Probabilistic Modelling of Expression Variation in Modern eQTL Studies

Max Zwießele

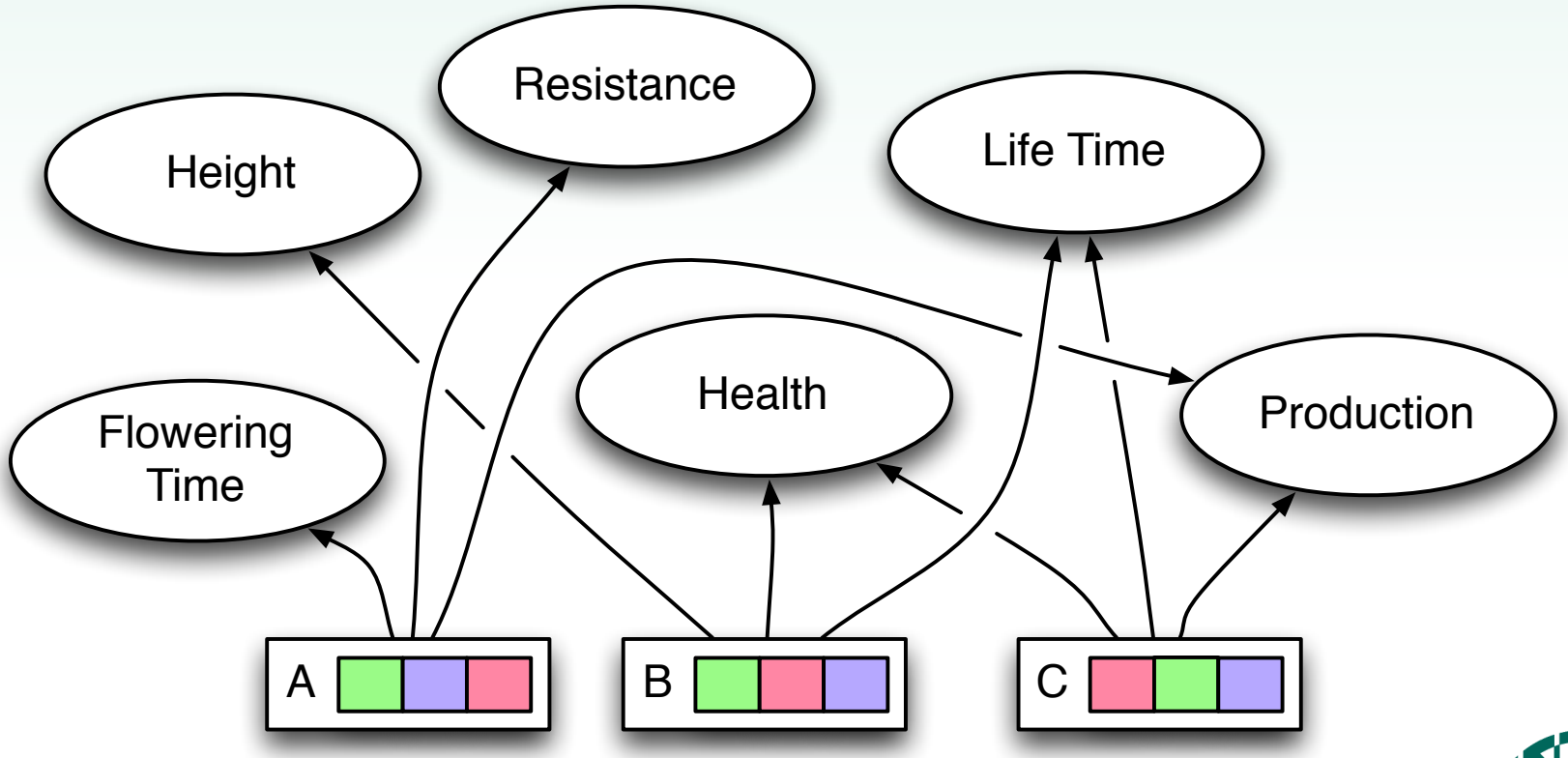
Eberhard Karls Universität Tübingen

March 24, 2013

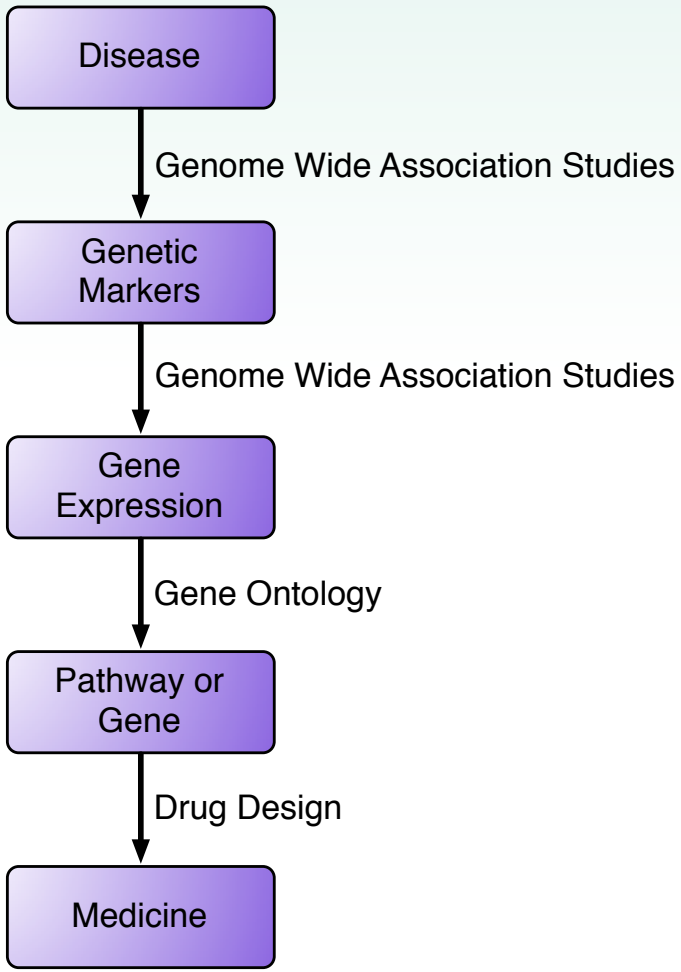
MAX-PLANCK-GESELLSCHAFT

BIOLOGISCHE KYBERNETIK

Why Genome-wide Association Studies?



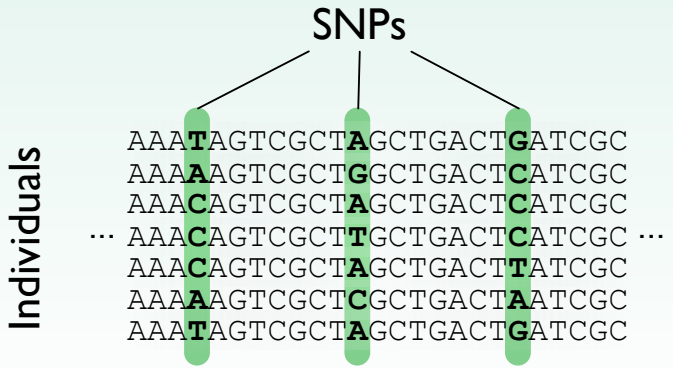
Why Genome-wide Association Studies?



Outline

Outline

Genome-wide Association Studies



Find Function f, which maps

- Genotype Variants **S**
- Gene Expression **Y**

$\Rightarrow$ $\mathbf{Y} = f(\mathbf{S})$.

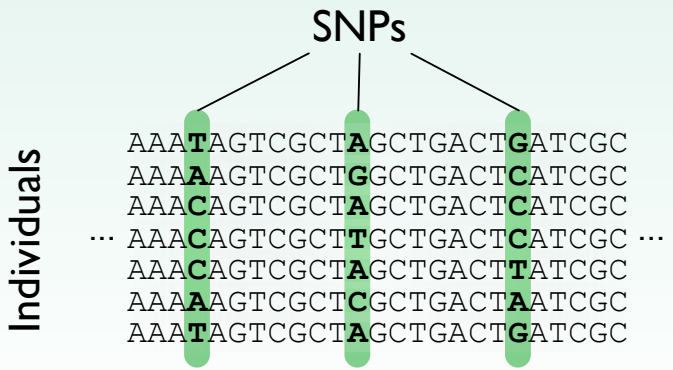
BIOLOGISCHE KYBERNETIK

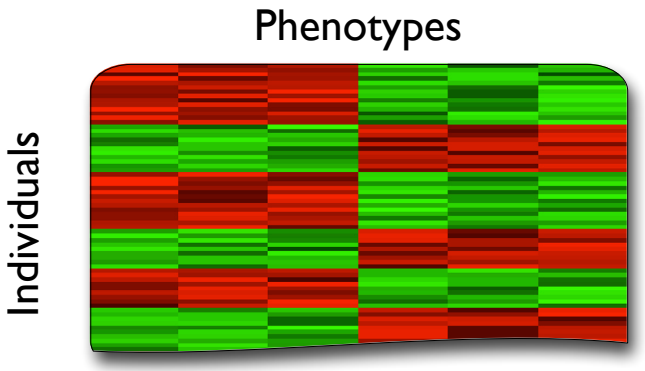
Genome-wide Association Studies

Find Function f, which maps

- Genotype Variants **S**
- Gene Expression **Y**

$\Rightarrow \mathbf{Y} = f(\mathbf{S})$.





