Introduction to Bayesian Statistics and an Application

Unconfounding the Confounded: Separating Treatment and Batch Effects in Confounded Microarray Experiments

Timothy M. Bahr

Department of Statistics
Brigham Young University

March 16, 2009

Who am I?

Tim Bahr, Undergrad...

Who am I?

Tim Bahr, Undergrad...

- 22, B.S. in Statistics, emphasis: Biostat

Who am I?

Tim Bahr, Undergrad...

- 22, B.S. in Statistics, emphasis: Biostat
- My first intro to Statistics in High School

Who am I?

Tim Bahr, Undergrad...

- 22, B.S. in Statistics, emphasis: Biostat
- My first intro to Statistics in High School
- Fascination with the Numerical Patterns in Science

Who am I?

Tim Bahr, Undergrad...

- 22, B.S. in Statistics, emphasis: Biostat
- My first intro to Statistics in High School
- Fascination with the Numerical Patterns in Science
- Future Goals

Who are you?

Bioinformatics

- Majors?
- Math/Stat Background?
- Microarrays?
- Research?
- Why Bioinformatics?
- Can I tell you what I think about Bioinformatics?

Definitions

Definitions

- **Bayesian Statistics $>>>$ statistical inferences on experimental data $+$ prior knowledge.**

Definitions

- Bayesian Statistics $>>>$ statistical inferences on experimental data $+$ prior knowledge.
- Classical (Frequentist) Statistics $>>>$ data from observations or experiments only.

Definitions

- **Bayesian Statistics $>>>$ statistical inferences on experimental data $+$ prior knowledge.**
- **Classical (Frequentist) Statistics $>>>$ data from observations or experiments only.**
- **Prior Distribution: The distribution we assume our parameters come from.**

Definitions

- Bayesian Statistics $>>>$ statistical inferences on experimental data $+$ prior knowledge.
- Classical (Frequentist) Statistics $>>>$ data from observations or experiments only.
- Prior Distribution: The distribution we assume our parameters come from.
- Gibbs Sampling (simplification): An algorithm that allows us to give interatively infer point estimates for "random" parameters.

Definitions

Definitions

- **Biostatistics: The application of statistics to a wide range of topics in biology.**

Definitions

- **Biostatistics: The application of statistics to a wide range of topics in biology.**
- **Gene Expression Microarray: A high-throughput technology in molecular biology used to detect gene expression levels in a cellular sample.**

Definitions

- **Biostatistics: The application of statistics to a wide range of topics in biology.**
- **Gene Expression Microarray: A high-throughput technology in molecular biology used to detect gene expression levels in a cellular sample.**
- **Confounded Experiment: when two or more variables vary together so that it is impossible to separate their unique effects.**

Bayesian Inference

Probabilistic inference that computes the distribution of the model parameters and gives prediction for previously unseen input values probabilistically.

Freqentist

Bayesian Inference

Probabilistic inference that computes the distribution of the model parameters and gives prediction for previously unseen input values probabilistically.

Freqentist

- $\boldsymbol{\theta}$, parameters, are fixed and unknown

Bayesian Inference

Probabilistic inference that computes the distribution of the model parameters and gives prediction for previously unseen input values probabilistically.

Freqentist

- $\boldsymbol{\theta}$, parameters, are fixed and unknown
- $\mathbf{X}$, random variables (data), are random

Bayesian

Bayesian Inference

Probabilistic inference that computes the distribution of the model parameters and gives prediction for previously unseen input values probabilistically.

Freqentist

- $\boldsymbol{\theta}$, parameters, are fixed and unknown
- $\mathbf{X}$, random variables (data), are random

Bayesian

- $\boldsymbol{\theta}$, parameters, are *random* and unknown

Bayesian Inference

Probabilistic inference that computes the distribution of the model parameters and gives prediction for previously unseen input values probabilistically.

