Detecting Rare and Faint Signals via Thresholding Maximum Likelihood Estimators

Song Xi Chen

Peking University

A joint work with Yumou Qiu and Dan Nettleton
Detecting Rare and Faint Nonzero Means

- \( X_1, \cdots, X_n \overset{i.i.d.}{\sim} N(\mu, I_p \times p) \), \( p \) dimensional random vectors
- \( \mu = \mathbb{E}(X_i) = (\mu_1, \cdots, \mu_p)' \)

- Donoho and Jin (2004) test for zero mean against rare and faint \( H_a \)

\[
H_0 : \mu_j = 0 \quad \text{for all } j \quad \text{vs.} \quad H_a : \mu_j \sim (1 - \epsilon)\nu_0 + \epsilon\nu_{\mu_a}
\]

- Rareness: few dimensions with nonzero mean, \( \epsilon = p^{-\kappa}, \kappa \in (1/2, 1) \)
- Faintness: weak nonzero mean, \( \mu_a = \sqrt{(2r \log p)/n}, \quad r \in (0, 1) \)

- Most challenging situation for signal detection
Optimal Detection Boundary

\[
DB(\kappa) = \begin{cases} 
\kappa - \frac{1}{2} & \text{if } \frac{1}{2} < \kappa \leq \frac{3}{4}, \\
\frac{\kappa - \frac{1}{2}}{(1 - \sqrt{1 - \kappa})^2} & \text{if } \frac{3}{4} < \kappa < 1, 
\end{cases}
\]

Phase Diagram

- Ingster (1997)
- \( r < DB(\kappa) \): Type I Error + Type II Error → 1 for ANY TEST
- \( r > DB(\kappa) \): there exists a test such that Type I Error + Type II Error → 0
Donoho and Jin (2004) proposed Higher Criticism (HC) test

\[ \{ p(j) \} : \text{ordered p-values for testing each dimension} \]

\[ HC^* = \max_{1 \leq j \leq p/2} \sqrt{p[j/p - p(j)]/[p(j)(1 - p(j))]}^{1/2} \]

Reject \( H_0 \) if \( HC^* > \sqrt{2 \log \log p} \)

HC test is able to attain the optimal detection boundary
Further Works on Testing Means

  - Apply HC on $\hat{\Sigma}^{-1}X_i$

- Zhong, Chen and Xu (2013): $L_2$ thresholding test, more powerful
  - $M_n(s) = \sum_{j=1}^{p} n(\bar{X}_j/s_j)^2 I(|\bar{X}_j|/s_j > \sqrt{(2s \log p)/n})$
  - $\mathcal{M}_n = \max_{0<s<1}\{M_n(s) - \mu_n(s)\}/\sigma_n(s)$

- Chen, Li and Zhong (2016): thresholding with data transformation
  - Apply $L_2$ thresholding on $\hat{\Sigma}^{-1}X_i$

- Fan (1996): goodness-of-fit test
Settings of the Problem

- Collect $p$ responses variables from $n$ observations under $m$ treatments
- $y_{ij}$: the value of the $j$th response variable in the $i$th observation
- $i = 1, \cdots, n$ and $j = 1, \cdots, p$
- Let $z_i$ be the explanatory variables for the $i$th observation
- $z_i = (z_{i,1}, \cdots, z_{i,m})'$

$m$ is the number of explanatory variables, **fixed**
Suppose $y_{ij}$ is within the exponential family

$\mu_{ij} = E(y_{ij})$ and $\phi_j$ is the dispersion parameter

**Link function:** $g(\mu_{ij}) = \eta_{ij} = z_i' \beta_j$

$\beta_j = (\beta_{j,1}, \cdots, \beta_{j,m})'$: treatment effect for the $j$th response variable
Any Signal in the Regression Coefficient?

