

# Notes on RMSD inference

for Alexander, Fábio, Ruth

Luca  <pgl@portamana.org>

16 June 2021; updated 20 August 2021

✚ Note: the probability-theoretic derivations and explanations below are very concise and likely not fully comprehensible. I'll add some sections with clearer explanations and fuller derivations later on.

## 1 Overview

We consider a quantity, the Root-Mean-Squared Distance, denoted by  $r$ , and a set of quantities called “features”, denoted by  $x$ . A value of RMSD  $r$  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 the value  $r$  of a new datapoint, given:

- features  $x$  for the new datapoint;
- known pairs  $(\bar{r}, \bar{x}) := ((r_1, x_1), (r_2, x_1), \dots, (r_N, x_N))$  for other  $N$  datapoints, called the “training set”;
- additional facts and hypotheses  $H$  (unavoidable since we are making an extrapolation).

Our uncertainty about  $r$  is expressed by the probability

$$p(r \mid x, \bar{r}, \bar{x}, H) \tag{1}$$

which we want to quantify using the probability calculus.

In the next section we will examine several possible underlying hypotheses and calculate the probabilities they lead to. In  §\*\*\* we will examine the consequences of the different hypotheses and their ensuing probabilities on a test dataset.

## 2 Underlying hypotheses

Quantifying the probability (1) requires two main assessments:

- The probability distribution for the long-run relative frequencies of RMSD and feature values that the training set would have if, hypothetically, it were augmented indefinitely.
- What kind of relevance the training set has for our inference about the new datapoint.

We discuss these assessments in the next two sections. Section 3 combines them into final probability formulae.

### 2.1 Probability for long-run frequencies

Let us imagine to extend the size of the training set indefinitely, in such a way that our probability assignments for any subset of it would be exchangeable (this condition is more general than an “i.i.d.” one). We could then measure the joint long-run relative frequencies  $F_{r,x}$  for all values  $r$  and  $x$  in such extended training set<sup>1</sup>, as well as the marginal relative frequencies  $F_{r\bullet}$  for  $r$ , and  $F_{\bullet x}$  for  $x$ ; and the conditional relative frequencies  $F_{r|x}$  for  $r$  given  $x$ , and  $F_{x|r}$  for  $x$  given  $r$ . These frequencies are related by

$$F_{r\bullet} = \sum_x F_{r,x}, \quad F_{\bullet x} = \sum_r F_{r,x}, \quad F_{r|x} = \frac{F_{r,x}}{F_{\bullet x}}, \quad F_{x|r} = \frac{F_{r,x}}{F_{r\bullet}}. \quad (2)$$

The collection of joint frequencies  $(F_{r,x})_{r,x}$  is denoted by  $F$ . For later use we denote by  $F_{|r}$  the collection of conditional frequencies  $(F_{x|r})_x$ , for each value of  $r$ . Note that from  $F$  we can obtain  $F_{|r}$  and  $F_{r\bullet}$  for each  $r$ , and vice versa.

We need to specify a probability density

$$p(F | H) dF \quad (3)$$

representing our belief and prior information as to what the long-run frequencies could be, before our observation of the training set. After the observation of the training set this probability is updated to

$$p(F | \tilde{r}, \tilde{x}, H) dF = \frac{p(F | H) \prod_{i=1}^N F_{r_i x_i}}{\int p(F | H) \prod_{i=1}^N F_{r_i x_i} dF} dF \quad (4)$$

<sup>1</sup> “But this *long run* is a misleading guide to current affairs. *In the long run* we are all dead.” (Keynes 2013 § 3.I, p. 65)

The product  $\prod_{i=1}^N F_{r_i x_i} = \prod_{r x} F_{r x}^N f_{r x}$ , where  $f_{r x}$  is the joint frequency of values  $r$  and  $x$  observed in the training set.

The choice of density (3) is constrained by analytic convenience and computational costs – such a density is defined over an infinite-dimensional manifold. Our specific choices and their motivations are discussed more in detail in [✂️appendix\\*\\*\\*](#). Here we briefly discuss two main possibilities:

I. An unfactorizable density:

$$p(F | H_I) dF . \quad (5)$$

II. A density factorizable in the conditional frequencies given  $r$ :

$$p(F | H_{II}) dF = \prod_r p(F_{|r} | H_{II}) p(F_{r\bullet} | H_{II}) dF_{|r} dF_{r\bullet} . \quad (6)$$

It represents the assumption that the conditional frequencies  $F_{|r}$  are completely uninformative about the marginal frequencies  $(F_{r\bullet})_r$ , and vice versa.


