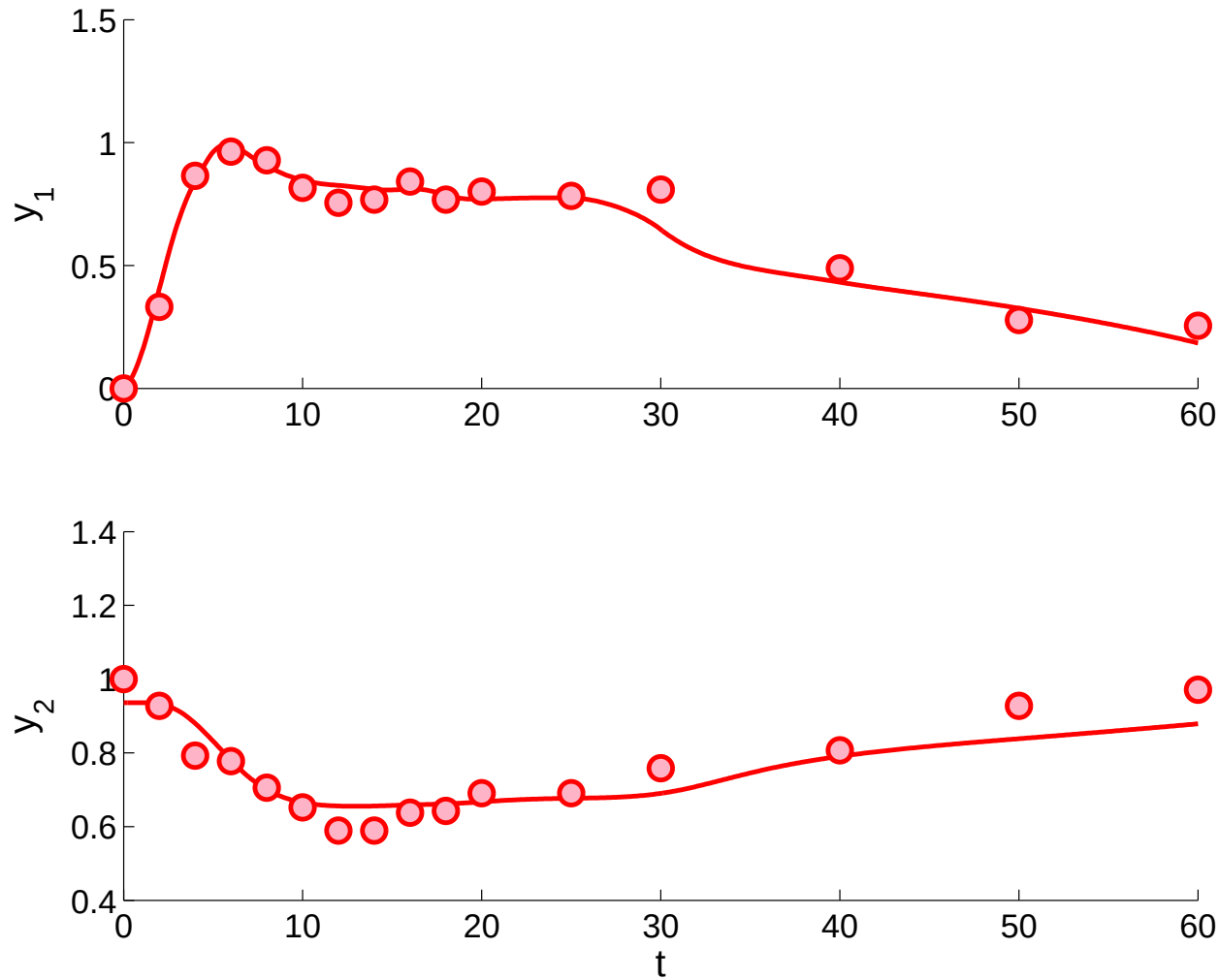
. . . and find an estimate of the model parameters, $\hat{\theta}$.



Motivating Example

We have a parametric model, $\mathcal{M}(t;\theta)$, which predicts the values of responses y_1 and y_2 at time t.

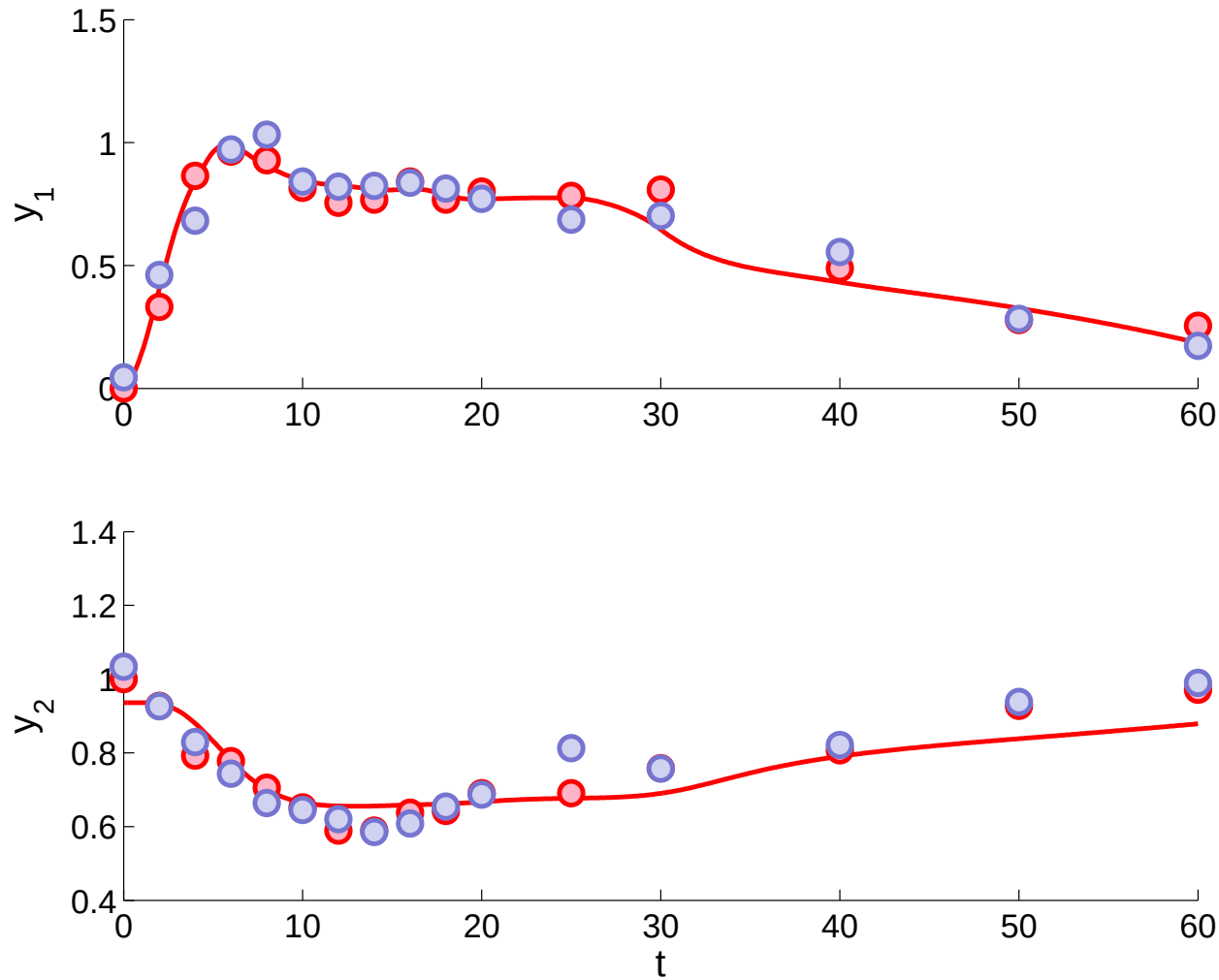
But what if we'd observed slightly different data?



