High Dimensional Data, Covariance Matrices and Application to Genetics

Samprit Banerjee, Ph.D
Div. of Biostatistics
Department of Public Health
Weill Medical College of Cornell University

UW-M 22-Apr-2010

Motivation

- High Dimensional Data
- Examples

Theoretical Underpinnings

- Random Matrices
- Shrinkage Estimation
- Decision Theory
- Bayesian Estimation

QTL Mapping

- Background
- Statistical Challenges
- Bayesian Solution
- Bayesian Multiple Traits

Data Deluge

"The coming century is surely the century of data"

David Donoho, 2000

"..industrial revolution of data."

The Economist, 2010

Sources of high dimensional data

- Genetics and Genomics
- Internet portals: e.g Netflix
- Financial data

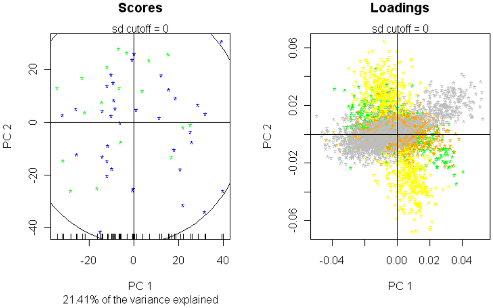
High Dimensional Data

In statistics,

- Observations: instances of a particular phenomenon
  - Example of instances ↔ human beings
  - Typically, n denotes the number of observations.
- Variable or Random variable is vector of values these observations are measured on
  - Example: blood pressure, weight, height.
  - Typically, p denotes the number of variables.
- Recent trend: explosive growth of p, ↔ dimensionality.
- $p \gg n$ classical methods of statistics fail!

Example 1: Principal Component Analysis

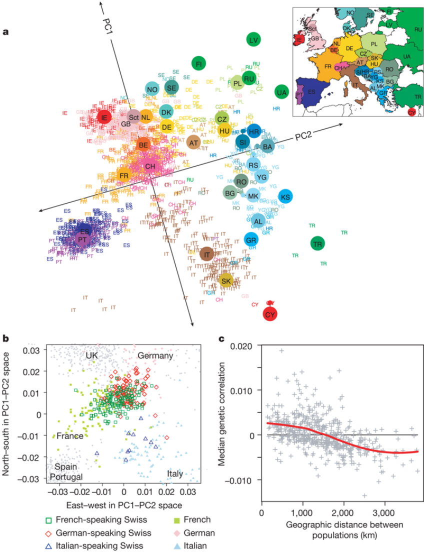
Let $\mathbf{X}_{n \times p} = [X_1 : X_2 : \cdots : X_p]$ be *i.i.d* variables.
Goal: reduce dimensionality by constructing a smaller number of “derived” variables.



$$w_1 = \arg \max_{||w||=1} var(W'X)$$

Spectral decomposition: $X'X = WLW'$, where
$L = diag\{\ell_1, ..., \ell_p\}$ are the eigenvalues.

Population Structure within Europe



[1]

[1] J Novembre et al. Nature 000, 1-4 (2008) doi:10.1038/nature07331

Example 2: Multivariate Regression

One of the most common use of the covariance matrix in statistics is during a multivariate regression.

$$\mathbf{Y}_{n\times p} = \mathbf{X}_{n\times q}\beta_{q\times p} + \mathbf{E}_{n\times p}$$

where $e_i \sim \mathcal{N}_p(0,\Sigma), i = 1, \cdots, n$ and Σ is $p \times p$.

- Historically $p < n$; High Dimensional data $p >> n$ or $q >> n$
- Estimates can be obtained by maximizing the likelihood

$$L(\beta,\Sigma|X,Y) \propto \prod_{i=1}^{n} |\Sigma|^{-1/2} exp \left\{ -\frac{1}{2}(Y_i - X_i\beta)'\Sigma^{-1}(Y_i - X_i\beta) \right\}$$

Seemingly Unrelated Regression

Zellner, 1962 introduced the Seemingly Unrelated Regression model.

$$\mathbf{Y}^*{}_{np\times 1} = \mathbf{X}^*{}_{np\times pq}\beta^*_{pq\times 1} + e^*_{np\times 1}$$

where $\mathbf{Y} = vec(\mathbf{Y})$, $\mathbf{X}^* = diag\{X_1, \cdots, X_p\}$, $\beta^* = vec(\beta)$,$e^* = vec(\mathbf{E})$ and $vec()$ is the vectorizing operator.

- $e^* \sim N(0, \Sigma \otimes I_n)$
- GLS estimates: $\hat{\beta} = (X^{*'}\Omega^{-1}X^*)^{-1}(X^{*'}\Omega^{-1}Y)$
- where $\Omega = \Sigma \otimes I_n$ and $\otimes$ is the Kronecker product.

Random Matrix Theory

- Covariance matrix $\Sigma_{p \times p}$ is a random matrix
- Eigenvalues of Σ, $\{\lambda_1, \cdots, \lambda_p\}$ are random
- Properties of interest: joint distribution of eigenvalues, number of eigenvalues falling below a given value
- Beginning in 1950s, physicists began to use random matrices to study energy levels of a system in quantum mechanics.
- Wigner proposed a statistical description of an "ensemble" of energy levels - properties empirical distribution and distribution of spacings.