The first possibility will be used with assumption [i.](#) of [§ 2.2](#), leading to simplified and computationally cheaper formulae. The factors  $p(F_{r\bullet} | H) dF_{r\bullet}$  in particular will become irrelevant. The updated density also factorizes in this case:

$$p(F | \bar{r}, \bar{x}, H_{II}) dF \propto \prod_r p(F_{r\bullet} | H_{II}) F_{r\bullet}^N f_{r\bullet} dF_{r\bullet} \times \prod_r p(F_{|r} | H_{II}) \left( \prod_{\substack{i=1 \\ r_i=r}}^N F_{x_i|r} \right) dF_{|r} , \quad (7)$$

where  $f_{r\bullet}$  is the marginal frequency of the  $r$  value in the training set. The marginal densities for the long-run conditional frequencies  $F_{|r}$ , for each  $r$ , are also found more easily:


$$p(F_{|r} | \bar{r}, \bar{x}, H_{II}) dF_{|r} \propto p(F_{|r} | H_{II}) \left( \prod_{\substack{i=1 \\ r_i=r}}^N F_{x_i|r} \right) dF_{|r} . \quad (8)$$

The prior and posterior probability densities for the long-run frequencies cannot be given in closed form. In practice they are represented by

means of a finite number  $T$  of typical samples,  $\{F^{(1)}, F^{(2)}, \dots, F^{(T)}\}$ , obtained by Markov-chain Monte Carlo methods. Integrations with respect to these densities is approximated by sums over these representative samples. These approximations are discussed in S\*\*\*.

## 2.2 Relevance of training set: exchangeability and subpopulations

The relevance of the training set to the new datapoint is usually expressed mathematically as a form of symmetry or *exchangeability* of the joint probability of the training set and the new datapoint. Such a symmetry intuitively tells us if and how the new datapoint could be considered as a member of the training set if the latter were hypothetically augmented indefinitely. Then the long-run frequencies considered in the previous section directly give us probabilities for the new datapoint.

The question of exchangeability is discussed in a brilliant paper by Lindley & Novick (1981), who show with examples its fundamental importance for making correct inferences. A short intuitive summary of this notion is given in appendix\*\*\*.

We consider three mutually exclusive cases of exchangeability, the first seeming the most appropriate to our inference:

- i. *Exchangeability in features  $x$  given RMSD  $r$* . The values of the features of the training set are relevant to the inference of those of the new datapoint, given the same RMSD value. The RMSDs of the training set, however, are not relevant for the RMSD of the new datapoint. This means that the probability for feature value  $x$  in the new datapoint, given RMSD  $r$ , would be equal to the long-run conditional frequency  $F_{x|r}$  if the latter were known:

$$p(x \mid r, F, H_i) = F_{x|r} . \quad (9)$$

- ii. *Exchangeability in RMSD  $r$  given features  $x$* . The RMSDs of the training set are relevant to the inference of the RMSD of the new datapoint, given the same feature values. The feature values of the training set, however, are not relevant for the inference of the feature values of the new datapoint. This means that the probability for a RMSD  $r$  in the new datapoint, given features  $x$ , would be equal to the long-run conditional frequency  $F_{r|x}$  if the latter were known:

$$p(r \mid x, F, H_{ii}) = F_{r|x} . \quad (10)$$

- iii. *Exchangeability in RMSD  $r$  and features  $x$ .* Both the RMSDs and the values of the features of the training set are relevant to the inference of those of the new datapoint. This is equivalent to the relevance of  $r$  given  $x$ , and of  $x$ ; and to the relevance of  $x$  given  $r$ , and of  $r$ . It also means that the probability for a RMSD  $r$  in the new datapoint, given features  $x$ , would be equal to the long-run conditional frequency  $F_{r|x}$  if the latter were known:

$$p(r \mid x, F, H_{\text{iii}}) = F_{r|x} . \quad (11)$$

Owing to the exchangeability in  $x$ , we also have that our probability for  $F$  is updated by our knowledge of  $x$  of the new datapoint:

$$p(F \mid x, \bar{r}, \bar{x}, H) = \frac{F_{\cdot x} p(F \mid \bar{r}, \bar{x}, H)}{\int F_{\cdot x} p(F \mid \bar{r}, \bar{x}, H) dF} . \quad (12)$$

In the first and second case, the features or RMSD of the training set could be irrelevant because the respective quantity of the new datapoint could be selected by hand or by some process different from that underlying the training set. Several authors connect or motivate these different kinds of relevance with the robustness of “causal relations” (which can be direct or indirect, stemming for example from a common cause) in the directions  $x \rightsquigarrow r$  or  $r \rightsquigarrow x$  (cf. Pearl 1988 §§ 2.1.2, 2.2.5).

In case ii. knowledge of the long-run frequencies  $F_{x|r}$  cannot give us directly the probability of  $r$  given  $x$ . Such probability can be obtained with Bayes’s theorem:

$$p(r \mid x, F, H_i) = \frac{p(x \mid r, F, H_i) p(r \mid H_i)}{\sum_r p(x \mid r, F, H_i) p(r \mid H_i)} = \frac{F_{x|r} p(r \mid H_i)}{\sum_r F_{x|r} p(r \mid H_i)} . \quad (13)$$

The prior probability  $p(r \mid H_i)$  required by the theorem cannot be obtained from the training set in this case. It needs to be assessed from other kinds of information or assumptions.