- Interested in testing the hypotheses
  \[ H_0 : D\beta_j = 0 \text{ for all } j \text{ vs. } H_a : \text{not all the } D\beta_j = 0. \]
- \( D_{d \times m} \) is a known matrix
Maize RNA-seq Experiment

- Maize gene expressions are measured by RNA-Seq
- 4 genotypes: B73, Mo17, B73×Mo17 (BM) and Mo17×B73 (MB)
- 4 tissues: cortex, stele, elongation zone, meristematic zone
- 4 barcodings: AR001, AR003, AR008 and AR009
- 4 replications
Genotypes and Tissues of Corn Root

- B73
- B73x Mo17
- Mo17xB73
- Mo17

- Cortex
- Stele
- Differentiation zone
- Elongation zone
- Meristematic zone

Primary root
Strip-plot Experimental Design

| B73-1-S-AR001 | Mo17-1-S-AR003 | BxM-1-S-AR008 | MxB-1-S-AR009 |
| B73-1-MZ-AR001 | Mo17-1-MZ-AR003 | BxM-1-MZ-AR008 | MxB-1-MZ-AR009 |
| B73-1-C-AR001  | Mo17-1-C-AR003  | BxM-1-C-AR008  | MxB-1-C-AR009  |
| B73-1-EZ-AR001 | Mo17-1-EZ-AR003 | BxM-1-EZ-AR008 | MxB-1-EZ-AR009 |
| B73-2-S-AR009  | Mo17-2-S-AR001  | BxM-2-S-AR003  | MxB-2-S-AR008  |
| B73-2-EZ-AR009 | Mo17-2-EZ-AR001 | BxM-2-EZ-AR003 | MxB-2-EZ-AR008 |
| B73-2-C-AR009  | Mo17-2-C-AR001  | BxM-2-C-AR003  | MxB-2-C-AR008  |
| B73-2-MZ-AR009 | Mo17-2-MZ-AR001 | BxM-2-MZ-AR003 | MxB-2-MZ-AR008 |

- S = stele (differentiation zone)
- C = cortex (differentiation zone)
- MZ = meristematic zone
- EZ = elongation zone

AR001-AR009: Adapter indices

- Replicate 1
- Replicate 2
- Replicate 3
- Replicate 4
Consider the tissue-specific analysis for the stele tissue

Denote $Y_j = (y_{1,j}, \cdots, y_{16,j})$ to be the expression level of the $j$th gene

NB model for $y_{i,j}$ with mean $\mu_{i,j}$ and dispersion parameter $\phi_j$ such that

$$\log(\mu_{i,j}) = \nu_j + X'_{g,i}\alpha_j + X'_{r,i}\xi_j + X'_{b,i}\gamma_j,$$

$\alpha_j, \xi_j, \gamma_j$: treatment effects for genotype, replication and barcoding

Satisfying $\alpha_{j,4} = \beta_{j,4} = \gamma_{j,4} = 0$, baseline
Hypotheses

- Any of the genes is differentially expressed across different genotypes?

- Genotype effect:

  \[ H_0 : \alpha_{1,j} = \alpha_{2,j} = \alpha_{3,j} = 0 \quad \text{for all} \ j \quad \text{vs.} \]
  \[ H_a : \text{at least one component of } \alpha_j \text{ not equal to 0 for some } j. \]

- Difference between B73×Mo17 (BM) and Mo17×B73 (MB):

  \[ H_0 : \alpha_{2,j} = \alpha_{3,j} \quad \text{for all} \ j \quad \text{vs.} \quad H_a : \alpha_{2,j} \neq \alpha_{3,j} \quad \text{for some } j. \]
Wald’s Test for each gene

Many small P-values

Strong evidence in favor of $H_a$

Some of the genes are differentially expressed under some genotypes
P-Value of the Difference Between BM and MB

How to detect rare and faint genetic signals in RNA-seq data?

- Few small P-values
- Not have strong evidence to reject $H_0$ based on this histogram
- If $H_a$ is true, few genes have different BM and MB genetic effects
Sparse and Weak $H_a$

Consider the hypotheses with rare and faint signals

$$H_0 : \beta_j = \beta_{j,0} \text{ such that } D\beta_{j,0} = 0 \text{ for all } j \text{ vs. }$$

$$H_a : \beta_j \sim (1 - \epsilon)\nu_{\beta_{j,0}} + \epsilon\nu_{\beta_{j,0} + \beta_{j,a}} \text{ for all } j,$$

