Machine Learning and Statistics in Genetics and Genomics

VII: Continuous latent variable models

Christoph Lippert

Microsoft Research
eScience group

Los Angeles , USA

Current topics in computational biology
UCLA
Winter quarter 2014

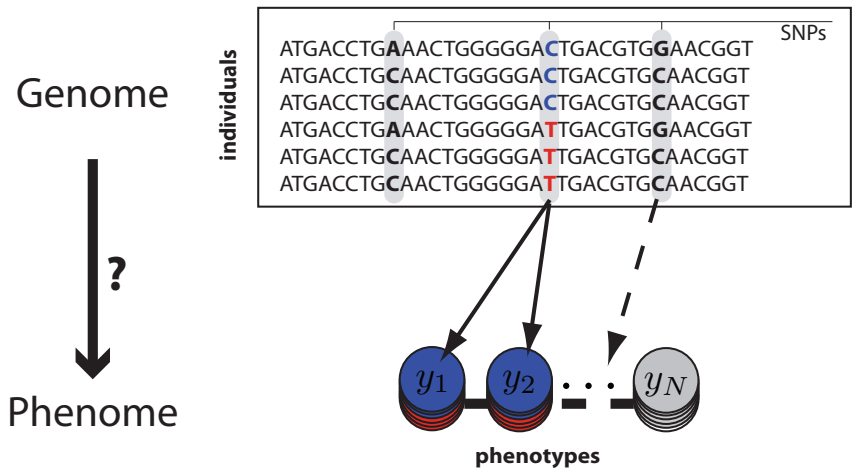
Outline

Outline

Why latent variables ?

Causal influences on phenotypes

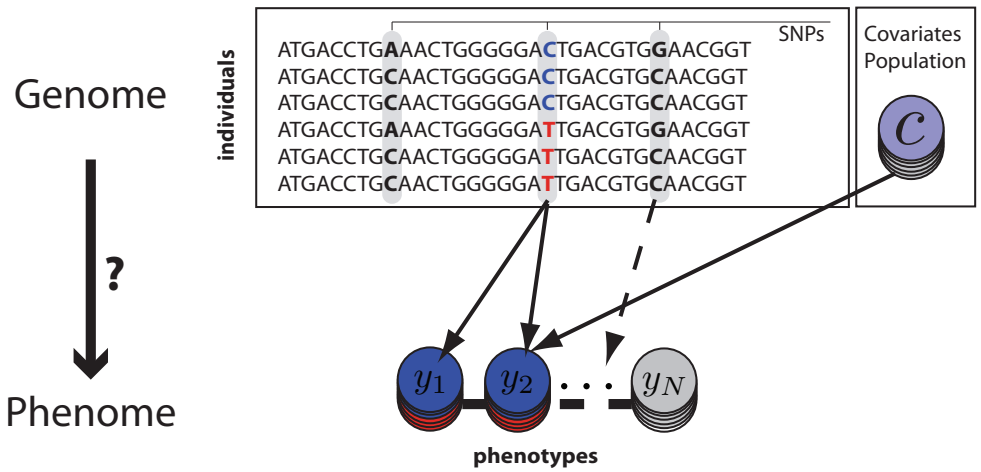
- Genotype
  - Primary variable of interest
- Known confounding factors
  - Covariates
  - Population structure
- Unknown (latent) confounders
  - Sample handling
  - Sample history
  - Subtle environmental perturbations



Why latent variables ?

Causal influences on phenotypes

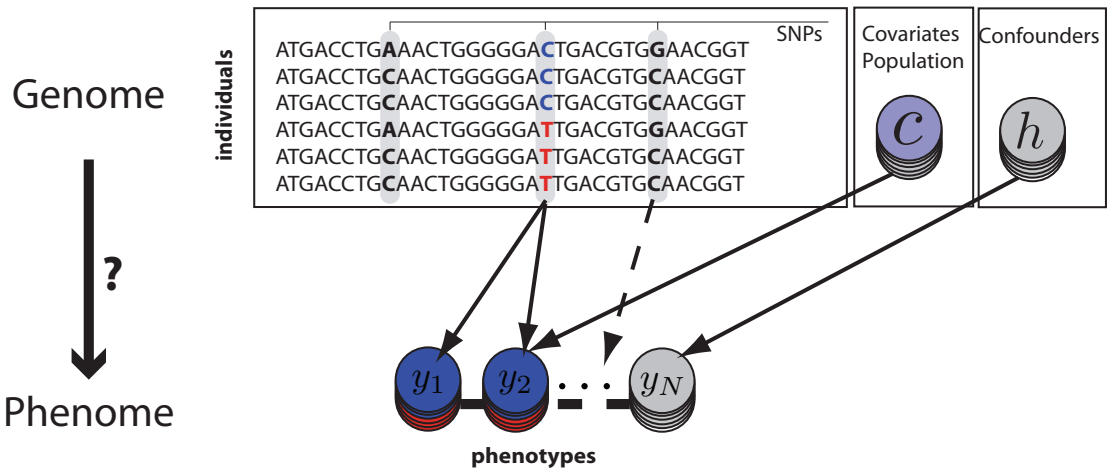
- Genotype
  - Primary variable of interest
- Known confounding factors
  - Covariates
  - Population structure
- Unknown (latent) confounders
  - Sample handling
  - Sample history
  - Subtle environmental perturbations



Why latent variables ?

Causal influences on phenotypes

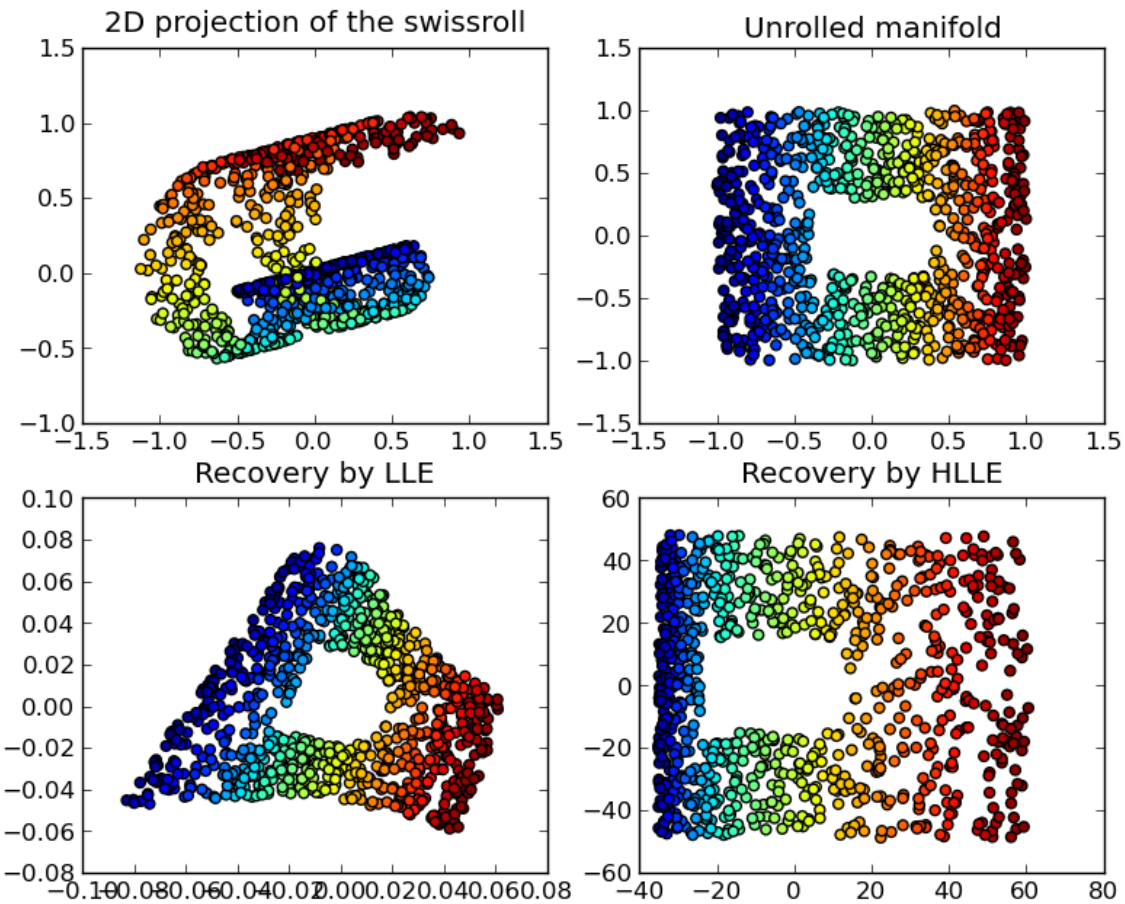
- Genotype
  - Primary variable of interest
- Known confounding factors
  - Covariates
  - Population structure
- Unknown (latent) confounders
  - Sample handling
  - Sample history
  - Subtle environmental perturbations



Outline

Outline

Manifolds and dimension reduction



(from Olivier Grisel, Generated using the Modular Data Processing toolkit and matplotlib.)

Linear dimension reduction

- Map G dimensional data on K dimensional manifold; $K << G$

$$\underbrace{\boldsymbol{Y}}_{NxG} = \underbrace{\boldsymbol{H}}_{NxK}\underbrace{\boldsymbol{W}}_{KxG} + \underbrace{\boldsymbol{\Psi}}_{NxG}$$

  - $\boldsymbol{H}$: latent factors in low-dimensional space
  - $\boldsymbol{W}$: weights for factors on data dimensions
  - $\boldsymbol{\Psi}$: noise, $\psi_{n,g} \sim \mathcal{N}(0, \sigma^2)$.
  - Challenge: neither $\boldsymbol{W}$ nor $\boldsymbol{H}$ known!
- Depending on assumptions on $\boldsymbol{W}$ and $\boldsymbol{H}$:
  - Principle component analysis (PCA)
  - Independent component analysis (ICA)
  - ...

Linear dimension reduction

- Map G dimensional data on K dimensional manifold; $K << G$

$$\underbrace{\boldsymbol{Y}}_{NxG} = \underbrace{\boldsymbol{H}}_{NxK}\underbrace{\boldsymbol{W}}_{KxG} + \underbrace{\boldsymbol{\Psi}}_{NxG}$$

  - $\boldsymbol{H}$: latent factors in low-dimensional space
  - $\boldsymbol{W}$: weights for factors on data dimensions
  - $\boldsymbol{\Psi}$: noise, $\psi_{n,g} \sim \mathcal{N}(0, \sigma^2)$.
  - Challenge: neither $\boldsymbol{W}$ nor $\boldsymbol{H}$ known!
- Depending on assumptions on W and H:
  - Principle component analysis (PCA)
  - Independent component analysis (ICA)
  - ...

Linear dimension reduction

- Map G dimensional data on K dimensional manifold; $K << G$

$$\underbrace{\boldsymbol{Y}}_{NxG} = \underbrace{\boldsymbol{H}}_{NxK}\underbrace{\boldsymbol{W}}_{KxG} + \underbrace{\boldsymbol{\Psi}}_{NxG}$$

  - $\boldsymbol{H}$: latent factors in low-dimensional space
  - $\boldsymbol{W}$: weights for factors on data dimensions
  - $\boldsymbol{\Psi}$: noise, $\psi_{n,g} \sim \mathcal{N}(0, \sigma^2)$.
  - Challenge: neither $\boldsymbol{W}$ nor $\boldsymbol{H}$ known!
- Depending on assumptions on $\boldsymbol{W}$ and $\boldsymbol{H}$:
  - Principle component analysis (PCA)
  - Independent component analysis (ICA)
  - ...

Linear dimension reduction

PCA

PCA is corresponds to a noise-free version of the model

$$\underbrace{\boldsymbol{Y}}_{NxG} = \underbrace{\boldsymbol{H}}_{NxK}\underbrace{\boldsymbol{W}}_{KxG}$$

- PCA components ($\boldsymbol{H}$) correspond to directions of maximum data variance in the original dataset:
  - Covariance matrix: $\boldsymbol{C} = \boldsymbol{Y}\boldsymbol{Y}^\top$
  - Eigenvalue/Eigen vectors $\boldsymbol{C}\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$
  - Projection matrix $\boldsymbol{P} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K]$
  - Principle components $\boldsymbol{H}_n = \boldsymbol{P} \cdot \boldsymbol{Y}_n$.

Linear dimension reduction

PCA

PCA is corresponds to a noise-free version of the model

$$\underbrace{\boldsymbol{Y}}_{NxG} = \underbrace{\boldsymbol{H}}_{NxK}\underbrace{\boldsymbol{W}}_{KxG}$$

- PCA components ($\boldsymbol{H}$) correspond to directions of maximum data variance in the original dataset:
  - Covariance matrix: $\boldsymbol{C} = \boldsymbol{Y}\boldsymbol{Y}^\top$
  - Eigenvalue/Eigen vectors $\boldsymbol{C}\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$
  - Projection matrix $\boldsymbol{P} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K]$
  - Principle components $\boldsymbol{H}_n = \boldsymbol{P} \cdot \boldsymbol{Y}_n$.

Linear dimension reduction

PCA

PCA is corresponds to a noise-free version of the model

$$\underbrace{\boldsymbol{Y}}_{NxG} = \underbrace{\boldsymbol{H}}_{NxK}\underbrace{\boldsymbol{W}}_{KxG}$$

- PCA components ($\boldsymbol{H}$) correspond to directions of maximum data variance in the original dataset:
  - Covariance matrix: $\boldsymbol{C} = \boldsymbol{Y}\boldsymbol{Y}^\top$
  - Eigenvalue/Eigen vectors $\boldsymbol{C}\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$
  - Projection matrix $\boldsymbol{P} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K]$
  - Principle components $\boldsymbol{H}_n = \boldsymbol{P} \cdot \boldsymbol{Y}_n$.

Linear dimension reduction

PCA

PCA is corresponds to a noise-free version of the model

$$\underbrace{\boldsymbol{Y}}_{NxG} = \underbrace{\boldsymbol{H}}_{NxK}\underbrace{\boldsymbol{W}}_{KxG}$$

- PCA components ($\boldsymbol{H}$) correspond to directions of maximum data variance in the original dataset:
  - Covariance matrix: $\boldsymbol{C} = \boldsymbol{Y}\boldsymbol{Y}^\top$
  - Eigenvalue/Eigen vectors $\boldsymbol{C}\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$
  - Projection matrix $\boldsymbol{P} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K]$
  - Principle components $\boldsymbol{H}_n = \boldsymbol{P} \cdot \boldsymbol{Y}_n$.

Linear dimension reduction

PCA

PCA is corresponds to a noise-free version of the model

$$\underbrace{\boldsymbol{Y}}_{NxG} = \underbrace{\boldsymbol{H}}_{NxK}\underbrace{\boldsymbol{W}}_{KxG}$$

- PCA components ($\boldsymbol{H}$) correspond to directions of maximum data variance in the original dataset:
  - Covariance matrix: $\boldsymbol{C} = \boldsymbol{Y}\boldsymbol{Y}^\top$
  - Eigenvalue/Eigen vectors $\boldsymbol{C}\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$
  - Projection matrix $\boldsymbol{P} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K]$
  - Principle components $\boldsymbol{H}_n = \boldsymbol{P} \cdot \boldsymbol{Y}_n$.

Linear dimension reduction

Bayesian PCA and GPLVM

Assumption: data dimensions or sample dimension independent **given** $\boldsymbol{H}$ and $\boldsymbol{W}$.

Probabilistic PCA

$$p(\boldsymbol{Y}|\boldsymbol{H},\boldsymbol{W}) = \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{y}_n \mid \boldsymbol{h}_n\boldsymbol{W}, \sigma^2\boldsymbol{I}\right)$$

$$p(\boldsymbol{H}) = \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{h}_n \mid 0, \sigma_h^2\boldsymbol{I}\right)$$

$$p(\boldsymbol{Y}|\boldsymbol{W}) = \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{y}_n \mid \boldsymbol{0}, \sigma_h^2\boldsymbol{W}\boldsymbol{W}^\top + \sigma^2\boldsymbol{I}\right)$$

GPLVM

$$p(\boldsymbol{Y}|\boldsymbol{H},\boldsymbol{W}) = \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{y}_{:,g} \mid \boldsymbol{H}\boldsymbol{w}_g, \sigma^2\boldsymbol{I}\right)$$

$$p(\boldsymbol{W}) = \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{w}_{:,g} \mid 0, \sigma_h^2\boldsymbol{I}\right)$$

$$p(\boldsymbol{Y}|\boldsymbol{H}) = \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{y}_{:,g} \mid \boldsymbol{0}, \underbrace{\sigma_h^2\boldsymbol{H}\boldsymbol{H}^\top}_{N \times N} + \sigma^2\boldsymbol{I}\right)$$

Linear dimension reduction

Bayesian PCA and GPLVM

Assumption: data dimensions or sample dimension independent **given** $\boldsymbol{H}$ and $\boldsymbol{W}$.

