3.27pt

Selected topics in the Analysis of Large Dimensional Data

Małgorzata Bogdan

University of Wroclaw

Bologna, 14-17/05/2019

Outline

- Testing for Global Null and Multiple Testing

Outline

- Testing for Global Null and Multiple Testing
- Model selection in multiple regression - Information Criteria

Outline

- Testing for Global Null and Multiple Testing
- Model selection in multiple regression - Information Criteria
- Regularization techniques (1)

Outline

- Testing for Global Null and Multiple Testing
- Model selection in multiple regression - Information Criteria
- Regularization techniques (1)
- Regularization techniques (2)

Classical example: Principle Components Analysis

Classical example: Principle Components Analysis

$X_{n \times p}$ - data matrix

Analysis of Large Data

Classical example: Principle Components Analysis

$X_{n\times p}$ - data matrix

Assumption - $X = M + E$, where M is a low rank matrix representing the signal and E is a random noise

Analysis of Large Data

Classical example: Principle Components Analysis

$X_{n \times p}$ - data matrix

Assumption - $X = M + E$, where M is a low rank matrix representing the signal and E is a random noise

Goal - recovering M, separating signal from the noise

Classical example: Principle Components Analysis

$X_{n \times p}$ - data matrix

Assumption - $X = M + E$, where M is a low rank matrix representing the signal and E is a random noise

Goal - recovering M, separating signal from the noise

Purpose - understanding the biological/economical etc phenomena which generate the data, data compression (few basis vectors [principal components] may contain most of the information in the data), missing values imputation

Analysis of Large Data

Classical example: Principle Components Analysis

$X_{n\times p}$ - data matrix

Assumption - $X = M + E$, where M is a low rank matrix representing the signal and E is a random noise

Goal - recovering M, separating signal from the noise

Purpose - understanding the biological/economical etc phenomena which generate the data, data compression (few basis vectors [principal components] may contain most of the information in the data), missing values imputation

General goal of large data analysis - separating the signal from noise, identifying the low dimensional structure spanning the noisy data

Classical example: Principle Components Analysis

$X_{n\times p}$ - data matrix

Assumption - $X = M + E$, where M is a low rank matrix representing the signal and E is a random noise

Goal - recovering M, separating signal from the noise

Purpose - understanding the biological/economical etc phenomena which generate the data, data compression (few basis vectors [principal components] may contain most of the information in the data), missing values imputation

General goal of large data analysis - separating the signal from noise, identifying the low dimensional structure spanning the noisy data

Major problem - multiple comparisons, multiple testing (in PCA selection of nonzero singular values)

Identifying genes associated with cancer

$X_{n_1 \times p}$ - expressions of p genes for n_1 healthy individuals

Identifying genes associated with cancer

$X_{n_1 \times p}$ - expressions of p genes for n_1 healthy individuals

$Y_{n_2 \times p}$ - expressions of p genes for n_2 cancer patients

Identifying genes associated with cancer

$X_{n_1 \times p}$ - expressions of p genes for n_1 healthy individuals

$Y_{n_2 \times p}$ - expressions of p genes for n_2 cancer patients

Assumption: X_{ij} for $i = 1, \ldots, n_1$ are iid with $E(X_{ij}) = \mu_{1j}$ and $Var(X_{ij}) = \sigma^2_{1j} < \infty$

Identifying genes associated with cancer

$X_{n_1 \times p}$ - expressions of p genes for n_1 healthy individuals

$Y_{n_2 \times p}$ - expressions of p genes for n_2 cancer patients

Assumption: X_{ij} for $i = 1, \ldots, n_1$ are iid with $E(X_{ij}) = \mu_{1j}$ and $Var(X_{ij}) = \sigma_{1j}^2 < \infty$

Y_{ij} for $i = 1, \ldots, n_2$ are iid with $E(Y_{ij}) = \mu_{2j}$ and $Var(Y_{ij}) = \sigma_{2j}^2 < \infty$

Identifying genes associated with cancer

$X_{n_1 \times p}$ - expressions of p genes for n_1 healthy individuals

$Y_{n_2 \times p}$ - expressions of p genes for n_2 cancer patients

Assumption: X_{ij} for $i = 1, \ldots, n_1$ are iid with $E(X_{ij}) = \mu_{1j}$ and $Var(X_{ij}) = \sigma^2_{1j} < \infty$

Y_{ij} for $i = 1, \ldots, n_2$ are iid with $E(Y_{ij}) = \mu_{2j}$ and $Var(Y_{ij}) = \sigma^2_{2j} < \infty$

Gene j is associated with cancer if $\mu_{1j} \neq \mu_{2j}$

Identifying genes associated with cancer

$X_{n_1 \times p}$ - expressions of p genes for n_1 healthy individuals

$Y_{n_2 \times p}$ - expressions of p genes for n_2 cancer patients

Assumption: X_{ij} for $i = 1, \ldots, n_1$ are iid with $E(X_{ij}) = \mu_{1j}$ and $Var(X_{ij}) = \sigma^2_{1j} < \infty$

Y_{ij} for $i = 1, \ldots, n_2$ are iid with $E(Y_{ij}) = \mu_{2j}$ and $Var(Y_{ij}) = \sigma^2_{2j} < \infty$

Gene j is associated with cancer if $\mu_{1j} \neq \mu_{2j}$

We test $H_{0j} : \mu_{1j} = \mu_{2j}$ with a t-test $t_j = \frac{\bar{X}_{.j} - \bar{Y}_{.j}}{S(\bar{X}_{.j} - \bar{Y}_{.j})}$, where $S(\bar{X}_{.j} - \bar{Y}_{.j})$ is the estimate of the standard deviation of $\bar{X}_{.j} - \bar{Y}_{.j}$