- For $x \in \mathbb{R}^m$, $\nu_x$ denotes the point mass distribution on $x$
- Under $H_a$, $\beta_j$ comes from a binary super-population
- $\epsilon = p^{-\kappa}$ for $\kappa \in (1/2, 1)$, $\kappa$ specify the sparsity
- $\beta_{j,a} = r_j \sqrt{2(\log p)/n}$ for $r_j \in (0, 1)^m$ and $Dr_j \neq 0$
- $\{r_j\}$ specify the signal strength
Regularize MLEs by Thresholding

- \((\hat{\beta}_j', \hat{\phi}_j)\)': MLEs to the \(j\)th response variable

- \(\hat{\beta}_j = (\hat{\beta}_{j,1}, \cdots, \hat{\beta}_{j,m})'\) the estimated treatment effects

- \(D\hat{\beta}_j\) MLE of \(D\beta_j\)

- Most of the responses have no signal

- Truncate those small \(\|D\hat{\beta}_j\|\) to 0 and keep the large ones
Thresholding Statistic

- Let $nI_j$ be the Fisher Information matrix of $\beta_j$
- $\hat{I}_j = I_j(\hat{\beta}_j, \hat{\phi}_j)$ is the estimated Fisher Information matrix
- $\hat{V}_j/n = D\hat{I}_j^{-1}D'/n$ is the estimated asymptotic variance of $D\hat{\beta}_j$
- For a constant $s \in (0, 1)$, the thresholding statistic is

$$T_n(s) = \sum_{j=1}^{p} n(D\hat{\beta}_j)'\hat{V}_j^{-1}(D\hat{\beta}_j)I(\|\hat{V}_j^{-1/2}D\hat{\beta}_j\| > \sqrt{(2s \log p)/n})$$
Lemma 1. Suppose $X_1, \cdots, X_n$ are independent with pdf $f_i(x_i, \theta)$

- $\theta$ is $m$ dimensional unknown parameter, $\theta_0$ is the truth
- $\hat{\theta}$ be the MLE of $\theta$
- Under some regularity conditions, for $w_n \ll n^{1/6}$
  - The 2nd derivative of $\log f_i(x_i, \theta)$ is Lipschitz continuous
  - MGFs of the 1st and 2nd derivative of $\log f_i(X_i, \theta)$ exist around 0

$$P\left(\sqrt{n}\hat{\theta} - \theta_0 \geq w_n\right) = P\left(|N(0, I_m)| \geq w_n\right)(1 + O(w_n^3/\sqrt{n})).$$
Cramér moderate deviation for MLE under non identically distr data

Tail distribution of MLEs can be approximated by that of $N(0, I_m)$

For i.i.d. data, Inglot and Kallenberg (2003)

$$\lim_{n \to \infty} w_n^{-2} \log \left\{ P \left( \sqrt{n} |l_\theta^{1/2}(\hat{\theta} - \theta^0)| \geq w_n \right) \right\} = -1/2$$

We provide the error rate of the normal approximation

Needed for the analysis of the thresholding approach
Let \( Y_i = (y_{i1}, \ldots, y_{ip})' \) and \( \mathcal{F}_a^b(Y_i) = \sigma\{Y_{ij} : a \leq j \leq b\} \)

\( \rho \)-mixing coefficients: \( \rho_i(k) = \sup_{m \in \mathbb{Z}} \rho\{\mathcal{F}_{-\infty}^m(Y_i), \mathcal{F}_m^\infty(Y_i)\} \).

Assume \( Y_i \) is a \( \rho \)-mixing sequence, and

\[ \rho_i(k) \leq C \alpha^k \quad \text{for a constant } \alpha \in (0, 1) \]

Only require the condition holds for some ordering of response variables.
Mean and Variance of $T_n(s)$ under $H_0$

