Toward optimal graphic tests

Jason (Jixian) Wang and Ram Tiwari

Bristol Myers Squibb
jixian.wang@bms.com

ADMTP Workshop 2023, Basel

Acknowledgement

We thank the Scientific Working Group on Multiplicity in GBDS/BMS for their input:

- Shuyu Chu
- Min Vedal
- Arun Kumar
- Azedeh Shohoudi
- Arkendu Chatterjee
- Wencong Chen
- Samira Soleymani

Background

- Multiple comparisons play an important role in drug development (FDA, 2017).
- The graphical test (GT) procedures (Bretz et al., 2009; Dmitrienko et al., 2009) provide an intuitive view on the process of the test and how it relates to design parameters.
- In practice, an FAQ is how to determine the parameters in a GT.
- We take a Bayesian approach to find the optimal GT that maximizes the expected utility that reflects the overall utility of rejecting a set of hypotheses.
- We examine technical issues often occurring when finding the optimal GT and discuss possible solutions.
- We show some examples including optimal sequence for fixed sequence tests, and comparison of "optimal" GT with Holm test.

The graphical test procedure and its parameters

- A GT can be represented by a directional graph with nodes representing, e.g., K hypotheses to be tested.
- It has two sets of parameters, a transition matrix G; and a vector of relative weights $w = (w_1, ..., w_K)$ of the hypotheses, with $\sum_{k=1}^{K} w_k = 1$.
- For a given graph, the K nodes are tested at levels $w \times \alpha$, where α is the overall type I error to be controlled, and the most significant one is rejected.
- An element g_{jk} of G denotes the transition rate of α from node j to k, if j is rejected. We also need $\sum_{k=1}^{K} g_{jk} = 1$, and $g_{ll} = 0$, for all ls. Then, the graph is updated and tested again for hypotheses not yet rejected until no rejection can be made.
- The test is based on p-values, but we will consider the test statistics Y (a K-vector) directly, and will assume that $Y \sim N(\mu, \Sigma)$ with prior knowledge on μ, Σ.

.

Graphic test: three hypotheses example

- We take an example with $K = 3$ with test statistics $Y \sim N(\mu, \Sigma)$ with $\mu = (\mu_1, \mu_2, \mu_3)$.
- Local tests are based on p-values $p_k = 1 - \Phi(\mu_k)/2$, assuming $\Sigma = \text{diag}(3)$.
- The three hypotheses are $H_{0k} : \mu_k = 0 \;\; vs. \;\; \mu_k \neq 0, k = 1, 2, 3$
- For $K = 3$, we have $w = (w_1, w_2, 1 - w_1 - w_2)$ and

$$G = \begin{pmatrix} 0 & g_{12} & 1 - g_{12} \\ g_{21} & 0 & 1 - g_{21} \\ g_{13} & 1 - g_{13} & 0 \end{pmatrix}. \tag{1}$$

- Therefore, the graph parameters are $(g_{12}, g_{21}, g_{31}, w_1, w_2)$. In general, there are $K(K-2) + (K-1)$ parameters to determine.
- We can maximize, e.g., the mean number of H_ks rejected, which can be evaluated by simulation (e.g., as implemented in gMCP).

Test function and power

- We write the whole test in terms of a vector of test function $\phi(Y) = (\phi_1(Y), ..., \phi_K(Y))$.
- $\phi_k(Y) = 1$ if kth hypothesis is finally rejected, and $\phi_k(Y) = 0$ otherwise.
- As it depends on G and w so we write it $\phi(Y|G, w)$.
- For given μ, Σ, the frequentist powers of testing the K hypotheses are

$$P_f(G, w) = E_Y(\phi(Y|G, w)|\mu, \Sigma) \tag{2}$$

- For GT, the form of $\phi_k(Y)$ is not obvious, but the power can be calculated.

Bayesian power and utility I

- Suppose we have prior knowledge on μ, Σ in the form of prior distribution $\mu, \Sigma \sim F_0(\mu, \Sigma)$, where $F_0(.)$ can take different forms.
- We can define a Bayesian counterpart of $P_f(G, w)$ as

$$P_b(G, w) = \int E_Y(\phi(Y|G, w)|\mu, \Sigma) dF_0(\mu, \Sigma) \tag{3}$$

- As $P_b(G, w)$ is a vector, we need a measure for the overall consequence of rejecting a set of hypothesis.
- We take the additive linear utility with $\mathbf{u} = (u_1, ..., u_k)$ (so u_k is the "value" of rejecting H_{0k})

$$U_b(G, w) = \boldsymbol{u}^T P_b(G, w) \tag{4}$$

Here we use $\mathbf{u} = (1, ..., 1)$, will call $U_b(G, w)$ "Power", meaning expected number of rejections.