Probabilistic PCA

$$p(\boldsymbol{Y}|\boldsymbol{H},\boldsymbol{W}) = \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{y}_n \mid \boldsymbol{h}_n\boldsymbol{W}, \sigma^2\boldsymbol{I}\right)$$

$$p(\boldsymbol{H}) = \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{h}_n \mid 0, \sigma_h^2\boldsymbol{I}\right)$$

$$p(\boldsymbol{Y}|\boldsymbol{W}) = \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{y}_n \mid \boldsymbol{0}, \sigma_h^2\boldsymbol{W}\boldsymbol{W}^\top + \sigma^2\boldsymbol{I}\right)$$

GPLVM

$$p(\boldsymbol{Y}|\boldsymbol{H},\boldsymbol{W}) = \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{y}_{:,g} \mid \boldsymbol{H}\boldsymbol{w}_g, \sigma^2\boldsymbol{I}\right)$$

$$p(\boldsymbol{W}) = \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{w}_{:,g} \mid 0, \sigma_h^2\boldsymbol{I}\right)$$

$$p(\boldsymbol{Y}|\boldsymbol{H}) = \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{y}_{:,g} \mid \boldsymbol{0}, \underbrace{\sigma_h^2\boldsymbol{H}\boldsymbol{H}^\top}_{NxN} + \sigma^2\boldsymbol{I}\right)$$

GPLVM

- Marginal likelihood

$$p(\boldsymbol{Y}|\boldsymbol{H}) = \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{y}_{:,g} \mid \mathbf{0}, \sigma_h^2 \boldsymbol{H}\boldsymbol{H}^\top + \sigma^2 \boldsymbol{I}\right)$$

- Inference of most probable "hidden factors" $\boldsymbol{H}$:

$$\hat{\boldsymbol{H}} = \underset{\boldsymbol{H}}{\operatorname{argmax}} \log \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{y}_{:,g} \mid \mathbf{0}, \sigma_h^2 \boldsymbol{H}\boldsymbol{H}^\top + \sigma^2 \boldsymbol{I}\right)$$

GPLVM

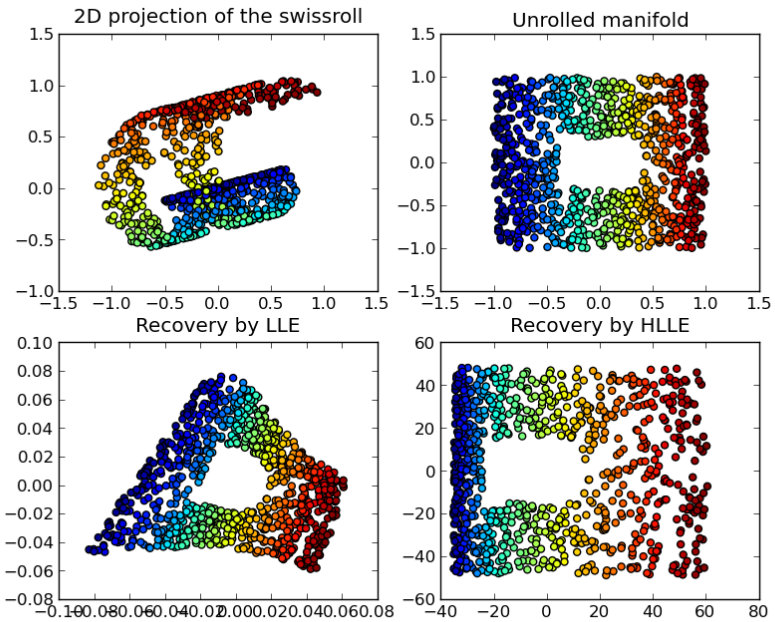
- Marginal likelihood

$$p(\boldsymbol{Y}|\boldsymbol{H}) = \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{y}_{:,g} \mid \mathbf{0}, \sigma_h^2 \boldsymbol{H}\boldsymbol{H}^\top + \sigma^2 \boldsymbol{I}\right)$$

- Inference of most probable "hidden factors" $\boldsymbol{H}$:

$$\hat{\boldsymbol{H}} = \underset{\boldsymbol{H}}{\operatorname{argmax}} \log \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{y}_{:,g} \mid \mathbf{0}, \sigma_h^2 \boldsymbol{H}\boldsymbol{H}^\top + \sigma^2 \boldsymbol{I}\right)$$

Are linear relationships sufficient?

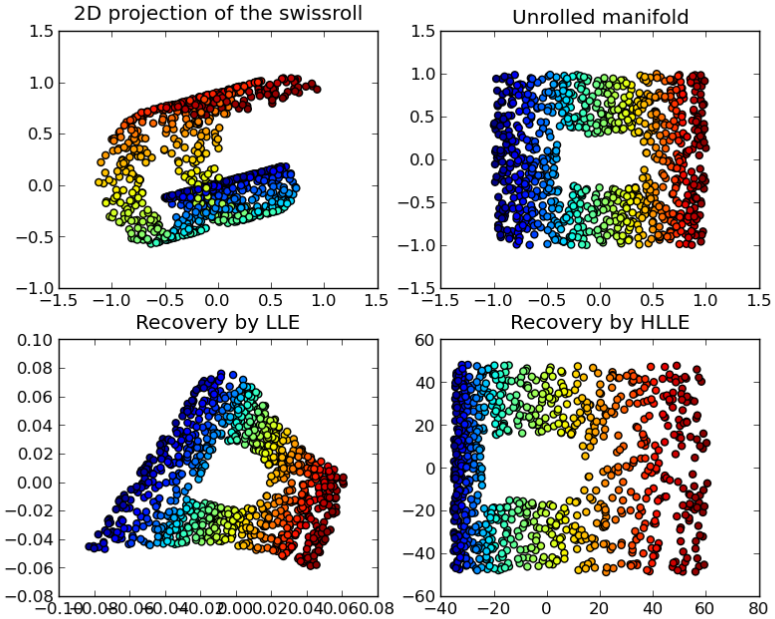


(from Olivier Grisel, Generated using the Modular Data Processing toolkit and matplotlib.)

- Non-linear generalizations, introducing general kernel function instead of linear covariance

$$\hat{\boldsymbol{H}} = \underset{\boldsymbol{H}}{\operatorname{argmax}} \log \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{y}_{:,g} \mid \boldsymbol{0}, \sigma_h^2 \boldsymbol{H}\boldsymbol{H}^\top + \sigma^2 \boldsymbol{I}\right)$$

Are linear relationships sufficient?

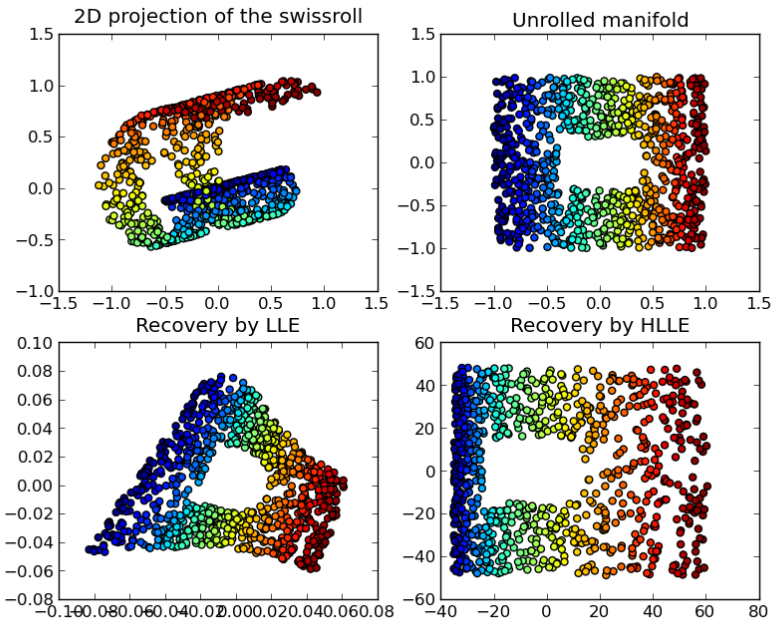


(from Olivier Grisel, Generated using the Modular Data Processing toolkit and matplotlib.)

- Non-linear generalizations, introducing general kernel function instead of linear covariance

$$\hat{\boldsymbol{H}} = \underset{\boldsymbol{H}}{\operatorname{argmax}} \log \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{y}_{:,g} \mid \mathbf{0}, \sigma_h^2 \boldsymbol{H}\boldsymbol{H}^\top + \sigma^2 \boldsymbol{I}\right)$$

Are linear relationships sufficient?



(from Olivier Grisel, Generated using the Modular Data Processing toolkit and matplotlib.)

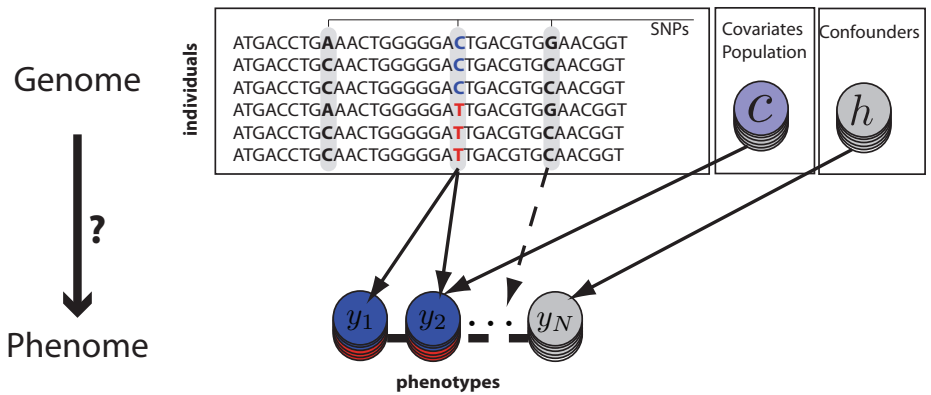
- Non-linear generalizations, introducing general kernel function instead of linear covariance

$$\hat{\boldsymbol{H}} = \underset{\boldsymbol{H}}{\operatorname{argmax}} \log \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{y}_{:,g} \mid \mathbf{0}, \sigma_h^2 \boldsymbol{K}_{\boldsymbol{H},\boldsymbol{H}} + \sigma^2 \boldsymbol{I}\right)$$

Outline

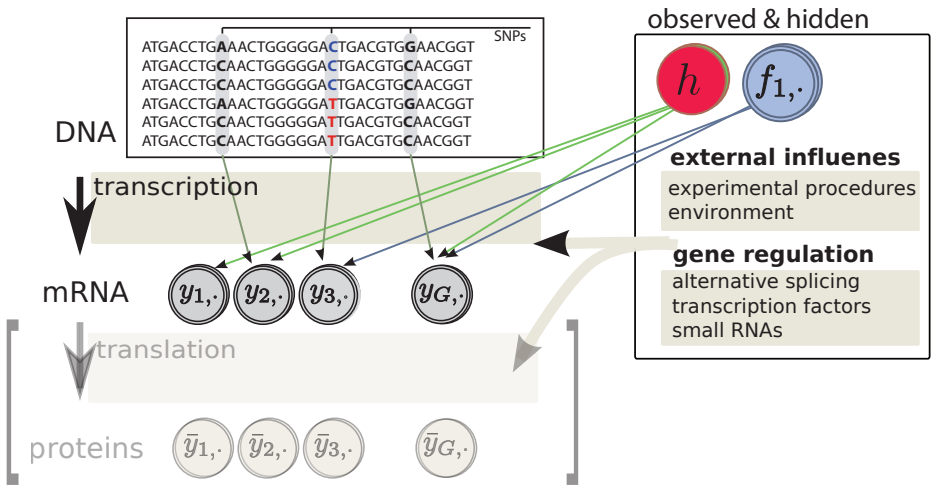
Confounders in eQTL studies

- Confounders in eQTL studies
  - **Experimental procedures**
  - Gene regulation
  - (Translation)
- Standard to take *known factors* (gender, population structure) into account.
- It is key to account for hidden factors as well.
  - Sample preparation
  - Sample history



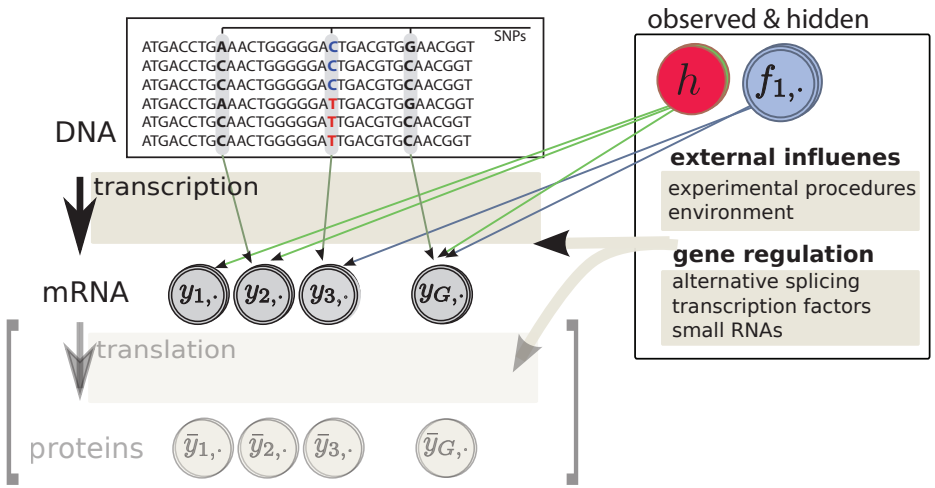
Confounders in eQTL studies

- Confounders in eQTL studies
  - **Experimental procedures**
  - Gene regulation
  - (Translation)
- Standard to take *known factors* (gender, population structure) into account.
- It is key to account for hidden factors as well.
  - Sample preparation
  - Sample history



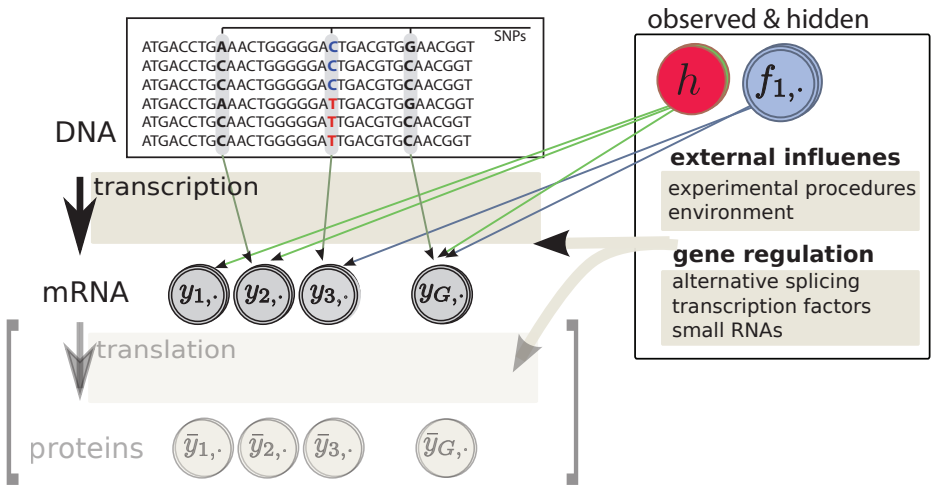
Confounders in eQTL studies

- Confounders in eQTL studies
  - **Experimental procedures**
  - Gene regulation
  - (Translation)
- Standard to take *known factors* (gender, population structure) into account.
- It is key to account for hidden factors as well.
  - Sample preparation
  - Sample history



Confounders in eQTL studies

- Confounders in eQTL studies
  - **Experimental procedures**
  - Gene regulation
  - (Translation)
- Standard to take *known factors* (gender, population structure) into account.
- It is key to account for hidden factors as well.
  - Sample preparation
  - Sample history

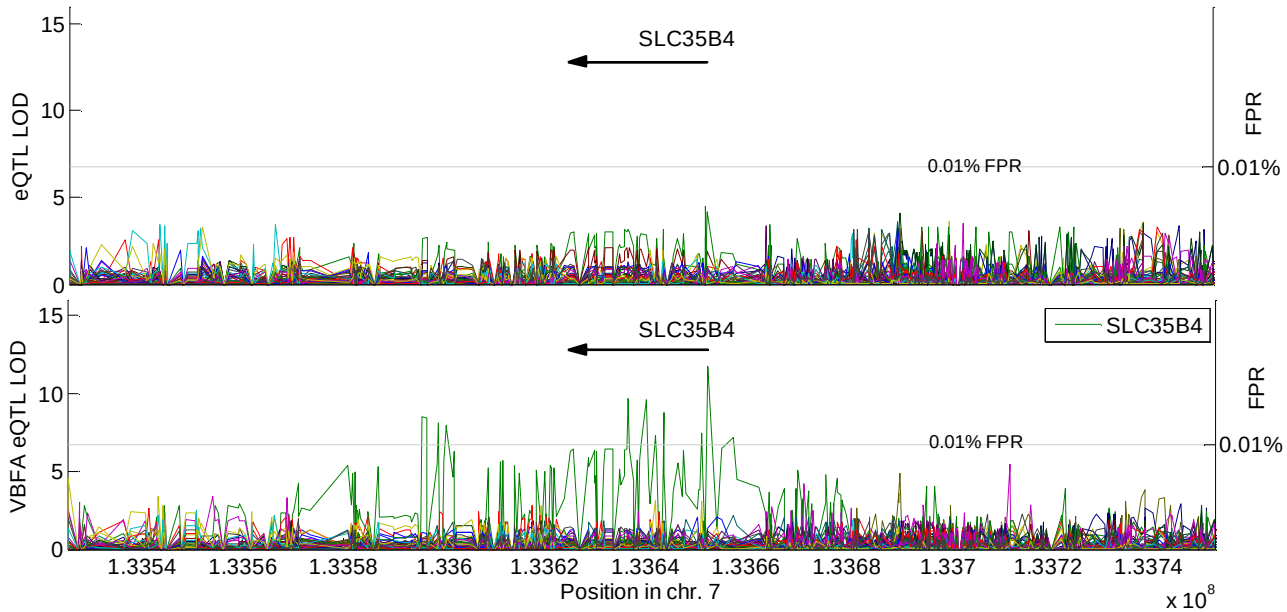


Motivating examples

HapMAP II, 3 populations, 90 individuals each, 40K genes, 3 million SNPs (here: small region in chromosome 7)

(a) Standard

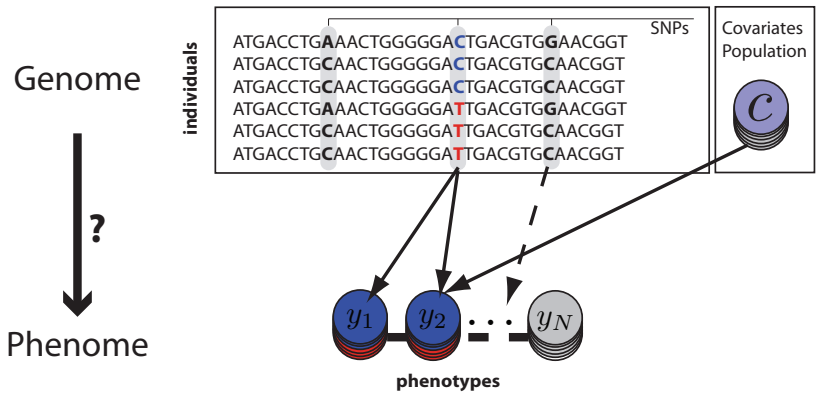
(b) VBFA accounting for hidden factors



Association model

Direct effects of confounders

- Start with standard association model.
- Include (K, few) global hidden factors (confounders) in the model.
- Factors $\boldsymbol{H} = \{\boldsymbol{h}_{:,1}, \ldots, \boldsymbol{h}_{:,K}\}$ need to be learned from the expression data.
- How to control model complexity, i.e. choose the number of factors?

