

A NEW DATA ADAPTIVE SOLUTION TO
THE NONPARAMETRIC TWO-SAMPLE
PROBLEM WITH HIGH POWER

Majnu John (john@ams.jhu.edu)

joint work with Carey E. Priebe.

SOME SCIENTIFIC QUESTIONS:

1) IN A CLINICAL STUDY SETTING:

— DRUG EFFECT / TREATMENT EFFECT

2) GENE EXPRESSION ARRAYS:

— FINDING DIFFERENTIALLY EXPRESSED GENES.

Two-sample problem:

Control Sample: X_1, \dots, X_n .

Treatment Sample: Y_1, \dots, Y_m .

$F(\cdot), G(\cdot)$ – Underlying C.D.F's of X's and Y's, respectively.

To test:

$$H_0 : F(x) = G(x), \forall x,$$

versus

$$H_A : F(x) \leq G(x), \forall x, \text{ with strict inequality}$$

for at least one x

Classical Mann-Whitney-Wilcoxon (MWW) test.
(Wilcoxon Rank sum test):

Test statistic:

$$\sum_{i=1}^n \sum_{j=1}^m I(X_i \leq Y_j) = \sum_{j=1}^m R_j - \frac{m(m+1)}{2}.$$

R_j : rank of Y_j in the pooled sample.

Classical Mann-Whitney-Wilcoxon (MWW) test statistic,

$$\frac{1}{mn} \sum_{i=1}^n \sum_{j=1}^m I(X_i \leq Y_j) = \int_{-\infty}^{\infty} F_n(x) dG_m(x)$$

is an empirical estimate of

$$\int_{-\infty}^{\infty} F(x) dG(x).$$

Here F_n , G_m are empirical distribution functions corresponding to F and G , respectively.

Xie and Priebe's generalization of MWW statistic:

Consider

$$\int_{-\infty}^{\infty} u(F(x))dv(G(x)), \quad (1)$$

u, v arbitrary increasing continuous real-valued functions on $[0, 1]$.

An approximation to (1) is

$$\int_{-\infty}^{\infty} u_r(F(x))dv_s(G(x)), \quad (2)$$

$u_r(\cdot), v_s(\cdot)$ - Bernstein polynomials corresponding to u and v respectively.

We may write

$$u_r(\cdot) = \sum_{k=1}^r \pi_k b_{k:r}(\cdot), v_s(\cdot) = \sum_{l=1}^s \mu_l b_{l:s}(\cdot),$$

where $b_{k:r}, b_{l:s}$ are tail binomial polynomials and

$$\sum_{k=1}^r \pi_k = 1 = \sum_{l=1}^s \mu_l, \pi_k, \mu_l \geq 0, \forall k, l.$$

Using this (2) becomes

$$\sum_{k=1}^r \sum_{l=1}^s \pi_k \mu_l \Pr(X_{k:r} < Y_{l:s}). \quad (3)$$

An empirical estimate of (3) is the Weighted Generalized MWW (WGMWW) statistic:

$$\left\{ \binom{n}{r} \binom{m}{s} \right\}^{-1} \times$$

$$\sum_C \sum_{k=1}^r \sum_{l=1}^s \pi_k \mu_l I(X_{k:r}(X_{i_1}, \dots, X_{i_r}) < Y_{l:s}(Y_{j_1}, \dots, Y_{j_s}))$$

\sum_C over all $1 \leq i_1 < \dots < i_r \leq n, 1 \leq j_1 < \dots < j_s \leq m$.

Pitman Asymptotic Efficacy (PAE) for the tests based on WGMWW statistics is:

$$\psi(\pi, \mu) = \frac{\varphi^2(\pi, \mu)}{\xi(\pi, \mu, \lambda)}.$$

Here $\lambda = \lim_{n \rightarrow \infty} \left(\frac{n}{n+m} \right)$,

$\varphi(\pi, \mu) =$

$$\sum_{k=1}^r \sum_{l=1}^s \frac{r!s!\pi_k\mu_l \int_{-\infty}^{\infty} F(x)^{k+l-2} (1-F(x))^{r+s-k-l} f^2(x) dx}{(k-1)!(r-k)!(l-1)!(s-l)!}.$$

$\xi(\pi, \mu, \lambda)$ does not depend on F or G .

$\pi = (\pi_1, \dots, \pi_r)$, $\mu = (\mu_1, \dots, \mu_s)$.

PERFORMANCE OF TESTS:

Consider a model indexed by parameters θ . To test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

type 1 error: reject the null, when it is true.

type 2 error: accept the null, when it is not true.

We fix, $\alpha = \text{Pr}(\text{type 1 error})$.