Bayesian power and utility II

- General $F_0(\mu, \Sigma)$ is difficult to specify. Sometimes $F_0(.)$ is a finite-mixture distribution, $\mu = \mu_l, \Sigma = \Sigma_l, l = 1, ..., L.$ with probability P_l. With this, we can write (5) as

$$U_b(G, w) = \sum_{l=1}^{L} P_l \boldsymbol{u}^T E_Y(\phi(Y|G, w)|\mu_l, \Sigma_l) \quad (5)$$

- This is the finite representation of Dirichlet process prior: $(P_1, ..., P_L) \sim Dir(a_0/L, ..., a_0/L)$ where $a_0 > 0$ is the precision parameter; and $(\mu_l, \Sigma_l) \sim N(\mu_0, \boldsymbol{\tau}^2)$.
- Beyond additive linear utility: u_k may change depending on if H_j is rejected, so the the utility of rejecting H_k and H_j are not additive.
- To count for this, we need to calculate $P(\text{Reject } H_j \cap H_k)$ as an "interaction" term.
- Assigning a utility to this term could be difficult.

Finding design parameters for maximizing the expected utility

- With the above, we can find optimal G, w that maximize $U_b(G, w)$ in principle.
- Searching for them is generally difficult, as $U_b(G, w)$ is not a continuous function of G and w and perhaps with multiple local maximums.
- The following approaches may find approximate ones:
  1. A grid approach when K is small (e.g., 3).
  2. A stochastic search.
  3. Approximate $U_b(G, w)$ with a differentiable function using, eg, deep neural network (Zhan, 2022).
  4. Alternate optimizations of G and w.
  5. Start with an optimal (sub-)graph.
  6. Efficient optimizer for specific tasks (e.g., optimal order of hypotheses in fixed sequence tests)
  7. Re-parameterization of GT parameters.
  8. Sample reuse simulation.

Optimal order of hypotheses in fixed sequence tests I

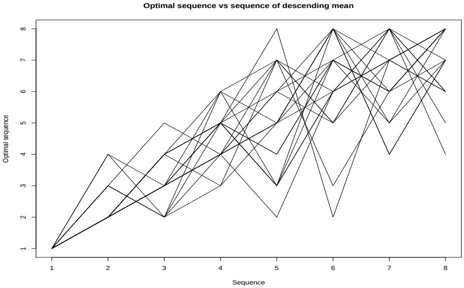
- Fixed sequence tests are a much simplified method from the general closed test procedures. But even for them, finding the optimal order of hypotheses is not easy.
- Zhang et al (2015) proposed using either a greedy algorithm which is optimal when the correlation is compound symmetry, or simulated annealing in general situations.
- This problem is a special case of optimal graphic test, when all weights are given to a single hypothesis and the transit matrix has one and only one unit element each row.
- This can be considered as an integer programming problem, which may be more efficient than simulated annealing.

Optimal order of hypotheses in fixed sequence tests II

- In practice, we often order the hypotheses according to their importance and/or individual power, but Σ may also play an important role.
- For a small number of secondary hypotheses is large, it is possible to find the exact optimal sequence by comparing the full permutations.
- Otherwise, we propose to compare only local permutations that are not too different from the proposed order.
- Using efficient sample reuse simulation algorithm (see later) make it feasible for most practical scenarios (eg, < 10 secondary hypotheses).

Optimal sequence and correlation matrix

- An example with 8 hypotheses: equal distance in the mean 2.64, 2.50, 2.36, 2.21, 2.07, 1.93, 1.79, 1.64, with either fixed or random Σ.
- Right figure shows optimal position of each H_k vs the order by the means in 20 simulation with random Σ (using R function **randcorr**).
- Right table gives Different correlation matrices: compound symmetry and 3 random ones .
- 10000 simulation runs with sample reuse. 80-90 seconds user time in R each scenario.



ρ	Opt. Sequence	Power(# Rej.)
0.0	1 2 3 4 5 6 7 8	2.058
0.3	1 2 3 4 5 6 7 8	2.616
0.6	1 2 3 4 5 6 7 8	3.256
Rand 1	1 2 3 4 7 6 8 5	2.162
Rand 2	1 2 3 4 5 6 7 8	2.187
Rand 3	1 2 3 4 6 5 7 8	2.567

A comparison between optimal GT and Holm's tests for 4 H_ks

- Bonforroni-Holm test is a GT with symmetric graph, equal weights and transit rate.
- We compare the optimal and Holm's tests for testing 4 hypotheses with varying μ and correlation ρ in compound symmetry Σ.
- 20000 simulation runs for each scenario.
- Power of both tests are given in the right table.
- The gain of optimal test varies depending on μ and ρ, except 2nd row, in which Holm is optimal.
- These are empirical power, hence the optimal ones may slightly over estimated.