### 3 Required probability and approximations

#### 3.1 Combination of assumptions and final formulae

The assumptions  $H_i, H_i, H_{\text{iii}}$  of § 2.2 give us the probability  $p(r \mid x \dots)$  of  $r$  given  $x$  if the long-run frequencies were known. The assumptions

$H_I$ ,  $H_{II}$  of e§ 2.1 give us probabilities for these long-run frequencies. By combining the two kinds of assumptions we finally arrive at the probability of interest  $p(r \mid x, \bar{r}, \bar{x}, H)$ .

We consider four combinations of assumptions:

1.  $H_I$  and  $H_{II}$ : exchangeability in  $r$  given  $x$ , mutually informative  $F_{r\cdot}$  and  $F_{I|r}$ . We find

$$p(r \mid x, \bar{r}, \bar{x}, H_{II I}) = \int F_{r|x} p(F \mid \bar{r}, \bar{x}, H_I) dF \quad (14)$$

to be combined with eq. (4).

2.  $H_I$  and  $H_{III}$ : exchangeability in  $r$  and  $x$ , mutually informative  $F_{r\cdot}$  and  $F_{I|r}$ . We find

$$p(r \mid x, \bar{r}, \bar{x}, H_{III I}) = \frac{\int F_{rx} p(F \mid \bar{r}, \bar{x}, H_I) dF}{\int F_{\cdot x} p(F \mid \bar{r}, \bar{x}, H_I) dF} \quad (15)$$

to be combined with eq. (4).

3.  $H_I$  and  $H_I$ : exchangeability in  $x$  given  $r$ , mutually informative  $F_{r\cdot}$  and  $F_{I|r}$ . Considering eq. (13) we find

$$p(r \mid x, \bar{r}, \bar{x}, H_{II I}) = \frac{p(r \mid H_I) \int F_{x|r} p(F \mid \bar{r}, \bar{x}, H_I) dF}{\sum_r p(r \mid H_I) \int F_{x|r} p(F \mid \bar{r}, \bar{x}, H_I) dF} \quad (16)$$

to be combined with eq. (4).

4.  $H_{II}$  and  $H_I$ : exchangeability in  $x$  given  $r$ , no mutual information between  $F_{r\cdot}$  and  $F_{I|r}$ . Considering eqs (13) and (8) we find

$$p(r \mid x, \bar{r}, \bar{x}, H_{II II}) = \frac{p(r \mid H_I) \int F_{x|r} p(F_{I|r} \mid \bar{r}, \bar{x}, H_{II}) dF_{I|r}}{\sum_r p(r \mid H_I) \int F_{x|r} p(F_{I|r} \mid \bar{r}, \bar{x}, H_{II}) dF_{I|r}} \quad (17)$$

to be combined with eq. (8).

Combination 1. is used for example in Müller et al. (1996) and discussed in Quintana et al. (2020 § 4).

Combination 2. differs from 1. in that its probability for the long-run frequencies  $F$  is updated with the knowledge of  $x$ . When the training set is large there should be little difference between the two cases.

Computationally combination 4. is the least costly, because it requires densities defined in one less dimension, and allows for parallel use of Markov-chain Monte Carlo samplers, one for each value  $r$ .

Finally, combinations 3. and 4. can be used only if  $r$  is a discrete variable.

### 3.2 Monte Carlo approximations

As mentioned in § 2.1, the probability densities  $p(F | \bar{r}, \bar{x}, H_I) dF$  and  $p(F_{|r} | \bar{r}, \bar{x}, H_{II}) dF_{|r}$  are effectively represented by a finite number  $T$  of samples  $\{F^{(t)}\}$  and  $\{F_{|r}^{(t)}\}$  drawn from them via Markov-chain Monte Carlo methods, and integration with respect to them is approximated by summation over such samples. The formulae obtained in the previous section then take the following approximate forms:

1'.  $H_I$  and  $H_{II}$ :

$$p(r | x, \bar{r}, \bar{x}, H_{II}) \approx \sum_t F_{r|x}^{(t)}, \quad F^{(t)} \text{ drawn from } p(F | \bar{r}, \bar{x}, H_I) dF. \quad (18)$$

2'.  $H_I$  and  $H_{III}$ :

$$p(r | x, \bar{r}, \bar{x}, H_{III}) \approx \frac{\sum_t F_{rx}^{(t)}}{\sum_t F_{\bullet x}^{(t)}}, \quad F^{(t)} \text{ drawn from } p(F | \bar{r}, \bar{x}, H_I) dF. \quad (19)$$

3'.  $H_I$  and  $H_i$ :

$$p(r | x, \bar{r}, \bar{x}, H_{II}) \approx \frac{p(r | H_i) \sum_t F_{x|r}^{(t)}}{\sum_r p(r | H_i) \sum_t F_{x|r}^{(t)}}, \quad F^{(t)} \text{ drawn from } p(F | \bar{r}, \bar{x}, H_I) dF. \quad (20)$$

