Tail probability of linear combinations of chi-square variables and its application to influence analysis in QTL detection

Satoshi Kuriki and Xiaoling Dou
(Inst. Statist. Math., Tokyo)

ISM Cooperative Research Symposium:
Extreme value theory and applications

Fri 27 July, 2012

Contents of talk

1. Some theoretical results
2. Application to the statistical genetics.

Summary

1. Some theoretical results

Quadratic form of a Gaussian vector

- Canonical form:

$$T = \sum_{i=1}^{n} a_i \xi_i^2, \quad \xi_i \sim N(0,1) \ \text{ i.i.d.}$$

  Note that:
  - a_i's are not necessarily positive.
  - Some of a_i's take the same values.
- Our purpose is to obtain the tail probability:

$$\bar{F}(x) = \Pr(T > x) \quad (x \to \infty)$$

- We propose a PP-plot for this tail probability.
- For the case where the numbers of the same a_i's are all even, i.e., T is a linear combination of chi-square distributions with 2 d.f., see Imhof (1961, Biometrika).

The case where a_i's are positive

- Proposition 1
  Let $a_1 = \cdots = a_m > a_{m+1} \geq \cdots \geq a_n > 0$. Then,

$$\bar{F}(x) = \Pr\left(\sum_{i=1}^{n} a_i \xi_i^2 > x\right)$$

$$\sim \Pr\left(\chi_m^2 > \frac{x}{a_1}\right) \times \prod_{i \geq m+1} \left(1 - \frac{a_i}{a_1}\right)^{-\frac{1}{2}} \quad (x \to \infty)$$

  Note that

$$\Pr\left(\chi_m^2 > x\right) \sim \frac{1}{2^{\frac{m}{2}-1}\Gamma\left(\frac{m}{2}\right)} x^{\frac{m}{2}-1} e^{-\frac{x}{2}} \quad (x \to \infty)$$

- For $m = 1$, e.g., Beran (1975, AS)

An intuitive explanation of Proposition 1

- Let

$$a_1 = \cdots = a_m (= 1) > a_{m+1} \geq \cdots \geq a_n > 0$$

for simplicity. We want to prove

$$\bar{F}(x) \sim \frac{C}{\Gamma(\frac{m}{2})} x^{\frac{m}{2}-1} e^{-\frac{x}{2}} \quad (x \to \infty)$$

where

$$C = \frac{1}{2^{\frac{m}{2}-1}} \prod_{i \geq m+1} (1 - a_i)^{-\frac{1}{2}}$$

Equivalently

$$e^{\frac{x}{2}} \bar{F}(x) \sim \frac{C}{\Gamma(\frac{m}{2})} x^{\frac{m}{2}-1} \quad (x \to \infty)$$

- By Tauberian theorem, it suffices to show that

$$\int_0^\infty e^{-sx} e^{\frac{x}{2}} \bar{F}(x) dx \sim C s^{-\frac{m}{2}} \quad (s \to 0)$$

if the regularity condition (ultimate monotonicity) is ensured.

An intuitive explanation of Proposition 1 (contd)

- By integration by parts,

$$\begin{aligned}\text{LHS} &= \int_0^\infty e^{-(s-\frac{1}{2})x}\bar{F}(x)dx \\ &= \frac{-1}{s-\frac{1}{2}}\left[e^{-(s-\frac{1}{2})x}\bar{F}(x)\Big|_0^\infty + \int_0^\infty e^{-(s-\frac{1}{2})x}dF(x)\right] \\ &= \frac{1-\phi(s-\frac{1}{2})}{s-\frac{1}{2}}, \qquad \phi(s) = \int_0^\infty e^{-sx}dF(x)\end{aligned}$$

- Actually, in our case,

$$\phi(s) = E\left[e^{-s\sum a_i\xi_i^2}\right] = \prod(1+2sa_i)^{-\frac{1}{2}},$$

and

$$\begin{aligned}\text{LHS} &= \frac{1-\{1+2(s-\frac{1}{2})\}^{-\frac{m}{2}}\prod_{i\geq m+1}(1+2(s-\frac{1}{2})a_i)^{-\frac{1}{2}}}{s-\frac{1}{2}} \\ &\sim \frac{s^{-\frac{m}{2}}}{2^{\frac{m}{2}-1}}\prod_{i\geq m+1}(1-a_i)^{-\frac{1}{2}} = Cs^{-\frac{m}{2}} = \text{RHS} \quad (s\to 0)\end{aligned}$$

