Computational functional genomics
(Spring 2001: Lecture 6)

David K. Gifford
(Adapted from a lecture by Tommi S. Jaakkola)
MIT LCS and AI Lab

gifford@mit.edu
Data normalization

- The experimental setup [affymetrics slide]

- Different measurements have to be put on the same “scale” to permit comparison
Normalization

• We can exploit additional “spiked controls” to normalize the measurements from each experiment (chip)

• Measured spiked controls \((m\) raw measurements per experiment), \(n\) experiments

\[
\begin{bmatrix}
x^{(1)}_1 & \ldots & x^{(1)}_m \\
\vdots & \ddots & \vdots \\
x^{(n)}_1 & \ldots & x^{(n)}_m
\end{bmatrix}
\]

• We need to construct a model over these observations that disentangles the experiment dependent scaling and the underlying (fixed) control levels
Normalization cont’d

• Here is a particular model that attempts to disentangles the experiment dependent scaling and the underlying (fixed) control levels

\[
X_{t1} = c_1 \times r_t \times e_{t1} \\
\ldots \\
X_{tm} = c_m \times r_t \times e_{tm}
\]  

(1)

where

- \(X_{tj}\) is the random variable corresponding to the \(j^{th}\) control measurement in experiment \(t\)
- \(c_j\) is the fixed control amount
- \(r_t\) is the unknown experiment dependent scaling
- \(e_{tj}\) is random multiplicative noise

• Note: same \(r_t\) for all controls
Normalization cont’d

- Log-transform all the variables

\[
\begin{align*}
\log(Y_{t1}) &= \log(c_1) + \log(r_t) + \log(e_{t1}) \\
\ldots \\
\log(X_{tm}) &= \log(c_m) + \log(r_t) + \log(e_{tm}) \\
\end{align*}
\]

After the transform we can express the model in the simple form

**Observation** = **Model** + **noise**

\[
Y_{tj} = (\mu_j + l_t) + \epsilon_{tj}, \quad \epsilon_{tj} \sim N(0, \sigma_j^2)
\]

Additional assumptions?
Normalization cont’d

• Log-transform all the variables

\[
\begin{align*}
\log(Y_{t1}) &= \mu_1 + l_t + \epsilon_{t1} \\
\log(X_{t1}) &= \log(c_1) + \log(r_t) + \log(e_{t1}) \\
\ldots \\
\log(X_{tm}) &= \log(c_m) + \log(R_t) + \log(e_{tm})
\end{align*}
\]  

(4)

After the transform we can express the model in the simple form

Observation = Model + noise

\[
Y_{tj} = (\mu_j + l_t) + \epsilon_{tj}, \quad \epsilon_{tj} \sim N(0, \sigma_j^2)
\]

(5)

Additional assumptions?

1. measurement noise \( \epsilon_{tj} \) is independent across controls/experiments
2. the noise is Gaussian
3. the noise variance does not depend on the experiment
Normalization cont’d

- We want to fit our model $Y_{tj} \sim N(\mu_j + l_t, \sigma_j^2)$ to the (log transformed) raw data

\[
\begin{bmatrix}
y_{11} & \cdots & y_{1m} \\
\vdots & & \vdots \\
y_{n1} & \cdots & y_{nm}
\end{bmatrix}
\]  

(6)

- We need maximum likelihood estimates of $l_t$, $\mu_j$, and $\sigma_j^2$

\[
\text{Likelihood} = P(y_{11}|\mu_1 + l_1, \sigma_1^2) \cdots P(y_{n1}|\mu_1 + l_n, \sigma_1^2) \\
P(y_{1m}|\mu_m + l_1, \sigma_m^2) \cdots P(y_{nm}|\mu_m + l_n, \sigma_m^2)
\]  

(7)
Normalization cont’d

Likelihood function

\[
\mathcal{L} = \log \left( \prod_{i=1}^{M} \prod_{j=1}^{N} P(y_{ji}|\mu_i, l_j, \sigma_i^2) \right) 
\]  \hspace{1cm} (8)

\[
= \sum_{i=1}^{M} \sum_{j=1}^{N} -\frac{1}{2} \left( \log(2\pi\sigma_i^2) + \frac{(y_{ji} - \mu_i - l_j)^2}{\sigma_i^2} \right) \hspace{1cm} (9)
\]
Normalization cont’d

\[ Y_{tj} \sim N(\mu_j + l_t, \sigma^2_j) \]

\[
\begin{bmatrix}
y_{11} & \cdots & y_{1m} \\
\vdots \\
y_{n1} & \cdots & y_{nm}
\end{bmatrix}
\]  

(10)

• Iterative solution:

\[
\hat{\mu}_j \leftarrow \frac{1}{n} \sum_{t=1}^{n} (y_{tj} - l_i)
\]  