Motivating Example

We have a parametric model, $\mathcal{M}(t;\theta)$, which predicts the values of responses y_1 and y_2 at time t.

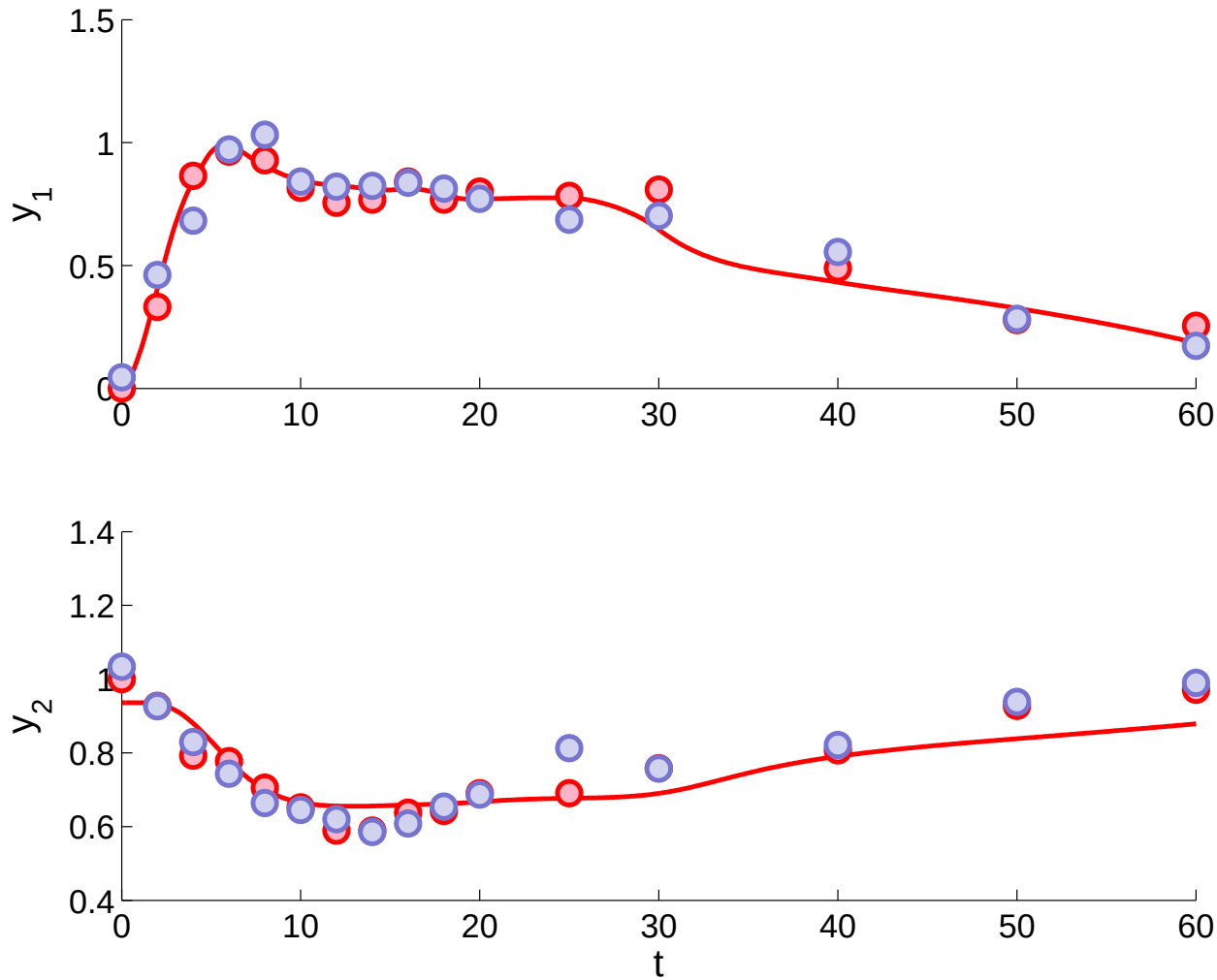
For example . . .



Motivating Example

We have a parametric model, $\mathcal{M}(t;\theta)$, which predicts the values of responses y_1 and y_2 at time t.

Would we have obtained a similar $\hat{\theta}$?



The Bootstrap

The Bootstrap: A statistical resampling technique used to assess properties of quantities or statistics inferred from a data set.

Overview

Main requirement: An approximating distribution from which samples may be drawn.