An approach to prove Proposition 1

- Recall that

$$T = \sum_{i=1}^{n} a_i \xi_i^2, \quad \xi_i \sim N(0,1) \ \text{ i.i.d.}$$

- Define a Gaussian process on $\mathbb{S}^{n-1}$ (the set of unit vectors in $\mathbb{R}^n$) by

$$Z(h) = \sum_{i=1}^{n} h_i \sqrt{a_i}\, \xi_i, \quad h = (h_i) \in \mathbb{S}^{n-1}.$$

  Then,

$$\max_{h \in \mathbb{S}^{n-1}} Z(h) = \sqrt{T}.$$

- Various methods for approximating the tail probability of the maximum of a Gaussian process are applicable.

An approach to prove Proposition 1 (contd)

- One approach is Euler-characteristic heuristic (volume-of-tube method) is

$$\Pr\left(\max_{h\in\mathbb{S}^{n-1}} Z(h) \geq x\right) \sim E[\chi(A_x)] \quad (x \to \infty)$$

where

$$A_x = \left\{h \in \mathbb{S}^{n-1} \mid Z(h) \geq x\right\} \quad \text{(excursion set)}$$

$\chi(\cdot)$: Euler characteristic.

- Thanks to Morse's theorem (see, e.g., Worsley, 1995; K & Takemura, 2009),

$$E[\chi(A_x)] = \int_{\mathbb{S}^{n-1}} E\Big[1(Z(h) \geq x)\det\big(-\ddot{Z}(h)\big) \mid \dot{Z}(h) = 0\Big] \times \theta(0)\, d\mathbb{S}^{n-1}(h)$$

where $\theta(0)$ is the density function of $\dot{Z}(h)$ evaluated at $\dot{Z}(h) = 0$. Details are omitted.

The case where a_i's are not necessarily positive

- Proposition 2
  Let $a_1 = \cdots = a_m > a_{m+1} \geq \cdots \geq a_n > -\infty$. Then,

$$\bar{F}(x) = \Pr\Big(\sum_{i=1}^{n} a_i \xi_i^2 > x\Big)$$

$$\sim \Pr\Big(\chi_m^2 > \frac{x}{a_1}\Big) \times \prod_{i \geq m+1} \Big(1 - \frac{a_i}{a_1}\Big)^{-\frac{1}{2}} \quad (x \to \infty)$$

  Note that

$$\Pr\big(\chi_m^2 > x\big) \sim \frac{1}{2^{\frac{m}{2}-1}\Gamma(\frac{m}{2})} x^{\frac{m}{2}-1} e^{-\frac{x}{2}} \quad (x \to \infty)$$

- Of the same form as Proposition 1.

Proof of Proposition 2

- Assume that

$$a_1 = \cdots = a_m > a_{m+1} \geq \cdots > 0 > \cdots \geq b_{m'+1} > b_{m'} = \cdots = b_1$$

Let

$$T = \sum a_i \xi_i^2 - \sum |b_j| \, \xi_j^2 =: Y - Z$$

We evaluate

$$\bar{F}(x) = \Pr(T > x) = E^Z[\Pr(Y - Z > x \mid Z)] = E^Z[\bar{F}_Y(x+Z)]$$

where $\bar{F}_Y(x) = \Pr(Y > x)$

- Lemma
  Let Z be a nonnegative r.v. If $\bar{F}_1(x) \sim \bar{F}_2(x)$ $(x \to \infty)$, then $E^Z[\bar{F}_1(x+Z)] \sim E^Z[\bar{F}_2(x+Z)]$.

