Explaining Confounding Factors in eQTL studies using a Dictionary of Latent Variables

Nicolò Fusi

Joint work with Oliver Stegle and Neil Lawrence

1. Department of Neuro- and Computer Science, University of Sheffield, UK
2. Department of Empirical Inference, Max Planck Institutes Tubingen, Germany
eQTL mapping

statistical technique with the goal to identify causal associations between variable genetic loci and the expression levels of individual genes.
eQTL mapping

Confounders introduce artifactual correlation in the expression levels of set of genes
The good news

Some of these confounders are known

- gender
- age
- ethnicity
The **bad** news

Most of them are completely **unknown** or **unmeasurable**

- optical effects
- laboratory conditions
- in humans, exposure to diesel fumes
- in humans, the stress of taking exams
PANAMA

Probabilistic ANalysis of MicroArray data

A non-parametric probabilistic model, that:

• can account for both known and unknown confounders
• is based on a linear additive model
• greatly improves the quality of the results
• fast
$Y$

\[
\begin{bmatrix}
\vdots \\
(N \times D)
\end{bmatrix}
\end{align*}
\]

Gene expression
Noise
Genotype
Latent Confounders
Gene expression
Gene expression
Latent Confounders
Genotype
Noise

\[ \mathbf{Y} = \mathbf{X} \mathbf{W} + \mathbf{S} \mathbf{V} + \epsilon \]

- \( \mathbf{Y} \): Gene expression (N x D)
- \( \mathbf{X} \): Latent Confounders (N x Q)
- \( \mathbf{W} \): (Q x D)
- \( \mathbf{S} \): Genotype (N x K)
- \( \mathbf{V} \): Noise (K x D)
- \( \epsilon \): Noise (N x D)
The likelihood is:

\[ P(Y \mid W, X, S) = \prod_{j=1}^{D} N(y_j \mid Wx_j + Vs_j, \sigma^2 I). \]

If we put spherical Gaussian priors over \( V \) and \( W \),

\[ P(W) = \prod_{i=1}^{D} N(w_i \mid 0, \alpha_w I) \]

\[ P(V) = \prod_{i=1}^{D} N(v_i \mid 0, \alpha_v I) \]
We can obtain the marginal likelihood

\[
P(Y \mid X) = \prod_{j=1}^{D} N(y_j \mid 0, C),
\]

Where

\[
C = \alpha_w XX^\top + \alpha_v SS^\top + \sigma^2 I.
\]

We can determine the parameters \( \theta \) and the latent variables \( X \) from the data by maximum likelihood

\[
\{\hat{\theta}, \hat{X}\} = \arg \max_{\theta, X} P(Y \mid X, \theta).
\]
PANAMA’s action is divided into two phases:

1. it learns a **dictionary** of latent variables that capture the main components of confounding variation

2. for each pair gene-SNPs it refits the weight parameters
Association testing

We determine the presence or the absence of an association by comparing two models:

\[
C = \alpha_w XX^\top + \alpha_v SS^\top + \sigma^2 I.
\]

The significance cutoff is determined by computing the positive False Discovery Rate (Storey, 2003):

\[
pFDR_{k,d} = \frac{\pi_0 \cdot p(y_d \mid H_0)}{\pi_0 \cdot p(y_d \mid H_0) + \pi_1 \cdot p(y_d \mid H_1)}.
\]
Experimental results
Simulated dataset

80 diploid individuals, 100 SNPs with a minor allele frequency of 0.4, 400 genes

<table>
<thead>
<tr>
<th></th>
<th>FDR 0.01</th>
<th></th>
<th>FDR 0.05</th>
<th></th>
<th>FDR 0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cis</td>
<td>trans</td>
<td>cis</td>
<td>trans</td>
<td>cis</td>
</tr>
<tr>
<td>linear</td>
<td>0.16</td>
<td>0.11</td>
<td>0.16</td>
<td>0.15</td>
<td>0.16</td>
</tr>
<tr>
<td>SVA</td>
<td>0.74</td>
<td>0.53</td>
<td>0.74</td>
<td>0.53</td>
<td>0.74</td>
</tr>
<tr>
<td>ICE</td>
<td>0.85</td>
<td>0.30</td>
<td>0.85</td>
<td>0.35</td>
<td>0.85</td>
</tr>
<tr>
<td>PANAMA</td>
<td>0.83</td>
<td>0.72</td>
<td>0.85</td>
<td>0.73</td>
<td>0.85</td>
</tr>
</tbody>
</table>
We want to allow an individual weighting of the latent variables.

\[ C = \alpha_w XX^\top + \alpha_v SS^\top + \sigma^2 I. \]

We constrain \( X \) to be orthonormal \((X^\top X = I)\) and modify the structure of the covariance

\[ C = XMX^\top + \alpha_v SS^\top + \sigma^2 I. \]

Where \( M \) is a matrix

\[
M = \begin{pmatrix}
\alpha_{w_1} & 0 & \cdots & 0 \\
0 & \alpha_{w_2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \alpha_{w_n}
\end{pmatrix}
\]
Experimental results
(again)
Simulated dataset

80 diploid individuals, 100 SNPs with a minor allele frequency of 0.4, 400 genes

<table>
<thead>
<tr>
<th></th>
<th>FDR 0.01</th>
<th></th>
<th>FDR 0.05</th>
<th></th>
<th>FDR 0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cis</td>
<td>trans</td>
<td>cis</td>
<td>trans</td>
<td>cis</td>
</tr>
<tr>
<td>linear</td>
<td>0.16</td>
<td>0.11</td>
<td>0.16</td>
<td>0.15</td>
<td>0.16</td>
</tr>
<tr>
<td>SVA</td>
<td>0.74</td>
<td>0.53</td>
<td>0.74</td>
<td>0.53</td>
<td>0.74</td>
</tr>
<tr>
<td>ICE</td>
<td>0.85</td>
<td>0.30</td>
<td>0.85</td>
<td>0.35</td>
<td>0.85</td>
</tr>
<tr>
<td>PANAMA</td>
<td>0.83</td>
<td>0.72</td>
<td>0.85</td>
<td>0.73</td>
<td>0.85</td>
</tr>
<tr>
<td>PANAMA-ARD</td>
<td><strong>0.87</strong></td>
<td><strong>0.80</strong></td>
<td><strong>0.89</strong></td>
<td><strong>0.81</strong></td>
<td><strong>0.89</strong></td>
</tr>
</tbody>
</table>
Yeast dataset

Smith and Kruglyak,

• genotypes and expression profiles of 108 yeast segregants

• grown in two environmental conditions: sugar and ethanol

• very strong environmental influence

• known, but not included in the model
Cis associations
Trans associations

![Graph showing number of genes against FDR for different methods: PANAMA-ARD, PANAMA, LINEAR, ICE, SVA.](image)
Overall associations
Validation

We include the environmental condition as a known covariate and measure the “overlap” between the calls made in the two settings.
Conclusions

• Confounding factors are a serious threat to the significance of eQTL studies and an accurate modeling is necessary

• the predominant assumption of a global set of confounders is clearly suboptimal

• PANAMA-ARD, through the individual reweighting of the inferred confounding factors is able to greatly improve the results of eQTL association studies

• Additional work is needed to improve the performance