1. Ideally, we would repeat the experiment.
2. Nonparametric bootstrap: Draw samples with replacement from original data set.
   - Time course data typically have few replicates, so hard to apply.
3. Parametric bootstrap: Fit a parametric probability model to the original data.
   - How can we fit such a probability model to our time course data?

The Bootstrap

The Bootstrap: A statistical resampling technique used to assess properties of quantities or statistics inferred from a data set.

Overview

Main requirement: An approximating distribution from which samples may be drawn.

1. Ideally, we would repeat the experiment.
2. Nonparametric bootstrap: Draw samples with replacement from original data set.
   - Time course data typically have few replicates, so hard to apply.
3. Parametric bootstrap: Fit a parametric probability model to the original data.
   - How can we fit such a probability model to our time course data?

The Bootstrap

The Bootstrap: A statistical resampling technique used to assess properties of quantities or statistics inferred from a data set.

Overview

Main requirement: An approximating distribution from which samples may be drawn.

1. Ideally, we would repeat the experiment.
2. Nonparametric bootstrap: Draw samples with replacement from original data set.
   - Time course data typically have few replicates, so hard to apply.
3. Parametric bootstrap: Fit a parametric probability model to the original data.
   - How can we fit such a probability model to our time course data?

The Bootstrap

The Bootstrap: A statistical resampling technique used to assess properties of quantities or statistics inferred from a data set.

Overview

Main requirement: An approximating distribution from which samples may be drawn.

1. Ideally, we would repeat the experiment.
2. Nonparametric bootstrap: Draw samples with replacement from original data set.
   - Time course data typically have few replicates, so hard to apply.
3. Parametric bootstrap: Fit a parametric probability model to the original data.
   - How can we fit such a probability model to our time course data?

The Bootstrap

The Bootstrap: A statistical resampling technique used to assess properties of quantities or statistics inferred from a data set.

Overview

Main requirement: An approximating distribution from which samples may be drawn.

1. Ideally, we would repeat the experiment.
2. Nonparametric bootstrap: Draw samples with replacement from original data set.
   - Time course data typically have few replicates, so hard to apply.
3. Parametric bootstrap: Fit a parametric probability model to the original data.
   - How can we fit such a probability model to our time course data?

Gaussian process regression (GPR)

Regression model:

$$y(t) = f(t) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2).$$

How do we choose f?

GPR: place a Gaussian process prior over f:

Gaussian process prior

- For any finite collection of times, $s_1, \ldots, s_n$, the function outputs $f(s_1), \ldots, f(s_n)$ are jointly distributed according to a multivariate Gaussian:

$$\forall n \in \mathbb{N} \text{ and } s_i \in \mathbb{R}_{\geq 0}, \quad [f(s_1), \ldots, f(s_n)]^\top \sim \mathcal{N}(\mathbf{m}, S).$$

- $\mathbf{m}_i = m(s_i)$ – mean function.
- $S_{ij} = k(s_i, s_j)$ – covariance function.
- Prior beliefs regarding properties of f expressed through m and k.

We write $f(t) \sim \mathcal{GP}(m, k)$.

Gaussian process regression (GPR)

Regression model:

$$y(t) = f(t) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2).$$

How do we choose f?

GPR: place a Gaussian process prior over f:

Gaussian process prior

- For any finite collection of times, $s_1, \ldots, s_n$, the function outputs $f(s_1), \ldots, f(s_n)$ are jointly distributed according to a multivariate Gaussian:

$$\forall n \in \mathbb{N} \text{ and } s_i \in \mathbb{R}_{\geq 0}, \quad [f(s_1), \ldots, f(s_n)]^\top \sim \mathcal{N}(\mathbf{m}, S).$$

- $\mathbf{m}_i = m(s_i)$ – mean function.
- $S_{ij} = k(s_i, s_j)$ – covariance function.
- Prior beliefs regarding properties of f expressed through m and k.

We write $f(t) \sim \mathcal{GP}(m, k)$.

Gaussian process regression (GPR)... continued

We have the following:

$$\begin{array}{ll} \text{Observed data:} & y_1, \ldots, y_p. \\ \text{Times:} & t_1, \ldots, t_p. \end{array}$$

We may update our GP prior in light of the observed data

Gaussian process posterior

1. According to our GP prior, $[f(s_1), \ldots, f(s_n), f(t_1), \ldots f(t_p)]^\top$ are jointly distributed according to a multivariate Gaussian.
2. Also, $y(t) = f(t) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$.
3. It follows that $[f(s_1), \ldots, f(s_n)]^\top | y_1, \ldots, y_p$ are jointly distributed according to a multivariate Gaussian.
4. We hence have a Gaussian process posterior, $f(t) \sim \mathcal{GP}(m_{post}, k_{post})$.

GPR Bootstrapping

1. From previously, $f(t) \sim \mathcal{GP}(m_{post}, k_{post})$.
2. So, the posterior distribution of

$$[f(t_1), \ldots, f(t_p)]^\top \text{ is } \sim \mathcal{N}(\mu, \Sigma).$$

3. As $y(t) = f(t) + \epsilon$, the posterior distribution of

$$[y(t_1), \ldots, y(t_p)]^\top \text{ is } \mathcal{N}(\mu, \Sigma + \sigma_\epsilon^2 I).$$