Proof of Proposition 2 (contd)

- Applying Lemma together with the result of Proposition 1

$$\bar{F}_Y(x) \sim D_m x^{\frac{m}{2}-1} e^{-\frac{x}{2a_1}} \times \prod_{i \geq m+1} \left(1 - \frac{a_i}{a_1}\right)^{-\frac{1}{2}}$$

$(D_m^{-1} = (2a_1)^{\frac{m}{2}-1}\Gamma(\frac{m}{2}))$, we have

$$\begin{aligned}
\bar{F}(x) &\sim D_m e^{-\frac{x}{2a_1}} E^Z\left[(x+Z)^{\frac{m}{2}-1} e^{-\frac{Z}{2a_1}}\right] \times \prod_{i \geq m+1} \left(1 - \frac{a_i}{a_1}\right)^{-\frac{1}{2}} \\
&\sim D_m e^{-\frac{x}{2a_1}} x^{\frac{m}{2}-1} E^Z\left[e^{-\frac{Z}{2a_1}}\right] \times \prod_{i \geq m+1} \left(1 - \frac{a_i}{a_1}\right)^{-\frac{1}{2}} \\
&\sim \Pr\left(\chi_m^2 > \frac{x}{a_1}\right) \times \prod_{j \geq 1} \left(1 + \frac{|b_j|}{a_1}\right)^{-\frac{1}{2}} \prod_{i \geq m+1} \left(1 - \frac{a_i}{a_1}\right)^{-\frac{1}{2}}
\end{aligned}$$

□

Example

- Double exponential distribution:

$$f(x) = \frac{1}{2}e^{-|x|}$$

$$\bar{F}(x) = \begin{cases} \frac{1}{2}e^{-x} & (x \geq 0) \\ 1 - \frac{1}{2}e^{-|x|} & (x < 0) \end{cases}$$

- On the other hand,

$$T = Y - Z, \quad Y, Z \sim \mathsf{Exp}(1)$$

$$= \sum_{i=1}^{4} a_i \xi_i^2, \quad (a_i) = \left(\frac{1}{2}, \frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}\right),$$

$$\bar{F}(x) \sim \Pr\left(\chi_2^2 > \left(\frac{x}{1/2}\right)\right) \times \left(1 - \frac{(-1/2)}{(1/2)}\right)^{-\frac{2}{2}} = \frac{1}{2}e^{-x} \quad (x \to \infty)$$

PP-plot

- Let
$$X_1, \ldots, X_N \sim \mathcal{L}\left(\sum \lambda_i \xi_i^2\right) \text{ i.i.d.}$$
Assume that
$$\lambda_{\max} = \max \lambda_i > 0 > \lambda_{\min} = \min \lambda_i,$$
the multiplicities of $\max \lambda_i$ and $\min \lambda_i$ are 1.
- The order statistics
$$X_{(1)} < \cdots < X_{(N)}$$

PP-plot (contd)

- PP-plot:

$$\left(-2\log\bar{G}_1\left(\frac{X_{(i)}}{\lambda_{\max}}\right)+\log\prod_{i\neq\max}\left(1-\frac{\lambda_i}{\lambda_{\max}}\right),-2\log\left(1-\frac{i}{N+1}\right)\right)$$

for i such that $X_{(i)}>0$

$$\left(2\log\bar{G}_1\left(\frac{|X_{(i)}|}{|\lambda_{\min}|}\right)-\log\prod_{i\neq\min}\left(1-\frac{\lambda_i}{\lambda_{\min}}\right),2\log\left(1-\frac{i}{N+1}\right)\right)$$