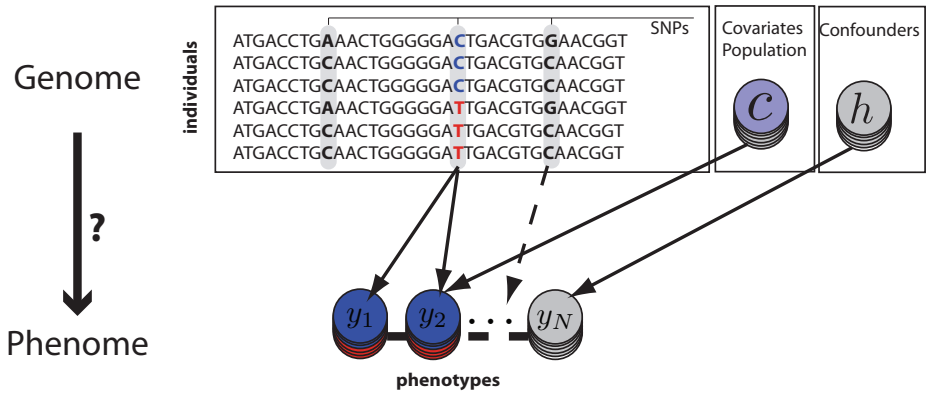


$$\boldsymbol{y}_{:,g} = \overbrace{\sum_{n=1}^{N} (\theta_{n,g} \boldsymbol{s}_n)}^{\text{genetic}} + \underbrace{\boldsymbol{\psi}}_{\text{noise}}$$

Association model

Direct effects of confounders

- Start with standard association model.
- Include (K, few) global hidden factors (confounders) in the model.
- Factors $\boldsymbol{H} = \{\boldsymbol{h}_{:,1}, \ldots, \boldsymbol{h}_{:,K}\}$ need to be learned from the expression data.
- How to control model complexity, i.e. choose the number of factors?

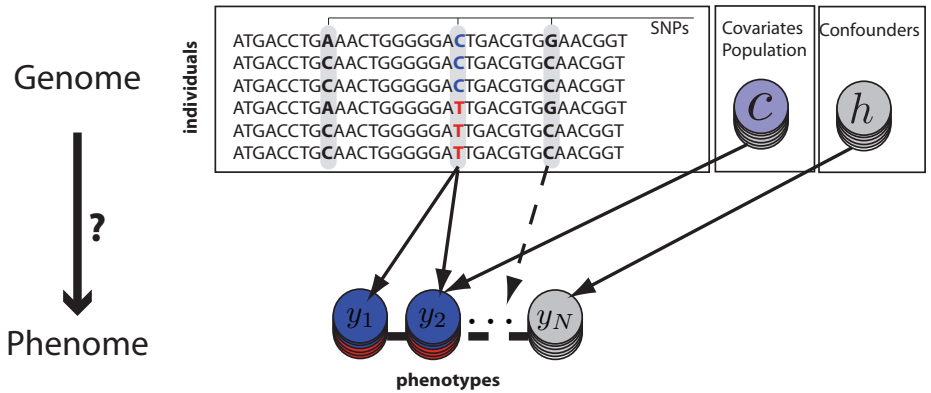


$$\boldsymbol{y}_{:,g} = \overbrace{\sum_{n=1}^{N} (\theta_{n,g} \boldsymbol{s}_n)}^{\text{genetic}} + \underbrace{\sum_{k=1}^{K} w_{g,k} \boldsymbol{h}_k}_{\text{hidden factors}} + \underbrace{\boldsymbol{\psi}}_{\text{noise}}$$

Association model

Direct effects of confounders

- Start with standard association model.
- Include (K, few) global hidden factors (confounders) in the model.
- Factors $\boldsymbol{H} = \{\boldsymbol{h}_{:,1}, \ldots, \boldsymbol{h}_{:,K}\}$ need to be learned from the expression data.
- How to control model complexity, i.e. choose the number of factors?



$$\boldsymbol{y}_{:,g} = \overbrace{\sum_{n=1}^{N} (\theta_{n,g}\boldsymbol{s}_n)}^{\text{genetic}} + \underbrace{\sum_{k=1}^{K} w_{g,k}\boldsymbol{h}_k}_{\text{hidden factors}} + \underbrace{\boldsymbol{\psi}}_{\text{noise}}$$

Association model

Gaussian process formulation

- Data likelihood of the linear generative model assuming Gaussian noise

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{H}, \boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{\Theta}_{\mathsf{K}}) = \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{y}_{:,g} \mid \sum_{s=1}^{S} \boldsymbol{x}_s \theta_s \; + \sum_{k=1}^{K} w_{g,k} \boldsymbol{h}_k, \sigma^2 \boldsymbol{I} \right)$$

- Specify Gaussian priors on the factor weight

$$P(w_{g,k}) = \mathcal{N}\left(w_{g,k} \mid 0, \sigma_h^2 \right)$$

- Integrating out the weights yields a Gaussian process marginal likelihood, factorizing over genes given $\boldsymbol{H}$

$$P(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{H}, \boldsymbol{\Theta}_{\mathsf{K}}) = \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{y}_{:,g} \mid \sum_{s=1}^{S} \boldsymbol{x}_s \theta_s, \sigma_h^2 \sum_{k=1}^{K} \boldsymbol{h}_k \boldsymbol{h}_k^{\top} + \sigma^2 \boldsymbol{I} \right)$$

Association model

Gaussian process formulation

- Data likelihood of the linear generative model assuming Gaussian noise

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{H}, \boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{\Theta}_{\mathsf{K}}) = \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{y}_{:,g} \mid \sum_{s=1}^{S} \boldsymbol{x}_s \theta_s \; + \sum_{k=1}^{K} w_{g,k} \boldsymbol{h}_k, \sigma^2 \boldsymbol{I} \right)$$

- Specify Gaussian priors on the factor weight

$$P(w_{g,k}) = \mathcal{N}\left(w_{g,k} \mid 0, \sigma_h^2 \right)$$

- Integrating out the weights yields a Gaussian process marginal likelihood, factorizing over genes given $\boldsymbol{H}$

$$P(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{H}, \boldsymbol{\Theta}_{\mathsf{K}}) = \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{y}_{:,g} \mid \sum_{s=1}^{S} \boldsymbol{x}_s \theta_s, \sigma_h^2 \sum_{k=1}^{K} \boldsymbol{h}_k \boldsymbol{h}_k^{\top} + \sigma^2 \boldsymbol{I} \right)$$

Association model

Gaussian process formulation

- Data likelihood of the linear generative model assuming Gaussian noise

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{H}, \boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{\Theta}_{\mathsf{K}}) = \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{y}_{:,g} \mid \sum_{s=1}^{S} \boldsymbol{x}_s \theta_s \; + \sum_{k=1}^{K} w_{g,k} \boldsymbol{h}_k, \sigma^2 \boldsymbol{I} \right)$$

- Specify Gaussian priors on the factor weight

$$P(w_{g,k}) = \mathcal{N}\left(w_{g,k} \mid 0, \sigma_h^2 \right)$$

- Integrating out the weights yields a Gaussian process marginal likelihood, factorizing over genes given $\boldsymbol{H}$

$$P(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{H}, \boldsymbol{\Theta}_{\mathsf{K}}) = \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{y}_{:,g} \mid \sum_{s=1}^{S} \boldsymbol{x}_s \theta_s, \sigma_h^2 \sum_{k=1}^{K} \boldsymbol{h}_k \boldsymbol{h}_k^{\top} + \sigma^2 \boldsymbol{I} \right)$$

Association model

Including population structure

$$P(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{H}, \boldsymbol{\Theta}_{\mathsf{K}}) = \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{y}_{:,g} \mid \sum_{s=1}^{S} \boldsymbol{x}_s \theta_s, \underbrace{\sigma_h^2 \sum_{k=1}^{K} \boldsymbol{h}_k \boldsymbol{h}_k^\top + \sigma^2 \boldsymbol{I}}_{NxN} \right)$$

- Including additional covariance term for population structure.
- Association test for single SNP.
- Challenge: Refitting σ_{g}^2, σ_h^2, $\boldsymbol{H}$ for every SNP.

Association model

Including population structure

$$P(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{H}, \boldsymbol{\Theta}_{\mathsf{K}}) = \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{y}_{:,g} \mid \sum_{s=1}^{S} \boldsymbol{x}_s \theta_s, \underbrace{\sigma_h^2 \sum_{k=1}^{K} \boldsymbol{h}_k \boldsymbol{h}_k^\top + \sigma_{\mathsf{g}}^2 \boldsymbol{X}\boldsymbol{X}^\top + \sigma^2 \boldsymbol{I}}_{NxN} \right)$$

- Including additional covariance term for population structure.
- Association test for single SNP.
- Challenge: Refitting σ_{g}^2, σ_h^2, $\boldsymbol{H}$ for every SNP.

Association model

Including population structure

$$P(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{H}, \boldsymbol{\Theta}_{\mathsf{K}}) = \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{y}_{:,g} \mid \boldsymbol{x}_i \theta_i, \underbrace{\sigma_h^2 \sum_{k=1}^{K} \boldsymbol{h}_k \boldsymbol{h}_k^\top + \sigma_{\mathsf{g}}^2 \boldsymbol{X}\boldsymbol{X}^\top + \sigma^2 \boldsymbol{I}}_{NxN} \right)$$

- Including additional covariance term for population structure.
- Association test for single SNP.
- Challenge: Refitting σ_{g}^2, σ_h^2, $\boldsymbol{H}$ for every SNP.

Association model

Including population structure

$$P(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{H}, \boldsymbol{\Theta}_{\mathsf{K}}) = \prod_{g=1}^{G} \mathcal{N}\left(\boldsymbol{y}_{:,g} \mid \boldsymbol{x}_i \theta_i, \underbrace{\sigma_h^2 \sum_{k=1}^{K} \boldsymbol{h}_k \boldsymbol{h}_k^\top + \sigma_{\mathsf{g}}^2 \boldsymbol{X}\boldsymbol{X}^\top + \sigma^2 \boldsymbol{I}}_{NxN} \right)$$

- Including additional covariance term for population structure.
- Association test for single SNP.
- Challenge: Refitting σ_{g}^2, σ_h^2, $\boldsymbol{H}$ for every SNP.

Association model

- Fit parameter once on null model

$$\hat{\boldsymbol{H}}, \hat{\sigma^2_{\mathrm{g}}}, \hat{\sigma^2_h}, \hat{\sigma^2} = \underset{\hat{\boldsymbol{H}}, \hat{\sigma^2_h}, \hat{\sigma^2_{\mathrm{g}}}, \hat{\sigma^2}}{\operatorname{argmax}} P(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{H}, \boldsymbol{\Theta}_{\mathsf{K}})$$

- Association testing for fixed hyperparameters

$$p(\boldsymbol{y}_{:,g} \mid \boldsymbol{x}_i, \boldsymbol{H}, \boldsymbol{\Theta}_{\mathsf{K}}) = \mathcal{N}\left(\boldsymbol{y}_{:,g} \mid \boldsymbol{x}_i\theta_i, \alpha^2(\hat{\sigma^2}\hat{\boldsymbol{H}}\hat{\boldsymbol{H}}^\top + \hat{\sigma^2_{\mathrm{g}}}\boldsymbol{X}\boldsymbol{X}^\top) + \sigma^2\boldsymbol{I}\right)$$

Association model

- Fit parameter once on null model

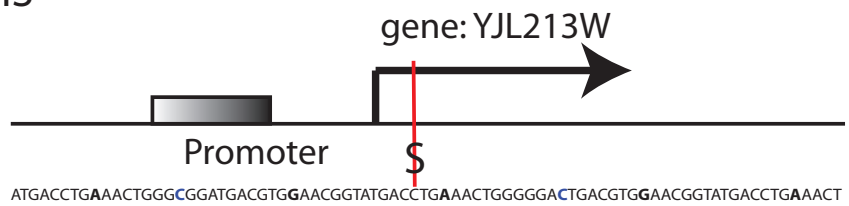
$$\hat{\boldsymbol{H}}, \hat{\sigma^2_{\mathrm{g}}}, \hat{\sigma^2_h}, \hat{\sigma^2} = \underset{\hat{\boldsymbol{H}}, \hat{\sigma^2_h}, \hat{\sigma^2_{\mathrm{g}}}, \hat{\sigma^2}}{\operatorname{argmax}} P(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{H}, \boldsymbol{\Theta}_{\mathsf{K}})$$

- Association testing for fixed hyperparameters

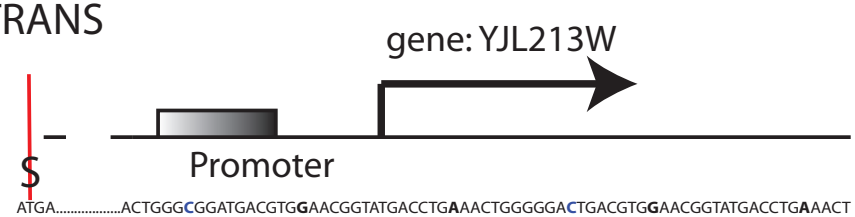
$$p(\boldsymbol{y}_{:,g} \mid \boldsymbol{x}_i, \boldsymbol{H}, \boldsymbol{\Theta}_{\mathsf{K}}) = \mathcal{N}\left(\boldsymbol{y}_{:,g} \mid \boldsymbol{x}_i\theta_i, \alpha^2(\hat{\sigma^2}\hat{\boldsymbol{H}}\hat{\boldsymbol{H}}^\top + \hat{\sigma^2_{\mathrm{g}}}\boldsymbol{X}\boldsymbol{X}^\top) + \sigma^2\boldsymbol{I}\right)$$

