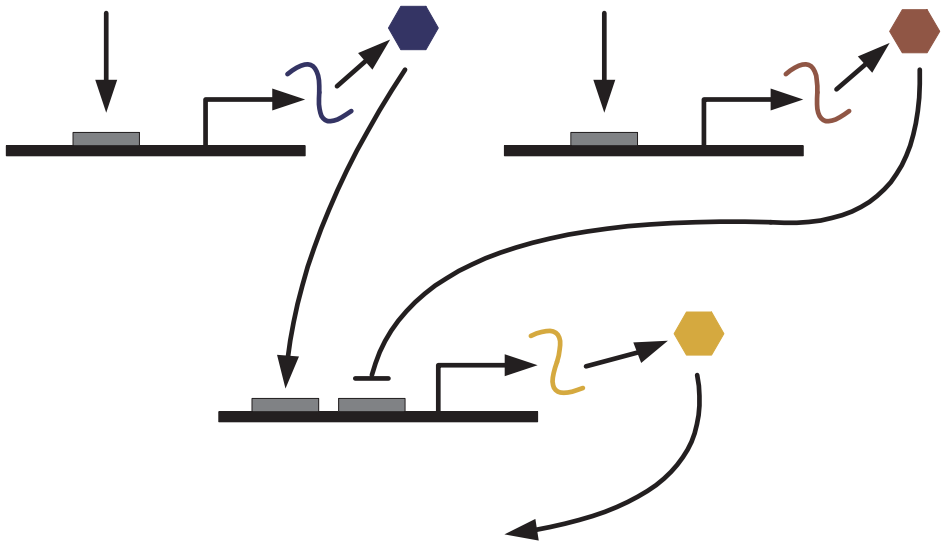
Reconstructing hidden protein activity by nonparameteric methods

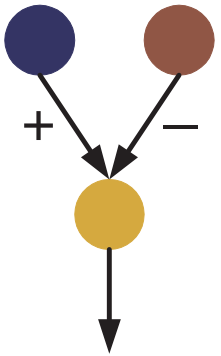
Lorenz Wernisch
with Iosifina Pournara, Yi Zhang

MRC Biostatistics Unit

Cambridge

Gene networks



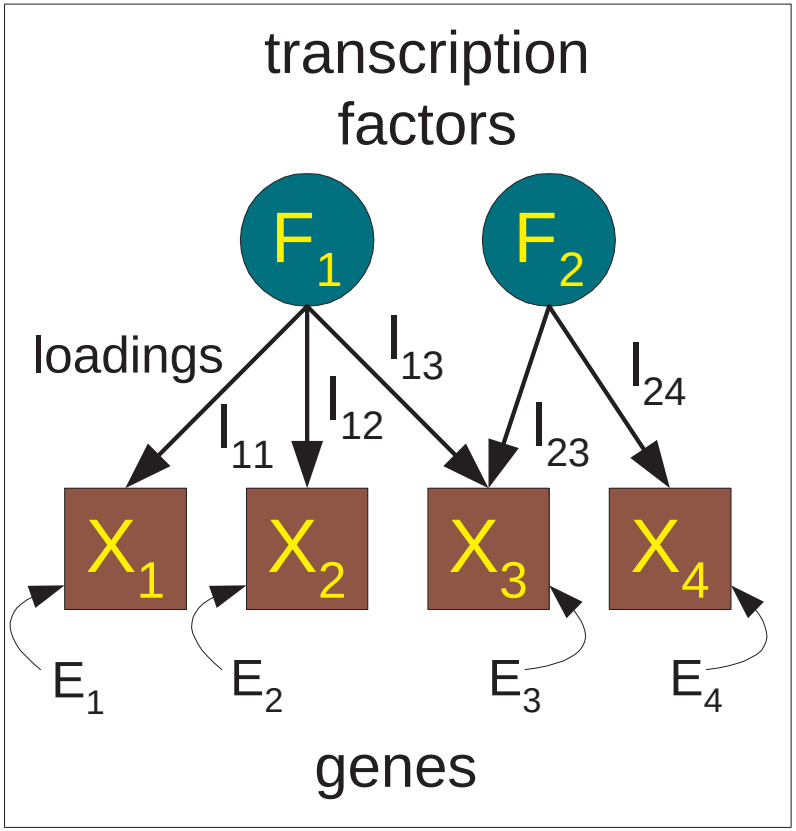


- Protein levels depend on mRNA transcription
- mRNA transcription levels depend on activation/inhibition by transcription factor proteins
- mRNA abundance does not necessarily reflect gene activity

Latent transcription factor activity

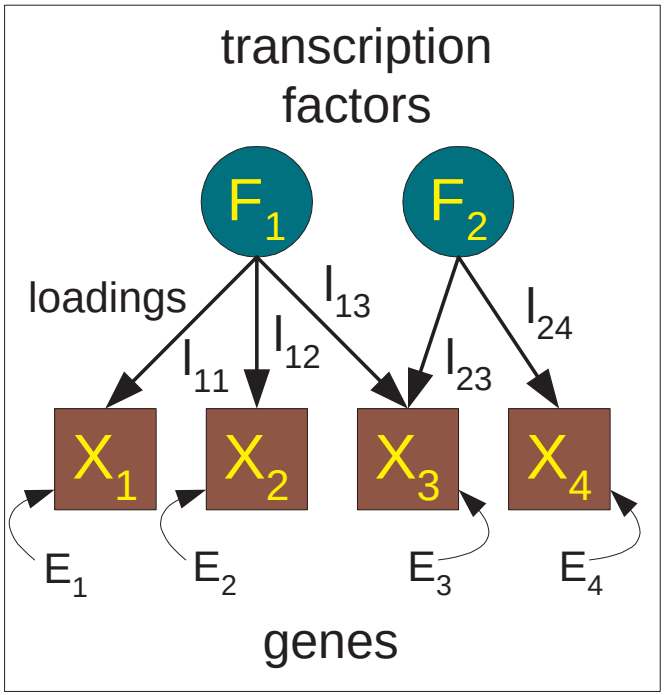
- How can protein activity be reconstructed from indirect effects on gene expression
- Of interest: not only inference of gene activity but also regulatory connectivity between hidden factors
- Options:
  - Factor analysis
  - Gaussian processes
  - Time-series models, Hidden Markov models

Simple two-level model

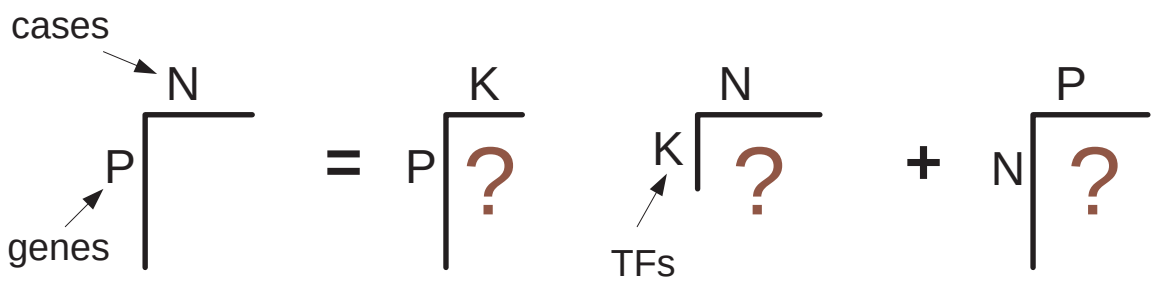


- Each gene potentially regulated by several transcription factors (TFs)
- Each TF potentially regulates several genes
- Measurement noise (E) on gene expression (X)
- No connection between factors
- **Sparse connectivity**

Factor analysis model



expression data		connectivity matrix		transcription factors matrix		noise
X	**=**	**L**		**F**	**+**	**E**



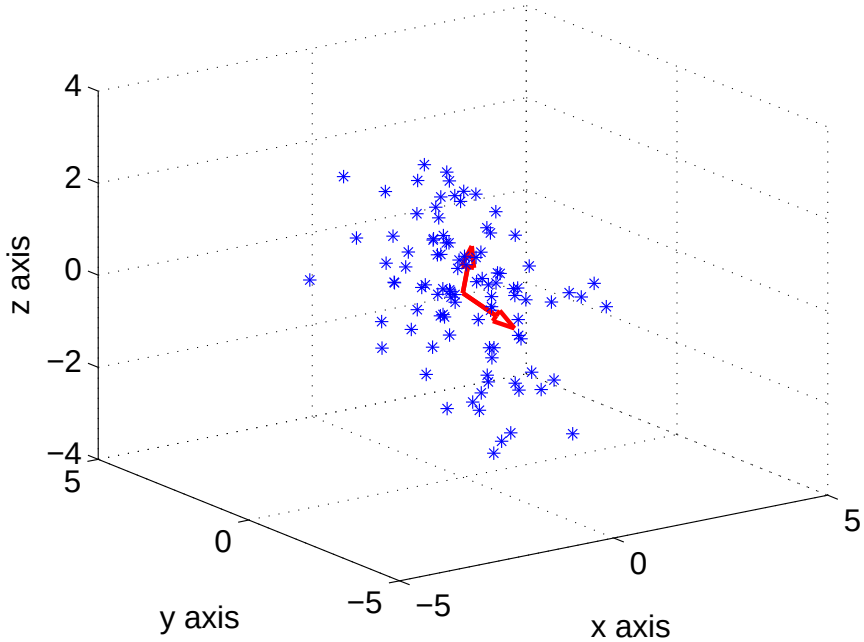
$$
\begin{aligned}
x_1 &= l_{11} f_1 + \epsilon_1 \\
x_2 &= l_{12} f_1 + \epsilon_2 \\
x_3 &= l_{13} f_1 + l_{23} f_2 + \epsilon_3 \\
x_4 &= l_{14} f_2 + \epsilon_4
\end{aligned}
$$

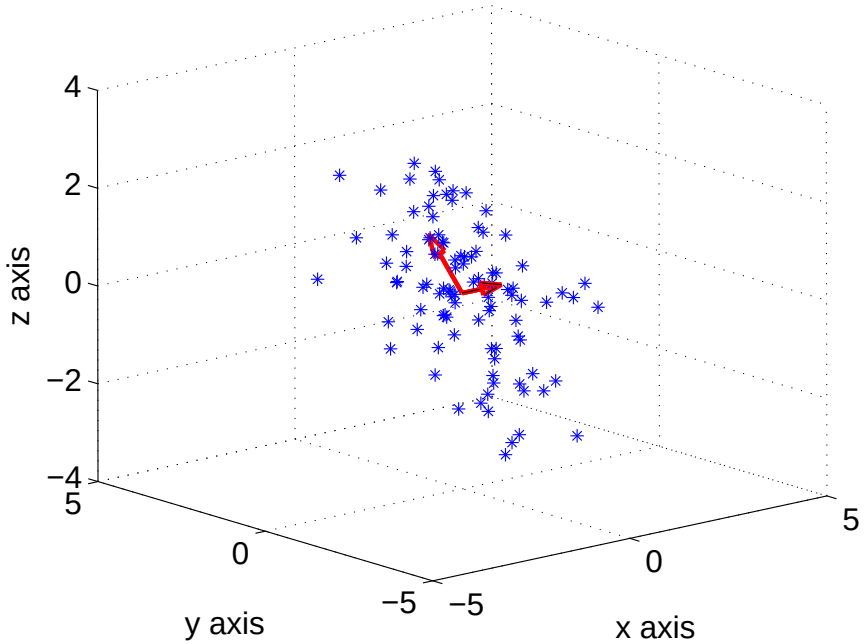
TFs

Genes

0	0	1	0	0	0
0	1	0	0	1	0
0	0	0	0	0	1
0	1	0	0	0	0
1	0	0	0	0	0
0	0	0	1	0	1
0	0	0	0	0	1
0	1	0	0	0	0

Problem of factor analysis





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

Factors determined up to rotation

Sparse loadings matrix?

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 alternatingly from posterior of factors and loadings
- Allows sparsity to be incorporating by using sparsity priors on the loadings
- Parameters can be summarised by averaging over their posterior distributions

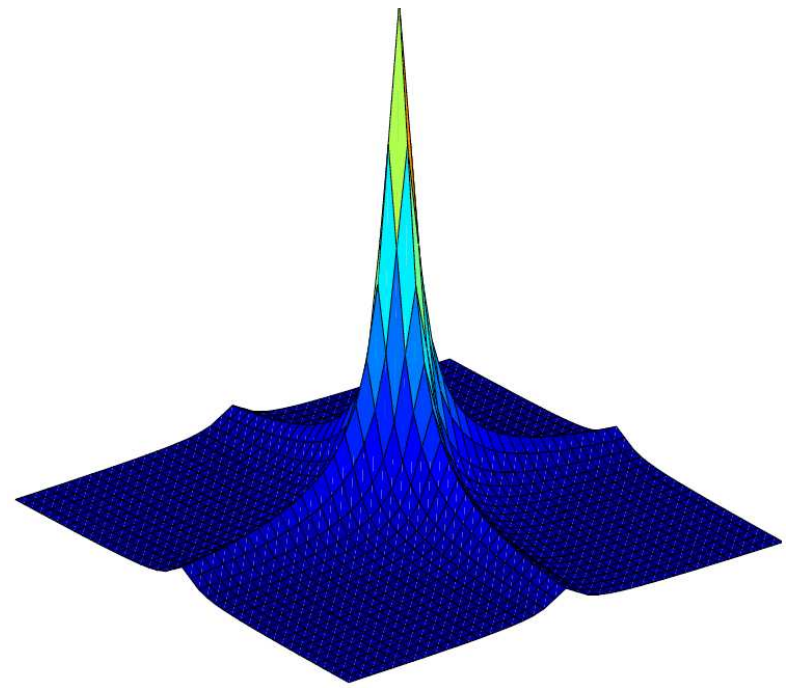
Sparsity priors on loadings

- A mixture prior on L

$$
\begin{aligned}
l_{pk}|z_{pk} = 0 &\sim \delta_0(l_{pk}) \\
l_{pk}|z_{pk} = 1 &\sim \mathcal{N}(0, \sigma^2) \\
z_{pk} &\sim \mathrm{Bern}(\pi)
\end{aligned}
$$

- A Gamma prior on L

$$
\begin{aligned}
l_{pk} &\sim \mathcal{N}(0, \delta_{pk}^{-1}) \\
\delta_{pk} &\sim \mathcal{G}(a, b)
\end{aligned}
$$