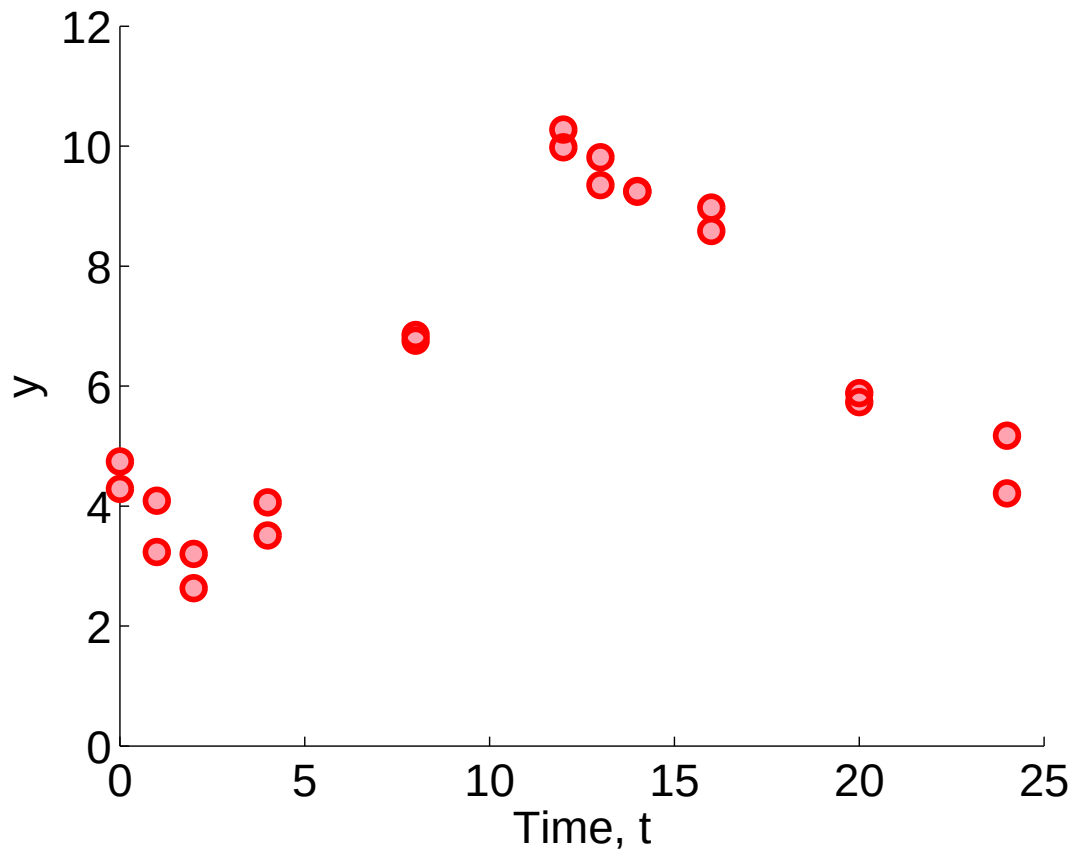
We hence have a parametric probability model for our time course data.

We may use this to obtain bootstrap samples.

(see Kirk and Stumpf, 2009, *GPR bootstrapping*, **Bioinformatics**.)

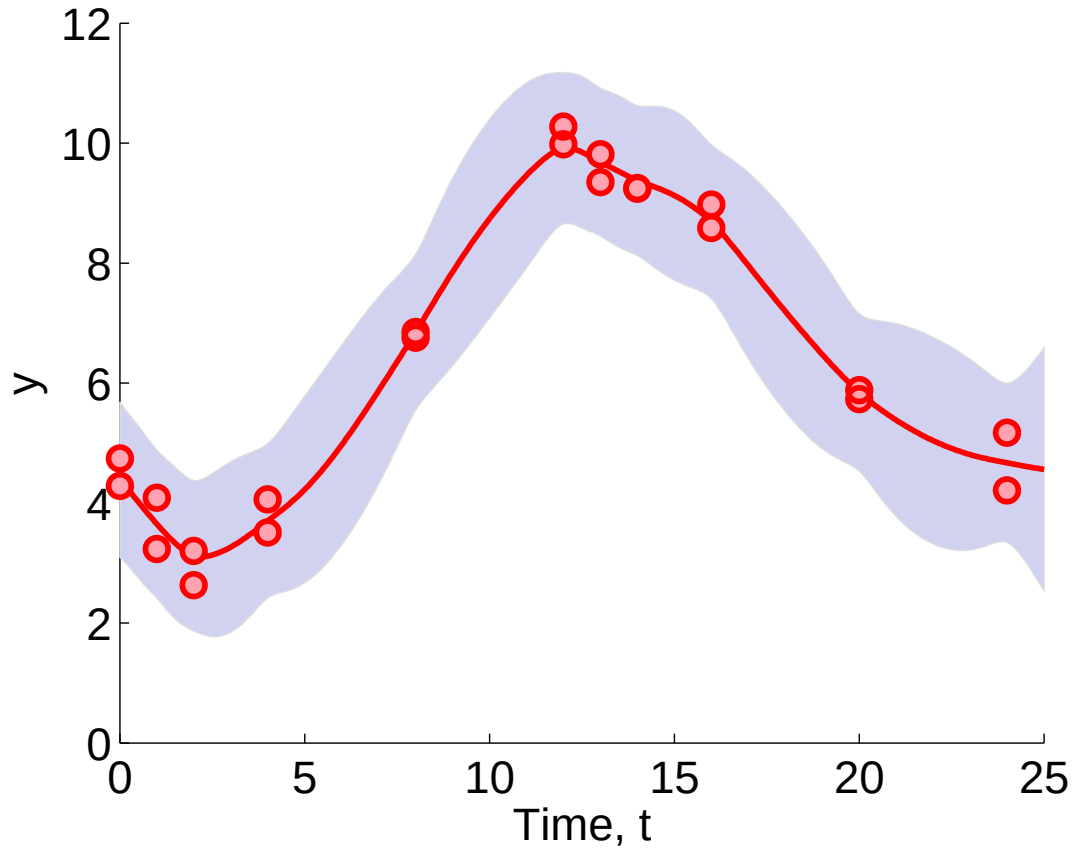
GPR bootstrapping

- First fit a GP regressor to the data to obtain a GP posterior.
- Draw bootstrap samples from resulting multivariate Gaussian.
- Infer quantity of interest for all data sets & assess variability.