Evaluation on real data

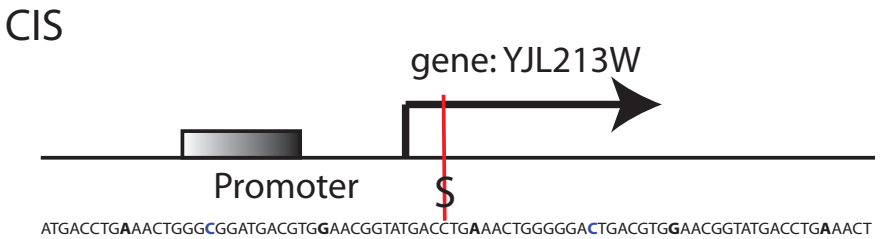
CIS

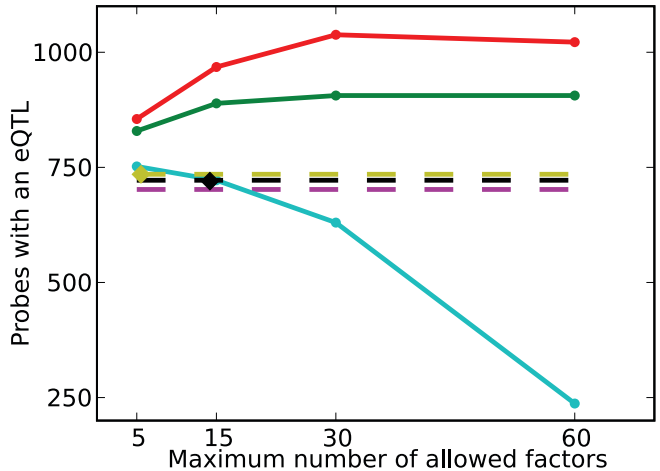


TRANS

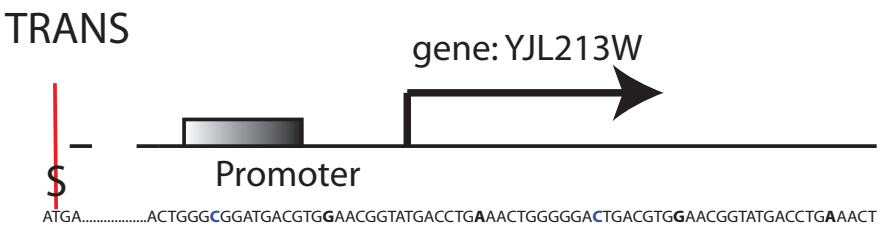


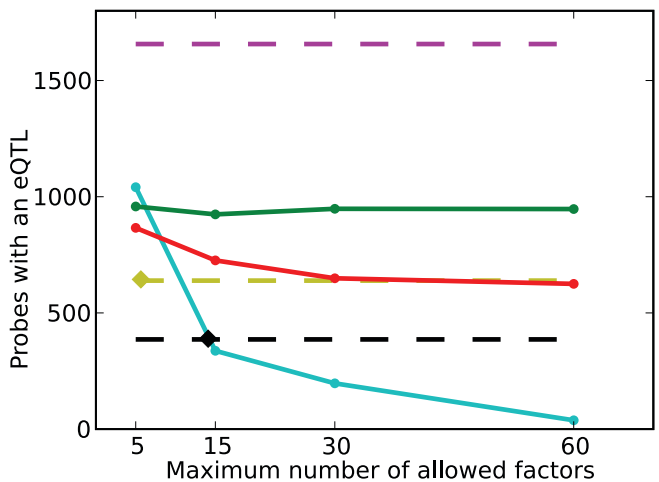
Evaluation on real data





(a) Yeast *cis* eQTLs

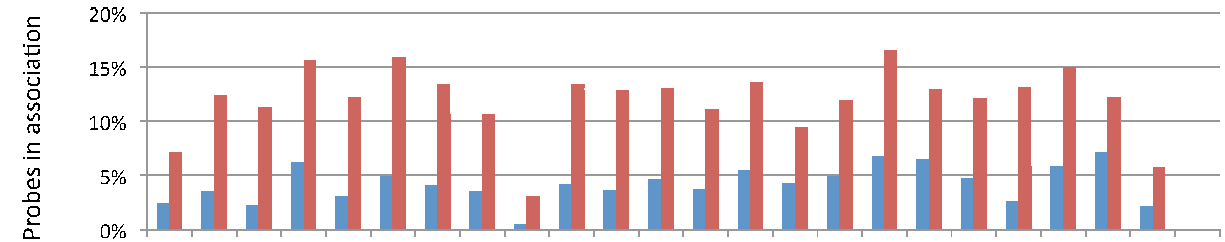




(b) Yeast *trans* eQTLs

Evaluation on real data

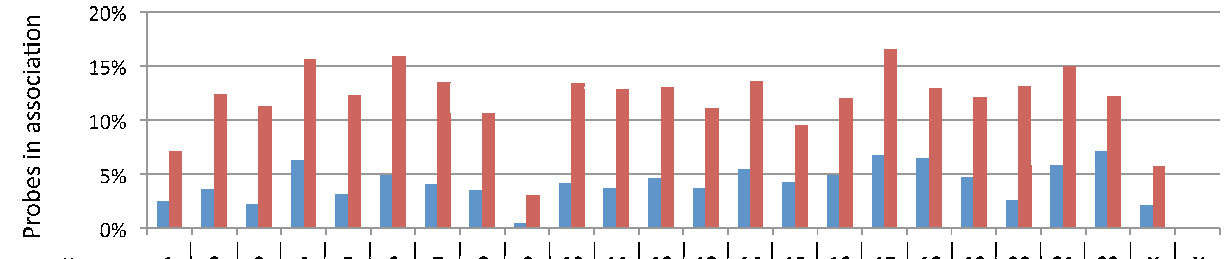
Application to HapMap II expression analysis



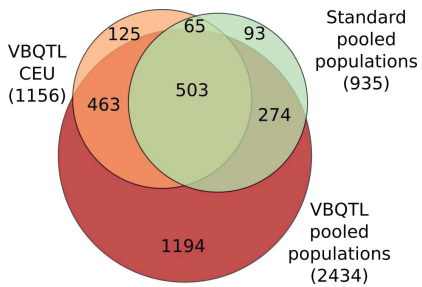
Chr.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	Y
Probes	1009	644	540	384	449	571	468	338	387	380	545	520	189	330	348	426	549	154	618	266	120	238	328	15
Standard	25	23	12	24	14	28	19	12	2	16	20	24	7	18	15	21	37	10	29	7	7	17	7	0
VBQTL	72	80	61	60	55	91	63	36	12	51	70	68	21	45	33	51	91	20	75	35	18	29	19	0

Evaluation on real data

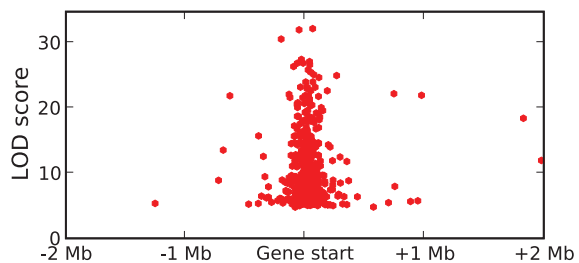
Application to HapMap II expression analysis



Chr.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	Y
Probes	1009	644	540	384	449	571	468	338	387	380	545	520	189	330	348	426	549	154	618	266	120	238	328	15
Standard	25	23	12	24	14	28	19	12	2	16	20	24	7	18	15	21	37	10	29	7	7	17	7	0
VBQTL	72	80	61	60	55	91	63	36	12	51	70	68	21	45	33	51	91	20	75	35	18	29	19	0



(a) Probes with a VBeQTL in pooled population



(f) *cis* VBeQTL location and strength relative to gene start

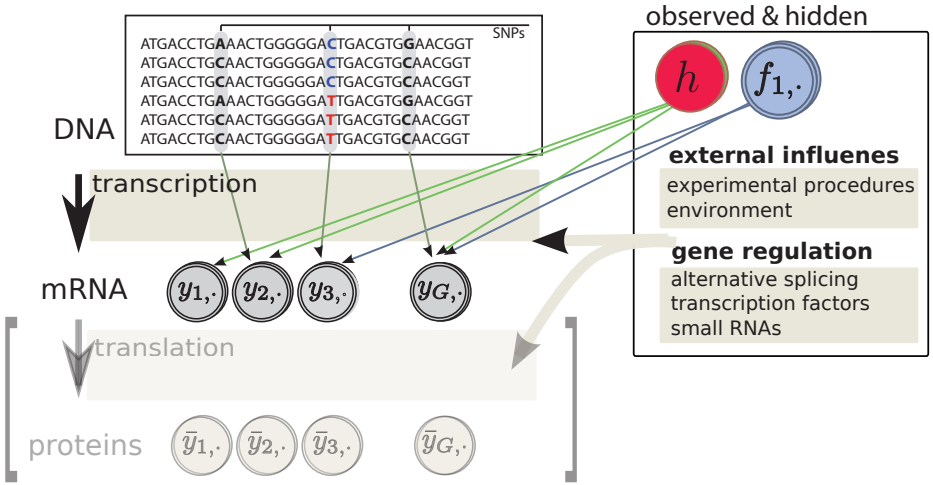
Outline

Association studies

Regulatory and external factors

- Confounding factors
  - Accounting for hidden confounders

$$y_{g,j} = \overbrace{b_{n,g}\left(\boldsymbol{\theta}_{n,g}\boldsymbol{s}_{n,j}\right)}^{\text{genetic}} + \underbrace{\boldsymbol{v}_g \boldsymbol{f}_j}_{\text{known factors}} + \underbrace{\boldsymbol{w}_g \boldsymbol{x}_j}_{\text{hidden factors}} + \underbrace{\psi_{n,g}}_{\text{noise}} .$$



- Accounting for regulatory factors?

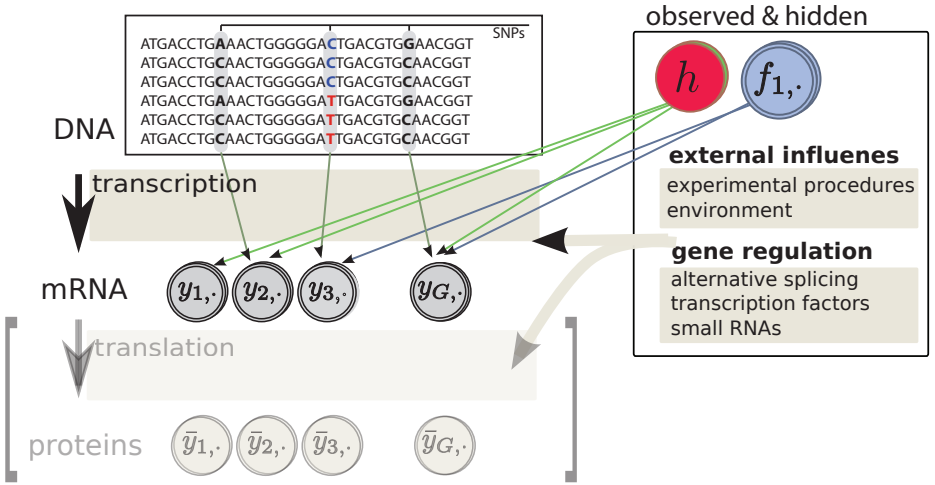
Association studies

Regulatory and external factors

- Confounding factors
  - Accounting for hidden confounders

$$y_{g,j} = \overbrace{b_{n,g}\left(\boldsymbol{\theta}_{n,g}\boldsymbol{s}_{n,j}\right)}^{\text{genetic}} + \underbrace{\boldsymbol{v}_g \boldsymbol{f}_j}_{\text{known factors}} + \underbrace{\boldsymbol{w}_g \boldsymbol{x}_j}_{\text{hidden factors}} + \underbrace{\psi_{n,g}}_{\text{noise}}.$$

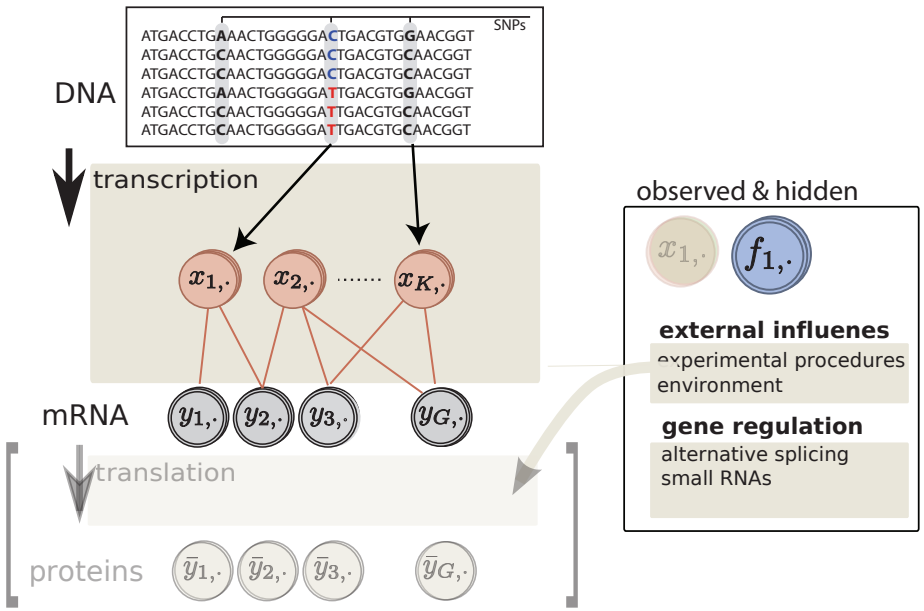
- Accounting for regulatory factors?



Association studies

Regulatory factors

- Account for regulatory factors:
  - Transcription factors
  - Pathway components
- Hypothesis: intermediate cellular factors mediate the observed association signals of target genes.
- Measuring X?
  - Difficult and expensive.
- Learn the unobserved factors X.

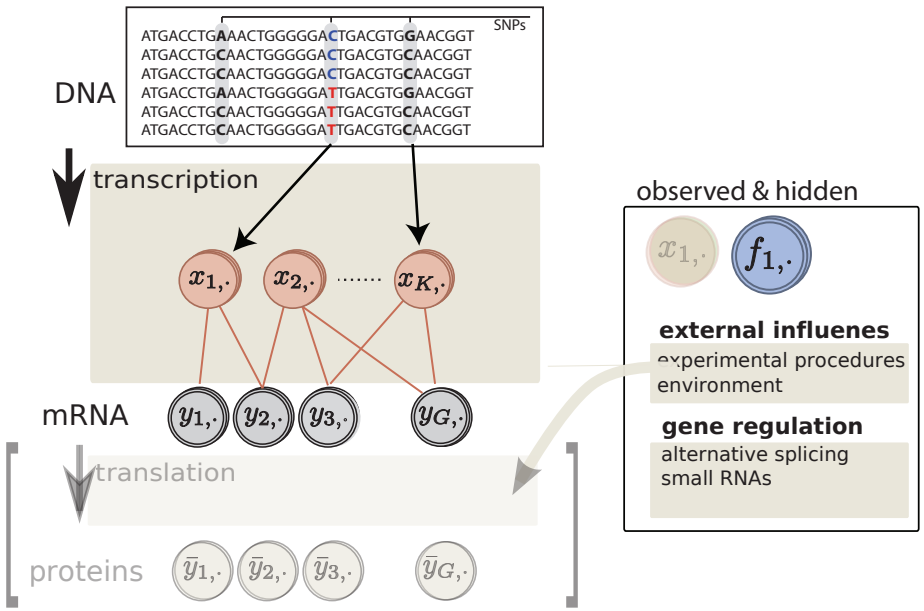


[Parts et al. 2011]

Association studies

Regulatory factors

- Account for regulatory factors:
  - Transcription factors
  - Pathway components
- Hypothesis: intermediate cellular factors mediate the observed association signals of target genes.
- Measuring $\boldsymbol{X}$?
  - Difficult and expensive.
- Learn the unobserved factors $\boldsymbol{X}$.

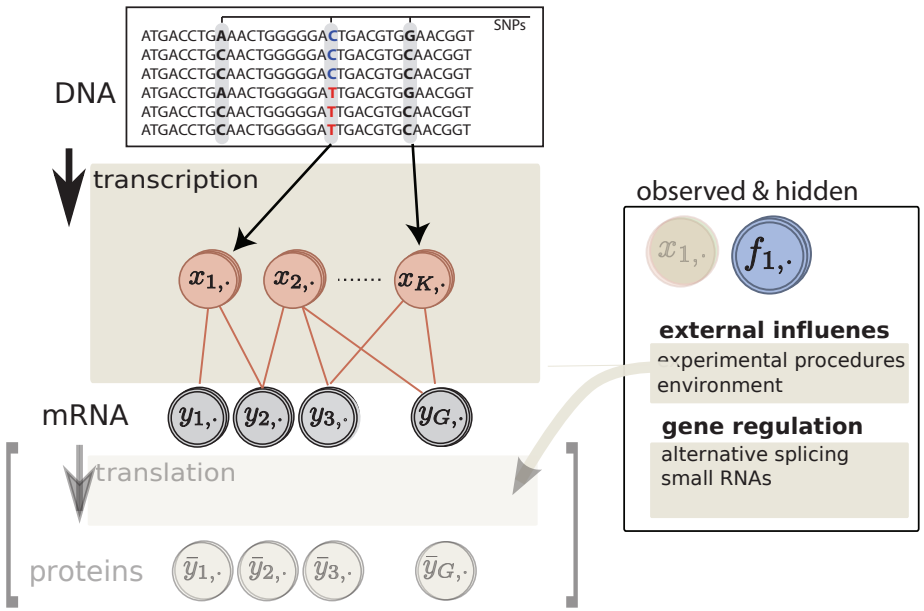


[Parts et al. 2011]

Association studies

Regulatory factors

- Account for regulatory factors:
  - Transcription factors
  - Pathway components
- Hypothesis: intermediate cellular factors mediate the observed association signals of target genes.
- Measuring $\boldsymbol{X}$?
  - Difficult and expensive.
- Learn the unobserved factors $\boldsymbol{X}$.



[Parts et al. 2011]

Association studies

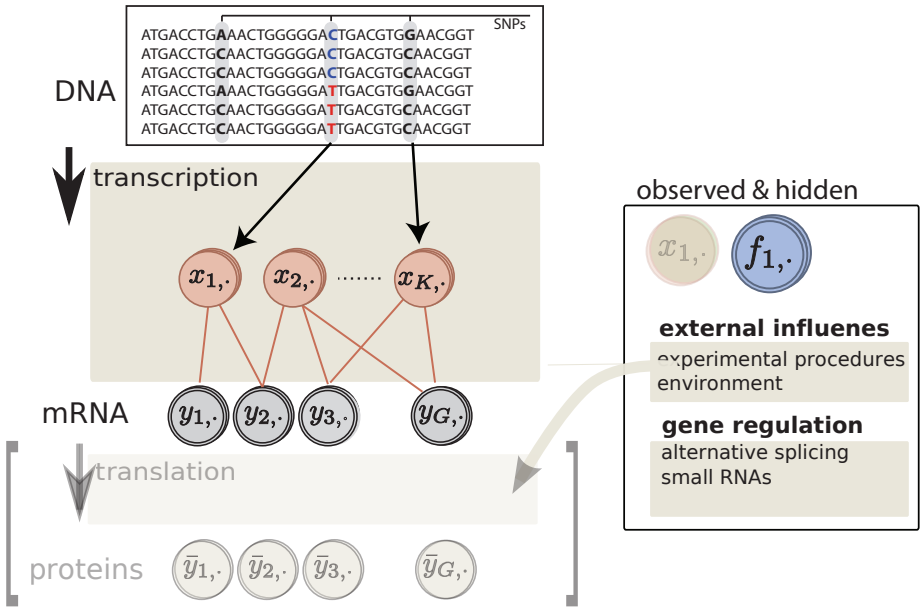
Transcription factor effects

- Inference of regulatory factors using a bilinear model

$$\underbrace{\boldsymbol{Y}}_{\text{Expr.}} = \underbrace{\boldsymbol{W}}_{\text{Weights}} \cdot \underbrace{\boldsymbol{X}}_{\text{Factors}} + \underbrace{\boldsymbol{\Psi}}_{\text{Noise}}.$$

  - $\boldsymbol{W}$ is sparse; each factor regulates only a subset of all genes.

