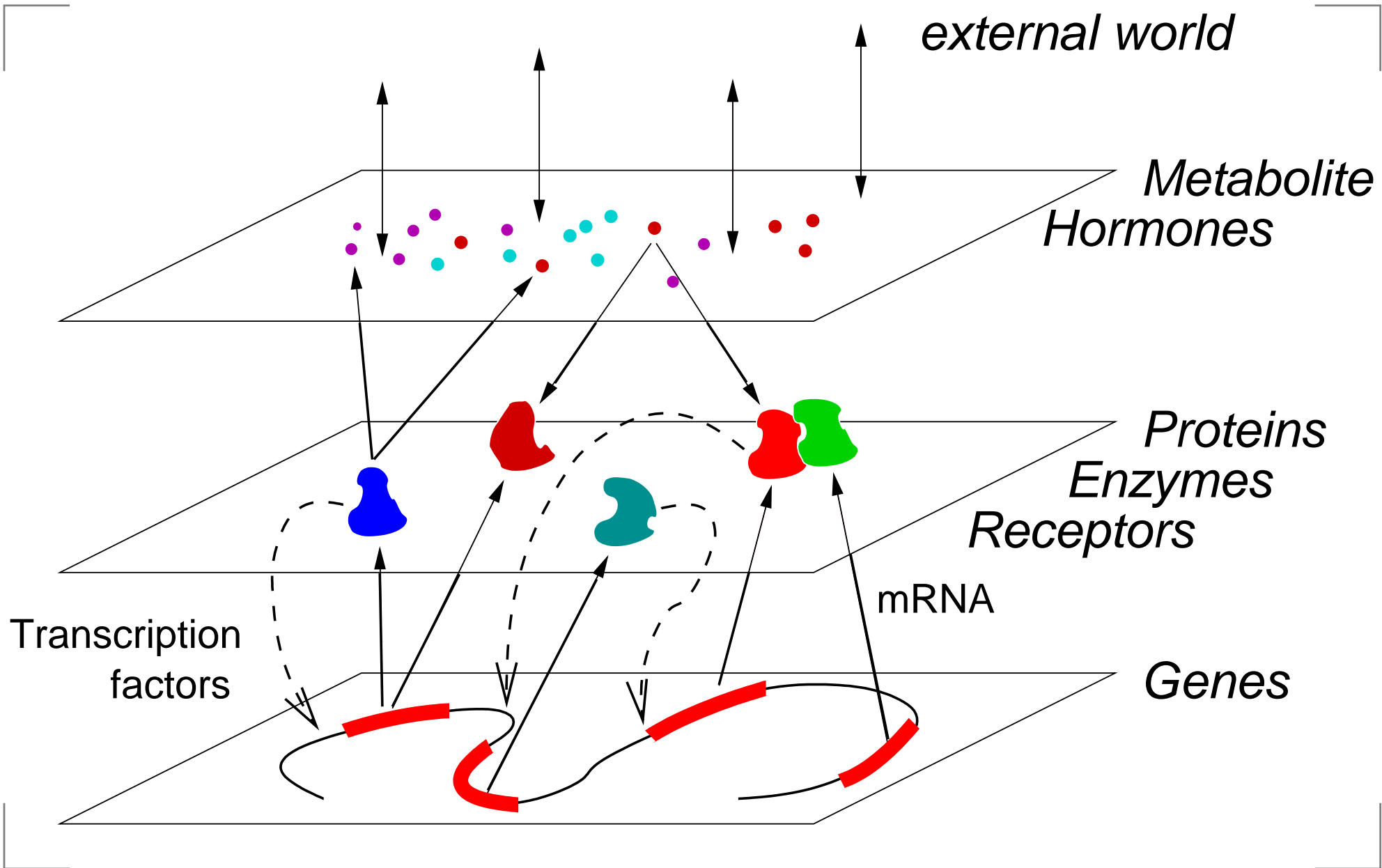


Modelling the hidden world of protein activities

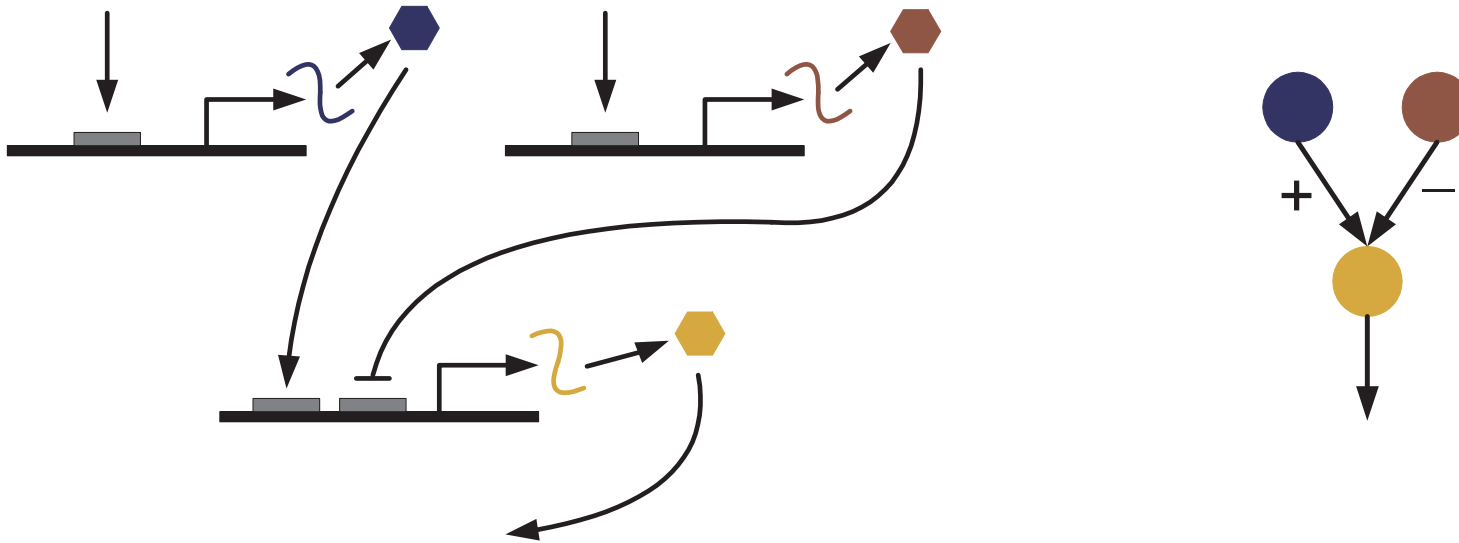
Lorenz Wernisch
with Iosifina Pournara, Yi Zhang, Nuria Domedel-Puig

MRC Biostatistics Unit
Cambridge

Cellular regulation



Gene networks

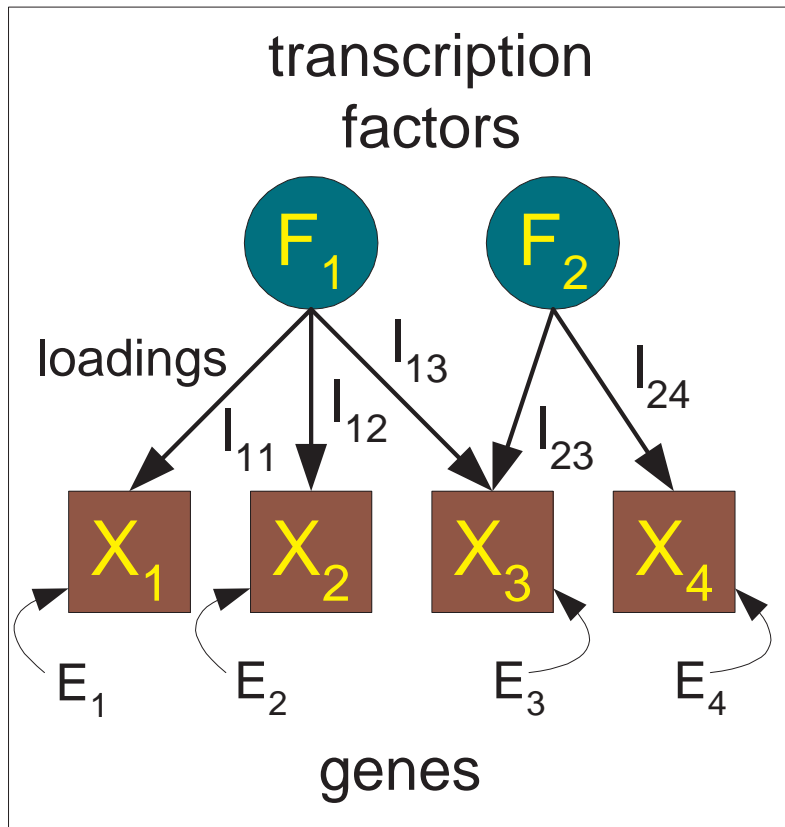


- Protein levels depend on mRNA transcription
- mRNA transcription levels depend on activation/inhibition by transcription factor proteins
- mRNA abundance does not necessarily reflect gene activity

Latent transcription factor activity

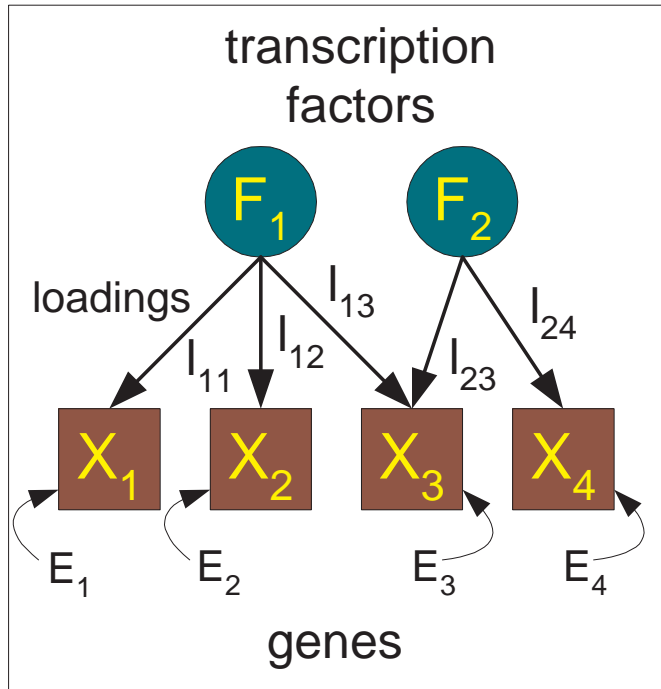
- How can protein activity be reconstructed from indirect effects on gene expression
- Of interest: not only inference of gene activity but also regulatory connectivity between hidden factors
- Options:
 - Factor analysis
 - Time-series models, Hidden Markov models
 - Differential equation models

Simple two-level model



- Each gene potentially regulated by several transcription factors (TFs)
- Each TF potentially regulates several genes
- Measurement noise (E) on gene expression (X)
- No connection between factors
- **Sparse connectivity**

Factor analysis model



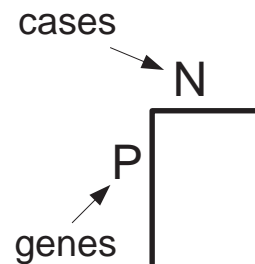
expression data

connectivity matrix

transcription factors matrix

noise

$$\mathbf{X} = \mathbf{L} \mathbf{F} + \mathbf{E}$$



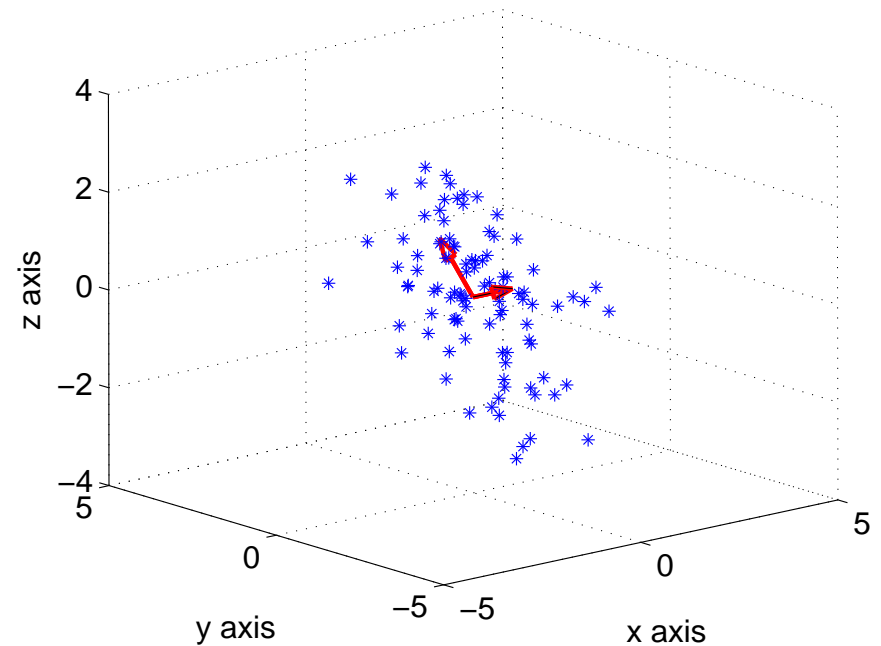
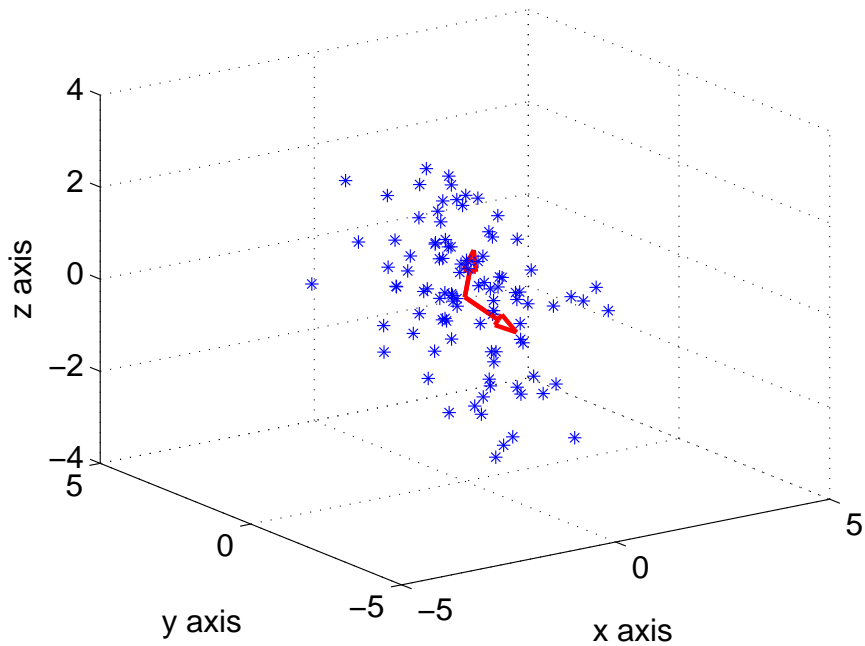
$$= \begin{matrix} \text{P} \\ \text{?} \end{matrix} \begin{matrix} \text{K} \\ \text{?} \end{matrix} + \begin{matrix} \text{N} \\ \text{?} \end{matrix}$$

TFs

	TFs					
	0	0	1	0	0	0
G	0	1	0	0	1	0
e	0	0	0	0	0	1
n	0	1	0	0	0	0
e	1	0	0	0	0	0
s	0	0	0	1	0	1
	0	0	0	0	0	1
	0	1	0	0	0	0

$$\begin{aligned} x_1 &= l_{11}f_1 + \epsilon_1 \\ x_2 &= l_{12}f_1 + \epsilon_2 \\ x_3 &= l_{13}f_1 + l_{23}f_2 + \epsilon_3 \\ x_4 &= l_{14}f_2 + \epsilon_4 \end{aligned}$$

Problem of factor analysis



Factors determined
up to rotation

Sparse loadings
matrix?

