Notes on RMSD inference

for Alexander, Fábio, Ruth

Luca ● <pgl@portamana.org>

16 June 2021; updated 6 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 $(\bar{y}, \bar{x}) \coloneqq \big((y_1, x_1), (y_2, x_1), \ldots, (y_N, x_N)\big)$ 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)

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

which we want to quantify.

The first question we must ask in approaching this problem is: how much 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 $\mathrm{p}(x_0 \mid y_0,\ \bar{y}, \bar{x},\ H)$ *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. 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.

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.

The final goal is to compare the results of the present principled approach with those of machine-learning algorithms in a case where both can be used. More about this in 🔧 §***.

2   Exchangeability, approximations, machine learning

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.

A probability distribution $p(z_0, z_1, z_2, \ldots \mid H)$ is called *infinitely exchangeable* if it's invariant under permutations of the values $z_0, z_1, \ldots,$ no matter what and how many they are. In other words, the ordering of the observations doesn't matter. Usually one says, somewhat improperly, that "$z_0, z_1, \ldots$ are exchangeable" for short. In our case each z is actually a pair of values (y, x), where x itself is multidimensional.

Exchangeability has the following important consequence: if we knew the infinite-limit long-run relative frequencies $\boldsymbol{F} := (F_{y,x})$ 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:

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

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

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

If the long-run frequencies $\boldsymbol{F}$ are unknown, they can be marginalized, i.e. integrated out, provided that we have a probability distribution $p(\boldsymbol{F} \mid H)$ for them. For our examples above,

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

and

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

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

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

The distribution $p(\boldsymbol{F} \mid H)$, usually called the "prior", embodies the assumptions that we make for the extrapolation. It therefore cannot be determined by the known datapoints. Given enough known datapoints $(\bar{y}, \bar{x})$, the probability distribution for y_0 given x_0 eventually does become equal to their limit conditional frequency. This sets the ultimate uncertainty with which the prediction can be made, and cannot be surpassed by any algorithm. So this approach eventually does yield an optimal

inference. The quickness with which the limit conditional frequency is reached, however, depends heavily on the choice of prior. The prior should therefore be chosen in a well-reasoned manner. In practice the choice is limited by computational constraints.

All machine-learning algorithms calculate equations such as the one above or approximations to it, for some choice of the prior $p(F \mid H)$ (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 manage to deal with larger dimensionality and datapoints by parametrizing the F-space in a clever way (for example, deep nets parametrize the F as nested compositions of some simple functions) and 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 cannot quantify the uncertainty of the inference.

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, as we do in testing situations for example. The values $(\bar{y}, \bar{x})$ are considered exchangeable.

The probability $p(y_0 \mid x_0, \bar{y}, \bar{x}, H)$ is then given by eq. (7), which can be rewritten this way:

$$p(y_0 \mid x_0, \bar{y}, \bar{x}, H) = \int F_{y_0|x_0} \, p(F \mid \bar{y}, \bar{x}, H) \, dF$$

$$\text{with} \quad p(F \mid \bar{y}, \bar{x}, H) = \frac{F_{y_1,x_1} \cdots F_{y_N,x_N} \, p(F \mid H)}{\int F_{y_1,x_1} \cdots F_{y_N,x_N} \, p(F \mid H) \, dF} ,$$

$$(8)$$

which is the main formula of the direct case.

The integrals are over the set of long-run joint frequency distributions, which is an infinite-dimensional manifold. A currently popular way

to parametrize it and at the same time choose a prior p($F \mid H$) is by means of so-called Dirichlet-process mixtures. In short, a generic $F_{y,x}$ is represented as a countable weighted sum of a simpler distribution K, called the kernel, with different parameters: $F_{y,x} = \sum_i w_i K(y, x \mid \theta_i)$. The prior is therefore defined over the possible infinite tuples (w_i, θ_i). A Dirichlet process is chosen as such prior. In the present case the kernel is the product of a multivariate normal distribution for y and for the continuous features in x, and a Dirichlet distribution for the discrete features in x (Liverani et al. 2015 § 3.3). An example of this approach is analysed and used by Müller et al. (1996). A discussion of why such a choice of prior may not be sensible is discussed by Petrone (2017) and Quintana et al. (2020 § 4).

The first integral in eq. (8) is numerically approximated by a sum of values of $F_{y_0|x_0}$ sampled from the distribution p($F \mid \bar{y}, \bar{x}, H$) via Markov-chain Monte Carlo methods. Details about the sampling algorithm are given in 🔧 §***.

4 Reverse case: methodology

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

5 Evaluation metrics and calibration

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.

Liverani, S., Hastie, D. I., Azizi, L., Papathomas, M., Richardson, S. (2015): *PReMiuM: an R package for profile regression mixture models using Dirichlet processes*. J. Stat. Soft. **64**, 7. DOI:10.18637/jss.v064.i07.

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.

Müller, P., Erkanli, A., West, M. (1996): *Bayesian curve fitting using multivariate normal mixtures*. Biometrika **83**^{1}, 67–79. DOI:10.1093/biomet/83.1.67.

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.

Petrone, S. (2017): *On Bayesian nonparametric regression*. Talk at the workshop *Building Bridges*. https://www.mn.uio.no/math/english/research/projects/focustat/workshops%20and%20conference/workshop-2017/oslo_sonia.pdf.

Quintana, F. A., Mueller, P., Jara, A., MacEachern, S. N. (2020): *The dependent Dirichlet process and related models*. arXiv:2007.06129.