Which method is more accurate? or Errors have error bars

Jan H. Jensen[1,*]

[1]Department of Chemistry, University of Copenhagen, Copenhagen, Denmark
[*]E-mail: jhjensen@chem.ku.dk; Twitter: @janhjensen

January 3, 2017

Abstract

This document is my attempt at distilling some of the information in two papers published by Anthony Nicholls (*J. Comput. Aided Mol. Des.* 2014, 28, 887; ibid 2016, 30, 103). Anthony also very kindly provided some new equations, not found in the papers, in response to my questions. The paper describes how one determines whether the difference in accuracy of two methods in predicting some properties for the same data set is statistically significant using root-mean-square errors, mean absolute errors, mean errors, and Pearsons r values.

Introduction

This document is my attempt at distilling some of the information in two papers published by Anthony Nicholls [1, 2]. Anthony also very kindly provided some new equations, not found in the papers, in response to my questions.

Errors also have error bars

Say you have two methods, A and B, for predicting some property and you want to determine which method is more accurate by computing the property using both methods for the same set of N different molecules for which reference values are available. You evaluate the error (for example the RMSE) of each method relative to the reference values and compare. The point of this post is that these errors have uncertainties (error bars) that depend on the number of data points (N, more data less uncertainty) and you have to take these uncertainties into consideration when you compare errors.

The most common error bars reflect 95% confidence and that's what I'll use here.

The expression for the error bars assume a large N where in practice "large" in this context means roughly 10 or more data points. If you use fewer points or would like more accurate estimates please see the Nicholls papers for what to do.

Root-Mean-Square-Error (RMSE)

The error bars for the RMSE are asymmetric. The lower and higher error bar on the RMSE for method X ($RMSE_X$) is

$$L_X = RMSE_X - \sqrt{RMSE_X^2 - \frac{1.96\sqrt{2}RMSE_X^2}{\sqrt{N-1}}}$$

$$= RMSE_X \left(1 - \sqrt{1 - \frac{1.96\sqrt{2}}{\sqrt{N-1}}}\right)$$

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2693v1 | CC BY 4.0 Open Access | rec: 3 Jan 2017, publ: 3 Jan 2017

$$U_X = RMSE_X \left(\sqrt{1 + \frac{1.96\sqrt{2}}{\sqrt{N-1}}} - 1 \right)$$

Mean Absolute Error (MAE)

The error bars for the MAE is also asymetric. The lower and higher error bar on the MAE for method X (MAE_X) is

$$L_X = MAE_X \left(1 - \sqrt{1 - \frac{1.96\sqrt{2}}{\sqrt{N-1}}} \right)$$

$$U_X = MAE_X \left(\sqrt{1 + \frac{1.96\sqrt{2}}{\sqrt{N-1}}} - 1 \right)$$

Mean Error (ME)

The error bars for the mean error are symmetric and given by

$$L_X = U_X = \frac{1.96 s_N}{\sqrt{N}}$$

where s_N is the standard population deviation (e.g. STDEVP in Excel).

Pearson's correlation coefficient, r

The first thing to check is whether your r values themselves are statistically significant, i.e. $r_X > r_{significant}$ where

$$r_{significant} = \frac{1.96}{\sqrt{N - 2 + 1.96^2}}$$

The error bars for the Pearson's r value are asymmetric and given by

$$L_X = r_X - \frac{e^{2F_-} - 1}{e^{2F_-} + 1}$$

$$U_X = \frac{e^{2F_+} - 1}{e^{2F_+} + 1} - r_X$$

where

$$F_\pm = \frac{1}{2} \ln \frac{1 + r_X}{1 - r_X} \pm r_{significant}$$

Comparing two methods

If $error_X$ is some measure of the error, RMSE, MAE, etc, and $error_A > error_B$ then the difference is statistically significant only if

$$error_A - error_B > \sqrt{L_A^2 + U_B^2 - 2 r_{AB} L_A U_B}$$

where r_{AB} is the Pearson's r value of method A compared to B, not to be confused with r_A which compares A to the reference value. Conversely, if this condition is not satisfied then you cannot say that method B is not more accurate than method A with 95% confidence because the error bars are too large.

Note also that if there is a high degree of correlation between the predictions ($r_{AB} \approx 1$) and the error bars are similar in size $L_A \approx U_B$ then even small differences in error could be significant.

Usually one can assume that $r_{AB} > 0$ so if $error_A - error_B > \sqrt{L_A^2 + U_B^2}$ or $error_A - error_B > L_A + U_B$ then the difference is statistically significant, but it is better to evaluate r_{AB} to be sure.

The meaning of 95% confidence

Say you compute errors for some property for 50 molecules using method A ($error_A$) and B ($error_B$) and observe that Eq 11 is true.

Assuming no prior knowledge on the performance of A and B, if you repeat this process an additional 40 times using all new molecules each time then in 38 cases ($38/40 = 0.95$) the errors observed for method A will likely be between $error_A - L_A$ and $error_A + U_A$ and similarly for method B. For one of the remaining two cases the error is expected to be larger than this range, while for the other remaining case it is expected to be smaller. Furthermore, in 39 of the 40 cases $error_A$ is likely larger than $error_B$, while $error_A$ is likely smaller than $error_B$ in the remaining case.

Computer code

Python code that determines statistical significance using RMSEs is available at https://github.com/jensengroup/statsig

References

[1] A. Nicholls. Confidence limits, error bars and method comparison in molecular modeling. part 1: The calculation of confidence intervals. *Journal of Computer-Aided Molecular Design*, 28(9):887–918, jun 2014.

[2] A. Nicholls. Confidence limits, error bars and method comparison in molecular modeling. part 2: comparing methods. *Journal of Computer-Aided Molecular Design*, 30(2):103–126, feb 2016.