.7	.7
-.7	.7
-.7	.7

1	0
0	1
0	1

Network component analysis NCA

Minimize distance between data X and decomposition (Liao et al., 2003):

$$\min_{\Lambda F} \|X - \Lambda F\|^2$$

Constraints on non-zero entries of loadings Λ necessary for unique decomposition of X :

- Λ must have full column rank
- Each column of Λ must have at least $K - 1$ zeros

In practice: almost complete network needs to be known in advance

Classical ML estimation

- Numerical maximization of likelihood
- Sparsity on the loadings matrix as separate step:
 - **Varimax** (each factor is associated with a small number of genes)
 - **Quartimax** (each gene is regulated by a small number of factors)
 - **Equimax** (something in between)
 - **tanh** (add as many zeros as possible)
 - **Procrustes** best rotation that matches MLE factors to true factors **if they are known**

Bayesian factor analysis

- Prior distributions on parameters
- Sample alternately from posterior of factors and loadings
- Allows sparsity to be incorporated by using sparsity priors on the loadings
- Parameters can be summarised by averaging over their posterior distributions

Sparsity priors on loadings

- A mixture prior on L

$$l_{pk} | z_{pk} = 0 \sim \delta_0(l_{pk})$$

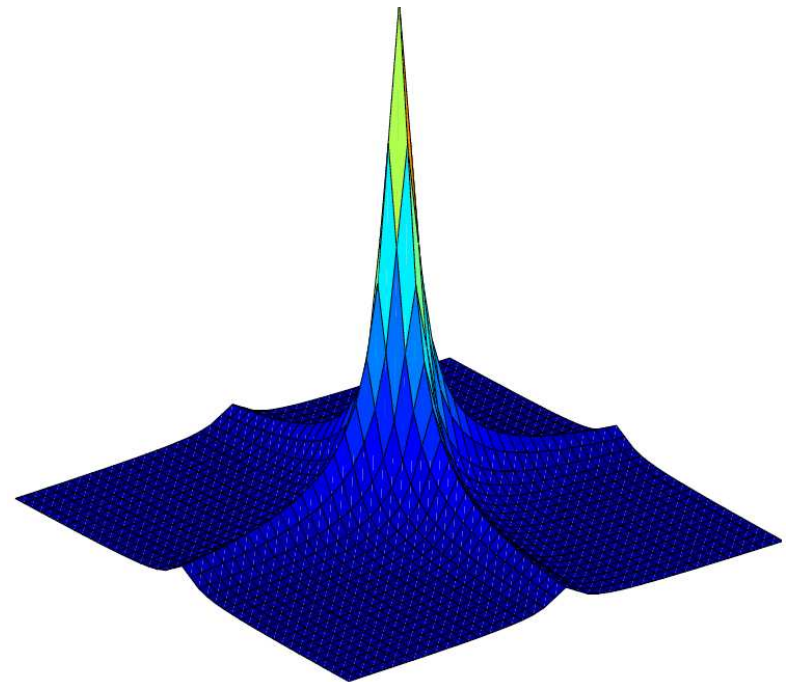
$$l_{pk} | z_{pk} = 1 \sim \mathcal{N}(0, \sigma^2)$$

$$z_{pk} \sim \text{Bern}(\pi)$$

- A Gamma prior on L

$$l_{pk} \sim \mathcal{N}(0, \delta_{pk}^{-1})$$

$$\delta_{pk} \sim \mathcal{G}(a, b)$$

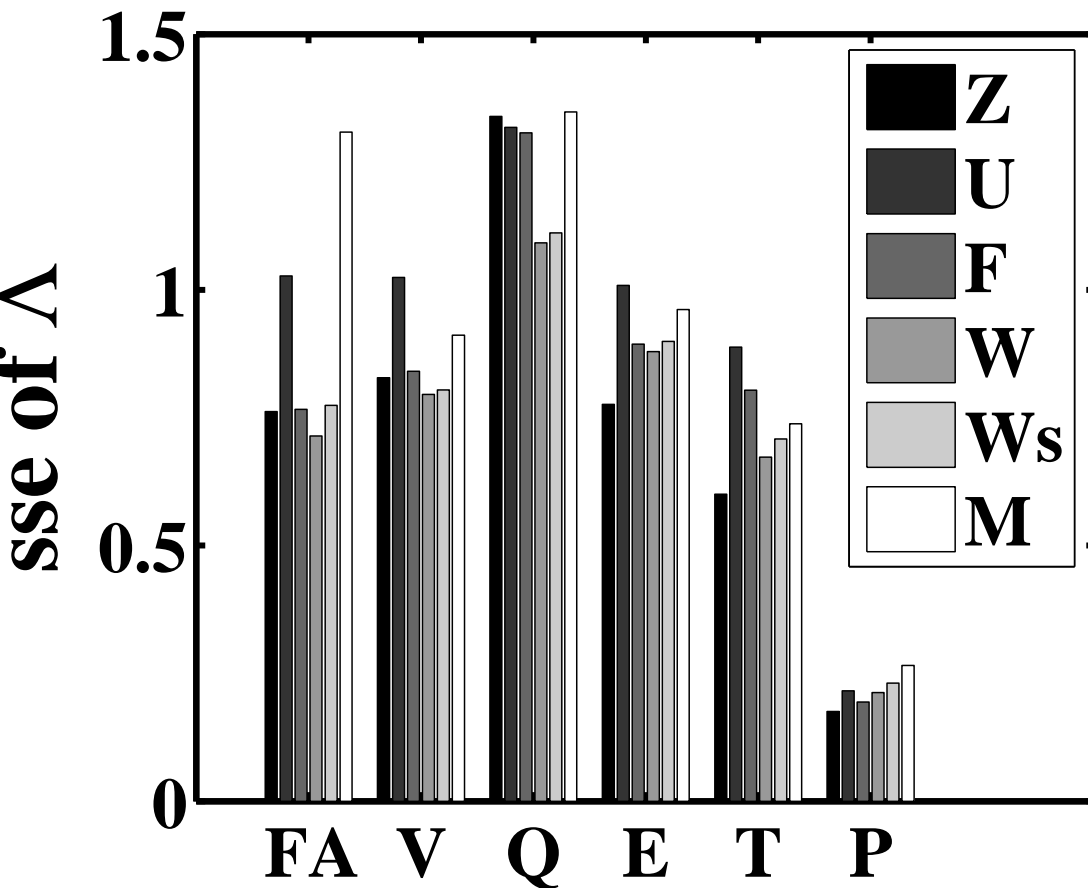


Methods for sparse FA

Factor and loadings matrix determined up to rotation:
identifiable by sparsity requirement

- (F) E Fokoue (2004): Gamma priors on precision
- (W) M West (2003): Sparsity (Bernoulli) priors on connectivity: $l_{ij} \neq 0 \sim \text{Bern}(p)$, $p \ll 1$
- (M) Classical FA (varimax, quartimax)
- (U) A Utsugi, T Kumagai (2001): Gibbs sampling for mixture of FAs, Gamma priors
- (Z) Z Ghahramani, G Hinton (1997): EM algorithm for mixture of FAs, Gamma priors

Simulation: SSE after rotation



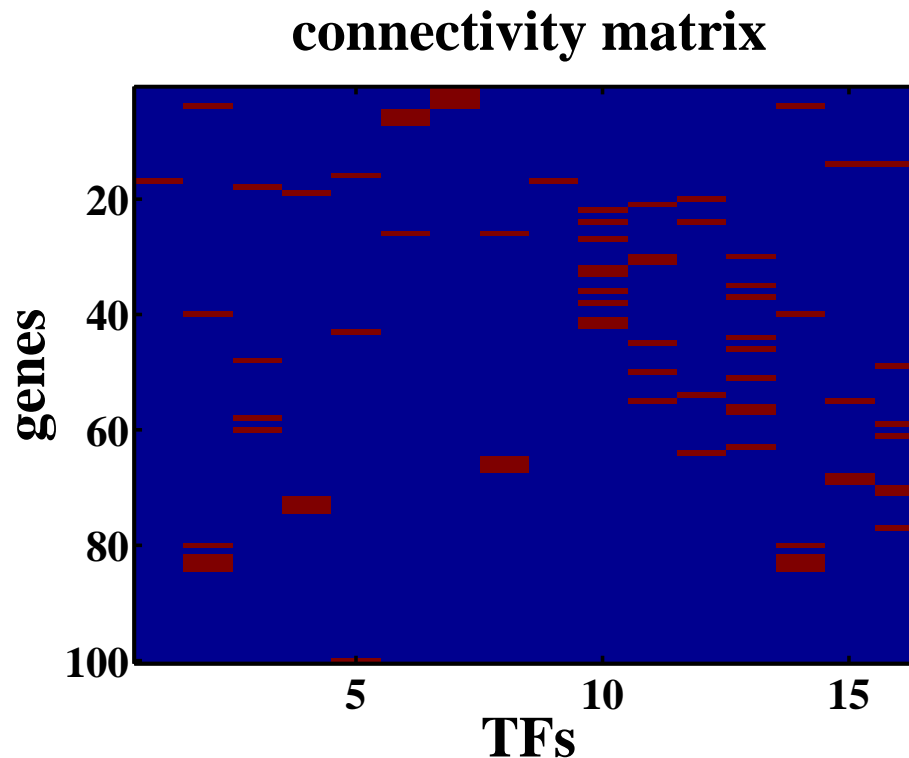
Loadings
reconstruction

rotation methods:
varimax V,
quartimax Q,
equamax E, tanh T,
procrustes P

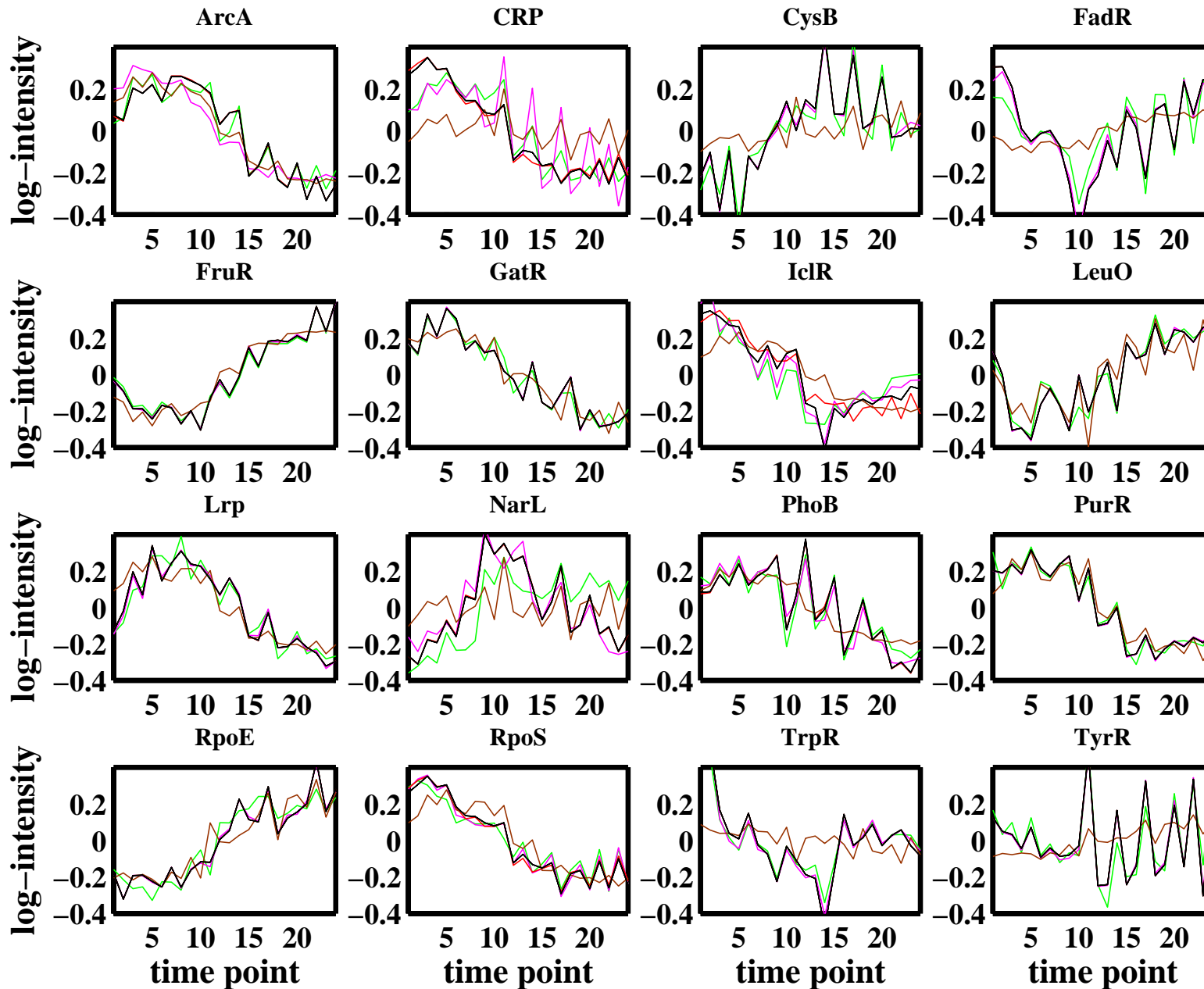
Bayesian methods slightly better than classical FA
Careful rotation (tanh) can improve results
Factor structure correct (up to rotation)