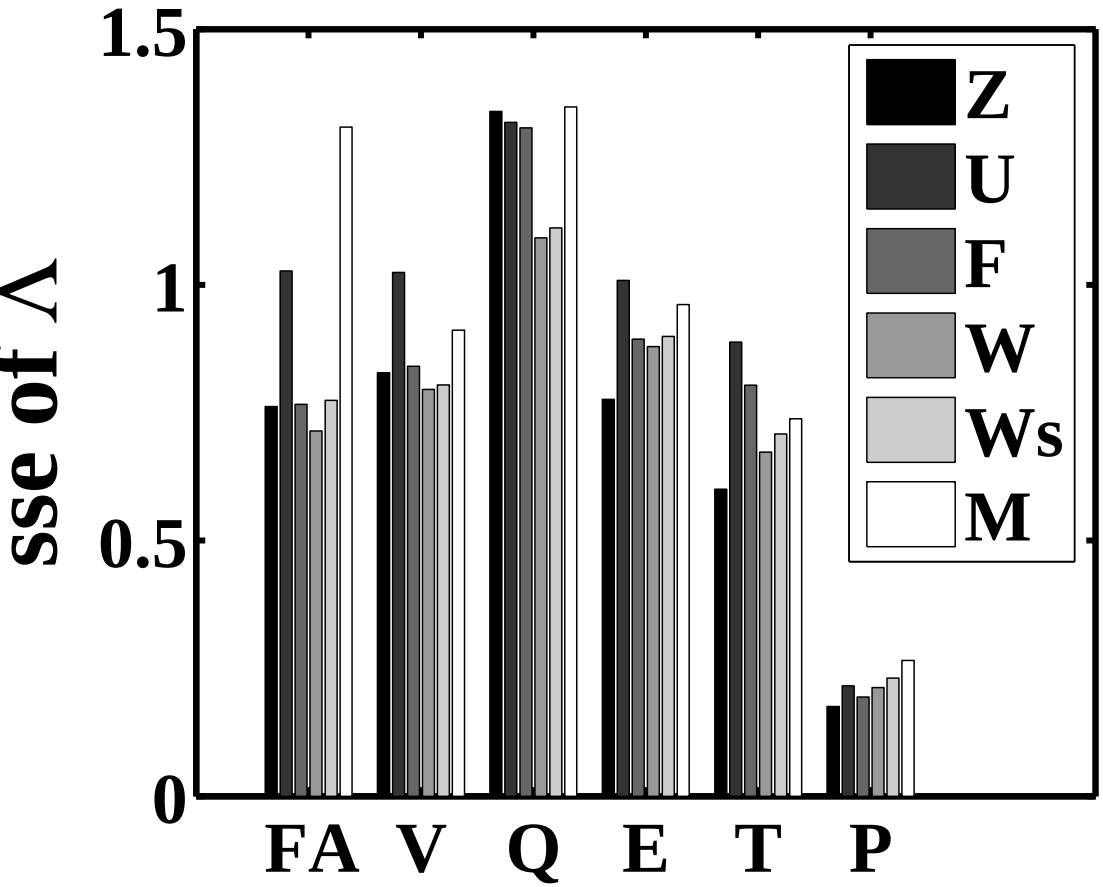


Methods for sparse FA

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: $l_{ij} \neq 0 \sim$ Bern(p), $p << 1$
- (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



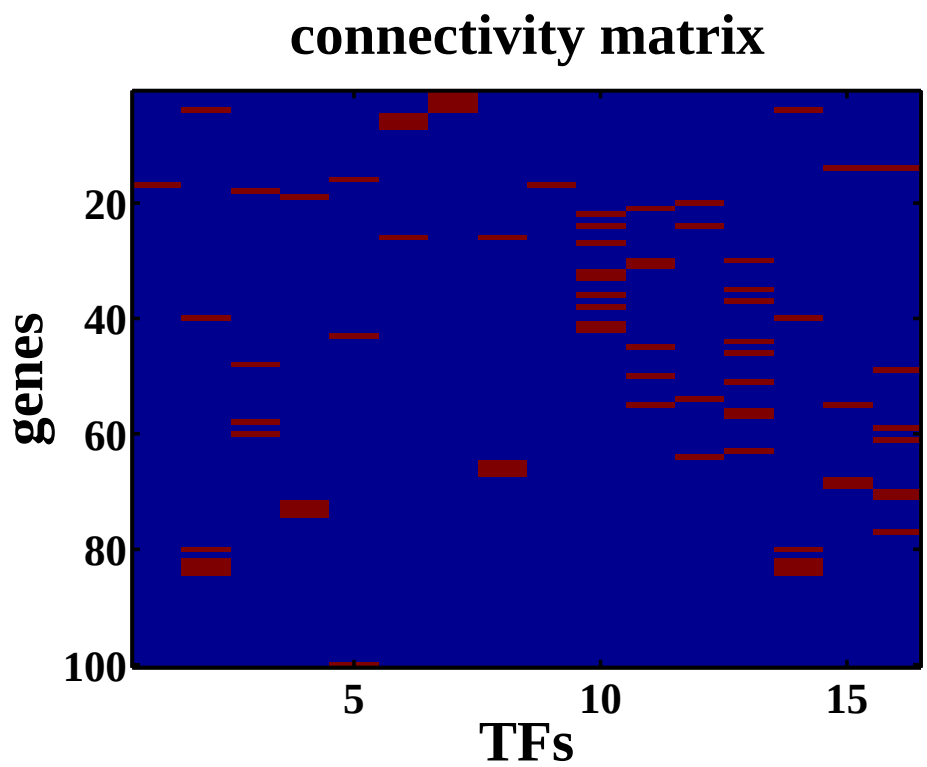
Loadings reconstruction

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

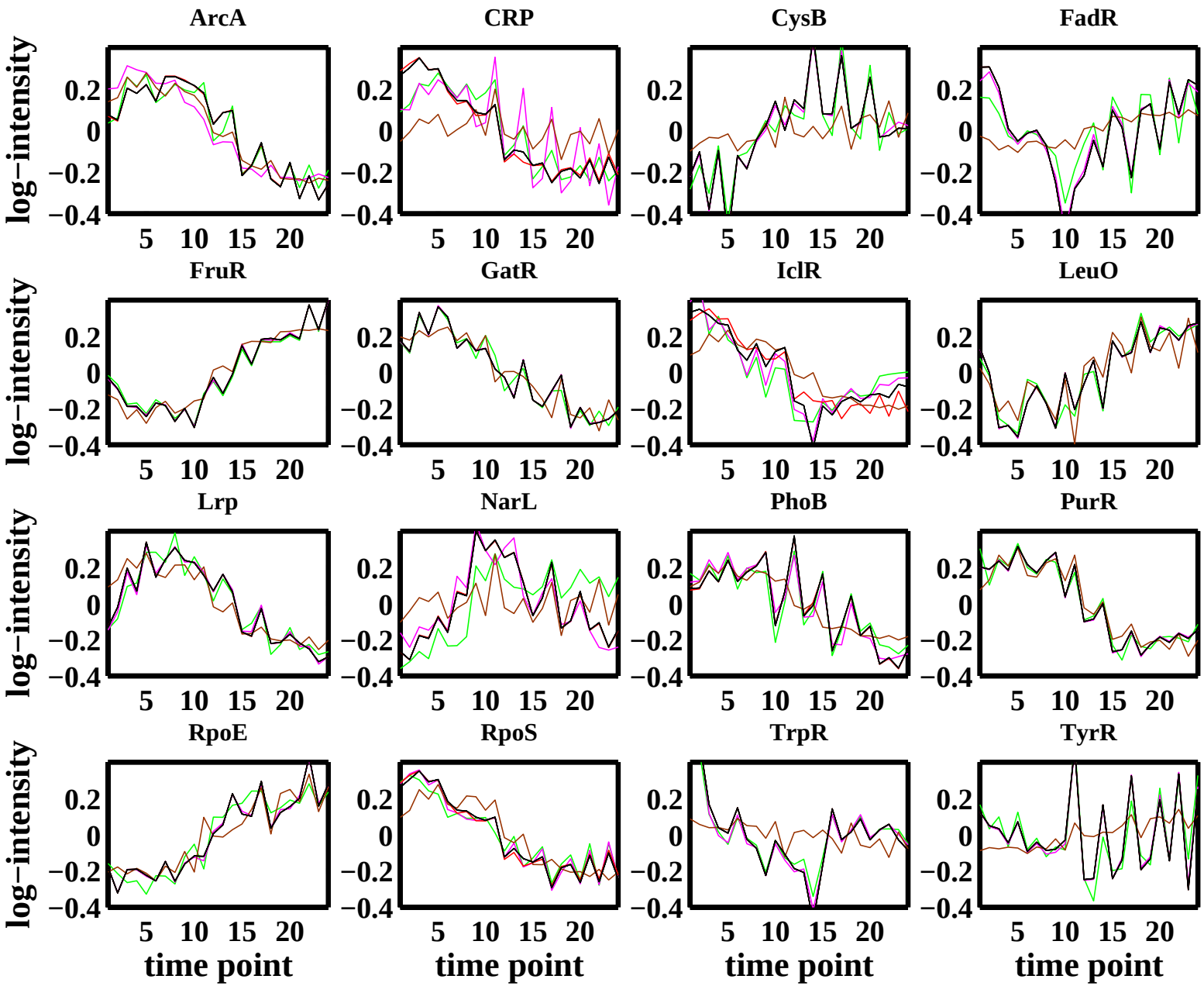
Bayesian methods slightly better than classical FA
Careful rotation (tanh) can improve results
Factor structure correct (up to rotation)

E. coli dataset

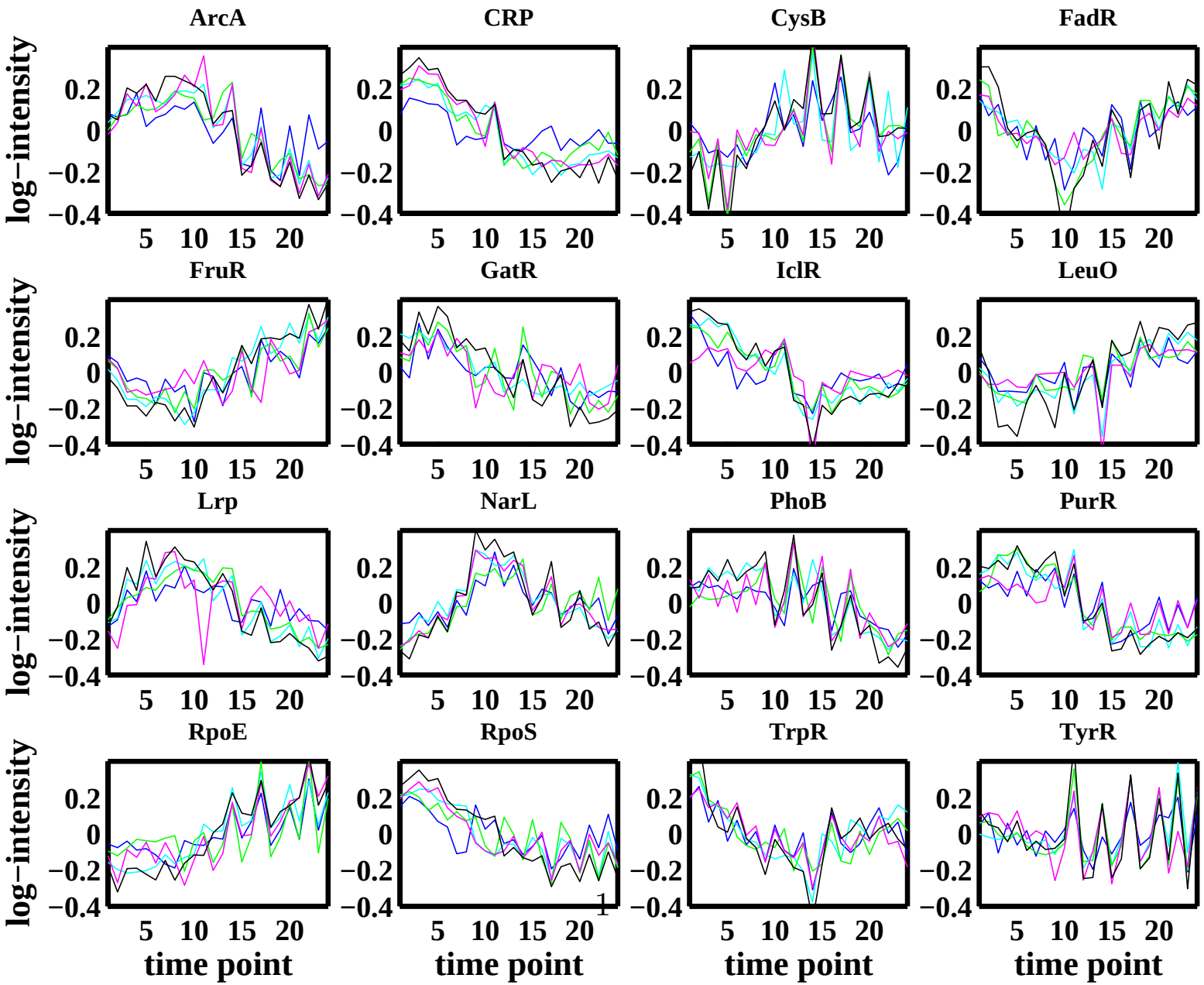
- Microarray time series data (23 time points) 100 genes, 16 putative transcription factors (Kao et al., PNAS 2003)
- Connectivity matrix based on RegulonDB and literature



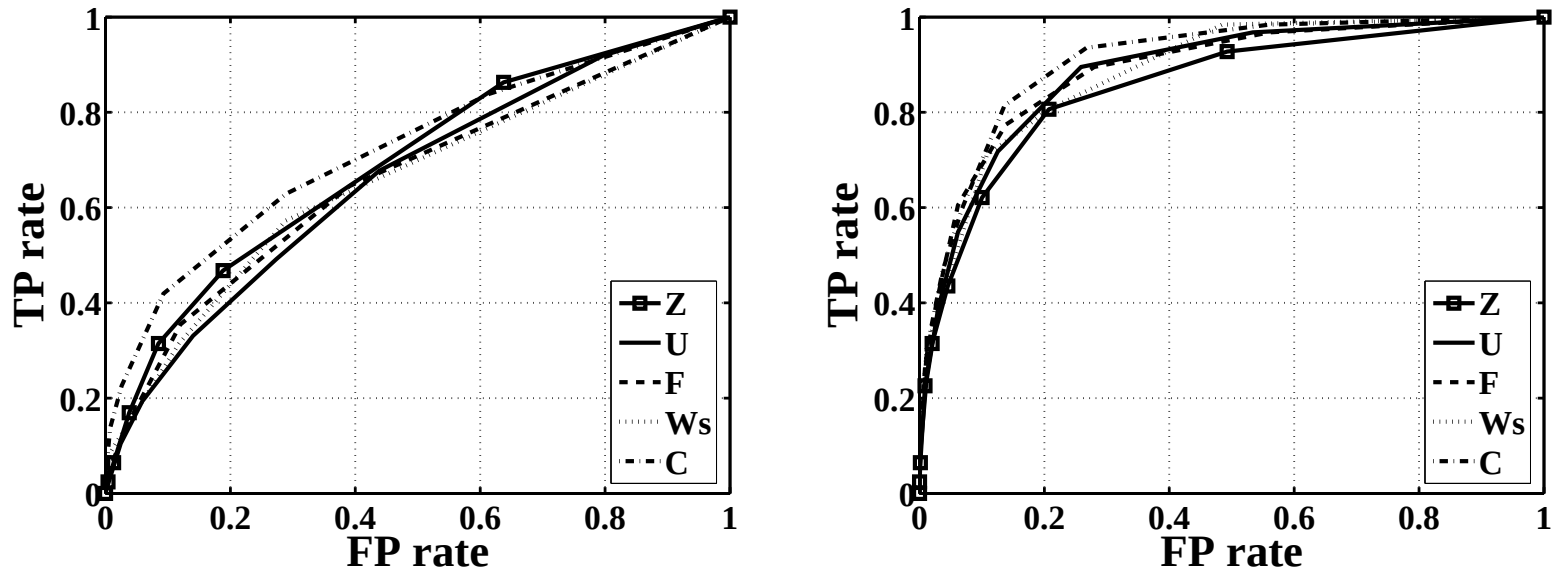
E. coli TFs given connectivity matrix



E. coli TFs without connectivity matrix



ROC for *E. coli* loadings (connectivity)