4'.  $H_{II}$  and  $H_i$ :

$$p(r | x, \bar{r}, \bar{x}, H_{II}) \approx \frac{p(r | H_i) \sum_t F_{x|r}^{(t)}}{\sum_r p(r | H_i) \sum_t F_{x|r}^{(t)}}, \quad F_{|r}^{(t)} \text{ drawn from } p(F_{|r} | \bar{r}, \bar{x}, H_{II}) dF_{|r} \text{ for each } r. \quad (21)$$

 Notes on the Markov-chain Monte Carlo adopted

## 4 Selection of features and training data

## 5 Evaluation

### Section still in progress

A short informal discussion on the “validation” and “testing” of probability models.

What does it mean to “test” a predictive probability model? It does not make sense to test against the truth of falsity of what’s being predicted for two reasons.

First, probability models ordinarily do not give truths; they gives probabilities. And we can’t compare a probability and a truth. It is also wrong to require that a model give highest probability to the true case, as a simple example shows.

Suppose you’re going to roll a regular die, and wonder about two exclusive outcomes: outcome  $A$  is 1 to 2, outcome  $B$  is 3 to 6. You adopt the probability model  $M$ , motivated by symmetry arguments, that assigns  $P(A | M) = 2/6$  and  $P(B | M) = 4/6$ . Imagine also that a friend of yours – who, like you, knows that the die and the roll procedure are regular – uses a different probability model  $M'$ , with  $P(A | M') = 0.999$  and  $P(B | M') = 0.001$ . I believe that you would ask your friend about the reason of this peculiar probability model: the prior information that you both have does not seem to support such an extremely high probability for  $A$ . Now you roll the die, and obtain 2. The “truth” is outcome  $A$ . Your model  $M$  assigned the lower probability to  $A$ . Your friend’s model  $M'$  assigned an extremely high probability to  $A$  instead. Would you then say that your friend’s model was “correct”, and yours “wrong”?

I hope you agree, by simple common sense, that your model  $M$  was the most *reasonable*, because all evidence available to you and your friend made  $A$  the least probable outcome of the two. This example shows that we cannot judge a model against the truth of the final outcome. We can only judge it against the prior evidence we had when it was formulated.

Secondly, we consider a probability model because we don’t know the truth. So we don’t have any truth to test it against. If we had that truth, we would not be using a probability model. You might say “I can at least test the model in cases where I know the truth” – but how do you know that the truth of your test cases is the same as the real case? You don’t. If you say “I suspect them to be similar to the real case”, it



means that you have some kind of relevant prior information about the latter. Then you should use this prior information in formulating your model (if you don't, you're misusing probability theory).

One danger of judging a model against test cases is that we can end up modifying it in order to fit them better – ending up in giving unreliable predictions for the real case, which may turn out to be very different from the test ones.

Such “tests”, however, can also be viewed from a different and more reasonable perspective. It is often difficult to translate our prior information into a mathematical formula. We can apply a candidate mathematical formula thus obtained to cases in which we know what the reasonable predictions would be, given our prior information. If the mathematical formula leads to different predictions, then it means that our mathematical translation was incorrect, and we must find a better one. Good (1950 § 4.3 p. 35) called this procedure the “device of imaginary results”.

It is in this last sense that we shall now “validate” the hypotheses discussed in § 2.

#### 🔧 Validation plan:

- Set sampled from the original dataset. This represents a real case in which the frequencies of  $r$ ,  $x$ , and all the conditional ones are expected to be as in our dataset.
- Set sampled from the original one but with proportions of  $r$  values different from those appearing in the original one. This represent a real case in which the conditional frequencies  $F_{x|r}$  are expected to be as in our dataset, but the frequencies of the different  $r$  values and the conditional frequencies  $F_{r|x}$  are expected to be different.

If we assume  $M$  different values of  $r$  and  $N$  different values of  $x$ , the set of frequencies  $\{F_{r\bullet}, F_{\bullet x}, F_{r|x}, F_{x|r}\}$  has  $M+N+MN+NM$  components constrained by the  $MN$  identities  $\{F_{x|r} F_{r\bullet} = F_{r|x} F_{\bullet x}\}$  and  $N+M-1$  independent normalization conditions. This gives  $MN-1$

## Appendices

### 🔧 Sections still in progress

## A Exchangeability, approximations, machine learning

Exchangeability is the basic fact or assumption upon which machine-learning algorithms and the calculations presented here are based.

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).

A probability distribution  $p(z_0, z_1, z_2, \dots | H)$  is called *infinitely exchangeable* if it's invariant under permutations of the values  $z_0, z_1, \dots$ , 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 " $z_0, z_1, \dots$  are exchangeable" for short. In our case each  $z$  is actually a pair of values  $(r, x)$ , where  $x$  itself is multidimensional.