for i such that $X_{(i)}<0$

where $\bar{G}_1(x)=\Pr\left(\chi_1^2>x\right)$

2. Application to the influence analysis in QTL detection

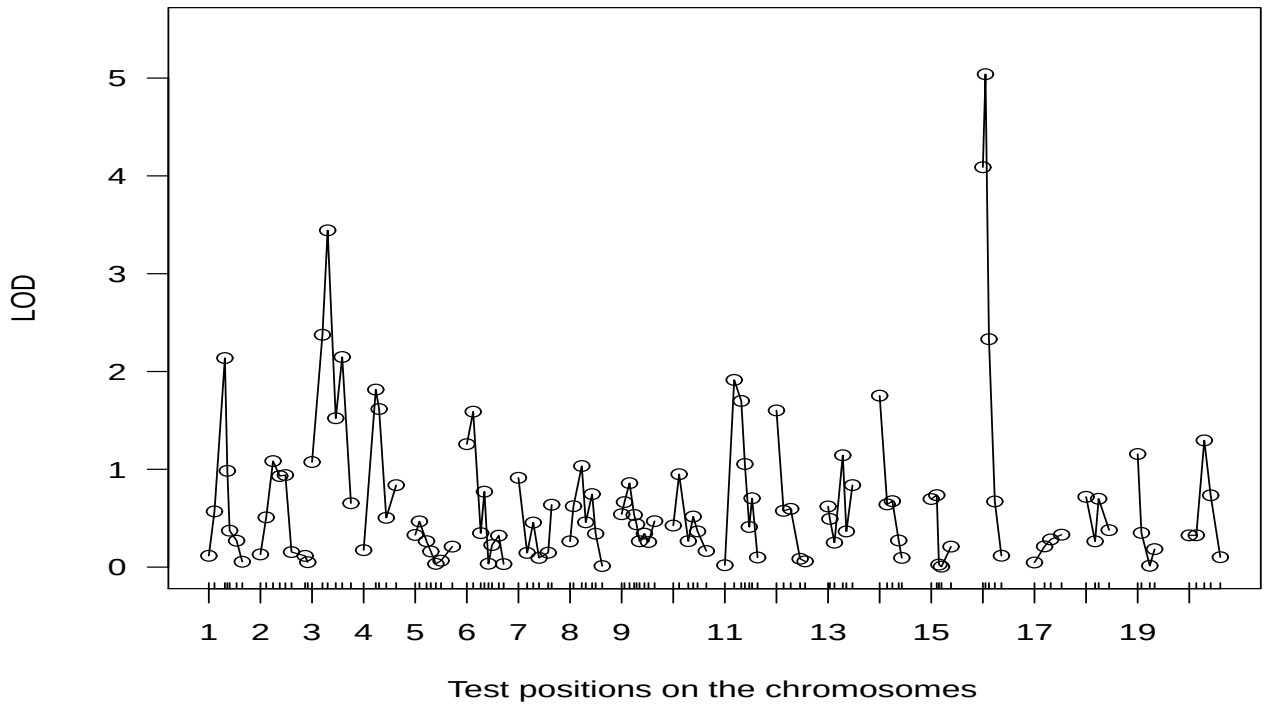
What is QTL analysis?

- N individuals (e.g., mice) data:

#	phenotype	genotype
1	y_1	$z_{i1}, \ldots, z_{1M}$
$\vdots$	$\vdots$	$\ldots$
i	y_i	$z_{i1}, \ldots, z_{iM}$
$\vdots$	$\vdots$	$\ldots$
N	y_N	$z_{N1}, \ldots, z_{NM}$

- Phenotype y_i: The measurement of interesting feature of individual i.
- Genotype z_{ij}: The type of gene at the locus j of individual i.
- Purpose of the analysis: To identify j (index of loci) such that z_{ij} is highly "correlated" to y_i.
  Such locus j is called QTL.

LOD Score



- H_j (QTL at j) : $y_i \sim N(\mu + \alpha z_{ij}, \sigma^2)$
  H_0 (no QTL) : $y_i \sim N(\mu, \sigma^2)$

$$\mathrm{LOD}(j) = \mathrm{const} \times \log \frac{\widehat{\sigma}^2_{(H_0)}}{\widehat{\sigma}^2_{(H_j)}} \quad \text{(LRT } H_0 \text{ vs. } H_j\text{)}$$

Influence function

- Empirical influence function of $\mathrm{LOD}(j)$ for the individual i (Dou, et al., 2012):

$$\mathrm{EIF}_i(j) = \mathrm{const} \times \left\{ \frac{\widehat{\varepsilon}^2_{i\,(H_0)}}{\widehat{\sigma}^2_{(H_0)}} - \frac{\widehat{\varepsilon}^2_{i\,(H_j)}}{\widehat{\sigma}^2_{(H_j)}} \right\}$$

  where

$$\begin{aligned} \widehat{\varepsilon}_{i\,(H_j)} &= y_i - \widehat{\mu}_{(H_j)} - \widehat{\alpha}_{(H_j)} z_{ij} \\ \widehat{\varepsilon}_{i\,(H_0)} &= y_i - \widehat{\mu}_{(H_0)} \end{aligned}$$