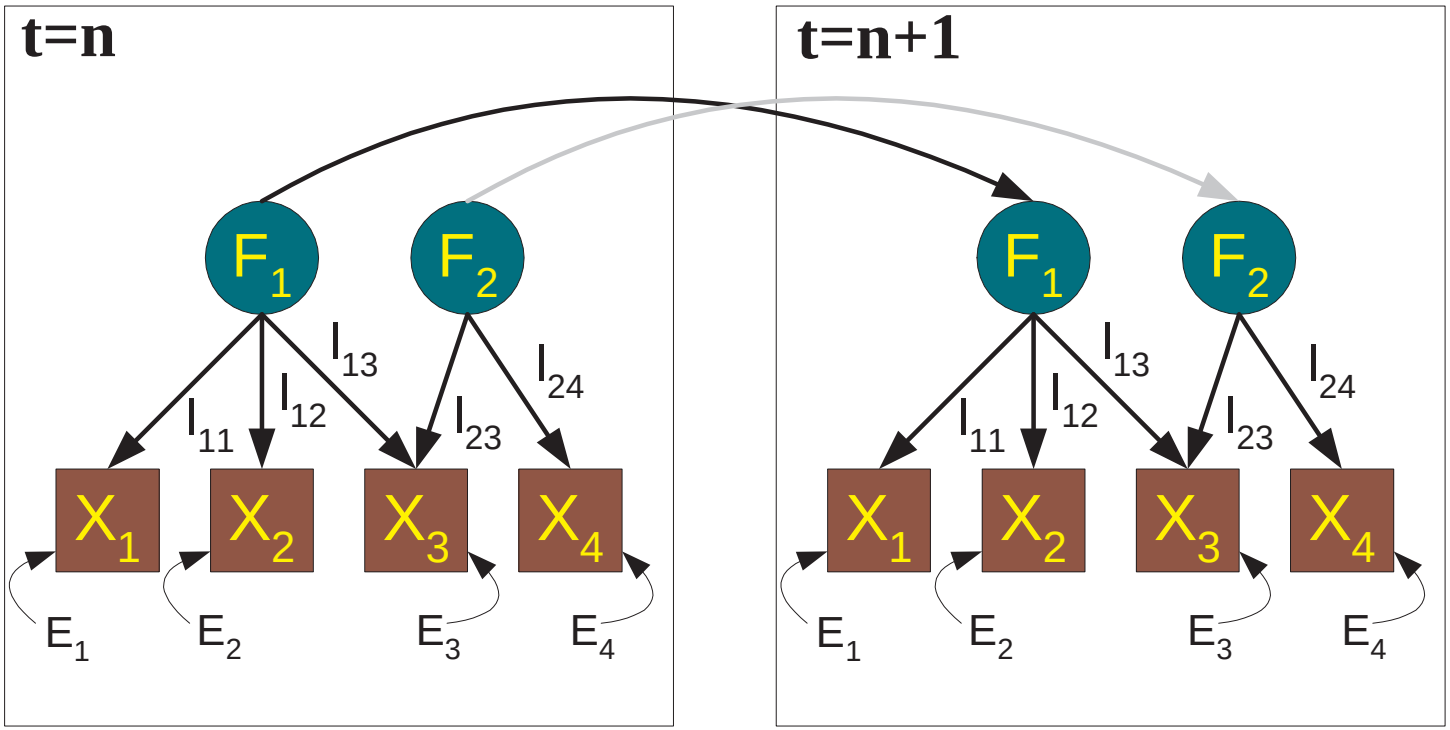
(a) (b)

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



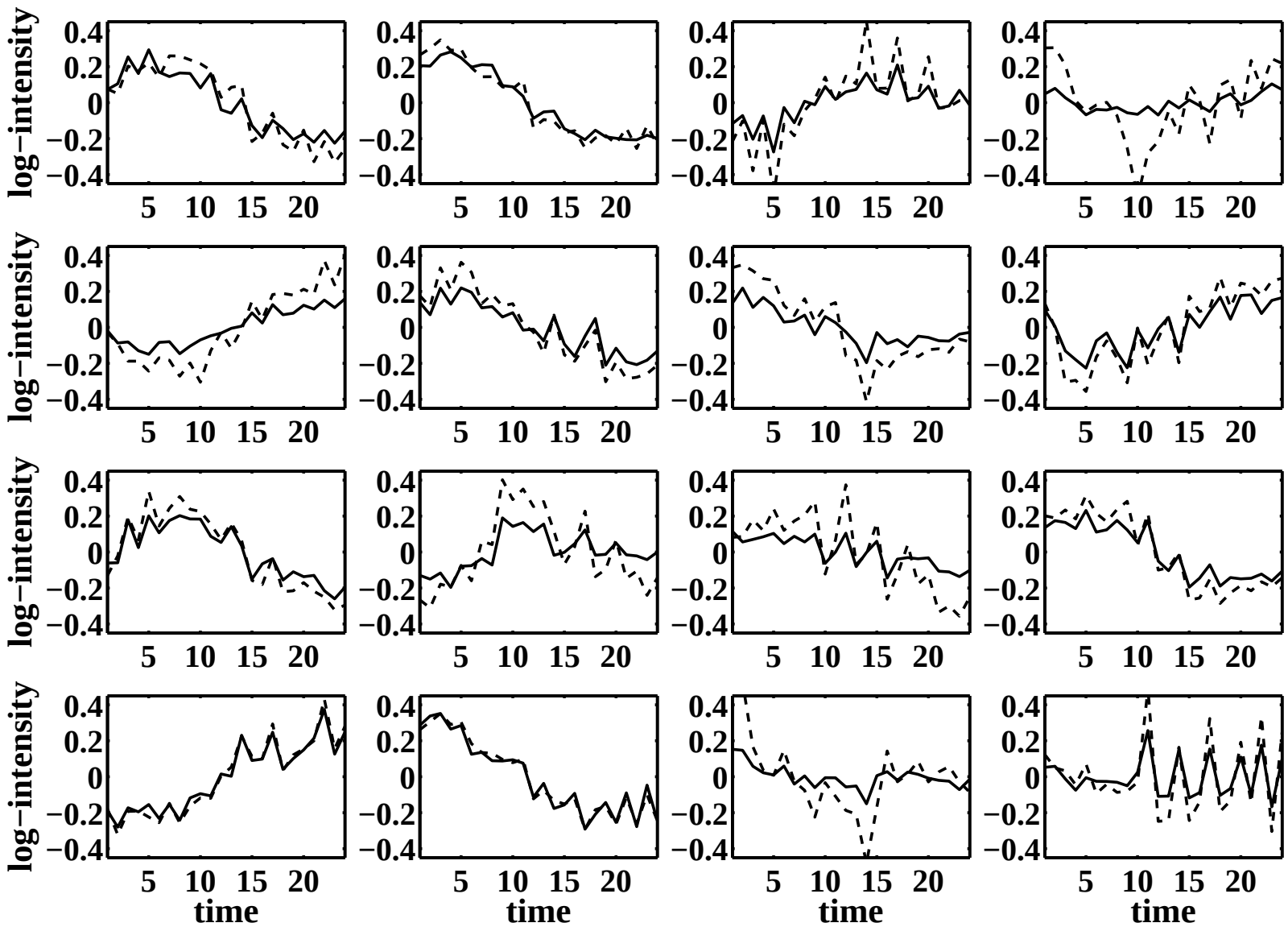
Correlation between time points

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

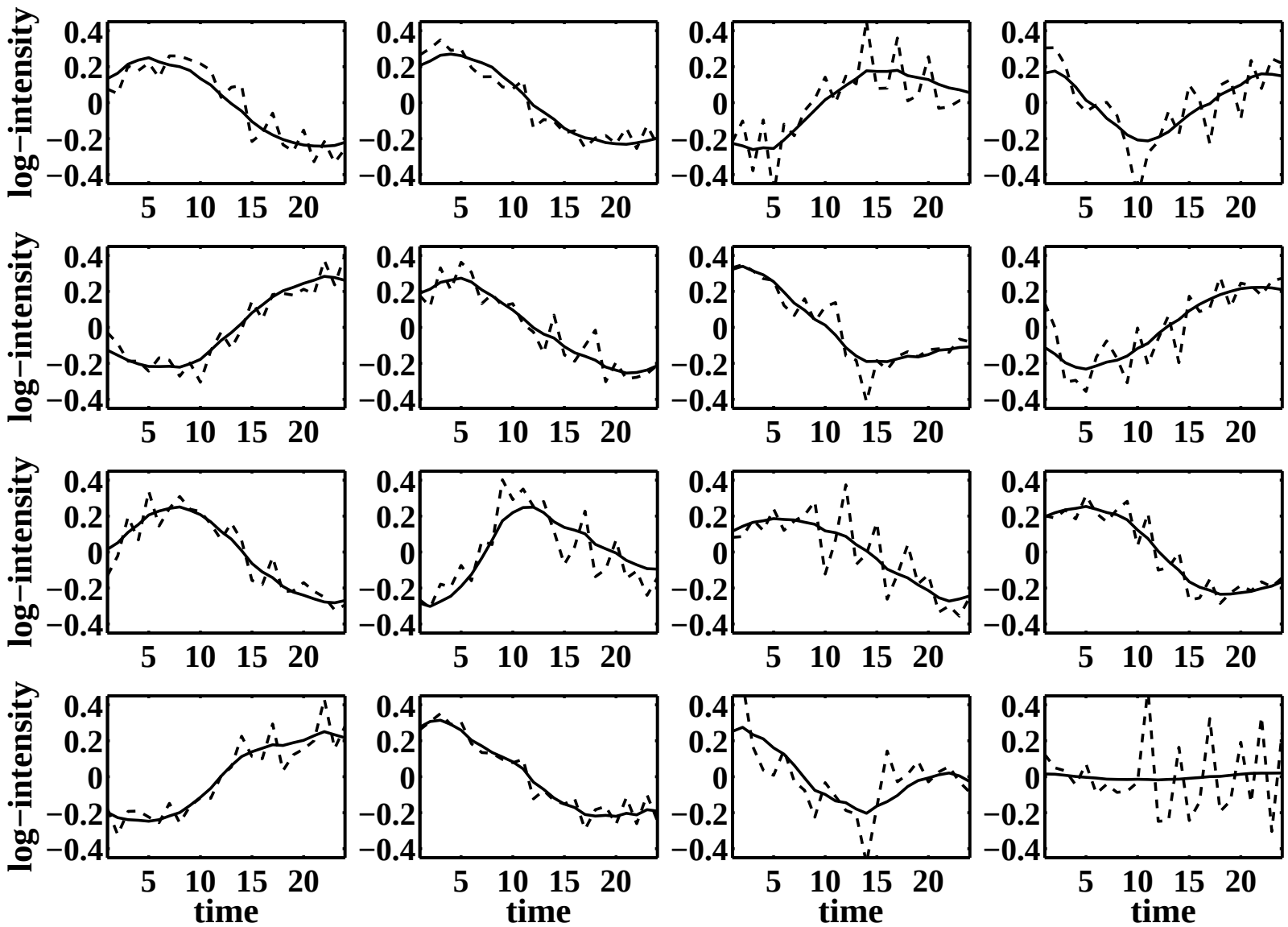
$$\begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{N-1} \\ \rho & 1 & \rho & \dots & \rho^{N-2} \\ \rho^2 & \rho & 1 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{N-1} & \rho^{N-2} & \dots & \dots & 1 \end{pmatrix}$$