- Incorporate biological prior knowledge to that factor are interpretable:
  - Transcription factor binding affinities.
- Inferred factors summarize co-expression clusters.



Association studies

Transcription factor effects

- Inference of regulatory factors using a bilinear model

$$\underbrace{\boldsymbol{Y}}_{\text{Expr.}} = \underbrace{\boldsymbol{W}}_{\text{Weights}} \cdot \underbrace{\boldsymbol{X}}_{\text{Factors}} + \underbrace{\boldsymbol{\Psi}}_{\text{Noise}}.$$

  - $\boldsymbol{W}$ is sparse; each factor regulates only a subset of all genes.
- Incorporate biological prior knowledge to that factor are interpretable:
  - Transcription factor binding affinities.
- Inferred factors summarize co-expression clusters.

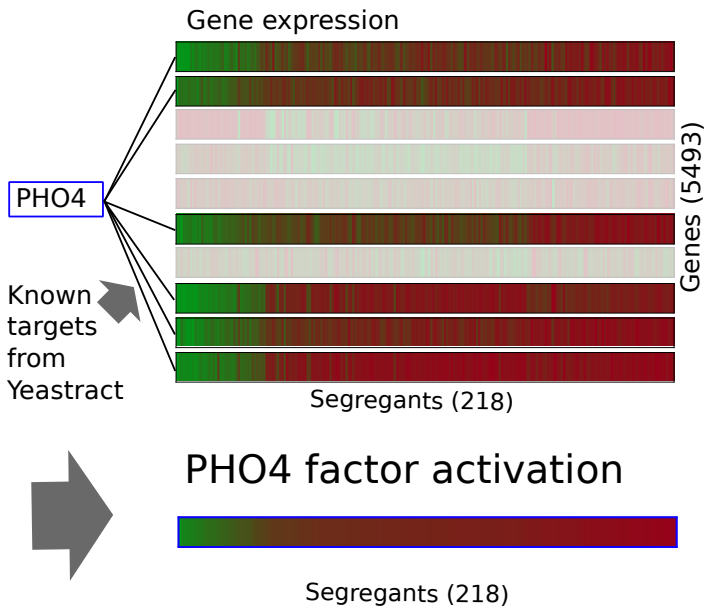
Association studies

Transcription factor effects

- Inference of regulatory factors using a bilinear model

$$\underbrace{\boldsymbol{Y}}_{\text{Expr.}} = \underbrace{\boldsymbol{W}}_{\text{Weights}} \cdot \underbrace{\boldsymbol{X}}_{\text{Factors}} + \underbrace{\boldsymbol{\Psi}}_{\text{Noise}}.$$

  - $\boldsymbol{W}$ is sparse; each factor regulates only a subset of all genes.
- Incorporate biological prior knowledge to that factor are interpretable:
  - Transcription factor binding affinities.
- Inferred factors summarize co-expression clusters.



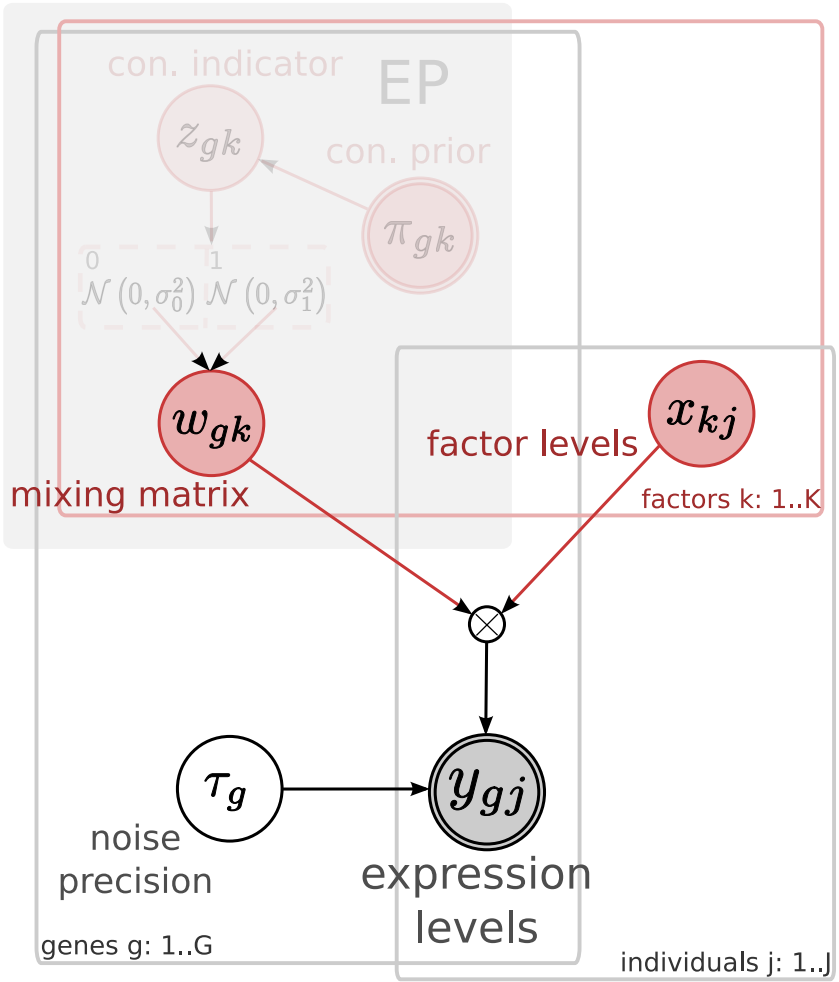
Sparse factor analysis

Probabilistic model

- Graphical model for $\boldsymbol{Y} = \boldsymbol{W} \cdot \boldsymbol{X} + \boldsymbol{\Psi}$.
- Indicators $z_{g,k}$ determine the sparsity pattern:

$$P(w_{g,k} \,|\, z_{g,k} = 0) = \mathcal{N}\left(w_{g,k} \mid 0, \sigma_0^2\right)$$
$$P(w_{g,k} \,|\, z_{g,k} = 1) = \mathcal{N}\left(w_{g,k} \mid 0, \sigma_1^2\right).$$

- Prior knowledge is encoded in $\pi_{g,k} = P(z_{g,k} = 1)$.
- Standard conjugate priors for the remaining random variables.



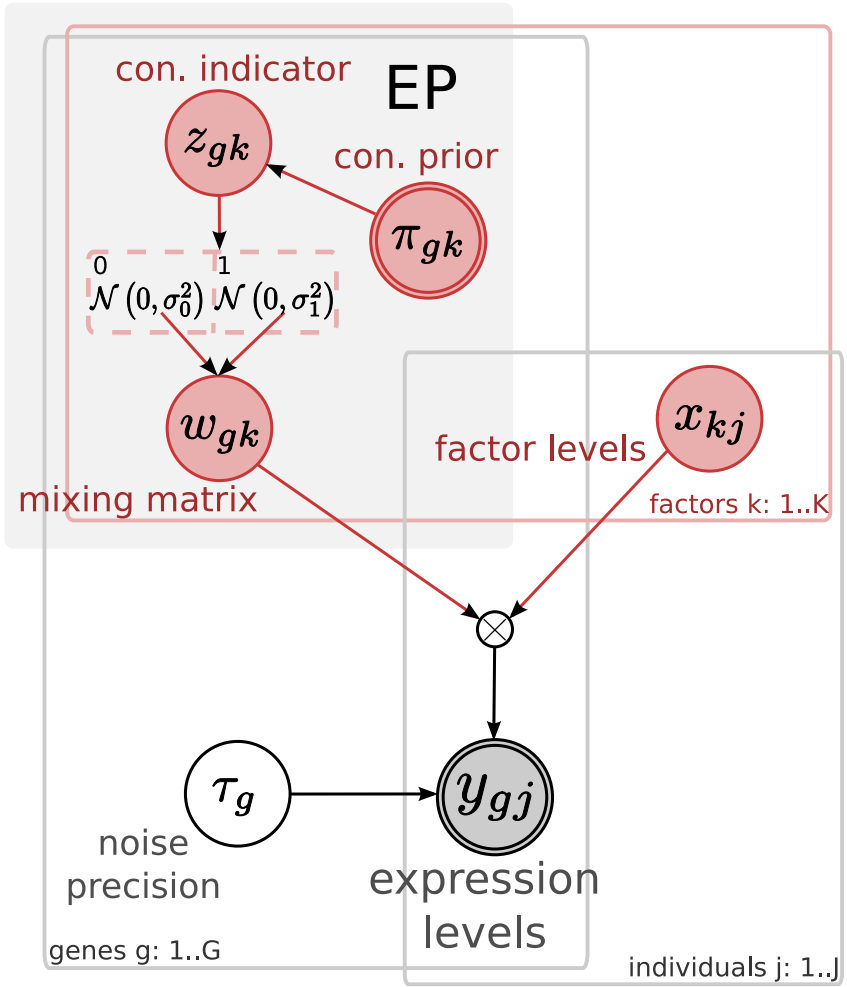
Sparse factor analysis

Probabilistic model

- Graphical model for $\boldsymbol{Y} = \boldsymbol{W} \cdot \boldsymbol{X} + \boldsymbol{\Psi}$.
- Indicators $z_{g,k}$ determine the sparsity pattern:

$$P(w_{g,k} \,|\, z_{g,k} = 0) = \mathcal{N}\left(w_{g,k} \mid 0, \sigma_0^2\right)$$
$$P(w_{g,k} \,|\, z_{g,k} = 1) = \mathcal{N}\left(w_{g,k} \mid 0, \sigma_1^2\right).$$

- Prior knowledge is encoded in $\pi_{g,k} = P(z_{g,k} = 1)$.
- Standard conjugate priors for the remaining random variables.



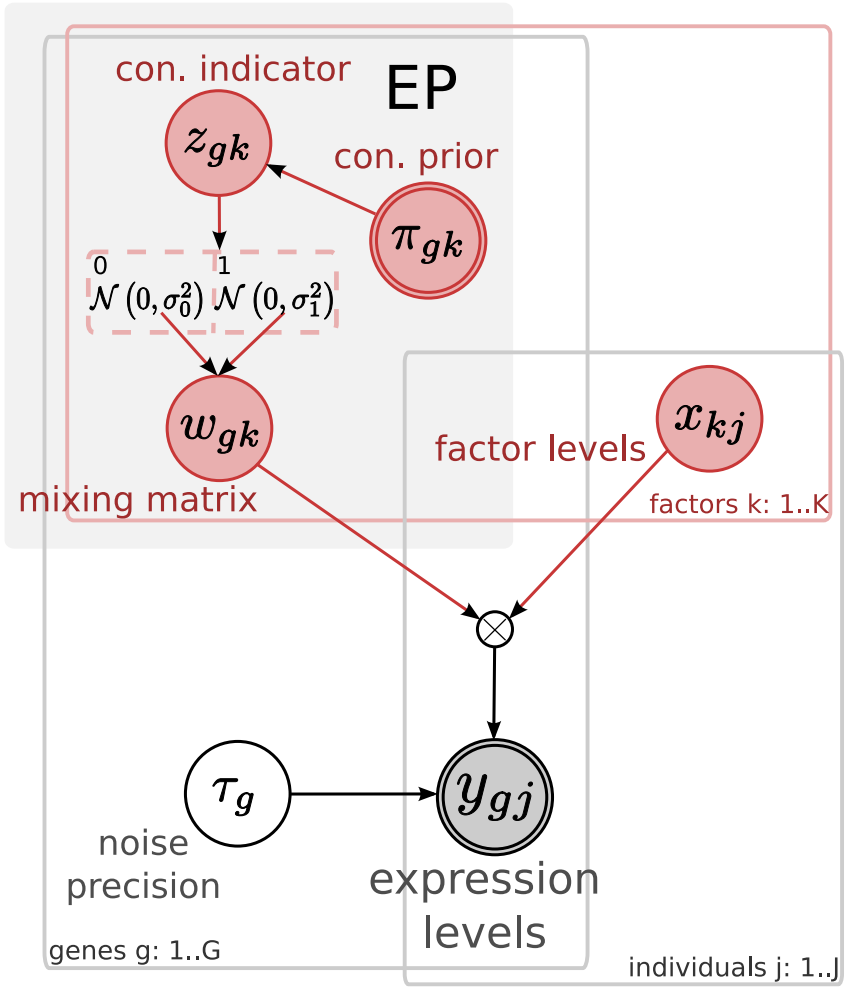
Sparse factor analysis

Probabilistic model

- Graphical model for $\boldsymbol{Y} = \boldsymbol{W} \cdot \boldsymbol{X} + \boldsymbol{\Psi}$.
- Indicators $z_{g,k}$ determine the sparsity pattern:

$$P(w_{g,k} \,|\, z_{g,k} = 0) = \mathcal{N}\left(w_{g,k} \mid 0, \sigma_0^2\right)$$
$$P(w_{g,k} \,|\, z_{g,k} = 1) = \mathcal{N}\left(w_{g,k} \mid 0, \sigma_1^2\right).$$

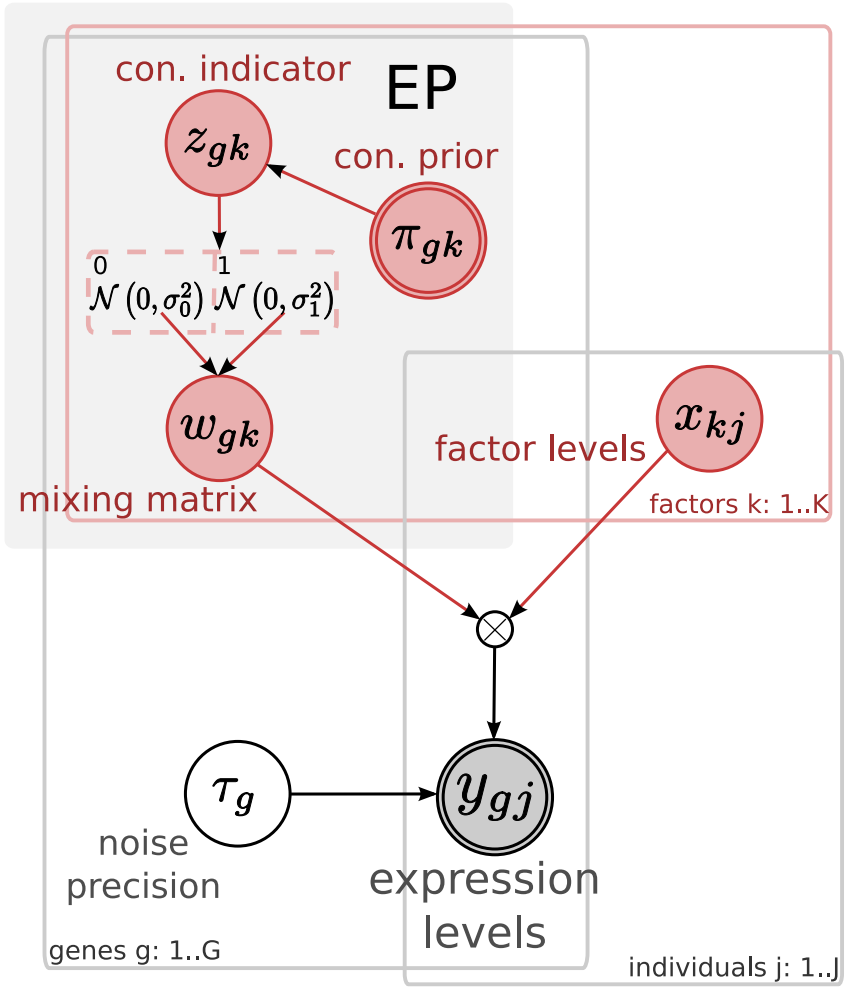
- Prior knowledge is encoded in $\pi_{g,k} = P(z_{g,k} = 1)$.
- Standard conjugate priors for the remaining random variables.



Sparse factor analysis

Inference

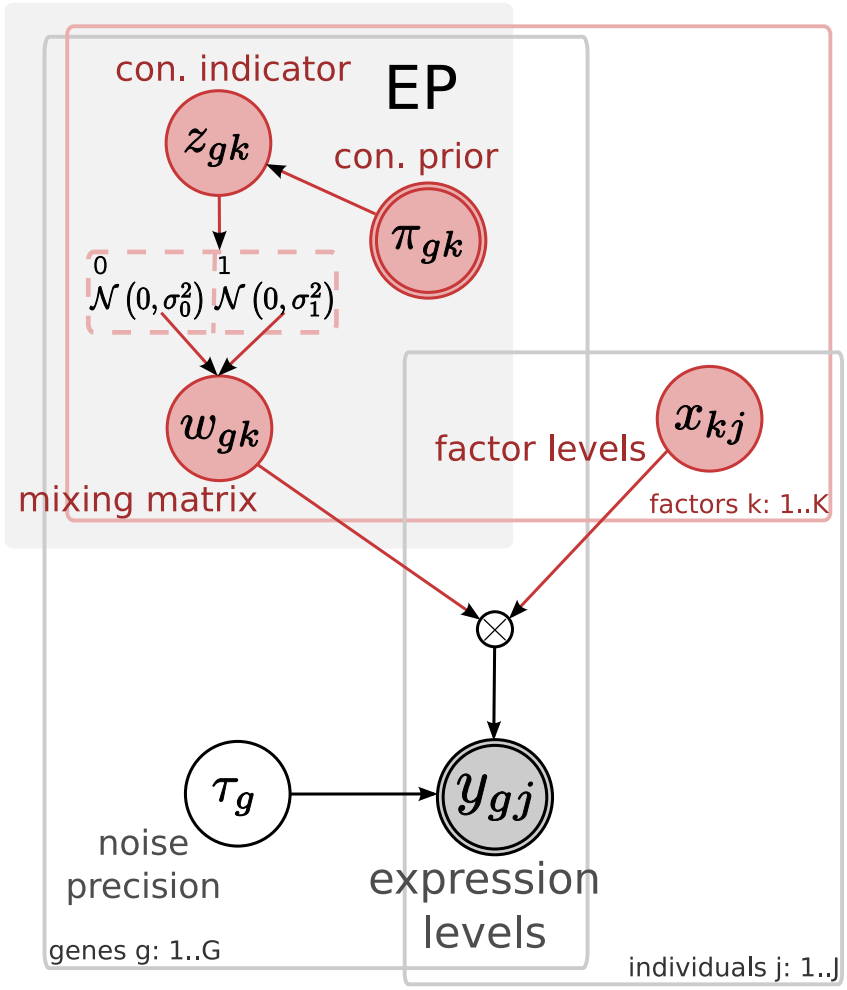
- Scale is challenging: thousands of genes, hundreds of TFs.
- Efficient deterministic approximate inference:
  1. Variational Bayesian inference for the core model.
  2. Expectation Propagation for the sparsity submodel.