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

$$p(r_1, x_1, r_2, x_2 | F, H) = F_{r_1, x_1} F_{r_2, x_2} . \quad (22)$$

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

$$p(r_1, r_2 | x_1, x_2, F, H) = F_{r_1|x_1} F_{r_2|x_2} . \quad (23)$$

If the long-run frequencies  $F$  are unknown, they can be marginalized, i.e. integrated out, provided that we have a probability distribution  $p(F | H)$  for them. For example:

$$p(r_1, x_1, r_2, x_2 | H) = \int F_{r_1, x_1} F_{r_2, x_2} p(F | H) dF , \quad (24)$$

from which we can also obtain conditional probabilities.

There are several other types of exchangeability. For example, we may have *conditional* or *partial* exchangeability of  $r$  given  $x$  if eq. (??) holds, but not eq. (??). If we have a probability distribution  $p(F_{r|x} | H)$  for the conditional frequencies, we obtain by marginalization

$$p(r_1, r_2 | x_1, x_2, H) = \int F_{r_1|x_1} F_{r_2|x_2} p(F_{r|x} | H) dF_{r|x} . \quad (25)$$


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 (??), by simply using the three rules of the probability calculus. For example, if  $r$  is assumed to be exchangeable given  $x$ , but the latter is not assumed to be exchangeable with the other  $\bar{x}$  because hand-picked, we find

$$p(r \mid x, \bar{r}, \bar{x}, H) = \frac{\int F_{r|x} F_{r_1, x_1} \cdots F_{r_N, x_N} p(F \mid H) dF}{\int F_{r_1, x_1} \cdots F_{r_N, x_N} p(F \mid H) dF} . \quad (26)$$

The distribution  $p(F \mid H)$ , 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  $(\bar{r}, \bar{x})$ , the probability distribution for  $r$  given  $x$  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  $r$  that locally maximizes the integrand in the numerator of eq. (??). This means, though, that they cannot quantify the uncertainty of the inference.

## B Exchangeability assumptions in formulae

- i. *Exchangeability of RMSD given features.* The RMSDs of the training set are relevant to the inference of the RMSD of the new datapoint, given the same feature values. The features of the training set, however, Exchangeability of  $r$  given  $x$ : The joint probability of the RMSDs of new datapoint and training set is symmetric under exchanges of datapoints, given the same features. That is,

$$p(r = a_0, r_1 = a_1, r_2 = a_2, \dots \mid x_1 = b, x_2 = b, x_3 = b, \dots, H) =$$

$$p(r = a_1, r_1 = a_0, r_2 = a_2, \dots \mid x_1 = b, x_2 = b, x_3 = b, \dots, H) =$$

$$p(r = a_2, r_1 = a_1, r_2 = a_0, \dots \mid x_1 = b, x_2 = b, x_3 = b, \dots, H) = \dots$$

and so on, for all simultaneous permutations of  $a_i$  and any  $b$ .  
(27)

- ii. *Exchangeability of  $x$  given  $r$ :* The joint probability of the features of new datapoint and training set is symmetric under exchanges of datapoints, given the same RMSD. That is,

$$p(x = b_0, x_1 = b_1, x_2 = b_2, \dots \mid r_1 = a, r_2 = a, r_3 = a, \dots, H) =$$

$$p(x = b_1, x_1 = b_0, x_2 = b_2, \dots \mid r_1 = a, r_2 = a, r_3 = a, \dots, H) =$$

$$p(x = b_2, x_1 = b_1, x_2 = b_0, \dots \mid r_1 = a, r_2 = a, r_3 = a, \dots, H) = \dots$$

and so on, for all simultaneous permutations of  $b_i$  and any  $a$ .  
(28)

- iii. The RMSDs and features of the training set are jointly relevant to the inference of those of the new datapoint. We then have full exchangeability of  $(r, x)$ : the joint probability of new datapoint and training set is symmetric under exchanges of datapoints. That is,

$$p(r = a_0, x = b_0, r_1 = a_1, x_1 = b_1, r_2 = a_2, x_2 = b_2, \dots \mid H) =$$

$$p(r = a_1, x = b_1, r_1 = a_0, x_1 = b_0, r_2 = a_2, x_2 = b_2, \dots \mid H) =$$

$$p(r = a_2, x = b_2, r_1 = a_1, x_1 = b_1, r_2 = a_0, x_2 = b_0, \dots \mid H) = \dots$$

and so on, for all simultaneous permutations of  $a_i$  and  $b_i$ . (29)

### C Selection of prior

The prior densities  $p(F | \bar{r}, \bar{x}, H_I) dF$  and  $p(F_{|r} | \bar{r}, \bar{x}, H_{II}) dF_{|r}$  for each  $r$  are represented by a Dirichlet-process mixture with product kernel of a multivariate normal for the continuous quantities and Dirichlet distributions for the discrete ones. As the average measure resulting from the Dirichlet process we choose the product of a normal-inverse-Wishart distribution and Dirichlet distributions; this choice leads to faster computation.

Let us first focus on the continuous features “tanimoto” and “sasa”.