Covariance Matrices

In statistics: $X_1, \cdots, X_n \sim N_p(0, \Sigma)$ and $X_{n \times p} = [X_1, \cdots, X_n]'$ The usual estimator is

Sample Covariance Matrix

$$S = X'X/n$$

Bayesian Estimation

$$\begin{aligned} \pi(\Sigma|X) &\propto p(X|\Sigma)\pi(\Sigma) \\ \hat{\Sigma} &= E_{\pi(.|X)}(\Sigma) \end{aligned}$$

Gaussian and Wishart Distributions

If $X_1, X_2, \cdots, X_n$ are n $i.i.d$ samples from a p-variate or p-dimensional Gaussian distribution $N_p(0, \Sigma)$ with density.

$$f(X) = |\sqrt{2\pi}\Sigma|^{-1/2} exp\left\{-\frac{1}{2}X'\Sigma^{-1}X\right\}$$

$S = X'X$ follows a Wishart distribution (named after John Wishart, 1928)

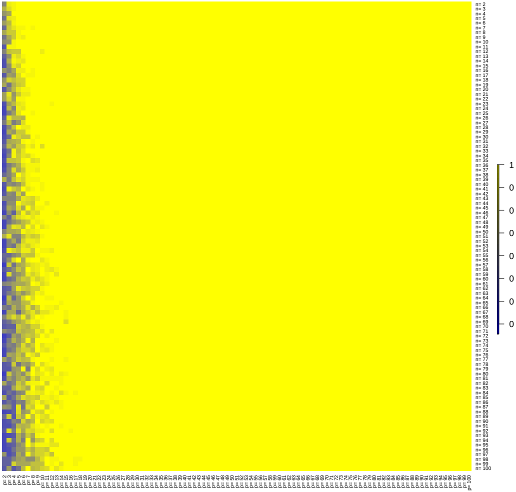
$$f(S) = c_{n,p}|\Sigma|^{-n/2}|S|^{(n-p-1)/2} etr\left\{-\frac{1}{2}\Sigma^{-1}S\right\}$$

where $etr() = exp(tr())$

Eigenstructure of sample covariance matrix

It is well known that the eigenvalues of the sample covariance matrix are more spread out compared to the true eigenvalues of the population covariance matrix

Spread of Sample Eigenvalues



- Counting the number of times the sample eigenvalues are spread.
- $\ell_1 < \lambda_1 | \ell_p > \lambda_p$
- $\ell_1 > \ell_2 > \cdots > \ell_p$ are the eigenvalues of the sample covariance matrix S
- $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ are the eigenvalues of the population covariance matrix Σ

Joint Distribution of Eigen Values

Fisher (Cambridge), Girshik (Columbia), Hsu (London), Mood (Princeton) and Roy (Calcutta)

Theorem

If S is $W_p(n, \Sigma)$ with $n \geq p$ the joint density function of the eigenvalues $\ell_1, \ell_2, \cdots, \ell_p$ of S is

$$\propto \prod_{j=1}^{p} \ell_j^{(n-p-1)/2} \prod_{j<k} (\ell_j - \ell_k) \times \int_{\mathbb{O}(p)} etr\left\{ -\frac{1}{2} \Sigma^{-1} H L H' \right\} dH$$

where $\mathbb{O}_p$ is the orthogonal group of $p \times p$ matrices, dH is the normalized Haar measure and L is the diagonal matrix $diag(\ell_1, \ell_2, \cdots, \ell_p)$. Assume $\ell_1 > \ell_2 > \cdots > \ell_p$.

The integral over $\mathbb{O}_p$ can be expanded by zonal polynomials. If

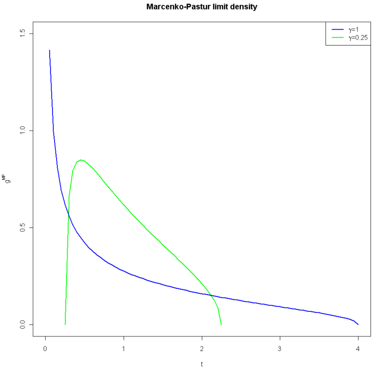
$\Sigma = I$ then the joint density simplifies

$$\propto \prod_{j=1}^{p} \ell_j^{(n-p-1)/2} \prod_{j<k} (\ell_j - \ell_k) exp\left(-\frac{1}{2} \sum_j \ell_j \right)$$

Eigenspectrum

- **Empirical Spectrum**: how many eigenvalues fall below a given value.
- **Wigner's Semi-Circle Law**: Wigner showed the limiting density of the "*empirical spectrum*" of real symmetric matrices A with $i.i.d$ entries is a semi-circle
- **Marčenko-Pastur** gave the limiting density of the "*empirical spectrum*" of the sample eigenvalues for a special case $A \sim W_p(n, I)$

Eigenspectrum



Study of eigenvalue distributions can be distinguished into

- **Bulk**: Refers to the properties of the full set $\ell_1, \ell_2, \cdots, \ell_p$
- **Extremes**: Addresses the (first few) largest and smallest eigenvalues

Largest Eigenvalue

Theorem (Johnstone, 2001)

Let $\ell_1 >, \cdots, > \ell_p$ denote the eigenvalues of the sample covariance matrix $X'X \sim W_p(n, I)$. Then

$$\frac{\ell_1 - \mu_{np}}{\sigma_{np}} \mathfrak{D}{\rightarrow} W_1 \sim F_1$$

where

$$\mu_{np} = (\sqrt{n-1} + \sqrt{p})^2$$
$$\sigma_{np} = (\sqrt{n-1} + \sqrt{p}) \left(\frac{1}{\sqrt{n-1}} + \frac{1}{\sqrt{p}} \right)^{1/3}$$

F_1 is the Tracy-Widom law of order 1 and has the distribution function defined by

$$F_1(s) = exp \left\{ -\frac{1}{2} \int_s^{\infty} q(x) + (x-s)q^2(x)dx \right\}, \qquad s \epsilon \mathbb{R}$$

where q solves the (nonlinear) Painlevé II differential equation

$$q(x) = xq(x) + 2q^3(x),$$
$$q(x) \sim Ai(x) \quad as \quad x \rightarrow +\infty$$

where $Ai(x)$ denotes the Airy function.

Lessons learned

- The Vandermonde determinant $\prod_{j>k}(\ell_j - \ell_k)$ of the joint eigenvalue induces repulsion
- The eigenstructure of the sample covariance is more spread out compared to that of the population covariance matrix
- This is less pronounced when p is small
- Both Bulk and Extreme distribution of eigenvalues are complicated for computation.

