Notes on RMSD inference

for Alexander, Fábio, Ruth

Luca <pgl@portamana.org>

16 June 2021; updated 4 July 2021

1 Overview

We consider a quantity, the Root-Mean-Squared Distance, denoted by y, and a set of quantities called "features", denoted by x. A value of RMSD y and values for the features x can be associated with each target-ligand pair, which I'll call a "datapoint".

Our problem is to infer y_0 given: x_0 for a new datapoint, known pairs $(\bar{y}, \bar{x}) \coloneqq \bigl((y_1, x_1), (y_2, x_1), \dots, (y_N, x_N)\bigr)$ for other N datapoints, and additional facts and hypotheses H. Additional hypotheses are unavoidable since we are making an extrapolation. Our uncertainty about y_0 is expressed by a probability (density)

$$\mathrm{p}(y_0 \mid x_0,\ \bar{y}, \bar{x},\ H)\,, \tag{1}$$

which we want to quantify.

The first question we must ask in approaching this problem is how much our uncertainty about y_0 depends on the relative frequencies for previous datapoints $(\bar{y}, \bar{x})$. For example, if y_0 is hand-picked by someone, then knowledge of the frequencies for $(\bar{y}, \bar{x})$ is irrelevant. But even in such a case it might be that the reversed probability $\mathrm{p}(x_0 \mid y_0,\ \bar{y}, \bar{x},\ H)$ *does* depend on the frequencies of previous datapoints, and it can be used to calculate (1) as

$$\mathrm{p}(y_0 \mid x_0,\ \bar{y}, \bar{x},\ H) = \frac{\mathrm{p}(x_0 \mid y_0,\ \bar{y}, \bar{x},\ H)\, \mathrm{p}(y_0 \mid H)}{\sum_{y_0} \mathrm{p}(x_0 \mid y_0,\ \bar{y}, \bar{x},\ H)\, \mathrm{p}(y_0 \mid H)}\,. \tag{2}$$

The important difference between eqs (1) and (2) is that in the latter we must also quantify $\mathrm{p}(y_0 \mid H)$ using extra-data facts or hypotheses.

In technical terms we are asking whether y_0 is *exchangeable* given x_0, or vice versa. Or using Fisherian (1956 §§ II.4, IV.1) parlance we are

This document is designed for screen reading and two-up printing on A4 or Letter paper asking whether y_0 should be considered as belonging to a *subpopulation* determined by x_0 or vice versa. (Or neither, in which case our study would simply end here, so we won't consider this third possibility.) Another way of seeing our question is whether the "causal connections" mainly go in the direction $x_0 \rightsquigarrow y_0$ or $y_0 \rightsquigarrow x_0$ (cf. Pearl 1988 §§ 2.1.2, 2.2.5).

This question is discussed in a brilliant paper by Lindley & Novick (1981), who show with examples its fundamental importance for making correct inferences.

In this study we'll consider both possibilities.

2 Direct case: methodology

For the direct case (1) we shall consider the RMSD y as a continuous variable, mapped to a log-scale.

The assumption in this case is that y_0 is exchangeable given x_0, but x_0 is not exchangeable given the known datapoints $\bar{x}$. This a sensible assumption – a fact indeed – if we are actually *choosing* the specific value of x_0, as we do in testing for example.

The calculation of $\mathrm{p}(y_0|x_0,\ \bar{y}, \bar{x},\ H)$ is actually done via the calculation of the joint probability $\mathrm{p}(y_0, x_0 \mid \bar{y}, \bar{x},\ H)$ through

$$\mathrm{p}(y_0 \mid x_0,\ \bar{y}, \bar{x},\ H) = \frac{\mathrm{p}(y_0, x_0 \mid \bar{y}, \bar{x},\ H)}{\mathrm{p}(x_0 \mid \bar{y}, \bar{x},\ H)} \tag{3}$$

with the assumption that the marginal probability for x_0 is independent of the datapoints and equal to unity: $\mathrm{p}(x_0 \mid \bar{y}, \bar{x},\ H) = \mathrm{p}(x_0 \mid H) = 1$, which is a sensible assumptions if we are actually *choosing* the specific value of x_0, as we for example do in testing.

3 Reverse case: methodology

For the reverse case (2) we shall consider a binned RMSD, divided into the three categories "good", "uncertain", "bad".

Bibliography

("de *X*" is listed under D, "van *X*" under V, and so on, regardless of national conventions.)

Fisher, R. A. (1956): *Statistical Methods and Scientific Inference*. (Oliver and Boyd, Edinburgh). https://archive.org/details/in.ernet.dli.2015.134555.

Lindley, D. V., Novick, M. R. (1981): *The role of exchangeability in inference*. Ann. Stat. **9**^{1}, 45–58. doi:10.1214/aos/1176345331.

Pearl, J. (1988): *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, rev. 2nd pr. (Kaufmann, San Francisco). doi:10.1016/C2009-0-27609-4.