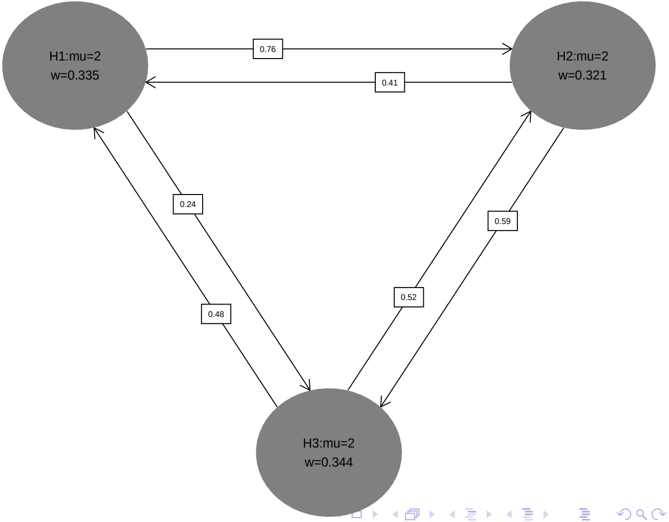
μ	ρ	Power(Opt)	Power(Holm)
(3,3,2,2)	0.0	2.379	2.326
(2,2,2,2)*	0.0	1.430	1.428
(3,2.5,2,1.5)	0.0	1.995	1.904
(3,2.5,2,2)	0.0	2.149	2.110
(3,2.7,2.4,2.1)	0.0	2.465	2.439
(3,3,2,2)	0.3	2.423	2.321
(3,3,2,2)	0.6	2.492	2.320

Optimal GT for 3 hypotheses

Graphic representation of optimal GT parameters for $\Sigma = diag(3)$ and $\mu = (2, 2, 2)$.

Holm test (all weights are $1/3$) and all non-zero elements in G are $1/2$) is optimal.

The optimal weights are close to $1/3$, but the parameters in G vary considerably.



The structure and parameter of optimal GTs

- Power and optimal GT parameters for $\Sigma = diag(3)$ and $\mu = (2, 2, 2)$, 20 simulation starting at Holm's test.
- Although the differences in power is minimum (Power of Holm is 1.194), the optimal weights have moderate change, while the G parameters are quite unstable.

Power	g_{12}	g_{21}	g_{31}	w_1	w_2	w_3
1.194	0.67	0.44	0.29	0.34	0.32	0.34
1.194	0.75	0.31	0.51	0.35	0.33	0.32
1.194	0.43	0.44	0.56	0.32	0.35	0.33
1.194	0.43	0.44	0.61	0.33	0.34	0.33
1.194	0.63	0.72	0.59	0.31	0.36	0.33
1.195	0.53	0.34	0.52	0.35	0.31	0.34
1.194	0.48	0.81	0.59	0.33	0.31	0.36
1.195	0.49	0.37	0.53	0.34	0.33	0.33
1.194	0.53	0.72	0.59	0.35	0.34	0.31
1.194	0.59	0.28	0.54	0.35	0.30	0.35
1.194	0.56	0.50	0.55	0.34	0.32	0.34
1.194	0.37	0.25	0.70	0.35	0.32	0.33
1.195	0.63	0.34	0.56	0.35	0.33	0.32
1.193	0.44	0.54	0.49	0.34	0.31	0.35
1.194	0.66	0.57	0.34	0.31	0.33	0.36
1.193	0.91	0.59	0.42	0.30	0.28	0.42
1.195	0.57	0.39	0.57	0.34	0.33	0.33
1.195	0.49	0.40	0.53	0.33	0.34	0.33
1.194	0.48	0.40	0.34	0.35	0.32	0.33
1.194	0.55	0.46	0.45	0.35	0.32	0.33

