

“Model” Selection Bias in Genome-Wide Genetic Mapping Studies

Lei Sun

University of Toronto

Model Selection Workshop - Vienna

July 24, 2008

Genetic Background in 1 Slide

- ➡ **SNP**: a type of *genetic marker* with only 2 variations in the population; \sim 10 million common SNPs in the genome
- ➡ **Allele**: a particular *variation* of a genetic marker; coded as a and A ; appearing in the population with certain **allele frequency**
- ➡ **Genotype**: the two alleles of an individual at a genetic marker
3 unordered genotypes for a SNP: aa Aa AA
- ➡ **Case-Control Genetic Association Test of a SNP**
e.g. 2 df genotype test, 1 df Armitage trend test, 1 df allele freq test, etc.

	aa	Aa	AA	Total
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

Genome-Wide Genetic Mapping Studies

- ⇒ A disease/trait of interest, e.g. type 1 diabetes, height
- ⇒ Two primary questions
 - t : where is the **location** of the gene on the genome?
 - μ : what is the **effect** of the gene, e.g. heritability, OR?
- ⇒ Genome-wide linkage/association studies; *1K to 1M markers* (surrogates for genes) placed across the genome, *for each marker t*
 - Testing** $H_0 : \mu_t = 0$ (\iff no genetic effect)
- ⇒ Gene-effect estimation; *for the most significant marker(s)*

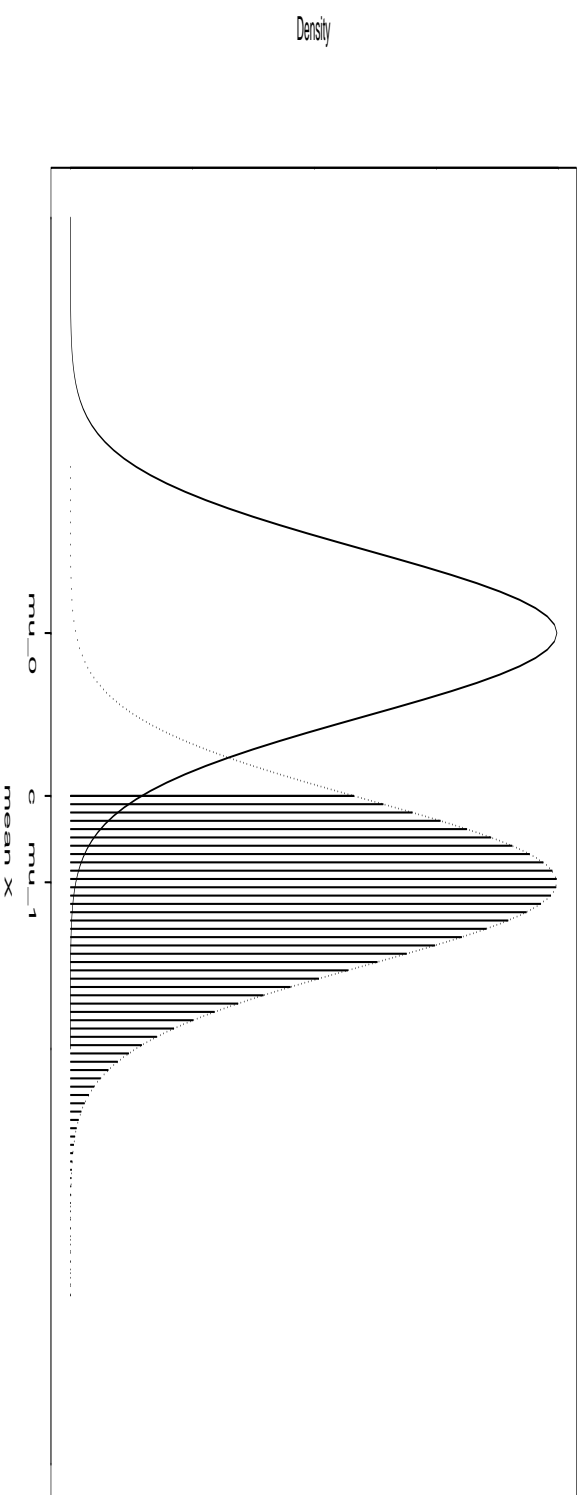
Estimating μ_t

Understanding the Bias

If $X_i \sim N(\mu, \sigma), i = 1, \dots, n$, is \bar{X} an unbiased estimator of μ ?