The normal-inverse-Wishart has four parameters  $\mu, \Sigma, \kappa, \delta$ , and leads to a multivariate t-distribution for the quantities. The mean of the t-distribution is equal to  $\mu$ ; the expected value of the covariance matrix is  $\frac{1}{\delta-2} \Sigma$ ; the covariance of the t-distribution is  $\frac{\kappa+1}{\kappa(\delta-2)} \Sigma$ ; the covariance of the mean is  $\frac{1}{\kappa(\delta-2)} \Sigma$ ; the degrees of freedom of the t-distribution equal  $\delta$ .

We choose these parameters, the concentration parameter of the Dirichlet process, and the coordinate systems on the space of the continuous quantities in order to build a prior distribution that represents as well as possible our knowledge before seeing the data. Here are some requirements:

- (a) The “sasa” quantities  $x$  have a range in  $[0, +\infty[$ . We consider them as scale quantities (even if the value 0 is included) and transform them to a log-scale:

$$x \mapsto r = \frac{1}{4} \ln x . \quad (30)$$

Their prior distribution is chosen as almost uniform on this scale:  $p[\ln(x) | H] d \ln(x) \propto d \ln(x)$ , approximated by a t-distribution with large standard deviation.

- (b) We set the mean of the t-distribution for the “sasa” quantities equal to the mean observed in the full dataset, in order to represent theoretical prior knowledge on this kind of quantities.
- (c) The “tanimoto” quantities  $x$  have a range in  $[0, 1]$ . In this coordinate system we assume that they have a Jeffreys prior  $p(x | H) dx \propto \frac{dx}{x(1-x)}$ . We transform them to  $\mathbf{R}$  with a logit function

$$x \mapsto r = \frac{1}{2} \text{logit}(x) := \frac{1}{2} \ln \frac{x}{1-x} . \quad (31)$$

In this coordinate system the prior is uniform, approximated by a t-distribution with large standard deviation.

- (d) We want the priors for the “sasa” and “tanimoto” quantities to be approximately independent; this is obtained by choosing a large degree-of-freedom parameter for the multivariate t-distribution.
- (e) We choose the expected variances for the “sasa” and “tanimoto” quantities mapped to  $\mathbf{R}$  to be approximately equal to the variances observed in the full dataset, in order to represent theoretical prior knowledge on this kind of quantities.
- (f) We choose the standard deviations for the means of the “sasa” and “tanimoto” quantities to be approximately equal to their ranges observed in the full dataset, in order to represent theoretical prior knowledge on this kind of quantities.

$$\begin{aligned}
 \frac{\kappa + 1}{\kappa (\delta - 2)} \Sigma &= 5^2 \quad (\text{t-distr. variance}) \\
 \frac{1}{\kappa (\delta - 2)} \Sigma &= 1^2 \quad (\text{variance of mean}) \\
 \frac{1}{\delta - 2} \Sigma &= \frac{1}{2^2} \quad (\text{exp. variance}) \\
 \mu_{\text{sasa}} &= 1 \\
 \delta &= 30
 \end{aligned} \tag{32}$$

\*\*\*\*\*

based on a Dirichlet-process mixture with product kernels of multivariate normal and Dirichlet distributions.

The “sasa” quantities  $x$  have a range in  $[0, +\infty[$ . We consider them as scale quantities (even if the value 0 is included) and transform them to a log-scale. Their prior distribution is chosen as almost uniform on this scale:  $p(\ln(x) | H) d\ln(x) \propto d\ln(x)$ , approximated by a normal with large standard deviation.

The “tanimoto” quantities  $x$  have a range in  $[0, 1]$ . We transform them to  $\mathbf{R}$  using the cumulative normal distribution:

$$x \mapsto \phi(x) := \frac{1}{2} \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt, \tag{33}$$

which we call an ‘erf-scale’. The prior is chosen as uniform in  $dx$ , and therefore has a standard normal density in the erf-scale.

Choosing a normal-inverse-Wishart distribution with parameters  $\mu^*, \Delta^*, \kappa^*, \nu^*$  as the mean for the Dirichlet process leads to a t-distribution as the predictive distribution for a quantity  $x \in \mathbf{R}^d$ , with  $\nu^* - n + 1$  degrees of freedom, mean  $\mu^*$ , covariance  $\frac{\kappa^* + 1}{\kappa^* (\nu^* - d - 1)} \Delta^*$ .

This means that we want  $\nu^* - n + 1$  large, say  $\sim 10$ , in order to approximate a normal,  $\mu^*$  equal to the zero vector, and  $\frac{\kappa^* + 1}{\kappa^* (\nu^* - d - 1)} \Delta^*$  diagonal and equal to 1 in the direction of the “tanimoto” quantities, and equal to some larger value, say  $\sim 4$ , in the direction of the “sasa” quantities. These requirements leave the parameter  $\kappa^*$  still undefined.

