Appendix to Nonparametric Assessment of Contamination in Multivariate Data Using Generalized Quantile Sets and FDR, published in the Journal of Computational and Graphical Statistics

Clayton Scott* and Eric Kolaczyk†

January 18, 2010

Proofs

The proofs of Propositions 1 and 3 rely on certain ROCs (or CDFs) which we discuss here in more detail. Consider the optimal test for the null hypothesis $X \sim P$ against the alternative $X \sim \mu$. By definition of $G_{P,\beta}$, and since these sets are unique under **[B]** and **[C]**, the critical region $G^c_{P,\beta}$ is the most powerful test of size $P(G^c_{P,\beta}) = 1 - P(G_{P,\beta}) = 1 - \beta$, with power equal to $\mu(G^c_{P,\beta}) = 1 - \mu(G_{P,\beta})$. Thus, $\{(1-\beta, 1-\mu(G_{P,\beta})) : 0 \leq \beta \leq 1\}$ traces out the ROC of the optimal test. In functional form, the ROC is given by

$$C(s) := 1 - \mu(G_{P,1-s}).$$

In a similar way, we can associate

$$\tilde{C}(s) = 1 - \mu(G_{Q,1-s})$$

*Department of Electrical Engineering and Computer Science, University of Michigan, 1301 Beal Avenue, Ann Arbor, MI 48105 (email: cscott-at-eecs-dot-umich-dot-edu).

†Department of Mathematics and Statistics, Boston University, 111 Cummington Street, Boston, MA 02215 (email: kolaczyk-at-math-dot-bu-edu).

with the optimal test for $X \sim Q$ versus $X \sim \mu$.

The estimation of π is facilitated by consideration of what might be called the *dual* ROCs to the *primal* ROCs above. In particular, we now view μ as the null distribution and P as the alternative. While this is the opposite of the scenario considered throughout the paper, it will be a useful analytical device. By definition of $G_{P,\beta}$, the critical region $G_{P,\beta}$ gives the most powerful test of size $\mu(G_{P,\beta})$ with power equal to $P(G_{P,\beta}) = \beta$. Thus, $\{(\mu(G_{P,\beta}), \beta) : 0 \leq \beta \leq 1\}$ traces out the ROC of the optimal test. In functional form, the ROC is given by

$$D(t) := \inf\{\beta : \mu(G_{P,\beta}) \leq t\}.$$

Note that the dual ROC can be obtained by reflecting $C(s)$ about the anti-diagonal of the unit square.

Similarly, the dual ROC corresponding to the optimal test of the null $X \sim \mu$ versus the alternative $X \sim Q$ (again, this test is viewed as purely an analytical device) is given by

$$\tilde{D}(t) := \inf\{\tilde{\beta} : \mu(G_{Q,\tilde{\beta}}) \leq t\},$$

and is traced out by the curve $\{(\mu(G_{Q,\tilde{\beta}}), \tilde{\beta}) : 0 \leq \tilde{\beta} \leq 1\}$. Again, this curve may be obtained by reflecting $\tilde{C}(s)$ about the anti-diagonal of the unit square.

Proof of Proposition 1.

For any pair of indices i, i', we wish to show $\beta_i \leq \beta_{i'}$ iff $\gamma_i \leq \gamma_{i'}$. Note that

$$\gamma_i = 1 - \mathrm{pFDR}(G_{P,\beta_i}) = \frac{\pi\mu(G^c_{P,\beta_i})}{Q(G^c_{P,\beta_i})} = \left[1 + \frac{1-\pi}{\pi}\frac{P(G^c_{P,\beta_i})}{\mu(G^c_{P,\beta_i})}\right]^{-1} = \left[1 + \frac{1-\pi}{\pi}\frac{1-\beta_i}{C(1-\beta_i)}\right]^{-1}. \quad (1)$$

So $\gamma_i \leq \gamma_{i'}$ iff $C(1-\beta_{i'})/(1-\beta_{i'}) \geq C(1-\beta_i)/(1-\beta_i)$, which is true by assumption. □

Proof of Proposition 2:

To establish the first statement, that $G_{P,\beta}$ is the Q-GQ set at level $\tilde{\beta}$, we must establish (a)

$Q(G_{P,\beta}) \geq \tilde{\beta}$ and (b) if $Q(G) \geq \tilde{\beta}$, then $\mu(G) \geq \mu(G_{P,\beta})$. To establish (a), observe

$$Q(G_{P,\beta}) = \pi\mu(G_{P,\beta}) + (1-\pi)P(G_{P,\beta}) \geq \pi\mu(G_{P,\beta}) + (1-\pi)\beta = \tilde{\beta}.$$

To establish (b), assume it does not hold. That is, assume there exists G such that $Q(G) \geq \tilde{\beta}$ and $\mu(G) < \mu(G_{P,\beta})$. Then

$$P(G) = \frac{Q(G) - \pi\mu(G)}{1-\pi} \geq \frac{\tilde{\beta} - \pi\mu(G)}{1-\pi} \geq \frac{\tilde{\beta} - \pi\mu(G_{P,\beta})}{1-\pi} = \beta,$$

which contradicts the definition of $G_{P,\beta}$ as the P-GQ set at level β.