E. coli dataset

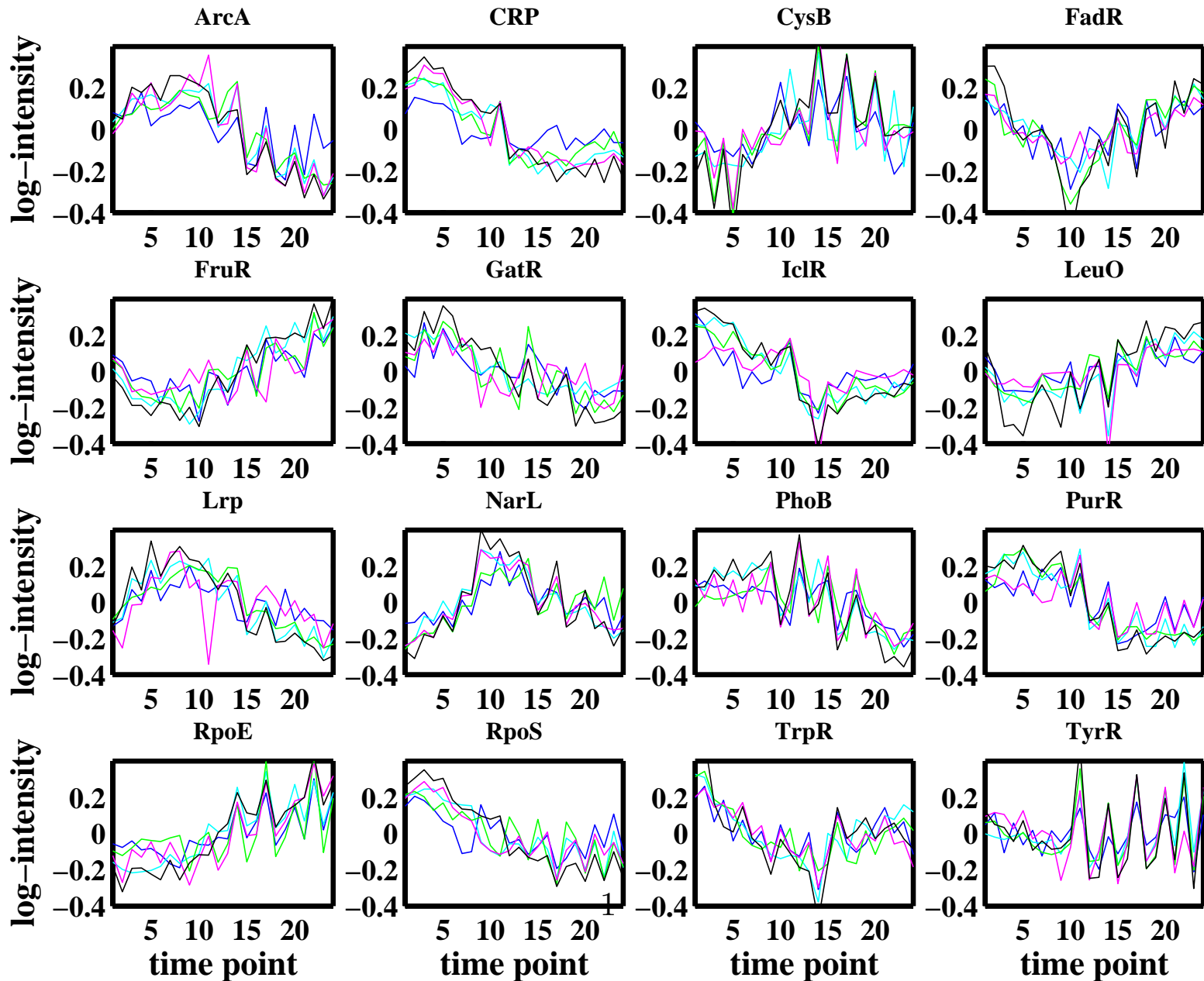
- Microarray time series data (23 time points) 100 genes, 16 putative transcription factors (Kao et al., PNAS 2003)
- Connectivity matrix based on RegulonDB and literature



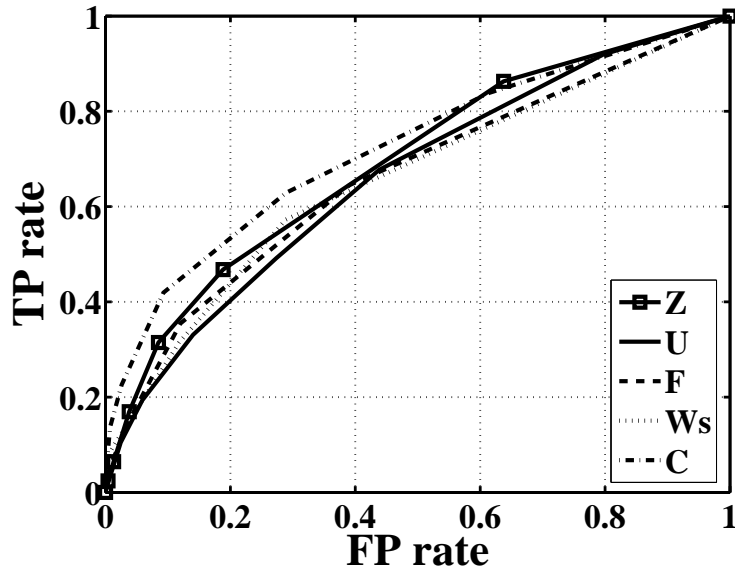
E. coli TFs given connectivity matrix



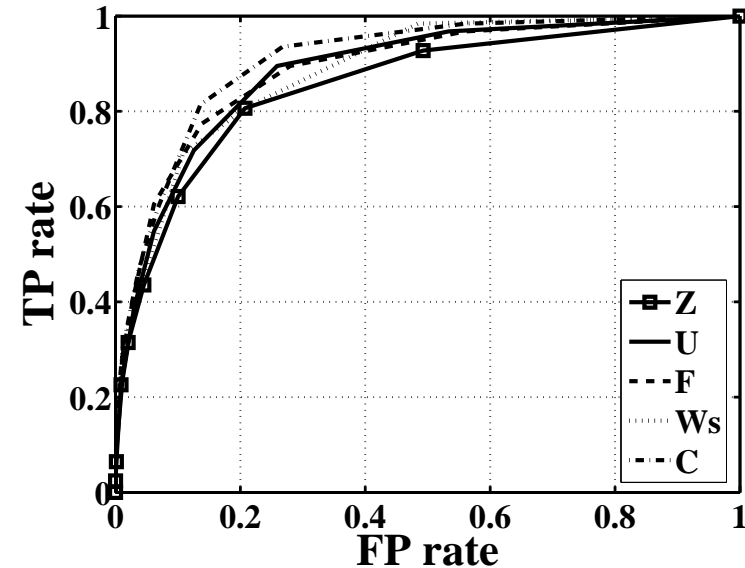
E. coli TFs without connectivity matrix



ROC for *E. coli* loadings (connectivity)



(a)



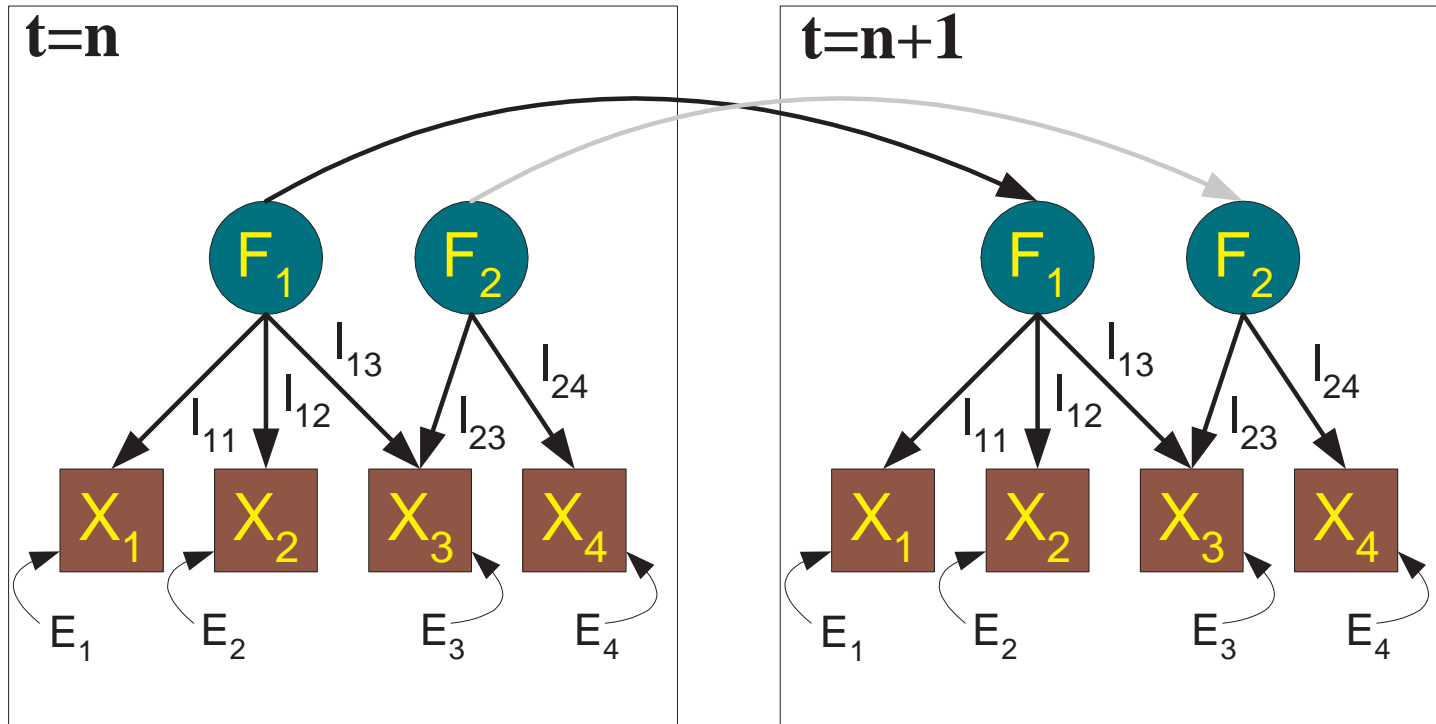
(b)

Vary cutoff for score in loadings matrix to find connection between gene and regulator

Combination of all methods improves result

Procrustes rotation (to true factors) shows (right): factors correct, but rotation problematic

Time-series factor analysis



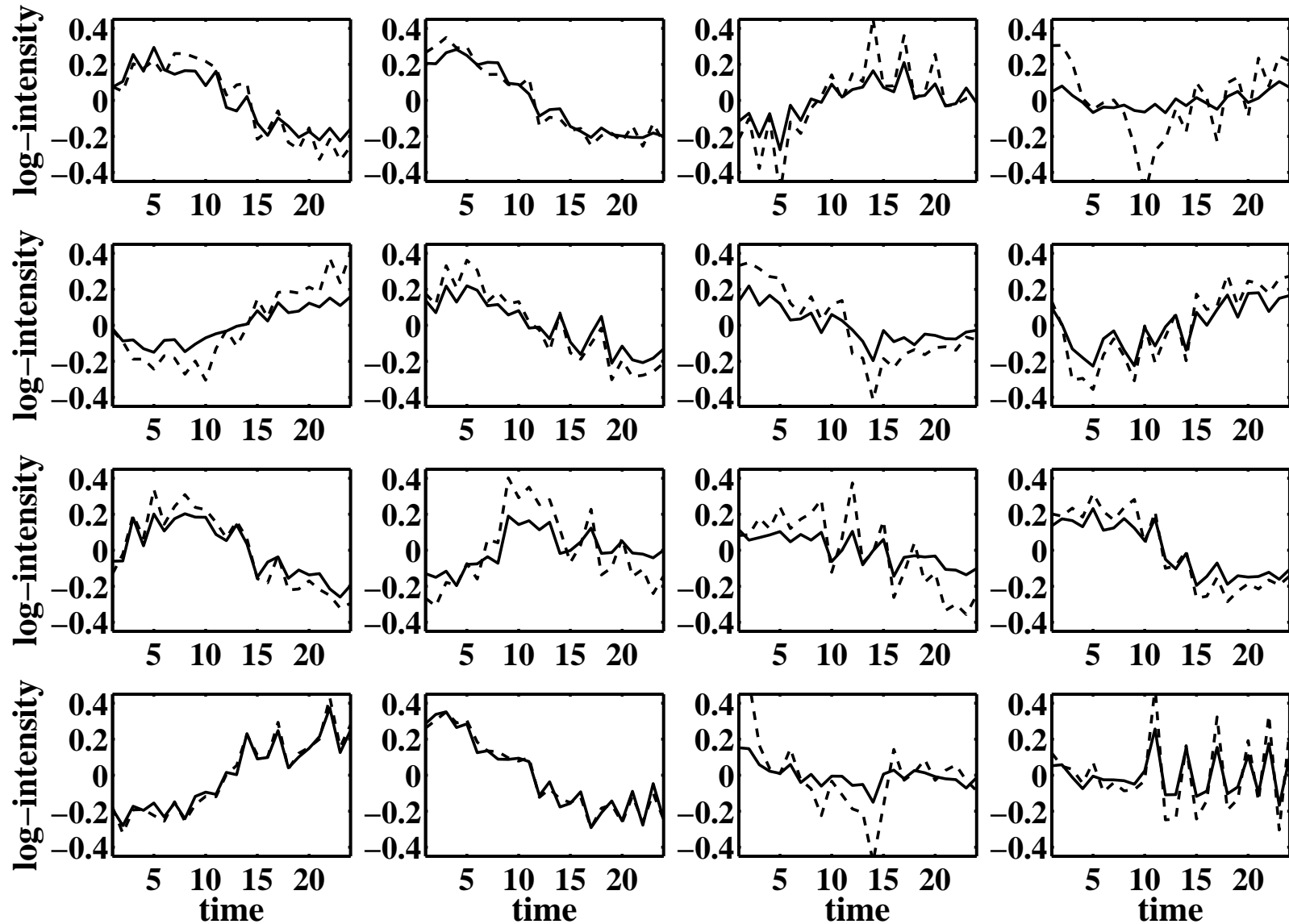
Correlation between time points

Correlation ρ between time points induces a correlation structure across all times:

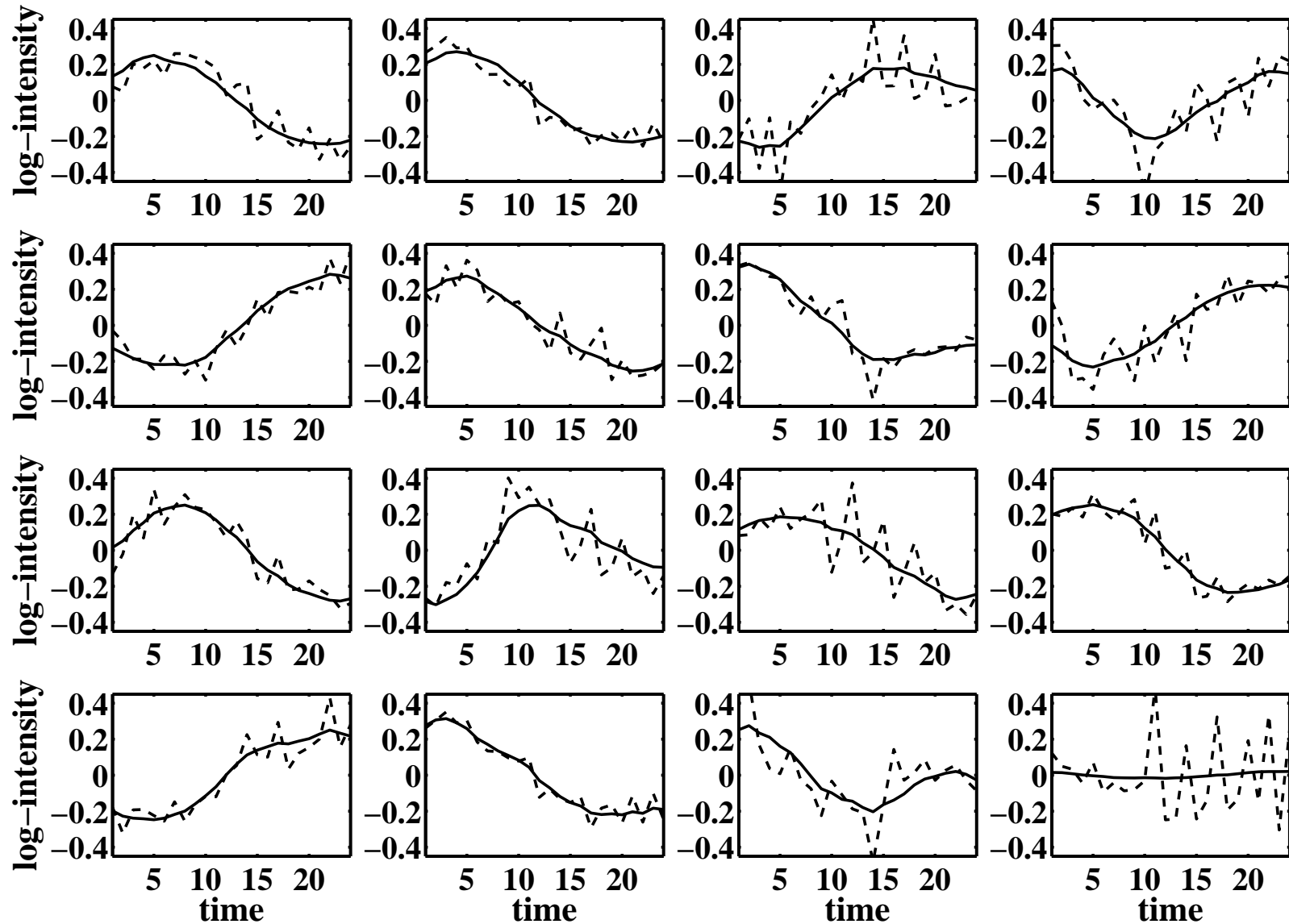
$$\begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{N-1} \\ \rho & 1 & \rho & \dots & \rho^{N-2} \\ \rho^2 & \rho & 1 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{N-1} & \rho^{N-2} & \dots & \dots & 1 \end{pmatrix}$$

Impose time correlation on factors in Bayesian inference, estimate ρ as well

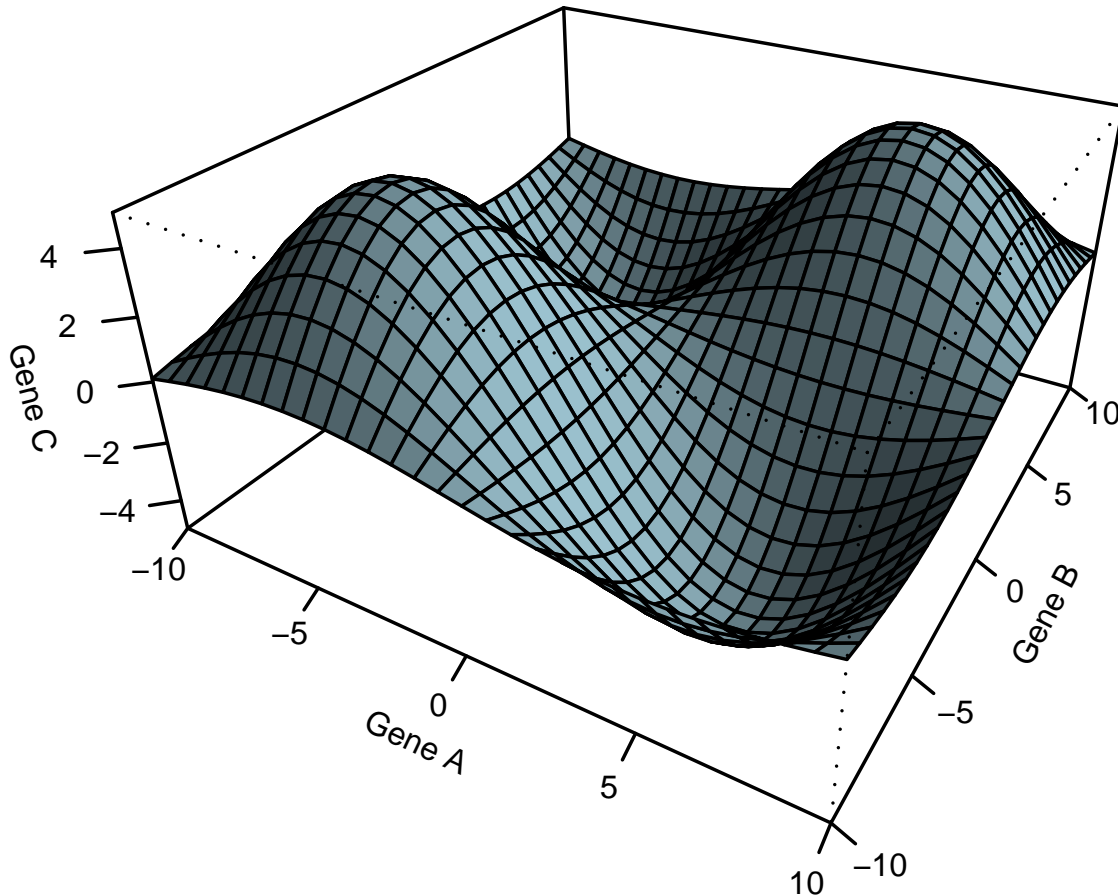
E. coli TFs without time correlation



E. coli TFs with time correlation



Nonlinear dependencies

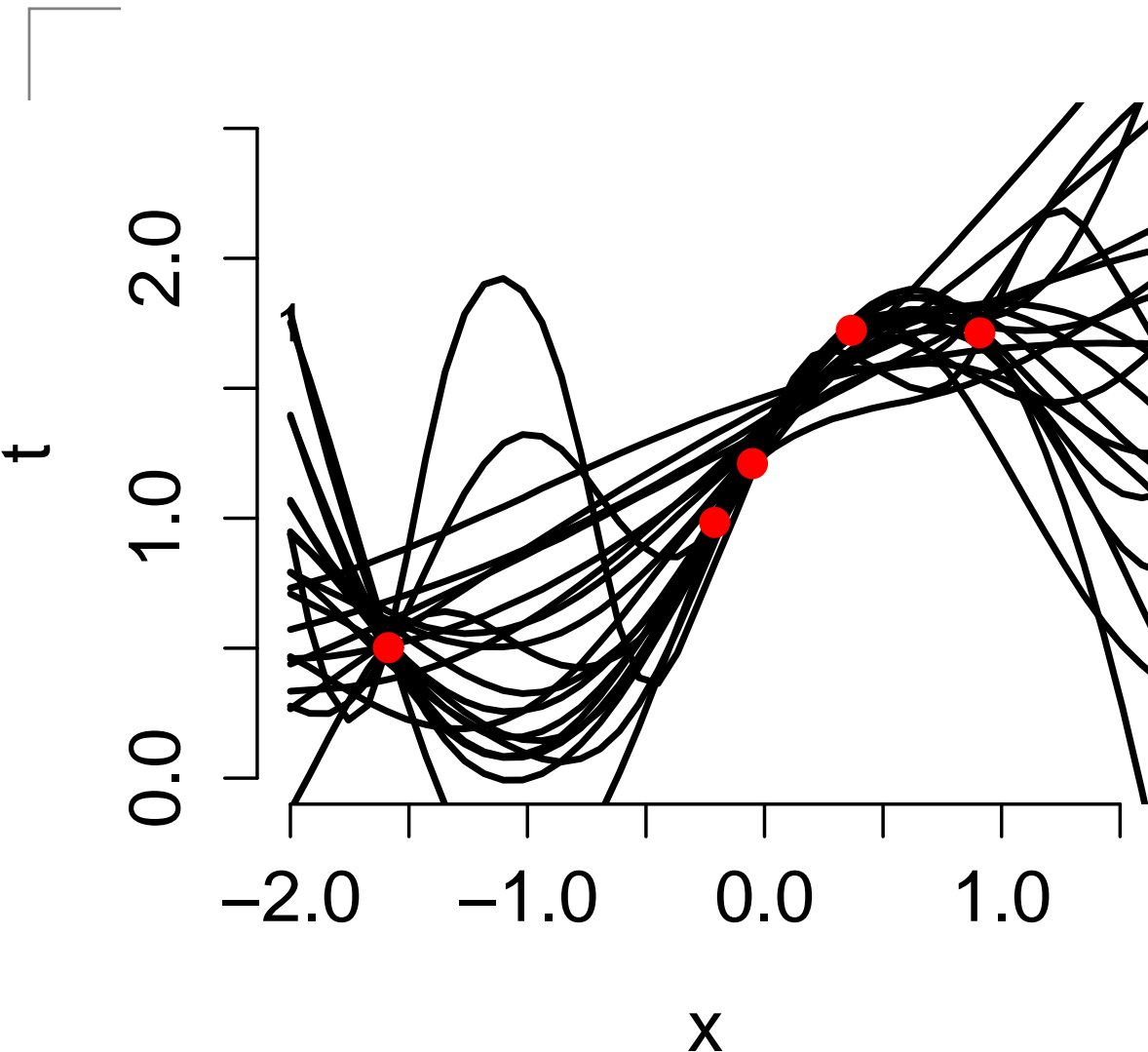


Assumed **linear** dependencies of level of gene A on other gene levels

Genes often operate as switches and complex gates with nonlinear interactions (eg exclusive or)

Nonlinear models: differential equations, Bayesian spline models, Gaussian processes (GPs)

Nonparametric methods



We don't want to exclude any of these curves a priori

Need some measure of adequacy:

- Curve should fit reasonably well
- It should be defined by highly flexible model family

Solution: family defined by **covariance structure**

Gaussian process

- N input vectors $x^{(1)}, \dots, x^{(N)}, x^{(p)} \in \mathbb{R}^d$
- Target values $t = (t^{(1)}, \dots, t^{(N)}), t^{(p)} \in \mathbb{R}$
- Joint distribution of the output t is multivariate Gaussian $N(0, K)$
- Covariance matrix K

$$K_{pq} = \beta_0 + C_L(x^{(p)}, x^{(q)}) + C_G(x^{(p)}, x^{(q)}) + \sigma_\epsilon^2 I(p = q)$$

β_0 overall constant

σ_ϵ^2 noise term along diagonal of K

$I()$ indicator function

Covariance components

Linear covariance part

$$C_L(x^{(p)}, x^{(q)}) = \sum_k \beta_k x_k^{(p)} x_k^{(q)}$$

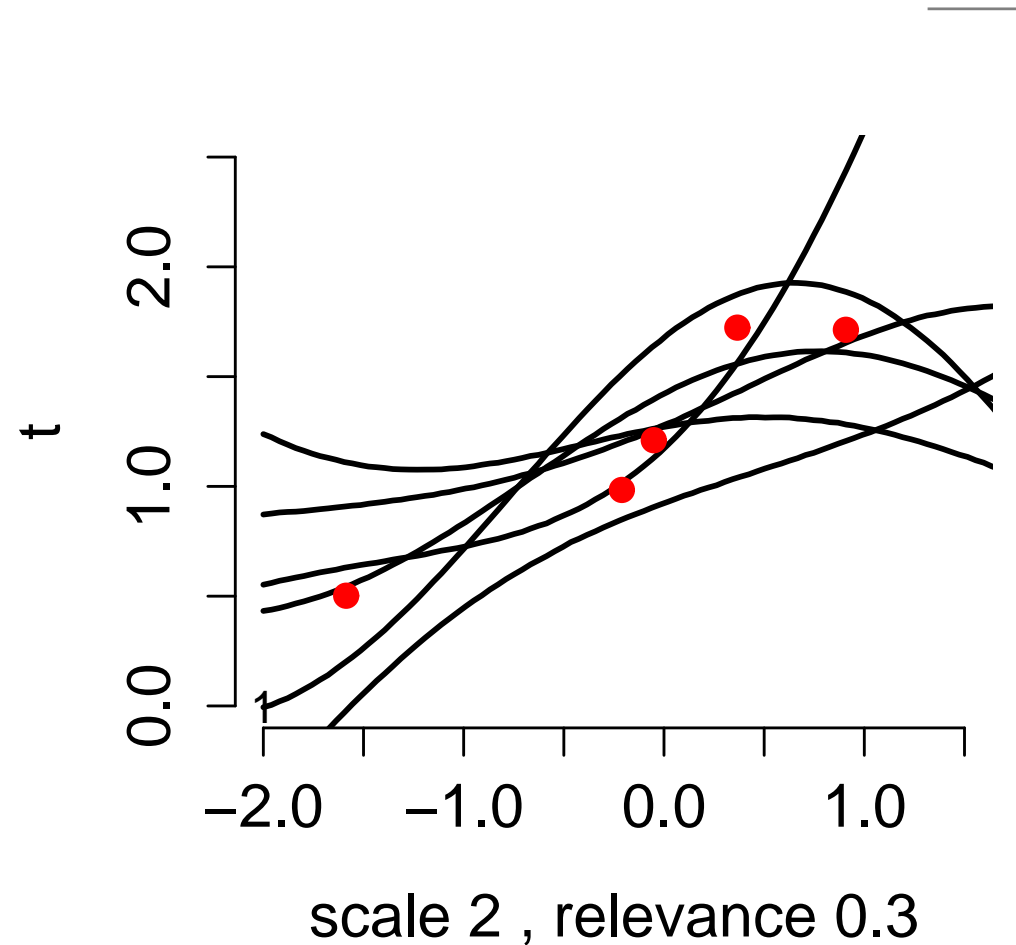
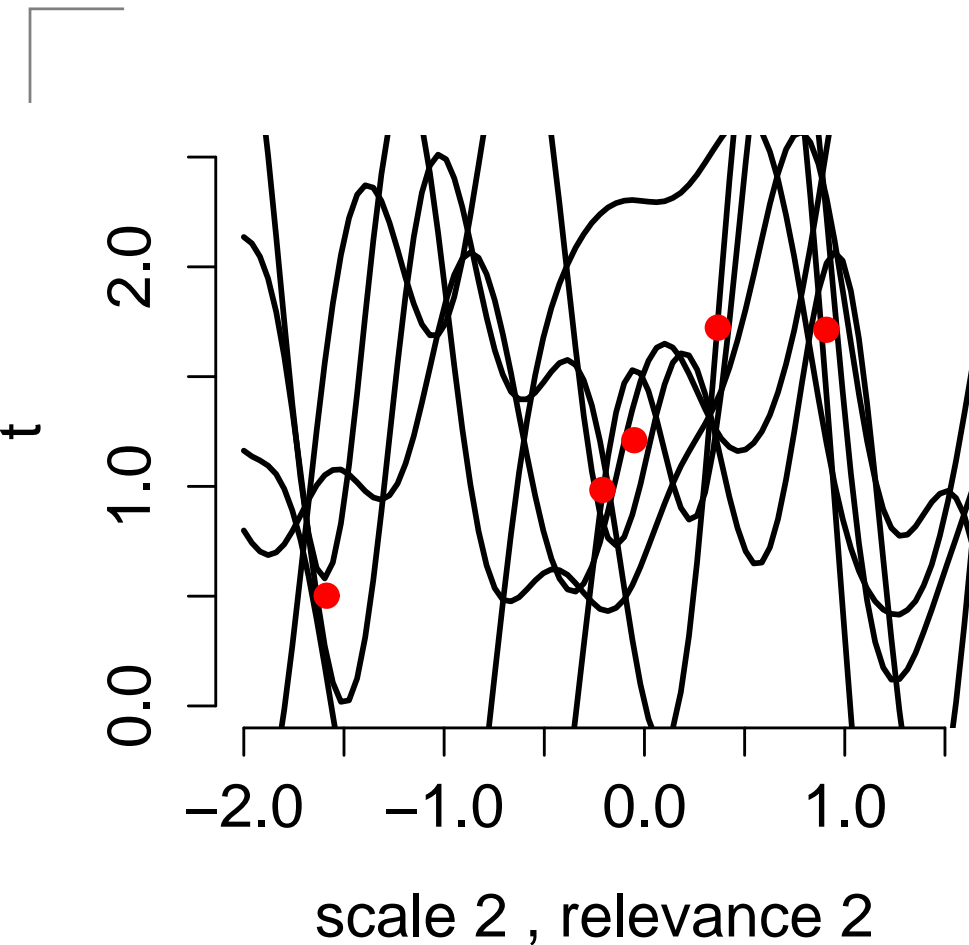
with **linear relevance parameters** β_1, \dots, β_d

