Measuring the similarity of forecasts provided in a quantile format

Johannes Bracher, johannes.bracher@kit.edu

(Please excuse that this document contains a lot of hand waving. I can write things up more formally if this becomes of interest.)

Short version

Consider two predictive distributions F and G. Their Cramer distance is defined as

$$\text{CD}(F, G) = \int_{-\infty}^{\infty} (F(x) - G(x))^2 dx$$

where $F(x)$ and $G(x)$ denote the cumulative distribution functions. Now assume that for each of the distributions we only know K quantiles at equally spaced levels $1/(K+1), 2/(K+1), \dots, K/(K+1)$. Denote these quantiles by $q_1^F, \dots, q_K^F$ and $q_1^G, \dots, q_K^G$, respectively. We introduce the following notations:

- $\mathbf{q}$ is a vector of length $2K$. It is obtained by pooling the $q_k^F, q_k^G, k = 1, \dots, K$ and ordering them in increasing order (ties can be ordered in an arbitrary manner).
- $\mathbf{a}$ is a vector of length $2K$ containing the value 1 wherever $\mathbf{q}$ contains a quantile of F and -1 wherever it contains a value of G.
- $\mathbf{b}$ is a vector of length $2K$ containing the absolute cumulative sums of $\mathbf{a}$, i.e. $b_i = \left|\sum_{j=1}^{i} a_j\right|$.

Then a quantile-based approximation of the Cramer distance is given by

$$\text{CD}(F, G) \approx \frac{1}{K(K+1)} \times \sum_{i=1}^{2K-1} b_i(b_i + 1)(q_{i+1} - q_i).$$

Actually I think that the approximation for small K is better if we use

$$\text{CD}(F, G) \approx \frac{1}{(K+1)^2} \times \sum_{i=1}^{2K-1} b_i^2(q_{i+1} - q_i),$$

which can be motivated much more directly, too (without the detour of the WIS approximation). It's just a bit further from the WIS-type approximation, but I don't see why that would be a problem.

Long version

Relationship between Cramer von Mises distance, CRPS and WIS

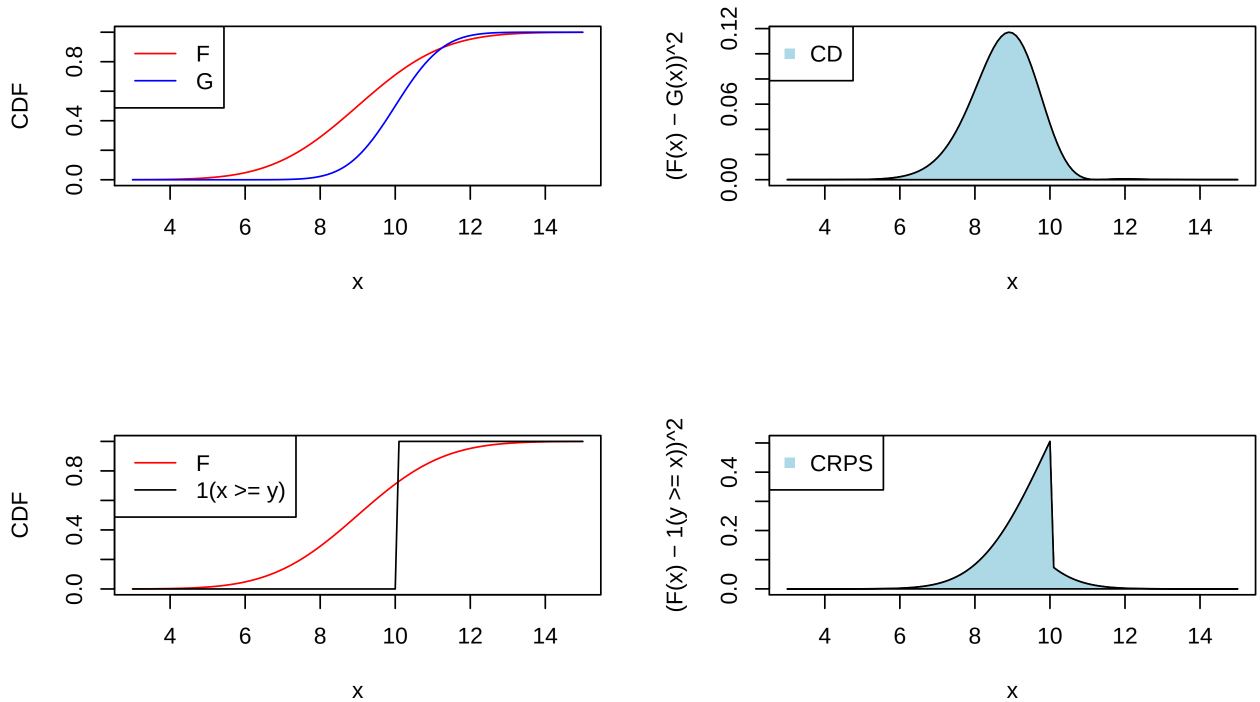
Consider two predictive distributions F and G and an observed value y. In the following we use F and G also to denote the respective cumulative distribution functions (CDFs). The similarity of the definitions of the Cramer von Mises distance (CD)

