GENOME-WIDE SCANS FOR QUANTITATIVE TRAIT LOCI IN EXPERIMENTAL POPULATIONS - ISSUES OF MULTIPLE TESTING AND MODEL SELECTION

Małgorzata Bogdan

Institute of Mathematics and Computer Science, Wrocław University of Technology, Poland

Poznań, 18.05.2010

Data for QTL mapping in backcross population and recombinant inbred lines

Y_i, $1 \leq i \leq n$ - trait values

Data for QTL mapping in backcross population and recombinant inbred lines

Y_i, $1 \leq i \leq n$ - trait values

Only two genotypes possible at a given locus

Data for QTL mapping in backcross population and recombinant inbred lines

Y_i, $1 \leq i \leq n$ - trait values

Only two genotypes possible at a given locus

X_{ij}, $1 \leq i \leq n$, $1 \leq j \leq m$ - dummy variables encoding genotypes at m markers

$X_{ij} \in \{-1/2, 1/2\}$

Data for QTL mapping in backcross population and recombinant inbred lines

Y_i, $1 \leq i \leq n$ - trait values

Only two genotypes possible at a given locus

X_{ij}, $1 \leq i \leq n$, $1 \leq j \leq m$ - dummy variables encoding genotypes at m markers

$X_{ij} \in \{-1/2, 1/2\}$

Multiple regression model:

$$Y_i = \mu + \sum_{j \in I} \beta_j X_{ij} + \sum_{(u,v) \in U} \gamma_{uv} X_{iu} X_{iv} + \varepsilon_i, \qquad (0.1)$$

I - a subset of $N = \{1, \ldots, m\}$, U - a subset of $N \times N$,

$\epsilon_i \sim N(0, \sigma^2)$

Notation

Task : estimation of the number of influential genes and interaction effects

Notation

Task : estimation of the number of influential genes and interaction effects

M_i - i-th linear model (set of explanatory variables),
k_i - number of main effects, q_i - number of interactions, $k_i + q_i < n$

Notation

Task : estimation of the number of influential genes and interaction effects

M_i - i-th linear model (set of explanatory variables),
k_i - number of main effects, q_i - number of interactions, $k_i + q_i < n$
$\theta_i = (\beta_0, \beta_1, \ldots, \beta_{k_i}, \gamma_1, \ldots, \gamma_{q_i}, \sigma)$ - vector of parameters

Notation

Task : estimation of the number of influential genes and interaction effects

M_i - i-th linear model (set of explanatory variables),
k_i - number of main effects, q_i - number of interactions, $k_i + q_i < n$
$\theta_i = (\beta_0, \beta_1, \ldots, \beta_{k_i}, \gamma_1, \ldots, \gamma_{q_i}, \sigma)$ - vector of parameters

$$L(Y|M_i, \theta_i) = \prod_{i=1}^{n} f_i(Y_i) = \frac{1}{(\sqrt{2\pi}\sigma)^n} exp\left(-\frac{RSS_{M_i,\theta_i}}{2\sigma^2}\right)$$

$$RSS_{M_i,\theta_i} = \sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{k_i} \beta_j X_{ij} - \sum_{l=1}^{q_i} \gamma_j X_{iu(l)} X_{iv(l)})^2$$

Model Selection Criteria

Akaike Information Criterion (Akaike, 1973) – maximize
$AIC = \log L(Y|M_i, \hat{\theta}_i) - (k_i + q_i)$

Model Selection Criteria

Akaike Information Criterion (Akaike, 1973) – maximize $AIC = \log L(Y|M_i, \hat{\theta}_i) - (k_i + q_i)$

Bayesian Information Criterion (Schwarz, 1978) – maximize $BIC = \log L(Y|M_i, \hat{\theta}_i) - \frac{1}{2}(k_i + q_i) \log n$

Model Selection Criteria

Akaike Information Criterion (Akaike, 1973) – maximize $AIC = \log L(Y|M_i, \hat{\theta}_i) - (k_i + q_i)$

Bayesian Information Criterion (Schwarz, 1978) – maximize $BIC = \log L(Y|M_i, \hat{\theta}_i) - \frac{1}{2}(k_i + q_i) \log n$

If m is fixed, $n \to \infty$ and $X'X/n \to Q$, where Q is a positive definite matrix, then BIC is consistent - the probability of choosing the proper model converges to 1.

Model Selection Criteria

Akaike Information Criterion (Akaike, 1973) – maximize $AIC = \log L(Y|M_i, \hat{\theta}_i) - (k_i + q_i)$

Bayesian Information Criterion (Schwarz, 1978) – maximize $BIC = \log L(Y|M_i, \hat{\theta}_i) - \frac{1}{2}(k_i + q_i) \log n$

If m is fixed, $n \to \infty$ and $X'X/n \to Q$, where Q is a positive definite matrix, then BIC is consistent - the probability of choosing the proper model converges to 1.

When $n \geq 8$ BIC never chooses more regressors than AIC and is usually considered as one of the most restrictive model selection criteria.