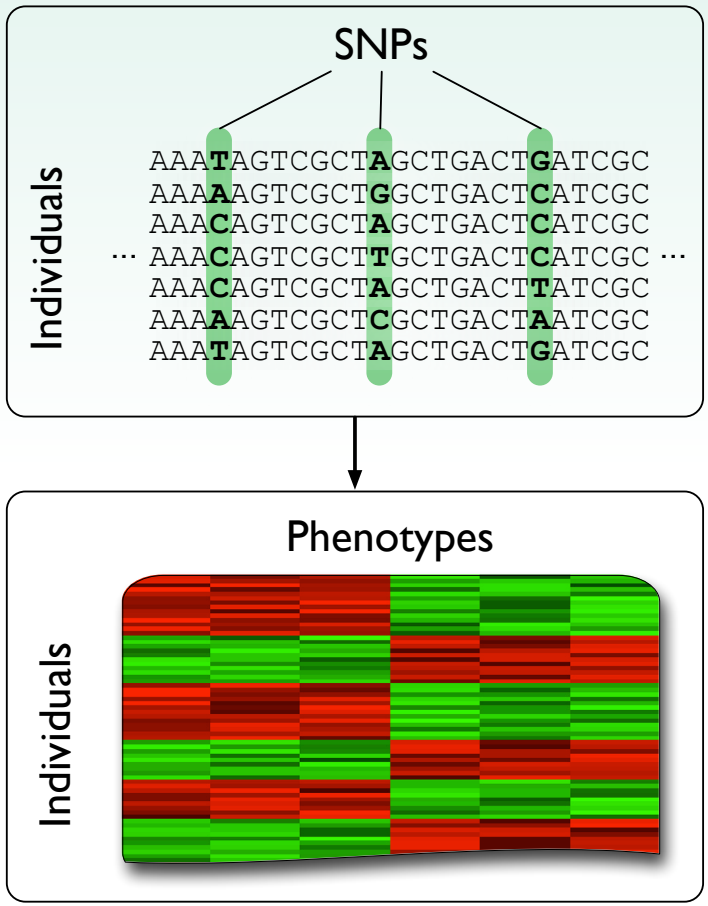
BIOLOGISCHE KYBERNETIK

Genome-wide Association Studies

Find Function f, which maps

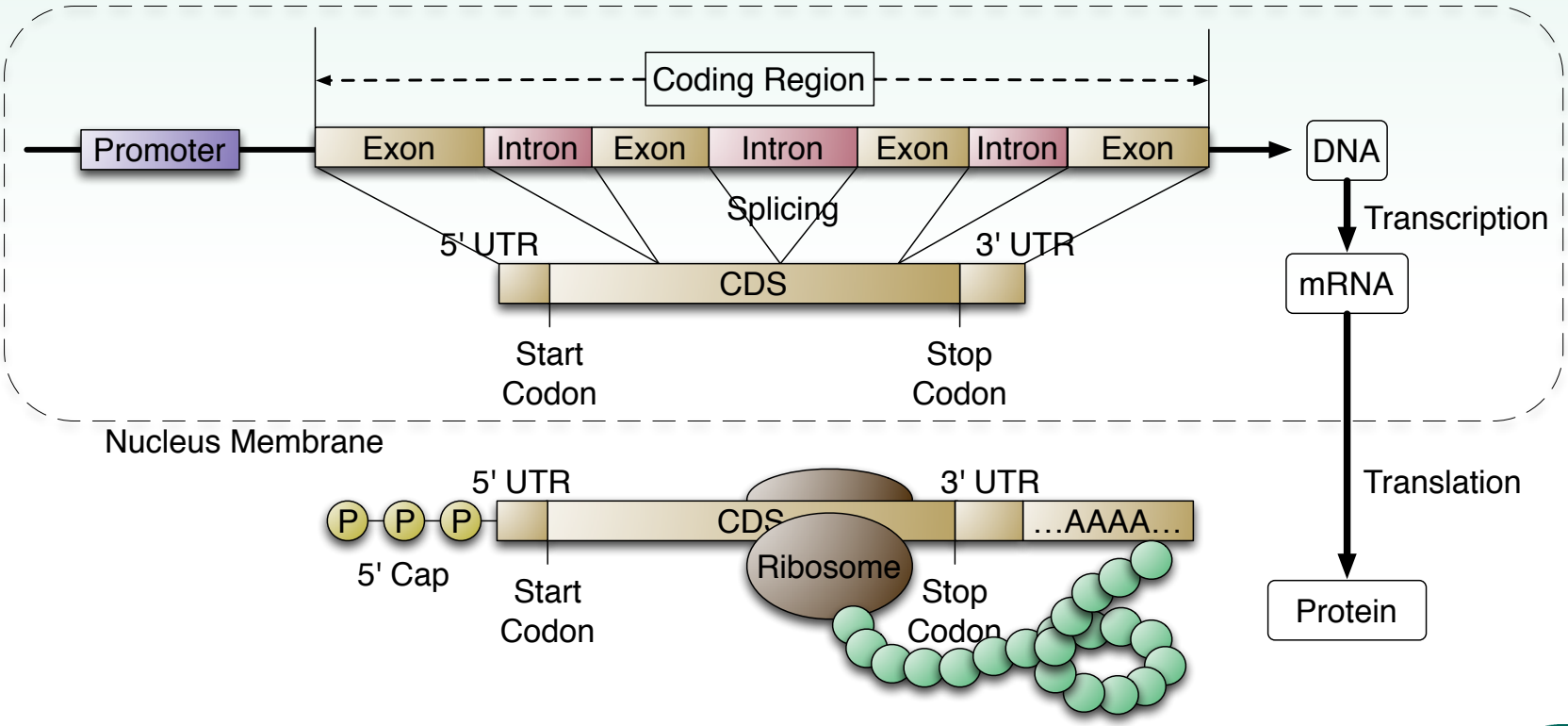
- Genotype Variants $\mathbf{S}$
- Gene Expression $\mathbf{Y}$

$\Rightarrow \mathbf{Y} = f(\mathbf{S})$.



BIOLOGISCHE KYBERNETIK

Gene Expression as quantitative Trait



BIOLOGISCHE KYBERNETIK

Confounding Variation (for more see [3])

Confounders:

"Non Genotype Variant Influences on Gene Expression"

Genetically driven:

- Combinatorial Effects of Expressed Genes
- Population Stratification
- Gender

Environmentally driven:

- Treatment Control Experiments
- Exposure to Stress:
  - Physical (pressure, height...)
  - Chemical
  - Human Made (Sea Water quality)

Confounding Variation (for more see [3])

Confounders:

"Non Genotype Variant Influences on Gene Expression"

Genetically driven:

- Combinatorial Effects of Expressed Genes
- Population Stratification
- Gender

Environmentally driven:

- Treatment Control Experiments
- Exposure to Stress:
  - Physical (pressure, height...)
  - Chemical
  - Human Made (Sea Water quality)

Confounding Variation (for more see [3])

Confounders:

"Non Genotype Variant Influences on Gene Expression"

Genetically driven:

- Combinatorial Effects of Expressed Genes
- Population Stratification
- Gender

Environmentally driven:

- Treatment Control Experiments
- Exposure to Stress:
  - Physical (pressure, height...)
  - Chemical
  - Human Made (Sea Water quality)

Confounding Variation (for more see [3])

Confounders:

"Non Genotype Variant Influences on Gene Expression"

Genetically driven:

- Combinatorial Effects of Expressed Genes
- Population Stratification
- Gender

Environmentally driven:

- Treatment Control Experiments
- Exposure to Stress:
  - Physical (pressure, height...)
  - Chemical
  - Human Made (Sea Water quality)

Confounding Variation (for more see [3])

Confounders:

"Non Genotype Variant Influences on Gene Expression"

Genetically driven:

- Combinatorial Effects of Expressed Genes
- Population Stratification
- Gender

Environmentally driven:

- Treatment Control Experiments
- Exposure to Stress:
  - Physical (pressure, height...)
  - Chemical
  - Human Made (Sea Water quality)

BIOLOGISCHE KYBERNETIK

Confounders and Probabilistic Modelling

Confounding Variation not known:

- Number Q (Dimension) of Confounders
- Abundance of Variation $\mathbf{X}$ itself
- Combinatorial Effects uncertain

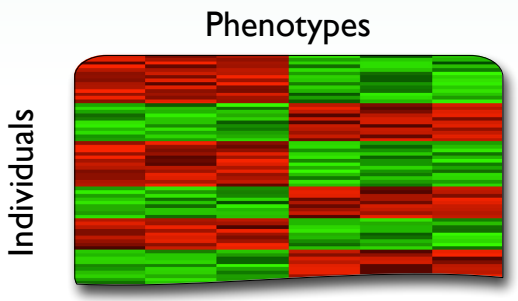
$\Rightarrow$ Bayesian Modelling of Confounders (Latent Variables)