⇒ *Estimation w/o testing: unbiased*

⇒ **Estimation after testing (using the same sample): biased**
 such naïve estimator in fact is $\mu_N = \bar{X} \mid \bar{X} > \text{critical value}$



Understanding the Bias, cont'd

- ⇨ Source of bias I - **tail selection due to testing**
 - The same data is used for both testing and estimation
 - *Estimation is performed only if the testing was significant*
 - Bias increases as power decreases (often low power in GWA)
 - ⇨ Source of bias II - **marker(s)/variable(s) selection due to maximization or considering ordered statistics**
 - *We are looking at the top SNP(s) with the maximum test statistic(s) among all the SNPs in the genome*
- $$X_{it} \sim N(\mu_t, \sigma), i = 1, \dots, n, t = 1, \dots, T$$
- $$\bar{X}_{(T)} = \bar{X}_{max} = \max\{\bar{X}_{.1}, \dots, \bar{X}_{.T}\}, \text{ or } \bar{X}_{(T)} \geq \bar{X}_{(T-1)} \geq \dots$$
- With or without the gw significance criterion
 - A more subtle issue; and more difficult to model

Estimation Bias and Its Implications

- ➡ Linkage: Goring et al. (2001) Association: Garner (2007)
- ➡ Overestimate the importance of genetic effect
- ➡ Underestimate the sample size needed for replication study

	Stage 1		Stage 2	
	GW, 500 Tests	Replicate the Most Sig. Finding		
True OR	$\alpha_1 = 0.0001$	$\alpha_2 = 0.05, 1 - \beta = 80\%$		
μ	N_1	$\hat{\mu}$ (bias)	$\hat{N}_2 \mu$	$\hat{N}_2 \hat{\mu}$
1.1	1,000	1.33 (0.23)	~ 3,500	~ 400
	5,000	1.15 (0.05)		~ 1,600
1.2	1,000	1.35 (0.15)	~ 1,000	~ 350
	5,000	1.202 (0.002)		~ 1,000

Methods Proposed

⇨ Resampling based method

- Sun and Bull (2005; Genetic Epidemiology)
- Wu et al. (2005; BMC genetics)
- Wu et al. (2006; Human Heredity)
- Yu et al. (2007; American Journal of Human Genetics)
- Jeffries (2007; Biostatistics)

⇨ Likelihood based method

- Zollner and Pritchard (2007; American Journal of Human Genetics)
- Zhong and Prentice (2008; Biostatistics)
- Ghosh et al. (2008; American Journal of Human Genetics)

Connection w/ Regression Prediction Error

⇒ *Overfitting leads to Underestimation of PE*: the same data was used for both model building and evaluation $\widehat{PE} = \sum (y_i - \hat{y}_i)^2 / n$

⇒ **Resampling-based methods: e.g. cross-validation and bootstrap.**

$$\widehat{PE}_{\text{resampling}} = \sum (y_i - \hat{y}_i^{\{-i\}})^2 / n.$$

- *model building in part of the sample,*
- *model evaluation in the remaining (independent) sample,*
- repeat this “sample splitting” many times in the original sample

⇒ **Similar approaches can be proposed**

model building \iff gene locus detection

model evaluation \iff Effect Size estimation

optimistic PE (downward bias) \iff optimistic ES (upward bias)

Conceptual Bootstrap Algorithm

- ⇒ Original data: $X_i \sim N(\mu, \sigma), i = 1, \dots, n$
- ⇒ Naive procedure, *estimation subject to selection*

$$\mu_N = \bar{X} | \bar{X} > c$$

- ⇒ **Obtain B (e.g. B=100) bootstrap datasets**
- ⇒ For each b_{th} bootstrap dataset, consider
- ⇒ **Detection sample**
 - *Bootstrap sample*, $X_i^b \sim N(\mu, \sigma), i = 1, \dots, n$
 - *Mimic the original naive procedure*, $\mu_D^b \sim \mu_N$

$$\mu_D^b = \bar{X}^b | \bar{X}^b > c$$

⇒ **Estimation sample**

- *The data points not sampled, $X_i^{-b} \sim N(\mu, \sigma), i = 1, \dots, \sim 37\%n$*
- *Mimic estimation with independent data, a direct estimation*

$$\mu_E^b = \overline{X}^{-b}$$

⇒ **Estimation of the bias**

$$\mu_D^b - \mu_E^b$$

⇒ **Shrinkage bias-reduced estimator**

$$\mu_N - \overline{(\mu_D^b - \mu_E^b)}$$

⇒ **Out-of-sample estimator: $\overline{\mu_E^b}$**

⇒ **Weighted estimator: $(1 - w) \mu_N + w \overline{\mu_E^b}$**

Likelihood-based Methods

⇒ The naïve MLE (ignoring the fact that $\bar{X} > c$):

$$L_N(\mu) = \text{Prob}\{\bar{X}; \mu\}$$

⇒ The correct *conditional* MLE:

$$L_C(\mu) = \text{Prob}\{\bar{X} | \bar{X} > c; \mu\} = \frac{L(\mu)}{\text{Power}}$$

Bootstrap vs. Likelihood-based Methods

⇒ OR estimates in a real GW association study

- The Wellcome Trust Case Control Consortium (WTCCC)
- Type 1 Diabetes
- ~2,000 cases and ~3,000 controls, and ~500K SNPs ($\alpha \sim 10^{-6}$)