Outline

Motivation
High Dimensional Data
Examples

Theoretical Underpinnings
Random Matrices
Shrinkage Estimation
Decision Theory
Bayesian Estimation

QTL Mapping
Background
Statistical Challenges
Bayesian Solution
Bayesian Multiple Traits

Stein's Estimator

The sample covariance matrix S can be decomposed into VLV', where V is an orthogonal matrix and $L = diag(\ell_1, \cdots, \ell_p)$ with $\ell_1 \geq \ell_2 \geq \cdots \geq \ell_p$. Stein (1975) considered the orthogonal invariant estimator:

$$\hat{\Sigma} = V\Phi(L)V'$$

where $\Phi(L) = diag(\phi_1, \cdots, \phi_p)$ with $\phi_i = \ell_i/\alpha_i$,

$$\alpha_i = (n - p + 1) + 2\ell_i \sum_{j \neq i} \frac{1}{\ell_i - \ell_j}$$

Outline

Motivation

High Dimensional Data

Examples

Theoretical Underpinnings

Random Matrices

Shrinkage Estimation

Decision Theory

Bayesian Estimation

QTL Mapping

Background

Statistical Challenges

Bayesian Solution

Bayesian Multiple Traits

Stein's Estimator contd...

Issues with Stein's estimator:

- The intuitive ordering of $\phi_1 \geq \phi_2 \geq \cdots \phi_p$ is frequently violated.
- Sometimes ϕ_i can be negative
  - Stein suggested an isotonizing algorithm to avoid this problem by pooling adjacent pairs.

Haff (1991) formally minimized the Bayes risk for an orthogonally invariant prior by a variational technique.

Decision Theoretic Tools

Definition (Decision Theory)

Decision theory in philosophy, mathematics and statistics is concerned with identifying the values, uncertainties and other issues relevant in a given decision, its rationality, and the resulting optimal decision. It is very closely related to the field of game theory. (source: Wikipedia)

Definition (Loss function)

A loss function is any function L from $\Theta \times \mathcal{D}$ in $[0, +\infty)$

We will consider the following Loss functions for Σ

- **Stein's Loss**: $L_1(\Sigma, \hat{\Sigma}) = tr(\hat{\Sigma}\Sigma^{-1}) - log|\hat{\Sigma}\Sigma^{-1}| - p.$
- **Quadratic Loss**: $L_2(\Sigma, \hat{\Sigma}) = tr(\hat{\Sigma}\Sigma^{-1} - I)^2$

Decision Theoretic Tools contd...

Frequentist Risk

$$R(\theta, \delta) = \int_{\mathcal{X}} L(\theta, \delta(x)) f(x|\theta) dx$$

Bayesian Paradigm

- Posterior Expected Loss

$$\rho(\pi, d|x) = \int_{\Theta} L(\theta, \delta(x)) \pi(\theta|x) d\theta$$

- Integrated Risk

$$r(\pi, \delta) = \int_{\Theta} \int_{\mathcal{X}} L(\theta, \delta(x)) f(x|\theta) dx \pi(\theta) d\theta$$

- Bayes estimator δ^{π} is that which minimized $r(\pi, \delta)$ and the corresponding risk is the Bayes risk.

... "To average over all possible values of x, when we know the observed value of x, seems to be a waste of information"

... "Such an evaluation may be satisfactory for the statistician, but it is not so appealing for a client, who wants optimal results for her data x, not that of another's"

Christian Robert, 2007 (The Bayesian Choice)

Bayesian Paradigm

$$\pi(\Sigma|X) \propto p(X|\Sigma)\pi(\Sigma)$$

- Posterior mean, maximum *a posteriori*
- Decision theoretic approach
- Bayes estimator: minimize the integrated risk based on a certain prior and loss function

Jeffreys Prior

Jeffreys' invariant principle: Sir Harold Jeffreys (1961) suggested any non-informative prior distribution should be justified on the grounds of its invariance under parameter transformation. So, if $\theta \sim \pi$ a priori, for any one-to-one transformation $\phi = \phi(\theta)$ the prior on ϕ should be $\pi(\phi)$.

$$\pi(\theta) \propto \mathcal{I}(\theta)^{1/2} \textit{ where } \mathcal{I}(\theta) = E_{x|\theta}\left(-\frac{\partial^2 L}{\partial\theta^2}\right)$$

This is easy to see since $\mathcal{I}(\phi) = \mathcal{I}(\theta)(d\theta/d\phi)^2$

- Jeffreys prior for the covariance matrix is

$$\pi(\Sigma) \propto |\Sigma|^{-(p+1)/2}$$

- Under Stein's loss (L_1), the Bayes estimator for the covariance matrix is the usual unbiased estimator, the sample covariance matrix S/n

Reference Prior

Reference Prior Principle: (Bernardo, 1992) Let x be the result of an experiment $\epsilon = \{\mathcal{X}, \Theta, p(x|\theta)\}$ and let C be the class of admissible priors. The reference posterior of θ after x has been observed is defined to be $\pi(\theta|x) = lim\ \pi_k(\theta|x)$ where $\pi_k(\theta|x) \propto p(x|\theta)\pi_k(\theta)$ is the posterior density corresponding to the prior $\pi_k(\theta)$ which maximizes $\mathcal{I}^{\theta}\{\epsilon(k), p(\theta)\} = \int p(x) \int p(\theta|x) log \frac{p(\theta|x)}{p(\theta)} d\theta dx$ the expected information (expected Kullback-Leibler divergence of the posterior with respect to the prior) about θ.

The Reference prior was derived by Yang and Berger (1995). Let $\Sigma = O\Lambda O'$ where O is an orthogonal matrix and Λ is a diagonal matrix. The reference prior formulation is as follows

$$
\begin{aligned}
\pi(\Lambda, O)(d\Lambda)(dO) &\propto \frac{1}{|\Lambda|}(d\Lambda)(dH) \\
&\propto \frac{1}{|\Sigma| \prod_{i<j}(\lambda_i - \lambda_j)}(d\Sigma)
\end{aligned}
$$

where (dH) is the conditional invariant Haar measure over the space of orthogonal matrices.