Outline

Notation

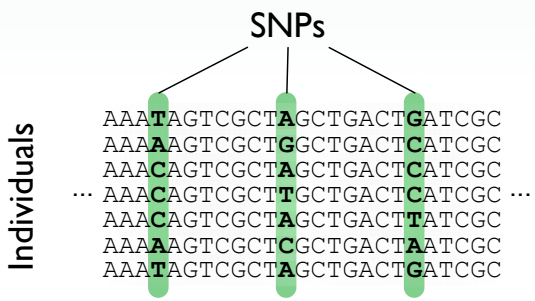
Expression Matrix:

$$\mathbf{Y} \in \mathbb{R}^{N \times D}$$



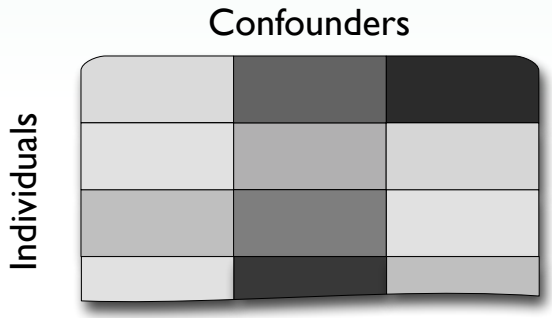
SNP Matrix:

$$\mathbf{S} \in \mathbb{R}^{N \times K}$$



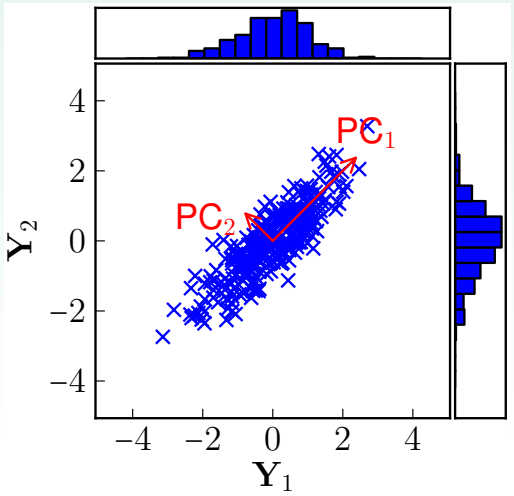
Confounder Matrix:

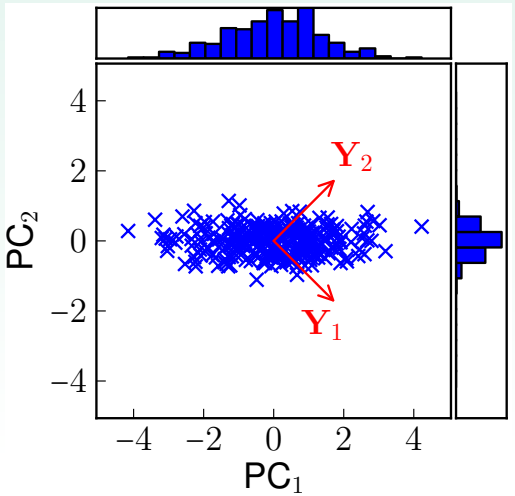
$$\mathbf{X} \in \mathbb{R}^{N \times Q}$$

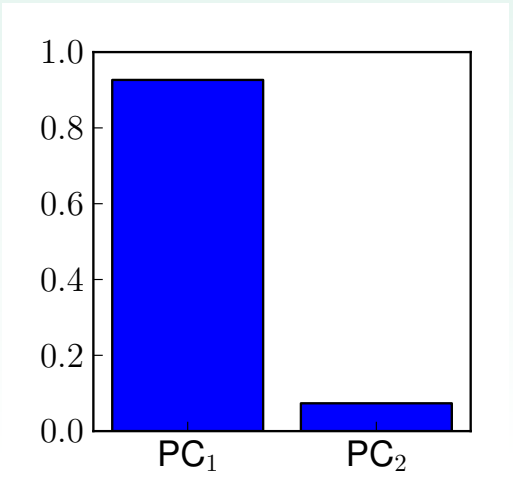


N : Samples D : Genes K : SNPs Q : Confounders

Principal Component Analysis







$$\mathbf{Y} = \mathbf{XV} \quad , \text{ where}$$

$$\mathbf{V}^{-1} = \text{eig}(\mathbf{Y}^\top \mathbf{Y})$$

$$\mathbf{X} = \mathbf{Y}\mathbf{V}_Q^{-1} \quad , \text{ where}$$

$$\mathbf{V}_Q^{-1} = \{\mathbf{v}_q : 1 \leq q \leq Q \text{ largest eigenvalues}\}$$

Gaussian Process Latent Variable Model [4, 5]

Consider Linear Generative Model:

$$\mathbf{Y} = \mathbf{X}\mathbf{V} \tag{1}$$

Prior over $\mathbf{V}$, instead of $\mathbf{X}$:

$$p(\mathbf{V}) = \prod_{d=1}^{D} \mathcal{N}(\mathbf{v}_d | \mathbf{0}, \mathbf{I}_Q) \tag{2}$$

Integrate out $\mathbf{V}$ to get:

$$p(\mathbf{Y}|\mathbf{X}) = \int p(\mathbf{Y}|\mathbf{X}, \mathbf{V}) p(\mathbf{V}|\mathbf{X})\, d\mathbf{V} \tag{3}$$

$$= \prod_{d=1}^{D} \mathcal{N}(\mathbf{y}_d | \mathbf{0}, \mathbf{K}^{\mathrm{NN}} + \beta^{-1}\mathbf{I}_N) \quad , \text{ where} \tag{4}$$

$\mathbf{K}_{ij}^{NN} = k(\mathbf{x}_i, \mathbf{x}_j)$ Covariance Function.

BIOLOGISCHE KYBERNETIK

Gaussian Process Latent Variable Model [4, 5]

Learn Latent Variables $\mathbf{X}$ through Maximum A Posteriori (MAP):

$$\{\hat{\mathbf{X}}, \hat{\boldsymbol{\theta}}\}_{\text{MAP}} = \arg\max_{\mathbf{X},\boldsymbol{\theta}} \ln p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) \quad , \text{ where} \tag{5}$$

$$\ln p(\mathbf{Y}|\mathbf{X}) = -\frac{DN}{2}\ln(2\pi) - \frac{D}{2}\ln|\mathbf{K}| - \frac{1}{2}\text{tr}(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^{\mathsf{T}}) \quad , \text{ with} \tag{6}$$

$$\mathbf{K} = \mathbf{K}^{NN} + \beta^{-1}\mathbf{I}_N \tag{7}$$

But, we want Bayesian estimate

$$p(\mathbf{Y}) = \int p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})\, d\mathbf{X} \tag{8}$$

BIOLOGISCHE KYBERNETIK

Variational Bayesian GPLVM [1]

$$p(\mathbf{Y}) = \int p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})\, d\mathbf{X} \tag{9}$$

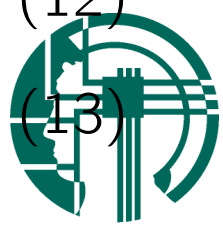
Intractable in $\mathbf{X} \Rightarrow$ Variational Approximation $q(\mathbf{X})$:

$$\ln p(\mathbf{Y}) \geq \mathcal{F}(q) = \int q(\mathbf{X}) \ln \frac{p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})}{q(\mathbf{X})} d\mathbf{X} \tag{10}$$

$$= \int q(\mathbf{X}) \ln p(\mathbf{Y}|\mathbf{X}) d\mathbf{X} - \int q(\mathbf{X}) \ln \frac{p(\mathbf{X})}{q(\mathbf{X})} d\mathbf{X} \tag{11}$$

$$= \int q(\mathbf{X}) \ln p(\mathbf{Y}|\mathbf{X}) d\mathbf{X} - \mathrm{KL}(q||p) \tag{12}$$