Model Selection Criteria

Akaike Information Criterion (Akaike, 1973) – maximize $AIC = \log L(Y|M_i, \hat{\theta}_i) - (k_i + q_i)$

Bayesian Information Criterion (Schwarz, 1978) – maximize $BIC = \log L(Y|M_i, \hat{\theta}_i) - \frac{1}{2}(k_i + q_i) \log n$

If m is fixed, $n \to \infty$ and $X'X/n \to Q$, where Q is a positive definite matrix, then BIC is consistent - the probability of choosing the proper model converges to 1.

When $n \geq 8$ BIC never chooses more regressors than AIC and is usually considered as one of the most restrictive model selection criteria.

Surprise ? : - Broman and Speed (JRSS, 2002) report that BIC overestimates the number of regressors when applied to QTL mapping.

Explanation - Bayesian roots of BIC (1)

$f(\theta_i)$ – prior density of θ_i, $\pi(M_i)$ – prior probability of M_i

Explanation - Bayesian roots of BIC (1)

$f(\theta_i)$ – prior density of θ_i, $\pi(M_i)$ – prior probability of M_i

$m_i(Y) = \int L(Y|M_i, \theta_i) f(\theta_i) d\theta_i$ – integrated likelihood of the data given the model M_i

Explanation - Bayesian roots of BIC (1)

$f(\theta_i)$ – prior density of θ_i, $\pi(M_i)$ – prior probability of M_i

$m_i(Y) = \int L(Y|M_i, \theta_i) f(\theta_i) d\theta_i$ – integrated likelihood of the data given the model M_i

posterior probability of M_i : $P(M_i|Y) \propto m_i(Y)\pi(M_i)$

Explanation - Bayesian roots of BIC (1)

$f(\theta_i)$ – prior density of θ_i, $\pi(M_i)$ – prior probability of M_i

$m_i(Y) = \int L(Y|M_i,\theta_i)f(\theta_i)d\theta_i$ – integrated likelihood of the data given the model M_i

posterior probability of M_i : $P(M_i|Y) \propto m_i(Y)\pi(M_i)$

BIC neglects $\pi(M_i)$ and uses Laplace approximation

$$\log m_i(Y) \approx \log L(Y|M_i,\hat{\theta}_i) - 1/2(k_i + q_i + 2)\log n + R_i,$$

where R_i is bounded in n.

Explanation - Bayesian roots of BIC (2)

neglecting $\pi(M_i) \equiv$ assigning the same probability to all models

Explanation - Bayesian roots of BIC (2)

neglecting $\pi(M_i) \equiv$ assigning the same probability to all models $\equiv$ the prior on the number of main effects is K is $B(m, \frac{1}{2})$

Explanation - Bayesian roots of BIC (2)

neglecting $\pi(M_i) \equiv$ assigning the same probability to all models

$\equiv$ the prior on the number of main effects is K is $B(m, \frac{1}{2})$

$E(K) = \frac{m}{2}$, $std(K) = \frac{\sqrt{m}}{2}$

Explanation - Bayesian roots of BIC (2)

neglecting $\pi(M_i) \equiv$ assigning the same probability to all models

$\equiv$ the prior on the number of main effects is K is $B(m, \frac{1}{2})$

$E(K) = \frac{m}{2}$, $std(K) = \frac{\sqrt{m}}{2}$

distribution concentrated almost entirely on

$[m/2 - 2\sqrt{m}, m/2 + 2\sqrt{m}]$

Explanation - Bayesian roots of BIC (2)

neglecting $\pi(M_i) \equiv$ assigning the same probability to all models

$\equiv$ the prior on the number of main effects is K is $B(m, \frac{1}{2})$

$E(K) = \frac{m}{2}$, $std(K) = \frac{\sqrt{m}}{2}$

distribution concentrated almost entirely on

$[m/2 - 2\sqrt{m}, m/2 + 2\sqrt{m}]$

for $m = 400$ the prior distribution on K is almost entirely concentrated on $[160, 240]$

Modified version of BIC, mBIC

M. Bogdan, J.K. Ghosh,R.W. Doerge, *Genetics* (2004)
Solution - using an informative prior distribution on the number of main and interaction effects

Modified version of BIC, mBIC

M. Bogdan, J.K. Ghosh,R.W. Doerge, *Genetics* (2004)
Solution - using an informative prior distribution on the number of main and interaction effects

Prior distribution on the number of main effects: $B(m, p_1)$

Modified version of BIC, mBIC

M. Bogdan, J.K. Ghosh,R.W. Doerge, *Genetics* (2004)

Solution - using an informative prior distribution on the number of main and interaction effects

Prior distribution on the number of main effects: $B(m, p_1)$

Prior distribution on the number of interactions: $B(N_e, p_2)$, where $N_e = m(m-1)/2$

Modified version of BIC, mBIC

M. Bogdan, J.K. Ghosh,R.W. Doerge, *Genetics* (2004)

Solution - using an informative prior distribution on the number of main and interaction effects