The structure and parameter of optimal GTs

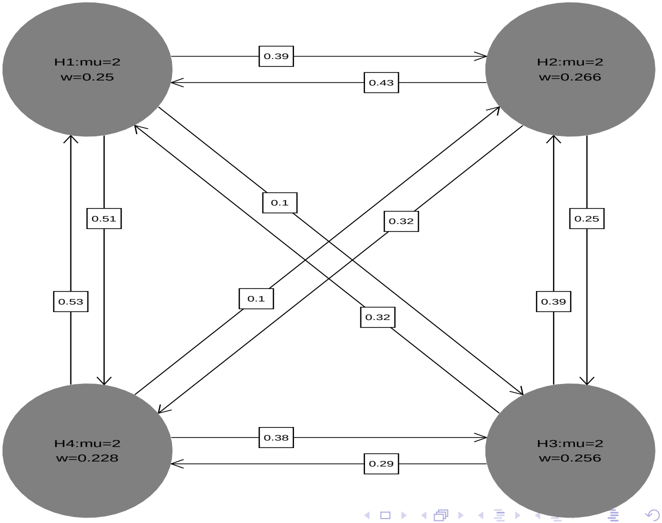
- Power and optimal GT parameters for $\Sigma = diag(3)$ and $\mu = (3, 2.5, 2)$, 20 simulation
- There is almost no differences in power , the optimal weights have moderate change, while the G parameters can be quite unstable.
- The less the local power, the less stable of the g_{jk} going out.
- Starting with Holm's test parameters leads to very similar power.

Power	g_{12}	g_{21}	g_{31}	w_1	w_2	w_3
1.875	0.67	0.74	0.17	0.61	0.29	0.10
1.875	0.65	0.84	0.33	0.58	0.32	0.10
1.875	0.59	0.71	0.65	0.59	0.33	0.08
1.875	0.70	0.82	0.10	0.63	0.26	0.11
1.875	0.70	0.70	0.16	0.60	0.29	0.10
1.875	0.73	0.57	0.56	0.63	0.25	0.12
1.875	0.71	0.71	0.63	0.57	0.32	0.11
1.875	0.64	0.62	0.93	0.58	0.35	0.07
1.875	0.66	0.73	0.55	0.62	0.30	0.08
1.875	0.68	0.74	0.94	0.66	0.26	0.08
1.875	0.72	0.88	0.67	0.60	0.28	0.12
1.875	0.72	0.59	0.96	0.60	0.29	0.11
1.875	0.68	0.92	0.62	0.56	0.31	0.13
1.875	0.67	0.83	0.28	0.58	0.31	0.11
1.875	0.65	0.72	0.79	0.60	0.32	0.08
1.875	0.69	0.72	0.42	0.65	0.27	0.08
1.875	0.75	0.71	0.36	0.64	0.24	0.11
1.875	0.75	0.74	0.13	0.60	0.32	0.08
1.875	0.76	0.72	0.95	0.62	0.26	0.11
1.875	0.69	0.70	0.34	0.66	0.25	0.10

Optimal GT for 4 hypotheses

Graphic representation of optimal GT parameters for $\Sigma = diag(4)$ and $\mu = (2, 2, 2, 2)$.

The weights are similar, but the G matrix is rather different from those of Holm test.



Finite mixed prior

- Often we are uncertain about the value of μ and Σ
- One possibility to pass this uncertainty on to optimal tests is via finite mixture priors.
- The optimal mixture GT maximizes the average power over the mixture.
- Suppose we believe that the means in the 3-hypothesis case are either (3,2.5,2) or (2,2,2), with equal chance.
- The optimal mixture GT has power **1.529**, while separate optimal GTs for the two μs have power **1.875, 1.194**, respectively, hence the average power is **1.535**, slightly higher than **1.529**.
- Suppose we have means (3,2.5,2), but with either 0 or 0.5 correlation with equal chance.
- The optimal mixture GT has power **1.890**, while separate optimal GTs for the two Σs have power **1.874, 1.921**, respectively, hence the average power is **1.898**, slightly higher than **1.890**.

Technical issues in optimization I

- The major technical issues in finding optimal GT is the existence of discontinuous points and local optimums in the power function.
- Zhan et al. (2022) proposed to use forward deep network to fit the power function such that the fitted model is well behaved.
- Then optimal parameters can be found using efficient algorithm, eg, with gradients.
- One needs to control the machine learning error, as well as the error due to local optimums.
- They also reported results by stochastic search and a genetic algorithm, all inferior than their deep network approach.