$$= \tilde{\mathcal{F}}(q) - \mathrm{KL}(q||p) \tag{13}$$

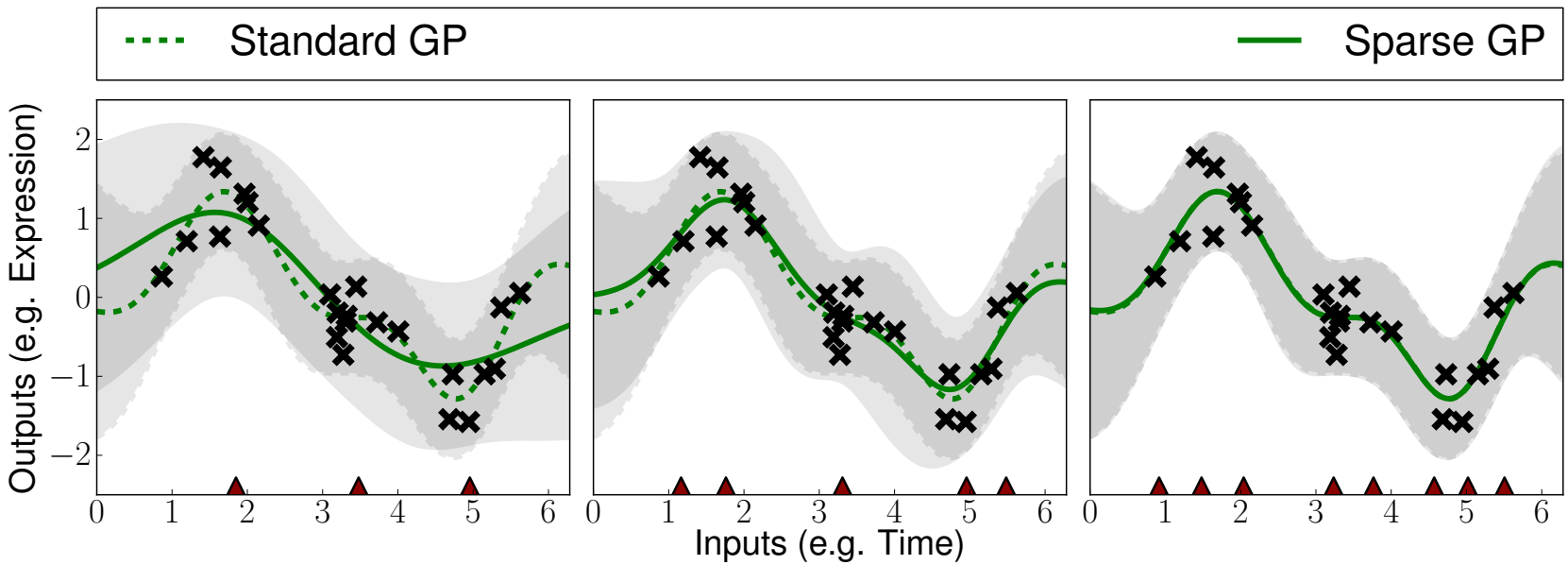


Variational Bayesian GPLVM [1]

$$\ln p(\mathbf{Y}) \geq \mathcal{F}(q) = \tilde{\mathcal{F}}(q) - \mathrm{KL}(q||p) \tag{14}$$

$$\tilde{\mathcal{F}}(q) = \int q(\mathbf{X}) \ln p(\mathbf{Y}|\mathbf{X}) d\mathbf{X} \tag{15}$$

Sparse Approximation for $p(\mathbf{Y}|\mathbf{X})$:

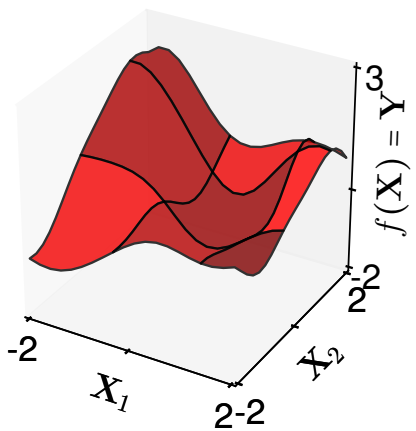


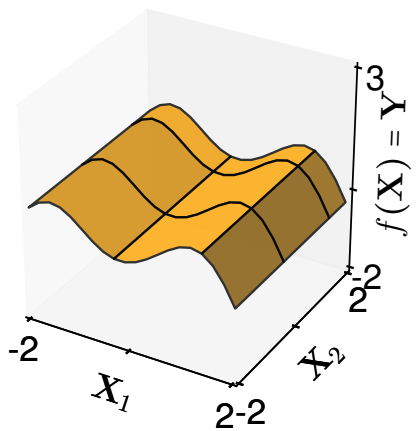
Automatic Relevance Determination [6]

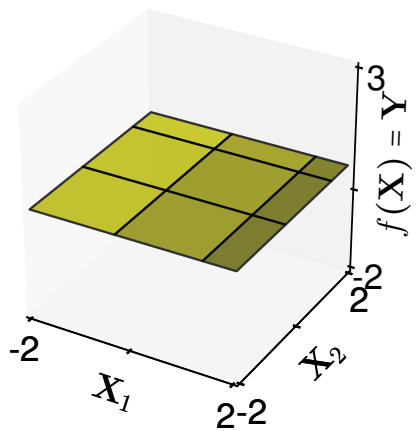
Consider Squared Exponential Covariance Function

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left\{\frac{1}{2}\boldsymbol{\alpha}^\top ||\mathbf{x} - \mathbf{x}'||^2\right\} \tag{16}$$

$\boldsymbol{\alpha} \in \mathbb{R}^Q$ contains Relevance for each Dimension $1 \leq q \leq Q$ of $\mathbf{X}$
$\mathbf{X} \in \mathbb{R}^{N \times Q}$ with rows $\mathbf{x} \in \mathbb{R}^Q$