Identifying genes associated with cancer

$X_{n_1 \times p}$ - expressions of p genes for n_1 healthy individuals

$Y_{n_2 \times p}$ - expressions of p genes for n_2 cancer patients

Assumption: X_{ij} for $i = 1, \ldots, n_1$ are iid with $E(X_{ij}) = \mu_{1j}$ and $Var(X_{ij}) = \sigma^2_{1j} < \infty$

Y_{ij} for $i = 1, \ldots, n_2$ are iid with $E(Y_{ij}) = \mu_{2j}$ and $Var(Y_{ij}) = \sigma^2_{2j} < \infty$

Gene j is associated with cancer if $\mu_{1j} \neq \mu_{2j}$

We test $H_{0j} : \mu_{1j} = \mu_{2j}$ with a t-test $t_j = \frac{\bar{X}_{.j} - \bar{Y}_{.j}}{S(\bar{X}_{.j} - \bar{Y}_{.j})}$, where $S(\bar{X}_{.j} - \bar{Y}_{.j})$ is the estimate of the standard deviation of $\bar{X}_{.j} - \bar{Y}_{.j}$

If n_1 and n_2 are large enough then $t_j \sim N(\mu_j, 1)$ with $\mu_j = \frac{\mu_{1j} - \mu_{2j}}{\sigma_{1j}/\sqrt{n_1} + \sigma_{2j}/\sqrt{n_2}}$ and $H_{0j} : \mu_j = 0$

Multiple testing (1)

$$X_i \sim N(\mu_i, 1), \quad i = 1, \ldots, p$$

Multiple testing (1)

$$X_i \sim N(\mu_i, 1), \quad i = 1, \dots, p$$

$$H_{0i} : \mu_i = 0 \quad \text{vs} \quad \mu_i \neq 0$$

Multiple testing (1)

$X_i \sim N(\mu_i, 1), \quad i = 1, \ldots, p$

$H_{0i} : \mu_i = 0 \quad \text{vs} \quad \mu_i \neq 0$

Reject H_{0i} when $|X_i| > c$

Multiple testing (1)

$X_i \sim N(\mu_i, 1), \quad i = 1, \ldots, p$

$H_{0i} : \mu_i = 0 \quad \text{vs} \quad \mu_i \neq 0$

Reject H_{0i} when $|X_i| > c$

Multiple comparison problem: if all μ_is are equal to zero than $max(|X_1|, \ldots, |X_p|) = \sqrt{2 \log p}(1 + o_p)$

Multiple testing (1)

$X_i \sim N(\mu_i, 1), \quad i = 1, \ldots, p$

$H_{0i} : \mu_i = 0 \quad \text{vs} \quad \mu_i \neq 0$

Reject H_{0i} when $|X_i| > c$

Multiple comparison problem: if all μ_is are equal to zero than $max(|X_1|, \ldots, |X_p|) = \sqrt{2 \log p}(1 + o_p)$

Thus to separate signal from noise we need $c = c(p) \to \infty$ as $p \to \infty$.

Testing for global null, Bonferroni procedure

$$X_i \sim N(\mu_i, 1), \quad i = 1, \ldots, p$$

Testing for global null, Bonferroni procedure

$X_i \sim N(\mu_i, 1), \quad i = 1, \ldots, p$

$H_{0i} : \mu_i = 0 \quad \text{vs} \quad \mu_i \neq 0$

Testing for global null, Bonferroni procedure

$X_i \sim N(\mu_i, 1), \quad i = 1, \dots, p$

$H_{0i} : \mu_i = 0 \quad \text{vs} \quad \mu_i \neq 0$

$$H_0 : \bigcap_{i=1}^{p} H_{0i}$$

Testing for global null, Bonferroni procedure

$X_i \sim N(\mu_i, 1), \quad i = 1, \ldots, p$

$H_{0i} : \mu_i = 0 \quad \text{vs} \quad \mu_i \neq 0$

$$H_0 : \bigcap_{i=1}^{p} H_{0i}$$

Bonferroni procedure: Reject H_0 when

$\max(|X_1|, \ldots, |X_p|) \geq \Phi^{-1}\left(1 - \frac{\alpha}{2p}\right) = c_{Bon}$

Testing for global null, Bonferroni procedure

$$X_i \sim N(\mu_i, 1), \quad i = 1, \dots, p$$
$$H_{0i} : \mu_i = 0 \quad \text{vs} \quad \mu_i \neq 0$$

$$H_0 : \bigcap_{i=1}^{p} H_{0i}$$

Bonferroni procedure: Reject H_0 when

$$\max(|X_1|, \dots, |X_p|) \geq \Phi^{-1}\left(1 - \frac{\alpha}{2p}\right) = c_{Bon}$$

Probability of type I error:

$$P_{H_0}\left(\bigcup_{j=1}^{p}\{|X_j| > c_{Bon}\}\right) \leq \sum_{j=1}^{p} P(\{|X_j| > c_{Bon}\}) = \alpha$$

Exact type I error of Bonferroni

Due to independence

$$
\begin{aligned}
P(\mathit{Type\ I\ Error}) &= 1 - P_{H_0}\left(\bigcap_{j=1}^{p}\{|X_j| < c_{Bon}\}\right) \\
&= 1 - \left(1 - \frac{\alpha}{p}\right)^p \to 1 - e^{-\alpha} = \alpha + o(\alpha)
\end{aligned}
$$

Exact type I error of Bonferroni

Due to independence

$$
\begin{aligned}
P(\mathit{Type\ I\ Error}) &= 1 - P_{H_0}\left(\bigcap_{j=1}^{p}\{|X_j| < c_{Bon}\}\right) \\
&= 1 - \left(1 - \frac{\alpha}{p}\right)^{p} \rightarrow 1 - e^{-\alpha} = \alpha + o(\alpha)
\end{aligned}
$$

$\alpha = 0.05 \quad, n = 30000, P(\mathit{Type\ I\ Error}) \approx 0.0488$

Needle in haystack

Needle in haystack: for some $i \in \{1, , p\}$, $\mu_i = \mu^p$ and for all $j \neq i$, $\mu_j = 0$

How long needle can be found ?

$$C_{Bon} = \sqrt{2 \log p}(1 + o_p)$$

How long needle can be found ?

$$C_{Bon} = \sqrt{2\log p}(1 + o_p)$$

If $\mu^p = (1+\epsilon)\sqrt{2\log p}$ then power of Bonferroni converges to 1

How long needle can be found ?

$$C_{Bon} = \sqrt{2\log p}(1 + o_p)$$

If $\mu^p = (1+\epsilon)\sqrt{2\log p}$ then power of Bonferroni converges to 1

If $\mu^p = (1-\epsilon)\sqrt{2\log p}$ then power of Bonferroni converges to $q(\alpha) < \alpha$.