$$\text{CD}(F, G) = \int_{-\infty}^{\infty} (F(x) - G(x))^2 dx$$

and the CRPS

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(x) - \mathbf{1}(x \geq y))^2 dx$$

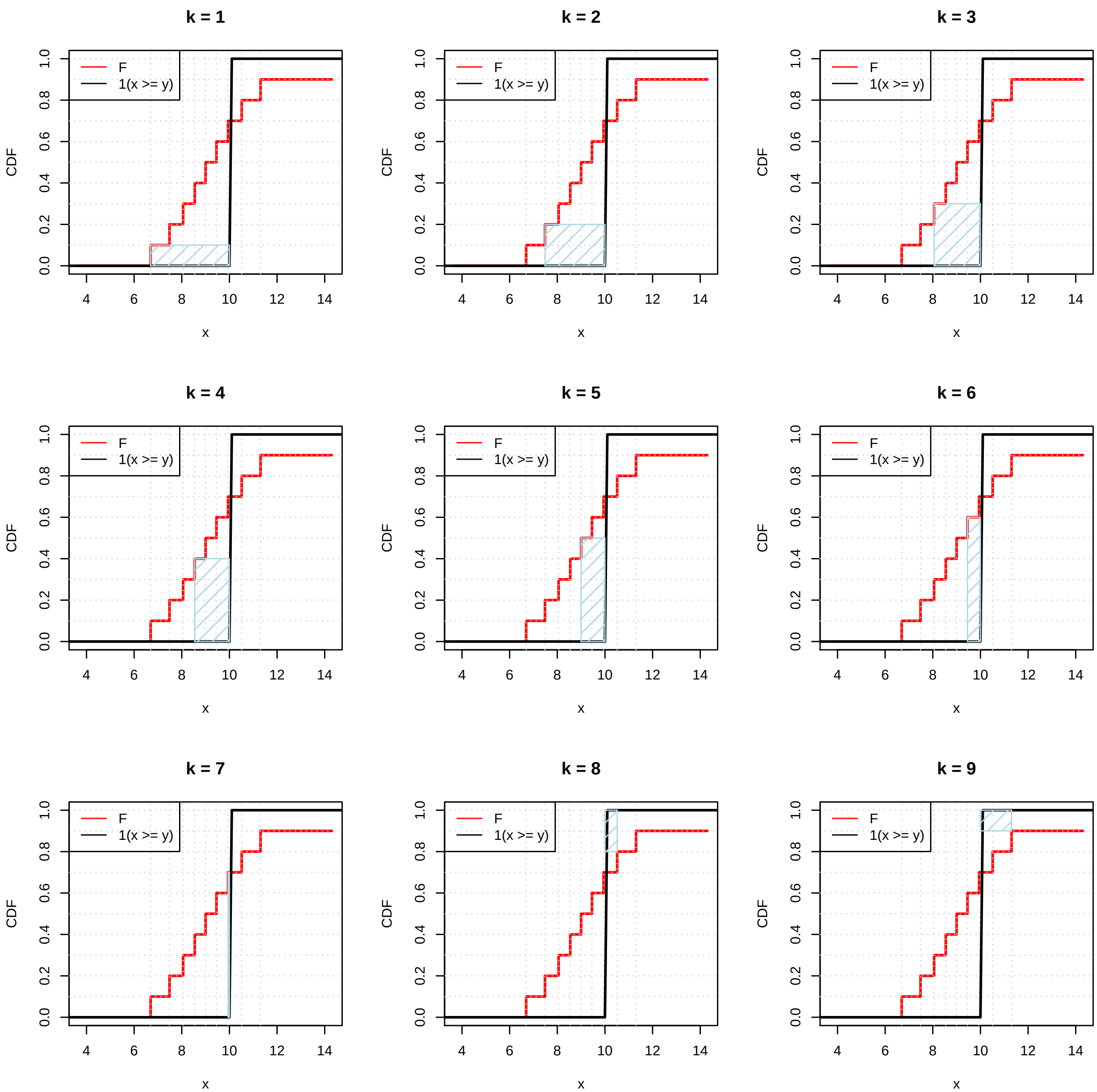
is obvious. Indeed, we can interpret $\mathbf{1}(x \geq y)$ as the CDF of a random variable which always takes the value y. In a sense the CRPS is thus a special case of the CD. The two integrals are illustrated in the figure below.