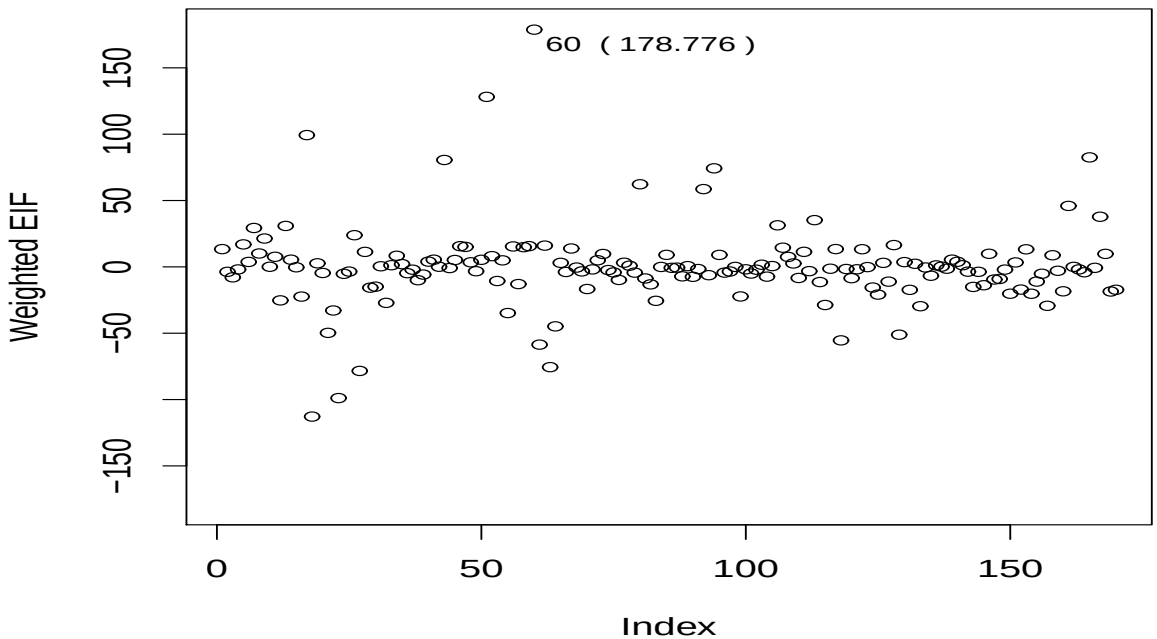
  are residuals under H_j and H_0.
- EIF of the weighted LOD score $\sum_{j \in J} c_j \mathrm{LOD}(j)$:

$$\sum_{j \in J} c_j \mathrm{EIF}_i(j) \quad \text{(Weighted EIF)}$$

  Available for detecting individuals that affect the shape of LOD score specified by the coefficients (c_i).

Influence function (contd)

- $(c_j) = (1.042, -2.356, 1.314)$



We want to make sure whether No. 60 mouse is influential.

Influence function (contd)

- Approximation:
  Suppose that in

$$\mathrm{EIF}_i(j) = \mathrm{const} \times \left\{ \frac{\widehat{\varepsilon}^2_{i\,(H_0)}}{\widehat{\sigma}^2_{(H_0)}} - \frac{\widehat{\varepsilon}^2_{i\,(H_j)}}{\widehat{\sigma}^2_{(H_j)}} \right\},$$