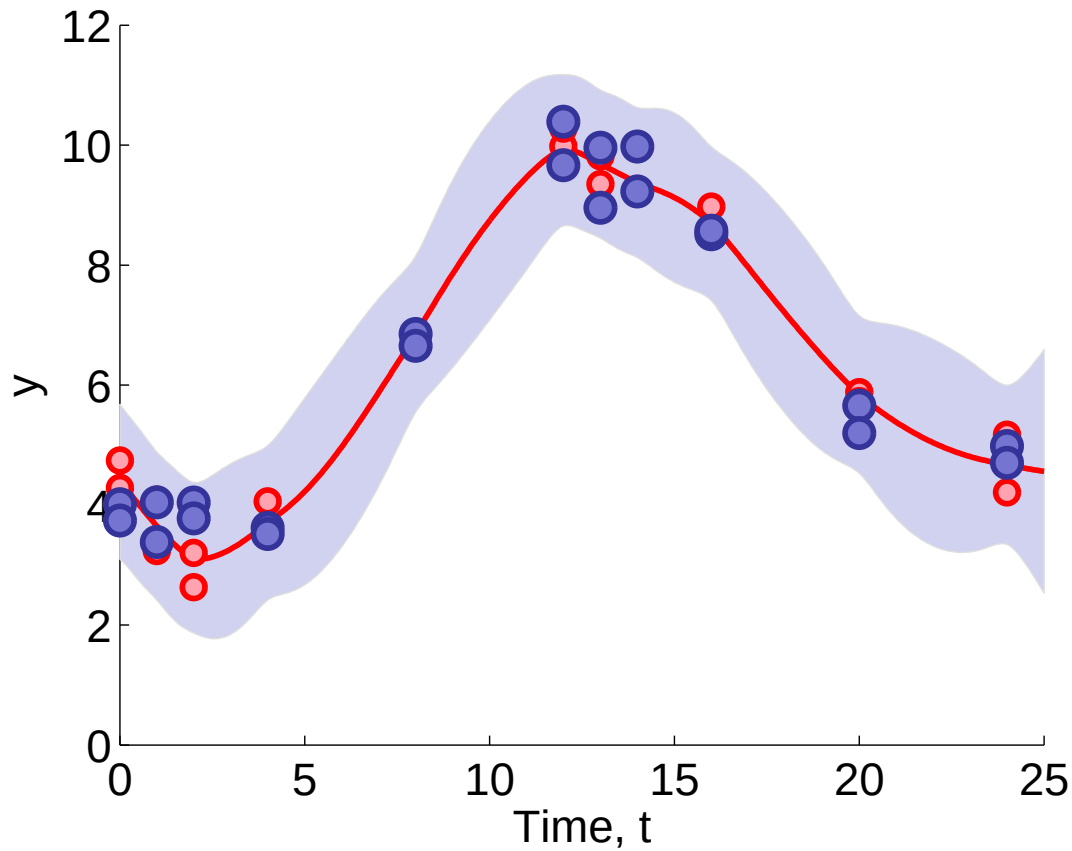
GPR bootstrapping

- First fit a GP regressor to the data to obtain a GP posterior.
- Draw bootstrap samples from resulting multivariate Gaussian.
- Infer quantity of interest for all data sets & assess variability.



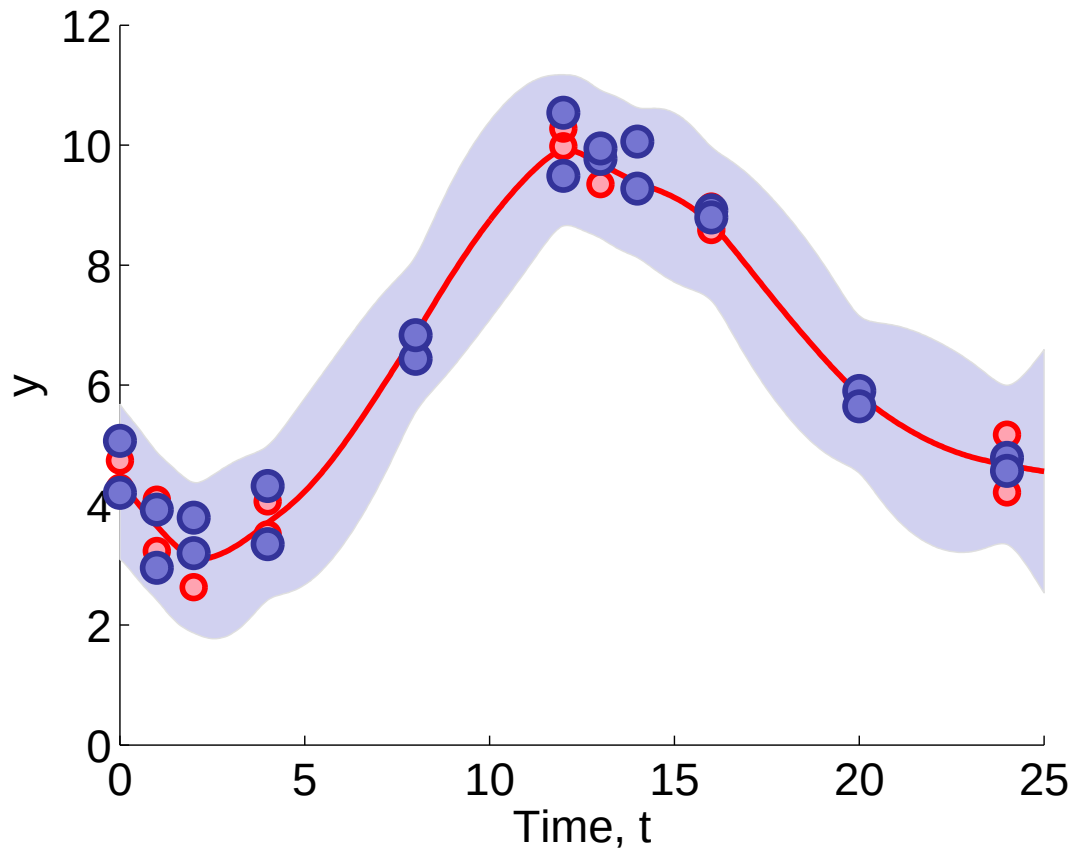
GPR bootstrapping

- First fit a GP regressor to the data to obtain a GP posterior.
- Draw bootstrap samples from resulting multivariate Gaussian.
- Infer quantity of interest for all data sets & assess variability.