## 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>.

Good, I. J. (1950): *Probability and the Weighing of Evidence*. (Griffin, London).

Keynes, J. M. (2013): *A Tract on Monetary Reform*, repr. of 2nd ed. (Cambridge University Press, Cambridge). First publ. 1923.

Lindley, D. V., Novick, M. R. (1981): *The role of exchangeability in inference*. *Ann. Stat.* **9**<sup>1</sup>, 45–58. [DOI:10.1214/aos/1176345331](https://doi.org/10.1214/aos/1176345331).

Liverani, S., Hastie, D. I., Azizi, L., Papatthomas, 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](https://doi.org/10.18637/jss.v064.i07).

MacKay, D. J. C. (1992a): *Bayesian interpolation*. *Neural Comp.* **4**<sup>3</sup>, 415–447. <http://www.inference.phy.cam.ac.uk/mackay/PhD.html>, [DOI:10.1162/neco.1992.4.3.415](https://doi.org/10.1162/neco.1992.4.3.415).

— (1992b): *A practical Bayesian framework for backpropagation networks*. *Neural Comp.* **4**<sup>3</sup>, 448–472. <http://www.inference.phy.cam.ac.uk/mackay/PhD.html>, [DOI:10.1162/neco.1992.4.3.448](https://doi.org/10.1162/neco.1992.4.3.448).

Müller, P., Erkanli, A., West, M. (1996): *Bayesian curve fitting using multivariate normal mixtures*. *Biometrika* **83**<sup>1</sup>, 67–79. [DOI:10.1093/biomet/83.1.67](https://doi.org/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](https://doi.org/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](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*.

## 🔑 OLD TEXT BELOW

\*\*\*\*\*

We introduce the long-run, joint relative frequencies  $F_{rx}$  that would be observed if the training dataset could hypothetically be extended to an infinite size. With these we also have marginal and conditional frequencies  $F_{r\bullet}, F_{\bullet x}, F_{r|x}, F_{x|r}$ . These frequencies are obviously unknown but they play a pivotal role in the inferences we are interested in. The collection of joint frequencies  $(F_{rx})_{rx}$  will be denoted by  $F$ .

In the following  $r$  will be treated as a discrete quantity with three possible values: ‘good’ ( $0 \text{ \AA} \leq r \leq 2 \text{ \AA}$ ), ‘inconclusive’ ( $2 \text{ \AA} < r < 3 \text{ \AA}$ ), ‘bad’ ( $3 \text{ \AA} \leq r$ ).

We denote with  $f_{rx}$  the joint relative frequency for values  $r$  and  $x$  in the training dataset; with  $f_{r\bullet}$  the marginal relative frequency for  $r$ , and with  $f_{\bullet x}$  that for  $x$ ; with  $f_{r|x}$  the conditional relative frequency for  $r$  given  $x$ , and vice versa for  $f_{x|r}$ . These frequencies are related by

$$f_{r\bullet} = \sum_x f_{rx}, \quad f_{\bullet x} = \sum_r f_{rx}, \quad f_{r|x} = f_{rx}/f_{\bullet x}, \quad f_{x|r} = f_{rx}/f_{r\bullet}. \quad (34)$$

We introduce the long-run, joint relative frequencies  $F_{rx}$  that would be observed if the training dataset could hypothetically be extended to an infinite size<sup>2</sup>. With these we also have marginal and conditional frequencies  $F_{r\bullet}, F_{\bullet x}, F_{r|x}, F_{x|r}$ . These frequencies are obviously unknown but they play a pivotal role in the inferences we are interested in. The collection of joint frequencies  $(F_{rx})_{rx}$  will be denoted by  $F$ .

The three cases of exchangeability listed above lead to three different equations for the calculation of the probability (1). Suppose that our uncertainty about the long-run frequencies is quantified by the probability density  $p(F | \bar{r}, \bar{x}, H) dF$ . Then:

I. In case ii. we have

$$p(r | x, \bar{r}, \bar{x}, H) = \int F_{r|x} p(F | \bar{r}, \bar{x}, H) dF. \quad (35)$$

II. In case i. we have

$$p(r | x, \bar{r}, \bar{x}, H) = \frac{p(r | H) \int F_{x|r} p(F | \bar{r}, \bar{x}, H) dF}{\sum_{r'} p(r' | H) \int F_{x|r'} p(F | \bar{r}, \bar{x}, H) dF}. \quad (36)$$

<sup>2</sup> “But this *long run* is a misleading guide to current affairs. *In the long run* we are all dead.” (Keynes 2013 § 3.I, p. 35)



III. In case [iii](#), we have

$$p(r \mid x, \bar{r}, \bar{x}, H) = \frac{\int F_{rx} p(F \mid \bar{r}, \bar{x}, H) dF}{\int F_{\cdot x} p(F \mid \bar{r}, \bar{x}, H) dF} . \quad (37)$$


For example, if  $r$  is hand-picked by someone, then knowledge of the frequencies for  $(\bar{r}, \bar{x})$  is irrelevant. But even in such a case it might be that the reversed probability  $p(x \mid r, \bar{r}, \bar{x}, H)$  *does* depend on the frequencies of previous datapoints, and it can be used to calculate [\(1\)](#) as

$$p(r \mid x, \bar{r}, \bar{x}, H) = \frac{p(x \mid r, \bar{r}, \bar{x}, H) p(r \mid H)}{\sum_r p(x \mid r, \bar{r}, \bar{x}, H) p(r \mid H)} . \quad (38)$$

The important difference between eqs [\(1\)](#) and [\(??\)](#) is that in the latter we must also quantify  $p(r \mid H)$  using extra-data facts or hypotheses.