Sampling from the Reference Posterior

The posterior resulting from the reference prior is

$$\pi_R(\Sigma|S)(d\Sigma) \propto \frac{etr(-\frac{1}{2}\Sigma^{-1}S)}{|\Sigma|^{n/2+1}\prod_{i<j}(\lambda_i - \lambda_j)}(d\Sigma)$$

A Metropolis-Hastings Sampler:

- Generate $\Sigma^{new} \sim W_p(n, S)$
- Accept Σ^{new} with probability
- $\alpha = min\left\{1, \frac{|\Sigma^{old}|^{(p+1)/2}\prod_{i<j}(\lambda_i^{old}-\lambda_j^{old})}{|\Sigma^{new}|^{(p+1)/2}\prod_{i<j}(\lambda_i^{new}-\lambda_j^{new})}\right\}$

Reference and Jeffreys comparison

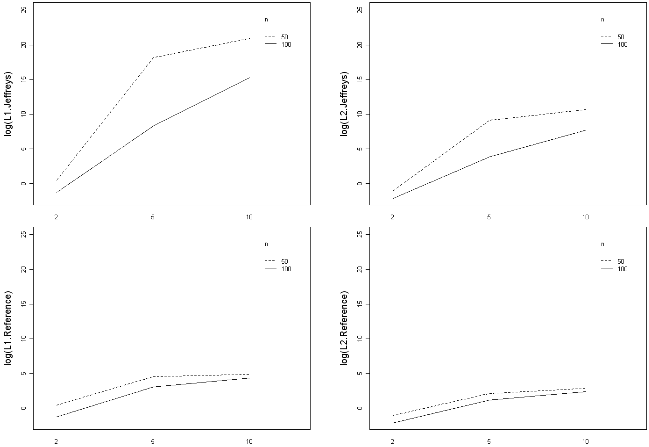
Simulation

- n=50,100
- p=2,5,10
- correlation structure: correlated and independent
- 50 replicated

Frequentist Risks of the posterior mean are approximated by average Loss under the following Loss functions.

- **Stein's Loss**: $L_1(\Sigma, \hat{\Sigma}) = tr(\hat{\Sigma}\Sigma^{-1}) - log|\hat{\Sigma}\Sigma^{-1}| - p$
- **Quadratic Loss**: $L_2(\Sigma, \hat{\Sigma}) = tr(\hat{\Sigma}\Sigma^{-1} - I)^2$

Reference and Jeffreys comparison



What is QTL Mapping?

Quantitative Trait Loci (QTL) Mapping

What is QTL Mapping?

Quantitative Trait Loci (QTL) Mapping

QT
y_1
y_2
y_3
y_4
y_5
y_6
y_7
y_8
y_9
y_{10}

- Quantitative Traits *e.g.* Blood pressure, BMI, FatMass, complex diseases (Alzhiemers) etc.

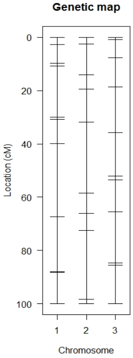
What is QTL Mapping?

Quantitative Trait Loci (QTL) Mapping

QT

y_1
y_2
y_3
y_4
y_5
y_6
y_7
y_8
y_9
y_{10}

L



- Loci → Genomic positions influencing the traits

What is QTL Mapping?

Quantitative Trait Loci (QTL) Mapping

QT	L	Mapping

y_1

y_2

y_3

y_4

y_5

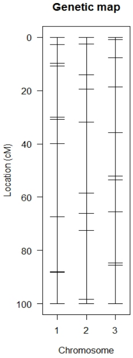
y_6

y_7

y_8

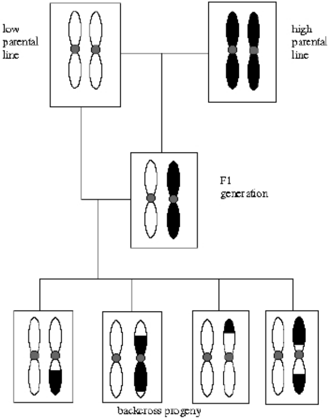
y_9

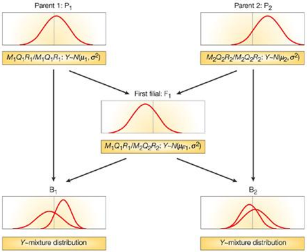
y_{10}



- Information from Quantitative traits combined with genetic information
- Try to map the positions of the genome influencing the traits

Genetic Design (Backcross Experiment)




Parent 1: P_1
Parent 2: P_2
$M_1Q_1R_1/M_1Q_1R_1$: $Y \sim N(\mu_1, \sigma^2)$
$M_2Q_2R_2/M_2Q_2R_2$: $Y \sim N(\mu_2, \sigma^2)$
First filial: F_1
$M_1Q_1R_1/M_2Q_2R_2$: $Y \sim N(\mu_{F1}, \sigma^2)$
B_1
B_2
Y~mixture distribution
Y~mixture distribution

Nature Reviews | Genetics

- Broman, 1997

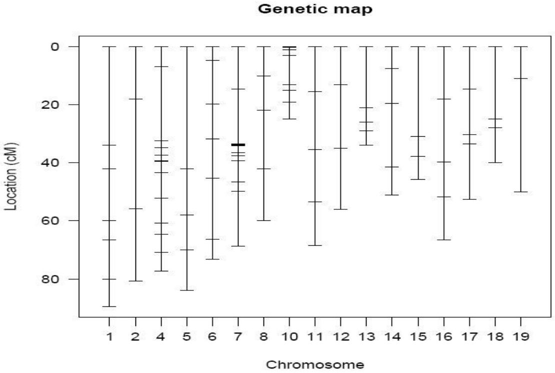
- Controlled experiments → not possible with humans
- Example of traits: BMI, fatmass, Obesity related traits etc.
- Big impact on public health

Importance of QTL Mapping

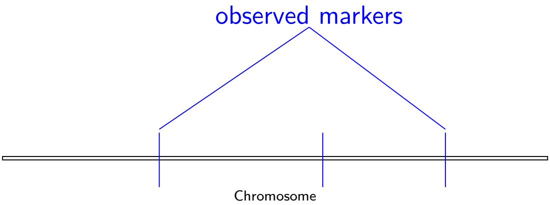
- Identifying QTL in experimental animals is critical for the understanding biochemical pathways underlying complex traits.
- These understanding translate to drug targets and eventual clinical trials.
- QTL mapping is also important for animal/plant breeding.