How long needle can be found ?

$$C_{Bon} = \sqrt{2 \log p}(1 + o_p)$$

If $\mu^p = (1+\epsilon)\sqrt{2\log p}$ then power of Bonferroni converges to 1

If $\mu^p = (1-\epsilon)\sqrt{2\log p}$ then power of Bonferroni converges to $q(\alpha) < \alpha$.

Is there a test procedure which can find a shorter needle ?

Neyman-Person test for the needle in haystack problem

H_0: $x = (x_1, \ldots, x_p)$ has the likelihood $L_0(x)$

Neyman-Person test for the needle in haystack problem

H_0: $x = (x_1, \dots, x_p)$ has the likelihood $L_0(x)$

H_A: x has the likelihood $L_A(x)$

Neyman-Person test for the needle in haystack problem

H_0: $x = (x_1, \ldots, x_p)$ has the likelihood $L_0(x)$

H_A: x has the likelihood $L_A(x)$

Neyman-Pearson optimal test (maximal power for a given type I error) has the form

Reject H_0 for large values of $L(x) = \frac{L_A(x)}{L_0(x)}$

Neyman-Person test for the needle in haystack problem

H_0: $x = (x_1, \ldots, x_p)$ has the likelihood $L_0(x)$

H_A: x has the likelihood $L_A(x)$

Neyman-Pearson optimal test (maximal power for a given type I error) has the form

Reject H_0 for large values of $L(x) = \frac{L_A(x)}{L_0(x)}$

Bayesian model for "needle in haystack" problem

$$H_0 : \mu = (\mu_1, \ldots, \mu_p) = 0$$

$$H_A : \mu \sim \frac{1}{p} \sum_{i=1}^{p} \delta_{\mu_i}$$

$$\delta_{\mu_i} : P\left(\{\mu_i = \mu^p, \mu_j = 0 \text{ for } j \neq i\}\right) = 1$$

Interpretation: under H_A there is just one signal of known magnitude μ^p but we do not know where and assume a uniform distribution over $i \in \{1, \ldots, p\}$.

Neyman-Pearson test for the needle in haystack problem (2)

$$L(x) = \frac{1}{p} \sum_{i=1}^{p} e^{x_i \mu^p - \frac{1}{2}(\mu^p)^2}$$

$$L(x) = \frac{1}{p} \sum_{i=1}^{p} e^{x_i \mu^p - \frac{1}{2}(\mu^p)^2}$$

If $\mu^p = (1-\epsilon)\sqrt{2 \log p}$ then the power of Neyman-Pearson test at the significance level α converges to α as $p \to \infty$

Neyman-Pearson test for the needle in haystack problem (2)

$$L(x) = \frac{1}{p}\sum_{i=1}^{p} e^{x_i \mu^p - \frac{1}{2}(\mu^p)^2}$$

If $\mu^p = (1-\epsilon)\sqrt{2\log p}$ then the power of Neyman-Pearson test at the significance level α converges to α as $p \to \infty$

Interpretation - Bonferroni correction has asymptotically optimal detection region under the needle in haystack problem

Chi-square test (1)

Reject H_0 when $||X||^2 = \sum_{i=1}^{p} X_i^2 > \chi_p^2(1-\alpha)$

Chi-square test (1)

Reject H_0 when $||X||^2 = \sum_{i=1}^{p} X_i^2 > \chi_p^2(1-\alpha)$

Under H_0: $T = \frac{||X||^2 - p}{\sqrt{2p}} \rightarrow N(0,1)$

Chi-square test (1)

Reject H_0 when $||X||^2 = \sum_{i=1}^p X_i^2 > \chi_p^2(1-\alpha)$

Under H_0: $T = \frac{||X||^2 - p}{\sqrt{2p}} \rightarrow N(0,1)$

Under H_1: $||X||^2 = \sum_{i=1}(\mu_i + Z_i)^2$, $Z_i \sim N(0,1)$

Chi-square test (1)

Reject H_0 when $||X||^2 = \sum_{i=1}^{p} X_i^2 > \chi_p^2(1-\alpha)$

Under H_0: $T = \frac{||X||^2 - p}{\sqrt{2p}} \to N(0,1)$

Under H_1: $||X||^2 = \sum_{i=1}(\mu_i + Z_i)^2$, $Z_i \sim N(0,1)$

$$E(\mu_i + Z_i)^2 = \mu_i^2 + 1, \quad Var(\mu_i + Z_i)^2 = 4\mu_i^2 + 2$$

Chi-square test (1)

Reject H_0 when $||X||^2 = \sum_{i=1}^{p} X_i^2 > \chi_p^2(1-\alpha)$

Under H_0: $T = \frac{||X||^2 - p}{\sqrt{2p}} \rightarrow N(0,1)$

Under H_1: $||X||^2 = \sum_{i=1}(\mu_i + Z_i)^2$, $Z_i \sim N(0,1)$

$$E(\mu_i + Z_i)^2 = \mu_i^2 + 1, \quad Var(\mu_i + Z_i)^2 = 4\mu_i^2 + 2$$

$$E\left(||X||^2\right) = ||\mu||^2 + p, \quad Var\left(||X||^2\right) = 4||\mu||^2 + 2p$$