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

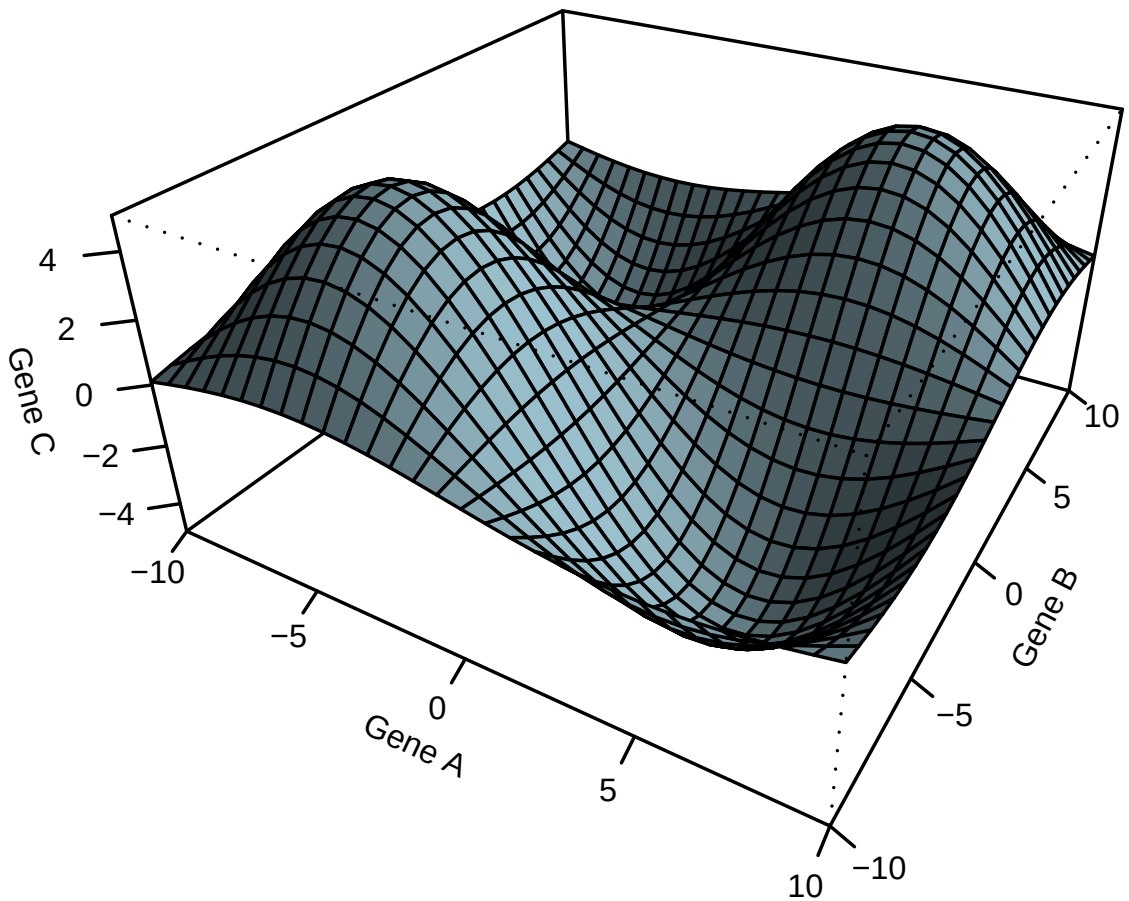
E. coli TFs without time correlation



E. coli TFs with time correlation



Nonlinear dependencies

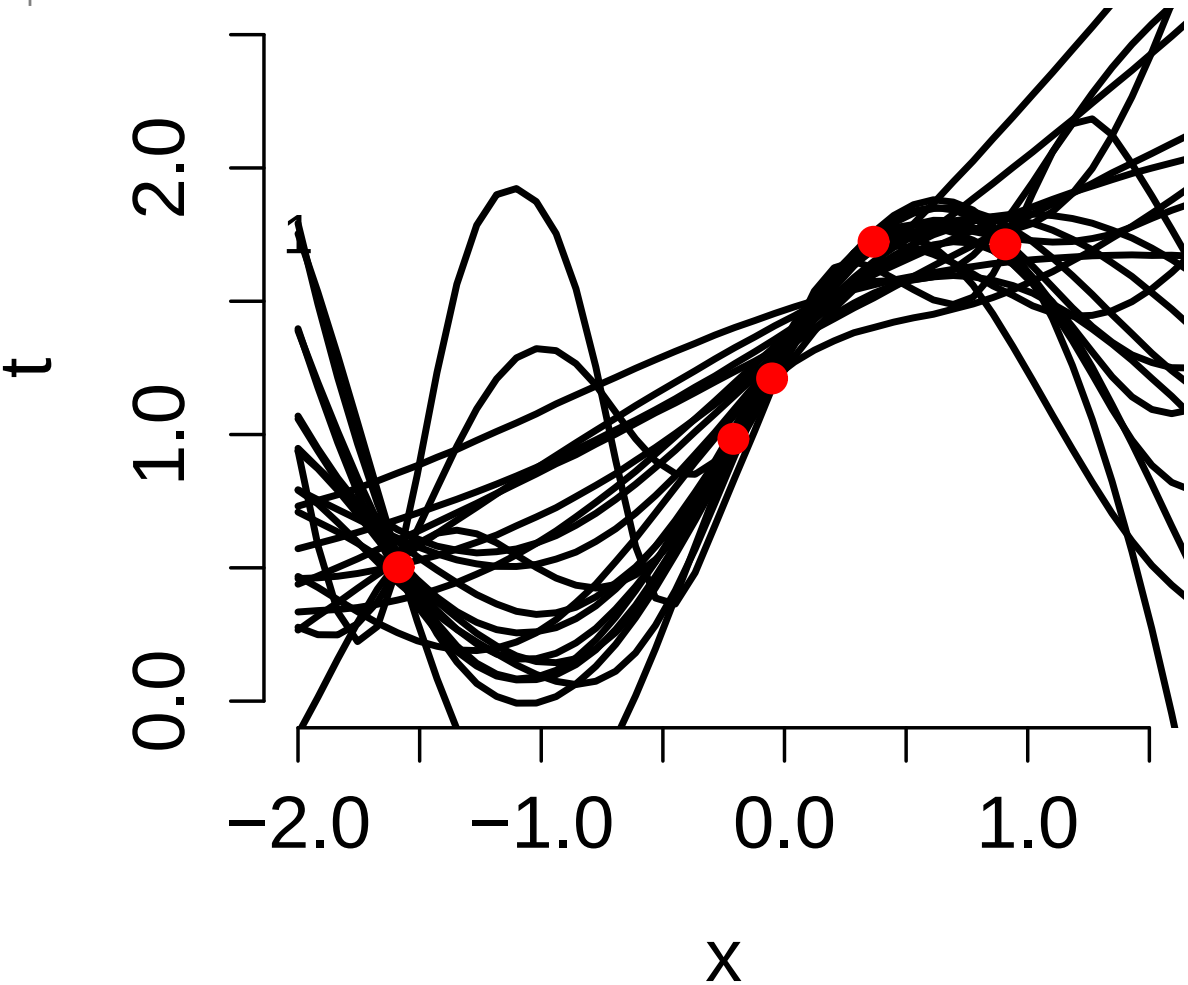


Assumed **linear** depencies of level of gene A on other gene levels

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 $x^{(1)}, \ldots, x^{(N)}$, $x^{(p)} \in \mathbb{R}^d$
- Target values $t = (t^{(1)}, \ldots, t^{(N)})$, $t^{(p)} \in \mathbb{R}$
- 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

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

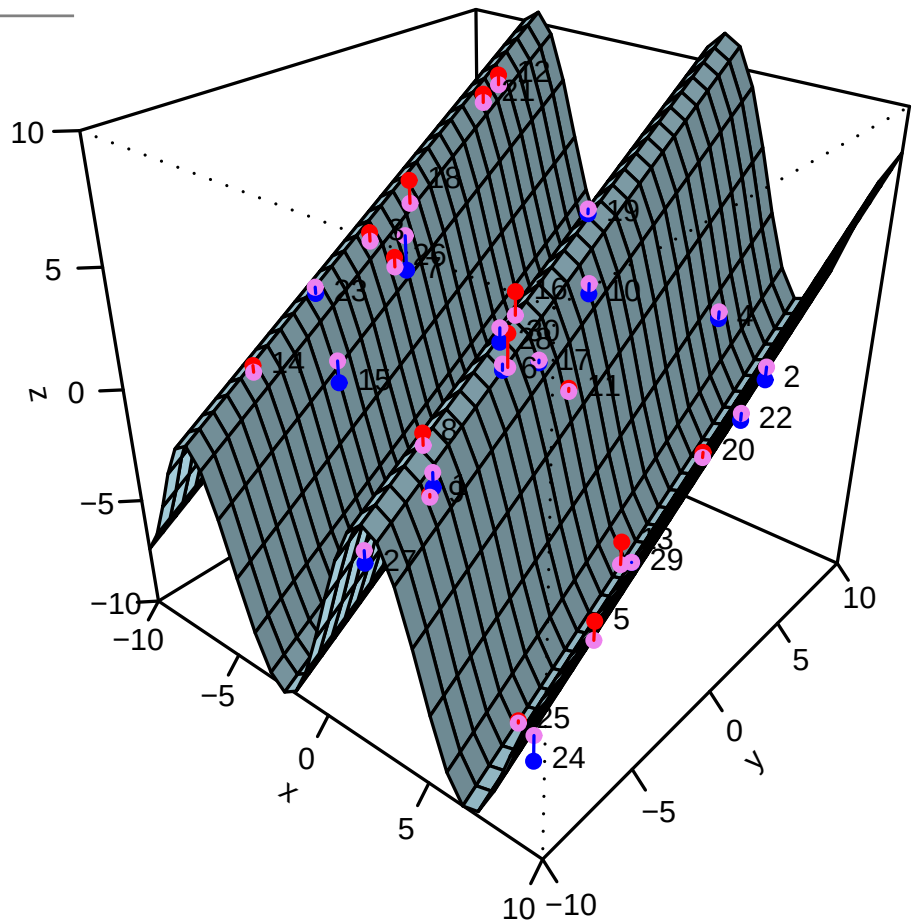
with **linear relevance parameters** $\beta_1, \ldots, \beta_d$

Squared exponential (Gaussian) covariance part

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

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

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

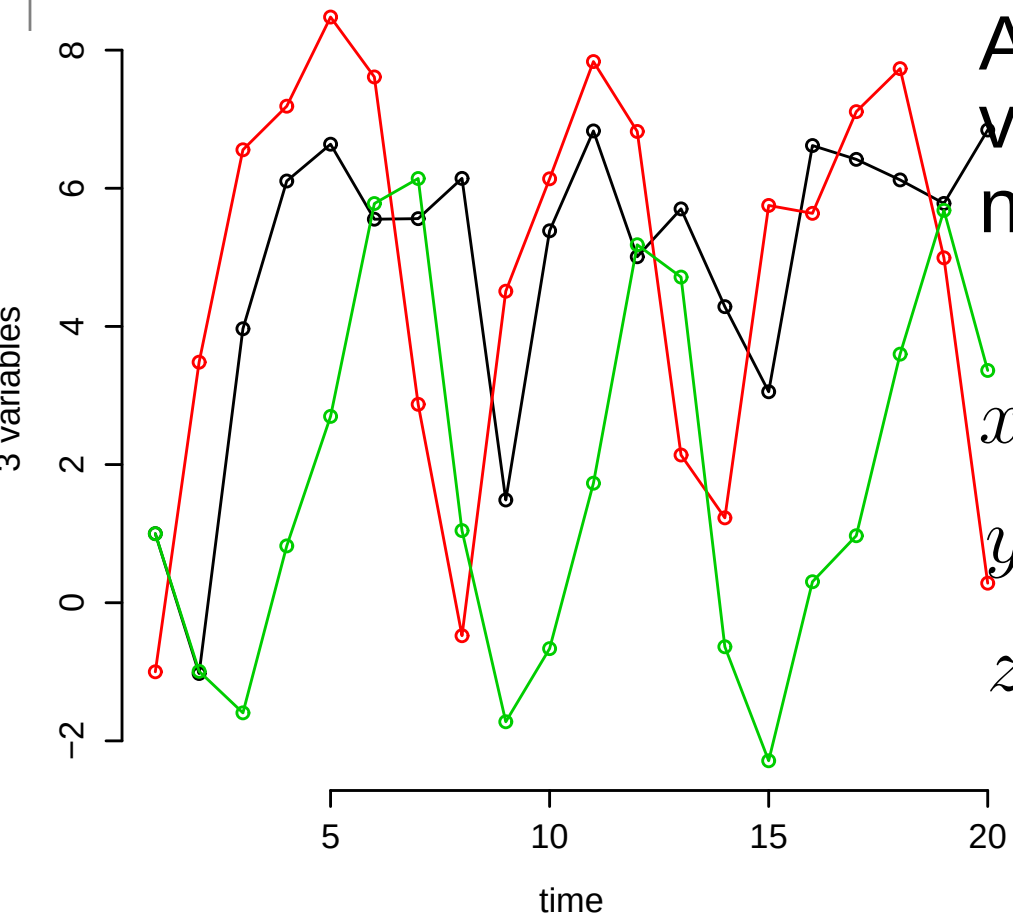
y_g^t: expression level of gene g at t

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

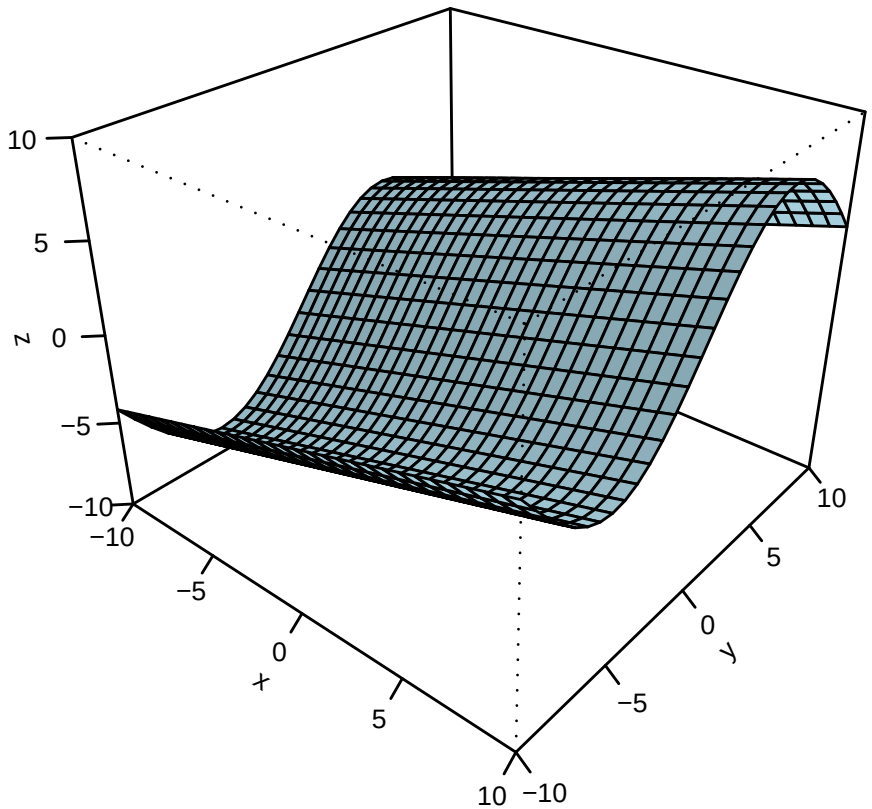
$$x_{t+1} = 0.35x_t + 5\sin(0.3y_t) + \epsilon_1$$

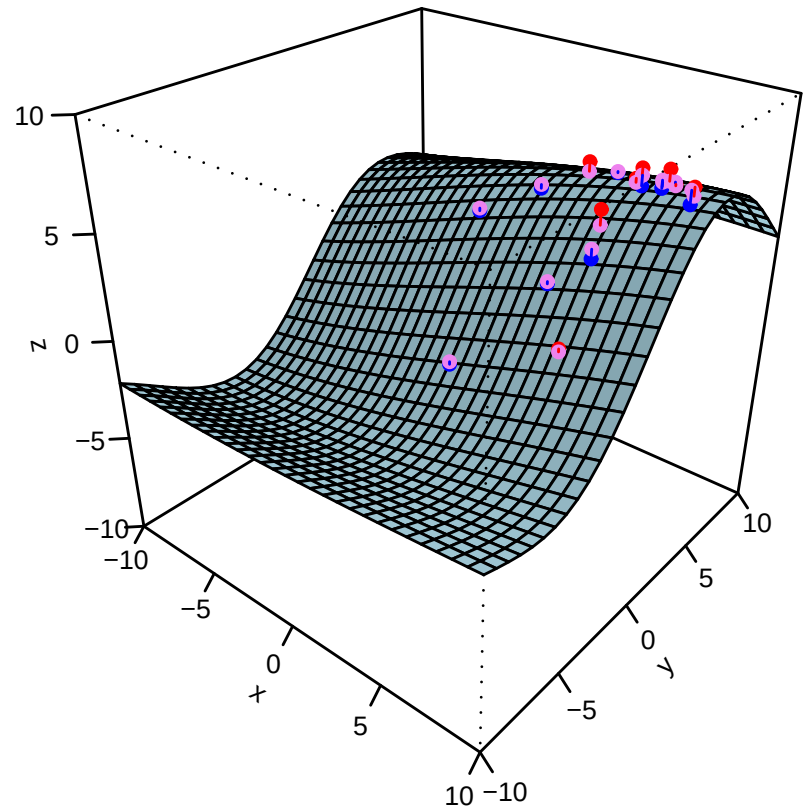
$$y_{t+1} = 0.4y_t + 5\cos(0.3z_t) + \epsilon_2$$