Data

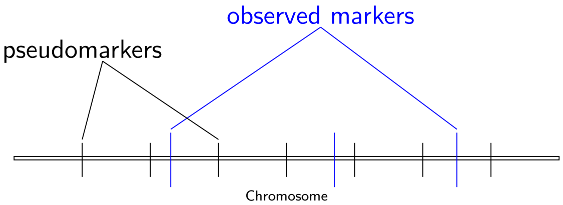
y_1	C1M1	C1M2	C2M1	C2M2	C15M2	C16M1	C19M1
8.8	AA	AA	AB	AA	AA	AB	AB
9.6	AA	AA	AB	AB	AB	AB	AB
10.6	AB	AB	AA	AA	AB	AA	AA
11.1	AB	AB	AA	AB	AB	AA	AA



Interval Mapping

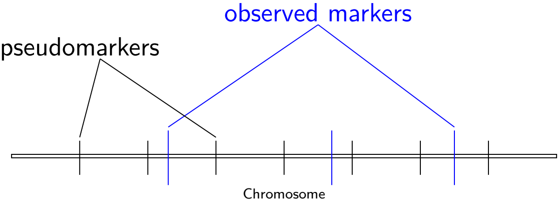


Interval Mapping



- Insert arbitrary positions (pseudomarkers) into marker intervals

Interval Mapping



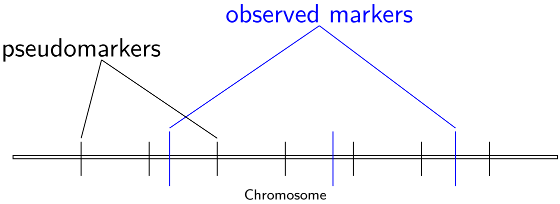
observed markers

pseudomarkers

Chromosome

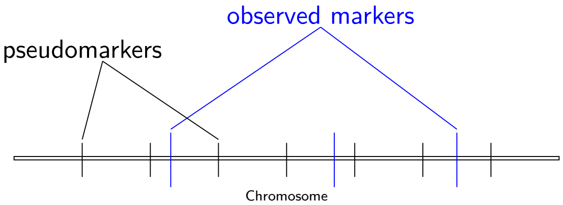
- Insert arbitrary positions (pseudomarkers) into marker intervals
- Enables us to detect QTL within marker intervals

Interval Mapping



- Insert arbitrary positions (pseudomarkers) into marker intervals
- Enables us to detect QTL within marker intervals
- Pseudomarkers and markers are considered as putative QTL

Interval Mapping



- Insert arbitrary positions (pseudomarkers) into marker intervals
- Enables us to detect QTL within marker intervals
- Pseudomarkers and markers are considered as putative QTL
- Pseudomarkers not observed – Hidden Markov Model to reconstruct genotypes

Challenges in QTL Mapping

Complex Traits

Challenges in QTL Mapping

Complex Traits

- interacting network of multiple genes and environmental factors

Challenges in QTL Mapping

Complex Traits

- interacting network of multiple genes and environmental factors
- small-to-moderate sized effects

Challenges in QTL Mapping

Complex Traits

- interacting network of multiple genes and environmental factors
- small-to-moderate sized effects
- high sample size required

Challenges in QTL Mapping

Complex Traits

- interacting network of multiple genes and environmental factors
- small-to-moderate sized effects
- high sample size required

Question

What combination of genes and interactions should one consider?

Challenges in QTL Mapping

Complex Traits

- interacting network of multiple genes and environmental factors
- small-to-moderate sized effects
- high sample size required

Question

What combination of genes and interactions should one consider?

Model Selection

Challenges in QTL Mapping

Complex Traits

- interacting network of multiple genes and environmental factors
- small-to-moderate sized effects
- high sample size required

Question

What combination of genes and interactions should one consider?

Model Selection

- For a BC (backcross) population with 40 genetic markers

Challenges in QTL Mapping

Complex Traits

- interacting network of multiple genes and environmental factors
- small-to-moderate sized effects
- high sample size required

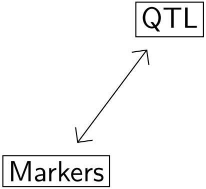
Question

What combination of genes and interactions should one consider?

Model Selection

- For a BC (backcross) population with 40 genetic markers
- $2^{40} = 10^{12} = 1,000,000,000,000$ models

Statistical structure



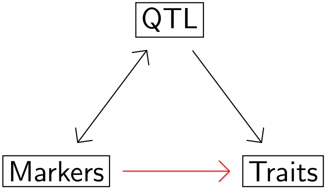
QTL

Markers

Two aspects of the QTL mapping problem

1. The missing data problem: Markers ↔ QTL

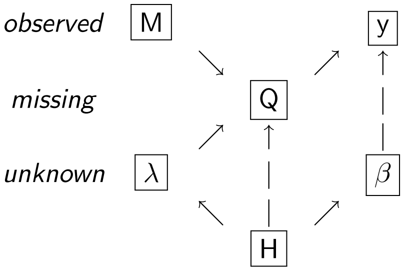
Statistical structure



Two aspects of the QTL mapping problem

1. The missing data problem: Markers ↔ QTL
2. The model selection problem: QTL → Traits

Bayesian Interval Mapping Framework



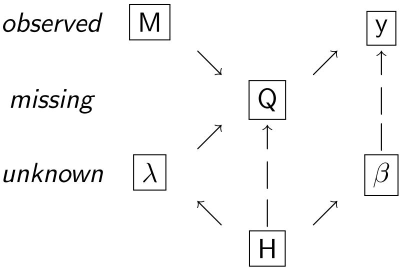
observed M y

missing Q

unknown λ β

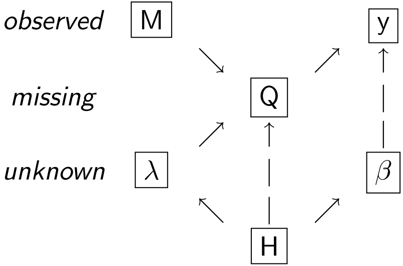
H

Bayesian Interval Mapping Framework



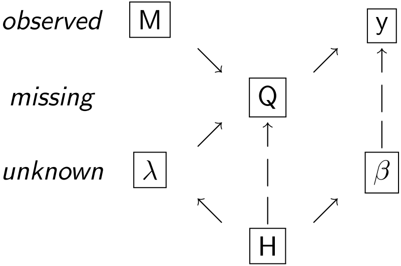
- Observed: y (traits) and M (marker and linkage map)