Freqentist

- $\boldsymbol{\theta}$, parameters, are fixed and unknown
- $\mathbf{X}$, random variables (data), are random

Bayesian

- $\boldsymbol{\theta}$, parameters, are *random* and unknown
- $\mathbf{X}$, random variables (data), are random

"If you want to work on really interesting problems [Bayesian Inference] is where those problems lie"
-Don Rubin, Ph.D., Dept. Chair, Harvard Statistics

The idea of a prior

The idea of a prior

- Frequentists assume a parameter is fixed:
  - For example $X \sim N(\mu, \sigma^2)$
  - μ is a *fixed* unknown value

The idea of a prior

- Frequentists assume a parameter is fixed:
  - For example $X \sim N(\mu, \sigma^2)$
  - μ is a *fixed* unknown value
- What if μ is not fixed? What if it too can assume a distribution with variation

The idea of a prior

- Frequentists assume a parameter is fixed:
  - For example $X \sim N(\mu, \sigma^2)$
  - μ is a *fixed* unknown value
- What if μ is not fixed? What if it too can assume a distribution with variation
- We assume a *prior* on μ. i.e. $\mu \sim N(m_\mu, s_\mu^2)$

Bayesian Inference

Bayes' Theorem, based on basic theories of probability:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta} \tag{1}$$

Bayes' Theorem, based on basic theories of probability:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta} \tag{1}$$

- $\pi(\theta|x)$ is the *posterior distribution* of our parameters, θ.

Bayes' Theorem, based on basic theories of probability:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta} \tag{1}$$

- $\pi(\theta|x)$ is the *posterior distribution* of our parameters, θ.
- $f(x|\theta)$ is the *likelihood* of the data

Bayesian Inference

Bayes' Theorem, based on basic theories of probability:

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int f(\mathbf{x}|\theta)\pi(\theta)d\theta} \tag{1}$$

- $\pi(\theta|\mathbf{x})$ is the *posterior distribution* of our parameters, $\boldsymbol{\theta}$.
- $f(\mathbf{x}|\theta)$ is the *likelihood* of the data
- $\pi(\theta)$ is the *prior* distribution assumed on our parameters, $\boldsymbol{\theta}$.

In the End: Estimate Parameters

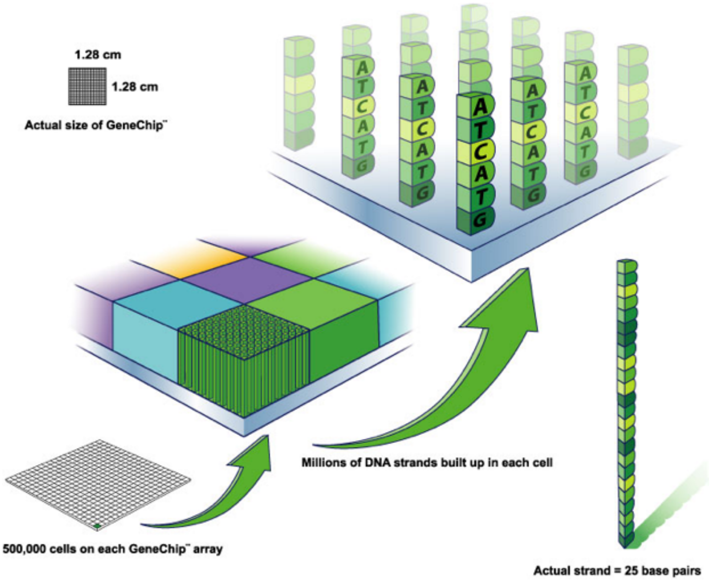
- We solve for the posterior of the parameters
- Use different methods to estimate an "optimum" value of our parameters.
  - Take the Expected Value of a Parameter
  - Gibbs Sampling
  - Metropolis-Hastings

What is a Microarray?

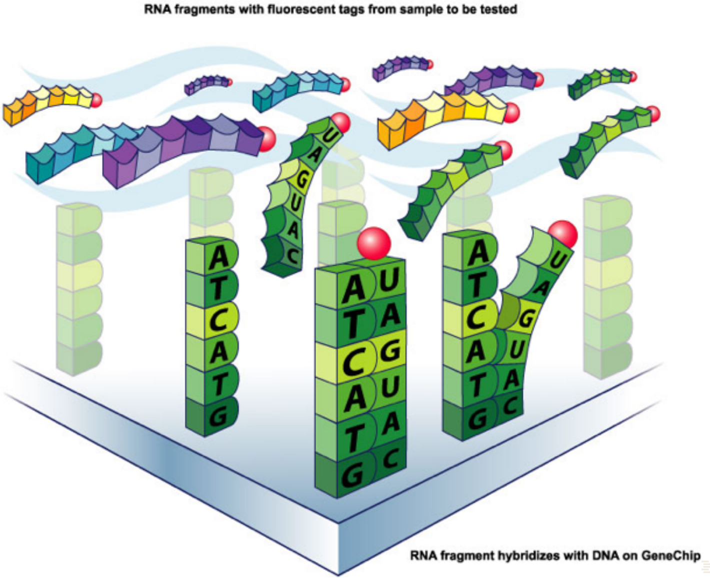
What is a Microarray?