Sparse factor analysis

Inference

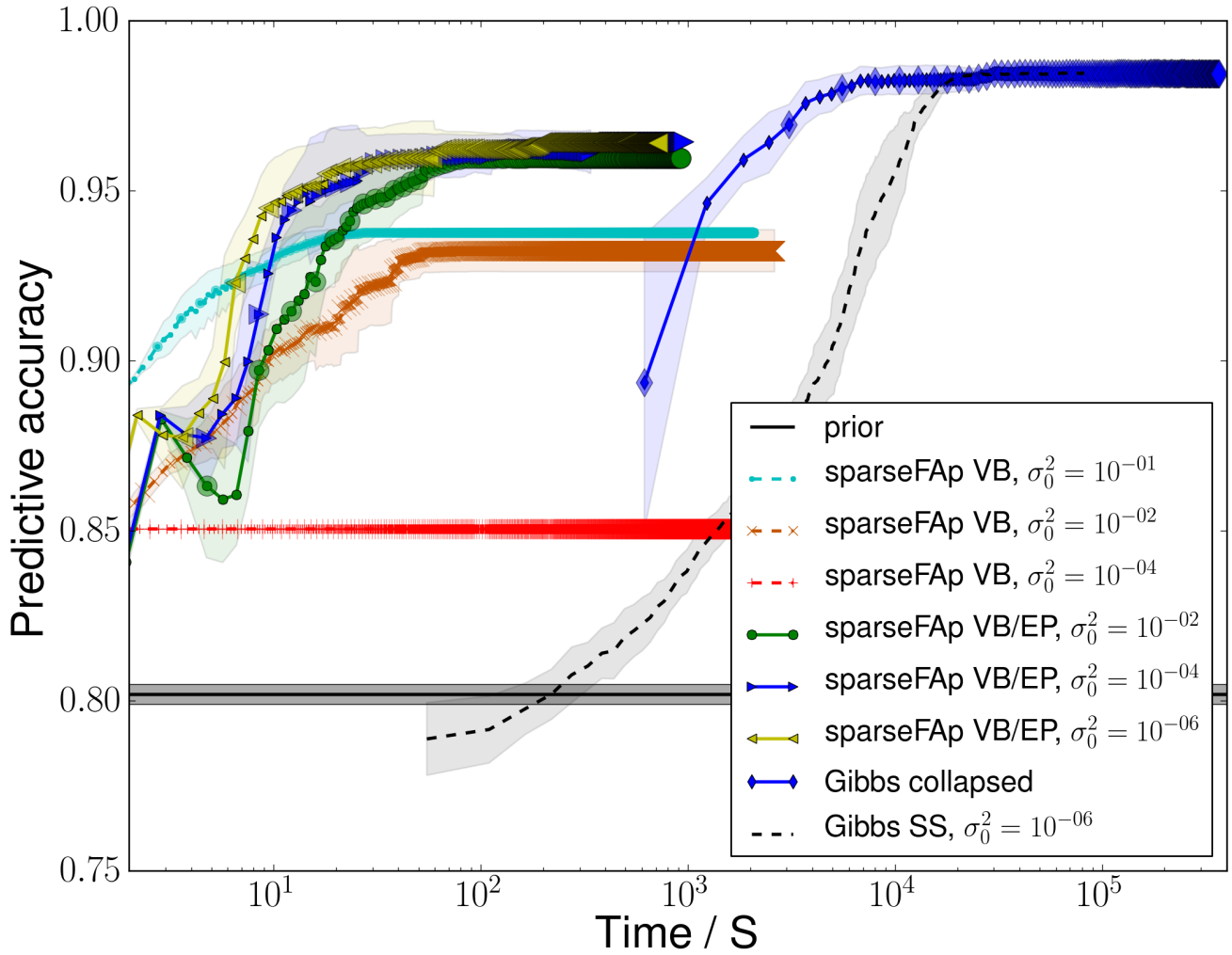
- Scale is challenging: thousands of genes, hundreds of TFs.
- Efficient deterministic approximate inference:
  1. Variational Bayesian inference for the core model.
  2. Expectation Propagation for the sparsity submodel.



Sparse factor analysis

Comparison to MCMC on (small!) simulated dataset

Recovery of the true simulated network structure.

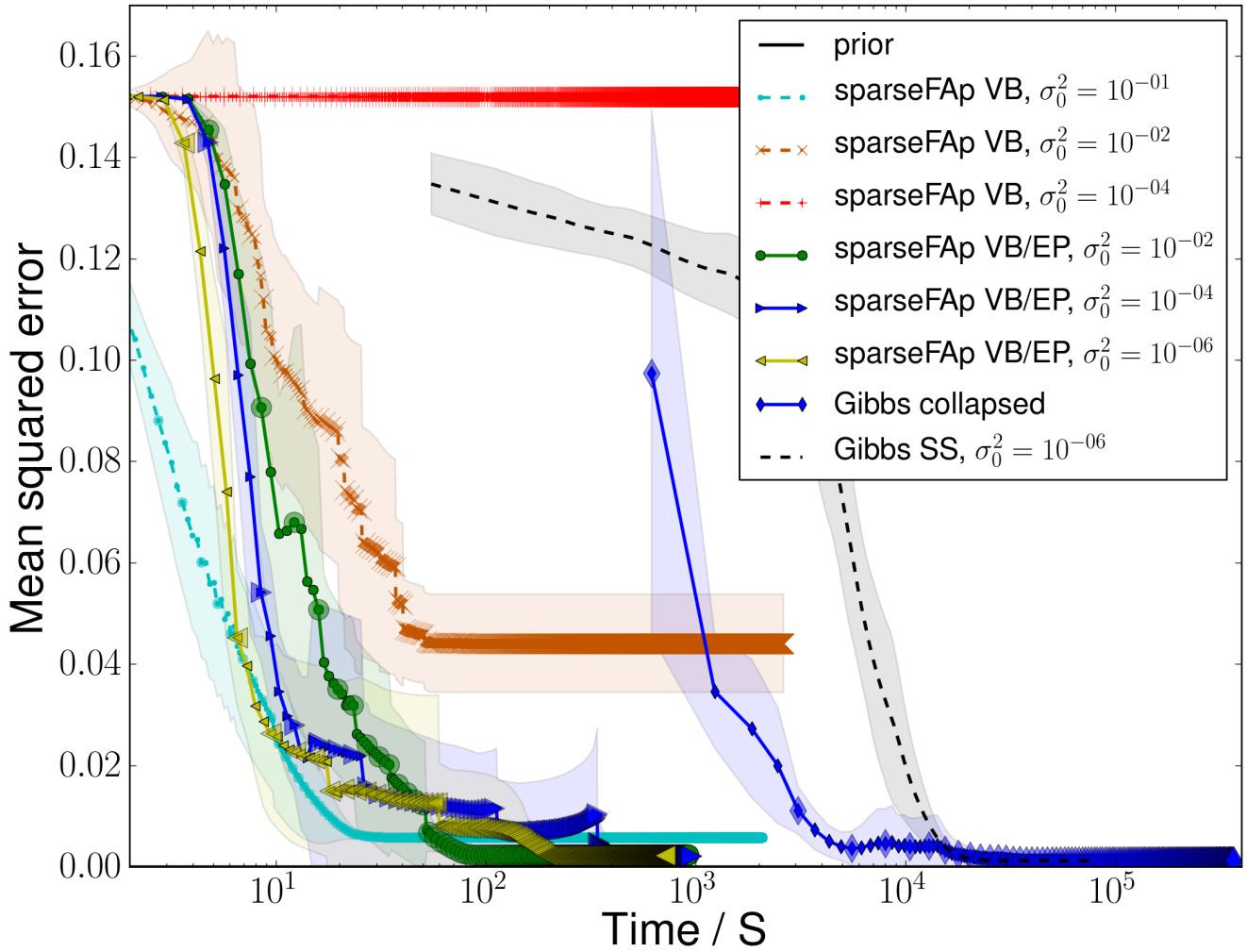


Stegle et al., in preparation

Sparse factor analysis

Comparison to MCMC on (small!) simulated dataset

Recovery of the true simulated factor activations.



Stegle et al., in preparation

Application to yeast

Dataset

- Applied the approach to 108 yeast strains (crosses), genotyped and expression profiled in 2 conditions (ethanol and glucose).
- Alternative types of prior information available
  - TF binding affinities (Yeastract).
  - Pathway information (KEGG).

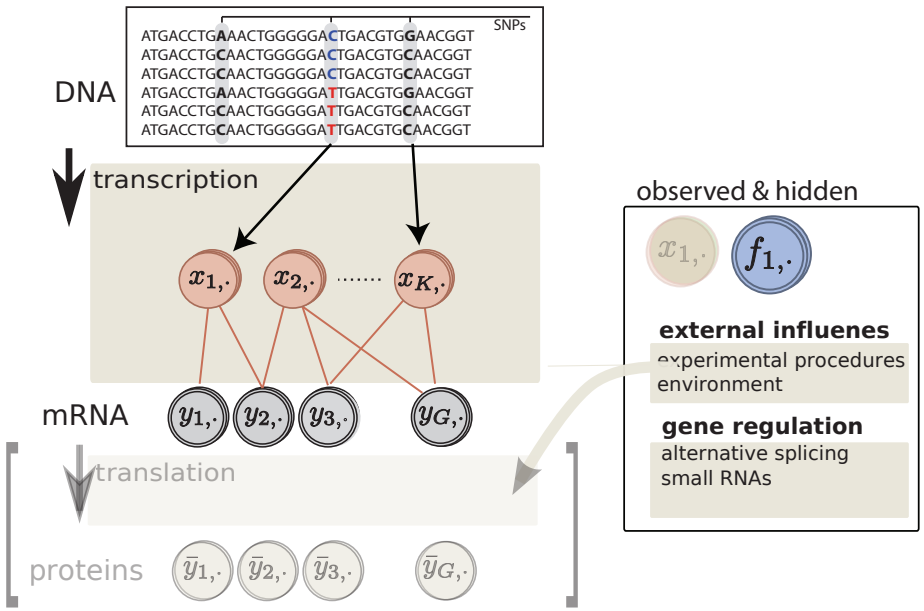
[Smith et al. 2008]

Application to yeast

Factor associations

Biological hypotheses

- Genetic variation (SNPs) may regulate factor activations.
- Similarly for the environment condition (glucose/ethanol).
- Interaction effects between SNPs, factors and genes.

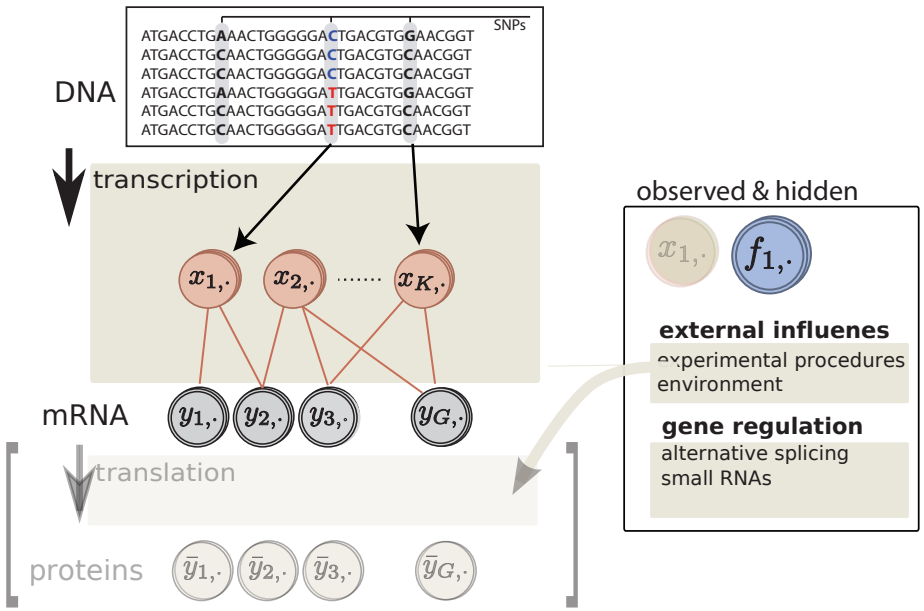


Application to yeast

Factor associations

Biological hypotheses

- Genetic variation (SNPs) may regulate factor activations.
- Similarly for the environment condition (glucose/ethanol).
- Interaction effects between SNPs, factors and genes.

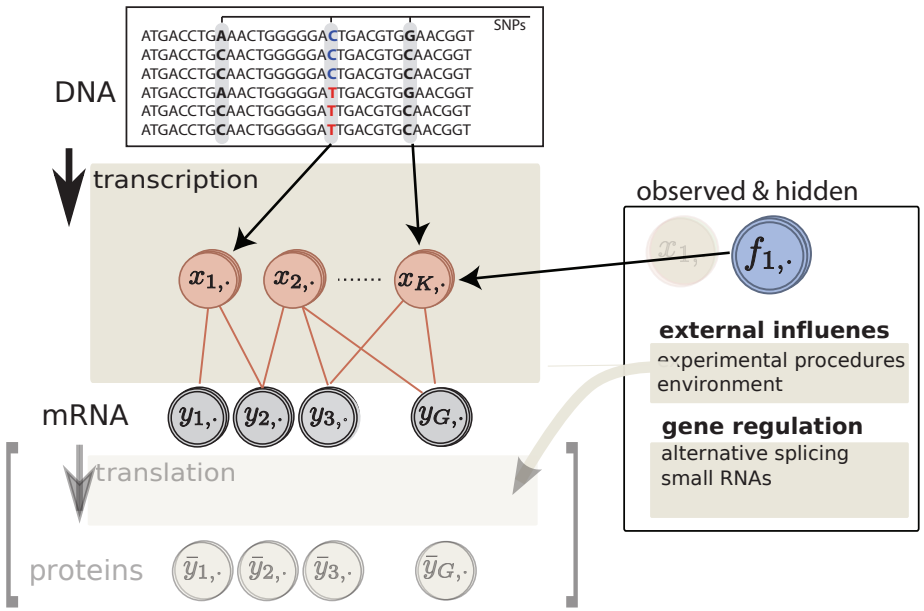


Application to yeast

Factor associations

Biological hypotheses

- Genetic variation (SNPs) may regulate factor activations.
- Similarly for the environment condition (glucose/ethanol).
- Interaction effects between SNPs, factors and genes.