Bayesian Interval Mapping Framework



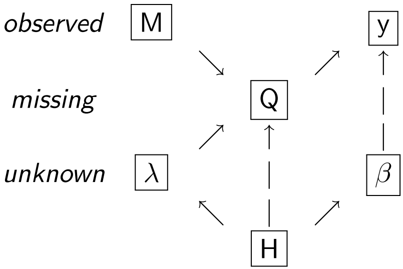
- Observed: y (traits) and M (marker and linkage map)
- Missing markers and QTL genotypes (Q)

Bayesian Interval Mapping Framework



- Observed: y (traits) and M (marker and linkage map)
- Missing markers and QTL genotypes (Q)
- Unknown parameters (λ, β, H, Q)

Bayesian Interval Mapping Framework



observed

M

y

missing

Q

unknown

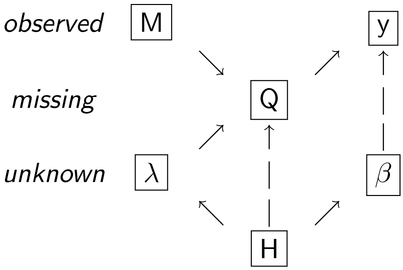
λ

β

H

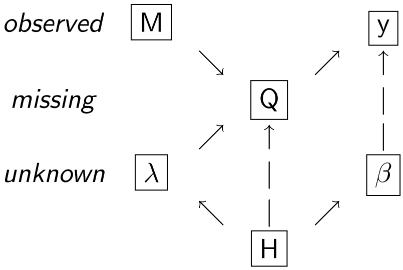
- Observed: y (traits) and M (marker and linkage map)
  - trait model $p(y \mid Q, \beta, \lambda, H)$
- Missing markers and QTL genotypes (Q)
- Unknown parameters (λ, β, H, Q)

Bayesian Interval Mapping Framework



- Observed: y (traits) and M (marker and linkage map)
  - trait model $p(y \mid Q, \beta, \lambda, H)$
- Missing markers and QTL genotypes (Q)
  - genetic model $p(Q \mid M, \lambda, H)$
- Unknown parameters (λ, β, H, Q)

Bayesian Interval Mapping Framework



- Observed: y (traits) and M (marker and linkage map)
  - trait model $p(y \mid Q, \beta, \lambda, H)$
- Missing markers and QTL genotypes (Q)
  - genetic model $p(Q \mid M, \lambda, H)$
- Unknown parameters (λ, β, H, Q)

posterior = likelihood × prior

$$p(\lambda, \beta, H, Q \mid y, M) \propto p(y \mid \beta, \lambda, Q, H) p(Q \mid M, \lambda, H) p(\beta, \lambda, H)$$

Why Multiple Traits?

y_1	y_2	C1M1	C1M2	C2M1	C2M2	C15M2	C16M1	C19M1
8.8	7.8	AA	AA	AB	AA	AA	AB	AB
9.6	10.1	AA	AA	AB	AB	AB	AB	AB
10.6	9.9	AB	AB	AA	AA	AB	AA	AA
11.1	10.9	AB	AB	AA	AB	AB	AA	AA

- Typically data on more than one phenotype (correlated) are collected *e.g.* BMI, fatmass etc.

Why Multiple Traits?

y_1	y_2	C1M1	C1M2	C2M1	C2M2	C15M2	C16M1	C19M1
8.8	7.8	AA	AA	AB	AA	AA	AB	AB
9.6	10.1	AA	AA	AB	AB	AB	AB	AB
10.6	9.9	AB	AB	AA	AA	AB	AA	AA
11.1	10.9	AB	AB	AA	AB	AB	AA	AA

- Typically data on more than one phenotype (correlated) are collected *e.g.* BMI, fatmass etc.
- Higher power to detect weak main and/or epistatic effects

Why Multiple Traits?

y_1	y_2	C1M1	C1M2	C2M1	C2M2	C15M2	C16M1	C19M1
8.8	7.8	AA	AA	AB	AA	AA	AB	AB
9.6	10.1	AA	AA	AB	AB	AB	AB	AB
10.6	9.9	AB	AB	AA	AA	AB	AA	AA
11.1	10.9	AB	AB	AA	AB	AB	AA	AA

- Typically data on more than one phenotype (correlated) are collected *e.g.* BMI, fatmass etc.
- Higher power to detect weak main and/or epistatic effects
- Higher precision of estimates

Why Multiple Traits?

y_1	y_2	C1M1	C1M2	C2M1	C2M2	C15M2	C16M1	C19M1
8.8	7.8	AA	AA	AB	AA	AA	AB	AB
9.6	10.1	AA	AA	AB	AB	AB	AB	AB
10.6	9.9	AB	AB	AA	AA	AB	AA	AA
11.1	10.9	AB	AB	AA	AB	AB	AA	AA

- Typically data on more than one phenotype (correlated) are collected *e.g.* BMI, fatmass etc.
- Higher power to detect weak main and/or epistatic effects
- Higher precision of estimates
- Separate close linkage from pleiotropy

Why Multiple Traits?

y_1	y_2	C1M1	C1M2	C2M1	C2M2	C15M2	C16M1	C19M1
8.8	7.8	AA	AA	AB	AA	AA	AB	AB
9.6	10.1	AA	AA	AB	AB	AB	AB	AB
10.6	9.9	AB	AB	AA	AA	AB	AA	AA
11.1	10.9	AB	AB	AA	AB	AB	AA	AA

- Typically data on more than one phenotype (correlated) are collected *e.g.* BMI, fatmass etc.
- Higher power to detect weak main and/or epistatic effects
- Higher precision of estimates
- Separate close linkage from pleiotropy
  - pleiotropy

Why Multiple Traits?