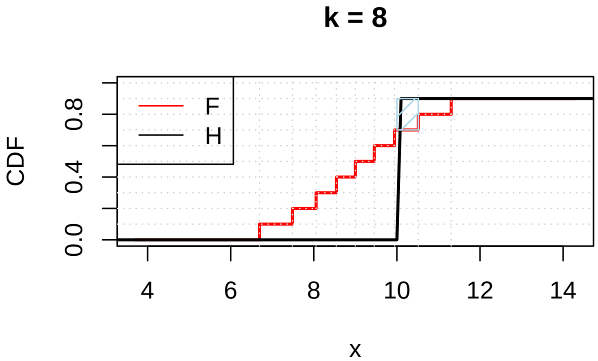
Now assume F is provided in a quantile format, with K quantiles at equally spaced levels $1/(K+1), 2/(K+1), \dots, K/(K+1)$. Denote these quantiles by $q_k^F, k = 1, ..., K$. The weighted interval score, or mean linear quantile score

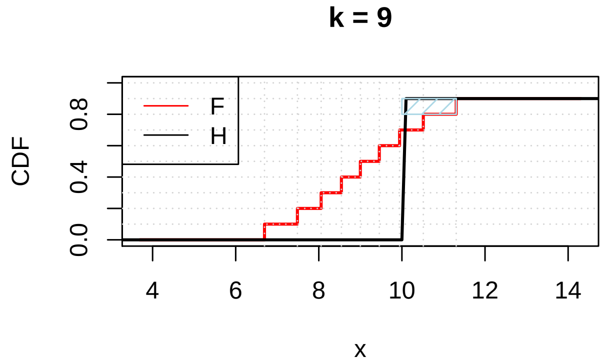
$$\text{WIS}(F, y) = \frac{2}{K} \times \sum_{k=1}^{K} \{\mathbf{1}(y \leq q_k^F)\} \times (q_k^F - y) \approx \text{CRPS}(F, y)$$

is a commonly used quantile-based approximation of the CRPS. We start by getting an intuition of what this approximation does and why it works. For the following we use $K = 9$ for illustration.



The above figure shows rectangles representing the $K = 9$ terms which are averaged to obtain the WIS. Note that in the above figure the black line for $\mathbf{1}(x \geq y)$ is a "full" CDF in the sense that it reaches the value one. This is a problem if we want to translate things to a more general setting with a step function with jumps at K quantiles of a distribution H. In this case, the right horizontal part of the black line would need to be at $K/(K + 1) = 0.9$ instead of 1, as is the case for the red line. Fortunately we can just shift down everything that happens right of y by $1/(K + 1) = 0.1$ and maintain a nice geometrical intuition (maybe even a nicer one):





The WIS is given by the average of the size of the K light blue boxes. To get a better understanding of what these represent we slice up the boxes vertically at the K quantiles $q_1^F, \dots, q_K^F$ of F and horizontally at the quantile levels $1/(K+1), \dots K/(K+1)$ (dotted lines in plots). Then obviously all of the resulting small boxes are the same height $1/(K+1)$ and we have the following (specific to the example shown in the plot):

- One box of width $(q_2^F - q_1^F)$ (the one from $k = 1$)
- Three boxes of width $(q_3^F - q_2^F)$ (one from $k = 1$ and two from $k = 2$)
- Six boxes of width $(q_4^F - q_3^F)$ (one from $k = 1$, two from $k = 2$ and three from $k = 3$)