Reparameterization I

- Parameterization of the GT parameters plays an important role.
- The parameters in the GT all have linear constraints $\sum_{i=1}^{K} w_i \leq 1$ and $\sum_{i=1}^{K} g_{jk} \leq 1$.
- Although we can consider our task as optimization with linear constraints, it is often less efficient and stable than the proposed reparameterization.
- Many software support boxed but not linear constraints.
- we use a reparameterization, with a similar idea as "stick breaking" prior in Bayesian analysis.
- For example, to get w_k with $\sum_{i=1}^{K} w_i \leq 1$, we make a non-decreasing sequence $0 \leq a_0 \leq a_1, ..., \leq a_K \leq 1$ then take $w_k = a_k - a_{k-1}$.
- Constraints $\sum_{i=1}^{K} w_i = a_K \leq 1$ is satisfied by construction.

Reparameterization II

- To ensure monotone a_ks we use

$$a_k = 1/(1+\exp(-\sum_{j=1}^{k} b_j)) \tag{6}$$

  where $b_j \geq 0, j > 1$.

- This reparameterization works well together with the Hooke-Jeeves algorithm for derivative-free optimization, implemented in R-package **dfoptim**.
- One can also find b_js given a_k:

$$b_k = \log(a_k/(1-a_k)) - \log(a_{k-1}/(1-a_{k-1})) \tag{7}$$

- This is particularly useful for specifying initial values for the optimizer, given a graph. For example, using the Holm test graph parameters as initial values often works well.

Sample reused simulation

- To mitigate the impact of simulation error, sample reuse is a way not only to mitigate this issue, but also reduce computing burden.
- The idea is to use the same set of random samples such that the optimization procedure is not affected by simulation error.
- The following algorithm is for optimal sequence search
  1. Generate n samples $Y \sim N(\mu, \Sigma)$ and calculate $R = I[Y > u_{1-\alpha/2}]$ ($n \times K$ matrix) with very large n.
  2. Generate permutations of sequence 1:K and delete non-local ones (too different from 1:K) and get a set of M permutations $Q_1, ..., Q_M$.
  3. Repeat for each permuted sequence Q_p, $m = 1, ..., M$ calculate the "survival function" $S_m(J) = \prod_{j=1}^{J} R_{Q_m(j)}$, $J = 1, ..., K$.
  4. Calculate the mean "survival time" (number of rejections) for each Q_m
  5. The one with the highest survival time is the optimal sequence.
- Can be fully vectorized in R-code, and is quicker than calcPower(.).

Discussion and further work

- Do we really need to use optimal GT? Probably not always, but it is a useful reference.
- Commonly used methods such as the Holm test are reasonably powerful for a wide range of setting, but it is still worthwhile to check.
- The structure of optimal GT is not stable, but the power is.
- Careful use of derivative-free approaches in combination with other tricks such as reparameterization provides feasible practical approaches, but more technical advance is still useful.
- Eliciting information for μ and Σ or the finite mixture prior is a practical challenge.
- Some of our approaches can be extended to optimal weighted tests (eg, Westfall & Krishen, 2001).

- Bretz, F., Maurer, W., Brannath, W., Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. Statistics in medicine, 28(4), 586-604.

- Dmitrienko A, Tamhane AC, Bretz F. (editors). Multiple Testing Problems in Pharmaceutical Statistics. Chapman and Hall/CRC Press, 2009, New York.

- Food and Drug Administration (2017), “Multiple Endpoints in Clinical Trials Guidance for Industry,” available at https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ ucm536750.pdf .

- Kramer, O., Ciaurri, D. E., and Koziel, S. (2011), “Derivative-Free Optimization,” in Computational Optimization, Methods and Algorithms, eds. S. Koziel and X-S. Yang, Berlin, Heidelberg: Springer-Verlag, pp. 61–83.

- Marcus, R., Eric, P., and Gabriel, K. R. (1976), “On closed Testing Procedures With Special Reference to Ordered Analysis of Variance,” Biometrika, 63, 655–660.

- Tianyu Zhan, Alan Hartford, Jian Kang Walter Offen (2022) Optimizing Graphical Procedures for Multiplicity Control in a Confirmatory Clinical Trial via Deep Learning, Statistics in Biopharmaceutical Research, 14:1, 92-102.

- Wang, J. (2002), Sample reuse simulation in optimal design for Tmax in pharmacokinetic experiments. Journal of the Royal Statistical Society: Series C (Applied Statistics), 51: 59-67.

- Westfall, P. H., and Krishen, A. (2001), "Optimally Weighted, Fixed Sequence and Gatekeeper Multiple Testing Procedures," Journal of Statistical Planning and Inference, 99, 25–40.

- Zhang Z, Wang C, Troendle JF. Optimizing the order of hypotheses in serial testing of multiple endpoints in clinical trials. Stat Med. 2015;34(9):1467-1482.