In technical terms we are asking whether  $r$  is *exchangeable* given  $x$ , or vice versa. Using Fisherian ([1956](#) §§ II.4, IV.1) parlance we are asking whether  $r$  should be considered as belonging to a *subpopulation* determined by  $x$  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 “causal connections” (which can originate from a common cause) are more robust in the direction  $x \rightsquigarrow r$  or  $r \rightsquigarrow x$  (cf. Pearl [1988](#) §§ 2.1.2, 2.2.5).

We study both possibilities mentioned above. At the end we'll compare the results and discuss which assumption makes more sense in various applications.

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  §\*\*\*.

## D Direct case: methodology

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

In the direct case the assumption is that  $r$  is exchangeable given  $x$ , but  $x$  is not exchangeable given the known datapoints  $\bar{x}$ . This is a sensible assumption – a fact indeed – if we are *choosing* the specific value of  $x$ , as we do in some testing situations for example. The values  $(\bar{r}, \bar{x})$  are considered exchangeable.


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

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

$$\text{with } p(F \mid \bar{r}, \bar{x}, H) = \frac{F_{r_1, x_1} \cdots F_{r_N, x_N} p(F \mid H)}{\int F_{r_1, x_1} \cdots F_{r_N, x_N} p(F \mid H) dF}, \quad (39)$$

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  $F_{r,x}$  is represented as a countable weighted sum of a simpler distribution  $K$ , called the kernel, with different parameters:  $F_{r,x} = \sum_i w_i K(r, x \mid \theta_i)$ . The prior is therefore defined over the possible infinite tuples  $(w_i, \theta_i)$ . A Dirichlet process is chosen as such prior. In the present case the kernel is the product of a multivariate normal distribution for  $r$  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. (??) is numerically approximated by a sum of values of  $F_{r|x}$  sampled from the distribution  $p(F \mid \bar{r}, \bar{x}, H)$  via Markov-chain Monte Carlo methods. Details about the sampling algorithm are given in  §\*\*\*.

## E Reverse case: methodology

For the reverse case (??) we consider a binned RMSD divided into three categories:  $r \in \{\text{'good'}, \text{'inconclusive'}, \text{'bad'}\}$ .


In the reverse case the assumption is that  $x$  is exchangeable given  $r$ , but  $r$  is not exchangeable given the known datapoints  $\bar{r}$ . This is again a sensible assumption if we are choosing the specific value of  $x$ , as we do in testing situations, because then  $r$  cannot be considered to come from some unsystematic process, if it has some causal connections with  $x$ . The values  $(\bar{r}, \bar{x})$  are considered exchangeable.

The probability  $p(x \mid r, \bar{r}, \bar{x}, H)$  is obtained analogously to eq. (??), reversing the roles of  $r$  and  $x$ . We make the additional assumption that the prior  $p(F_{x|r} \mid H)$  factorizes for the frequencies conditional on the three  $r$  categories:

$$p(F_{x|r} \mid H) = p(F_{x|\text{good}} \mid H) \cdot p(F_{x|\text{inconcl.}} \mid H) \cdot p(F_{x|\text{bad}} \mid H). \quad (40)$$

The result is three distinct conditional probabilities:

$$\begin{aligned} p(x \mid r, \bar{r}, \bar{x}, H) &= \int F_{x|r} p(F_{x|r} \mid \bar{x}, H) dF_{x|r} \\ p(F_{x|r} \mid \bar{x}, H) &= \frac{F_{x_{1'}|r} \cdots F_{x_{N'}|r} p(F_{x|r} \mid H)}{\int F_{x_{1'}|r} \cdots F_{x_{N'}|r} p(F_{x|r} \mid H) dF_{x|r}} \quad (41) \\ &\text{for each } r \in \{\text{'good'}, \text{'inconclusive'}, \text{'bad'}\}, \end{aligned}$$

where the conditionals contain only  $\bar{x}$  values associated with that specific  $r$  value ( I'll find clearer notation and explanations).

The fact that we have three distinct conditional probability distributions allows us to split the computation into three. They can be done in parallel and each has one less dimension and is being conditioned on fewer data, so the computation is somewhat faster.

The computation of each probability is analogous to that explained in § ??, based on a Dirichlet-process mixture with product kernels of multivariate normal and Dirichlet distributions.