6.1 Profile information criteria

Abramovich *et al.* (2000) **[36]** consider theoretically thresholding estimators of a sequence of Normal distribution means, where the threshold is determined by a lack of fit criterion (l^p distance) penalized by FDR. They show that the estimators are asymptotically minimax. Regarding massive multiple tests as the estimation problem described in Section 2, Cheng *et al.* (2004) **[24]** develop criteria to determine the significance threshold α for the *HT*(α) procedure (3). The *profile information* (I_p) criterion consists of a lack-of-fit term of the P value ensemble quantile function from *U*(0, 1) penalized by the expected number of false discoveries under model (1). Empirically, the lack-of-fit term is defined by $\tilde{D}(\alpha) = \sqrt{m}\left\{\int_0^{\alpha}[t-\tilde{Q}_m(t)]_+^2 dt\right\}^{1/2}, \alpha \in (0,1],$

where $\tilde{Q}_m(\cdot)$ is the P value EQF (cf. Section 4) and $[x]_+$ denotes the positive part of *x*, i.e., $[x]_+ = \max\{x, 0\}$. So $\tilde{D}(\alpha)$ measures how far are the P value sample quantiles below the diagonal line on the interval (0, α]. Empirically the profile information criterion I_p is given by

$$\tilde{I}_p(\alpha) = [\tilde{D}(\alpha)]^{-1} + \lambda(m, \hat{\pi}_0) m \hat{\pi}_0 \alpha, \qquad \alpha \in (0, 1)$$

Here $m\hat{\pi}_0\alpha$ is an estimate of the expected number of false positives, $\lambda(m, \hat{\pi}_0)$ is a penalty factor, and $[\tilde{D}(\alpha)]^{-1}$ measures the deviation of the P values from the *U*(0, 1) distribution. The more concentrated are the P values towardzero, the larger is $[\tilde{D}(\alpha)]$ and thus the smaller is $[\tilde{D}(\alpha)]^{-1}$; therefore one minimizes $\tilde{I}_p(\alpha)$ with respect to α. So the data-driven "optimal" significance threshold is the $\hat{\alpha}^*$ that minimizes $\tilde{I}_p(\alpha)$; and the $HT(\hat{\alpha}^*)$ procedure rejects H_{i0} if $P_i \le \hat{\alpha}^*$. Cheng (2006) **[15]** extends I_p by introducing the *adaptive profile information* (API) criterion based on the quantile model $Q_m^*(u) = I(0 \le u \le \tau)(nu^{\gamma} + \delta u) + I(\tau_m \le u \le 1)(\beta_0 + \beta_{1u})$ (cf. Section 4.3). API is defined as

$$API(\alpha) := \left[\int_0^{\alpha} (t - Q_m^*(t))^{\gamma} dt\right]^{-1/\gamma} + \lambda(m, \pi_0, \delta) m \pi_0 \alpha, \qquad \alpha \in (0, 1)$$

Here the major modification is on the lack-of-fit term: the L^2 norm is replace by the L^{γ} norm. Recall that $\gamma \geq 1$ is a parameter reflecting how far the P value quantiles are below the *U*(0, 1) quantiles in the vicinity of zero. The L^{γ} norm emphasis this deviation and makes the criterion more adaptive to the P value behavior around zero. Cheng (2006) **[15]** considers an approximation of the lack-of-fit term that simplifies both theoretical development and computation in practice, and proposes a procedure to estimate the parameters in API. The data-driven optimal significance threshold $\hat{\alpha}^*$ is the α that minimizes an approximate API with estimated parameters in $Q_m^*(\cdot)$.

A key issue is the choice of the penalty factor λ. Cheng *et al.* (2004) **[24]** and Cheng (2006) **[15]** consider a few conservative choices and show for $\pi_0 < 1$ the pERR of the $HT(\hat{\alpha}^*)$ procedure (3) diminishes to zero as $m \to \infty$ regardless the dependence among the P values; and for $\pi_0 \le 1$ the ERR diminishes to zero as $m \to \infty$ if the P values posses certain dependence structure. The simulation studies therein indicate that these choices perform well when there is substantial power to reject the false null hypothesis in a number of individual tests, and they tend to be conservative when the power is low. Moreover, in a range of scenarios API moderately outperforms I_p.

6.2 Total error proportion

Pounds and Morris (2003) **[28]** observe that given a threshold α, the area under the P value density function can be partitioned into four distinct regions corresponding to the four hypothesis testing outcomes resulted from the *HT*(α) procedure (3). More specifically, the area to the left of α corresponds to rejections and the area below π_0 can be attributed to the *U*(0, 1) distribution. Thus, assuming that the null distribution of the P values is *U*(0, 1), the area left of α and below π_0 corresponds to Type I errors, the area left of α and above π_0 corresponds to correct rejections, the area above π_0 and right of α corresponds to Type II errors, and the area below π_0 and right of α corresponds to correct non-rejections. In particular, under model (1) the expected proportion of tests resulting in a Type I error is given by $FP(\alpha) = \pi_0\alpha$. Additionally, the expected proportion of tests resulting in a Type II error is given by $FN(\alpha) = (1 - \pi_0)(1 - Hm(\alpha))$. The *total error proportion* is the sum TE(α) = FP(α) +FN(α), which is the expected proportion of tests resulting in a Type I or Type II error. Cheng *et al.* (2004) **[24]** use the term “total error criterion” and Genovese and Wasserman (2002) **[11]** use the term “total misclassification risk” to describe the total error proportion.

In practice, an estimate of the total error proportion can be used as a criterion to guide the selection of α. An estimate of TE(α) can be obtained by substituting estimates for the terms in FP and FN. Then, the value of α that minimizes this TE estimate can be easily determined. The TE estimators can be nonparametric **[24]** or parametric with the mixture models. **[27, 28]** Let $\hat{\alpha}_{TE}$ be the α so obtained.

Using $\hat{\alpha}_{TE}$ to declare significance has some useful operating characteristics. First, if the estimate of $F_m(\cdot), \hat{F}_m(t) = t$ for all t (indicating an all null case), then $\hat{\alpha}_{TE}$ = 0 (no rejections are made). Additionally, $\hat{\alpha}_{TE}$ corresponds to a 50% empirical Bayes probability that the null hypothesis is true. **[28]**