- We use microarrays to detect gene expression levels for a given cellular sample.

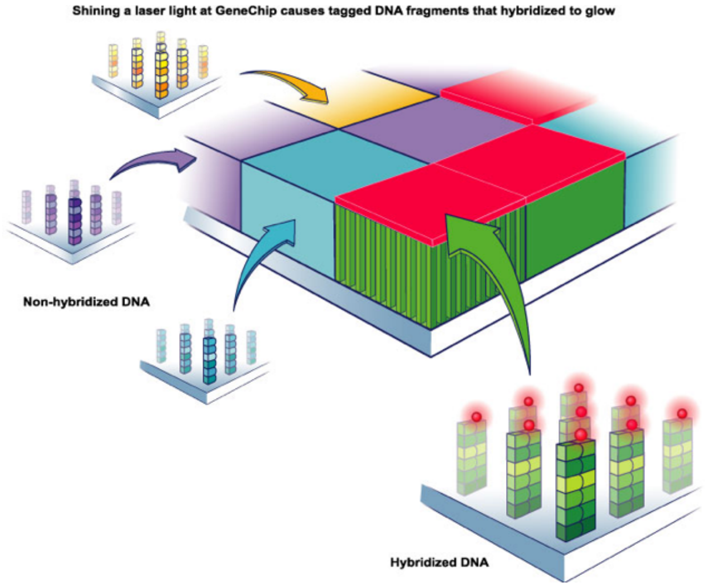
Microarrays



Microarrays



Microarrays



Confounded Experiments

- Consider a fertilizer experiment with corn:

First, An "unconfounded" experiment.

- 1 plot of corn; left half- control (no fertilizer), right half- treatment (Fertilizer)
- Differences in corn quality can be attributed to the treatment effect.

What is a Confounded Experiment?

- Consider a fertilizer experiment with corn:

What is a Confounded Experiment?

- Consider a fertilizer experiment with corn:

- Plot 1 (Batch 1)

What is a Confounded Experiment?

- Consider a fertilizer experiment with corn:

- Plot 1 (Batch 1)
- Control (no fertilizer)

What is a Confounded Experiment?

- Consider a fertilizer experiment with corn:

- Plot 1 (Batch 1)
- Control (no fertilizer)
- Plot 2 (Batch 2) - 1 mi. away

What is a Confounded Experiment?

- Consider a fertilizer experiment with corn:

- Plot 1 (Batch 1)
- Control (no fertilizer)

- Plot 2 (Batch 2) - 1 mi. away
- Treatment (New Fertilizer)

What is a Confounded Experiment?

- Consider a fertilizer experiment with corn:

If we observe a significant difference between the corn quality of the two plots (batches), can we attribute this difference to the fertilizer?

What is a Confounded Experiment?

- Consider a fertilizer experiment with corn:

No. The difference may be due to the treatment effect, the plot (batch effect), or a combination of the two.

What is a Confounded Experiment?

- Consider a fertilizer experiment with corn:

The Treatment Effect is *confounded* with the Plot or Batch Effect.

What is a Confounded Experiment?

- Consider a fertilizer experiment with corn:

The same principle applies to microarray experiments.

Confounded Experiments: Microarrays

Confounded Experiments: Microarrays

- Microarrays prepared at different times, in different places, by different people etc. ... are often confounded by batch effects.

Confounded Experiments: Microarrays

- Microarrays prepared at different times, in different places, by different people etc. ... are often confounded by batch effects.
- We are not interested in the the batch effect. We want to subtract it out.

Confounded Experiments: Microarrays

- Microarrays prepared at different times, in different places, by different people etc. ... are often confounded by batch effects.
- We are not interested in the the batch effect. We want to subtract it out.
- Our algorithm uses statistical methods to adjust for the Batch effect in confounded microarray experiments.