(11)

\[
\hat{\sigma}^2_j \leftarrow \frac{1}{n} \sum_{t=1}^{n} (y_{tj} - l_t - \hat{\mu}_j)^2
\]  

(12)

\[
\hat{l}_t \leftarrow \frac{1}{\left( \sum_{j=1}^{m} \hat{\sigma}^{-2}_j \right)} \sum_{j=1}^{m} \hat{\sigma}^{-2}_j (y_{tj} - \hat{\mu}_j)
\]  

(13)

• As a result, experiment \( t \) should be scaled down by a factor \( r_t = \exp(l_t) \)
MAP Normalization

- One spiked control can come to dominate the solution
- Use MAP regularization term over $\sigma_i^2$

$$P(\mu_j, l_t, \sigma_j^2|y_{tj}) \propto P(y_{tj}|\mu_j, l_t, \sigma_j^2) \cdot P(\sigma_j^2) \quad (14)$$

- Use a Wishart prior (conjugate to the Gaussian)

$$P(\sigma_j^2) = C(\alpha, k) \left( \frac{1}{\sigma_j^2} \right)^{\frac{\alpha-3}{2}} e^{-\frac{k}{2\sigma_j^2}} \quad (15)$$

- New iterative solution for $\sigma_j^2$:

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{N} (y_{tj} - \hat{\mu}_i - \hat{l}_t)^2 + k}{N + \alpha - 3} \quad (16)$$
Topics

• Normalization
  – MLE based estimation
  – MAP estimation

• **Statistical models**
  – linear regression
Statistical models

Observed data = Model + Noise

- linear regression
- additive models
- factor analysis
- graph models
- etc.
Statistical models: example

- Linear regression model

\[ X_2 = \left( X_1 \beta_1 + \beta_0 \right) + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \]  

(17)

where

\[ X_2 = \text{response variable} \]
\[ X_1 = \text{explanatory variable} \]
\[ \epsilon = \text{residual} \]
\[ \beta_1, \beta_0, \sigma^2 = \text{parameters} \]

in terms of \( n \) observations

\[
\begin{align*}
x_2^{(1)} &= (x_1^{(1)} \beta_1 + \beta_0) + \epsilon^{(1)} \\
&\vdots \\
x_2^{(n)} &= (x_1^{(n)} \beta_1 + \beta_0) + \epsilon^{(n)}
\end{align*}
\]
• The linear regression model

\[ X_2 = (X_1 \beta_1 + \beta_0) + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \quad (18) \]

can be written as a generative probability distribution over the response variable \( X_2 \)

\[ X_2 \sim N \left( (X_1 \beta_1 + \beta_0), \sigma^2 \right) \]

or

\[ P(x_2| (x_1 \beta_1 + \beta_0), \sigma^2) \]

• Are there any advantages from the probabilistic interpretation?
Statistical models: example cont’d

- The linear regression model

\[
X_2 = (X_1 \beta_1 + \beta_0) + \epsilon, \quad \epsilon \sim N(0, \sigma^2)
\]

(19)

can be written as a generative probability distribution over the response variable \(X_2\)

\[
X_2 \sim N((X_1 \beta_1 + \beta_0), \sigma^2)
\]

or

\[
P(x_2|(x_1 \beta_1 + \beta_0), \sigma^2)
\]

- Are there any advantages from the probabilistic interpretation?
  - we can fit the model using maximum likelihood criterion
Statistical models: example cont’d

- The linear regression model

\[ X_2 = \underbrace{(X_1 \beta_1 + \beta_0)}_{\text{Model}} + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \quad (20) \]

can be written as a generative probability distribution over the response variable \( X_2 \)

\[ X_2 \sim N \left( (X_1 \beta_1 + \beta_0), \sigma^2 \right) \]

or

\[ P(x_2 | (x_1 \beta_1 + \beta_0), \sigma^2) \]

- Are there any advantages from the probabilistic interpretation?
  - we can fit the model using **maximum likelihood criterion**
  - we can assess the **significance of the fit** relative to a simpler (or a more complex) model
Linear regression model

- Another interpretation of fitting (linear) regression models

We require that the residual or prediction error to have zero mean and be uncorrelated with the explanatory variable $X_1$:

zero mean:  \[
\sum_{t=1}^{n} \left[ x_2^{(t)} - (x_1^{(t)} \beta + \beta_0) \right] = 0
\]  \hspace{1cm} (21)

uncorrelated:  \[
\sum_{t=1}^{n} x_1^{(t)} \left[ x_2^{(t)} - (x_1^{(t)} \beta + \beta_0) \right] = 0
\]  \hspace{1cm} (22)