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

Stable cycling easy to achieve with nonlinear networks

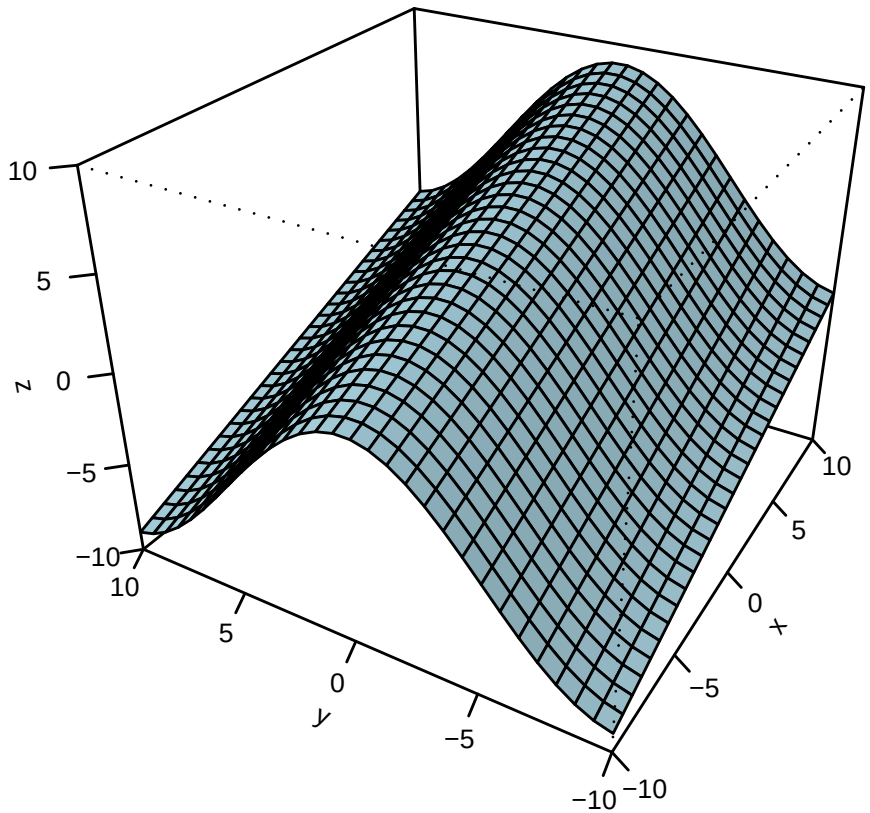
GP on time-series

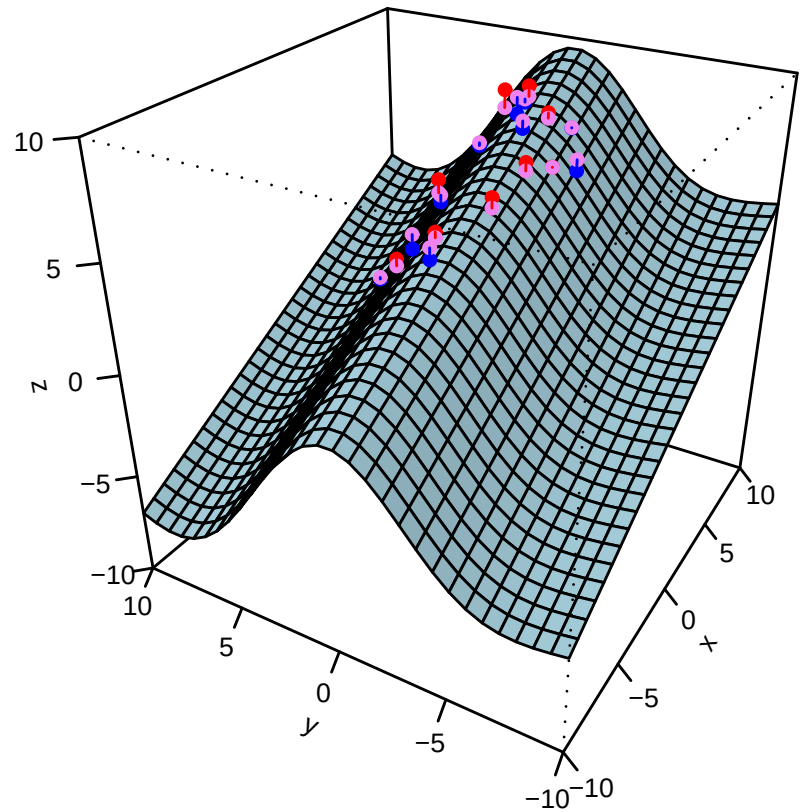




Variable 1: the linear and nonlinear relevance parameters for input 3 are both 0

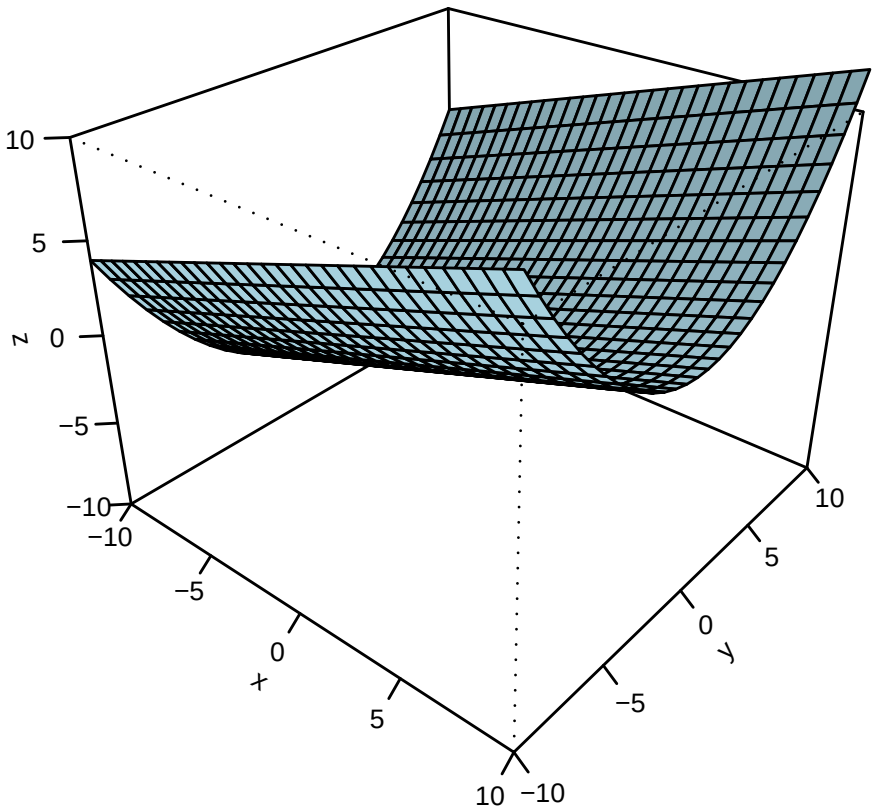
GP on time-series

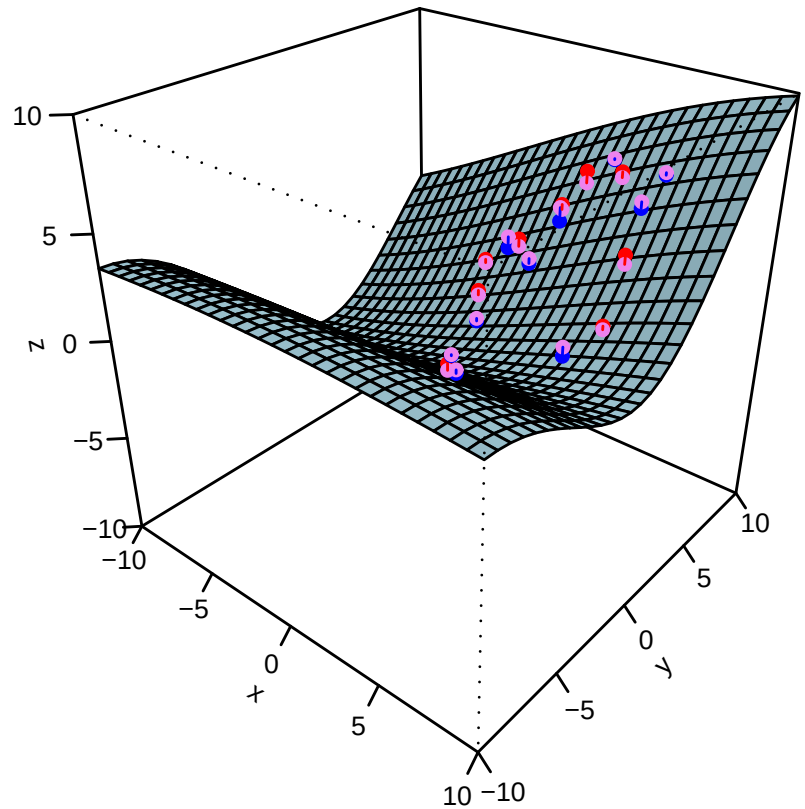




Variable 2: the linear and nonlinear relevance parameters for input 1 are both 0

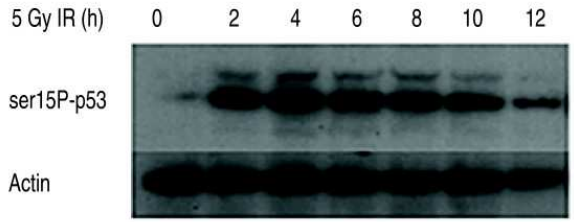
GP on time-series

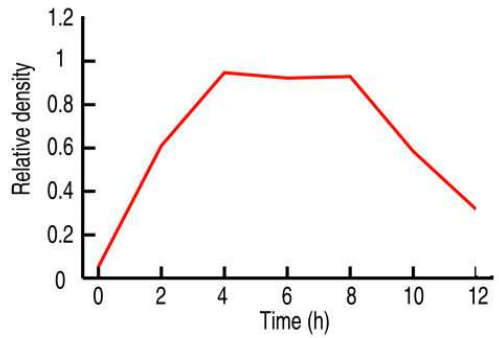


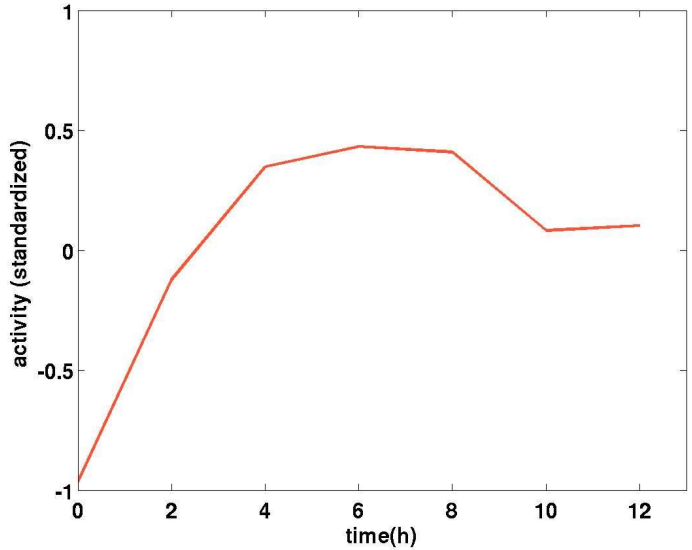


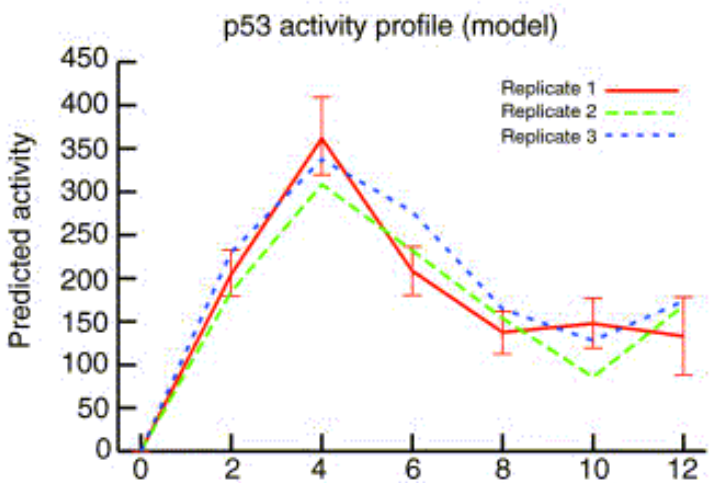
Variable 3: the linear and nonlinear relevance parameters for input 1 are both 0

Reconstruction of p53 activity







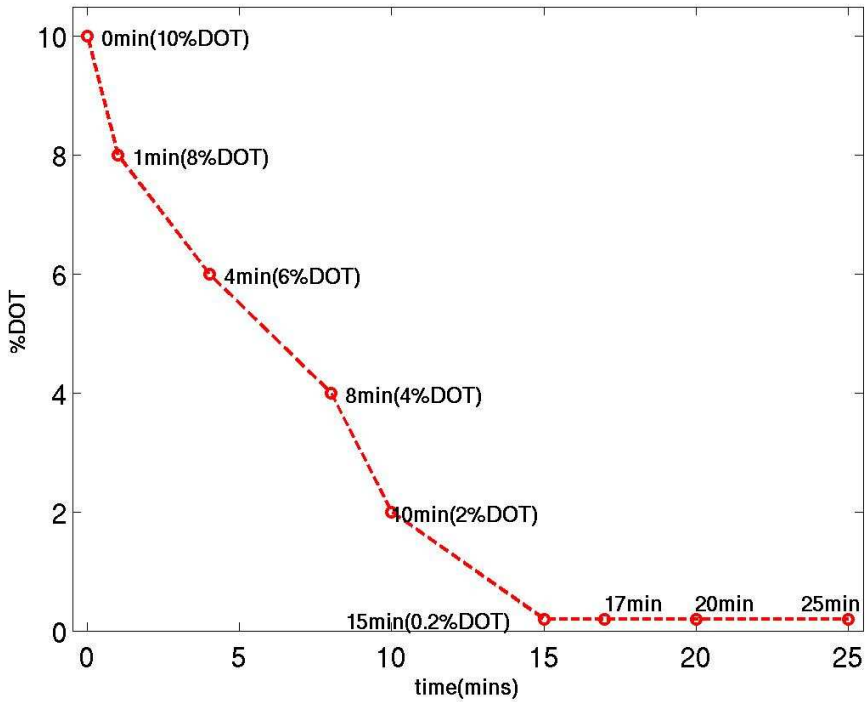


Reconstruction by Gaussian process dynamical model with 1 hidden factor

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

Left: reconstruction by *Barenco et al.*

DosR induction in *M. tuberculosis*

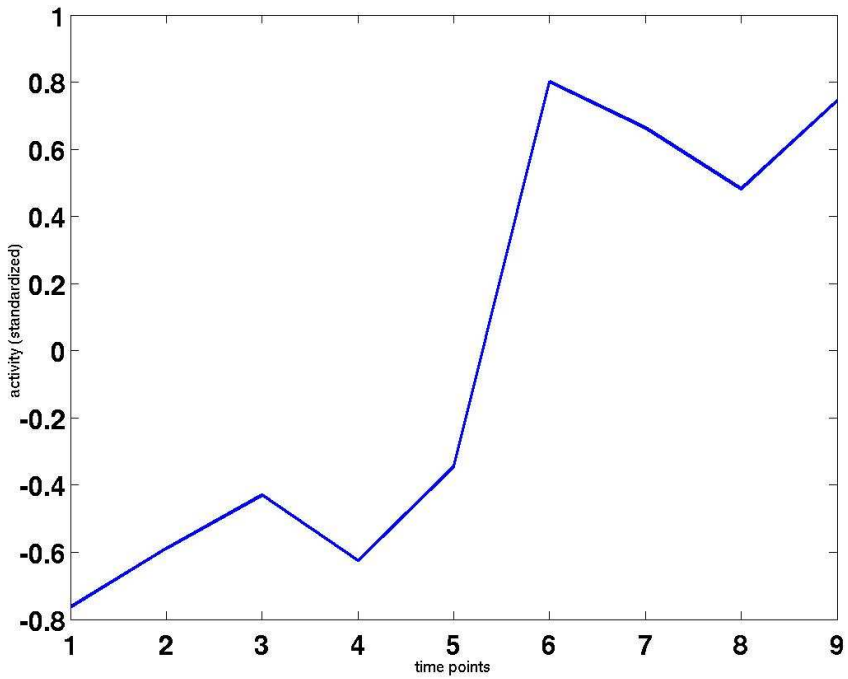


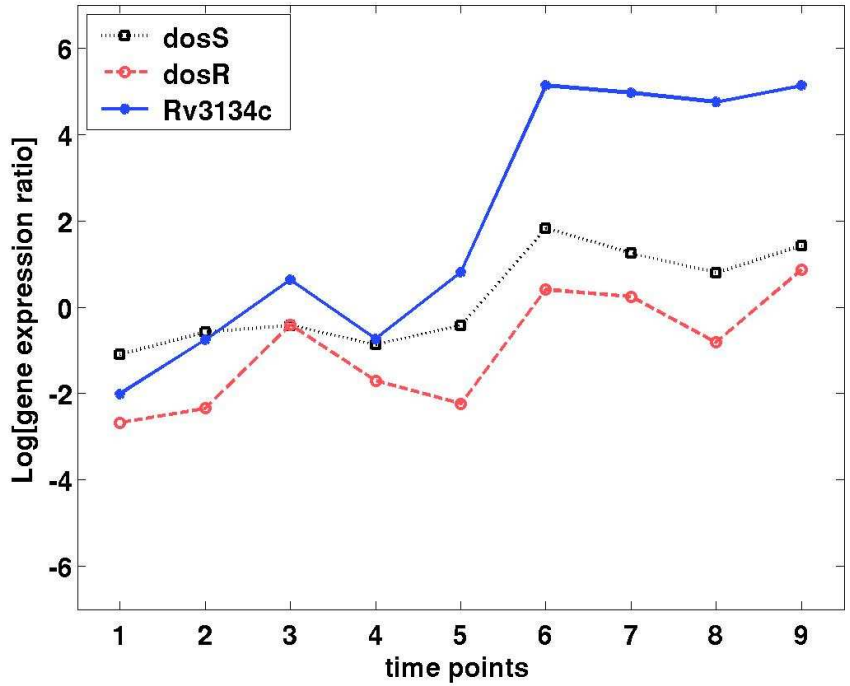
- DosR induced by various signals, including low oxygen
- Regulates about 30 other genes directly, including itself
- Known binding motif