Prior distribution on the number of main effects: $B(m, p_1)$

Prior distribution on the number of interactions: $B(N_e, p_2)$, where $N_e = m(m-1)/2$

$E(k) = mp_1 = c_1$, $E(q) = N_e p_2 = c_2$

Modified version of BIC, mBIC

M. Bogdan, J.K. Ghosh,R.W. Doerge, *Genetics* (2004)

Solution - using an informative prior distribution on the number of main and interaction effects

Prior distribution on the number of main effects: $B(m, p_1)$

Prior distribution on the number of interactions: $B(N_e, p_2)$, where $N_e = m(m-1)/2$

$E(k) = mp_1 = c_1$, $E(q) = N_e p_2 = c_2$

mBIC: maximize

$$\log L(Y|\hat{\theta}) - \frac{1}{2}(k+q)\log(n) - k\log\left(\frac{m}{c_1} - 1\right) - q\log\left(\frac{N_e}{c_2} - 1\right)$$

Modified version of BIC, mBIC

M. Bogdan, J.K. Ghosh,R.W. Doerge, *Genetics* (2004)

Solution - using an informative prior distribution on the number of main and interaction effects

Prior distribution on the number of main effects: $B(m, p_1)$

Prior distribution on the number of interactions: $B(N_e, p_2)$, where $N_e = m(m-1)/2$

$E(k) = mp_1 = c_1$, $E(q) = N_e p_2 = c_2$

mBIC: maximize

$$\log L(Y|\hat{\theta}) - \frac{1}{2}(k+q)\log(n) - k\log\left(\frac{m}{c_1} - 1\right) - q\log\left(\frac{N_e}{c_2} - 1\right)$$

Standard version of mBIC uses $c_1 = c_2 = 2.2$ to control the overall type I error at the level below 10% (for $n \geq 200$ and $m \geq 30$.)

Modified version of BIC, mBIC

M. Bogdan, J.K. Ghosh,R.W. Doerge, *Genetics* (2004)

Solution - using an informative prior distribution on the number of main and interaction effects

Prior distribution on the number of main effects: $B(m, p_1)$

Prior distribution on the number of interactions: $B(N_e, p_2)$, where $N_e = m(m-1)/2$

$E(k) = mp_1 = c_1$, $E(q) = N_e p_2 = c_2$

mBIC: maximize

$$\log L(Y|\hat{\theta}) - \frac{1}{2}(k+q)\log(n) - k\log\left(\frac{m}{c_1} - 1\right) - q\log\left(\frac{N_e}{c_2} - 1\right)$$

Standard version of mBIC uses $c_1 = c_2 = 2.2$ to control the overall type I error at the level below 10% (for $n \geq 200$ and $m \geq 30$.)

The overall type I error is approximately equally divided between main and interaction effects.

Relationship to multiple testing - only main effects (1)

Orthogonal design: $X^T X = nI_{(m+1)\times(m+1)}, \quad (1)$

Relationship to multiple testing - only main effects (1)

Orthogonal design: $X^T X = nI_{(m+1)\times(m+1)}, \quad (1)$

BIC chooses those X_j's for which

$$\frac{n\hat{\beta}_j^2}{\sigma^2} > \log n$$

Relationship to multiple testing - only main effects (1)

Orthogonal design: $X^T X = nI_{(m+1)\times(m+1)}, \quad (1)$

BIC chooses those X_j's for which

$$\frac{n\hat{\beta}_j^2}{\sigma^2} > \log n$$

Under $\mathcal{H}_{0j}:\ \beta_j = 0, \qquad Z_j = \frac{\sqrt{n}\hat{\beta}_j}{\sigma} \sim N(0,1)$

Relationship to multiple testing - only main effects (1)

Orthogonal design: $X^T X = nI_{(m+1)\times(m+1)}, \qquad (1)$

BIC chooses those X_j's for which

$$\frac{n\hat{\beta}_j^2}{\sigma^2} > \log n$$

Under $\mathcal{H}_{0j}: \ \beta_j = 0, \qquad Z_j = \frac{\sqrt{n}\hat{\beta}_j}{\sigma} \sim N(0,1)$

It holds that for large values of n

$$\alpha_n = 2P(Z_j > \sqrt{\log n}) \approx \sqrt{\frac{2}{\pi n \log n}}.$$

Relationship to multiple testing (2)

When the number of true signals $K << m$, the expected number of "false discoveries" is approximately equal to $E(FP) = m\sqrt{\frac{2}{\pi n \log n}}$.

Relationship to multiple testing (2)

When the number of true signals $K << m$, the expected number of "false discoveries" is approximately equal to $E(FP) = m\sqrt{\frac{2}{\pi n \log n}}$.

When $m = n = 200$ then $E(FP) \approx 5$

Relationship to multiple testing (2)

When the number of true signals $K << m$, the expected number of "false discoveries" is approximately equal to $E(FP) = m\sqrt{\frac{2}{\pi n \log n}}$.

