Supplementary Material to "Automatic Classifiers as Scientific Instruments: One Step Further Away from Ground-Truth"

J. Whitehill and A. Ramakrishnan

1 Proofs

1.1 Proof of Proposition 2

We prove the proposition for the case that $r < 0$; the case for $r > 0$ is similar.

From Section III, we have that

$$\rho(\widehat{\mathsf{u}}, \mathbf{v}) = qr + \hat{\mathsf{u}}_3 \sqrt{1 - r^2}$$

Since each $\hat{\mathsf{u}}_i$ ($i = 3, 4, \ldots, n$) is a coordinate on an $(n-3)$-sphere, it can be re-parameterized [Muller(1959)] by sampling $n - 2$ standard normal random variables and normalizing, i.e.:

$$\hat{\mathsf{u}}_i = \frac{\sqrt{1-q^2} \times \mathsf{z}_i}{\sqrt{\sum_{j=3}^n \mathsf{z}_j^2}}$$

where each $\mathsf{z}_i \sim \mathbb{N}(0,1)$. A false positive correlation thus occurs when $\hat{u}_3$ is at least $c = |qr|/\sqrt{1-r^2}$ more than its expected value qr:

$$\Pr[\hat{\mathsf{u}}_3 \geq c] = \Pr\left[\frac{\sqrt{1-q^2} \times \mathsf{z}_3}{\sqrt{\sum_{j=3}^n \mathsf{z}_j^2}} \geq c\right]$$

Due to the inequality, we must handle the cases that $\mathsf{z}_3 \geq 0$ and $\mathsf{z}_3 < 0$ separately. Note that the latter case contributes 0 probability since $c \geq 0$ and $q > 0$. Also, since z_3 is a standard normal random variable, $\Pr[\mathsf{z}_3 \geq 0] = 0.5$.

$$\begin{aligned}
&\Pr[\hat{u}_3 \geq c] \\
&= \Pr\left[\frac{\sqrt{1-q^2} \times \mathsf{z}_3}{\sqrt{\sum_{j=3}^n \mathsf{z}_j^2}} \geq c \;\middle|\; \mathsf{z}_3 \geq 0\right] \Pr[\mathsf{z}_3 \geq 0] + \\
&\quad \Pr\left[\frac{\sqrt{1-q^2} \times \mathsf{z}_3}{\sqrt{\sum_{j=3}^n \mathsf{z}_j^2}} \geq c \;\middle|\; \mathsf{z}_3 < 0\right] \Pr[\mathsf{z}_3 < 0] + \\
&= \frac{1}{2}\Pr\left[\frac{\sqrt{1-q^2} \times \mathsf{z}_3}{\sqrt{\sum_{j=3}^n \mathsf{z}_j^2}} \geq c \;\middle|\; \mathsf{z}_3 \geq 0\right] + 0 \\
&= \frac{1}{2}\Pr\left[(1-q^2)\mathsf{z}_3^2 \geq c^2 \sum_{j=3}^n \mathsf{z}_j^2\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\Pr\left[(1-q^2-c^2)\mathsf{z}_3^2 \geq c^2 \sum_{j=4}^{n} \mathsf{z}_j^2\right] \\
&= \frac{1}{2}\Pr\left[\mathsf{z}_3^2 \geq \frac{c^2}{(1-q^2-c^2)} \sum_{j=4}^{n} \mathsf{z}_j^2\right]
\end{aligned}
$$

For $n > 3$, each side of the inequality is a sum of squared normally distributed random variables, i.e., a χ^2-random variable (though with different degrees of freedom). We can thus rewrite this probability as

$$
\begin{aligned}
\Pr[\hat{u}_3 \geq c] &= \frac{1}{2}\Pr\left[\chi_1^2 \geq \left(\frac{c^2}{1-q^2-c^2}\right)\chi^2_{(n-3)}\right] \\
&= \frac{1}{2}\int_0^\infty f_1(t)F_{n-3}\left(\frac{1-q^2-c^2}{c^2}t\right)dt \\
&\doteq h(n,q,r)
\end{aligned}
$$

where χ_1^2 and $\chi^2_{(n-3)}$ are χ^2 random variables with 1 and $(n-3)$ degrees of freedom, respectively. The probability is equivalent to the integral because, for any value t of the χ_1^2 variable, we require that the χ^2_{n-3} variable be less than t (after applying a scaling factor). To our knowledge, there is no closed formula for this integral, but we can compute it numerically. For $n = 3$, we have

$$
\begin{aligned}
\Pr[\hat{u}_3 \geq c] &= \frac{1}{2}\Pr\left[(1-q^2-c^2)\mathsf{z}_3^2 \geq 0\right] \\
&= \frac{1}{2}\Pr[c^2 \leq 1-q^2]
\end{aligned}
$$

since a χ^2-random variable is non-negative, and where the probability of $c^2 \leq 1-q^2$ is 1 if the inequality is true and 0 otherwise.