Chi-square test (1)

Reject H_0 when $||X||^2 = \sum_{i=1}^{p} X_i^2 > \chi_p^2(1-\alpha)$

Under H_0: $T = \frac{||X||^2 - p}{\sqrt{2p}} \rightarrow N(0,1)$

Under H_1: $||X||^2 = \sum_{i=1}(\mu_i + Z_i)^2$, $Z_i \sim N(0,1)$

$$E(\mu_i + Z_i)^2 = \mu_i^2 + 1, \quad Var(\mu_i + Z_i)^2 = 4\mu_i^2 + 2$$

$$E\left(||X||^2\right) = ||\mu||^2 + p, \quad Var\left(||X||^2\right) = 4||\mu||^2 + 2p$$

$$\text{If } \ ||\mu||^2/p \rightarrow 0 \ \text{ then } \ \frac{||X||^2 - (p + ||\mu^2||)}{\sqrt{2p + 4||\mu||^2}} \rightarrow N(0,1)$$

Chi-square test (2)

For large p

$$T \sim N\left(\theta, 1 + \frac{\theta}{\sqrt{p/8}}\right) \text{ with } \theta = \frac{||\mu||^2}{\sqrt{2p}}$$

Chi-square test (2)

For large p

$$T \sim N\left(\theta, 1 + \frac{\theta}{\sqrt{p/8}}\right) \text{ with } \theta = \frac{||\mu||^2}{\sqrt{2p}}$$

Power of the chi-square test converges to 1 if $\frac{||\mu||^2}{\sqrt{p}} \to \infty$

Chi-square test (2)

For large p

$$T \sim N\left(\theta, 1 + \frac{\theta}{\sqrt{p/8}}\right) \text{ with } \theta = \frac{||\mu||^2}{\sqrt{2p}}$$

Power of the chi-square test converges to 1 if $\frac{||\mu||^2}{\sqrt{p}} \to \infty$

Power of the chi-square test converges to α if $\frac{||\mu||^2}{\sqrt{p}} \to 0$

Can we do better ?

Bayesian model for equally distributed signals:

$H_0 : \mu = 0, \quad H_A : \mu \sim \pi_\rho$- uniform distribution on the sphere of radius ρ

Can we do better ?

Bayesian model for equally distributed signals:

$H_0 : \mu = 0, \quad H_A : \mu \sim \pi_\rho$- uniform distribution on the sphere of radius ρ

Interpretation: We know L_2 norm of the vector of means under alternative but we do know how it is distributed between elements of this vector, so we assume a uniform distribution on the sphere

Can we do better ?

Bayesian model for equally distributed signals:

$H_0 : \mu = 0, \quad H_A : \mu \sim \pi_\rho$- uniform distribution on the sphere of radius ρ

Interpretation: We know L_2 norm of the vector of means under alternative but we do know how it is distributed between elements of this vector, so we assume a uniform distribution on the sphere

$$\mu = \rho u, \;\; L_A(X) = \int_{S^{p-1}} L(X|u)\pi(du)$$

Can we do better ?

Bayesian model for equally distributed signals:

$H_0 : \mu = 0, \quad H_A : \mu \sim \pi_\rho$- uniform distribution on the sphere of radius ρ

Interpretation: We know L_2 norm of the vector of means under alternative but we do know how it is distributed between elements of this vector, so we assume a uniform distribution on the sphere

$$\mu = \rho u, \;\; L_A(X) = \int_{S^{p-1}} L(X|u)\pi(du)$$

Neyman-Pearson test:

$$L = \int_{S^{p-1}} \frac{e^{\left(-\frac{1}{2}||x-\rho u||^2\right)}}{exp\left(-\frac{1}{2}||x||^2\right)}\pi(du)$$

Can we do better ?

Bayesian model for equally distributed signals:

$H_0 : \mu = 0, \quad H_A : \mu \sim \pi_\rho$- uniform distribution on the sphere of radius ρ

Interpretation: We know L_2 norm of the vector of means under alternative but we do know how it is distributed between elements of this vector, so we assume a uniform distribution on the sphere

$$\mu = \rho u, \;\; L_A(X) = \int_{S^{p-1}} L(X|u)\pi(du)$$

Neyman-Pearson test:

$$L = \int_{S^{p-1}} \frac{e^{\left(-\frac{1}{2}||x-\rho u||^2\right)}}{exp\left(-\frac{1}{2}||x||^2\right)}\pi(du)$$

Power of Neyman-Pearson test converges to α when $\frac{||\rho||^2}{\sqrt{p}} \to 0$

Relationship between signal strength and the sparsity

Only one mean different from 0 and equal to $2\sqrt{2\log p}$ - Bonferroni strong, chi-square test weak

Relationship between signal strength and the sparsity

Only one mean different from 0 and equal to $2\sqrt{2\log p}$ - Bonferroni strong, chi-square test weak

$k = \sqrt{p}$ means equal to 5 - Bonferroni weak, chi-square test strong

Relationship between signal strength and the sparsity

Only one mean different from 0 and equal to $2\sqrt{2\log p}$ - Bonferroni strong, chi-square test weak

$k = \sqrt{p}$ means equal to 5 - Bonferroni weak, chi-square test strong

What happens in the middle ?

Sparse mixture model

Model:

$$H_0 : \mu = 0, \quad H_A : \mu_1, \ldots, \mu_p \quad iid \quad (1-\epsilon)\delta_0 + \epsilon\delta_\mu$$

Sparse mixture model

Model:

$$H_0 : \mu = 0, \quad H_A : \mu_1, \ldots, \mu_p \quad iid \;\; (1-\epsilon)\delta_0 + \epsilon\delta_\mu$$

Interpretation: If alternative holds then all nonzero means are equal to fixed and known μ. The percentage of non-zero means is equal to ϵ and every μ_i has the same chance of being different from zero.