When $m = n = 200$ then $E(FP) \approx 5$

BIC is not consistent when $\frac{m}{\sqrt{n \log n}} \rightarrow \infty$

Relationship to multiple testing (3)

Bonferroni correction for multiple testing : $\alpha_{n,m} = \frac{\alpha_n}{m}$

Relationship to multiple testing (3)

Bonferroni correction for multiple testing : $\alpha_{n,m} = \frac{\alpha_n}{m}$

probability of detecting at least one “false positive”: FWER $\leq \alpha_n$

Relationship to multiple testing (3)

Bonferroni correction for multiple testing : $\alpha_{n,m} = \frac{\alpha_n}{m}$

probability of detecting at least one "false positive": FWER $\leq \alpha_n$

$2(1 - \Phi(\sqrt{c_{Bon}})) = \frac{\alpha_n}{m}$

Relationship to multiple testing (3)

Bonferroni correction for multiple testing : $\alpha_{n,m} = \frac{\alpha_n}{m}$

probability of detecting at least one "false positive": FWER $\leq \alpha_n$

$2(1 - \Phi(\sqrt{c_{Bon}})) = \frac{\alpha_n}{m}$

$$c_{Bon} = 2\log\left(\frac{m}{\alpha_n}\right)(1 + o_{n,m}) = (\log n + 2\log m)(1 + o_{n,m}) \ ,$$

where $o_{n,m}$ converges to zero when n or m tends to infinity.

Relationship to multiple testing (3)

Bonferroni correction for multiple testing : $\alpha_{n,m} = \frac{\alpha_n}{m}$

probability of detecting at least one “false positive”: FWER $\leq \alpha_n$

$2(1 - \Phi(\sqrt{c_{Bon}})) = \frac{\alpha_n}{m}$

$$c_{Bon} = 2\log\left(\frac{m}{\alpha_n}\right)(1 + o_{n,m}) = (\log n + 2\log m)(1 + o_{n,m}) \quad ,$$

where $o_{n,m}$ converges to zero when n or m tends to infinity.

$c_{mBIC} = \log n + 2\log\left(\frac{m}{c} - 1\right) \approx \log n + 2\log m - 2\log c$

Extended BIC, EBIC

J. Chen, Z. Chen, *Biometrika* (2008)

Extended BIC, EBIC

J. Chen, Z. Chen, *Biometrika* (2008)

Standard version - uniform prior on the number of main effects

Extended BIC, EBIC

J. Chen, Z. Chen, *Biometrika* (2008)

Standard version - uniform prior on the number of main effects

$$EBIC := 2\log(L(Y|\hat{\theta})) - k\log(n) - 2\log\binom{m}{k} \ .$$

Extended BIC, EBIC

J. Chen, Z. Chen, *Biometrika* (2008)

Standard version - uniform prior on the number of main effects

$$EBIC := 2\log(L(Y|\hat{\theta})) - k\log(n) - 2\log\binom{m}{k}\ .$$

When $K \leq 10$ EBIC works similarly to mBIC.

For larger K EBIC offers a substantially larger power then mBIC.

Caution - when $k > m/2$ the additional penalty becomes negative.

Extended BIC, EBIC

J. Chen, Z. Chen, *Biometrika* (2008)

Standard version - uniform prior on the number of main effects

$$EBIC := 2\log(L(Y|\hat{\theta})) - k\log(n) - 2\log\binom{m}{k} \ .$$

When $K \leq 10$ EBIC works similarly to mBIC.

For larger K EBIC offers a substantially larger power then mBIC.

Caution - when $k > m/2$ the additional penalty becomes negative.

$$mBIC2 := 2\log(L(Y|\hat{\theta})) - k\log(n) - 2k\log(m/4) + 2\log(k!)$$

Extended BIC, EBIC

J. Chen, Z. Chen, *Biometrika* (2008)

Standard version - uniform prior on the number of main effects

$$EBIC := 2\log(L(Y|\hat{\theta})) - k\log(n) - 2\log\binom{m}{k} .$$

When $K \leq 10$ EBIC works similarly to mBIC.

For larger K EBIC offers a substantially larger power then mBIC.

Caution - when $k > m/2$ the additional penalty becomes negative.

$$mBIC2 := 2\log(L(Y|\hat{\theta})) - k\log(n) - 2k\log(m/4) + 2\log(k!)$$

Related to the Benjamini-Hochberg correction for multiple testing

Properties of mBIC and mBIC2

Consistency and Bayesian asymptotic optimality under sparsity and orthogonal designs (asymptotics when $m \to \infty$ and $n \to \infty$):

Frommlet, Bogdan, Chakrabarti, Ghosh, Murawska, in preparation

Properties of mBIC and mBIC2

Consistency and Bayesian asymptotic optimality under sparsity and orthogonal designs (asymptotics when $m \to \infty$ and $n \to \infty$):

Frommlet, Bogdan, Chakrabarti, Ghosh, Murawska, in preparation

mBIC asymptotically optimal when $pm \approx C$,

mBIC2 asymptotically optimal when $p \to 0$ and $mp \to (0, \infty]$.