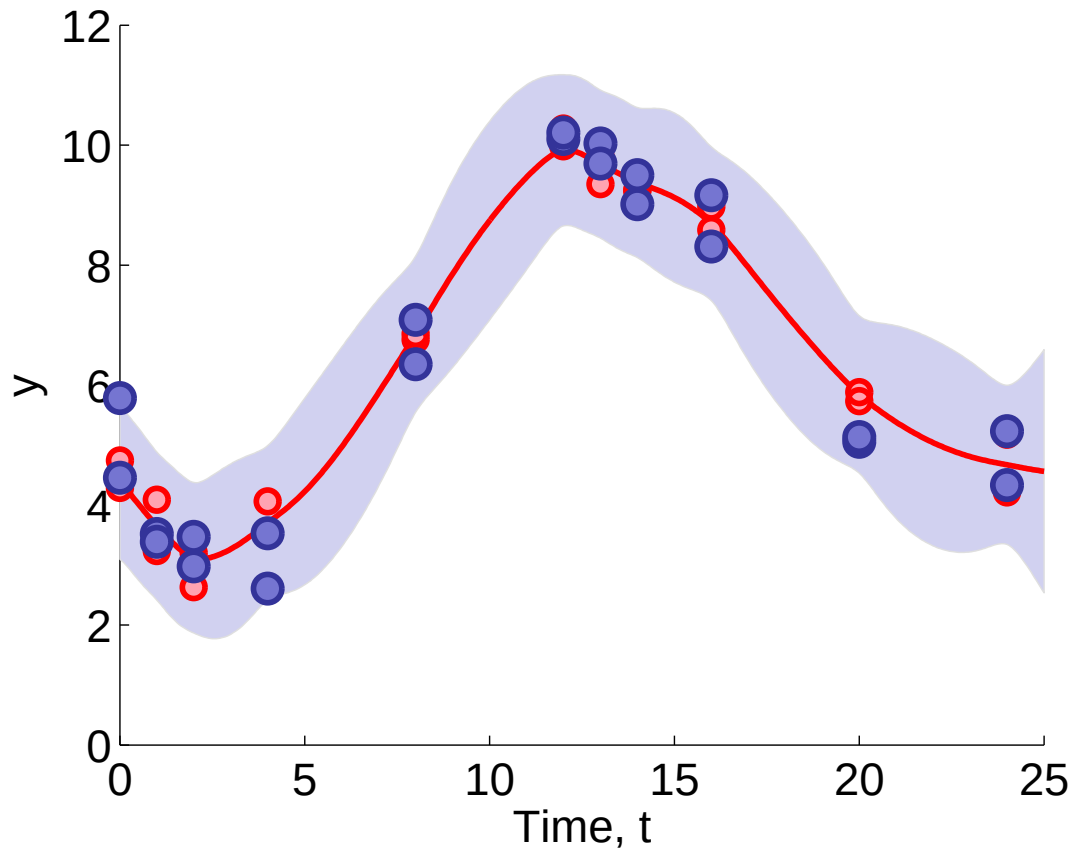
GPR bootstrapping

- First fit a GP regressor to the data to obtain a GP posterior.
- Draw bootstrap samples from resulting multivariate Gaussian.
- Infer quantity of interest for all data sets & assess variability.



GPR bootstrapping

- First fit a GP regressor to the data to obtain a GP posterior.
- Draw bootstrap samples from resulting multivariate Gaussian.
- Infer quantity of interest for all data sets & assess variability.



Example: JAK2-STAT5 Signalling Pathway

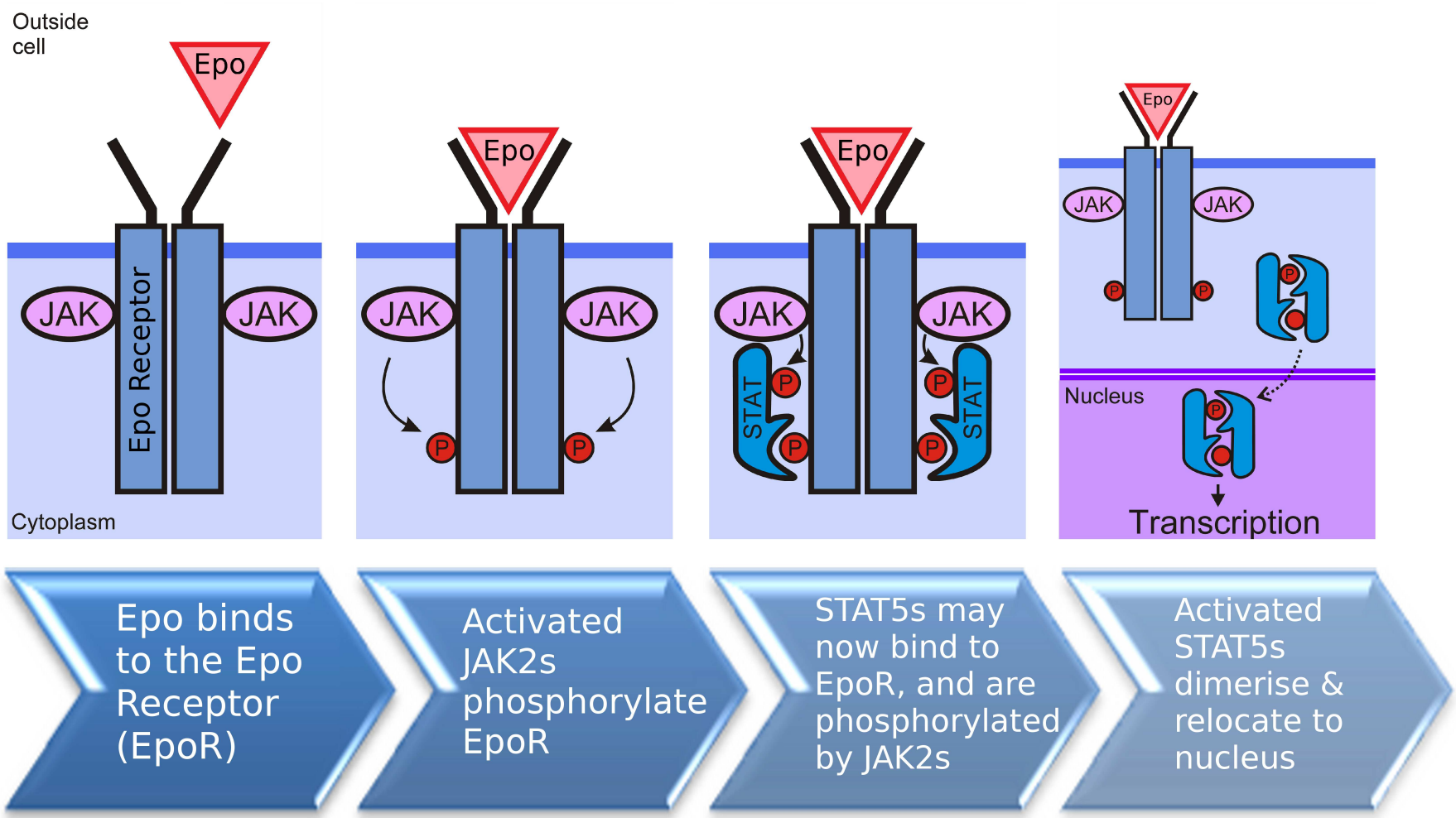


Figure: Adapted from Znamenkiy, 2006.