Sparse mixture model

Model:

$$H_0 : \mu = 0, \quad H_A : \mu_1, \ldots, \mu_p \quad iid \ \ (1-\epsilon)\delta_0 + \epsilon\delta_\mu$$

Interpretation: If alternative holds then all nonzero means are equal to fixed and known μ. The percentage of non-zero means is equal to ϵ and every μ_i has the same chance of being different from zero.

Neyman-Pearson test:

$$L = \Pi_{i=1}^{p}[(1-\epsilon) + \epsilon e^{\mu X_i - \mu^2/2}]$$

Ingster (1999) detection boundary

$$\epsilon^p = p^{-\beta}, \quad \frac{1}{2} < \beta < 1$$

when $\beta = 1/2$ we have about $k = p^{-1/2}p = \sqrt{p}$ signals,
when $\beta = 1$ then the number of signals $k = p^{-1}p = 1$ is equal to 1 (needle in the haystack)

$$\mu^p = \sqrt{2r \log p}$$

Ingster (1999) detection boundary

$$\epsilon^p = p^{-\beta}, \ \ \frac{1}{2} < \beta < 1$$

when $\beta = 1/2$ we have about $k = p^{-1/2}p = \sqrt{p}$ signals,
when $\beta = 1$ then the number of signals $k = p^{-1}p = 1$ is equal to 1 (needle in the haystack)

$$\mu^p = \sqrt{2r \log p}$$

$$\rho(\beta) = \begin{cases} \beta - 1/2 & \textit{for} \quad 1/2 < \beta \leq 3/4 \\ (1 - \sqrt{1-\beta})^2 & \textit{for} \quad 3/4 \leq \beta \leq 1) \end{cases}$$

Neyman-Pearson test has the full asymptotic power if $r > \rho(\beta)$ and no asymptotic power if $r < \rho(\beta)$.

Ingster (1999) detection boundary

$$\epsilon^p = p^{-\beta}, \quad \frac{1}{2} < \beta < 1$$

when $\beta = 1/2$ we have about $k = p^{-1/2}p = \sqrt{p}$ signals,
when $\beta = 1$ then the number of signals $k = p^{-1}p = 1$ is equal to 1 (needle in the haystack)

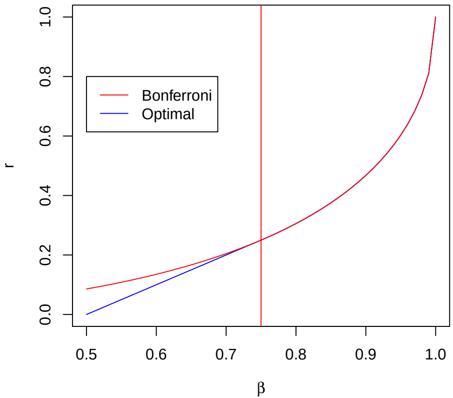
$$\mu^p = \sqrt{2r \log p}$$

$$\rho(\beta) = \begin{cases} \beta - 1/2 & \textit{for} \quad 1/2 < \beta \leq 3/4 \\ (1 - \sqrt{1-\beta})^2 & \textit{for} \quad 3/4 \leq \beta \leq 1) \end{cases}$$

Neyman-Pearson test has the full asymptotic power if $r > \rho(\beta)$ and no asymptotic power if $r < \rho(\beta)$.

Donoho and Jin (2004): Bonferroni detection boundary coincides with the optimal detection boundary if $\beta \geq 3/4$

Ingster detection boundary



Interpretation: When the number of needles increases than they can be substantially shorter than $\sqrt{2\log p}$ to be detected by Bonferroni.

Lesson learned from theory of testing the global hypothesis

The length of sparse needles to be detected needs to grow with p (amount of hay).

Lesson learned from theory of testing the global hypothesis

The length of sparse needles to be detected needs to grow with p (amount of hay).

The length of sparse needles to be detected depends on their number. If there are more of them, they can be shorter.

Lesson learned from theory of testing the global hypothesis

The length of sparse needles to be detected needs to grow with p (amount of hay).

The length of sparse needles to be detected depends on their number. If there are more of them, they can be shorter.

There is no a single optimal method for testing the global null hypothesis - the selection of the method should depend on the expectation on the signal sparsity.

Multiple testing

We now separately test each of hypotheses $H_{0i} : \mu_i = 0$

Multiple testing

We now separately test each of hypotheses $H_{0i} : \mu_i = 0$

	H_0 accepted	H_0 rejected	
H_0 true	U	V	p_0
H_0 false	T	S	p_1
	W	R	p

We now separately test each of hypotheses $H_{0i} : \mu_i = 0$

	H_0 accepted	H_0 rejected	
H_0 true	U	V	p_0
H_0 false	T	S	p_1
	W	R	p

$$FWER = P(V > 0), \qquad FDR = E\left(\frac{V}{R \vee 1}\right)$$

Multiple testing

We now separately test each of hypotheses $H_{0i} : \mu_i = 0$

	H_0 accepted	H_0 rejected	
H_0 true	U	V	p_0
H_0 false	T	S	p_1
	W	R	p

$$FWER = P(V > 0), \qquad FDR = E\left(\frac{V}{R \vee 1}\right)$$

$$E(V) = \alpha p_0$$

Multiple testing

We now separately test each of hypotheses $H_{0i} : \mu_i = 0$

	H_0 accepted	H_0 rejected	
H_0 true	U	V	p_0
H_0 false	T	S	p_1
	W	R	p

$$FWER = P(V > 0), \qquad FDR = E\left(\frac{V}{R \vee 1}\right)$$

$$E(V) = \alpha p_0$$

$$\alpha = 0.05, p_0 = 5000 \to E(V) = 250$$

Multiple testing procedures

Bonferroni correction: Use significance level $\frac{\alpha}{p}$.

Multiple testing procedures

Bonferroni correction: Use significance level $\frac{\alpha}{p}$.