Properties of mBIC and mBIC2

Consistency and Bayesian asymptotic optimality under sparsity and orthogonal designs (asymptotics when $m \to \infty$ and $n \to \infty$):

Frommlet, Bogdan, Chakrabarti, Ghosh, Murawska, in preparation

mBIC asymptotically optimal when $pm \approx C$,

mBIC2 asymptotically optimal when $p \to 0$ and $mp \to (0, \infty]$.

Consistency of EBIC, mBIC and mBIC2, under more general designs and a fixed bound on the maximal number of effects - Chen and Chen (Biometrika, 2008)

Applications of mBIC for QTL mapping

1. Extending to intercross + a two-step version of mBIC : Baierl, Bogdan, Frommlet, Futschik *Genetics, 2006*

2. Robust versions based on M-estimates: Baierl, Futschik, Bogdan, Biecek *CSDA, 2007*

3. Rank version: Żak, Baierl, Bogdan, Futschik *Genetics, 2007*

4. Application for dense markers and interval mapping: Bogdan, Frommlet, Biecek, Cheng, Ghosh, Doerge, *Biometrics*, 2008

Intercross

Cockerham model (Kao and Zeng, Genetics, 2002):

Additive Effect for individual i: $X_{aij} = \begin{cases} 1 & \text{if } g_{ij} = AA, \\ 0 & \text{if } g_{ij} = aA, \\ -1 & \text{if } g_{ij} = aa. \end{cases}$

Dominance Effect for individual i: $X_{dij} = \begin{cases} 1/2 & \text{if } g_{ij} = Aa, \\ -1/2 & \text{otherwise} \quad . \end{cases}$

Intercross

Cockerham model (Kao and Zeng, Genetics, 2002):

Additive Effect for individual i: $$X_{aij} = \begin{cases} 1 & \text{if } g_{ij} = AA, \\ 0 & \text{if } g_{ij} = aA, \\ -1 & \text{if } g_{ij} = aa. \end{cases}$$

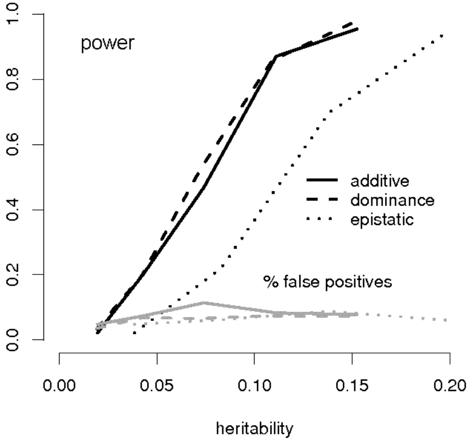
Dominance Effect for individual i: $$X_{dij} = \begin{cases} 1/2 & \text{if } g_{ij} = Aa, \\ -1/2 & \text{otherwise} \quad . \end{cases}$$

Four types of interaction effects: add-add $X_{aij}X_{aik}$, add-dom $X_{aij}X_{dik}$, dom-add $X_{dij}X_{aik}$ and dom-dom $X_{dij}X_{dik}$

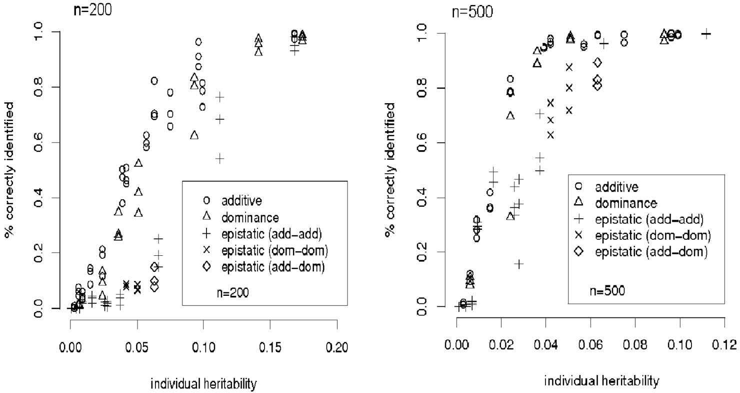
Simulations (1)

A. Baierl, M. Bogdan, F.Frommlet, A. Futschik, *Genetics* (2006) - intercross design

Simple models - one or two effects, n=200



Simulations (2)



Zero Inflated Generalized Poisson Regression (1)

Erhardt, Bogdan, Czado (2010) - submitted for publication

Zero Inflated Generalized Poisson Regression (1)

Erhardt, Bogdan, Czado (2010) - submitted for publication

Y - count variable

Zero Inflated Generalized Poisson Regression (1)

Erhardt, Bogdan, Czado (2010) - submitted for publication

Y - count variable

Poisson distribution: for $k \in 0, 1, ..$ $\quad P(Y = k) = \frac{\mu^k}{k!}\exp(-\mu)$