More generally, for segments where $|F(x) - H(x)| = k/(K+1)$ we have $\sum_{i=1}^k i = k(k+1)/2$ boxes. This means that we can re-write the WIS as

$$\text{WIS}(F, y) = \frac{2}{K} \times \sum_{k=1}^{K} \underbrace{\mu(\{x : |F(X) - H(x)| = k/(K+1)\})}_{\text{width of boxes}} \times \underbrace{\frac{1}{K+1}}_{\text{height of boxes}} \times \underbrace{\frac{k(k+1)}{2}}_{\text{number of boxes with respective width}}, \quad (1)$$

$$= \sum_{k=1}^{K} \mu(\{x : |F(x) - H(x)| = k/(K+1)\}) \times \underbrace{\frac{k}{K} \times \frac{k+1}{K+1}}_{\approx (F(x) - H(x))^2}$$

where $\mu(\{x : |F(x) - H(x)| = 1/k\})$ is the total length of the segments where $|F(x) - H(x)| = 1/k$. This makes it clear that for large K the approximation

$$\text{WIS}(F, y) \approx \int_{-\infty}^{\infty} (F(x) - H(x))^2 dx$$

indeed holds with

$$H(X) = \begin{cases} 0 \text{ if } x < y \\ K/(K+1) \text{ if } x \geq y. \end{cases}$$

Extension to comparison of two predictive distributions

Formulation (1) and its motivation are sufficiently general that they can also apply it to two "actual" CDFs which both have more than just one jump. Indeed, we can (1) it to approximate the Cramer von Mises distance of any pair of distributions F and G:

$$\text{CD}(F, G) \approx \sum_{k=1}^{K} \mu(\{x : |F(x) - H(x)| = k/(K+1)\}) \times \frac{k}{K} \times \frac{k+1}{K+1}$$

This expression could be evaluated directly (or at least approximated very closely) using a grid for the x values. However, a simpler way to compute this exists (writing down why it works is tedious and I need to clarify some details, but I've checked in numerous examples and it does work). Denote by $\mathbf{q}$ a vector of length $2K$ containing all quantiles $q_1^F, \dots, q_F^K$ of F and $q_1^G, \dots, q_G^K$ in increasing order; and denote by $\mathbf{a}$ a vector of length $2K$ with 1 wherever the corresponding entry of $\mathbf{q}$ comes from the quantiles of F and -1 if it comes from the quantiles of G. Then the above approximation can also be written as:

$$\mathrm{CD}(F, G) \approx \frac{1}{K(K+1)} \times \sum_{k=1}^{2K-1} b_k(b_k + 1)(q_{i+1} - q_i),$$

where

$$b_k = \left| \sum_{i=1}^{k} a_k \right|.$$

Some advantages of this measure:

- Can be computed quickly for a large number of pairs of forecasts.
- Follows the same philosophy as the WIS.
- Similarly to the WIS it can be decomposed into various components (e.g. according to whether $F(X)$ or $G(x)$ predicts larger values or whether one of them is more dispersed).
- If G is just an additively shifted version of F the CD corresponds to the (absolute value of the) shift. This means that the measure can again be interpreted on the natural scale of the data as a sort of generalized absolute difference.