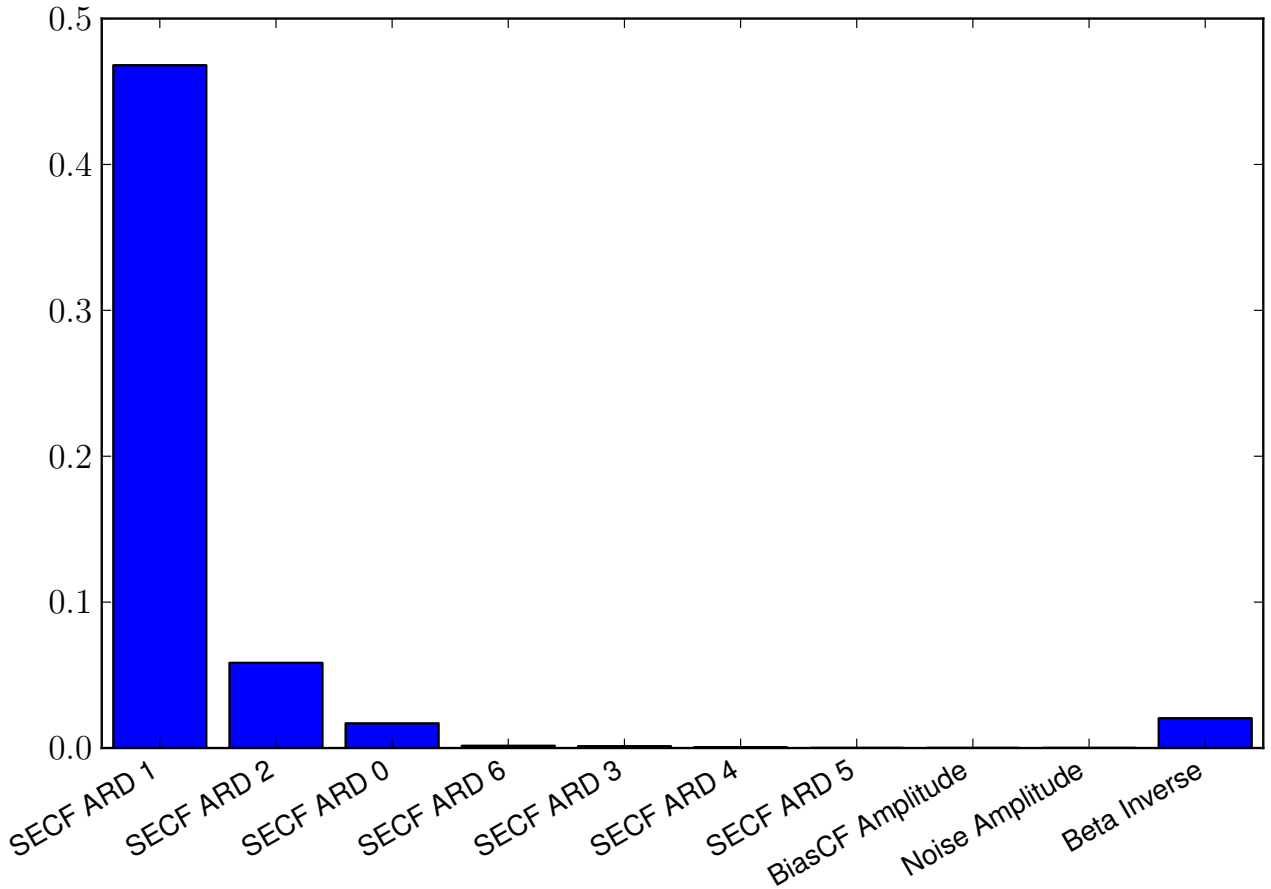






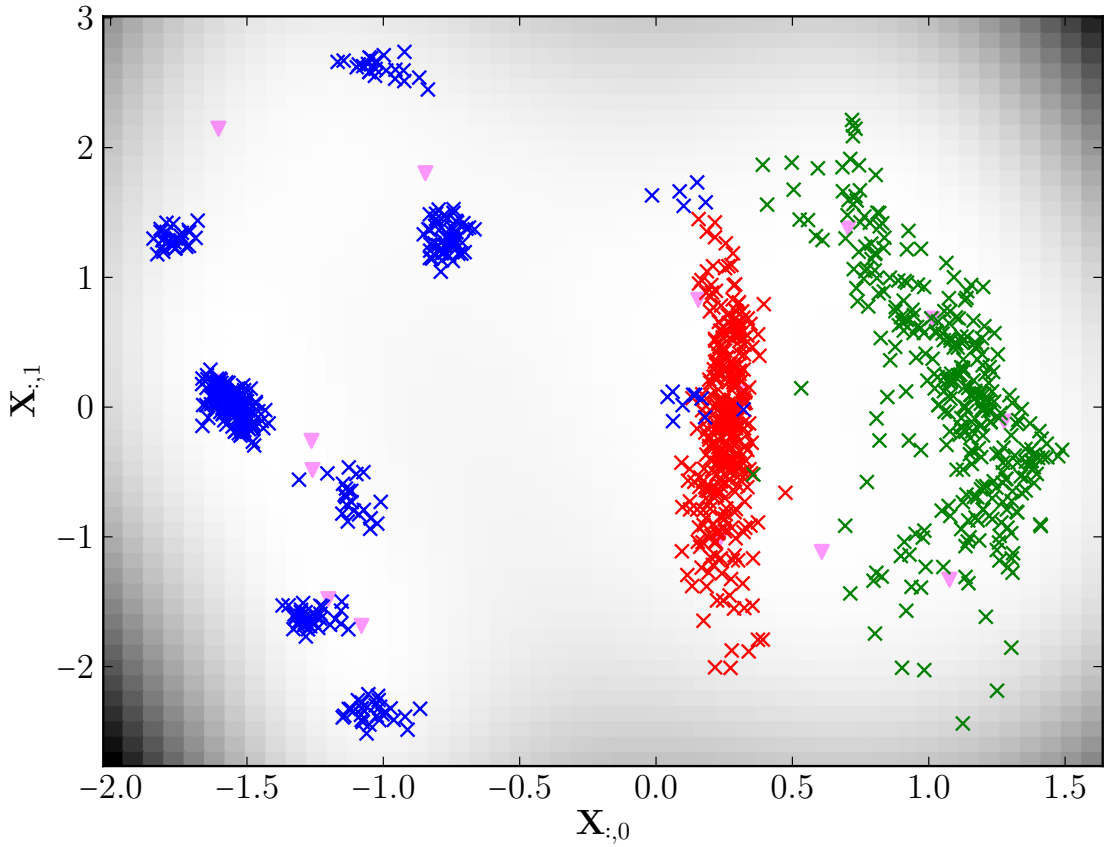
BIOLOGISCHE KYBERNETIK

Intuition

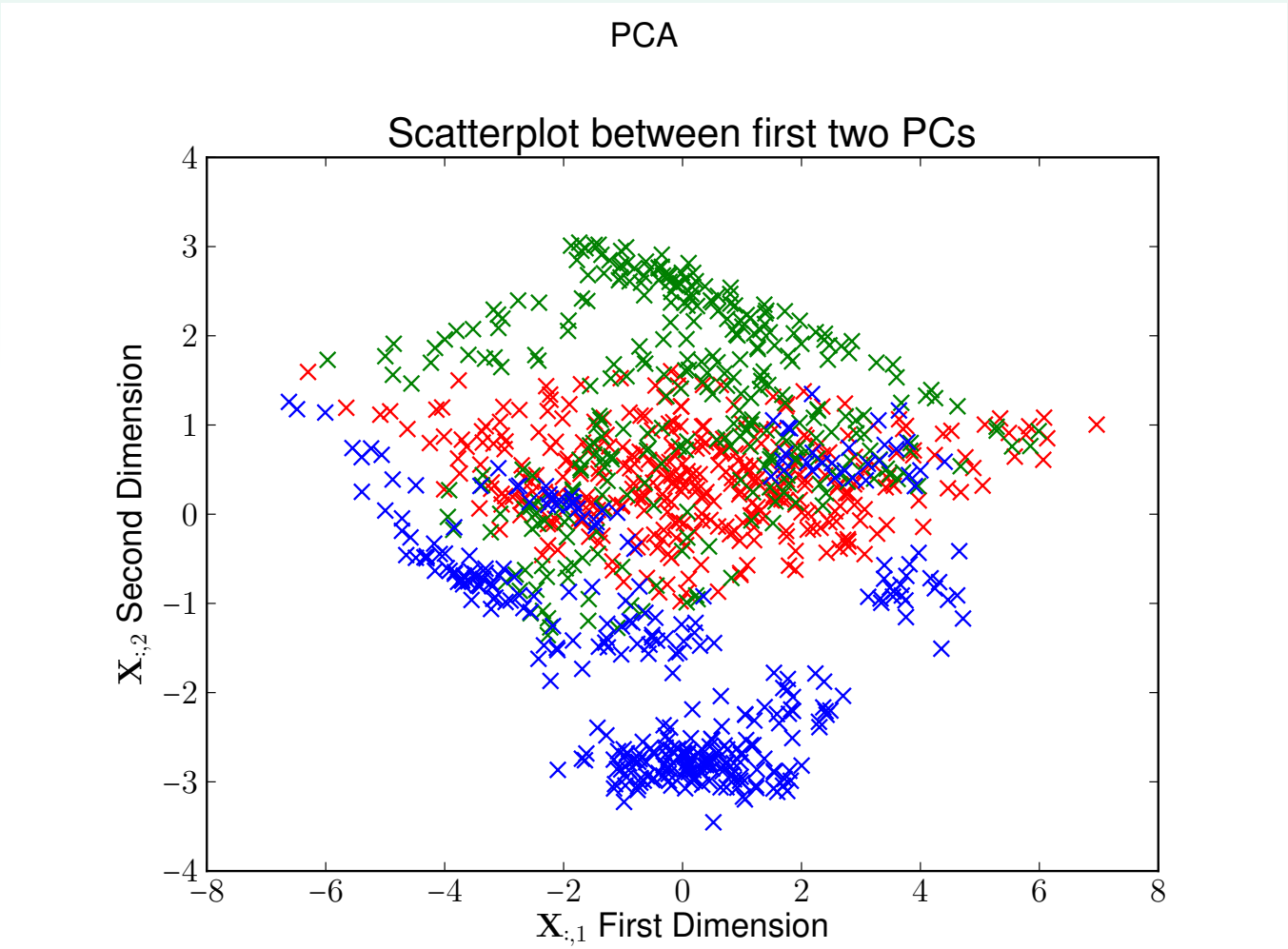


BIOLOGISCHE KYBERNETIK

Intuition

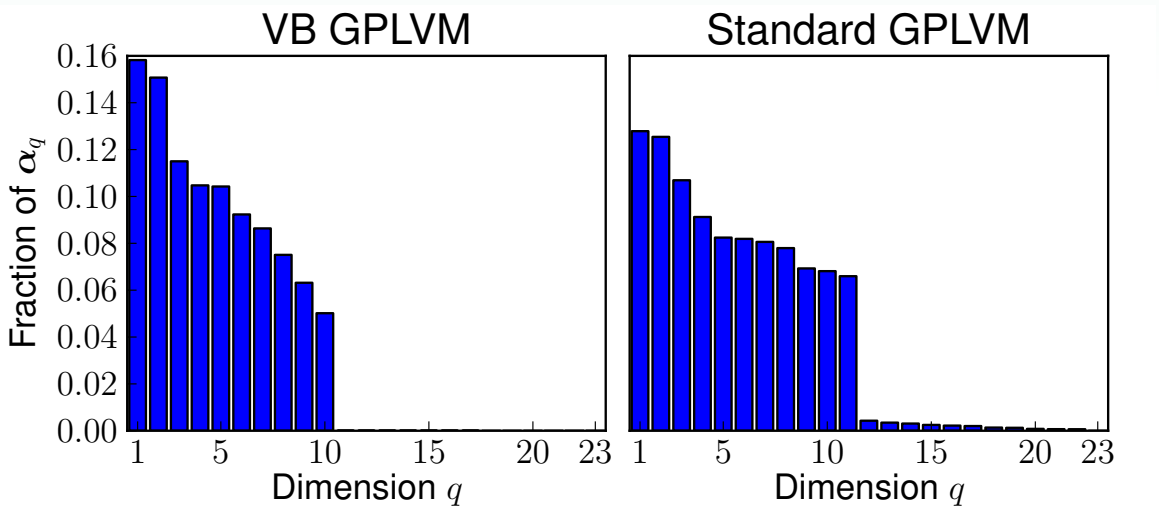


Intuition

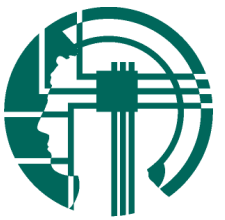

PCA
Scatterplot between first two PCs
$\mathbf{X}_{:,2}$ Second Dimension
$\mathbf{X}_{:,1}$ First Dimension
4
3
2
1
0
−1
−2
−3
−4
−8
−6
−4
−2
0
2
4
6
8

Learning the right Dimensionality

Simulated Dataset (from Linear Mixed Model) with $Q = 10$ Confounders:



Outline



PANAMA

Linear Mixed Model Assumed:

$$\overbrace{\mathbf{Y}}^{N\times D} = \overbrace{\mathbf{S}}^{N\times K}\mathbf{W} + \overbrace{\mathbf{X}}^{N\times Q}\mathbf{V} \tag{17}$$

N : Samples D : Genes K : SNPs Q : Confounders

Iterative Approach through GPLVM:

1. Learn Confounders by GPLVM
2. Add Confounders correlated with Genotype
3. Fit the new Model
4. Continue with 1. until no Confounders are added