Reject H_{0i} if $|X_i| \geq \Phi^{-1}\left(1 - \frac{\alpha}{2p}\right) = \sqrt{2\log p}(1 + o(1))$

Multiple testing procedures

Bonferroni correction: Use significance level $\frac{\alpha}{p}$.

Reject H_{0i} if $|X_i| \geq \Phi^{-1}\left(1 - \frac{\alpha}{2p}\right) = \sqrt{2\log p}(1 + o(1))$

Benjamini-Hochberg (1995) procedure:

(1) $|X|_{(1)} \geq |X|_{(2)} \geq \ldots \geq |X|_{(p)}$

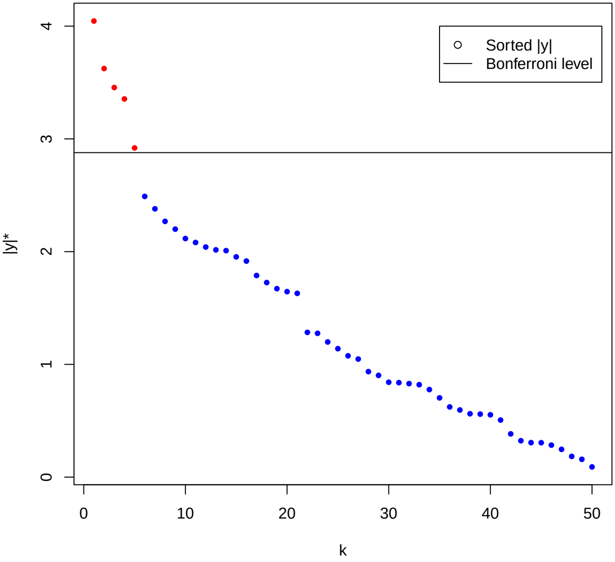
(2) Find the largest index i such that

$$|X|_{(i)} \geq \Phi^{-1}(1 - \alpha_i), \quad \alpha_i = \alpha\frac{i}{2p}, \tag{1}$$

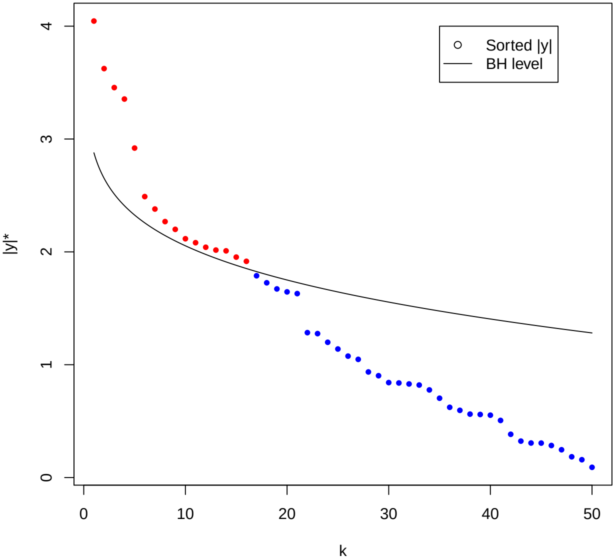
Call this index i_{SU}.

(3) Reject all $H_{(i)}$'s for which $i \leq i_{\mathrm{SU}}$

Bonferroni correction



Benjamini and Hochberg correction



FWER and FDR control

For Bonferroni correction $FWER \leq \alpha$

FWER and FDR control

For Bonferroni correction $FWER \leq \alpha$

(Benjamini,Hochberg, 1995) If $X_1, \ldots, X_p$ are independent then BH controls FDR at:

FWER and FDR control

For Bonferroni correction $FWER \leq \alpha$

(Benjamini,Hochberg, 1995) If $X_1, \ldots, X_p$ are independent then BH controls FDR at:

$$\text{FDR} = \mathbb{E}\left[\frac{V}{R \vee 1}\right] = \alpha \frac{p_0}{p}, \tag{2}$$

where p_0 is the number of true null hypotheses, $p_0 = |\{i : \mu_i = 0\}|$

FWER and FDR control

For Bonferroni correction $FWER \leq \alpha$

(Benjamini,Hochberg, 1995) If $X_1, \ldots, X_p$ are independent then BH controls FDR at:

$$\text{FDR} = \mathbb{E}\left[\frac{V}{R \vee 1}\right] = \alpha \frac{p_0}{p}, \qquad (2)$$

where p_0 is the number of true null hypotheses, $p_0 = |\{i : \mu_i = 0\}|$

(Benjamini, Yekutieli, 2001) When test statistics are "positively correlated"then BH controls FDR at or below the level $\alpha \frac{p_0}{p}$. Independently of the correlation structure FDR is controlled at or below the level $\alpha \frac{p_0}{p}$ if $|X|_{(j)}$ is compared to $\Phi^{-1}\left(1 - \frac{j\alpha}{2p \sum_{i=1}^{p} \frac{1}{i}}\right)$.

Estimating μ

$$X_i \sim N(\mu_i, \sigma^2), \quad X_1, \ldots, X_p \text{ are independent}$$

Estimating μ

$$X_i \sim N(\mu_i, \sigma^2), \quad X_1, \ldots, X_p \text{ are independent}$$

$$\hat{\mu}_{MLE} = X = (X_1, \ldots, X_p)$$

Estimating μ

$$X_i \sim N(\mu_i, \sigma^2), \quad X_1, \ldots, X_p \ \text{ are independent}$$

$$\hat{\mu}_{MLE} = X = (X_1, \ldots, X_p)$$

$$MSE(\hat{\mu}_{MLE}) = E||\hat{\mu}_{MLE} - \mu||^2 = \sum_{i=1}^{p} E(\hat{\mu}_i - \mu_i)^2 = p\sigma^2$$

Estimating μ

$$X_i \sim N(\mu_i, \sigma^2), \quad X_1, \ldots, X_p \text{ are independent}$$

$$\hat{\mu}_{MLE} = X = (X_1, \ldots, X_p)$$

$$MSE(\hat{\mu}_{MLE}) = E||\hat{\mu}_{MLE} - \mu||^2 = \sum_{i=1}^{p} E(\hat{\mu}_i - \mu_i)^2 = p\sigma^2$$

Can we do better ?