Zero Inflated Generalized Poisson Regression (1)

Erhardt, Bogdan, Czado (2010) - submitted for publication

Y - count variable

Poisson distribution: for $k \in 0, 1, ..$ $\quad P(Y = k) = \frac{\mu^k}{k!} \exp(-\mu)$

$EY = VarY = \mu$

Zero Inflated Generalized Poisson Regression (1)

Erhardt, Bogdan, Czado (2010) - submitted for publication

Y - count variable

Poisson distribution: for $k \in 0, 1, ..$ $\quad P(Y = k) = \frac{\mu^k}{k!} \exp(-\mu)$

$EY = VarY = \mu$

Generalized Poisson Regression:

$$P(Y = k) = \frac{\mu(\mu + (\varphi - 1)k)^{k-1}}{k!} \varphi^{-k} e^{-\frac{1}{\varphi}(\mu + (\varphi - 1)k)}$$

Zero Inflated Generalized Poisson Regression (1)

Erhardt, Bogdan, Czado (2010) - submitted for publication

Y - count variable

Poisson distribution: for $k \in 0, 1, ..$ $P(Y = k) = \frac{\mu^k}{k!} \exp(-\mu)$

$EY = VarY = \mu$

Generalized Poisson Regression:

$$P(Y = k) = \frac{\mu(\mu + (\varphi - 1)k)^{k-1}}{k!} \varphi^{-k} e^{-\frac{1}{\varphi}(\mu + (\varphi - 1)k)}$$

$E(Y) = \mu$ and $Var(Y) = \varphi^2 \mu$

Zero Inflated Generalized Poisson Regression

Zero Inflated Generalized Poisson Distribution :

$$ZIGP(\mu,\varphi,\omega) = \omega\delta_0 + (1-\omega)GP(\mu,\varphi) \quad ,$$

where $\omega \in [0,1]$ is the zero-inflation parameter.

Zero Inflated Generalized Poisson Regression

Zero Inflated Generalized Poisson Distribution :

$$ZIGP(\mu, \varphi, \omega) = \omega\delta_0 + (1-\omega)GP(\mu, \varphi) \quad ,$$

where $\omega \in [0, 1]$ is the zero-inflation parameter.

Zero Inflated Generalized Poisson Regression: $\omega = const$,

$$\log \mu_i = \beta_0 + \sum_{j \in I} \beta_j X_{ij} + \sum_{(u,v) \in U} \gamma_{uv} X_{iu} X_{iv}$$

Zero Inflated Generalized Poisson Regression

Zero Inflated Generalized Poisson Distribution :

$$ZIGP(\mu, \varphi, \omega) = \omega\delta_0 + (1-\omega)GP(\mu, \varphi) \quad ,$$

where $\omega \in [0,1]$ is the zero-inflation parameter.

Zero Inflated Generalized Poisson Regression: $\omega = const$,

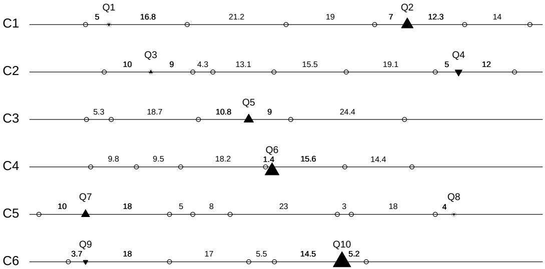
$$\log \mu_i = \beta_0 + \sum_{j \in I} \beta_j X_{ij} + \sum_{(u,v) \in U} \gamma_{uv} X_{iu} X_{iv}$$

Vinzenz Erhardt - implementation in R, available in CRAN (The Comprehensive R Archive Network)

Simulations

Lyons, M. A., H. Wittenburg, R. Li, K. A. Walsh, M. R. Leonard, G. A. Churchill, M. C. Carey, and B. Paigen (2003). New quantitative trait loci that contribute to cholesterol gallstone formation detected in an intercross of CAST/Ei and 129S1/SvImJ inbred mice. *Physiol. Genomics 14*(3), 225–239.

20 chromosomes, 100 markers, 10 QTL on 6 chromosomes



Characteristics

- true positives (TP): number of selected effects whose distance to the simulated QTL's was less or equal 20 cM; if more than one effect was caught in the interval around a certain QTL only one of them was counted
- false positives (FP): number of selected effects whose distance to the simulated QTL's was higher than 20 cM
- misclassification error, ME = false positives (FP) + false negatives (FN), where $FN = 10 - TP$
- power: $TP/10$
- observed false discovery rate : $FDR = FP/(FP + TP)$

Results (1)

	$\mathbf{n} = \mathbf{500}$, $\varphi = 2$, $\omega = 40\%$				
	mBIC				
	LM	PoiR	ZIPR	GPR	ZIGPR
FP	0.117	30.817	13.813	0.405	0.234
ME	5.949	31.847	15.088	8.381	4.047
Power	0.417	0.897	0.873	0.202	0.619
FDR	0.025	0.770	0.599	0.142	0.033
	EBIC				
FP	0.120	48.520	23.765	0.465	0.435
ME	5.815	49.110	24.665	8.490	3.940
Power	0.430	0.941	0.910	0.198	0.649
FDR	0.024	0.835	0.708	0.154	0.057

Results (2)

Data generated according to the Poisson Regression.