y_1	y_2	C1M1	C1M2	C2M1	C2M2	C15M2	C16M1	C19M1
8.8	7.8	AA	AA	AB	AA	AA	AB	AB
9.6	10.1	AA	AA	AB	AB	AB	AB	AB
10.6	9.9	AB	AB	AA	AA	AB	AA	AA
11.1	10.9	AB	AB	AA	AB	AB	AA	AA

- Typically data on more than one phenotype (correlated) are collected *e.g.* BMI, fatmass etc.
- Higher power to detect weak main and/or epistatic effects
- Higher precision of estimates
- Separate close linkage from pleiotropy
  - pleiotropy
    - one gene, affecting both traits indicating common biochemical pathways

Why Multiple Traits?

y_1	y_2	C1M1	C1M2	C2M1	C2M2	C15M2	C16M1	C19M1
8.8	7.8	AA	AA	AB	AA	AA	AB	AB
9.6	10.1	AA	AA	AB	AB	AB	AB	AB
10.6	9.9	AB	AB	AA	AA	AB	AA	AA
11.1	10.9	AB	AB	AA	AB	AB	AA	AA

- Typically data on more than one phenotype (correlated) are collected *e.g.* BMI, fatmass etc.
- Higher power to detect weak main and/or epistatic effects
- Higher precision of estimates
- Separate close linkage from pleiotropy
  - pleiotropy
    - one gene, affecting both traits indicating common biochemical pathways
  - close linkage

Why Multiple Traits?

y_1	y_2	C1M1	C1M2	C2M1	C2M2	C15M2	C16M1	C19M1
8.8	7.8	AA	AA	AB	AA	AA	AB	AB
9.6	10.1	AA	AA	AB	AB	AB	AB	AB
10.6	9.9	AB	AB	AA	AA	AB	AA	AA
11.1	10.9	AB	AB	AA	AB	AB	AA	AA

- Typically data on more than one phenotype (correlated) are collected *e.g.* BMI, fatmass etc.
- Higher power to detect weak main and/or epistatic effects
- Higher precision of estimates
- Separate close linkage from pleiotropy
  - pleiotropy
    - one gene, affecting both traits indicating common biochemical pathways
  - close linkage
    - two tightly linked genes resulting in collinear genotypes

QTL SUR Model

The QTL SUR Model:

$$y_{ti} = \mu_t + \mathbf{X}_{ti}\beta_t + e_{ti}, i = 1, \cdots, n; t = 1, \cdots, T$$

where t corresponds to the phenotypes or traits or dependent variables and i corresponds to the individuals. It is assumed the $\mathbf{e}_i = \{e_{1i}, \cdots, e_{Ti}\} \sim N_T(0, \Sigma)$

Model Parameters

Following Godsill (2001) fix the total number of loci/independent variables that can be selected to L Then define:

- Model Indicators : $\boldsymbol{\gamma} = \{\gamma_{t1}, \cdots, \gamma_{tL}\}$
- Locus Indices : $\boldsymbol{\lambda} = \{\lambda_{t1}, \cdots, \lambda_{tL}\}$

Following special cases of the SURd model can be obtained below:

- SURs : $\lambda_{ti} = \lambda_i \forall t = 1, \cdots, T$
- Tranditional Multivariate Model (TMV): $\gamma_{ti} = \gamma_t \forall t = 1, \cdots, T$
- Single Trait Analysis (STA): $\Sigma = I$ will reduce to univariate trait-by-trait analysis

Choice of Priors

Prior on β

- batches k=add,dom,add-add interaction etc.
- $\beta_k \sim \mathcal{N}(0, \sigma_k^2)$ and $\sigma_k^2 \sim Inv - \chi^2(\nu_k, s_k^2)$
- s_k^2 controls the prior heritability per effect $s_k^2 = (\nu_k - 2)E(h_j)V_p/(\nu_k V_j)$

Choice of Priors

Prior on β

- batches k=add,dom,add-add interaction etc.
- $\beta_k \sim \mathcal{N}(0, \sigma_k^2)$ and $\sigma_k^2 \sim Inv - \chi^2(\nu_k, s_k^2)$
- s_k^2 controls the prior heritability per effect $s_k^2 = (\nu_k - 2)E(h_j)V_p/(\nu_k V_j)$

Prior on number of QTL (ℓ)

- $\ell \sim Poission(\ell_0)$
- Choice of $L = \ell_0 + 3\sqrt{\ell_0}$

Choice of Priors

Prior on β

- batches k=add,dom,add-add interaction etc.
- $\beta_k \sim \mathcal{N}(0, \sigma_k^2)$ and $\sigma_k^2 \sim Inv - \chi^2(\nu_k, s_k^2)$
- s_k^2 controls the prior heritability per effect $s_k^2 = (\nu_k - 2)E(h_j)V_p/(\nu_k V_j)$

Prior on number of QTL (ℓ)

- $\ell \sim Poission(\ell_0)$
- Choice of $L = \ell_0 + 3\sqrt{\ell_0}$

Prior on λ and γ

- independent priors on QTL positions and indicators

Choice of Priors

Prior on β

- batches k=add,dom,add-add interaction etc.
- $\beta_k \sim \mathcal{N}(0, \sigma_k^2)$ and $\sigma_k^2 \sim Inv - \chi^2(\nu_k, s_k^2)$
- s_k^2 controls the prior heritability per effect $s_k^2 = (\nu_k - 2)E(h_j)V_p/(\nu_k V_j)$

Prior on Σ

- $p(\Sigma) \propto \frac{1}{|\Sigma| \prod_{i<j}(d_i - d_j)}$

Prior on number of QTL (ℓ)

- $\ell \sim Poission(\ell_0)$
- Choice of $L = \ell_0 + 3\sqrt{\ell_0}$

Prior on λ and γ

- independent priors on QTL positions and indicators

Composite Model Space Approach

- The idea is to circumvent the trans-dimensional character of the problem by modeling all parameters simultaneously.