To prove the second half of the proposition, consider $0 \leq \tilde{\beta} \leq \tilde{\beta}_{\max}$. Consider the function $\tau(\beta') := \pi\mu(G_{P,\beta'}) + (1-\pi)P(G_{P,\beta'})$. Since, by assumption **[C]**, f has no plateaus, $P(G_{P,\beta'}) = \beta'$. In addition, since μ is absolutely continuous with respect to Lebesgue measure, by assumption **[B]**, $\mu(G_{P,\beta'})$ is continuous and nondecreasing. Therefore τ is continuous and increasing as a function of $0 \leq \beta' \leq 1$, taking values between 0 and $\tilde{\beta}_{\max}$. By the intermediate value theorem, there exists β' such that $\tilde{\beta} = \tau(\beta') = \pi\mu(G_{P,\beta'}) + (1-\pi)\beta'$. Furthermore, this β' is unique since τ is increasing. By the first part of this theorem, we conclude $G_{Q,\tilde{\beta}} = G_{P,\beta'}$. Combining this fact with the equation

$$\pi\mu(G_{Q,\tilde{\beta}}) + (1-\pi)\beta = \pi\mu(G_{P,\beta'}) + (1-\pi)\beta',$$

which results from equating two different expressions for $\tilde{\beta}$, we conclude that $\beta = \beta'$. Since the P-GQ sets are unique, it follows that $G_{Q,\tilde{\beta}} = G_{P,\beta}$. □

Proof of Corollary 1

Consider first the case $X_i \in G_{P,1}$, which implies $\tilde{\beta}_i \leq \tilde{\beta}_{\max}$. By Proposition 2 we have that $G_{P,\beta_i} = G_{Q,\tilde{\beta}_i}$. By Bayes' rule,

$$\begin{aligned}
\gamma_i &= \Pr\left(Y = 1 | X \notin G_{P,\beta_i}\right) = \frac{\pi\mu(G^c_{P,\beta_i})}{Q(G^c_{P,\beta_i})} \\
&= \frac{\pi(1-\mu(G_{P,\beta_i}))}{1-Q(G_{P,\beta_i})} = \frac{\pi(1-\mu(G_{Q,\tilde{\beta}_i}))}{1-Q(G_{Q,\tilde{\beta}_i})}
\end{aligned}$$

$$= \quad \frac{\pi(1-\mu(G_{Q,\tilde{\beta}_i}))}{1-\tilde{\beta}_i} \; .$$

If $X \notin G_{P,1}$, then $\beta_i = 1$, and $G_{P,\beta_i} = G_{P,1}$ is a subset of $G_{Q,\tilde{\beta}_i}$. Thus

$$\begin{aligned}
\frac{\pi(1-\mu(G_{Q,\tilde{\beta}_i}))}{1-\tilde{\beta}_i} &= \frac{\pi(1-\mu(G_{Q,\tilde{\beta}_i}))}{1-Q(G_{Q,\tilde{\beta}_i})} = \frac{\pi(1-\mu(G_{Q,\tilde{\beta}_i}))}{1-(\pi\mu(G_{Q,\tilde{\beta}_i})+(1-\pi)P(G_{Q,\tilde{\beta}_i}))} \\
&= \frac{\pi(1-\mu(G_{Q,\tilde{\beta}_i}))}{\pi(1-\mu(G_{Q,\tilde{\beta}_i}))} = 1
\end{aligned}$$

which is the value of γ_i in this case. □

Proof of Proposition 3:

$$\begin{aligned}
\Pr(Z \le t | X \sim Q) &= \Pr(\mu(G(X)) \le t | X \sim Q) = Q(\{\mu(G(X)) \le t\}) \\
&= Q(G_{Q,\tilde{D}(t)}) = \tilde{D}(t).
\end{aligned}$$

Thus $\tilde{D}(t) = \pi \Pr(Z \le t | X \sim \mu) + (1-\pi)\Pr(Z \le t | X \sim P)$. Now

$$\begin{aligned}
\Pr(Z \le t | X \sim \mu) &= \Pr(\mu(G(X)) \le t | X \sim \mu) = \mu(\{\mu(G(X)) \le t\}) \\
&= \mu(G_{Q,\tilde{D}(t)}) = t.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\Pr(Z \le t | X \sim P) &= \Pr(\mu(G(X)) \le t | X \sim P) = P(\{\mu(G(X)) \le t\}) \\
&= P(G_{Q,\tilde{D}(t)}) = P(G_{P,D(t)}) = D(t).
\end{aligned}$$

The result follows by differentiating $\tilde{D}(t)$. □