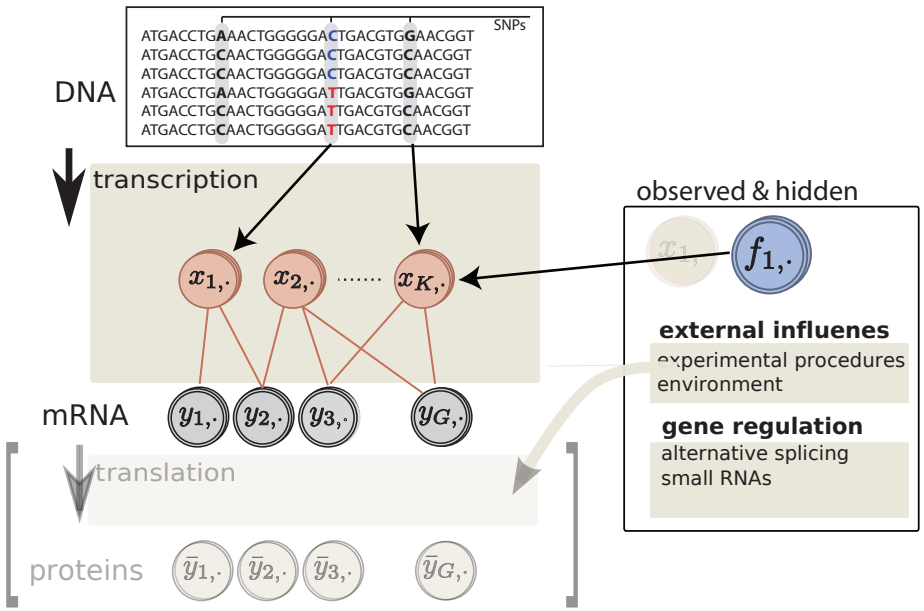


Application to yeast

Factor associations

Biological hypotheses

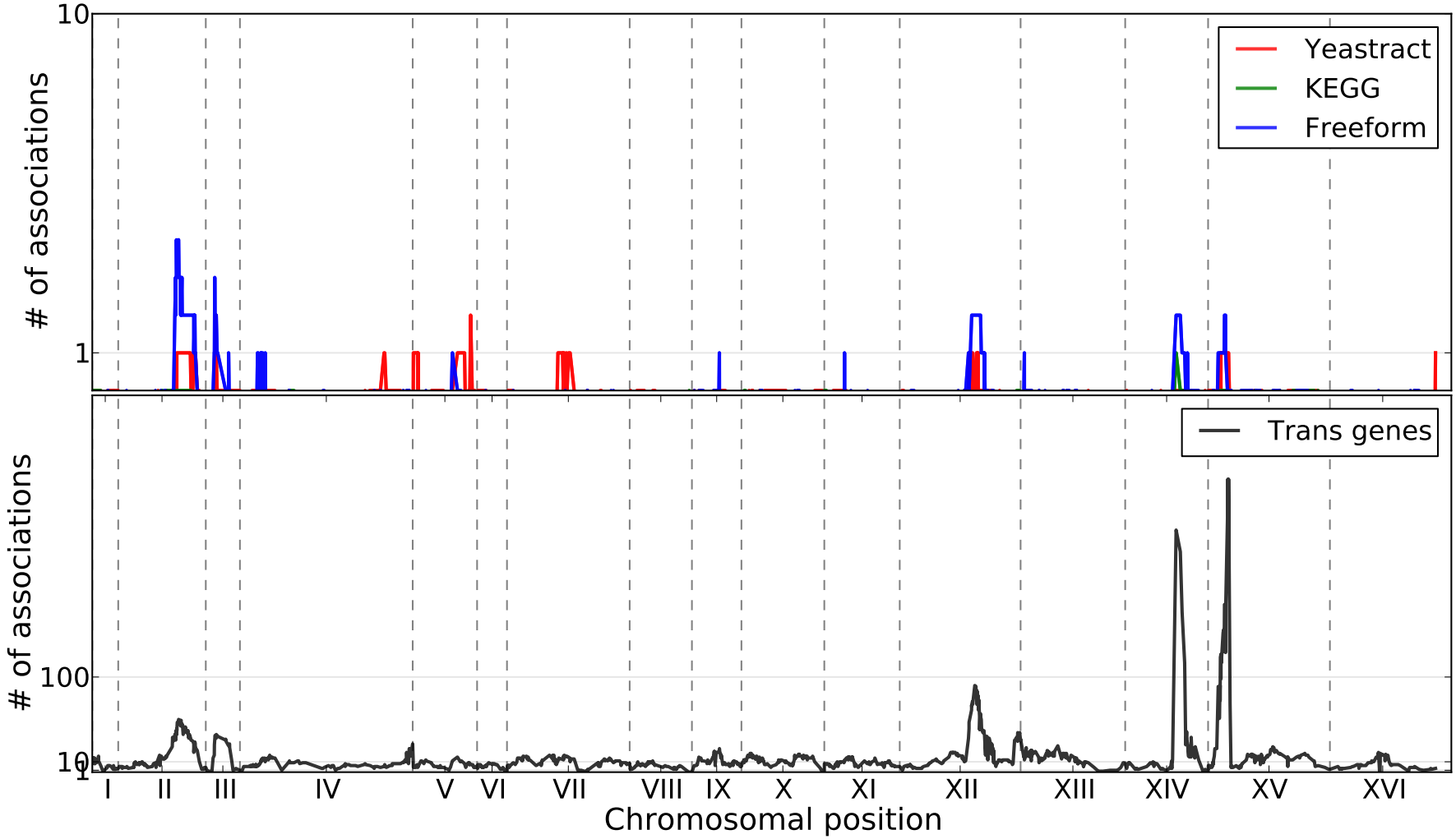
- Genetic variation (SNPs) may regulate factor activations.
- Similarly for the environment condition (glucose/ethanol).
- Interaction effects between SNPs, factors and genes.



Application to yeast

Factor associations

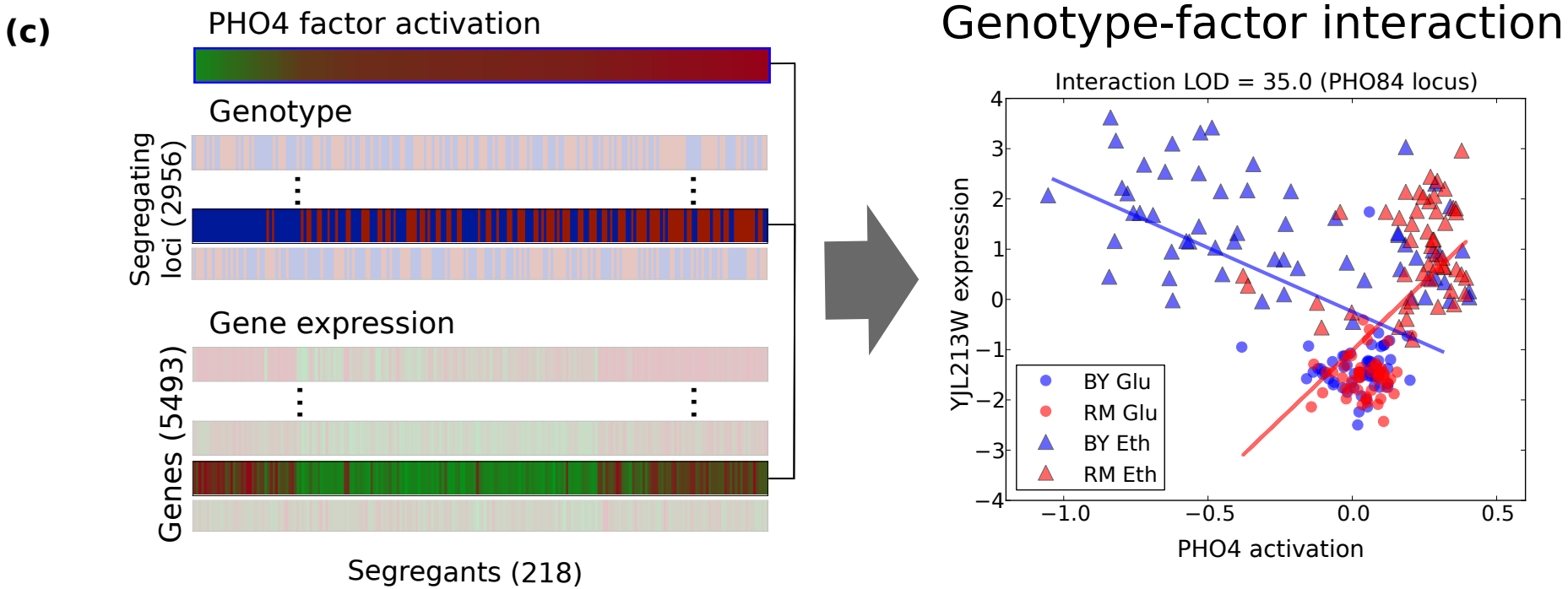
- Genome-wide association density.



Application to yeast

Factor interactions

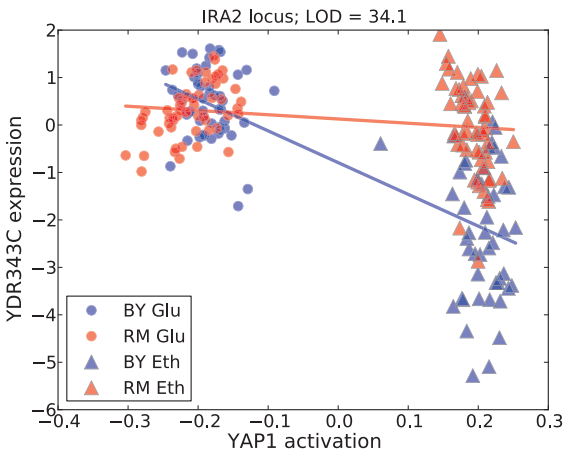
- Interaction tests between all gene/SNP/factor triplets.



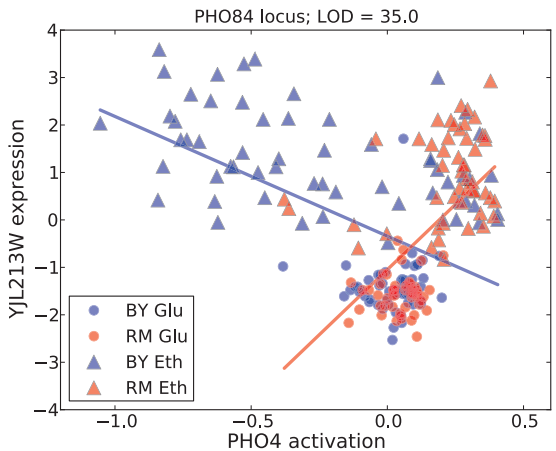
Application to yeast

Factor interactions

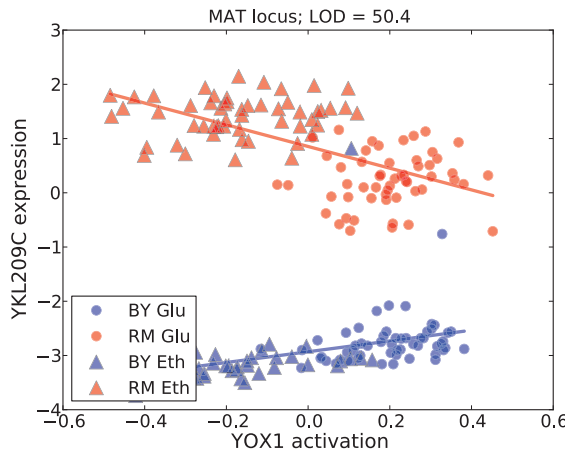
- Example interactions.



(a) YAP1-IRA2 interaction



(b) PHO4-PHO84 interaction

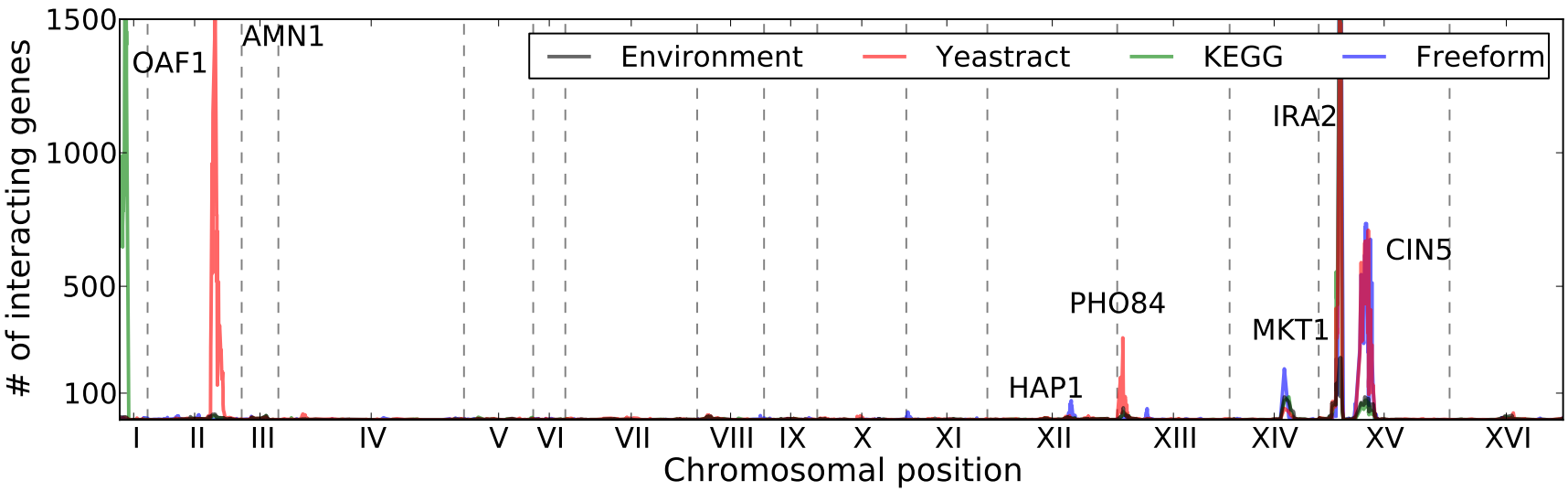


(c) MAT-YOX1 interaction

Application to yeast

Factor interactions

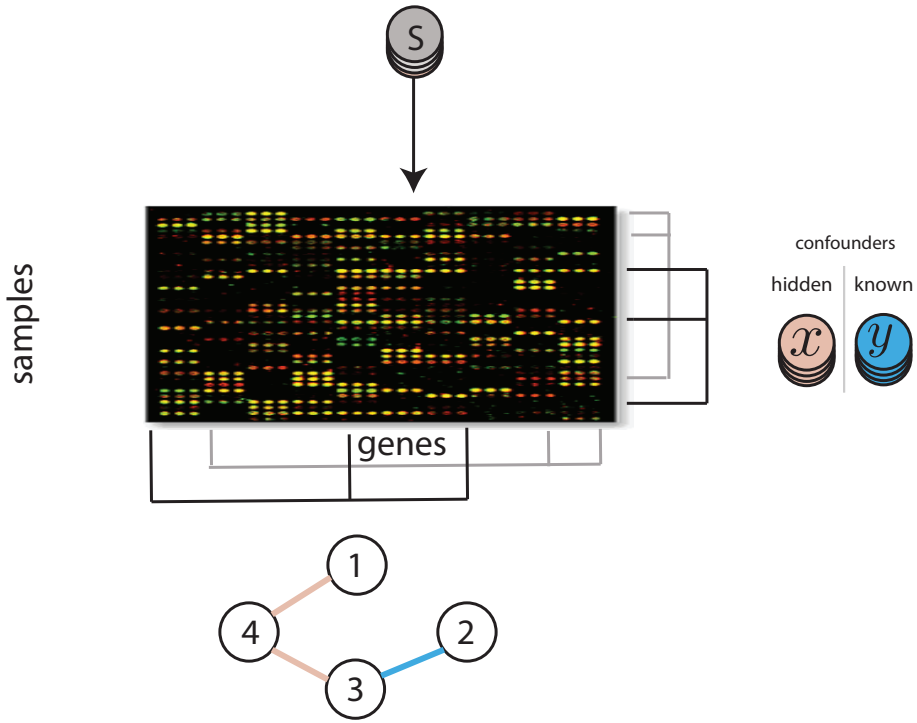
- Genome-wide interaction density.



Outline

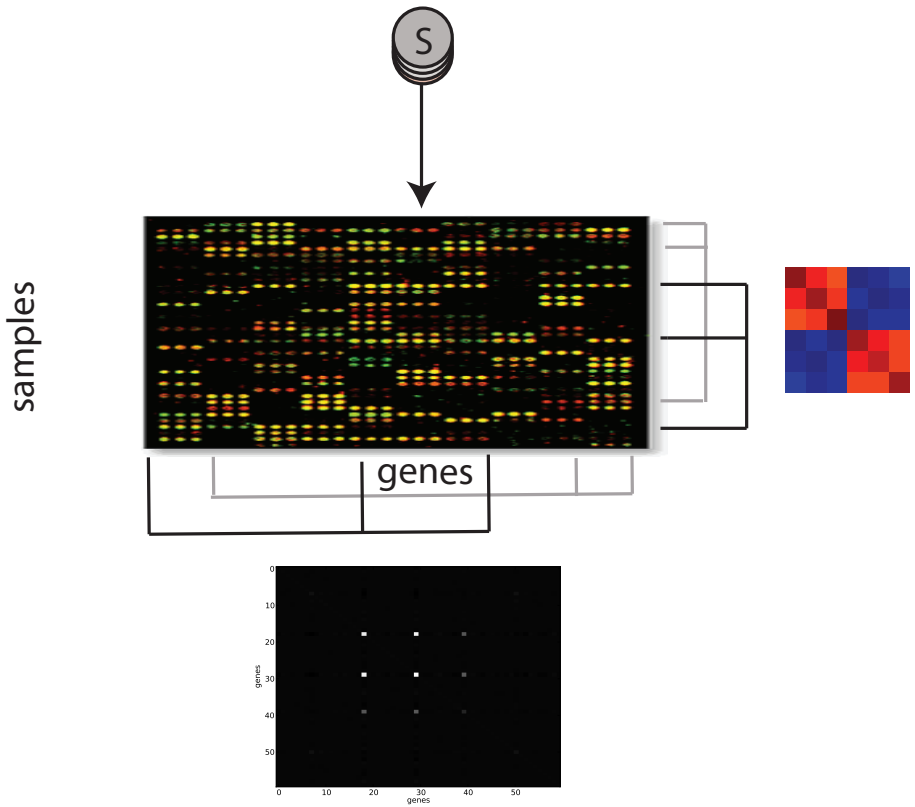
Matrix variate normal distributions

- Confounders induces structure between samples.
- Gene regulation induces structure between genes.
- Matrix variate models for joint correction.



Matrix variate normal distributions

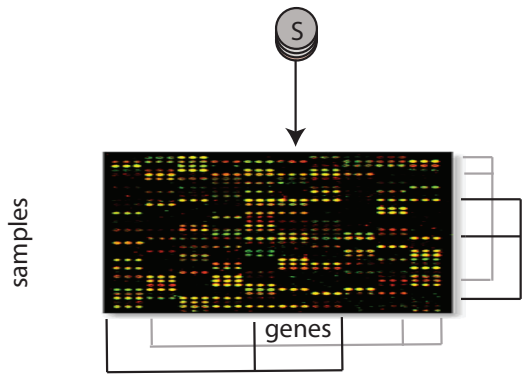
- Confounders induces structure between samples.
- Gene regulation induces structure between genes.
- Matrix variate models for joint correction.