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

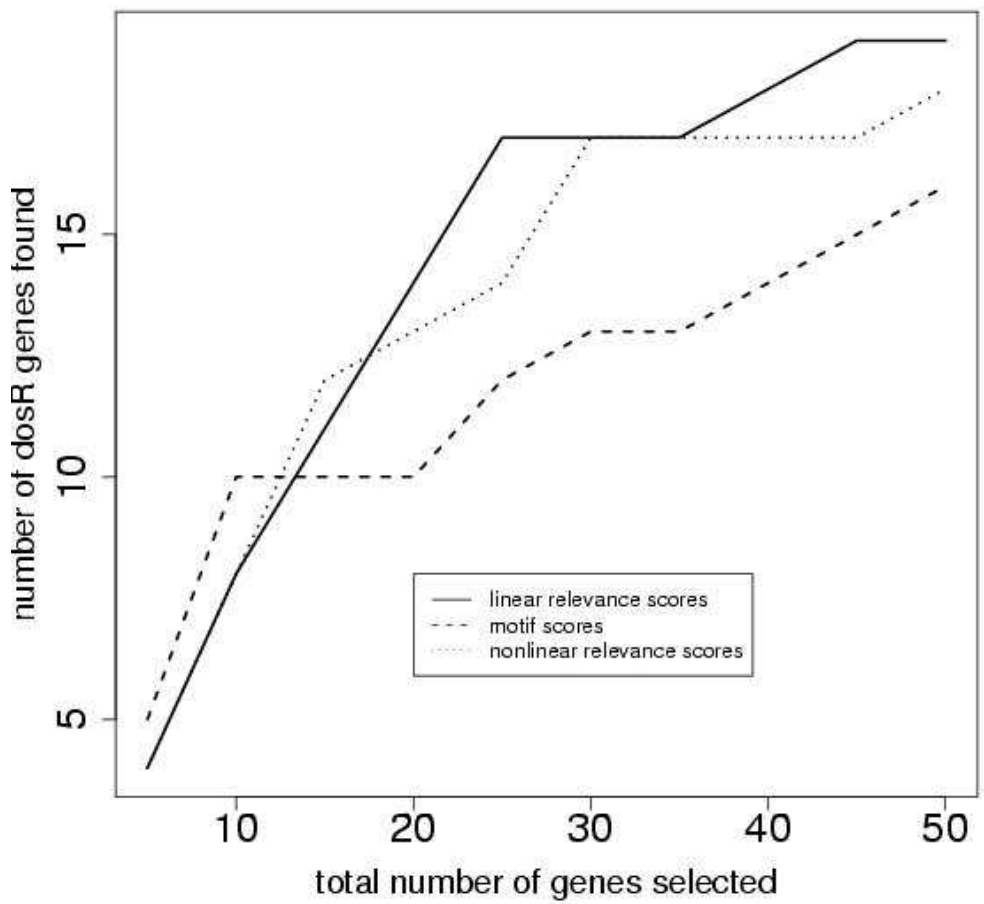
TFA and expression profile different (closer to RV3134c)

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:

$$\log \frac{p_g}{1 - p_g} = w_0 + w_1 \beta^g_{\text{lin}} + w_2 \beta^g_{\text{nonlin}} + w_3 x^g_{\text{motif}}$$

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

Simulated data for 3 hidden factors

General setup: 3 hidden variables (TFs), 10 observed variables (genes), 30 samples

Linear data:

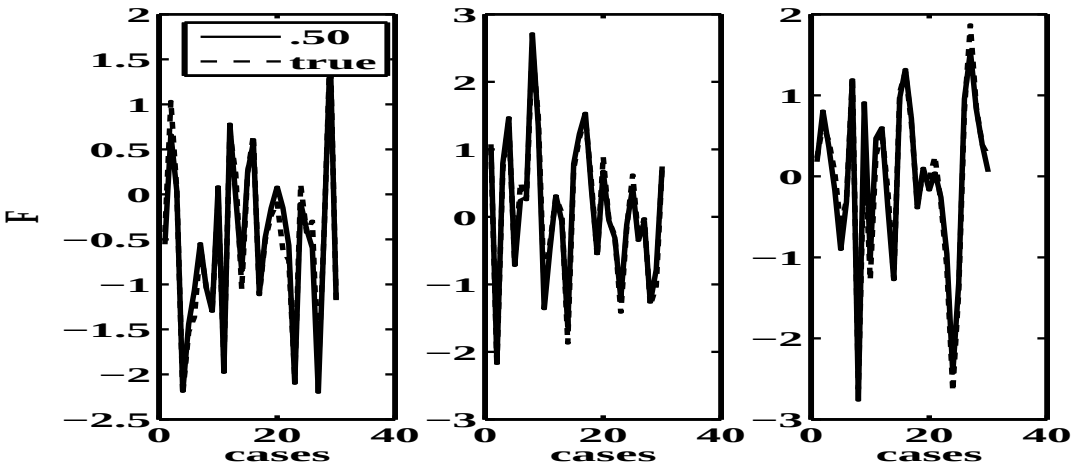
$$\begin{aligned} X &= LF + E \\ F &\sim \mathcal{N}(0,1), E \sim N(0, \mathrm{diag}(1/\mathrm{snr})) \end{aligned}$$

Nonlinear data by redefining:

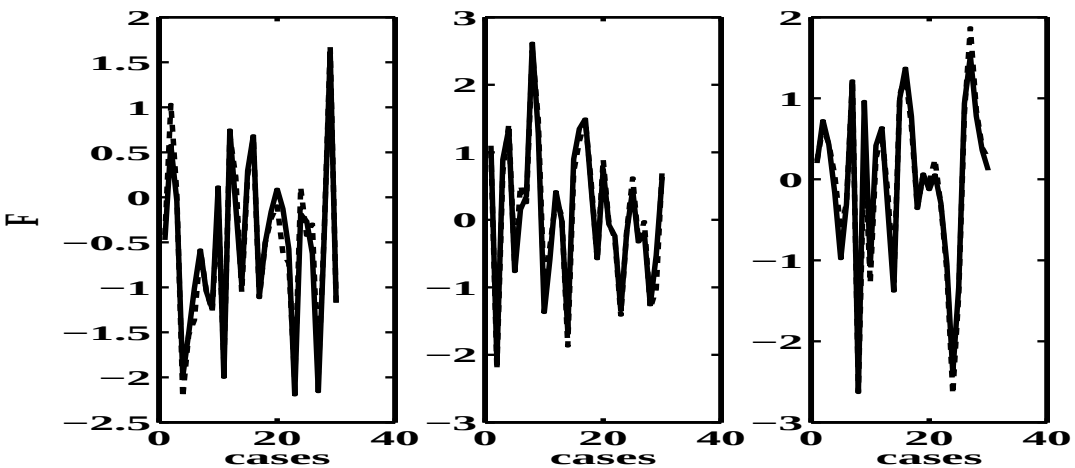
$$x_1 = f_1^3 + \epsilon_1, \quad x_4 = f_2^2 + \epsilon_4, \quad x_6 = f_2 + f_3^2 + \epsilon_6$$
$$x_7 = \cos(f_3) + \epsilon_7, \quad x_9 = f_1^2 + f_3^3 + \epsilon_9$$

From experience: every factor needs to be characterised by **at least one linear** relationship