Why?

Why?

- Often times biologists can save money by using data that was obtained in previous experiments.

Why?

- Often times biologists can save money by using data that was obtained in previous experiments.
- Inter-lab collaboration becomes much more reliable when batch effects are accounted for.

Our Solution

Our method allows precise estimation of the batch effect and the treatment effect.

Our Solution

Our method allows precise estimation of the batch effect and the treatment effect.

- A dynamic linear model

Our Solution

Our method allows precise estimation of the batch effect and the treatment effect.

- A dynamic linear model
- Novel yet Appropriate Assumptions

Our Solution

Our method allows precise estimation of the batch effect and the treatment effect.

- A dynamic linear model
- Novel yet Appropriate Assumptions
- Bayesian Statistical Methods

The Model:

$$y_{ig} = \mu_g + X_i\alpha_g + Z_i\tau_g + \varepsilon_{ig} \tag{2}$$

Our Model

The Model:

$$y_{ig} = \mu_g + X_i\alpha_g + Z_i\tau_g + \varepsilon_{ig} \tag{2}$$

- y_{ig} - the "expression level" for a sample i from gene g

The Model:

$$y_{ig} = \mu_g + X_i \alpha_g + Z_i \tau_g + \varepsilon_{ig} \tag{2}$$

- y_{ig} - the "expression level" for a sample i from gene g
- μ_g - an overall average for gene g

The Model:

$$y_{ig} = \mu_g + X_i\alpha_g + Z_i\tau_g + \varepsilon_{ig} \qquad (2)$$

- y_{ig} - the "expression level" for a sample i from gene g
- μ_g - an overall average for gene g
- α_g - the Treatment Effect for gene g

Our Model

The Model:

$$y_{ig} = \mu_g + X_i\alpha_g + Z_i\tau_g + \varepsilon_{ig} \tag{2}$$

- y_{ig} - the "expression level" for a sample i from gene g
- μ_g - an overall average for gene g
- α_g - the Treatment Effect for gene g
- τ_g - the Batch Effect for gene g

Our Model

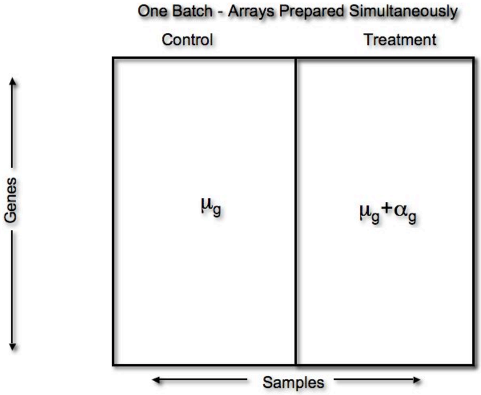
The Model:

$$y_{ig} = \mu_g + X_i\alpha_g + Z_i\tau_g + \varepsilon_{ig} \tag{2}$$

- y_{ig} - the "expression level" for a sample i from gene g
- μ_g - an overall average for gene g
- α_g - the Treatment Effect for gene g
- τ_g - the Batch Effect for gene g
- ε_{ig} - error for sample i from gene g

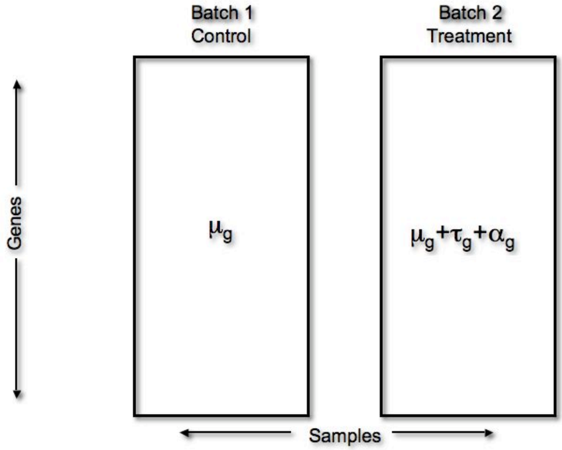
Solution Formulation and Assumptions

First, an "unconfounded" formulation.