BIOLOGISCHE KYBERNETIK

Variational Bayesian GPLVM for Confounder Detection

Same Linear Mixed Model Assumed:

$$\mathbf{Y} = \mathbf{SW} + \mathbf{XV} \tag{18}$$

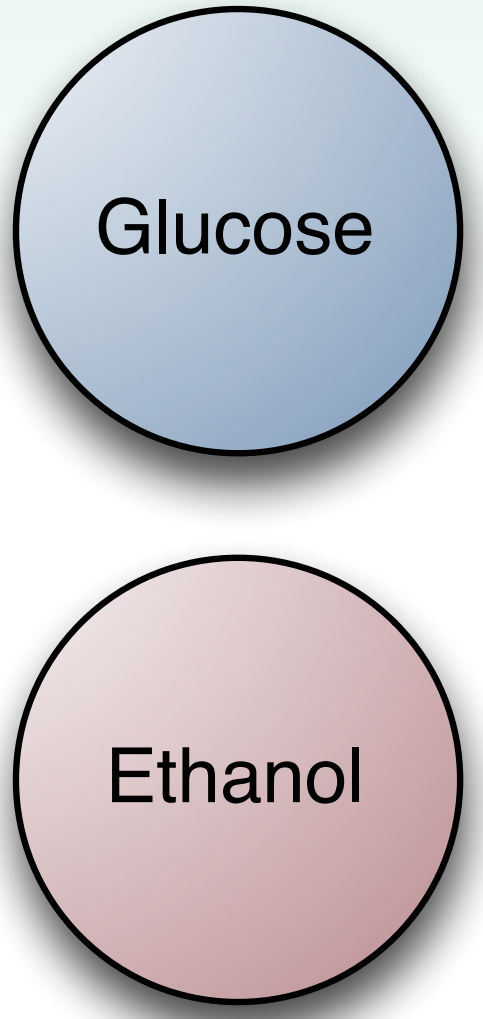
No Iterative Approach necessary
$\Rightarrow$ ARD + Bayesian Approximation

- Learn Confounders $\mathbf{X}$ through Variational Bayesian GPLVM
- Fit Linear Mixed Model with learned Confounders.

Yeast Dataset [7]

Saccharomyces Cerevisiae:

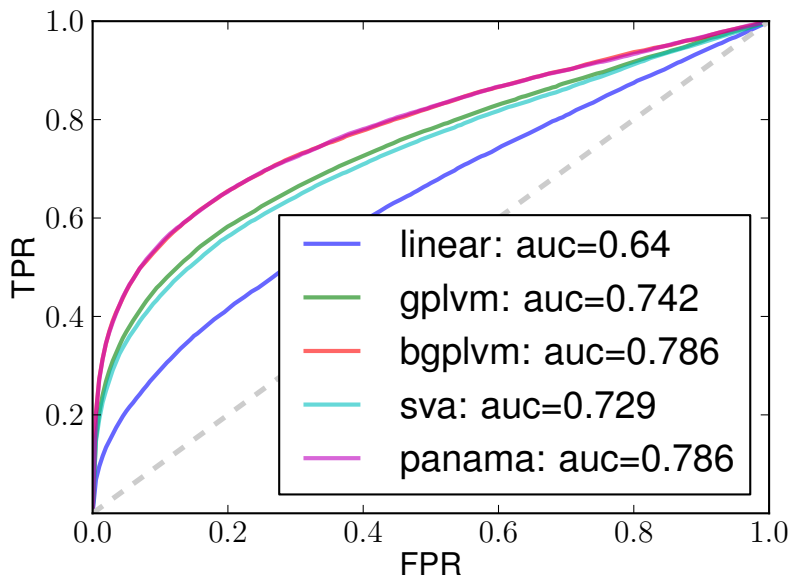
- Expression Profiled $N = 109$ Samples over $D = 5493$ Genes
- $K = 2956$ SNPs Genotyped
- 2 Environments: Glucose and Ethanol
- We only use Glucose Part.



Simulated Dataset by [2]

- Simulated Dataset founded on real Yeast [7] eQTL Experiment

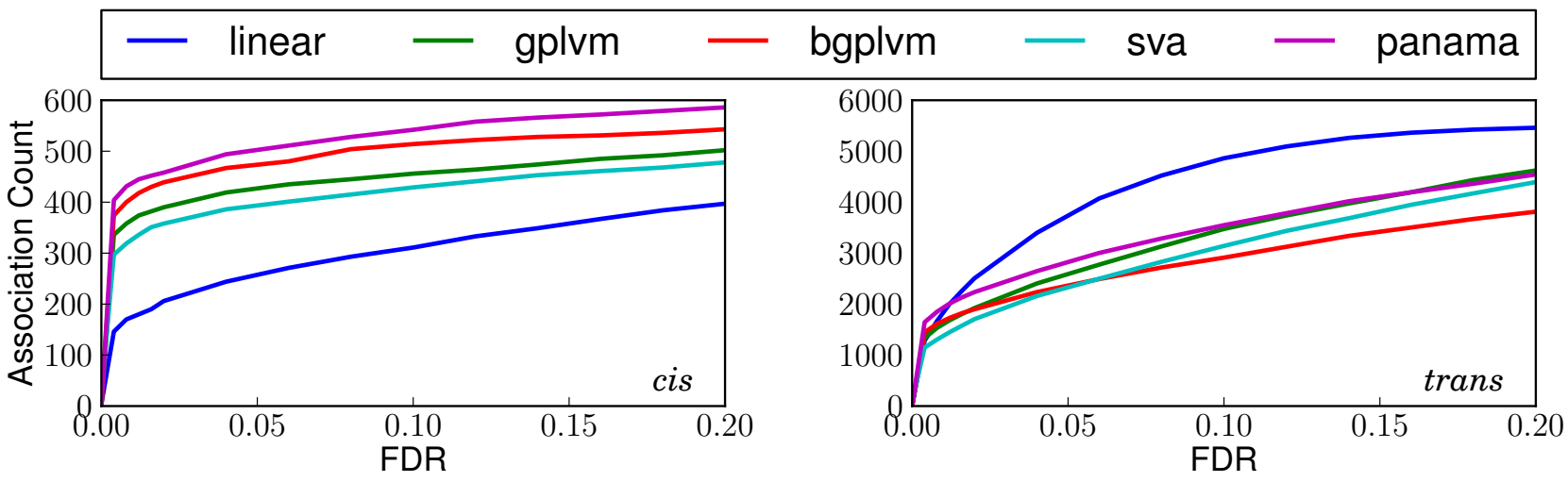
⇒ Ground Truth for Confounding Variation and Associations:



Simulated Dataset by [2]

- *cis* and *trans* Associations:

⇒ Less False Positives:



Manifold Relevance Determination for eQTL Studies

Combine Variational Bayesian GPLVMs to determine Private/Shared Information:

$$\ln p(\mathbf{Y}^{\kappa}) \geq \ln \sum_{\iota \in \kappa} \int q(\mathbf{X}) \ln p(\mathbf{Y}^{\iota}|\mathbf{X}) d\mathbf{X} \quad - \mathrm{KL}(q||p) \quad , \text{ where} \tag{19}$$

$$\mathbf{Y}^{\kappa} = \{\mathbf{Y}^{\iota} : \iota \in \kappa\} \tag{20}$$

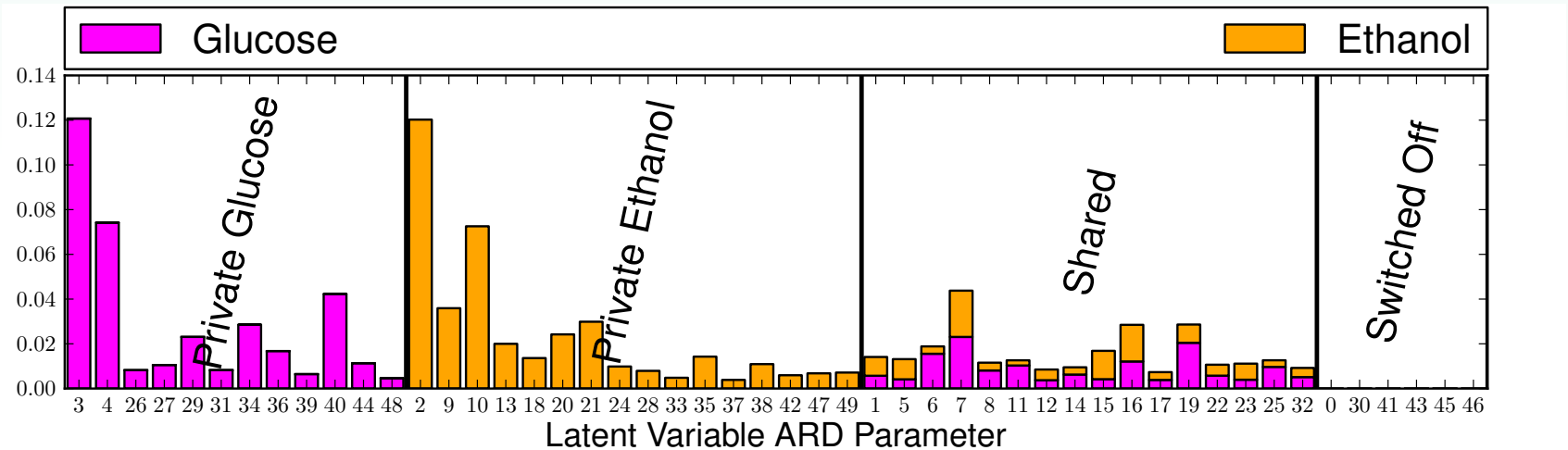
Thus,

- One Confounder Matrix $\mathbf{X}$
- ARD Parameters $\boldsymbol{\alpha}^{\iota}$ per Experiment $\mathbf{Y}^{\iota}$