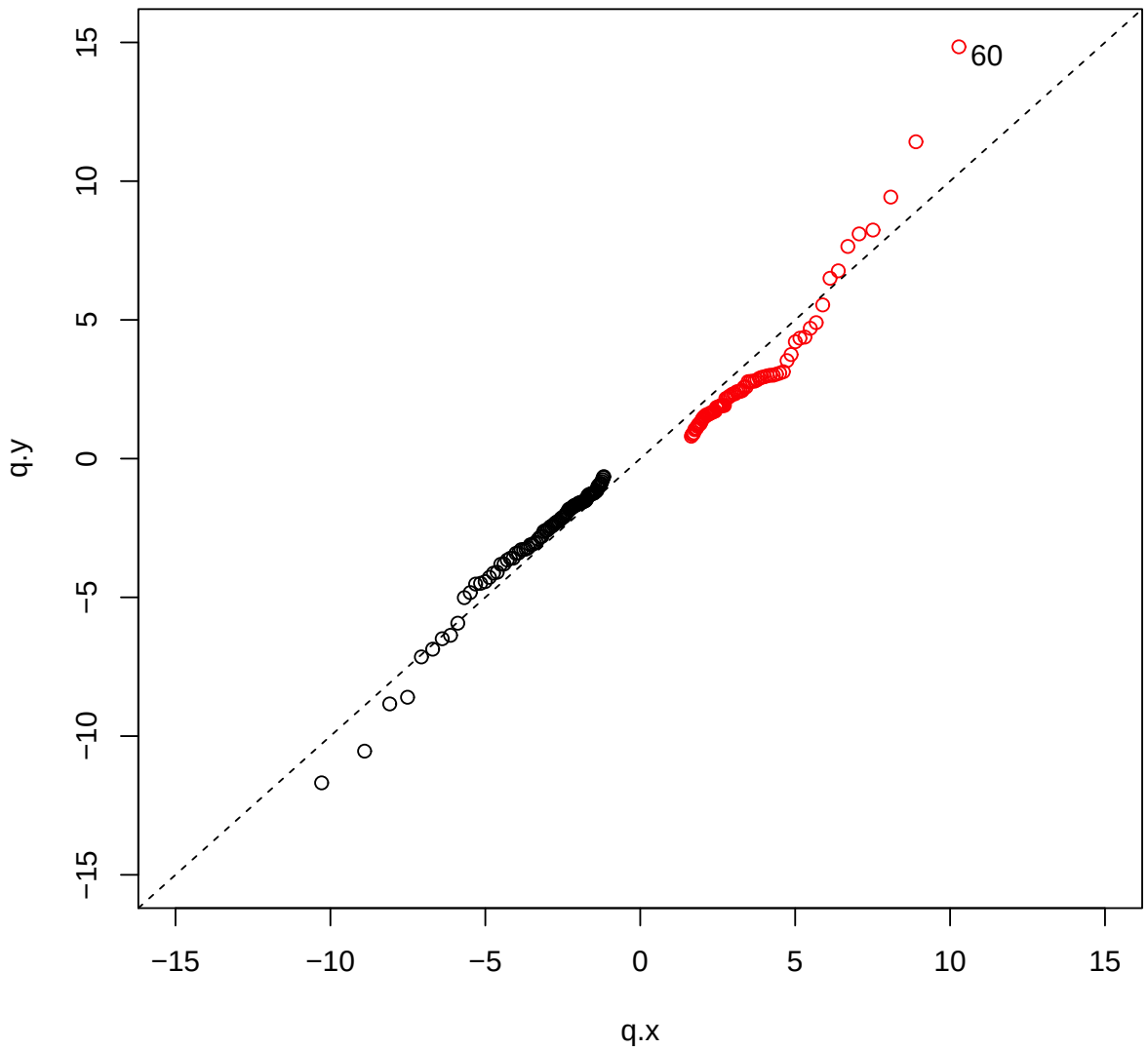
  $\widehat{\varepsilon}_{(H_j)}$ and $\widehat{\varepsilon}_{(H_0)}$ are Gaussian random variables, and $\widehat{\sigma}^2_{(H_j)}$, $\widehat{\sigma}^2_{(H_0)}$ are constants. Then,

$$\sum_{j \in J} c_j \mathrm{EIF}_i(j) \overset{d}{\approx} \sum_{j \in J} a_j \xi_j^2, \quad \xi_j \sim N(0,1)$$

- $(a_j) = (16.143, -2.629, -13.514)$

PP-plot

- The PP-plot suggests that No. 60 is influential.



Concluding remarks

- In Propositions 1 and 2, we provide the upper tail probability formula for a linear combination of chi-square random variables (a quadratic form of a Gaussian vector).
- We applied PP-plot to the influence analysis in QTL detection.
- We want to extend our result to the case where the number n of terms in T is infinite.
- Acknowledgment:
The authors thank Hsien-Kuei Hwang for his comments on the original version of slides.