Matrix variate normal distributions

- Confounders induces structure between samples.
- Gene regulation induces structure between genes.
- Matrix variate models for joint correction.
-

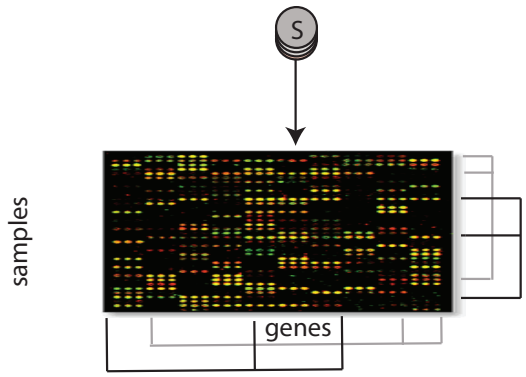
Matrix variate normal distributions



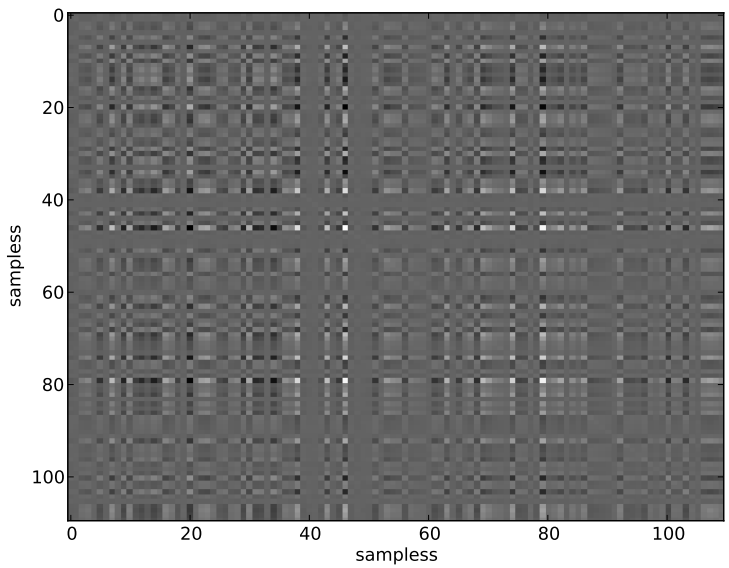
Row covariances (samples) $\boldsymbol{K}$

Column covariances (genes) $\boldsymbol{\Lambda}^{-1}$

Matrix variate normal distributions

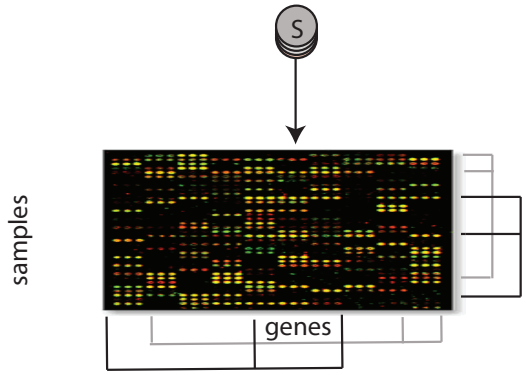


Row covariances (samples) $\boldsymbol{K}$

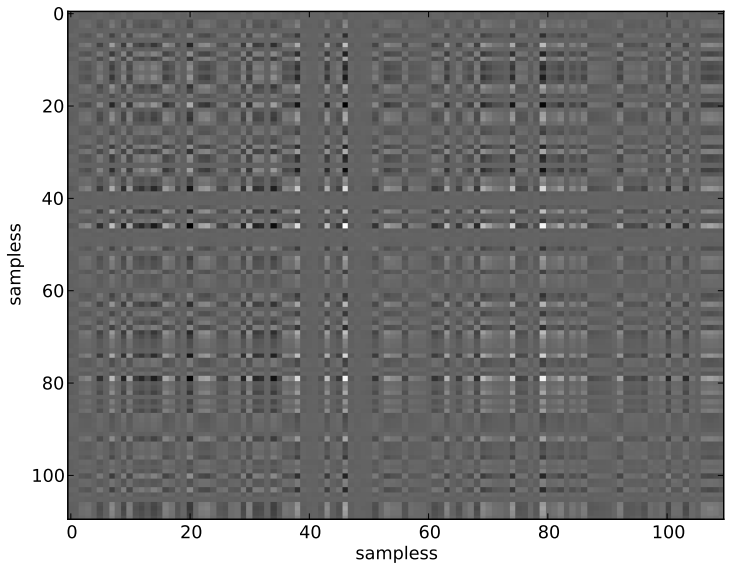


Column covariances (genes) $\boldsymbol{\Lambda}^{-1}$

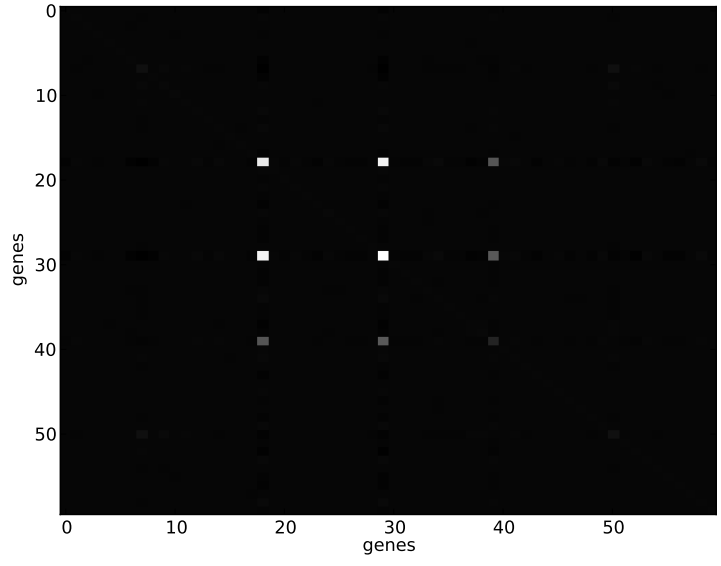
Matrix variate normal distributions



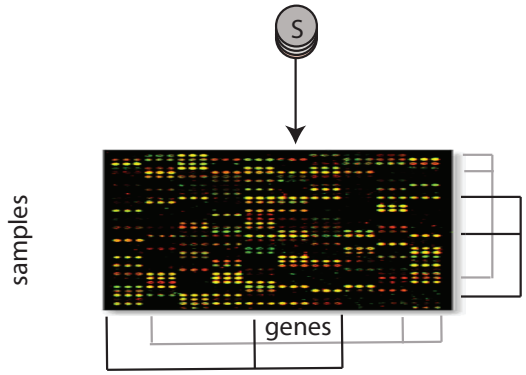
Row covariances (samples) $\boldsymbol{K}$



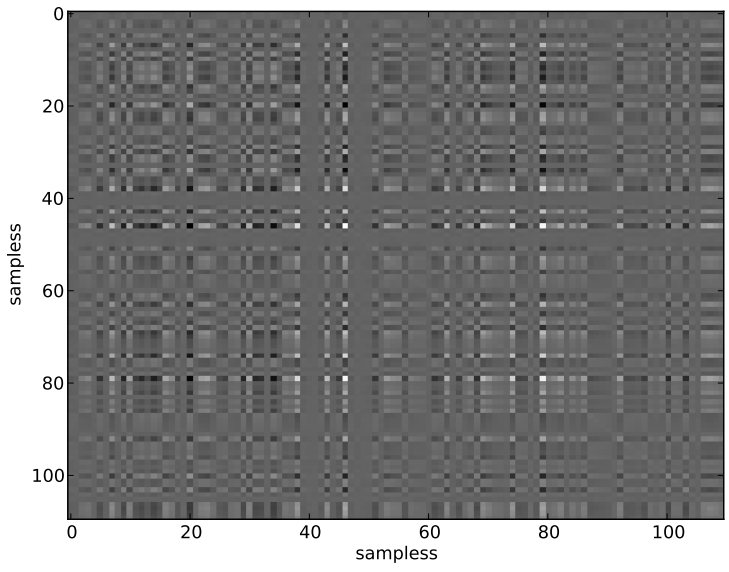
Column covariances (genes) $\boldsymbol{\Lambda}^{-1}$



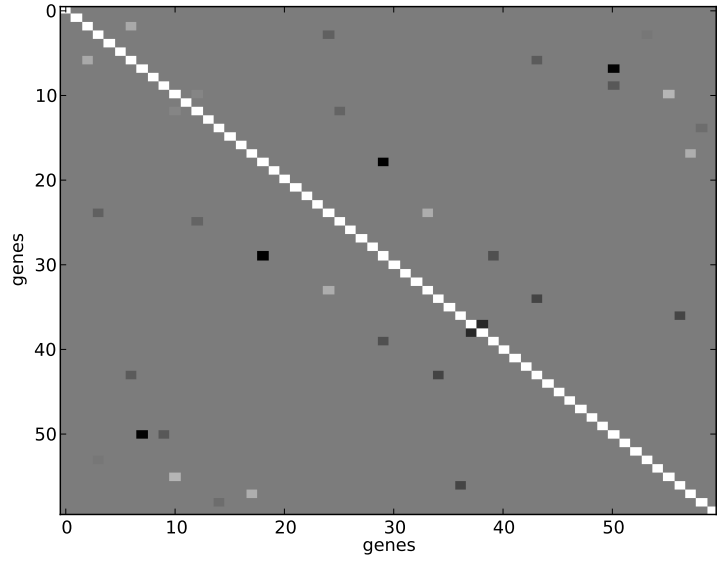
Matrix variate normal distributions



Row covariances (samples) $\boldsymbol{K}$



Column covariances (genes) $\boldsymbol{\Lambda}^{-1}$



Matrix variate distributions

- Matrix variate distribution with column covariance $\boldsymbol{\Lambda}^{-1}$ and row covariance $\boldsymbol{K}$

$$p(\boldsymbol{Y} \mid \boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Lambda}^{-1}, \boldsymbol{K}) \propto \exp\left(-0.5\text{tr}\left[\boldsymbol{\Lambda}\boldsymbol{Y}^\top \boldsymbol{K}^{-1}\boldsymbol{Y}\right]\right)$$

- Equivalent to Kronecker covariance structure on vectorized data

$$p(\text{vec}\boldsymbol{Y} \mid \boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Lambda}^{-1}, \boldsymbol{K}) = \mathcal{N}\left(\text{vec}\boldsymbol{Y} \mid \mathbf{0}, \boldsymbol{\Lambda}^{-1} \otimes \boldsymbol{K}\right)$$

Matrix variate distributions

- Matrix variate distribution with column covariance $\boldsymbol{\Lambda}^{-1}$ and row covariance $\boldsymbol{K}$

$$p(\boldsymbol{Y} \mid \boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Lambda}^{-1}, \boldsymbol{K}) \propto \exp\left(-0.5\text{tr}\left[\boldsymbol{\Lambda}\boldsymbol{Y}^\top \boldsymbol{K}^{-1}\boldsymbol{Y}\right]\right)$$

- Equivalent to Kronecker covariance structure on vectorized data

$$p(\text{vec}\boldsymbol{Y} \mid \boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Lambda}^{-1}, \boldsymbol{K}) = \mathcal{N}\left(\text{vec}\boldsymbol{Y} \mid \mathbf{0}, \boldsymbol{\Lambda}^{-1} \otimes \boldsymbol{K}\right)$$

Drawing samples from a MN

1. sample $Y \sim \prod_{r=1}^{R}\prod_{g=1}^{G}\mathcal{N}(0,1)$
2. $Y = \text{chol}(\boldsymbol{K}) \cdot Y$
3. $Y = Y \cdot \text{chol}(\boldsymbol{\Lambda}^{-1})^\top$

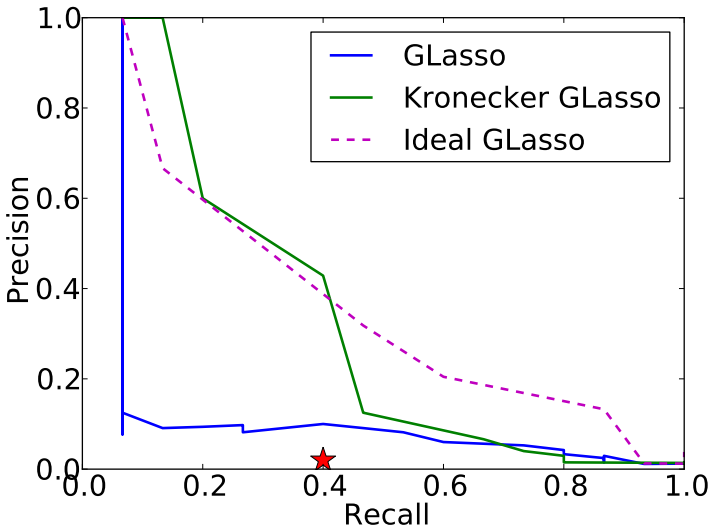
Kronecker matrix product

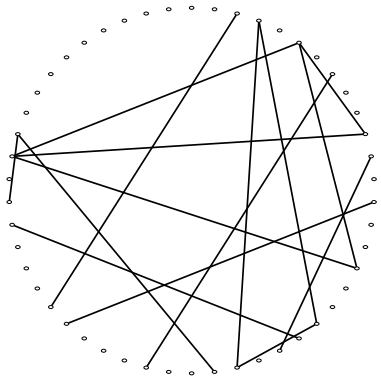
$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{pmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \dots & a_{1H}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \dots & a_{2H}\boldsymbol{B} \\ \dots & \dots & \dots & \vdots \\ a_{G1}\boldsymbol{B} & a_{G2}\boldsymbol{B} & \dots & a_{GH}\boldsymbol{B} \end{pmatrix}$$

Simulated example

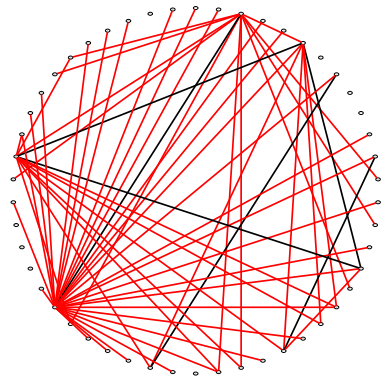
better recovery of the true graph

- Recovery of simulated network. Weak confounding variation (20% variance in row covariance).





ground truth

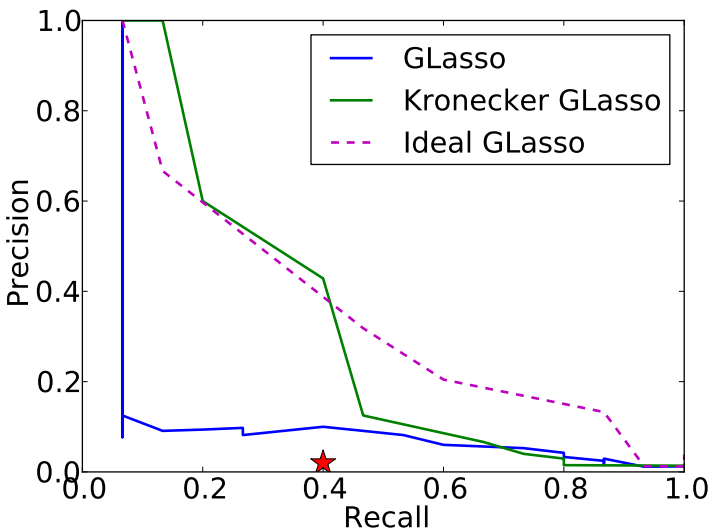


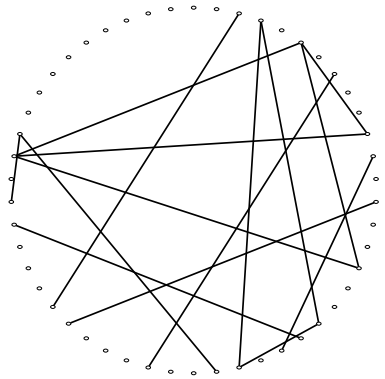
GLasso

Simulated example

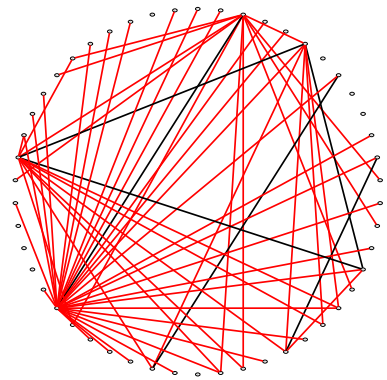
better recovery of the true graph

- Recovery of simulated network. Weak confounding variation (20% variance in row covariance).

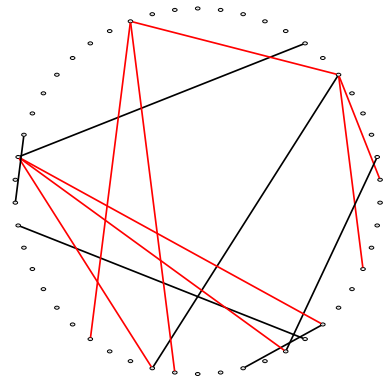




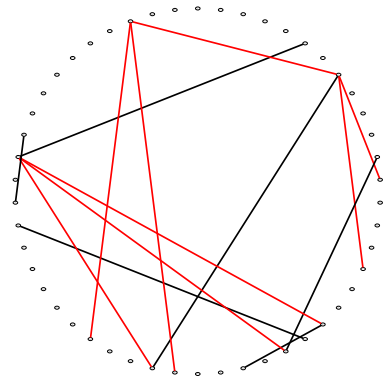
ground truth



GLasso



Kornecker GLasso



Ideal GLasso

Outline

Summary

- Latent variables can have a dramatic effect in GWAs.
- In expression studies, confounders can be estimated from data.
- Cellular features can be learnt from the expression profiles.
- Duality of regulatory relationships and confounders in matrix variate normal models.