BIOLOGISCHE KYBERNETIK

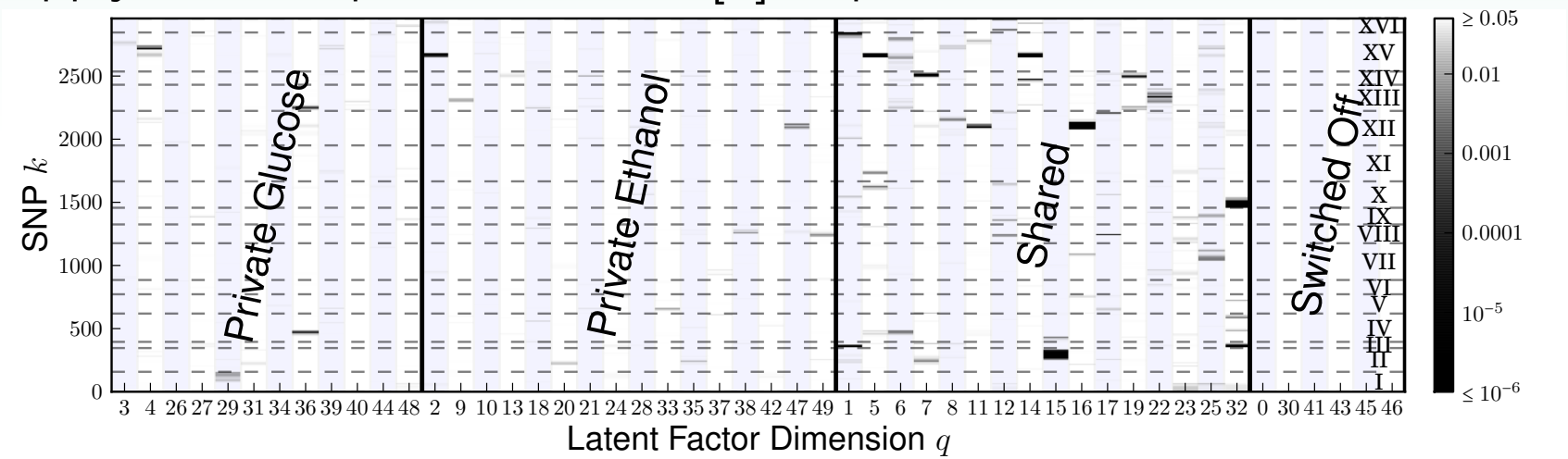
Manifold Relevance Determination for eQTL Studies

Apply to both parts of Smith [7] Experiment:



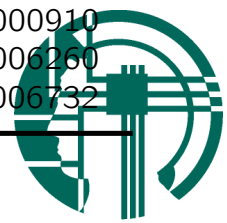
Manifold Relevance Determination for eQTL Studies

Apply to both parts of Smith [7] Experiment:

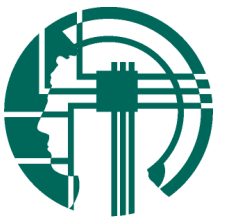


Manifold Relevance Determination for eQTL Studies

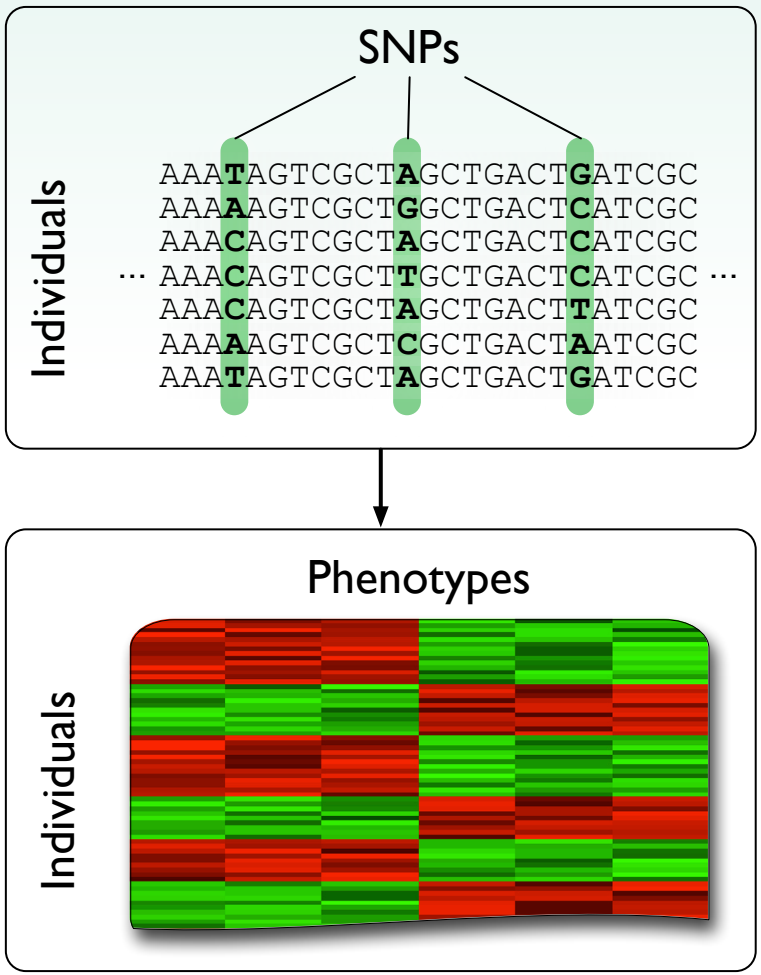
Private Glucose	#Hits	GO
cellular macromolecule catabolic process	7	0044265
macromolecule catabolic process	7	0009057
ncRNA processing	6	0034470
ncRNA metabolic process	6	0034660
tRNA metabolic process	4	0006399
Private Ethanol	**#Hits**	**GO**
response to abiotic stimulus	6	0009628
regulation of cell cycle	5	0051726
pseudohyphal growth	4	0007124
cell growth	4	0016049
growth of unicellular organism as a thread of attached cells	4	0070783
regulation of cell cycle process	4	0010564
Shared	**#Hits**	**GO**
cell cycle	42	0007049
cell cycle process	34	0022402
tRNA metabolic process	14	0006399
cytokinesis	13	0000910
DNA replication	13	0006260
coenzyme metabolic process	13	0006732



