# Notes on RMSD inference

## for Alexander, Fábio, Ruth

Luca <pgl@portamana.org>

16 June 2021; updated 5 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,
- additional facts and hypotheses $H$.

Additional hypotheses are unavoidable since we are making an extrapolation.

Our uncertainty about $y_0$ is expressed by the probability (density)

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

which we want to quantify.

The first question we must ask in approaching this problem is: how much does our uncertainty about $y_0$ depend 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 $\mathrm{p}(x_0 \,|\, y_0,\ \bar{y}, \bar{x},\ H)$ *does* depend on the frequencies of previous datapoints, and it can be used to calculate (1) as

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

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

This document is designed for screen reading and two-up printing on A4 or Letter paper

In technical terms we are asking whether $y_0$ is *exchangeable* given $x_0$, or vice versa. Using Fisherian (1956 §§ II.4, IV.1) parlance we are 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. Exchangeability is the basic fact or assumption upon which machine-learning algorithms and the calculations presented here are based. A short intuitive summary of this notion is given in the next section.

In this study we'll study both possibilities mentioned above. At the end we'll compare the results and discuss which assumption makes more sense in various circumstances, for example during testing.

## 2 Exchangeability, approximations, machine learning

Infinite exchangeability is the fact or assumption that the ordering of known and yet unknown datapoints is irrelevant for new inferences. Good reviews are given by Dawid (2013) and Bernardo & Smith (1994 § 4.2). It has the following important consequence. If we knew the infinite-limit long-run relative frequencies $\boldsymbol{F} \coloneqq (F_{y,x})$ of the values of past and future datapoints, then the probability of observing a sequence of new values would be equal to the long-run relative frequency of those values, for symmetry reasons. For example:

$$\mathrm{p}(y_1, x_1,\ y_2, x_2 \mid \boldsymbol{F}, H) = F_{y_1,x_1}\, F_{y_2,x_2}\ . \tag{3}$$

Another consequence is that conditional probabilities are also equal to the long-run *conditional* frequencies $\boldsymbol{F}_{|} \coloneqq (F_{y|x}) = (F_{y,x}/\sum_{y'} F_{y',x})$. For example:

$$\mathrm{p}(y_1, y_2 \mid x_1, x_2,\ \boldsymbol{F}, H) = F_{y_1|x_1}\, F_{y_2|x_2}\ . \tag{4}$$

If the long-run frequencies $\boldsymbol{F}$ are unknown, they can be marginalized, i.e. integrated out, provided that we have a probability distribution

$\mathrm{p}(\boldsymbol{F} \mid H)$ for them. For our examples above,

$$\mathrm{p}(y_1, x_1,\; y_2, x_2 \mid H) = \int F_{y_1,x_1}\, F_{y_2,x_2}\, \mathrm{p}(\boldsymbol{F} \mid H)\, \mathrm{d}\boldsymbol{F} \tag{5}$$

and

$$\mathrm{p}(y_1, y_2 \mid x_1, x_2,\; H) = \int F_{y_1|x_1}\, F_{y_2|x_2}\, \mathrm{p}(\boldsymbol{F} \mid H)\, \mathrm{d}\boldsymbol{F}\;. \tag{6}$$

There are several other types of exchangeability. For example, we may have *conditional* or *partial* exchangeability of $y$ given $x$ if eq. (6) holds, but not eq. (3). Lindley & Novick (1981) discuss the importance of conditional exchangeability for various inference problems, and the errors that arise if the wrong type of exchangeability is assumed.

From the generalization of the equations above we obtain other probabilities for various cases of regression, such as eqs (1) or (2), by simply using the three rules of the probability calculus. For example, if $y_0$ is assumed to be exchangeable given $x_0$, but the latter is not assumed to be exchangeable with the other $\bar{x}$ because hand-picked, we find

$$\mathrm{p}(y_0 \mid x_0,\; \bar{y}, \bar{x},\; H) = \frac{\int F_{y_0|x_0}\, F_{y_1,x_1}\, \cdots\, F_{y_N,x_N}\, \mathrm{p}(\boldsymbol{F} \mid H)\, \mathrm{d}\boldsymbol{F}}{\int F_{y_1,x_1}\, \cdots\, F_{y_N,x_N}\, \mathrm{p}(\boldsymbol{F} \mid H)\, \mathrm{d}\boldsymbol{F}}\;. \tag{7}$$

All machine-learning algorithms calculate equations such as the one above or approximations to it, for some choice of the probability $\mathrm{p}(\boldsymbol{F} \mid H)$ (MacKay 1992a; Bishop 2006).

The formulae above can be numerically implemented exactly or with a good approximation only if the features $x$ have low dimensions and the number $N$ of known data is small 🔧 add refs. Many machine-learning algorithms, such as deep nets, manage to deal with larger dimensionality and datapoints by finding an argument that maximizes the integrands above (MacKay 1992a,b), for example a value of $y_0$ that locally maximizes the integrand in the numerator of eq. (7). This means, though, that they lose the uncertainty quantification.

## 3 Direct case: methodology

For the direct case (1) we consider the RMSD $y$ as a continuous variable, mapped to a log-scale to avoid dealing with finite ranges.

In the direct case the assumption 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 $x_0$.The values $(\bar{y}, \bar{x})$ are exchangeable.

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

$$\mathrm{p}(y_0 \mid x_0,\ \bar{y}, \bar{x},\ H) = \frac{\mathrm{p}(y_0, x_0 \mid \bar{y}, \bar{x},\ H)}{\mathrm{p}(x_0 \mid \bar{y}, \bar{x},\ H)} \tag{8}$$

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.

## 4 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.)

Bernardo, J.-M., Smith, A. F. (1994): *Bayesian Theory*. (Wiley, Chichester).

Bishop, C. M. (2006): *Pattern Recognition and Machine Learning*. (Springer, New York).

Damien, P., Dellaportas, P., Polson, N. G., Stephens, D. A., eds. (2013): *Bayesian Theory and Applications*. (Oxford University Press, Oxford).

Dawid, A. P. (2013): *Exchangeability and its ramifications*. In: Damien, Dellaportas, Polson, Stephens (2013): ch. 2:19–29.

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.

MacKay, D. J. C. (1992a): *Bayesian interpolation*. Neural Comp. **4**$^3$, 415–447. http://www.inference.phy.cam.ac.uk/mackay/PhD.html, DOI:10.1162/neco.1992.4.3.415.

— (1992b): *A practical Bayesian framework for backpropagation networks*. Neural Comp. **4**$^3$, 448–472. http://www.inference.phy.cam.ac.uk/mackay/PhD.html, DOI:10.1162/neco.1992.4.3.448.

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.