- The joint posterior distribution:

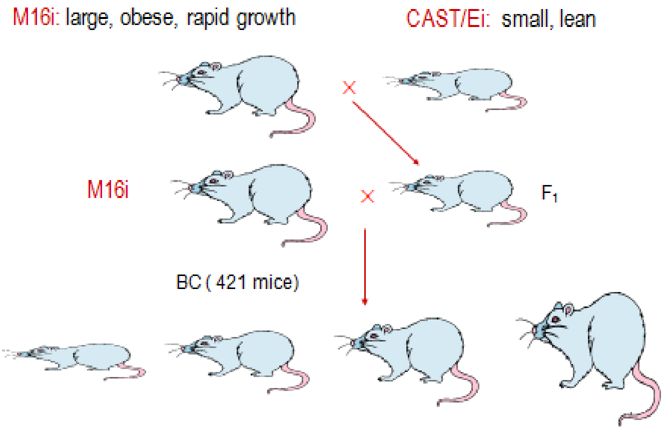
$$\begin{aligned} p(\gamma, \lambda, \theta, \Sigma | Y, X) &\propto p(Y|X, \gamma, \lambda, \theta, \Sigma) p(\lambda_\gamma, \theta_\gamma | \gamma, \Sigma) \\ \times &\quad p(\lambda_{-\gamma}, \theta_{-\gamma} | \gamma, \Sigma) p(\gamma) p(\Sigma, \theta) \end{aligned}$$

- where $\theta = \{\beta, \sigma^2\}$ and $\lambda_{-\gamma}$ is the collection of all λ_{ti}'s for which $\gamma_{ti} = 0$.

- Assume a priori independence

$$p(\lambda_{-\gamma}, \theta_{-\gamma} | \lambda_\gamma, \theta_\gamma, \gamma, \Sigma) \propto p(\lambda_{-\gamma}, \theta_{-\gamma} | \gamma, \Sigma)$$

Real Data Set



M16i: large, obese, rapid growth

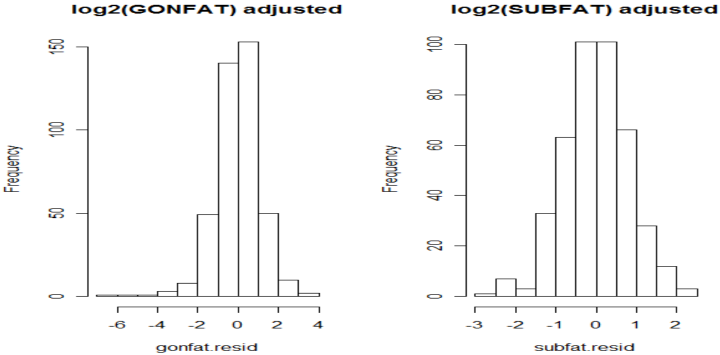
CAST/Ei: small, lean

M16i

F_1

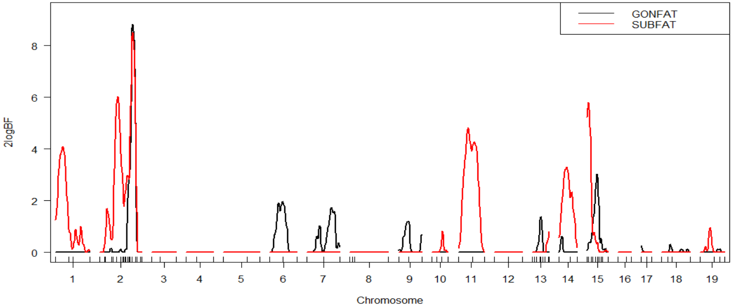
BC (421 mice)

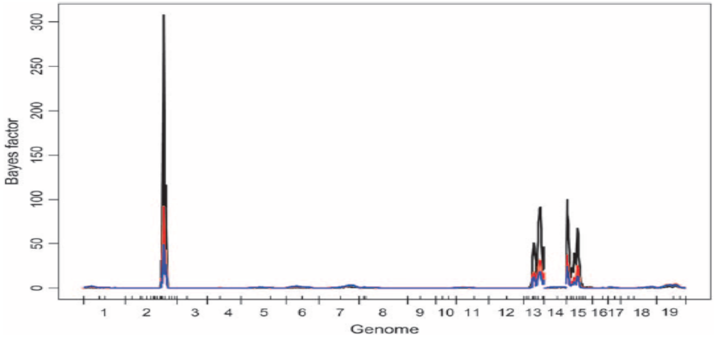
Trait Phenotype



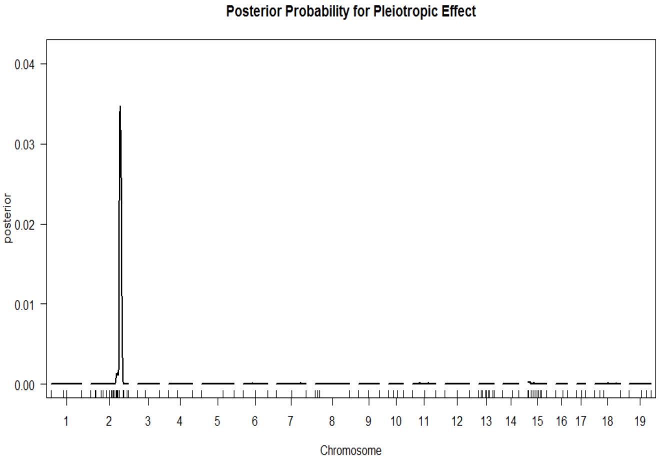
- GONFAT → Right Gonadal fat pad
- SUBFAT → Subcutaneous fat pad

Bayes Factor Profile for SUBFAT and GONFAT





Pleiotropic Effect



Future Research

Pleiotropy vs. Coincident linkage

- SURd: Models the coincident linkage hypothesis
- TMV: Models pleiotropy
- Bayes Factor comparison of pleiotropy vs coincident linkage

Variety of traits

- Ordinal traits using threshold model
- Survival traits

Future Research

eQTL (expression QTL)

- mRNA expression are considered traits
- Tens of thousands of traits (T)
- Lot of attention recently by researchers
- NIH RFAs
- http://grants.nih.gov/grants/guide/rfa-files/RFA-RM-09-006.html

Covariance matrix modeling

- Current implementation breaks down for large T
- Investigation of different priors

Acknowledgements

- Stefano Monni (Weill Cornell)
- Nengjun Yi (University of Alabama at Birmingham)
- Brian Yandell (University of Wisconsin - Madison)
- CTSC grant