Squared exponential (Gaussian) covariance part

$$C_G(x^{(p)}, x^{(q)}) = \alpha_0 \exp\left(-\sum_k \alpha_k (x_k^{(p)} - x_k^{(q)})^2\right)$$

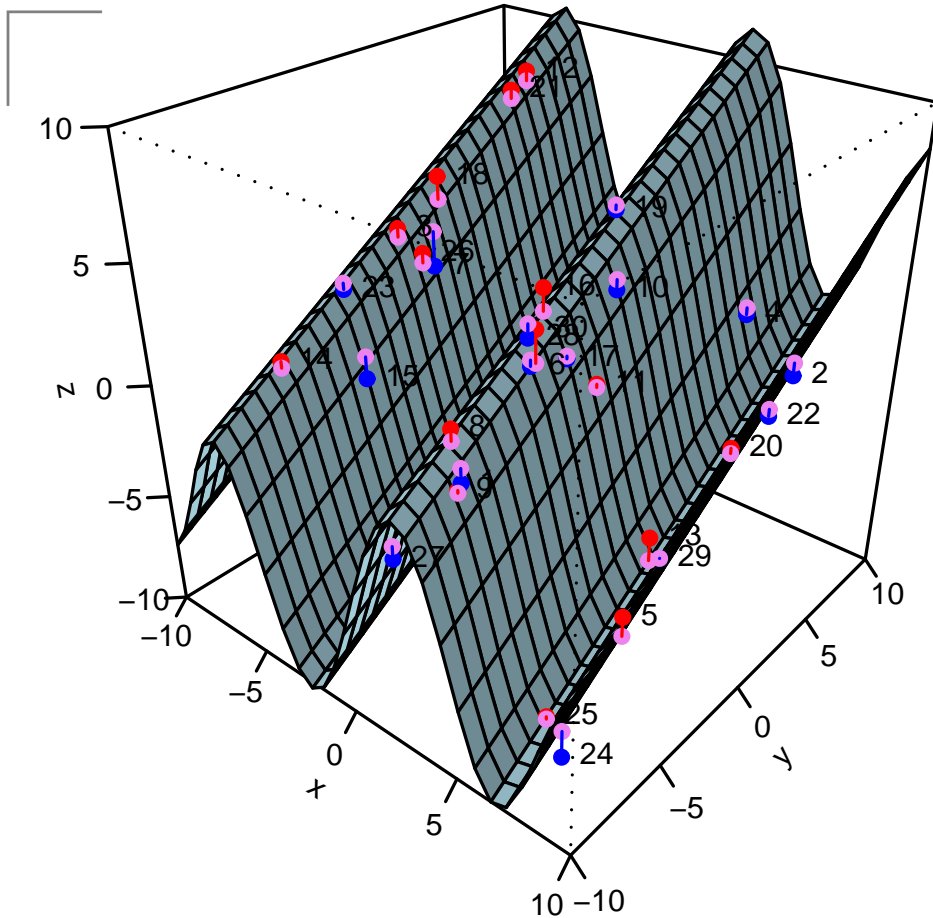
with **nonlinear relevance parameters** $\alpha_1, \dots, \alpha_d$
and scale parameter α_0

Relevance parameters



Relevance small: corresponding input less influential!

GP on simulated data



Relevance parameters:

	x_1	x_2	x_3
nonlinear	0.21	0	0
linear	0	0.35	0

estimated sd 0.92

30 data points with $f(x_1, x_2, x_3) = 5 \sin(0.7x_1) + 0.5x_2 + \epsilon$
where $\epsilon \sim N(0, 1)$

Dynamical model using GPs

$$x_i^t = f(x_i^{t-1}, \theta_x) + \epsilon_{x,t}$$

$$y_g^t = h(x^t, \theta_{y,g}) + \epsilon_{y,t}$$

x_i^t : hidden gene activity at time t

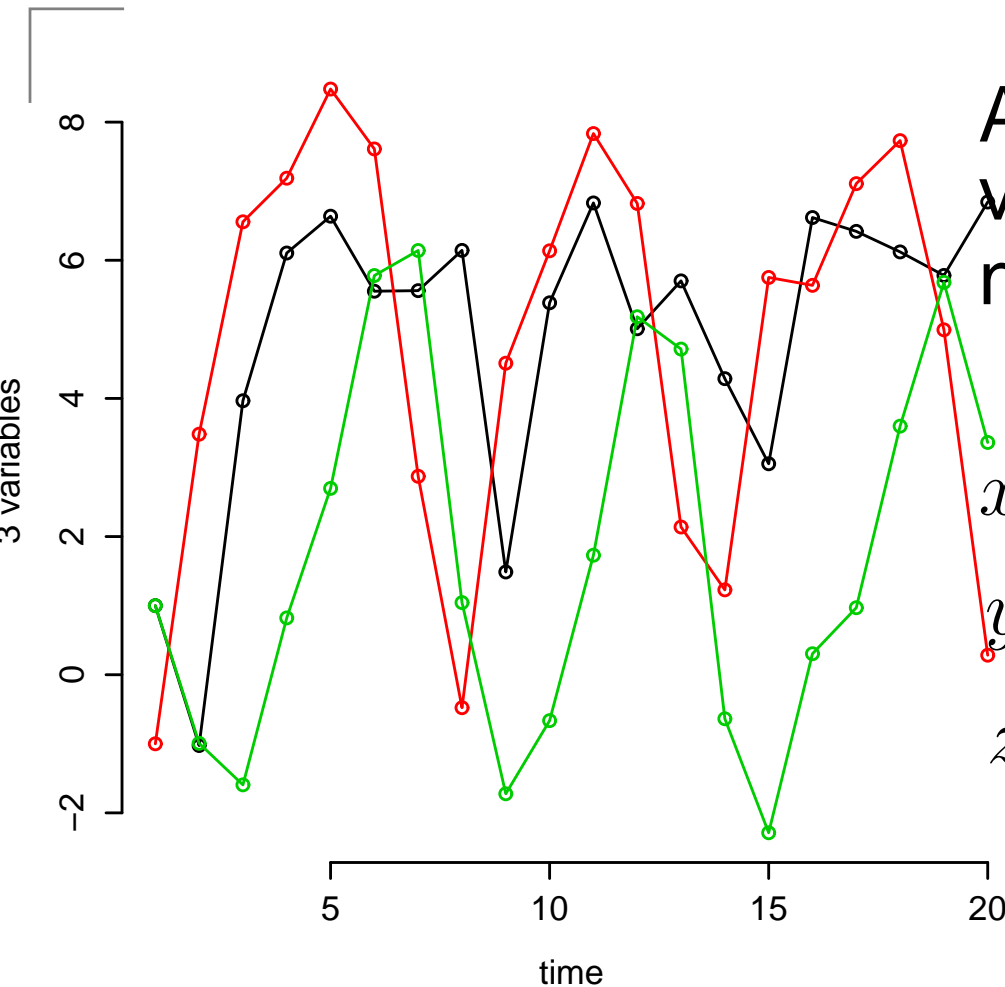
y_g^t : expression level of gene g at t

f : describes dynamics, modeled by GP

h : describes measurement, modeled by GP

Joint selection of θ and x to maximize likelihood

GP on simulated time-series data



Artificial network of 3 variables connected by nonlinear relationships

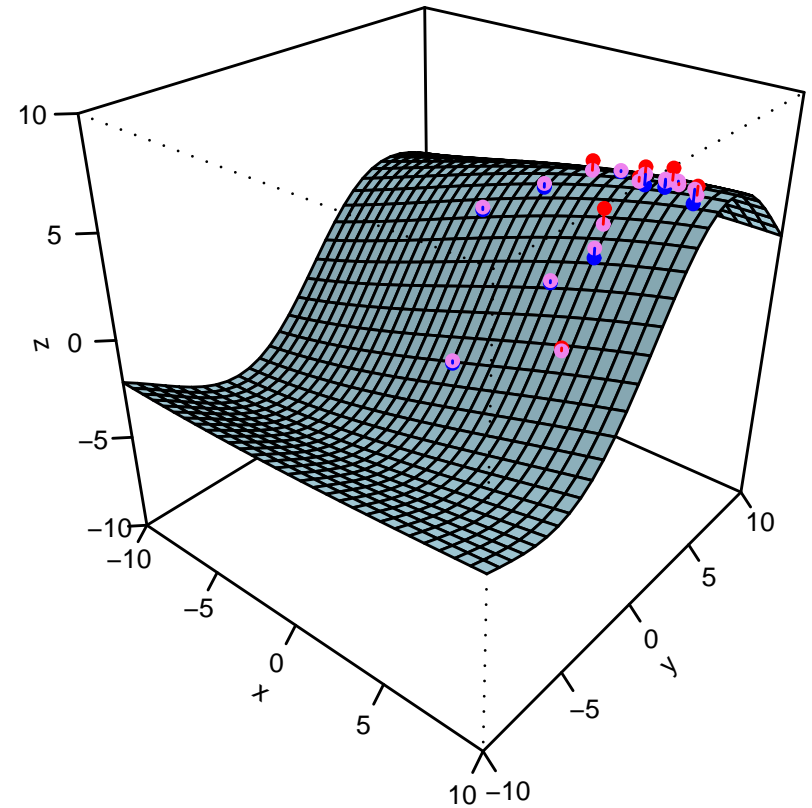
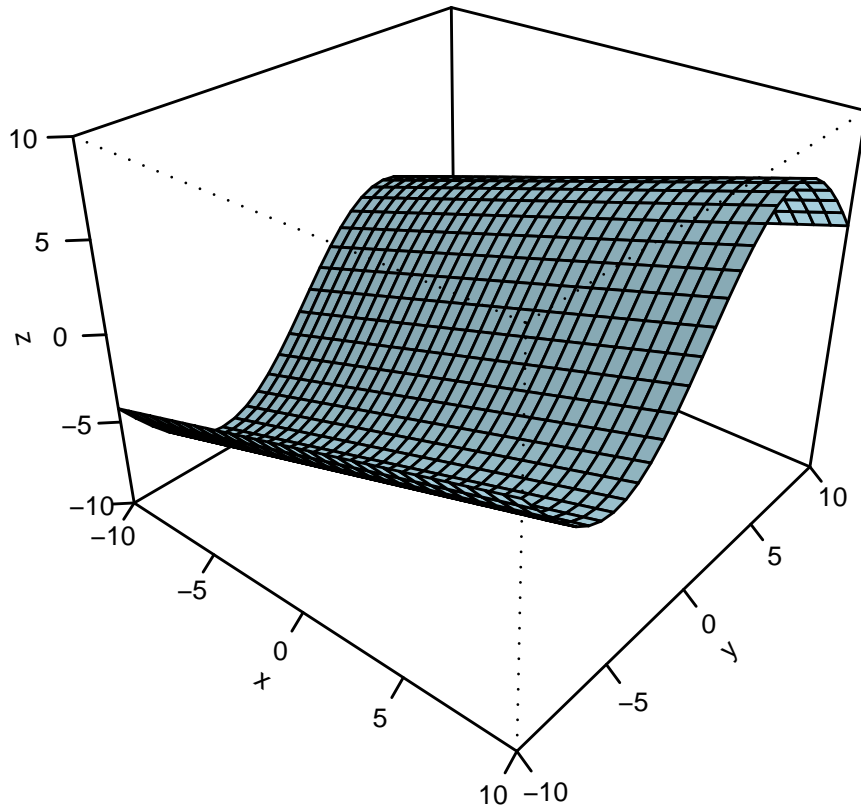
$$x_{t+1} = 0.35x_t + 5 \sin(0.3y_t) + \epsilon_1$$