1.2 Proof of Proposition 3

For convenience, define $\alpha = \frac{1-q^2-c^2}{c^2}$.

$$
\begin{aligned}
&h(n+1,q,r) - h(n,q,r) \\
&\quad = \int_0^\infty \left[f_1(t)F_{(n+1)-3}(\alpha t) - f_1(t)F_{n-3}(\alpha t)\right]dt \\
&\quad = \int_0^\infty f_1(t)\left[F_{n-2}(\alpha t) - F_{n-3}(\alpha t)\right]dt
\end{aligned}
$$

Ghosh [Ghosh(1973)] proved that, for any fixed $t > 0$, $\Pr[\chi_k^2 > t]$ is monotonically increasing in the degrees of freedom k; hence, $F_k(t)$ is monotonically decreasing in k. Therefore, $F_{n-2}(\alpha t) - F_{n-3}(\alpha t) < 0$ for all t. Since f_k is a non-negative function for all k, then the integral in Equation 1 must be negative; hence, h is monotonically decreasing in n for every $c > 0$ and $q \in (0, 1]$.

1.3 Proof of Proposition 4

First, we show that α is monotonically decreasing in q^2:

$$\begin{aligned}
\alpha(q) &= \frac{1-q^2-c^2}{c^2} \\
&= \frac{1-q^2-q^2r^2/(1-r^2)}{q^2r^2/(1-r^2)} \\
&= \frac{(1-r^2)(1-q^2)-q^2r^2}{q^2r^2} \\
&= \frac{1-r^2-q^2}{q^2r^2} \\
&= \frac{1-r^2}{q^2r^2} - \frac{1}{r^2}
\end{aligned}$$

The first term is monotonically decreasing in q^2, and the second term is constant in q^2.

Next, let ϵ be a positive real number such that $q+\epsilon \leq 1$:

$$\begin{aligned}
& h(n, q+\epsilon, r) - h(n, q, r) \\
&= \int_0^\infty f_1(t) F_{n-3}\left(\alpha(q+\epsilon)t\right) dt - \\
&\quad \int_0^\infty f_1(t) F_{n-3}\left(\alpha(q)t\right) dt \\
&= \int_0^\infty f_1(t) \left[F_{n-3}\left(\alpha(q+\epsilon)t\right) - F_{n-3}\left(\alpha(q)t\right)\right] dt
\end{aligned}$$

Since F_{n-3} is monotonically *increasing*, then the expression in brackets is negative. Since f_1 is non-negative, then the entire integral must be less than 0.

2 Sampling distribution $\mathbf{Pr}(\hat{\mathsf{q}} \mid q, n)$

The sampling distribution can be computed exactly [Fisher(1915)], but this is computationally feasible only for small n. Hence, we use the approximation from Soper [Soper(1913)]: Let q denote the population Pearson correlation coefficient, and let $\hat{\mathsf{q}}$ denote the sample correlation from n data. Then

$$\begin{aligned}
\Pr(\hat{\mathsf{q}} \mid q, n) &\propto (1-\hat{\mathsf{q}})^{m_1}(1+\hat{\mathsf{q}})^{m_2} \\
m_1 &= \frac{1}{2}(\lambda-1)(1-\mu_q) - 1 \\
m_2 &= \frac{1}{2}(\lambda-1)(1+\mu_q) - 1 \\
\lambda &= (1-\mu_q^2)/\sigma_q^2
\end{aligned}$$

$$
\begin{aligned}
\sigma_q &= \frac{(1-q^2)}{\sqrt{n}}\left(1+\frac{(1+5.5q^2)}{2n}\right) \\
\mu_q &= \sqrt{q^2-\frac{c}{n}-\frac{c(1+5q^2)}{2n^2}} \\
c &= q^2(1-q^2)
\end{aligned}
$$

References

[Fisher(1915)] Fisher, R. A. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

[Ghosh(1973)] Ghosh, B. Some monotonicity theorems for χ^2, F and t distributions with applications. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 480–492, 1973.

[Muller(1959)] Muller, M. E. A note on a method for generating points uniformly on n-dimensional spheres. *Communications of the ACM*, 2(4):19–20, 1959.

[Soper(1913)] Soper, H. On the probable error of the correlation coefficient to a second approximation. *Biometrika*, 9(1/2):91–115, 1913.