General mapper proof

Mohsen Sadatsafavi

2024-01-10

Consider a parametric family of probability distributions with support on [0,1], with the following characteristics:

- the CDF is smooth,
- the distribution is quantile-identifiable: being fully identifiable by knowing a pair of its quantile values.

We note that common two-parameter distributions for probabilities, such as the beta ($\pi \sim Beta(\alpha, \beta)$), logit normal($logit(\pi) \sim Normal(\mu, \sigma^2)$) and probit-normal ($\Phi^{-1}(\pi) \sim Normal(\mu, \sigma^2)$) satisfy the above criteria. All these distributions are smooth. The quantile-identifiability of the beta distribution is proven in Shih et al. For the logit-normal and probit-normal distributions, it is immediately deduced from the monotonical link to the normal distribution and the quantile-identifiablity of the latter.

Lemma: For a class of probability distirbutions with the above characteristics, the combination of expected value and c-statistic uniquly identifies the distirbution.

Proof

Let F be the CDF from the family of distributions of interest. Let m be its first moment, and c the c-statistic, defined as $c = P(\pi_2 > \pi_1 | Y_2 > Y_1)$ where $\pi_i \sim F$ and $Y_i \sim Bernoulli(\pi_i)$ a realization of response value given the probability.

We shall prove that Fs is uniquely identifiable from $\{m, c\}$.

First, we apply Bayes' theorem to the definition of c:

$c := P(Y_2 > Y_2 | \pi_2 > \pi_1) \frac{P(\pi_2 > \pi_1)}{P(Y_2 > Y_1)} = P(Y_2 = 1, Y_1 = 0 | \pi_2 > \pi_1) \frac{0.5}{P(Y_2 = 1, Y_1 = 0)} = \frac{\mathbf{E}\pi_{max}(1 - \mathbf{E}\pi_{min})}{2m(1-m)} = \frac{\mathbf{E}\pi_{max}(1 - 2m + \mathbf{E}\pi_{max})}{2m(1-m)}$,

where π_{max} and π_{min} are, respectively, the maximum and minimum of a pair of independent RVs from F. $\mathbf{E}\pi_{max}(1 - 2m + \mathbf{E}\pi_{max})$ is monotonical with respect to $\mathbf{E}\pi_{max}$. Further, let $f(x) := dF(x)/dx$ be the PDF of F. We have $\mathbf{E}\pi_{max} = \int_0^1 2xf(x)F(x)dx = xF^2(x)|_0^1 - \int_0^1 F^2(x)dx = 1 - \int_0^1 F^2(x)dx$; i.e., $\mathbf{E}\pi_{max}$ is monotonically related to $\int_0^1 F^2(x)dx$. As such, the goal is achieved by showing that $\{m, \int_0^1 F^2(x)dx\}$ uniquely identifies F.

The rest of the proof is by contradiction. We show that two different CDFs F_1 and F_2 with the same m cannot result in the same $\int_0^1 F^2(x)dx$.

To proceed, we note that for probability distributions with support on [0,1], the equality of means indicates the equality of the the area under CDFs, as (by integration by parts) $m = \int_0^1 xf(x)dx = xF(x)|_0^1 - \int_0^1 F(x)dx = 1 - \int_0^1 F(x)dx$.

Given that both CDFs are anchored at (0,0) and (1,1), are smooth, have the same area under the CDF but are not equal at all points, they must cross. However, they can only cross once, given the quantile-identifiability requirement (if they cross two or more times, any pairs of quantiles defined by the crossing points would fail to identify them uniquely).

Let z be the unique crossing point of the two CDFs (where $F_1(z) = F_2(z)$). We break $\int_0^1 F^2(x)dx$ into two parts around z:

$\int_1^1 (F_1^2(x) - F_2^2(x))dx = \int_1^z (F_1^2(x) - F_2^2(x))dx + \int_z^1 (F_1^2(x) - F_2^2(x))dx = \int_1^z (F_1(x) - F_2(x))(F_1(x) + F_2(x))dx + \int_z^1 (F_1(x) - F_2(x))(F_1(x) + F_2(x))dx$.

Without loss of generality, assume we label Fs such that $F_1(x) > F_2(x)$ when $x \in [0, z)$. In this region, barring $x = 0$, $F_1(x) - F_2(x) > 0$, and (due to Fs monotonically increasing) $0 < F_1(x) + F_2(x) < F_1(z) + F_2(z)$. As such, replacing $F_1(x) + F_2(x)$ by the larger positive quantity $F_1(z) + F_2(z)$ will increase this term. As well, in the $x \in [z, 1]$ region, barring the two points where $x \in z, 1$, $F_1(x) - F_2(x) < 0$, and $0 < F_1(x) + F_2(x) < F_1(z) + F_2(z)$. As such, replacing $F_1(x) + F_2(x)$ by the smaller positive quantity $F_1(z) + F_2(z)$ will also increase this term. Therefore we have

$\int_0^1 (F_1^2(x) - F_2^2(x))dx < \int_1^z (F_1(x) - F_2(x))(F_1(z) + F_2(z))dx + \int_z^1 (F_1(x) - F_2(x))(F_1(z) + F_2(z))dx$, and the term on the right hand side is zero because of the equality of the area under the CDFs. Therefore, $\int_0^1 (F_1^2(x) - F_2^2(x))dx < 0$, contradicting their equality.