- Tail distribution of $\sqrt{n} \hat{V}_j^{-1/2} D \hat{\beta}_j$ be approximated by that of $N(0, I_d)$
- Write $\lambda_n(s) = 2s \log p$, assuming $\log p = o(n^{1/3})$
- $\mu_0(s) = pE\{\chi_d^2 \mathbb{I}(\chi_d^2 \geq \lambda(s))\}$
- $\sigma_0^2(s) = p\text{Var}\{\chi_d^2 \mathbb{I}(\chi_d^2 \geq \lambda(s))\}$
- $E\{T_n(s)\mid H_0\} = \mu_0(s)\{1 + O(\lambda_n(s)^{3/2} / \sqrt{n})\}$
- $\text{Var}\{T_n(s)\mid H_0\} = \sigma_0^2(s)\{1 + o(1)\}$
Theorem 1. Under $H_0$ and some regularity conditions, if $\log p = o(n^{1/3})$,

$$\frac{T_n(s) - \mathbb{E}\{T_n(s) | H_0\}}{\sqrt{\text{Var}\{T_n(s) | H_0\}}} \xrightarrow{d} N(0, 1) \text{ as } n, p \to \infty.$$ 

- Reject $H_0$ if $T_n(s) - \mu_0(s) > \sigma_0(s)z_\alpha$
Multi-level Thresholding

- Standardized thresholding statistic
  \[ \hat{T}_n(s) = \sigma_0(s)^{-1} \{ T_n(s) - \hat{\mu}(s) \} \]

- For a small positive constant \( \omega \), maximize \( \hat{T}_n(s) \) over \( \mathcal{S}_n \)
  \[ \mathcal{S}_n = \{ s_j : s_j = n(D\hat{\beta}_j)' \hat{V}_j^{-1}(D\hat{\beta}_j)/(2 \log p), s_j \leq 1 - \omega, j \leq p \} \]

- Let \( \mathcal{T}_n = \max_{s \in \mathcal{S}_n} \hat{T}_n(s) \)
Multi-level Thresholding Test (MTT)

**Theorem 2.** Under $H_0$, some regularity conditions and $\log p = o(n^{1/3})$

\[ P\left( a_p T_n - b_p(\omega) \leq x \right) \rightarrow \exp\{-\exp(-x)\} \]

where $a_p$ and $b_p(\omega)$ are multi-log($p$) terms.

- $g_\alpha$: the upper $\alpha$ quantile of Gumbel distribution
- Reject $H_0$ if $T_n > a_p^{-1}(g_\alpha + b_p(\omega))$
Power Analysis

\[ H_0 : \beta_j = \beta_{j,0} \text{ such that } D\beta_{j,0} = 0 \text{ for all } j \text{ vs.} \]

\[ H_a : \beta_j \sim (1 - \epsilon)\nu_{\beta_{j,0}} + \epsilon\nu_{\beta_{j,0} + \beta_{j,a}} \text{ for all } j, \]

- \( \beta_{j,a} = r_j \sqrt{2(\log p)/n} \) and \( \epsilon = p^{-\kappa} \)

- the standardized signal strength:

\[ r_s = \max_j r_j' D' V_j^{-1} D r_j \text{ where } V_j = D I_j^{-1} D' \]
Detection Boundary of Multi-level Thresholding (MTT)

\[ DB(\kappa) = \begin{cases} 
\kappa - \frac{1}{2} \\
(1 - \sqrt{1 - \kappa})^2 
\end{cases} \quad \text{if } \frac{1}{2} < \kappa \leq \frac{3}{4}, \]
\[ \text{if } \frac{3}{4} < \kappa < 1, \]

Phase Diagram

- For type I error rates \( \alpha_n \to 0 \) slowly as \( n \to \infty \)
- If \( r_s < DB(\kappa) \), power \( \to 0 \)
- If \( r_s > DB(\kappa) \), power \( \to 1 \)
Detection Lower Bound - linear model

- Linear regression model
  \[ y_{ij} = z_i'\beta_j + \varepsilon_{ij} \quad \text{for} \quad \varepsilon_{ij} \sim_{i.i.d.} N(0, \sigma^2). \]

- Design matrix \( Z = (z_1, \cdots, z_n)' \)

- \( r_s = \lim_{n \to \infty} \max_j r_j' D' \{ D(Z'Z)^{-1} D' \}^{-1} Dr_j / (n\sigma^2) \)

- If \( r_s < DB(\kappa) \), Type I Error + Type II Error \( \to 1 \) for ANY TEST

- MTT can attain the optimal detection boundary
Detection Lower Bound - general models

- \( I_\beta = - \lim_{n \to \infty} \sum_{i=1}^{n} E \frac{\partial^2}{\partial \beta \partial \beta'} \log f(y_{ij}; z_i, \theta) / n \)

- \( r_0 = \max_j r_j I_\beta r_j \)

- \( r_s = \max_j r_j D' V_j^{-1} D r_j \) for \( V_j = D I_j^{-1} D' \)

- If \( r_0 < \text{DB}(\kappa) \), Type I Error + Type II Error \( \to 1 \) for ANY TEST

- \( r_0 \geq r_s \) with equality if \( \phi_j \) is known and \( D \) takes the identity matrix

- MTT can attain the optimal detection boundary under this case
Signal Identification

- Which genes are differentially expressed?

