Estimating the contribution of non-genetic factors to gene expression using Gaussian Process Latent Variable Models

Nicolò Fusi and Neil Lawrence

Learning and Inference in Computational Systems Biology

31st March 2010

1 eQTL mapping

2 Dataset

3 The model

4 Experiments

5 Conclusions

Outline

1 eQTL mapping

2 Dataset

3 The model

4 Experiments

5 Conclusions

Expression Quantitative Trait Loci - eQTL

- Transcript abudance is regulated by polymorphisms in the regulatory elements
- Statistical methods can be used to discover which polymorphism affects the expression levels of a gene
- This mapping sometimes is obfuscated by non-genetic factors

Expression Quantitative Trait Loci - eQTL

- Transcript abudance is regulated by polymorphisms in the regulatory elements
- Statistical methods can be used to discover which polymorphism affects the expression levels of a gene
- This mapping sometimes is obfuscated by non-genetic factors

Expression Quantitative Trait Loci - eQTL

- Transcript abudance is regulated by polymorphisms in the regulatory elements
- Statistical methods can be used to discover which polymorphism affects the expression levels of a gene
- This mapping sometimes is obfuscated by non-genetic factors

Outline

1. eQTL mapping
2. Dataset
3. The model
4. Experiments
5. Conclusions

Single Nucleotide Polymorphisms

A single nucleotide polymorphism is a variation in the DNA sequence that affects only one nucleotide.

- They make up about 90% of all human genetic variation
- They capture 84% of the total genetic variation in gene expression

Single Nucleotide Polymorphisms

A single nucleotide polymorphism is a variation in the DNA sequence that affects only one nucleotide.

- They make up about 90% of all human genetic variation
- They capture 84% of the total genetic variation in gene expression

Single Nucleotide Polymorphisms

A single nucleotide polymorphism is a variation in the DNA sequence that affects only one nucleotide.

- They make up about 90% of all human genetic variation
- They capture 84% of the total genetic variation in gene expression

The Hapmap dataset

- a multi-country effort to identify and catalog genetic similarities and differences in human beings
- 3.1 million human single nucleotide polymorphisms have been genotyped
- 270 individuals from 4 geographically diverse populations (Hapmap phase II)

The Hapmap dataset

- a multi-country effort to identify and catalog genetic similarities and differences in human beings
- 3.1 million human single nucleotide polymorphisms have been genotyped
- 270 individuals from 4 geographically diverse populations (Hapmap phase II)

The Hapmap dataset

- a multi-country effort to identify and catalog genetic similarities and differences in human beings
- 3.1 million human single nucleotide polymorphisms have been genotyped
- 270 individuals from 4 geographically diverse populations (Hapmap phase II)

Project GENEVAR - GENe Expression VARiation

- Gene expression data from EBV-transformed lymphoblastoid cell lines (Stranger et al., Nature Genetics 2007)
- 270 individuals from Hapmap phase I and II
- 47,293 gene probes

Project GENEVAR - GENe Expression VARiation

- Gene expression data from EBV-transformed lymphoblastoid cell lines (Stranger et al., Nature Genetics 2007)
- 270 individuals from Hapmap phase I and II
- 47,293 gene probes

Project GENEVAR - GENe Expression VARiation

- Gene expression data from EBV-transformed lymphoblastoid cell lines (Stranger et al., Nature Genetics 2007)
- 270 individuals from Hapmap phase I and II
- 47,293 gene probes

Outline

1 eQTL mapping

2 Dataset

3 The model

4 Experiments

5 Conclusions

Confounding factors

Several studies have shown that non-genetic factors can obfuscate associations:

- Known Factors: age, sex, ethnicity, ...
- Batch effects: optical effects
- Unknown factors

Confounding factors

Several studies have shown that non-genetic factors can obfuscate associations:

- Known Factors: age, sex, ethnicity, ...
- Batch effects: optical effects
- Unknown factors

Confounding factors

Several studies have shown that non-genetic factors can obfuscate associations:

- Known Factors: age, sex, ethnicity, ...
- Batch effects: optical effects
- Unknown factors

Modelling non-genetic factors

- Our model is inspired by *Stegle et al, Lecture notes in Computer Science (2006)*.
- We model non-genetic factors as unobserved latent variables.
- Gene expression levels are described as a linear function of SNP data and non-genetic factors

$$\mathbf{Y} = \mathbf{S}\mathbf{V} + \mathbf{X}\mathbf{W} + \mu\mathbf{1}^{\top} + \epsilon$$