$$y_{t+1} = 0.4y_t + 5 \cos(0.3z_t) + \epsilon_2$$

$$z_{t+1} = 0.4z_t + 0.1y_t^2 - 2 + \epsilon_3$$

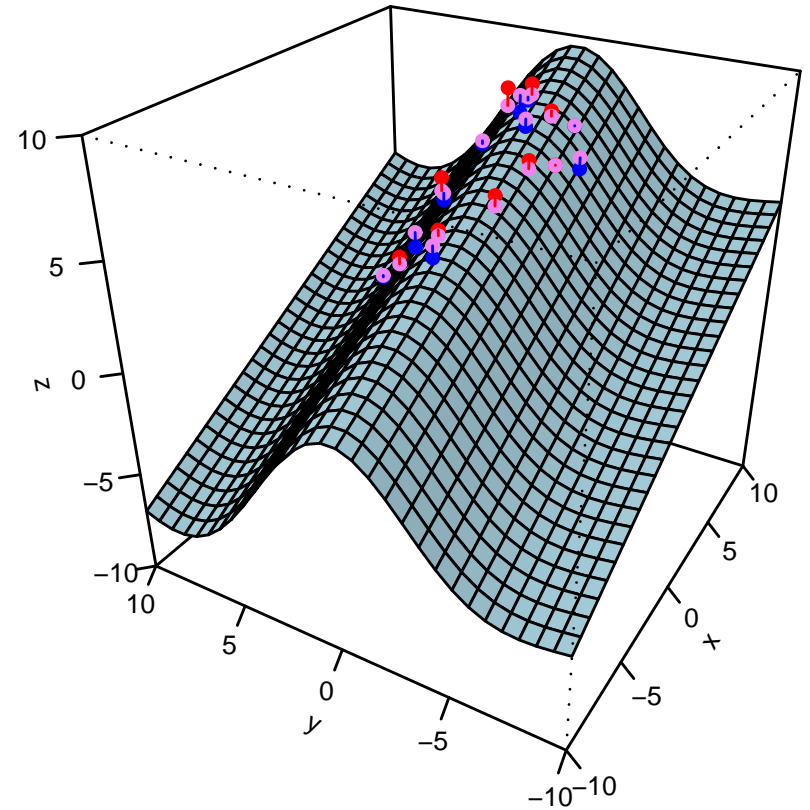
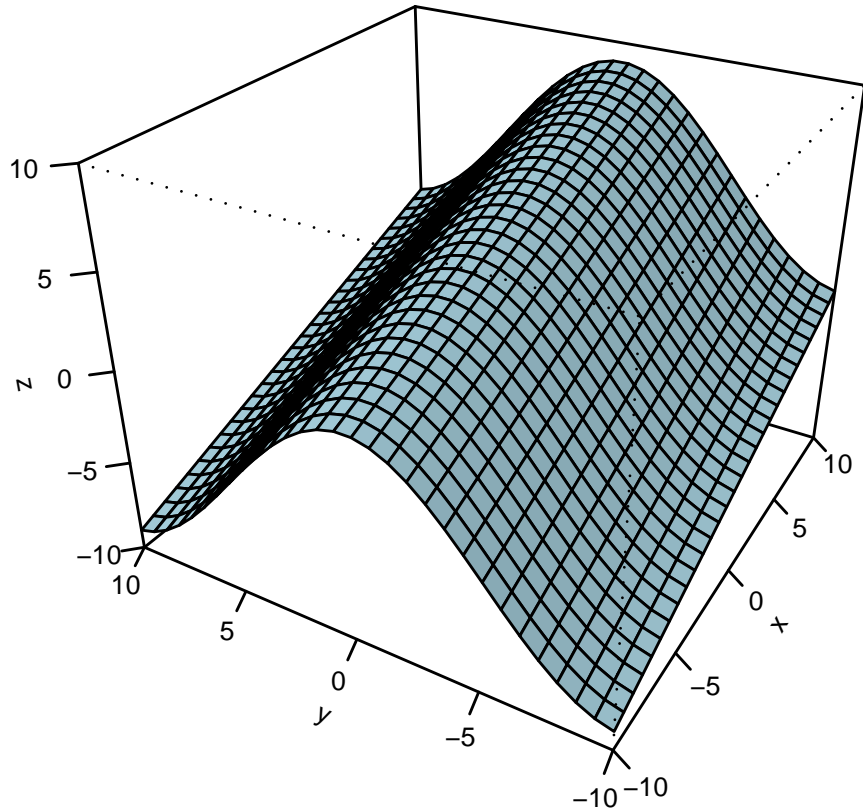
Stable cycling easy to achieve with nonlinear networks

GP on time-series



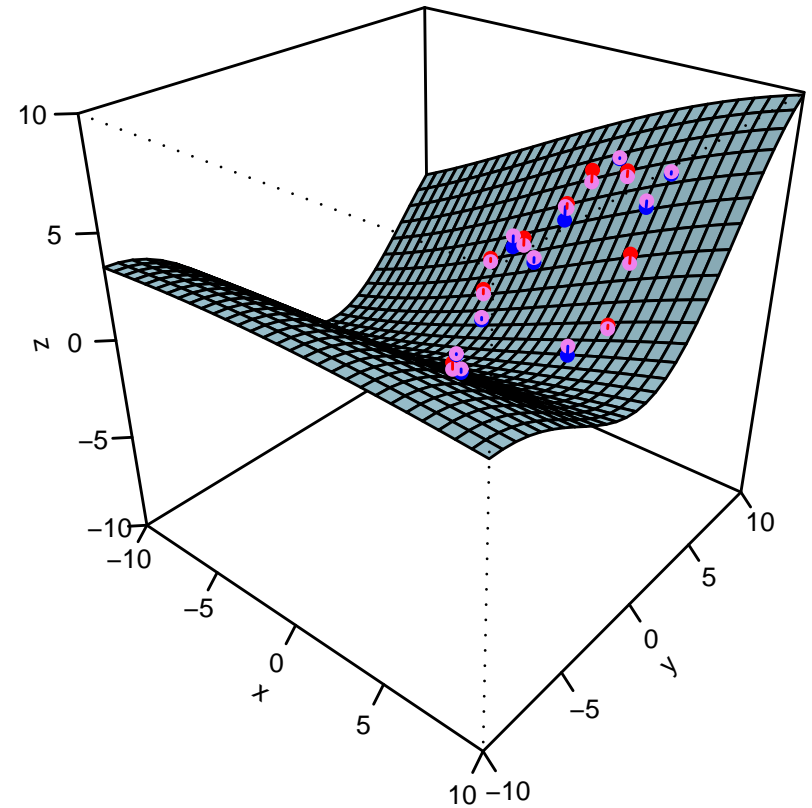
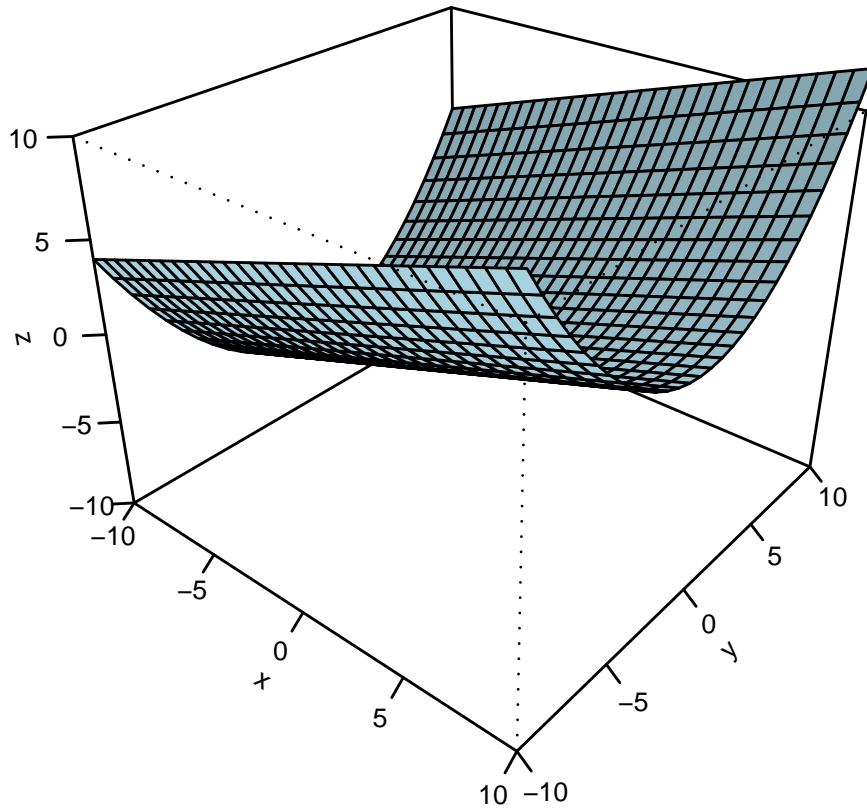
Variable 1: the linear and nonlinear relevance parameters for input 3 are both 0

GP on time-series



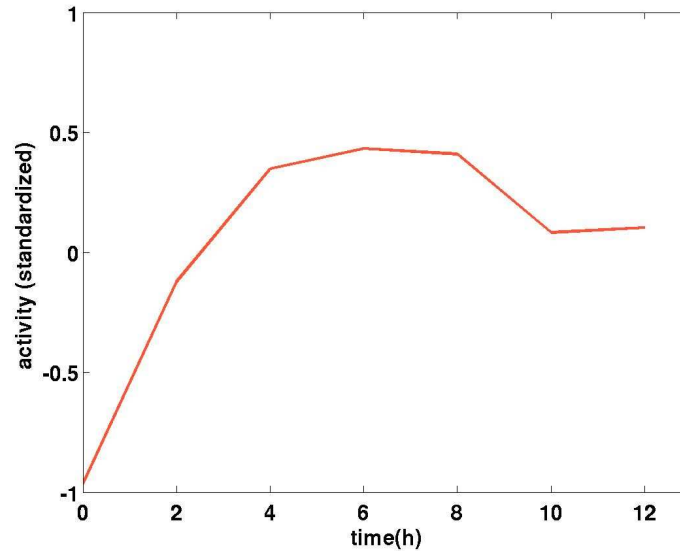
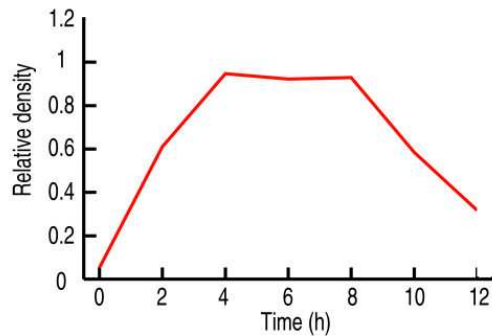
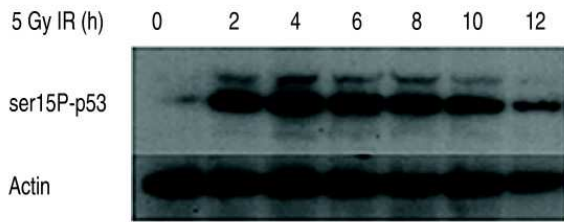
Variable 2: the linear and nonlinear relevance parameters for input 1 are both 0

GP on time-series



Variable 3: the linear and nonlinear relevance parameters for input 1 are both 0

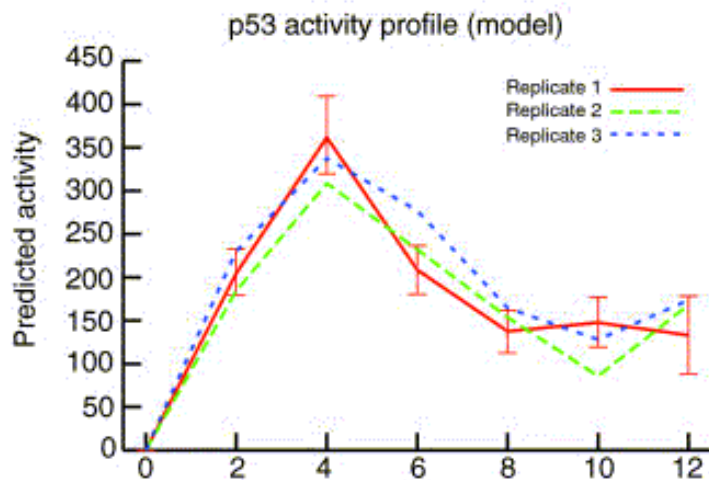
Reconstruction of p53 activity



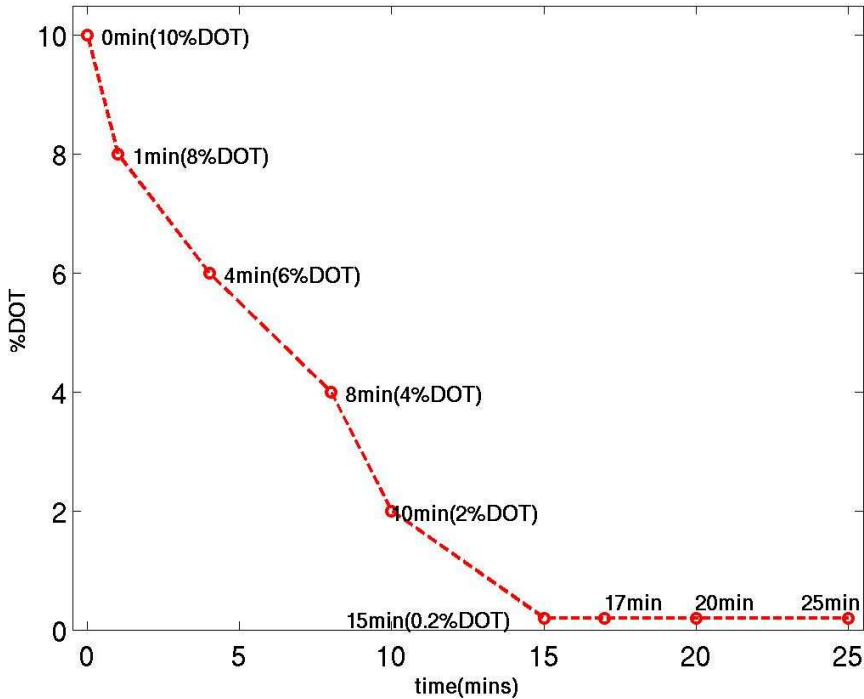
Reconstruction by Gaussian process dynamical model with 1 hidden factor

Comparison with western blot result from *Barenco et al., 2006*

Left: reconstruction by *Barenco et al.*



DosR induction in *M. tuberculosis*

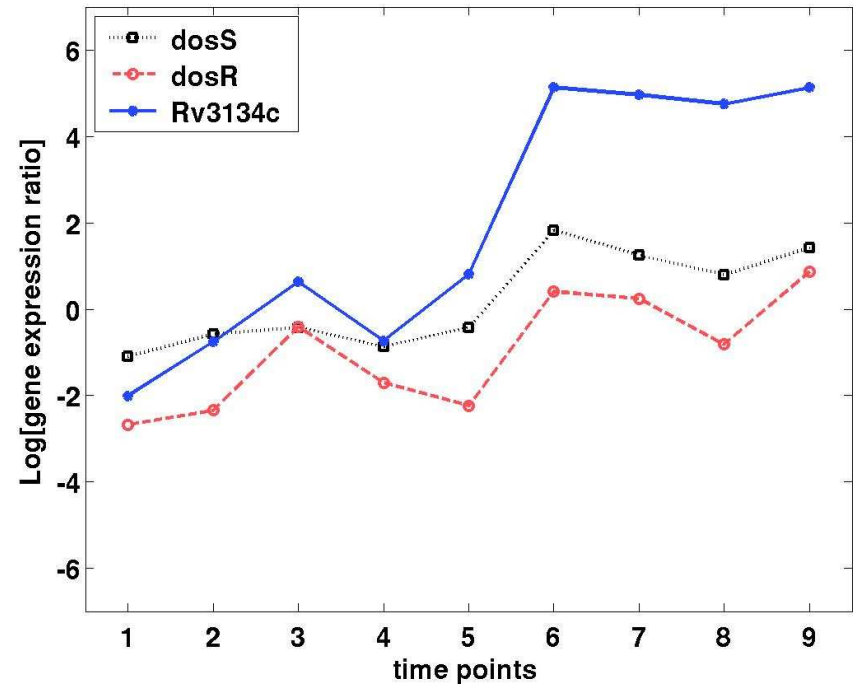
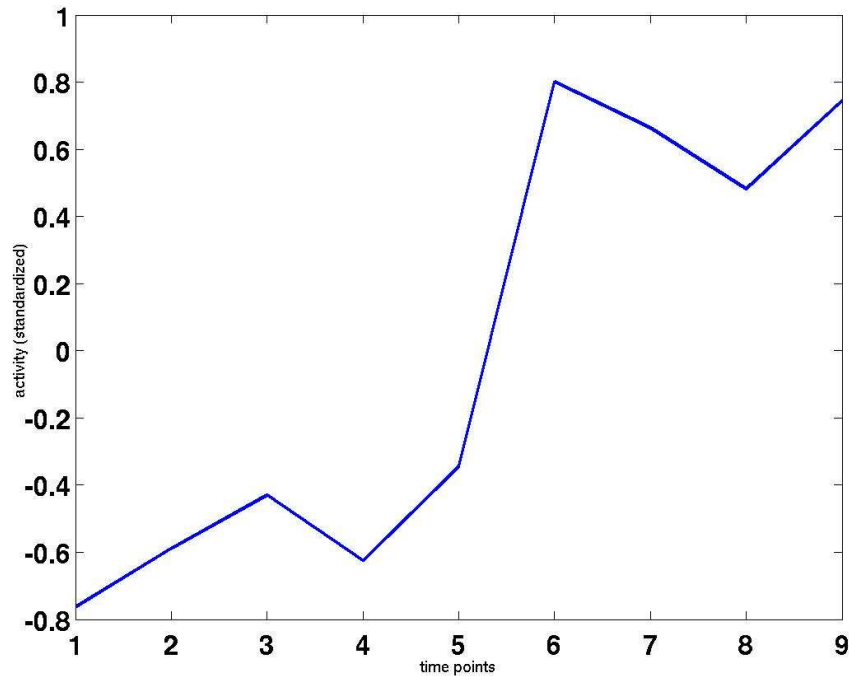