Example: A signalling pathway

JAK-STAT signalling pathway — The Model (Swameye et al, 2003)

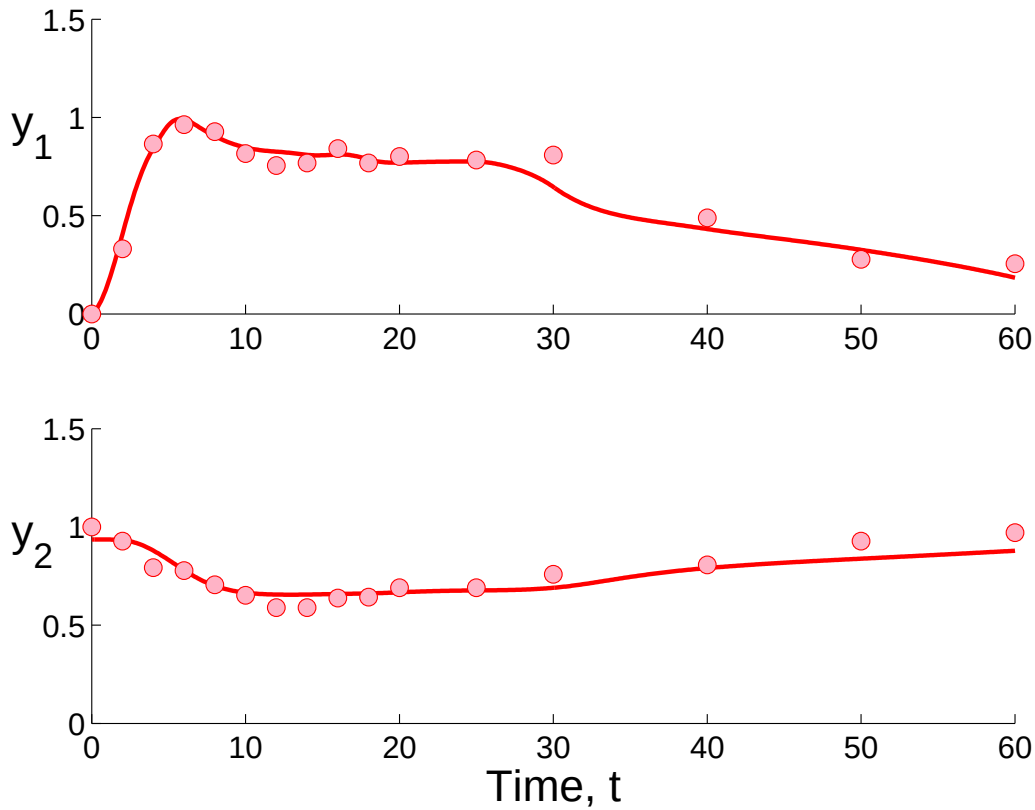
- Parametric ODE model:

$$\frac{dv_1}{dt} = -r_1 v_1 D + 2 r_4 v_4 \qquad \frac{dv_2}{dt} = r_1 v_1 D - v_2^2$$
$$\frac{dv_3}{dt} = -r_3 v_3 + 0.5 v_2^2 \qquad \frac{dv_4}{dt} = r_3 v_3 - r_4 v_4.$$
$$y_1 = r_5(v_2 + 2v_3) \qquad y_2 = r_6(v_1 + v_2 + 2v_3).$$

- v_1 — conc. unphosphorylated STAT5 in cytoplasm.
- v_2 — conc. phosphorylated monomeric STAT5 in cytoplasm.
- v_3 — conc. phosphorylated dimeric STAT5 in cytoplasm.
- v_4 — conc. STAT5 in nucleus.
- D — time-varying, experimentally determined quantity.
- r_i's — unknown parameters.