Modelling non-genetic factors

- Our model is inspired by *Stegle et al, Lecture notes in Computer Science (2006)*.
- We model non-genetic factors as unobserved latent variables.
- Gene expression levels are described as a linear function of SNP data and non-genetic factors

$$\mathbf{Y} = \mathbf{SV} + \mathbf{XW} + \mu\mathbf{1}^{\top} + \epsilon$$

Modelling non-genetic factors

- Our model is inspired by *Stegle et al, Lecture notes in Computer Science (2006)*.
- We model non-genetic factors as unobserved latent variables.
- Gene expression levels are described as a linear function of SNP data and non-genetic factors

$$\mathbf{Y} = \mathbf{SV} + \mathbf{XW} + \mu\mathbf{1}^\top + \epsilon$$

Modelling non-genetic factors

- Our model is inspired by *Stegle et al, Lecture notes in Computer Science (2006)*.
- We model non-genetic factors as unobserved latent variables.
- Gene expression levels are described as a linear function of SNP data and non-genetic factors

$$\mathbf{Y} = \mathbf{SV} + \mathbf{XW} + \mu\mathbf{1}^\top + \epsilon$$

Modelling non-genetic factors

- Our model is inspired by *Stegle et al, Lecture notes in Computer Science (2006)*.
- We model non-genetic factors as unobserved latent variables.
- Gene expression levels are described as a linear function of SNP data and non-genetic factors

$$\mathbf{Y} = \mathbf{S}\mathbf{V} + \mathbf{X}\mathbf{W} + \mu\mathbf{1}^{\top} + \epsilon$$

Modelling non-genetic factors

- Our model is inspired by *Stegle et al, Lecture notes in Computer Science (2006)*.
- We model non-genetic factors as unobserved latent variables.
- Gene expression levels are described as a linear function of SNP data and non-genetic factors

$$\mathbf{Y} = \mathbf{SV} + \mathbf{XW} + \mu\mathbf{1}^\top + \epsilon$$

Modelling non-genetic factors

- Our model is inspired by *Stegle et al, Lecture notes in Computer Science (2006)*.
- We model non-genetic factors as unobserved latent variables.
- Gene expression levels are described as a linear function of SNP data and non-genetic factors

$$\mathbf{Y} = \mathbf{S}\mathbf{V} + \mathbf{X}\mathbf{W} + \mu\mathbf{1}^{\top} + \epsilon$$

Modelling non-genetic factors

- Our model is inspired by *Stegle et al, Lecture notes in Computer Science (2006)*.
- We model non-genetic factors as unobserved latent variables.
- Gene expression levels are described as a linear function of SNP data and non-genetic factors

$$\mathbf{Y} = \mathbf{S}\mathbf{V} + \mathbf{X}\mathbf{W} + \mu\mathbf{1}^{\top} + \epsilon$$

Modelling non-genetic factors

- Our model is inspired by *Stegle et al, Lecture notes in Computer Science (2006)*.
- We model non-genetic factors as unobserved latent variables.
- Gene expression levels are described as a linear function of SNP data and non-genetic factors

$$\mathbf{Y} = \mathbf{SV} + \mathbf{XW} + \mu\mathbf{1}^{\top} + \epsilon$$

dual Probabilistic Principal Component Analysis

We learn the parameters by:

- Marginalizing $\mathbf{W}, \mathbf{V}, \mu, \epsilon$
- Maximizing the log-likelihood with respect to the latent variables ($\mathbf{X}$)

For a particular choice of priors over **W** and **V** this approach is equivalent to probabilistic Principal Component Analysis

dual Probabilistic Principal Component Analysis

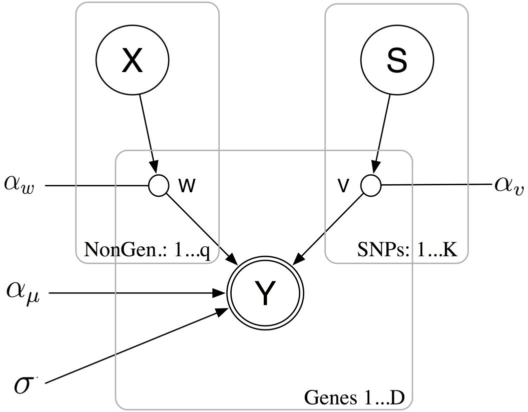
We put Gaussian priors over **W**, **V** and μ:

$$P(\mathbf{W}) = \prod_{i=1}^{D} N(\mathbf{w_i}|\mathbf{0}, \alpha_w \mathbf{I})$$

$$P(\mathbf{V}) = \prod_{i=1}^{D} N(\mathbf{v_i}|\mathbf{0}, \alpha_v \mathbf{I})$$

$$P(\mu) = N(\mu|\mathbf{0}, \alpha_\mu \mathbf{I})$$

dual Probabilistic Principal Component Analysis



dual Probabilistic Principal Component Analysis

The likelihood of **Y** can be then written as

$$P(\mathbf{Y}|\mathbf{W},\mathbf{X},\mathbf{S},\mu) = \prod_{j=1}^{D} N(\mathbf{y_j}|\mathbf{W}\mathbf{x}_j + \mathbf{V}\mathbf{s}_j + \mu, \sigma^2\mathbf{I})$$

Marginalizing $\mathbf{W}, \mathbf{V}, \mu, \epsilon$ we obtain the marginal likelihood

$$P(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^{D} N(\mathbf{y}_j|\mathbf{0}, \alpha_w \mathbf{X}\mathbf{X}^\top + \alpha_v \mathbf{S}\mathbf{S}^\top + \alpha_\mu + \sigma^2\mathbf{I})$$

dual Probabilistic Principal Component Analysis

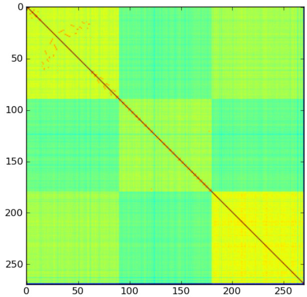
The likelihood of **Y** can be then written as

$$P(\mathbf{Y}|\mathbf{W},\mathbf{X},\mathbf{S},\mu) = \prod_{j=1}^{D} N(\mathbf{y_j}|\mathbf{W}\mathbf{x}_j + \mathbf{V}\mathbf{s}_j + \mu, \sigma^2\mathbf{I})$$

Marginalizing $\mathbf{W}, \mathbf{V}, \mu, \epsilon$ we obtain the marginal likelihood

$$P(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^{D} N(\mathbf{y}_j|\mathbf{0}, \alpha_w\mathbf{X}\mathbf{X}^\top + \alpha_v\mathbf{S}\mathbf{S}^\top + \alpha_\mu + \sigma^2\mathbf{I})$$

Population structure



Accounting for population structure

$$C = \alpha_w \mathbf{X}\mathbf{X}^\top + \alpha_v \mathbf{S}\mathbf{S}^\top + \alpha_\mu + \sigma^2 \mathbf{I}$$

$$C = \alpha_w \mathbf{X}\mathbf{X}^\top + \alpha_v \mathbf{S}\mathbf{S}^\top + \alpha_p \mathbf{P}\mathbf{P}^\top + \alpha_\mu + \sigma^2 \mathbf{I}$$

$$C = \alpha_w \mathbf{X}\mathbf{X}^\top + \alpha_v \mathbf{S}\mathbf{S}^\top + \alpha_p \mathbf{P}\mathbf{P}^\top + \alpha_g \mathbf{G}\mathbf{G}^\top + \alpha_\mu + \sigma^2 \mathbf{I}$$

Accounting for population structure

$$C = \alpha_w \mathbf{X}\mathbf{X}^\top + \alpha_v \mathbf{S}\mathbf{S}^\top + \alpha_\mu + \sigma^2 \mathbf{I}$$

$$C = \alpha_w \mathbf{X}\mathbf{X}^\top + \alpha_v \mathbf{S}\mathbf{S}^\top + \alpha_p \mathbf{P}\mathbf{P}^\top + \alpha_\mu + \sigma^2 \mathbf{I}$$

$$C = \alpha_w \mathbf{X}\mathbf{X}^\top + \alpha_v \mathbf{S}\mathbf{S}^\top + \alpha_p \mathbf{P}\mathbf{P}^\top + \alpha_g \mathbf{G}\mathbf{G}^\top + \alpha_\mu + \sigma^2 \mathbf{I}$$

Accounting for population structure

$$C = \alpha_w \mathbf{X}\mathbf{X}^\top + \alpha_v \mathbf{S}\mathbf{S}^\top + \alpha_\mu + \sigma^2 \mathbf{I}$$

$$C = \alpha_w \mathbf{X}\mathbf{X}^\top + \alpha_v \mathbf{S}\mathbf{S}^\top + \alpha_p \mathbf{P}\mathbf{P}^\top + \alpha_\mu + \sigma^2 \mathbf{I}$$

$$C = \alpha_w \mathbf{X}\mathbf{X}^\top + \alpha_v \mathbf{S}\mathbf{S}^\top + \alpha_p \mathbf{P}\mathbf{P}^\top + \alpha_g \mathbf{G}\mathbf{G}^\top + \alpha_\mu + \sigma^2 \mathbf{I}$$