Factor reconstruction: linear, noise 10%



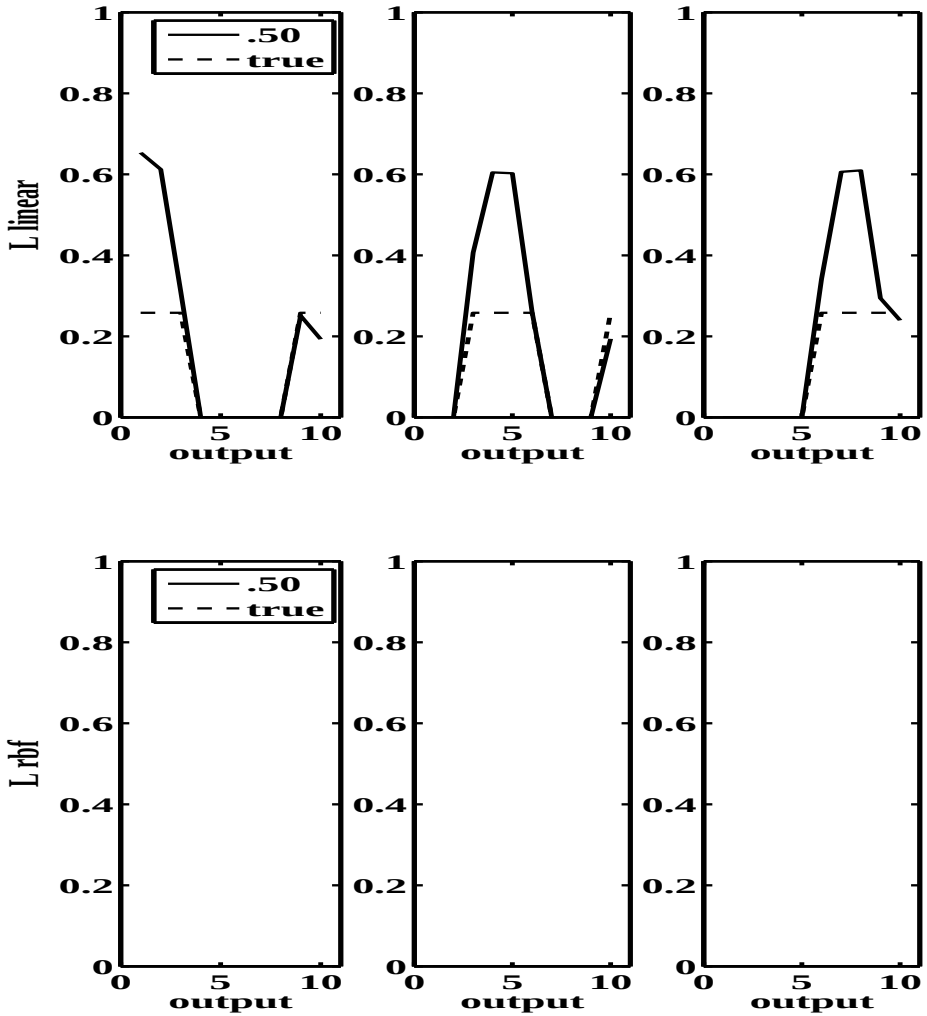
Gaussian process

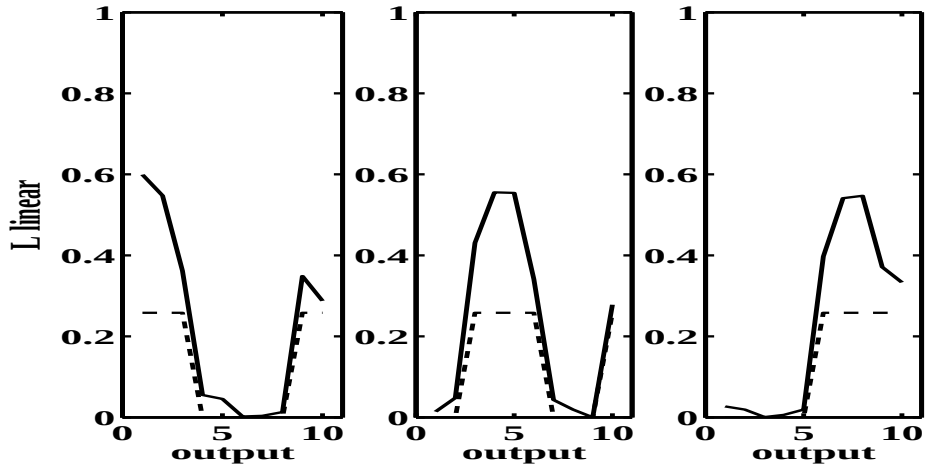


Bayesian factor analysis

very similar

Relevance: linear, noise 10%



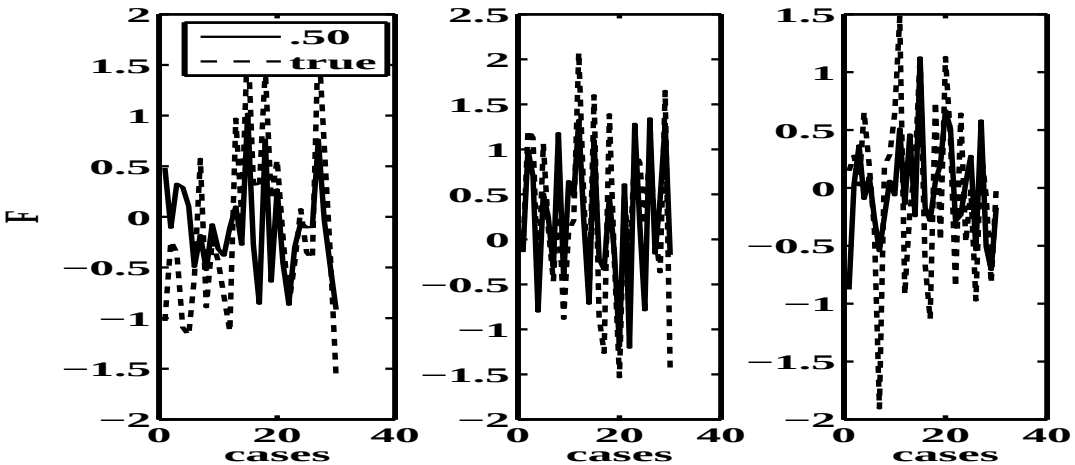


Above: Bayesian FA loadings

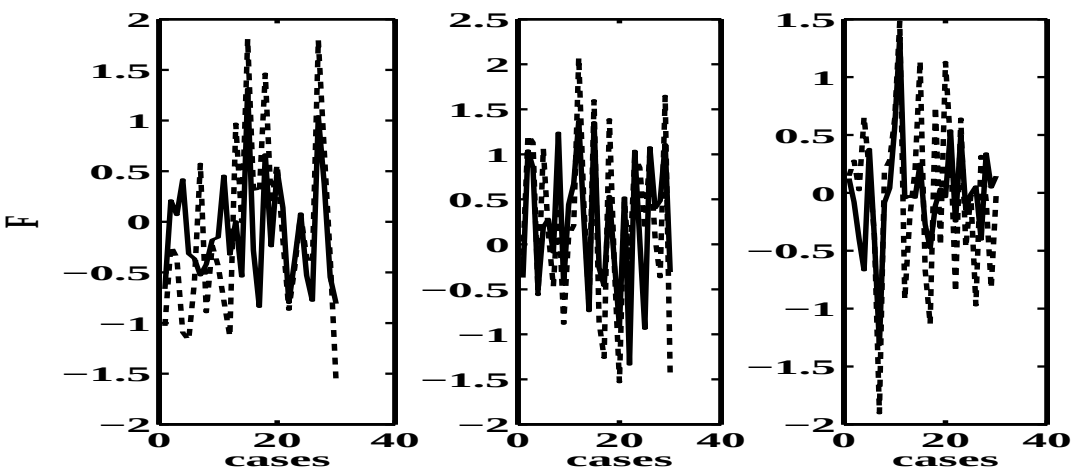
Left: GP linear and nonlinear relevance

GP slightly cleaner connectivity

Factor reconstruction: linear, noise 100%

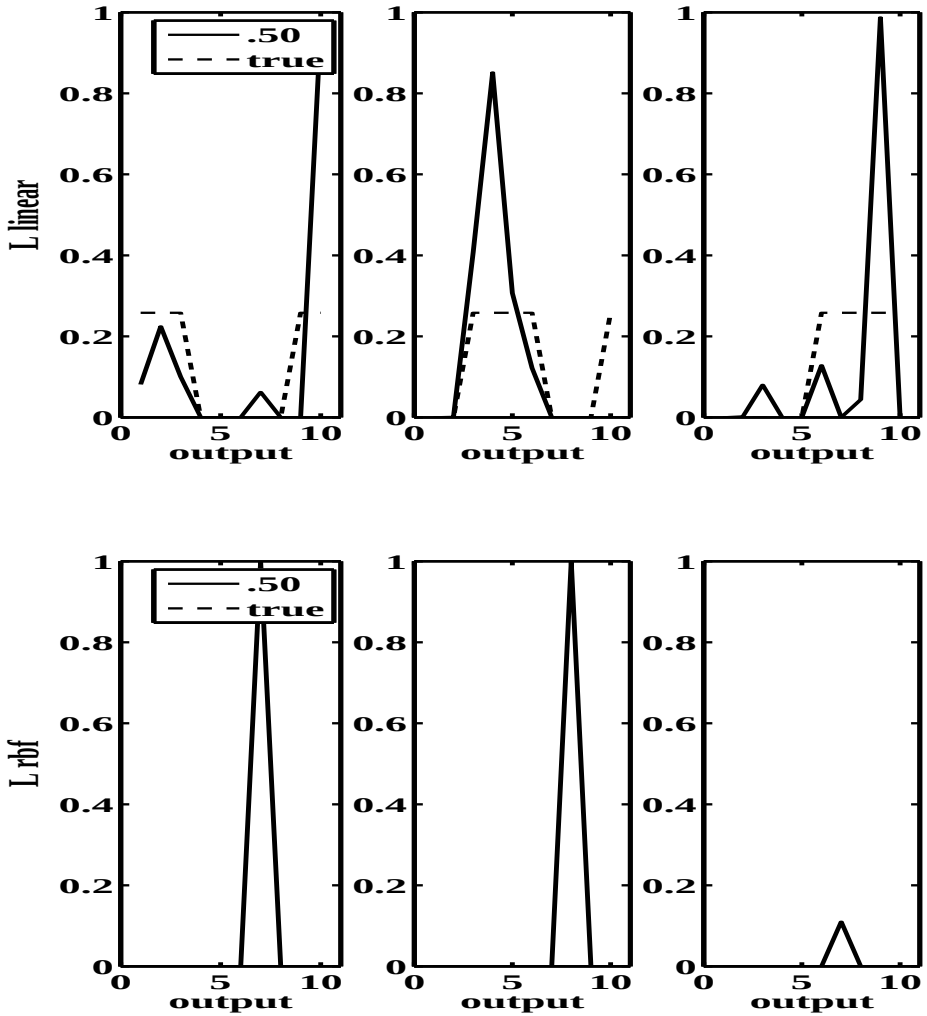


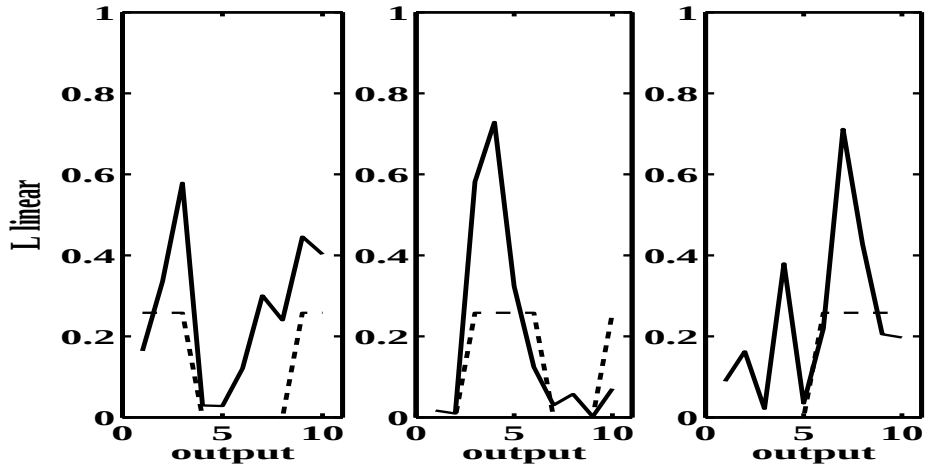
Gaussian process



Bayesian factor analysis

Relevance: linear, noise 100%

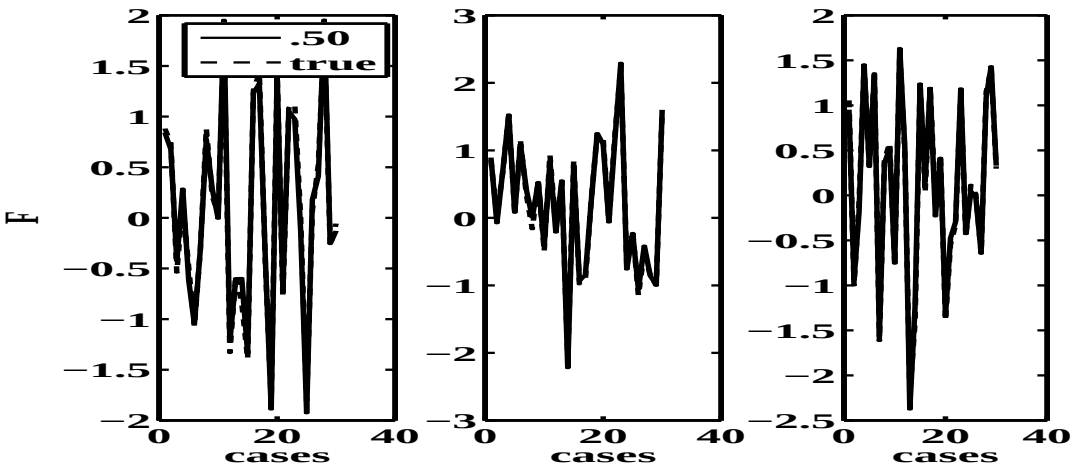




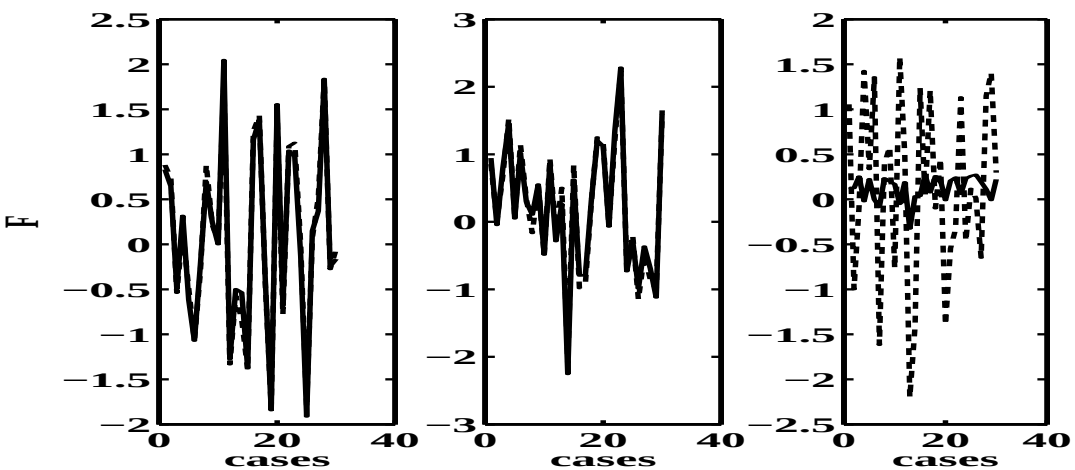
Above: Bayesian FA loadings

Left: GP linear and nonlinear relevance

Factor reconstruction: nonlinear, noise 1%

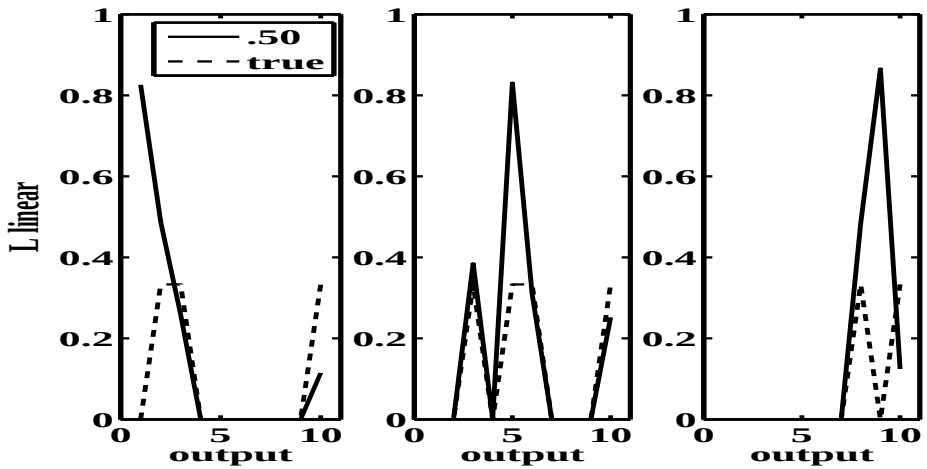


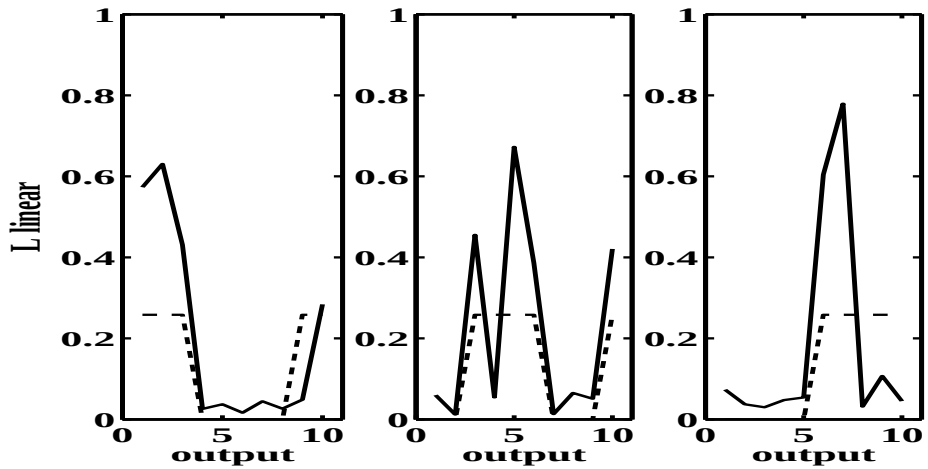
Gaussian process

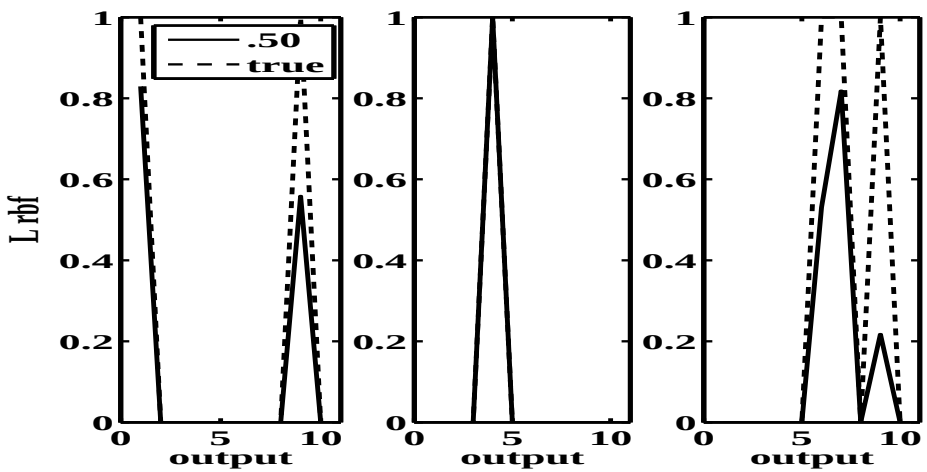


Bayesian factor analysis

Relevance: nonlinear, noise 1%



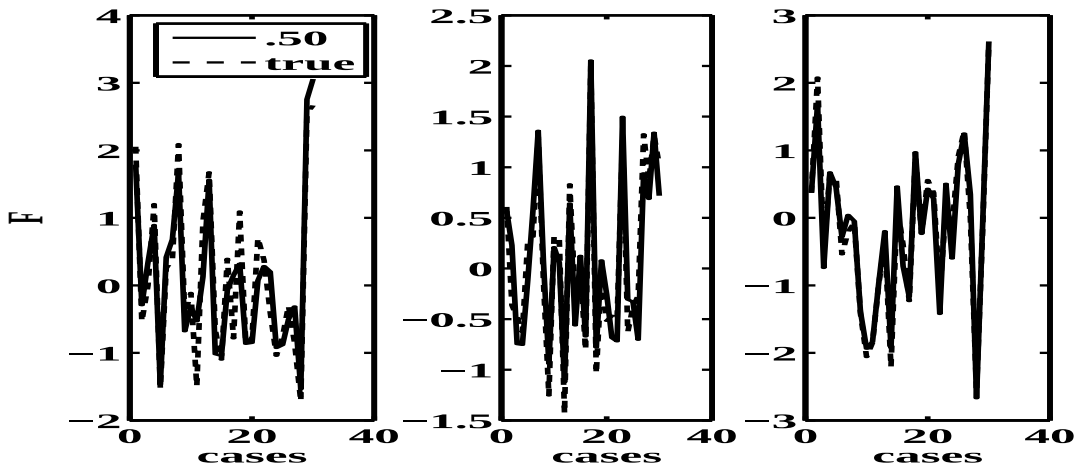




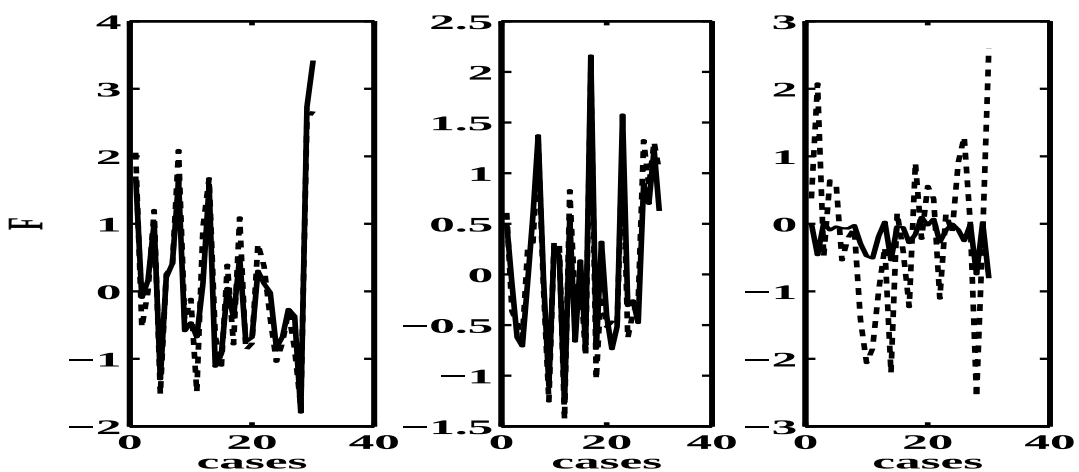
Above: Bayesian FA loadings

Left: GP linear and nonlinear relevance

Factor reconstruction: nonlinear, noise 10%