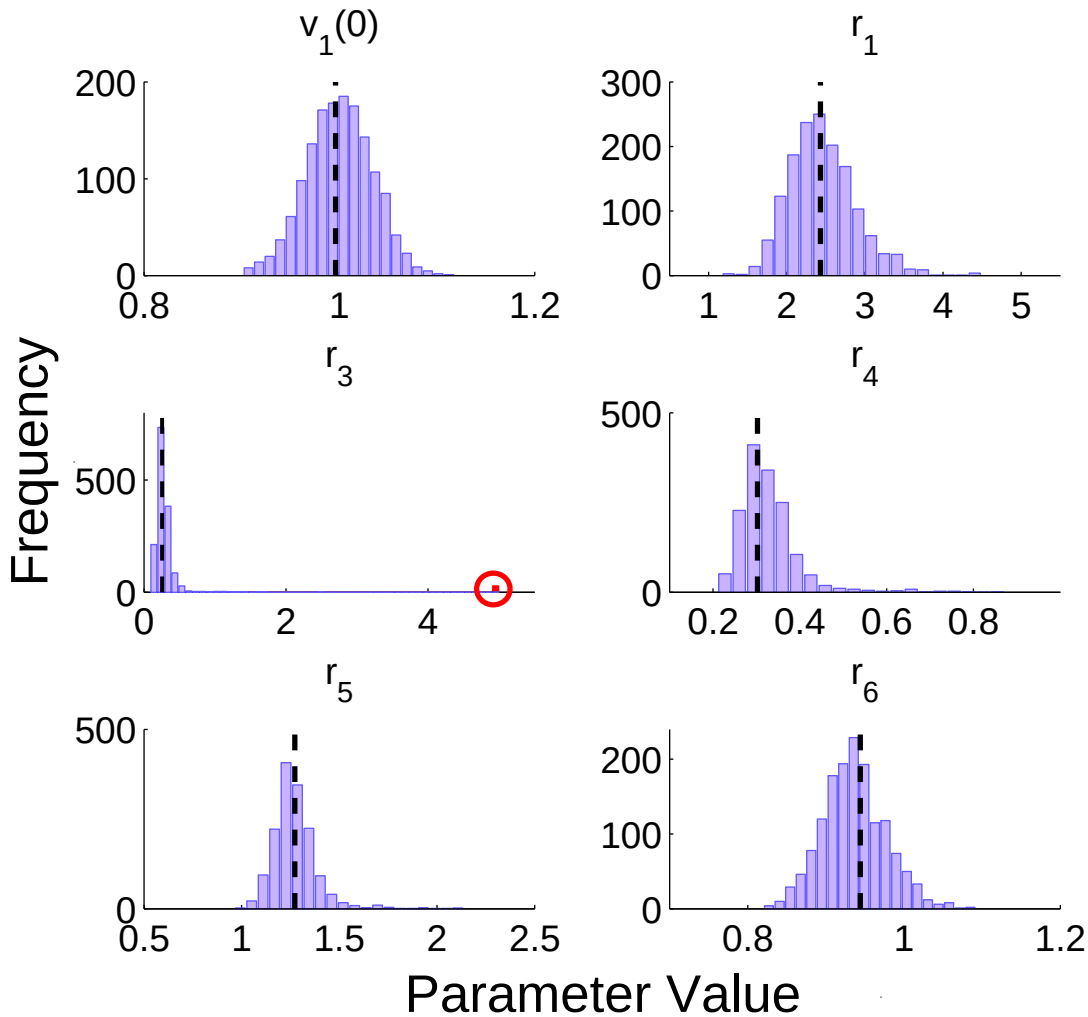
Original Data

Parameter estimates from original ("DATA1_Hall") data set:
$v_1(0) = 0.996,\ r_1 = 2.43,\ r_3 = 0.256,\ r_4 = 0.303,\ r_5 = 1.27,\ r_6 = 0.944$

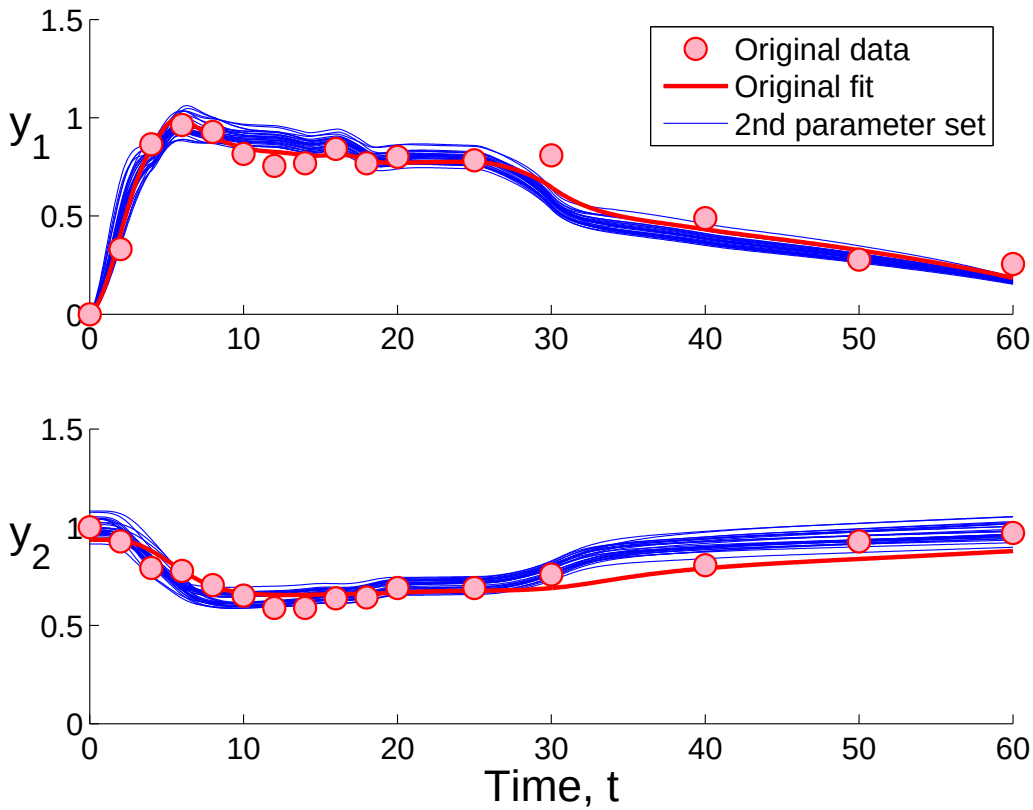


Example: A signalling pathway

Results of estimating parameters from GPR bootstrapped data:

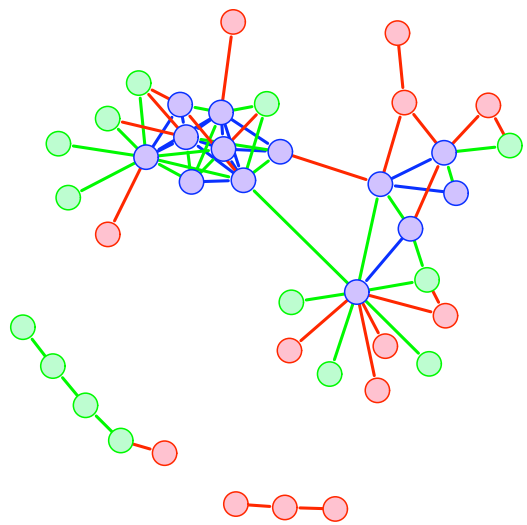


Second Parameter Set



Conclusions

- GPR can be used to bootstrap time-course data.
- JAK2-STAT5 model: identified 2nd set of plausible parameter estimates.
- Otherwise, parameter estimates relatively stable.
- Also considered gene networks: very sensitive!
  - . . . due to very high levels of noise in the data.



Acknowledgements

- Michael Stumpf and Sylvia Richardson
- The Stumpf group
- The Wellcome Trust