- DosR induced by various signals, including low oxygen
- Regulates about 30 other genes directly, including itself
- Known binding motif

Timecourse experiment with oxygen reduction in chemostat culture

Joanna Bacon (HPA Porton Down)

Reconstruction of DosR activity



Reconstructed DosR
TFA

Expression levels DosR,
DosS, Rv3134c

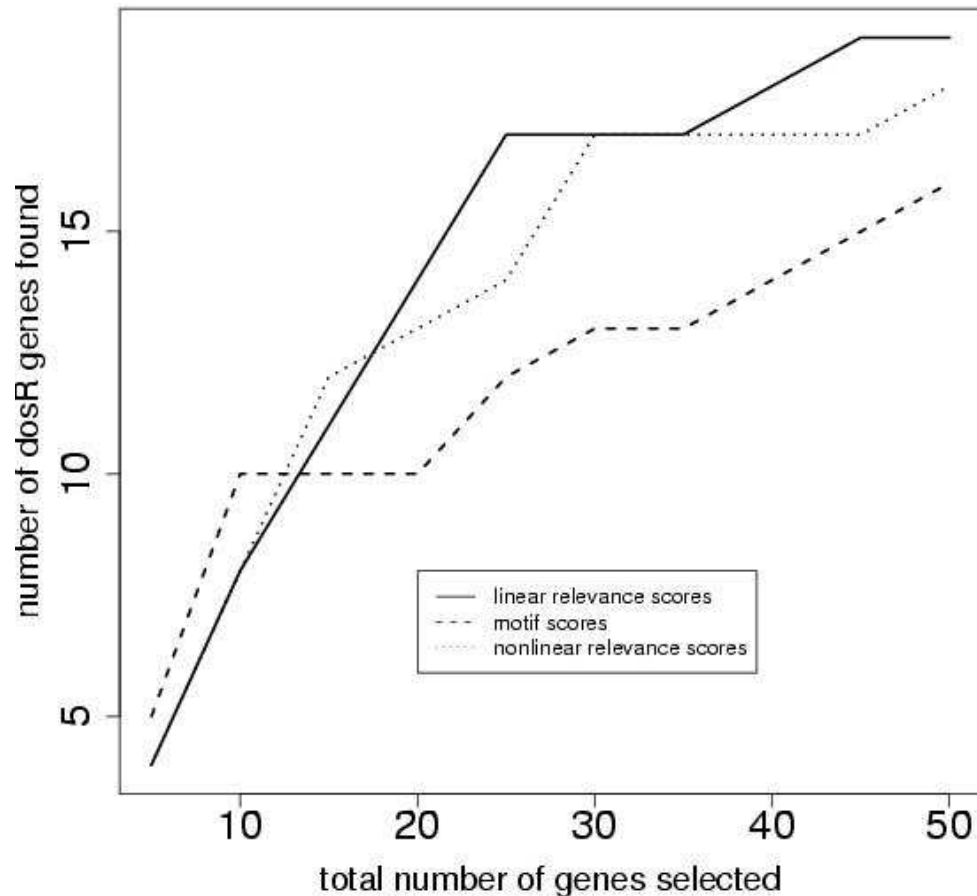
TFA and expression profile different (closer to
RV3134c)

Combining with motif information

- Binding motif (log odds scoring matrix) of DosR is known
- Sum positive log odds scores for upstream region of each gene
- Combine with linear and nonlinear relevance parameters of Gaussian process regression:

$$\log \frac{p_g}{1 - p_g} = w_0 + w_1 \beta_{\text{lin}}^g + w_2 \beta_{\text{nonlin}}^g + w_3 x_{\text{motif}}^g$$

Predicted DosR-regulated genes



Test case: 21 confirmed DosR dependent genes in leave-one-out CV

In top 50:

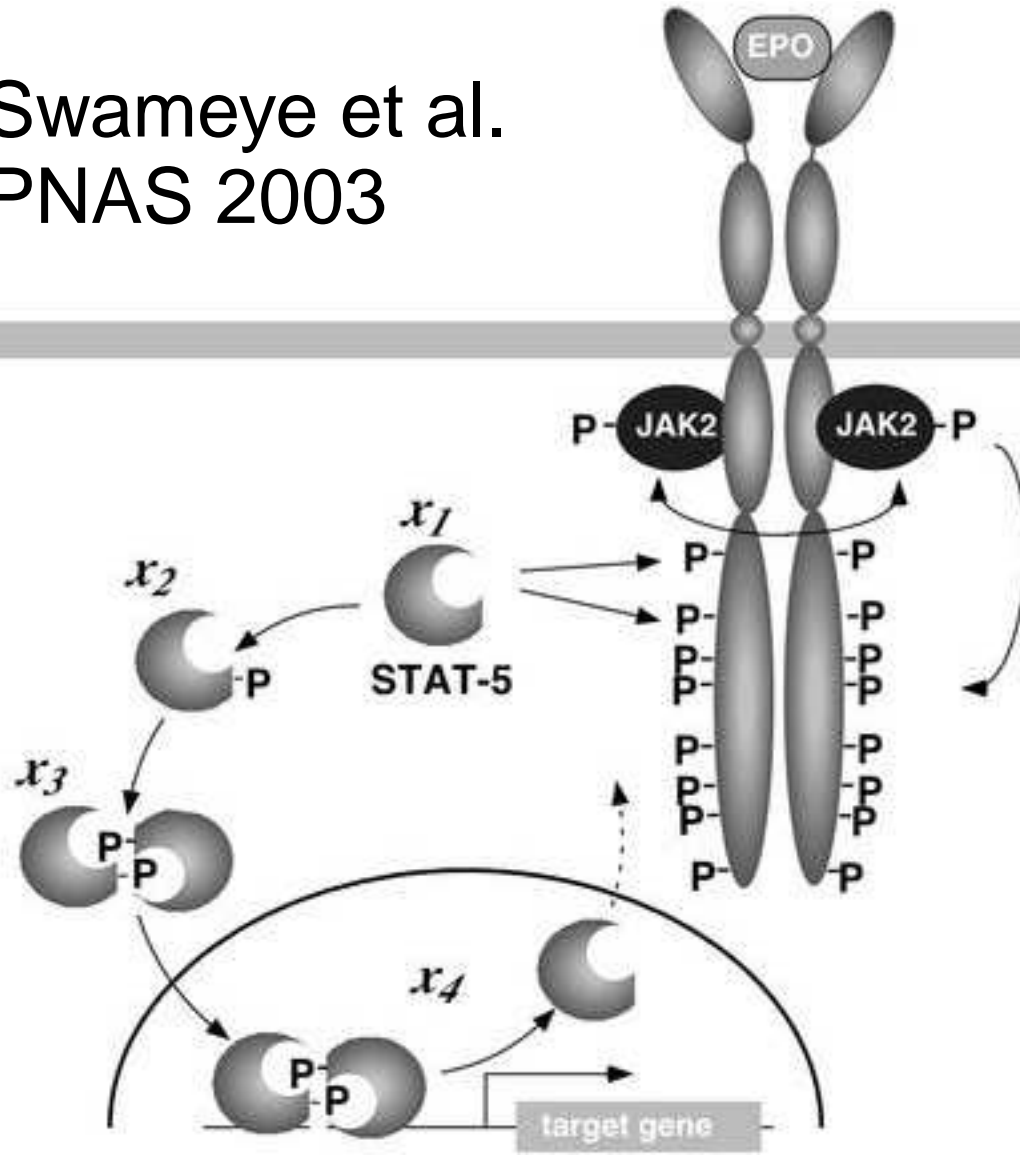
- Motifs: 15/21
- Linear GP: 19/21
- Nonlinear GP: 18/21

In top 34:

- Combined by logistic regression: 21/21

JAK-STAT5 pathway

Swameye et al.
PNAS 2003



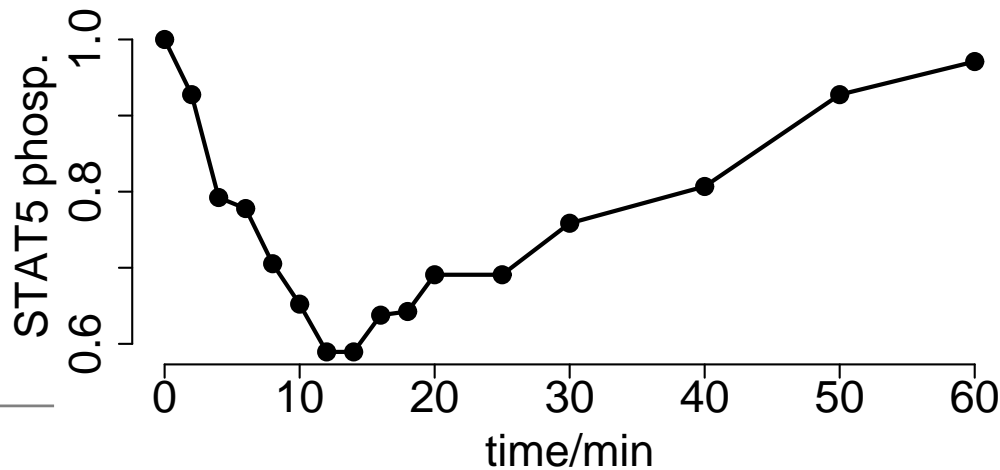
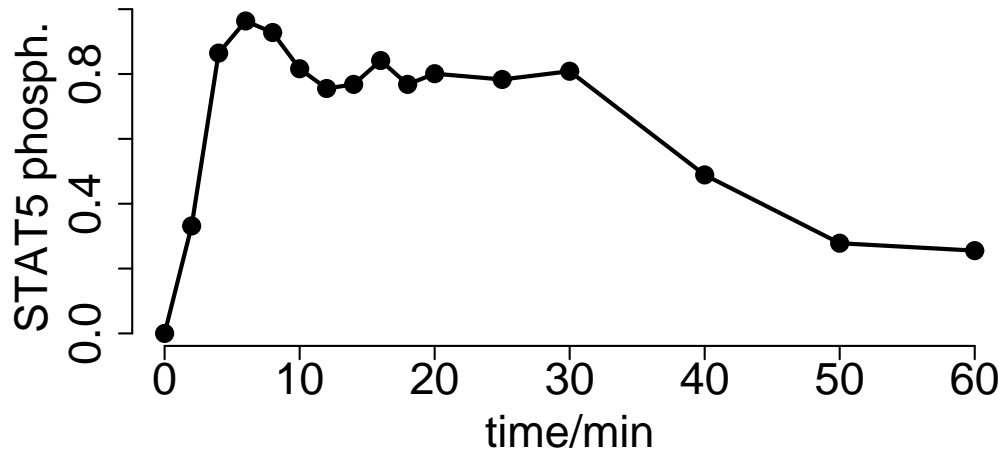
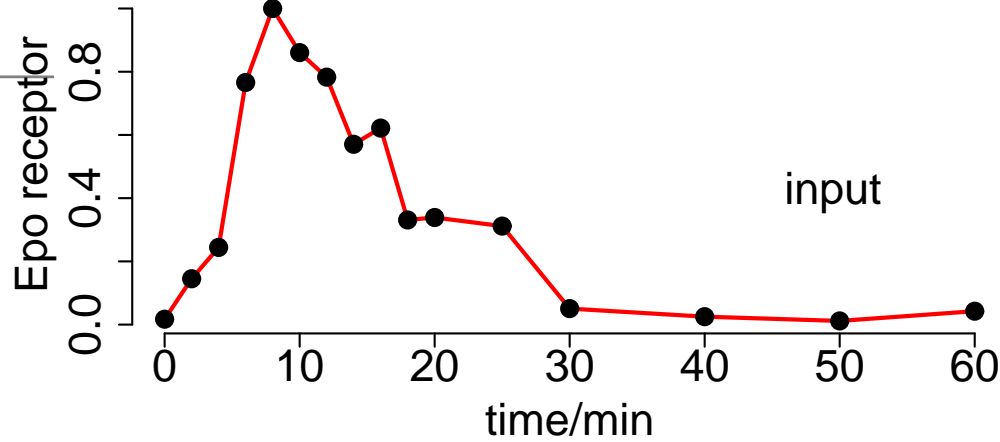
1) Receptor EpoR transforms unphosphorylated STAT_{unph} into phosphorylated STAT_{ph}

2) Phosphorylated STAT_{ph} dimerizes

3) Dimers move into nucleus and bind as TFs

Various models for next step: degradation, recycling with/without delay

Data for JAK STAT-5



input: EpoR

output1: STAT-5
unphosphorylated

output2: STAT-5
phosphorylated

(Quantitation by
chemiluminescence
and Lumilmager
detection,

lab of U. Klingmüller,
see Swameye et al.
PNAS 2003)