Example

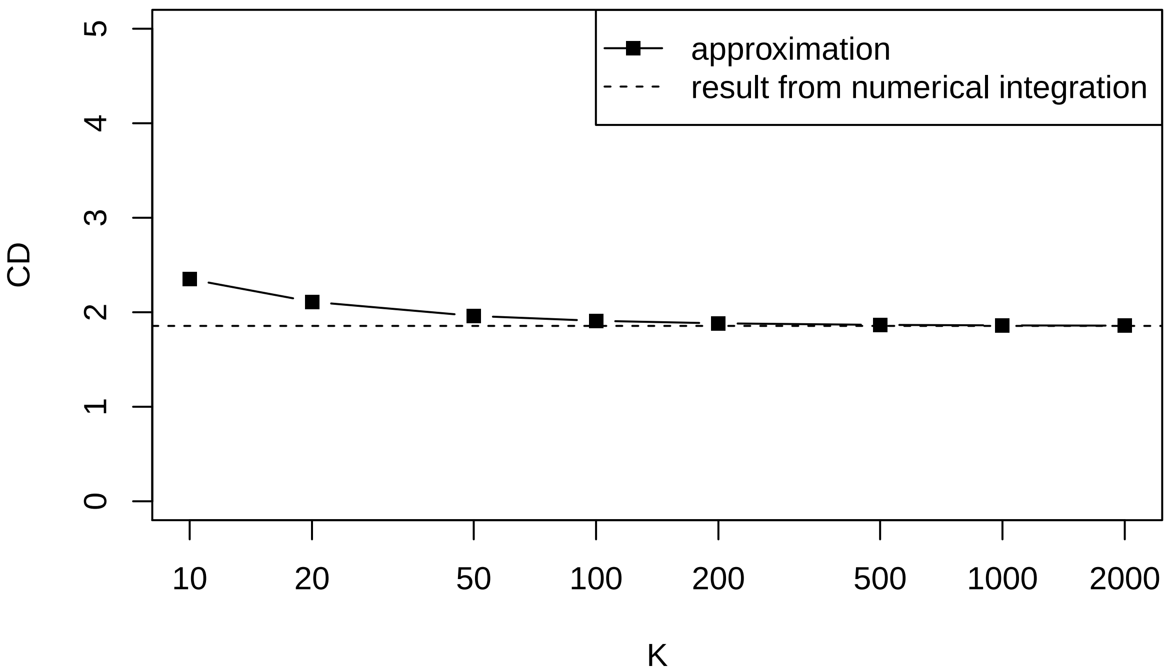
This is a simple R function to compute the approximation of the CD:

```r
# q_F: vector containing the (1:K)/(K + 1) quantiles of F
# q_G: vector containing the (1:K)/(K + 1) quantiles of G
approx_cd <- function(q_F, q_G){

  # compute quantile levels from length of provided quantile vectors:
  K <- length(q_F)
  if(length(q_G) != K) stop("q_F and q_G need to be of the same length")
  p <- (1:K)/(K + 1) # function assumes that the quantile levels are equally spaced

  # pool quantiles:
  q0 <- c(q_F, q_G)
  # vector of grouping variables, with 1 for values belonging to F, -1 for values
  # belonging to G
  a0 <- c(rep(1, length(q_F)), rep(-1, length(q_G)))

  # re-order both vectors:
  q <- q0[order(q0)]
  a <- a0[order(q0)]
  # and compute "how many quantiles ahead" F or G is at a given segment:
  b <- abs(cumsum(a))

  # compute the lengths of segments defined by sorted quantiles:
  diffs_q <- c(diff(q), 0) # zero necessary for indexing below, but we could put
  # anything (gets multiplied w zero)

  # and approximate CD
  cvm <- sum(diffs_q*b*(b + 1))/(K + 1)/(K)

  return(mean(cvm))
}
```

As an example we apply the approximation to the distributions F and G already used in the figures above. They are given by $N(9, 1.8)$ for F and $N(10, 1)$ for G. The figure below shows the value of the approximation with different values for K, with the "exact" value obtained by numerical integration via a very fine grid (dashed line).

```r
# define distributions:
mu_F <- 14
sigma_F <- 1.8

mu_G <- 10
sigma_G <- 2.5

# simple numerical integration using grid:
grid_x <- seq(from = -10, to = 25, by = 0.1)
p_F <- pnorm(grid_x, mu_F, sigma_F)
p_G <- pnorm(grid_x, mu_G, sigma_G)
exact_cd <- 0.1*sum((p_F - p_G)^2)

# values of K to check:
values_K <- c(10, 20, 50, 100, 200, 500, 1000, 2000)
```

```r
# vector to store results:
values_approx_cd <- numeric(length(values_K))

# apply approximation for each_
for(i in seq_along(values_K)){
  p_temp <- (1:(values_K[i] - 1))/values_K[i] # quantile levels
  q_F_temp <- qnorm(p_temp, mu_F, sigma_F) # quantiles of F
  q_G_temp <- qnorm(p_temp, mu_G, sigma_G) # quantiles of G
  values_approx_cd[i] <- approx_cd(q_F = q_F_temp, q_G = q_G_temp) # approximation
}

# plot:
plot(values_K, values_approx_cd, ylim = c(0, 5), xlab = "K", ylab = "CD",
     pch = 15, type = "b", log = "x")
abline(h = exact_cd, lty = 2)
legend("topright", c("approximation", "result from numerical integration"),
       pch = c(15, NA), lty = c(1, 2))
```



Useful links

- http://pages.stat.wisc.edu/~wahba/stat860public/pdf4/Energy/EnergyDistance10.1002-wics.1375.pdf
- https://en.wikipedia.org/wiki/Energy_distance]