SNP	Naive (WTCCC)	Likelihood (Ghosh et al.)	Bootstrap (Faye et al.)	Indep/Replication (Todd et al.)
rs2542151	1.33	1.09	1.24	1.29
rs2292239	1.30	1.25	1.32	1.28
rs17696736	1.37	1.37	1.39	1.16
rs17388568	1.27	1.20	1.23	1.08

- Other concerns: e.g. heterogeneity between the two samples

Simulation studies

- Both methods substantially reduce the estimation bias with similar performance (smaller variance for bootstrap shrinkage).
- Results for a one-locus normal model ($\alpha = 10^{-6}$)

Estimator	Power = 10%		Power = 50%	
	Relative Bias	RMSE	Relative Bias	RMSE
Naïve	0.491	0.025	0.166	0.015
Likelihood	-0.257	0.033	-0.195	0.035
Bootstrap	0.196	0.015	-0.001	0.015

- Both are *not unbiased*
- Both have impractically *large variance*, e.g.
CI of OR for rs2292239: (1.08, 1.42) implies sample size needed for a replication study (80% power at 0.05) ranges from 6,000 to 350

Why the Variance is So Large?

- ⇒ The *effective* sample size for estimation *accounting for Selection* is *smaller* than n
- ⇒ The original sample size n can be deceptively large for estimation
- ⇒ The loss of information is due to the use of the *same dataset* for hypothesis testing
- ⇒ The information loss is inversely proportional to the power of the testing component
- ⇒ Leeb and Pötscher (2006): *one could not estimate the unconditional distribution of a post-model-selection estimator with reasonable accuracy even asymptotically*

Bayesian Framework

- ⇨ *Incorporating prior information to further reduce the bias and decrease the variance*
- ⇨ Current implementation in the single-locus normal model
 - “*spike and slab*” type of priors used in Bayesian variable selection (e.g. Ishwaran and Rao, 2005)
 - $p(\mu) \propto \xi \delta_{\{0\}}(\mu) + (1 - \xi)g(\mu|\theta)$, the Dirac measure at zero representing the fact that the marker of interest may not be the gene
 - $g(\mu|\theta) \propto \text{Uniform}(0, a)$, a is a large number corresponding a realistic upper bound of the parameter of interest
 - $\xi \propto \text{Beta}(2/3, 2/3)$, a hyper-parameter
 - *MCMC techniques needed* for the posterior distribution of μ
- ⇨ Preliminary simulation studies: **a 25% reduction in RMSE**

Where Are We Now?

- ➡ Single-locus model
- ➡ Confidence Interval
 - Bootstrap: a double-bootstrap procedure; computationally intensive
 - Likelihood
 - Zollner and Pritchard (2007): use of χ_1^2 distribution; contain all μ values s.t. $2\log(L_c(\hat{\mu})/L_c(\mu)) \leq 1 - \alpha$ quantiles of χ_1^2 ; might not be accurate because the conditional MLE is not normally distributed.
 - Ghosh et al. (2008): $\alpha/2$ and $1 - \alpha/2$ quantiles of the conditional density, $L_c(\mu)$

➡ Multi-locus model: GW setting; multi-marker/variable selection

- Bootstrap: straightforward; already implemented for GW linkage data; computationally intensive
 - Heritability estimates in a real GW linkage study
 - The Framingham Heart Study
 - Trait of interest: Systolic Blood Pressure
 - ~ 330 families ($\sim 1,700$ individuals) and ~ 400 markers
 - 2 significant markers, i.e. LOD (\log_{10} -based LR) > 3

Marker	On Which Chromosome	Observed LOD	Naïve (Briollais et al.)	Bootstrap (Wu et al.)
1	8	3.43	0.55	0.13
2	17	3.29	0.46	0.11

- Likelihood: difficult to model the order statistic in the presence of correlation between markers/tests

▣▣▣ → Prior information

- Bayesian framework: more in-depth work needed for both the prior and the MCMC algorithm
 - Bootstrap framework: $\mu_N - k \overline{(\mu_D^b - \mu_E^b)}$, $k \sim \text{prior}$?
 - Can we show that the conditional MLE is biased?
 - Can we find a family of priors that leads to unbiased estimators?
- ▣▣▣ → Software development for the bootstrap methods
- BR-squared (Bias-Reduced estimates via Bootstrap Resampling)
 - Phase I: GW linkage data
 - Phase II: GW association data

Acknowledgement

⇨ Collaborators

- co-PIs: Shelley Bull, Radu Craiu
- Students: Laura Faye, Longyang Wu, Lizhen Xu

⇨ Fundings

- CIHR (Canadian Institutes of Health Research)
- NSERC (Natural Sciences and Engineering Research Council of Canada)