Bias-Variance Tradeoff

$$MSE(\hat{\mu}_i) = E(\hat{\mu}_i - \mu_i)^2 = B_i^2 + Var_i,$$

where $B_i = E\hat{\mu}_i - \mu_i$ is the bias of $\hat{\mu}_i$

and $Var_i = E(\hat{\mu}_i - E(\hat{\mu}_i))^2$ is the variance of $\hat{\mu}_i$.

Bias-Variance Tradeoff

$$MSE(\hat{\mu}_i) = E(\hat{\mu}_i - \mu_i)^2 = B_i^2 + Var_i,$$

where $B_i = E\hat{\mu}_i - \mu_i$ is the bias of $\hat{\mu}_i$

and $Var_i = E(\hat{\mu}_i - E(\hat{\mu}_i))^2$ is the variance of $\hat{\mu}_i$.

In our problem $E(\hat{\mu}_{MLE}) = \mu$ and $MSE(\hat{\mu}_{MLE}) = \sum_{i=1}^{p} Var_i$

Bias-Variance Tradeoff

$$MSE(\hat{\mu}_i) = E(\hat{\mu}_i - \mu_i)^2 = B_i^2 + Var_i,$$

where $B_i = E\hat{\mu}_i - \mu_i$ is the bias of $\hat{\mu}_i$

and $Var_i = E(\hat{\mu}_i - E(\hat{\mu}_i))^2$ is the variance of $\hat{\mu}_i$.

In our problem $E(\hat{\mu}_{MLE}) = \mu$ and $MSE(\hat{\mu}_{MLE}) = \sum_{i=1}^{p} Var_i$

Can we improve MSE by introducing some bias and reducing the variance ?

Shrinking towards zero

Consider the estimate $\hat{\mu}_c = c\hat{\mu}_{MLE}$

Shrinking towards zero

Consider the estimate $\hat{\mu}_c = c\hat{\mu}_{MLE}$

$$B_i(c) = c\mu_i - \mu_i = (c-1)\mu_i \text{ and } Var_i(c) = c^2\sigma^2$$

Shrinking towards zero

Consider the estimate $\hat{\mu}_c = c\hat{\mu}_{MLE}$

$$B_i(c) = c\mu_i - \mu_i = (c-1)\mu_i \quad \text{and} \quad Var_i(c) = c^2\sigma^2$$

$$MSE_i(c) = (c-1)^2\mu_i^2 + c^2\sigma^2$$

Shrinking towards zero

Consider the estimate $\hat{\mu}_c = c\hat{\mu}_{MLE}$

$$B_i(c) = c\mu_i - \mu_i = (c-1)\mu_i \ \text{ and } \ Var_i(c) = c^2\sigma^2$$

$$MSE_i(c) = (c-1)^2\mu_i^2 + c^2\sigma^2$$

$$MSE(c) = E||\hat{\mu}_c - \mu||^2 = (c-1)^2||\mu||^2 + c^2 p\sigma^2$$

Shrinking towards zero

Consider the estimate $\hat{\mu}_c = c\hat{\mu}_{MLE}$

$$B_i(c) = c\mu_i - \mu_i = (c-1)\mu_i \text{ and } Var_i(c) = c^2\sigma^2$$

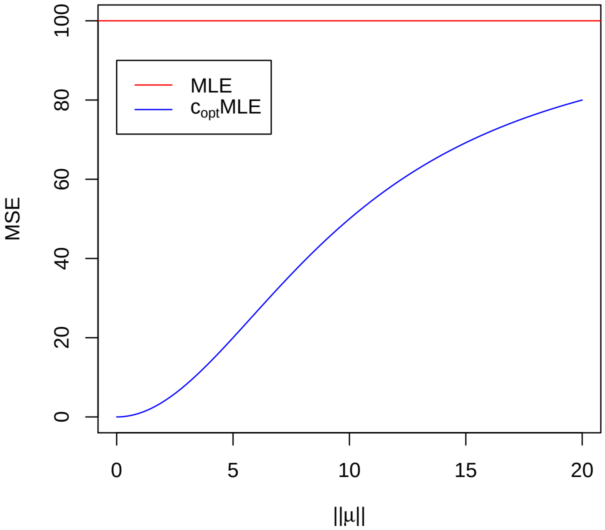
$$MSE_i(c) = (c-1)^2\mu_i^2 + c^2\sigma^2$$

$$MSE(c) = E||\hat{\mu}_c - \mu||^2 = (c-1)^2||\mu||^2 + c^2 p\sigma^2$$

Using elementary calculus we can show that the optimal value of c is equal to

$$c_{opt} = argmin_{c\in R} MSE(c) = \frac{||\mu||^2}{||\mu||^2 + p\sigma^2} \in [0,1) \ .$$

Improvement in MSE, $p = 100, \sigma = 1$



Shrinking towards the common mean

Consider an estimator

$$\hat{\mu}_d = (1 - d)\hat{\mu}_{MLE} + d\bar{X}$$

Shrinking towards the common mean

Consider an estimator

$$\hat{\mu}_d = (1 - d)\hat{\mu}_{MLE} + d\bar{X}$$

$$d_{opt} = \frac{\sigma^2}{\sigma^2 + Var(\mu)} \in (0, 1], \quad \text{with} \quad Var(\mu) = \frac{1}{p-1}\sum(\mu_i - \bar{\mu})^2.$$