Outline

1. eQTL mapping
2. Dataset
3. The model
4. Experiments
5. Conclusions

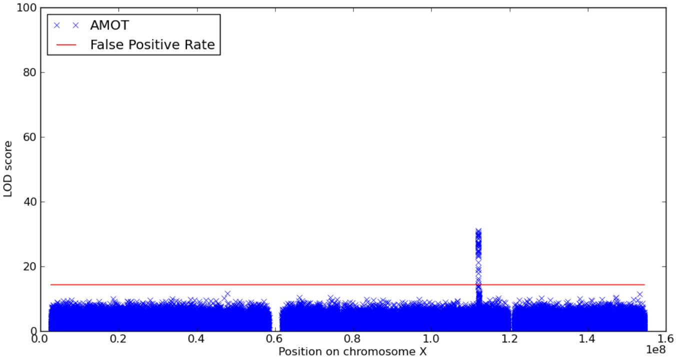
eQTL scan using data from Hapmap and GENEVAR

At each locus we compute the log-odds score:

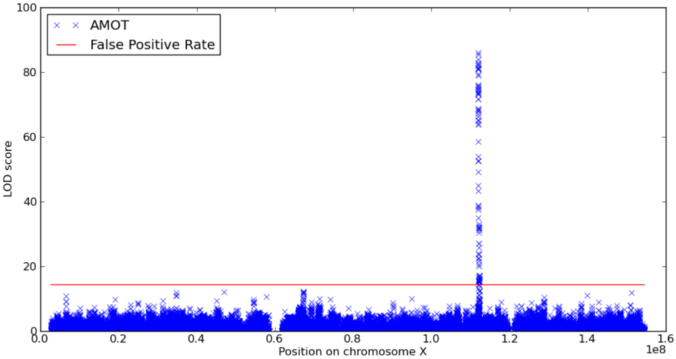
$$L_i = \log_{10} \left\{ \prod_n \frac{P(Y_m | s_{n,j}, \theta_{i,n})}{P(Y_m | \theta_{bkg})} \right\} \tag{1}$$

The significance of an association is evaluated via permutation testing.

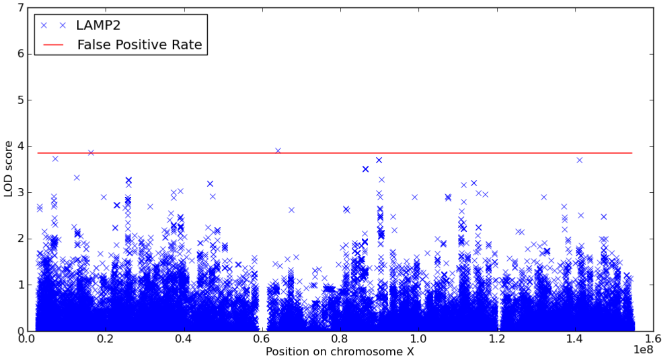
Traditional eQTL scan



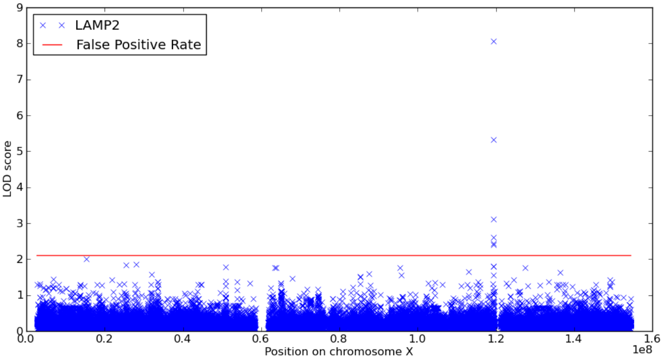
eQTL scan accounting for non-genetic factors



Traditional eQTL scan



eQTL scan accounting for non-genetic factors



Outline

1 eQTL mapping

2 Dataset

3 The model

4 Experiments

5 Conclusions

Conclusions

- We presented a model that explicitly accounts for non-genetic factors
- Using this model we can detect an higher number of significant associations
- Many extensions are possible (future work)