	$\mathbf{n = 200}$, mBIC				
	LM	PoiR	ZIPR	GPR	ZIGPR
FP	0.095	8.200	8.150	0.405	0.410
FP+FN	5.285	9.830	9.810	3.920	3.930
Power	0.481	0.837	0.834	0.648	0.648
FDR	0.018	0.476	0.475	0.053	0.053

Conclusion - Poisson Regression has a tendency to include spurious QTL to explain an increased data heterogeneity

Real Data Analysis (1)

Lyons, M. A., H. Wittenburg, R. Li, K. A. Walsh, M. R. Leonard, G. A. Churchill, M. C. Carey, and B. Paigen (2003). New quantitative trait loci that contribute to cholesterol gallstone formation detected in an intercross of CAST/Ei and 129S1/SvImJ inbred mice. *Physiol. Genomics 14*(3), 225–239.

Real Data Analysis (1)

Lyons, M. A., H. Wittenburg, R. Li, K. A. Walsh, M. R. Leonard, G. A. Churchill, M. C. Carey, and B. Paigen (2003). New quantitative trait loci that contribute to cholesterol gallstone formation detected in an intercross of CAST/Ei and 129S1/SvImJ inbred mice. *Physiol. Genomics 14*(3), 225–239.

$n = 277$, intercross, Y - number of gallstones

Real Data Analysis (2)

Additive effect at D5Mit183 (QTL previously identified at Lyons et al. (2003))

Real Data Analysis (2)

Additive effect at D5Mit183 (QTL previously identified at Lyons et al. (2003))

Interaction with a novel QTL at D4Mit42

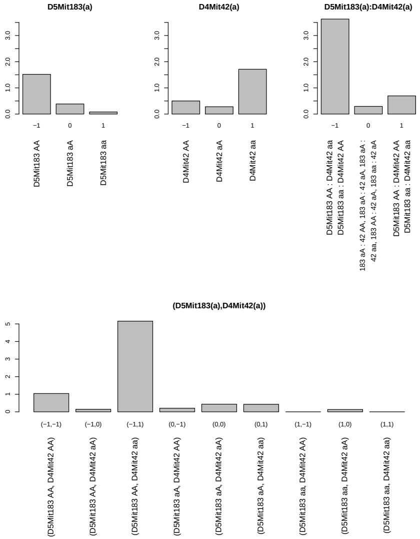
Real Data Analysis (2)

Additive effect at D5Mit183 (QTL previously identified at Lyons et al. (2003))

Interaction with a novel QTL at D4Mit42

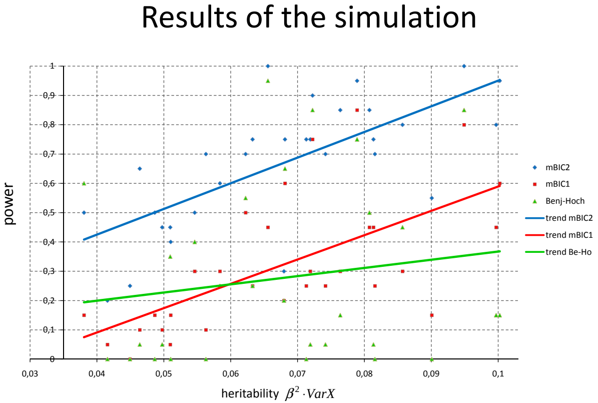
	Estimate	Std. Error	z value	$Pr(>\|z\|)$
Intercept	-0.864	0.573	-1.510	0.131
D5Mit183(a)	-1.244	0.442	-2.817	0.005
D4Mit42(a)	-0.215	0.476	-0.451	0.652
D5Mit183(a):D4Mit42(a)	-2.177	0.548	-3.973	$7.1 \cdot 10^{-5}$
φ	5.387	2.185	2.466	0.014
ω	0.458	0.163	2.809	0.005

Real Data Analysis (2)



Current research

Application for GWAS: Frommlet, Twaróg, Bogdan - in preparation $m \approx 200000$, 30 QTL, total heritability $= 66\%$, individual heritability $[1.3\%, 3.4\%]$