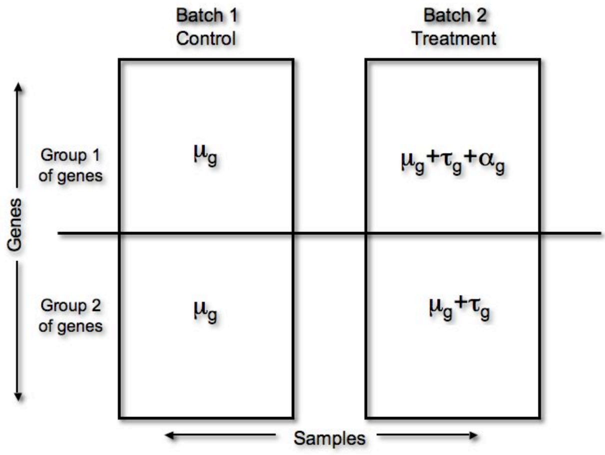
Difference between treatment and control can be attributed to "treatment effect."

Solution Formulation and Assumptions



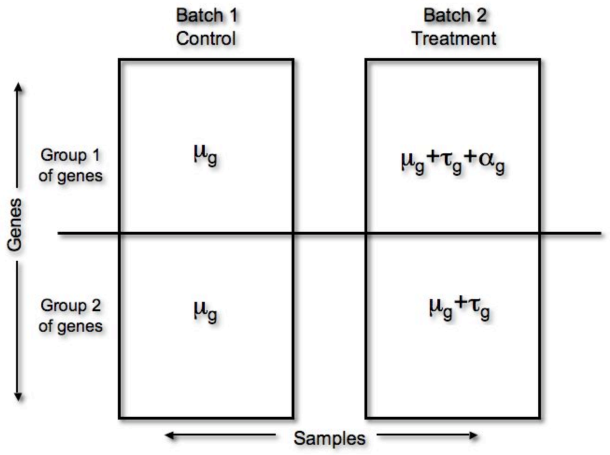
We can't differentiate the values of α_g and τ_g.

Solution Formulation and Assumptions



We assume treatment, α_g, has no effect on group 2 genes

Solution Formulation and Assumptions

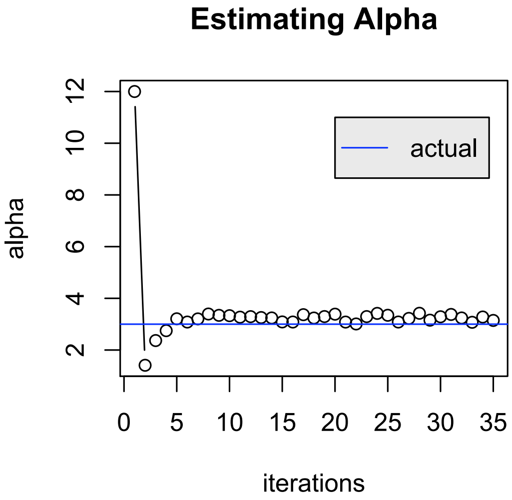


Determine which genes in each group $>>>$ estimate α_g and τ_g.

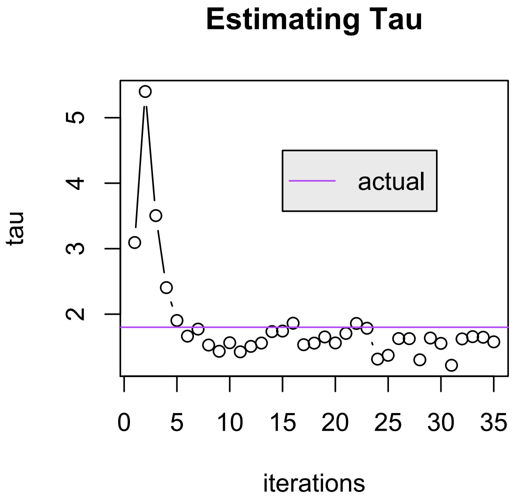
Gibbs Sampling

- A *Bayesian Method*
- Gives us the power to estimate which genes are in each group
- *Iteratively* estimates values until sequence converges

ESTIMATING α_g



Estimating τ_g



Possible Applications

- Microarrays in Cancer Research
- Clinical use of microarrays for diagnosis
- Possible applications in non-array experiments

Acknowledgments

- W. Evan Johnson, mentor
- Nathaniel Gustafson, programmer
- BYU Dept. of Statistics
- Johnson Lab

(old picture)