For an alternative value θ_1 ,

POWER = $1 - \text{Pr}(\text{type 2 error})$

$$= P_{\theta_1}(\text{null hypothesis is rejected}).$$

POWER FUNCTION (of a test of hypothesis relevant to this model) is

$$\beta(\theta) = P_{\theta}(\text{null hypothesis is rejected}).$$

ASYMPTOTIC RELATIVE EFFICIENCY (ARE):

$\{S_{n_i}\}, \{T_{n'_i}\}$ - sequences of statistics for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

Let $\{\theta_i\}$ be a sequence of alternatives such that $\lim_{i \rightarrow \infty} \theta_i = \theta_0$. Assume, $\{n_i\}, \{n'_i\}$ sequences of positive integers for which have $\{S_{n_i}\}, \{T_{n'_i}\}$ have the same limiting significance level α and $\alpha < \lim_{i \rightarrow \infty} \beta_{S_{n_i}}(\theta_i) = \beta_{T_{n'_i}}(\theta_i) < 1$.

$$ARE(S, T) = \lim_{i \rightarrow \infty} \frac{n'_i}{n_i}$$

Under a bunch of extra condtions,

$$\text{NOETHER'S THEOREM: ARE}(S, T) = \frac{K_S^2}{K_T^2}$$

where

$$K_S = \lim_{n \rightarrow \infty} \frac{\mu'_{S_n}(\theta_0)}{\sqrt{n\sigma_{S_n}^2(\theta_0)}}$$

is the PITMAN ASYMPTOTIC EFFICACY (PAE) of the test based on $\{S_n\}$.

PAE for the tests based on WGMWW statistics is $\psi(\pi, \mu) = \frac{\varphi^2(\pi, \mu)}{\xi(\pi, \mu)}$, where $\varphi(\pi, \mu) =$

$$\sum_{k=1}^r \sum_{l=1}^s \frac{r!s!\pi_k\mu_l \int_{-\infty}^{\infty} F(x)^{k+l-2} (1-F(x))^{r+s-k-l} f^2(x) dx}{(k-1)!(r-k)!(l-1)!(s-l)!}$$

$\xi(\pi, \mu)$ does not depend on F or G . $\pi = (\pi_1, \dots, \pi_r)$,
 $\mu = (\mu_1, \dots, \mu_s)$.

PAE for the tests based on WGMWW statistics

$$\text{is } \psi(\pi, \mu) = \frac{\varphi^2(\pi, \mu)}{\xi(\pi, \mu, \lambda)}.$$

Here $\lambda = \lim_{n \rightarrow \infty} \left(\frac{n}{n+m} \right)$,

$$\varphi(\pi, \mu) =$$

$$\sum_{k=1}^r \sum_{l=1}^s \frac{r!s! \pi_k \mu_l \int_{-\infty}^{\infty} F(x)^{k+l-2} (1-F(x))^{r+s-k-l} f^2(x) dx}{(k-1)!(r-k)!(l-1)!(s-l)!}.$$

$\xi(\pi, \mu, \lambda)$ does not depend on F or G . $\pi = (\pi_1, \dots, \pi_r)$, $\mu = (\mu_1, \dots, \mu_s)$.

Our estimator for ψ is $\hat{\psi} = \frac{\hat{\varphi}^2}{\xi}$

where $\hat{\varphi} =$

$$\frac{(k-1)}{n^2} \sum_{\nu=1}^{n-k} h(F_n(\rho X_{n-\nu+1:n} + (1-\rho)X_{n-(\nu+k)+1:n})) \times \\ (X_{n-\nu+1:n} - X_{n-(\nu+k)+1:n})^{-1},$$

$$h(y) = \sum_{i=1}^r \sum_{j=1}^s \pi_i \mu_j C(r, s, i, j) y^{i+j-2} (1-y)^{r+s-i-j}$$

$$\text{with } C(r, s, i, j) = \frac{r!s!}{(i-1)!(r-i)!(j-1)!(s-j)!},$$

$k > 2$, is an integer and $0 \leq \rho \leq 1$.

THEOREM: $\hat{\psi} \rightarrow \psi$ almost surely, as $n \rightarrow \infty$.

ASSUMPTION: F has a piecewise uniformly continuous density f , ultimately monotonically non-increasing as $x \rightarrow \pm\infty$.

COROLLARY: For fixed positive integers r and s , consider the following compact subset of $R^r \times R^s$:

$$S = \{(\pi_1, \dots, \pi_n, \mu_1, \dots, \mu_s) : \pi_i \geq 0, \mu_j \geq 0, \forall i, j$$

$$\text{and } \sum_{i=1}^r \pi_i = 1 = \sum_{j=1}^s \mu_j\}$$

Define

$$\hat{\psi} : S \rightarrow R, \text{ as } \hat{\psi} = \frac{\hat{\varphi}^2}{\xi}, \text{ and}$$

$$\psi : S \rightarrow R, \text{ as } \psi = \frac{\varphi^2}{\xi}, \text{ where}$$

$\hat{\varphi}$, φ and ξ are as before. Suppose, $y'_n \rightarrow y^0$ with some positive probability, where the y'_n and y^0 are in S and the y'_n satisfy

$$\hat{\psi}(y'_n) \geq \hat{\psi}(y), \text{ w.p.1, } \forall y \in S, n \geq 1;$$

then $\psi(y^0) \geq \psi(y)$, w.p.1, $\forall y \in S$.

Main idea in the proof:

the transformation: $V_{n:\nu} = -\log F(X_{n:n-\nu+1})$,
and Renyi's representation:

$$V_{n:\nu} = \sum_{i=1}^{\nu} \frac{Z_i}{n-i+1}.$$

where Z_i 's are i.i.d. random variables.

Conclusions:

Our nonparametric test is

1) data-adaptive

2) has substantially higher power than many existing nonparametric tests (especially when the underlying density 'highly non-Gaussian'.)