Explanation

$$\hat{\beta} = \frac{Cov(Y,X)}{VarX}$$

Explanation

$$\hat{\beta} = \frac{Cov(Y,X)}{VarX}$$

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \epsilon$$

Explanation

$$\hat{\beta} = \frac{Cov(Y,X)}{VarX}$$

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \epsilon$$

$$Cov(Y, X_1) = \sum_{i=1}^{k} \beta_i Cov(X_1, X_i) + Cov(X_1, \epsilon)$$

Explanation

$\hat{\beta} = \frac{Cov(Y,X)}{VarX}$

$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \epsilon$

$Cov(Y, X_1) = \sum_{i=1}^{k} \beta_i Cov(X_1, X_i) + Cov(X_1, \epsilon)$

Assume that for $i > 1$, $Cov(X_1, X_i) \sim N(0, \sigma^2)$

Explanation

$\hat{\beta} = \frac{Cov(Y,X)}{VarX}$

$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \epsilon$

$Cov(Y, X_1) = \sum_{i=1}^{k} \beta_i Cov(X_1, X_i) + Cov(X_1, \epsilon)$

Assume that for $i > 1$, $Cov(X_1, X_i) \sim N(0, \sigma^2)$

$E(Cov(Y, X_1)) = \beta_1 \sigma_{X_1}^2$

Explanation

$\hat{\beta} = \frac{Cov(Y,X)}{VarX}$

$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \epsilon$

$Cov(Y, X_1) = \sum_{i=1}^{k} \beta_i Cov(X_1, X_i) + Cov(X_1, \epsilon)$

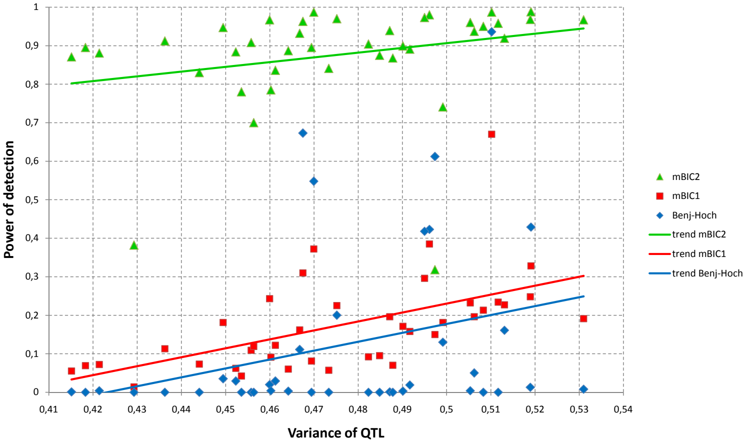
Assume that for $i > 1$, $Cov(X_1, X_i) \sim N(0, \sigma^2)$

$E(Cov(Y, X_1)) = \beta_1 \sigma_{X_1}^2$

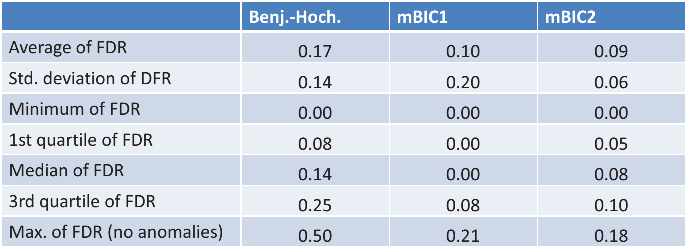
$Var[Cov(Y, X_1)] \approx \sum_{j=2}^{k} \beta_j \sigma^2$

Results

$h^2 \approx 80\%$



Results



	Benj.-Hoch.	mBIC1	mBIC2
Average of FDR	0.17	0.10	0.09
Std. deviation of DFR	0.14	0.20	0.06
Minimum of FDR	0.00	0.00	0.00
1st quartile of FDR	0.08	0.00	0.05
Median of FDR	0.14	0.00	0.08
3rd quartile of FDR	0.25	0.08	0.10
Max. of FDR (no anomalies)	0.50	0.21	0.18

References

1. Baierl, A., Bogdan, M., Frommlet, F., Futschik, A., 2006. On Locating multiple interacting quantitative trait loci in intercross designs. *Genetics* 173, 1693-1703.

2. Baierl, A., Futschik, A.,Bogdan, M.,Biecek, P., 2007. Locating multiple interacting quantitative trait loci using robust model selection, *Computational Statistics and Data Analysis* 51, 6423-6434.

3. Bogdan, M., Ghosh, J.K., Doerge, R.W., 2004. Modifying the Schwarz Bayesian Information Criterion to locate multiple interacting quantitative trait loci. *Genetics* 167, 989–999.

4. Bogdan, M., Frommlet, F., Biecek, P., Cheng, R., Ghosh, J. K., Doerge R. W. 2008 Extending the Modified Bayesian Information Criterion (mBIC) to dense markers and multiple interval mapping. *Biometrics*, doi: 10.1111/j.1541-0420.2008.00989.x.

5. Broman, K.W., Speed, T.P., 2002. A model selection approach for the identification of quantitative trait loci in experimental crosses. *J. Roy. Stat. Soc. B* 64, 641–656.

6. George, E.I., McCulloch, R.E., 1993. Variable Selection Via Gibbs Sampling. *J. Amer. Statist. Assoc.* 88 : 881-889.

7. Żak, M., Baierl, A., Bogdan, M., Futschik, A., 2007. Locating multiple interacting quantitative trait

loci using rank-based model selection. *Genetics* 176, 1845-1854.