Shrinking towards the common mean

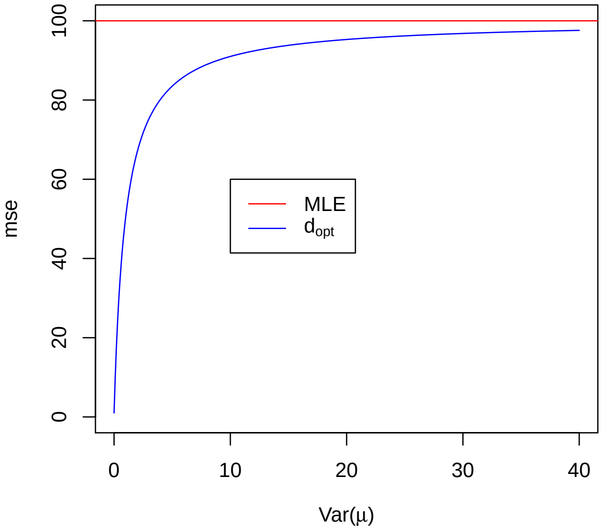
Consider an estimator

$$\hat{\mu}_d = (1-d)\hat{\mu}_{MLE} + d\bar{X}$$

$$d_{opt} = \frac{\sigma^2}{\sigma^2 + Var(\mu)} \in (0,1], \quad \text{with} \quad Var(\mu) = \frac{1}{p-1}\sum(\mu_i - \bar{\mu})^2.$$

$$d = 1 \text{ if and only if } \mu_1 = \ldots = \mu_p$$

Improvement in MSE, $p = 100, \sigma = 1$



James-Stein estimators (1961)

$$c_{opt} = \frac{||\mu||^2}{||\mu||^2 + p\sigma^2} = \left(1 - \frac{p\sigma^2}{||\mu||^2 + p\sigma^2} = \right) = \left(1 - \frac{p\sigma^2}{E||X||^2}\right)$$

James-Stein estimators (1961)

$$c_{opt} = \frac{||\mu||^2}{||\mu||^2 + p\sigma^2} = \left(1 - \frac{p\sigma^2}{||\mu||^2 + p\sigma^2} = \right) = \left(1 - \frac{p\sigma^2}{E||X||^2}\right)$$

$$c_{JS} = \left(1 - \frac{(p-2)\sigma^2}{||X||^2}\right)$$

James-Stein estimators (1961)

$$c_{opt} = \frac{||\mu||^2}{||\mu||^2 + p\sigma^2} = \left(1 - \frac{p\sigma^2}{||\mu||^2 + p\sigma^2} = \right) = \left(1 - \frac{p\sigma^2}{E||X||^2}\right)$$

$$c_{JS} = \left(1 - \frac{(p-2)\sigma^2}{||X||^2}\right)$$

$$d_{opt} = \frac{\sigma^2}{\sigma^2 + Var(\mu)}$$

James-Stein estimators (1961)

$$c_{opt} = \frac{||\mu||^2}{||\mu||^2 + p\sigma^2} = \left(1 - \frac{p\sigma^2}{||\mu||^2 + p\sigma^2} = \right) = \left(1 - \frac{p\sigma^2}{E||X||^2}\right)$$

$$c_{JS} = \left(1 - \frac{(p-2)\sigma^2}{||X||^2}\right)$$

$$d_{opt} = \frac{\sigma^2}{\sigma^2 + Var(\mu)}$$

$$d_{JS} = \frac{p-3}{p-1}\frac{\sigma^2}{Var(X)}$$

$$c_{opt} = \frac{||\mu||^2}{||\mu||^2 + p\sigma^2} = \left(1 - \frac{p\sigma^2}{||\mu||^2 + p\sigma^2} = \right) = \left(1 - \frac{p\sigma^2}{E||X||^2}\right)$$

$$c_{JS} = \left(1 - \frac{(p-2)\sigma^2}{||X||^2}\right)$$

$$d_{opt} = \frac{\sigma^2}{\sigma^2 + Var(\mu)}$$

$$d_{JS} = \frac{p-3}{p-1}\frac{\sigma^2}{Var(X)}$$

If $p > 3$ then for both J-S estimators it holds
$E||\hat{\mu}_{JS} - \mu||^2 < E||\hat{\mu}_{MLE} - \mu||^2$

Hard thresholding

When signal is sparse even better results can be obtained by hard thresholding

$$\hat{\mu}_i = \begin{cases} X_i & \text{when} \quad H_{0i} \text{ is rejected} \\ 0 & \text{when} \quad H_{0i} \text{ is not rejected} \end{cases}, \tag{3}$$

where the decisions are made by Bonferroni or BH multiple testing procedures. Bonferroni is optimal for very sparse signals while BH "adapts" to the unknown sparsity (see Abramovich, Benjamini, Donoho and Johnstone, Ann.Statist. 2006)

Facts to remember

In high dimensional problems unbiased estimators can often be improved by biased estimators with reduced variance.

Facts to remember

In high dimensional problems unbiased estimators can often be improved by biased estimators with reduced variance.

When $p > 2$ then the maximum likelihood estimator of the vector of means for the multivariate normal distribution with independent covariates is not admissible. It can be improved by James-Stein estimator.

Facts to remember

In high dimensional problems unbiased estimators can often be improved by biased estimators with reduced variance.

When $p > 2$ then the maximum likelihood estimator of the vector of means for the multivariate normal distribution with independent covariates is not admissible. It can be improved by James-Stein estimator.

In case when the signal is sparse this can be further improved by thresholding rules.

Facts to remember

In high dimensional problems unbiased estimators can often be improved by biased estimators with reduced variance.

When $p > 2$ then the maximum likelihood estimator of the vector of means for the multivariate normal distribution with independent covariates is not admissible. It can be improved by James-Stein estimator.

In case when the signal is sparse this can be further improved by thresholding rules.

Hard thresholded estimator of μ using BH multiple testing rule adapts to the unknown sparsity and is asymptotically optimal in the sense discussed in (Abramovich, Benjamini, Donoho and Johnstone, Ann.Statist. 2006)