Outline

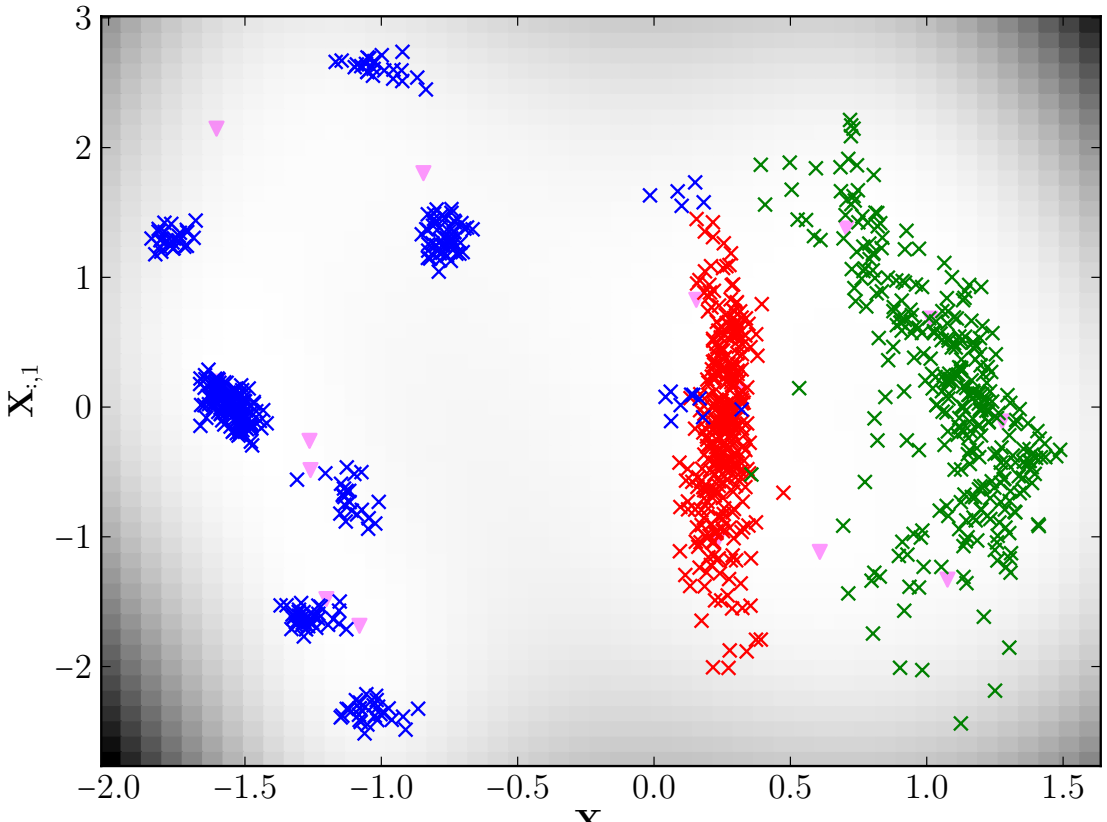


Summary

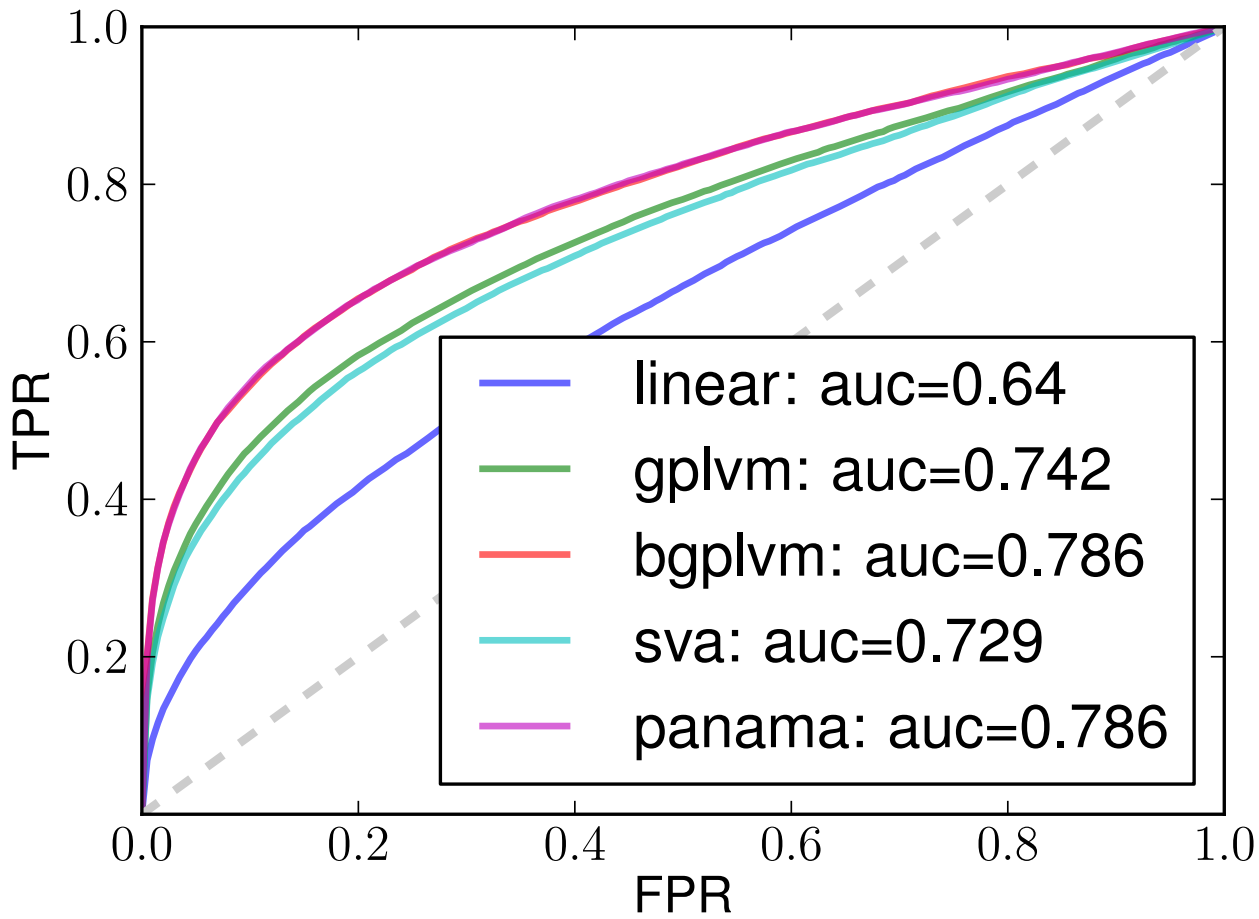

SNPs
Individuals
AAATAGTCGCTAGCTGACTGATCGC
AAAAAGTCGCTGGCTGACTCATCGC
AAACAGTCGCTAGCTGACTCATCGC
AAACAGTCGCTTGCTGACTCATCGC
AAACAGTCGCTAGCTGACTTATCGC
AAAAAGTCGCTCGCTGACTAATCGC
AAATAGTCGCTAGCTGACTGATCGC
Phenotypes
Individuals

Summary

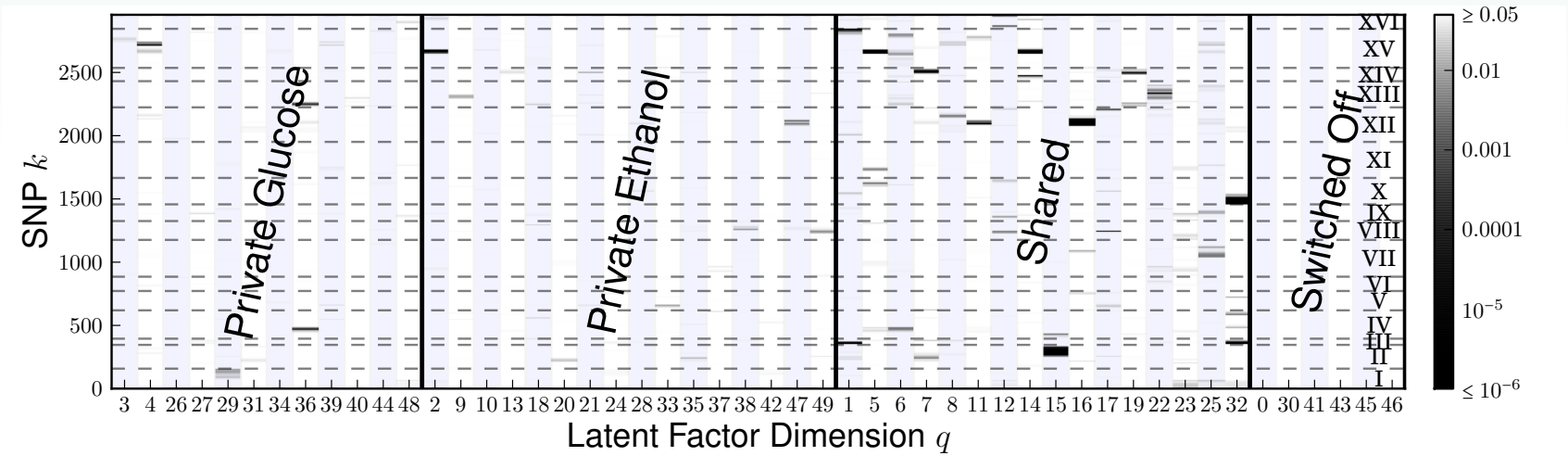
$$p(\mathbf{Y}) = \int p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})\, d\mathbf{X} \tag{21}$$



Summary


1.0
0.8
0.6
0.4
0.2
TPR
0.0
0.2
0.4
0.6
0.8
1.0
FPR
linear: auc=0.64
gplvm: auc=0.742
bgplvm: auc=0.786
sva: auc=0.729
panama: auc=0.786

Summary


SNP k
Latent Factor Dimension q
Private Glucose
Private Ethanol
Shared
Switched Off
≥ 0.05
0.01
0.001
0.0001
10^{-5}
≤ 10^{-6}

Discussion

References I


A.C. Damianou, M.K. Titsias, and N.D. Lawrence.
Variational Gaussian Process Dynamical Systems.
Neural Information Processing System (NIPS), 2011.


Nicolo Fusi, Oliver Stegle, and Neil D. Lawrence.
Accurate modeling of confounding variation in eQTL studies leads to a great increase in power to detect trans-regulatory effects.
Nature Precedings, 2011.


Joel N Hirschhorn and Mark J Daly.
Genome-wide association studies for common diseases and complex traits.
Nat Rev Genet, 6(2):95–108, Feb 2005.


N.D. Lawrence.
Gaussian process latent variable models for visualisation of high dimensional data.
Advances in neural information processing systems, 16:329–336, 2004.

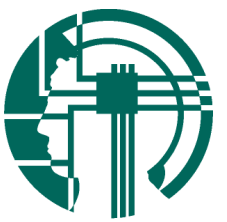

Neil Lawrence.
Probabilistic non-linear principal component analysis with Gaussian process latent variable models.
Journal of Machine Learning Research, November 2005.

BIOLOGISCHE KYBERNETIK

References II


Carl Edward Rasmussen.
Gaussian Processes for Machine Learning.
The MIT Press, 2006.


Erin N Smith and Leonid Kruglyak.
Gene-environment interaction in yeast gene expression.
PLoS Biol, 6(4):e83, Apr 2008.



From DNA to Protein