- Multiple testing: $H_{j,0} : D\beta_j = 0$ vs. $H_{j,a} : D\beta_j \neq 0$

- FDP: proportion of falsely rejected among all rejected nulls

- FDR: expectation of FDP

- Exceedance FDP rate: probability that FDP exceeds a specific value
Step-down Procedure

- Obtain p-value of Wald test for each dimension
- \( \pi(j) \): the dimension label of the \( j \)th smallest p-value
- \( W_j = \{\pi(j), \pi(j+1), \ldots, \pi(p)\} \)
- \( H_{W_j,0} : D_{=0} = 0 \) for all \( \ell \in W_j \)
- MTT on \( H_{W_j,0} \) for each \( W_j \) at level \( \alpha \) until no rejection
- \( J = \min\{j : T(W_j) \leq a_{p-j+1}^{-1}(g_{\alpha} + b_{p-j+1}(\omega))\} \)
- Step-down procedure rejects the first \( J - 1 \) significant dimensions
Augment the rejection set of the step-down procedure by rejecting the next \( \lceil (J - 1)c/(1 - c) \rceil \) significant dimensions.

Genovese and Wasserman (2006): maximum test for \( H_{W_j,0} \)

Let \( J^* = \min\{p, \lceil (J - 1)/(1 - c) \rceil \} \)

Reject the first \( J^* \) significant dimensions.
FDP Control and Power

- $p - p_0$: total number of signals; $S$: true positives
- Standardized signal strength homogenous across dimensions

**Theorem 3.** Under $H_a$, some regularity conditions and $\log p = o(n^{1/3})$

- **Proposed procedure:** $P(\text{FDP} > c) \leq \alpha$ as $n, p \to \infty$

- $S_{GW}/(p - p_0) \xrightarrow{p} 0$ when $r_s < 1$

- $S_{GW}/(p - p_0) \xrightarrow{p} 1$ at rate $p^{-(\sqrt{r_s} - \sqrt{\kappa})^2 + o(1)}$ when $r_s > 1$

- $S_{prop}/(p - p_0), S_{BH}/(p - p_0)$ converge to 1 at rate $p^{-(\sqrt{r_s} - \sqrt{\kappa})^2 + o(1)}$
Simulation

\[ H_0 : \beta_{j1} = \beta_{j2} \text{ for all } j \text{ vs. } H_a : \beta_{j2} \overset{i.i.d.}{\sim} (1 - \epsilon)\nu_{\beta_{j1}} + \epsilon\nu_{\beta_{j1} + \beta_a} \]

- Balanced designs with two treatments, \( z_i = (1, 0)' \) or \( (0, 1)' \)
- \( n = 40, \ p = 100, 400, 700, 1000 \), nominal size 5%
- \( \epsilon = p^{-\kappa} \) and \( \beta_a = \sqrt{(2r_a \log p)/n} \)
- \( \kappa = 0.6 \) (7, 11, 14 and 16) and 0.55 (8, 15, 19 and 22)
- \( r_a \) was chosen differently to make \( r_s \in (0, 1) \)
- \( \beta_{11}, \ldots, \beta_{p1} \) were randomly generated, and kept fixed in simulation
Poisson and Binomial Regression

**Poisson regression**

- Response $y_{ij}$ follow Poisson distribution with mean $\mu_{ij}$

- Log link: $\log(\mu_{ij}) = z_i^T \beta_j$ and $\phi = 1$

**Binomial regression**

- Response $y_{ij}$ follow Binomial($m_{ij}, p_{ij}$)