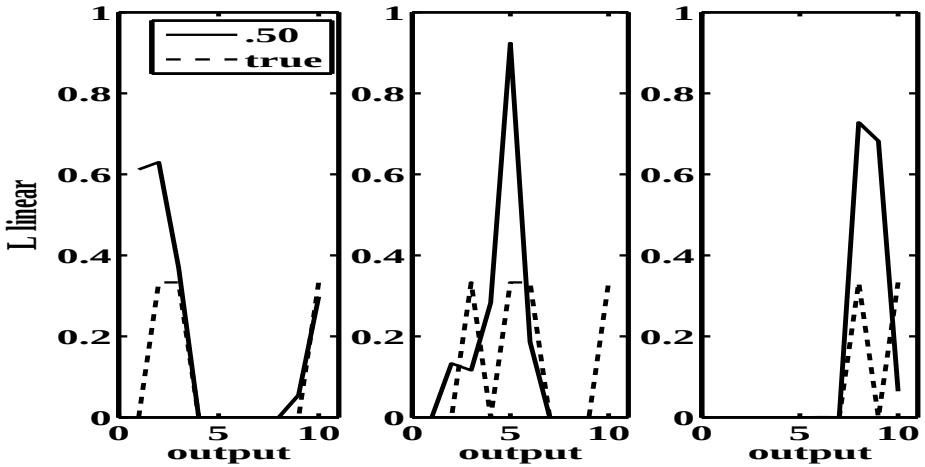


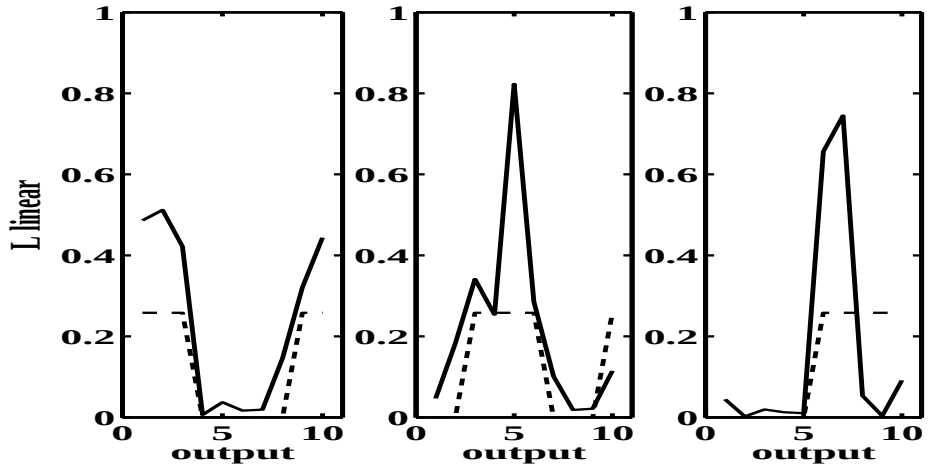
Gaussian process

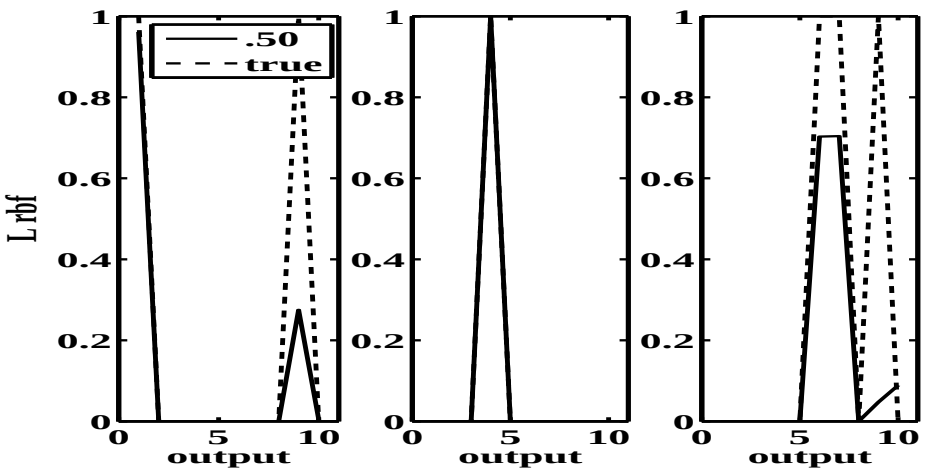


Bayesian factor analysis

Relevance: nonlinear, noise 10%



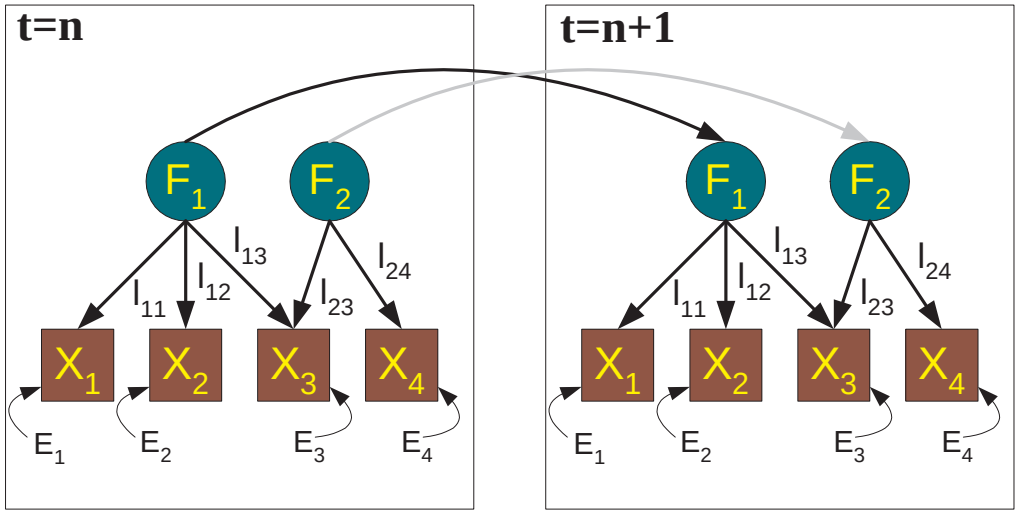




Above: Bayesian FA loadings

Left: GP linear and nonlinear relevance

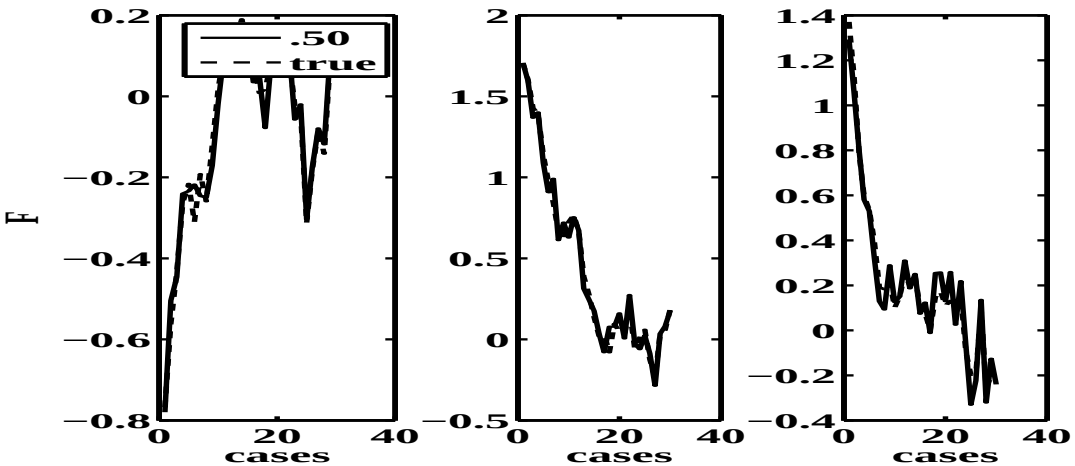
Dynamic linear model



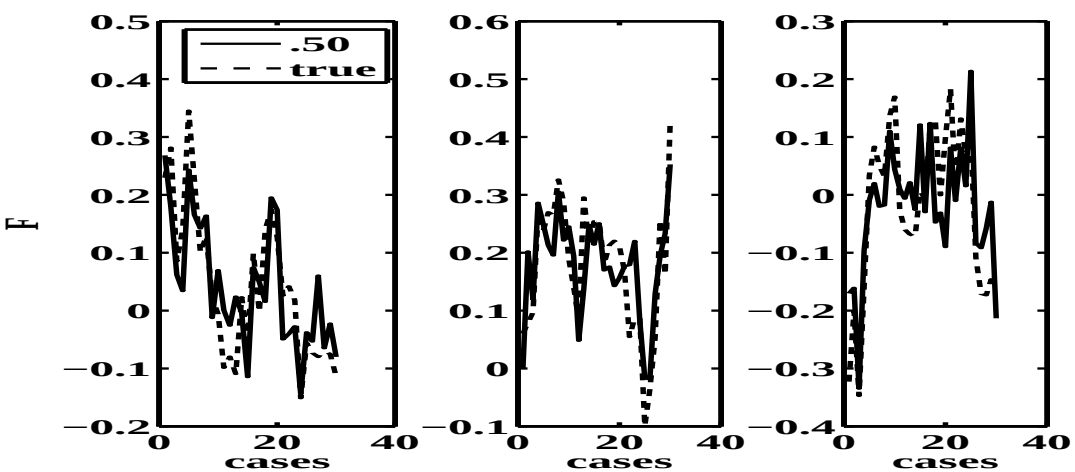
- 3 hidden variables (TFs), 10 observed variables (genes), 30 time points
- $\rho = [0.72\ 0.85\ 0.78]$

$$
\begin{aligned}
X &= LF + E \\
f_{t+1}^{k} &= \rho_k f_t^k + \upsilon
\end{aligned}
$$

Factor reconstruction: dynamic linear

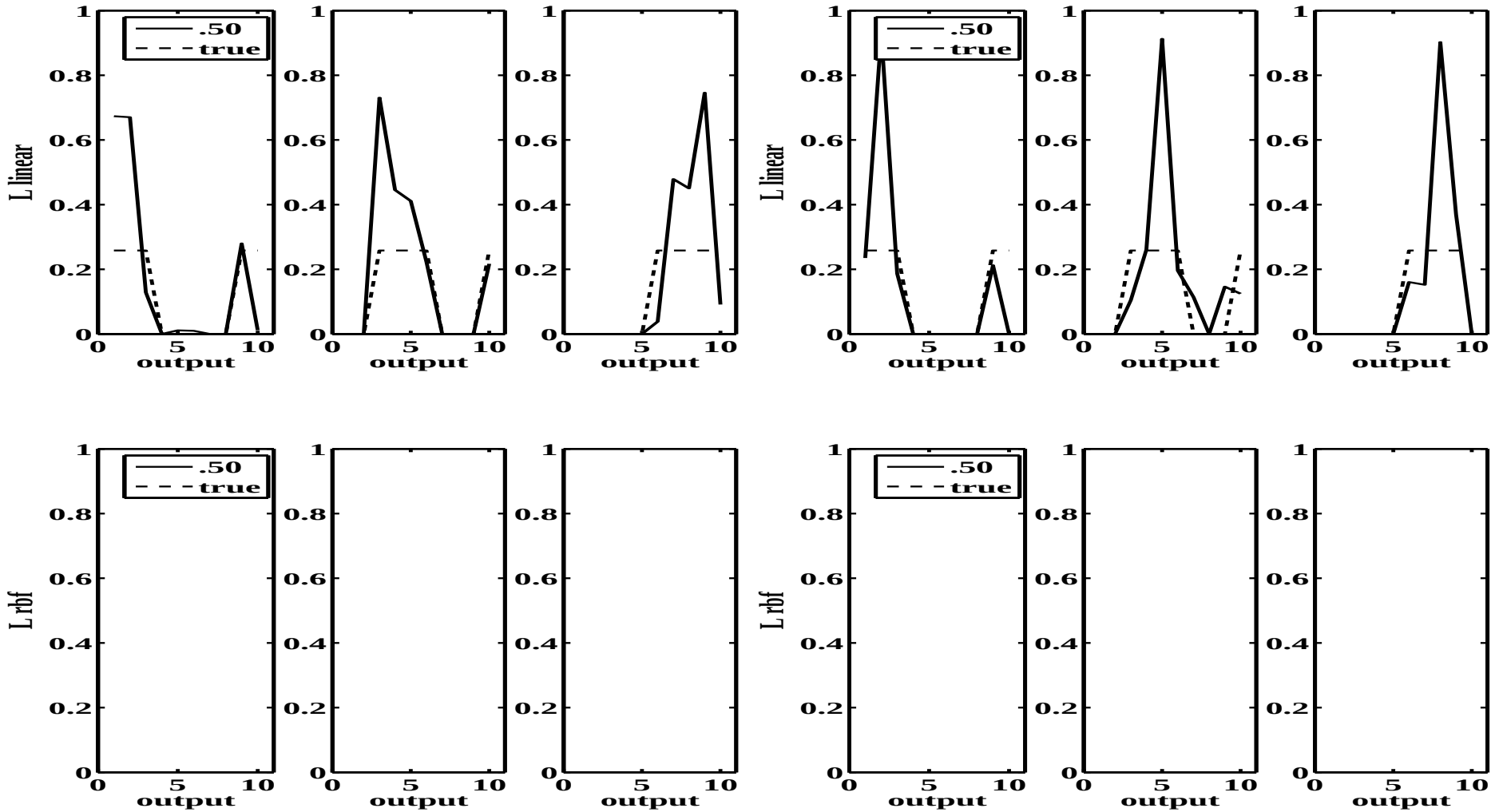


Gaussian process noise 10%



Gaussian process noise 100%

Relevance: dynamic linear



noise 10% noise 100%

Summary

- Methods for reconstructing hidden variables:
  - Factor analysis
  - Nonparametric (GP) FA
  - Nonparametric (GP) state space models
- Clear interpretation of GP **relevance parameters as TF–gene connectivity**
- **Time series** information **improves** performance dramatically
- **Identifiability** and estimation almost **impossible without linear relationships** (at least some)
- Highly flexible **Java program** developed