Differential equation model for JAK-STAT5

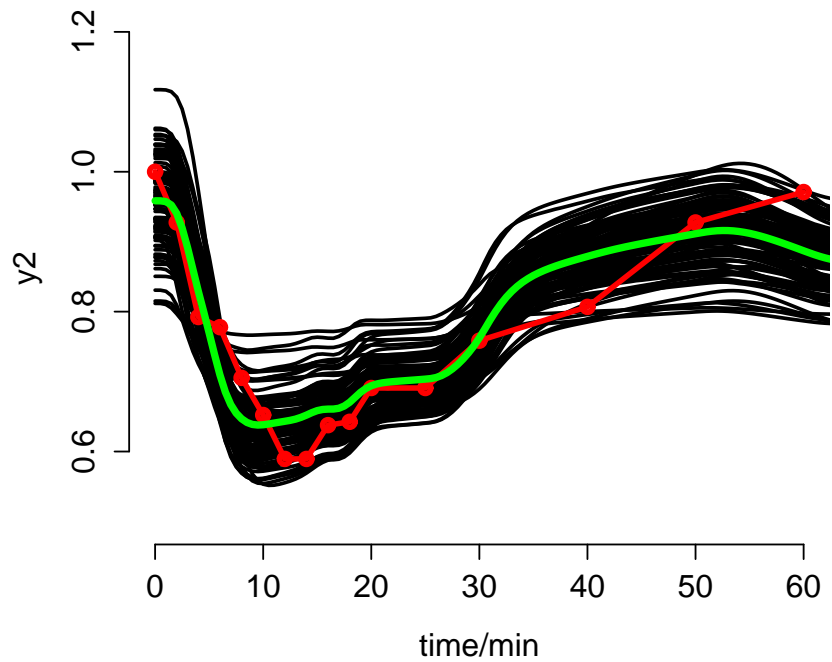
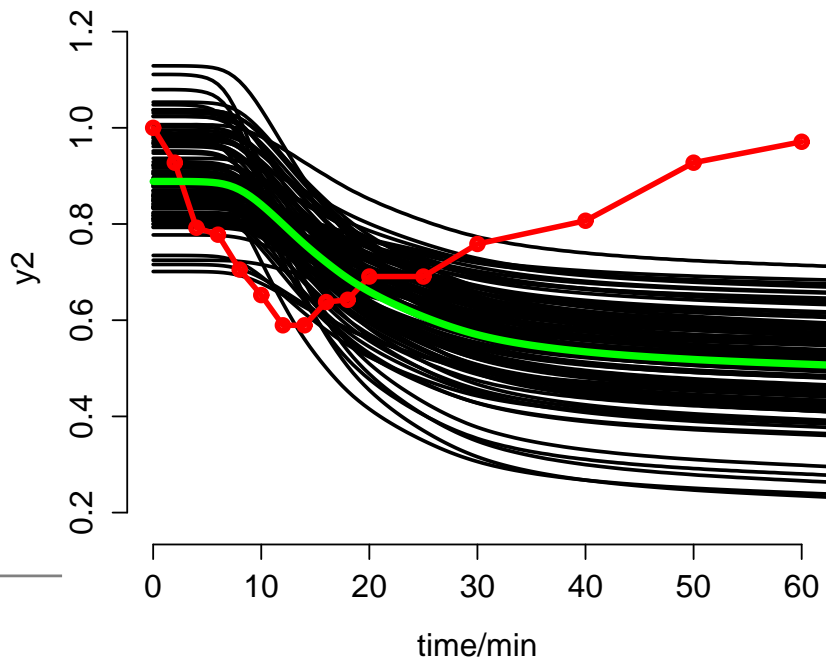
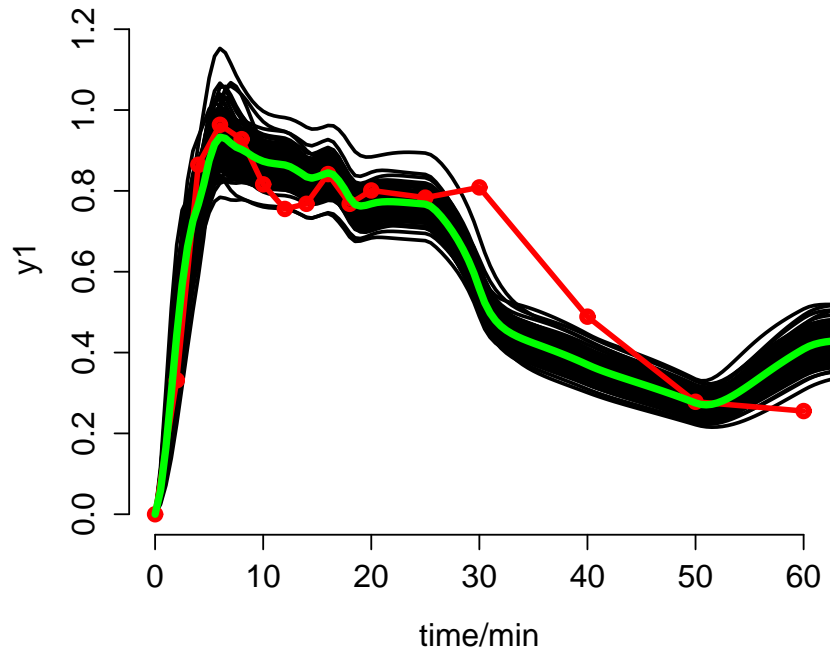
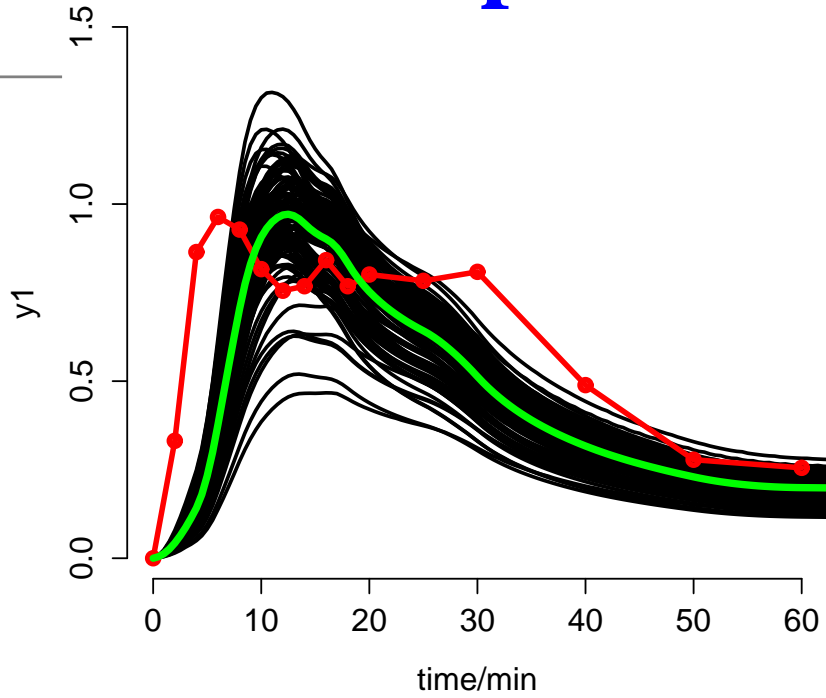
$$\frac{d \text{STAT}_{\text{unph}}}{dt} = -k_1 \text{STAT}_{\text{unph}} \text{EpoR}$$

$$\frac{d \text{STAT}_{\text{ph}}}{dt} = k_1 \text{STAT}_{\text{unph}} \text{EpoR} - k_2 (\text{STAT}_{\text{ph}})^2$$

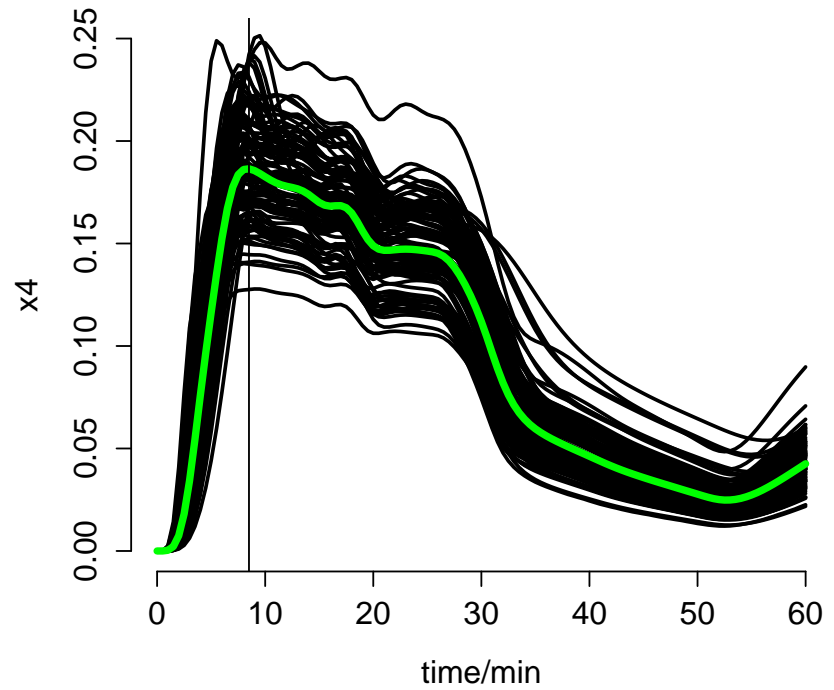
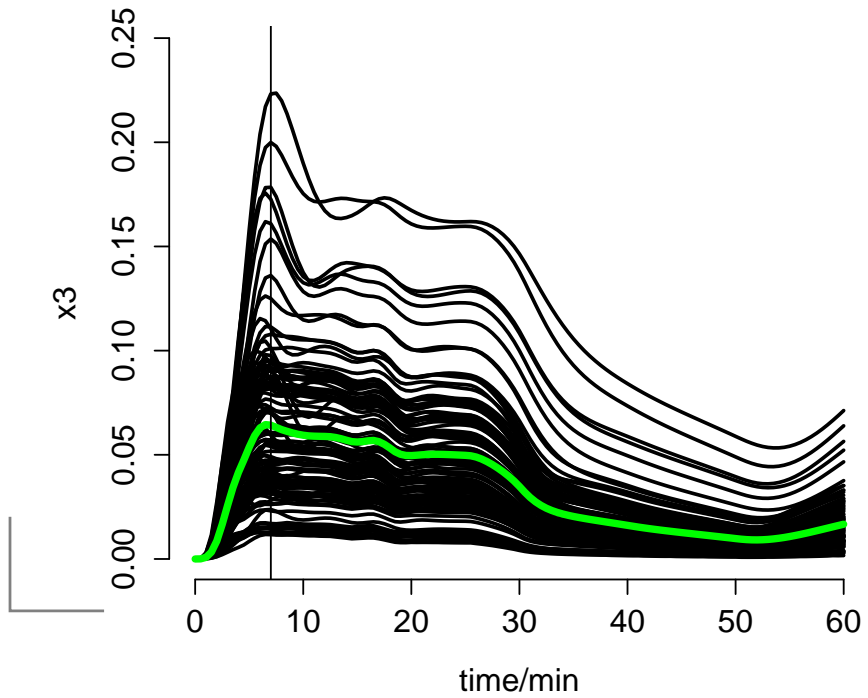
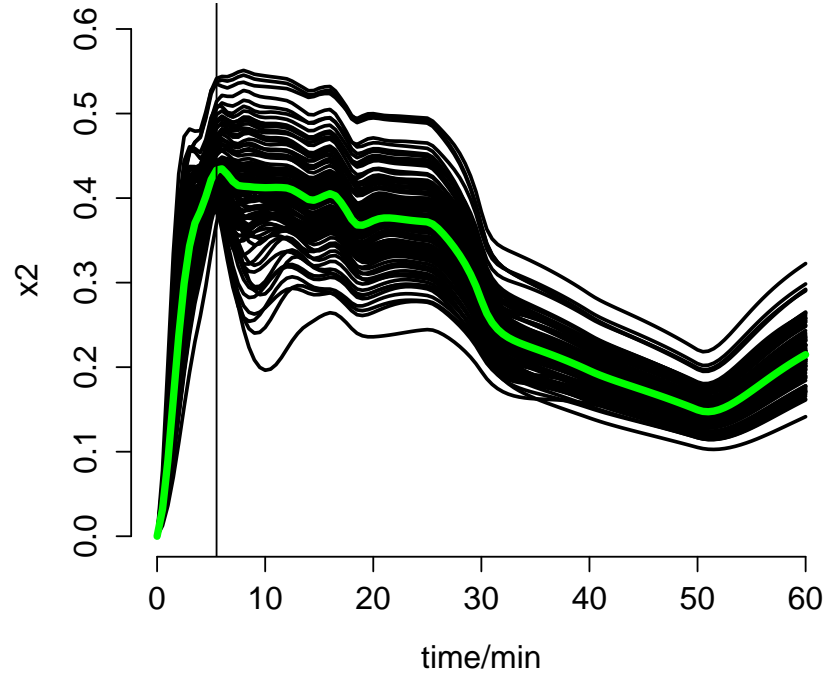
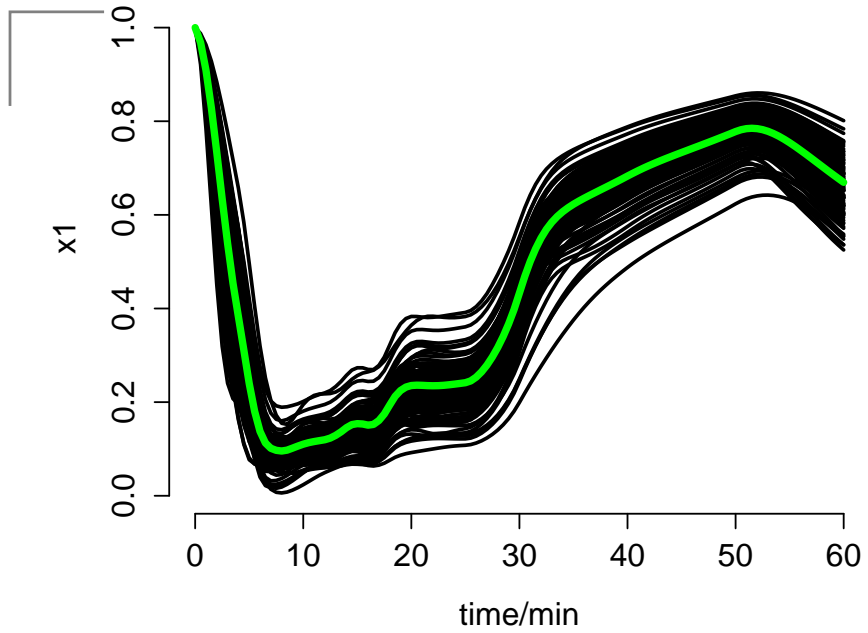
$$\frac{d \text{STAT}_{\text{dim}}}{dt} = 0.5 k_2 (\text{STAT}_{\text{ph}})^2 - k_3 \text{STAT}_{\text{dim}}$$

$$\frac{d \text{STAT}_{\text{nucl}}}{dt} = k_3 \text{STAT}_{\text{dim}}$$

Compare no recycling/recycling



Hidden variables



Summary

- Methods for reconstructing hidden variables:
 - Factor analysis
 - State space models (nonlinear)
 - DE models
- Joint estimation of nonparametric mapping and hidden factors rather challenging
- Problem of identifying reconstructed factor with gene
- Identifiability issues for factors, loadings, transformations