- Logistic link: $p_{ij} = \exp(z_i^T \beta_j) / \{\exp(z_i^T \beta_j) + 1\}$

- $m_{ij}$ is randomly chosen from the integers between 20 and 40.
Negative Binomial Regression

- \( E(y_{ij}) = \mu_{ij} \) and \( \text{Var}(y_{ij}) = \mu_{ij} + \frac{\mu_{ij}^2}{\phi_j} \)

- Log link: \( \log(\mu_{ij}) = z_i^T \beta_j \)

- The dispersion parameter \( \phi_j \overset{i.i.d.}{\sim} \text{Unif}(3, 5) \)

- The MLE \( \hat{\phi}_j \) over estimates \( \phi_j \Rightarrow \) size distortion

- Use parametric bootstrap to correct the bias of \( \hat{\phi}_j \)
Logistic-Normal Mixed Regression

- Single normal random effect

\[ y_{1ij} \sim \text{Bin}(m_{1ij}, p_{1ij}) \quad \text{and} \quad y_{2ij} \sim \text{Bin}(m_{2ij}, p_{2ij}), \]

\[ \text{Logit}(p_{1ij}) = \beta_{1j} + \gamma_{ij} \quad \text{and} \quad \text{Logit}(p_{2ij}) = \beta_{2j} + \gamma_{ij} \]

- \( \gamma_{ij} \sim N(0, \sigma_j^2) \) for \( i = 1, \cdots, n/2, \sigma_j = 0.1 \)

- \( m_{1ij}, m_{2ij} \) is randomly chosen from the integers between 20 and 40.
Poisson and Negative Binomial Regression

Under $H_0$, $\{\beta_j\}_{j=1}^p \overset{i.i.d.}{\sim} \text{unif}(2, 5)$

Poisson regression, $n = 40$

Negative binomial regression, $n = 40$
Binomial and Logistic-Normal Regression

Under $H_0$, $\{\beta_j\}_{j=1}^p \overset{i.i.d.}{\sim} \text{unif}(-1, 2.5)$

Binomial regression, $n = 40$

Logistic normal regression, $n = 40$
Power Comparison

multi-level thresholding, HC and minimum p-value test for $H_0$ under Poisson

$n = 20, p = 1000$

Index = (Number of Signals/10,10r_a)

Power

$n = 40, p = 1000$

Index = (Number of Signals/10,10r_a)

Power
Signal Identification: FDP and False Negatives

Negative binomial under $H_a$ with $n = 40$ and $p = 10000$

**False Discovery Rate**

- **Exceedance FDP Rate**

- **False Negative**

- **Non-Discovery Proportion**

Maize RNA-seq Study

- NB model $\log(\mu_{i,j}) = \nu_j + X'_{g,i} \alpha_j + X'_{r,i} \beta_j + X'_{b,i} \gamma_j$

- $\alpha_j$, $\beta_j$, $\gamma_j$: treatment effects for genotype, replication and barcoding

- Genotype effect:

  $H_0 : \alpha_{1,j} = \alpha_{2,j} = \alpha_{3,j} = 0$ for all $j$ vs.
  $H_a :$ at least one component of $\alpha_j$ not equal to 0 for some $j$.

- The difference between the genotypes $B73 \times Mo17$ and $Mo17 \times B73$:

  $H_0 : \alpha_{2,j} = \alpha_{3,j}$ for all $j$ vs. $H_a : \alpha_{2,j} \neq \alpha_{3,j}$ for some $j$. 
Multi-Threshold Test for the Overall Genotype Effect

- **Multi-Threshold Statistics**
  \[ T_n = 2826.1 \]

- At 5% nominal level,
  
  Reject \( H_0 \) if \( T_n > 3.09 \)

- Strong evidence in favor of \( H_a \)

- Consistent with the findings from the P-value histogram
Multi-Threshold Test for Difference between MB and BM

- Multi-Threshold Statistics
  \[ T_n = 22.5 \]

- Reject \( H_0 \) since \( T_n > 3.09 \)

- Some genes DE between genotypes MB and BM

- Identify 32 DE genes
Conclusion

- Rare and faint signal detection for count data
- Multi-level thresholding test under GLM
- Step down procedure for signal identification
- Extension to GLMM
**Reference**


