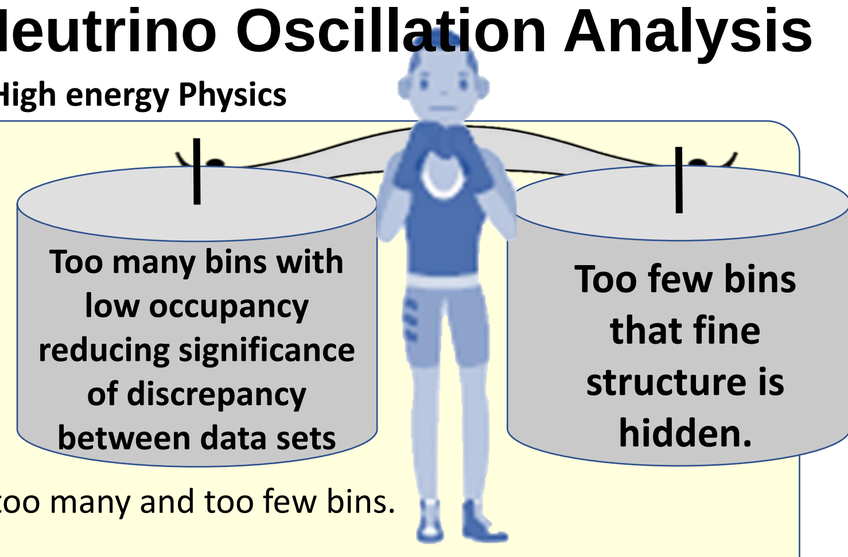
Investigating the efficacy of Novel Statistical Methods in Neutrino Oscillation Analysis

Luke Boyden & Roberto King **Supervisor: Professor Morgan Wascko** **Research Group: High energy Physics**

Why does this new statistical test think they have what it takes to go all the way?



- Multivariate analysis (multiple measurements made on each experimental unit) plays a key role in our understanding of neutrino oscillations as it is used to test the compatibility of experimental and simulated data.
- Since neutrinos are very difficult to detect, we have limited events and so the phase space in the multidimensions is sparsely populated [1]. Despite this, the compatibility is still often tested by binning the data and computing a test statistic called the **Pearson's χ^2** [2].
- Binning data always results in a loss of information and an unsatisfactory compromise must be made between too many and too few bins.
- For these reasons, it's expected that **methods that don't involve binning can perform better**, specifically the **"point to point dissimilarity method".**

What are the tests we are comparing?

In the red corner - Point to Point Dissimilarity Test

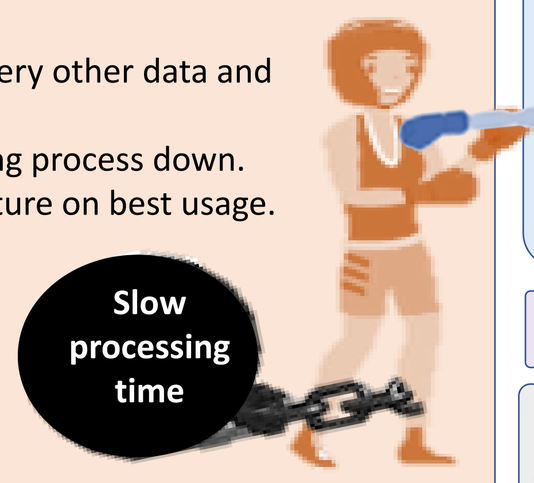
VS

In the blue corner - The Chi-Squared Test

$$T = \frac{1}{n_d(n_d-1)}\sum_{i,j>1}^{n_d}\psi\left(\left|\boldsymbol{x}_i^d - \boldsymbol{x}_j^d\right|\right) + \frac{1}{n_{mc}(n_{mc}-1)}\sum_{i,j>1}^{n_{mc}}\psi\left(\left|\boldsymbol{x}_i^{mc} - \boldsymbol{x}_j^{mc}\right|\right) - \frac{1}{n_d n_{mc}}\sum_{i,j>1}^{n_{mc}}\psi\left(\left|\boldsymbol{x}_i^d - \boldsymbol{x}_j^{mc}\right|\right)$$

Where the scripts d and mc refer to actual and MC simulated data, n refers to number of data events, and $\boldsymbol{x}$ is the event position vector [3].

- Based on the distance between a point and every other data and Monte-Carlo point.
- No loss of information - can slow the computing process down.
- Uncommonly used technique, not much literature on best usage.



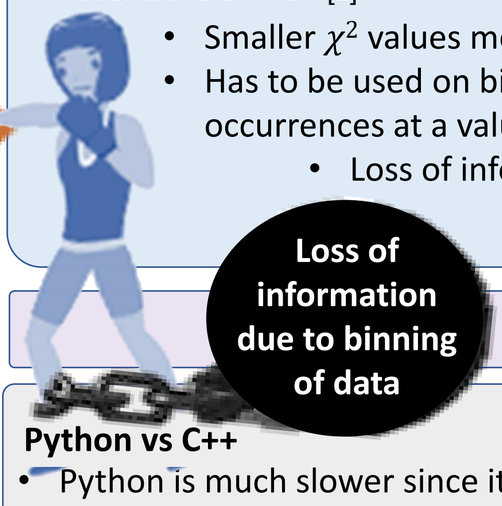
Kernel Function $\psi\left(\left|x_i - x_j\right|\right)$:
The kernel function is effectively a weighting function acting on every pair on points, we will be investigating:

- How 4D space can be rescaled to give all dimensions equal potential in affecting the value of T
- Whether the Kernel function would be more effective taking in 4 variables rather than a simple "distance" variable.

$$\chi^2 = \sum_i^{n_b}\frac{(O_i - E_i)^2}{E_i}$$

Where n_b is the number of bins, O_i is the observed and E_i is the expected number of data events in the i^{th} bin [2].

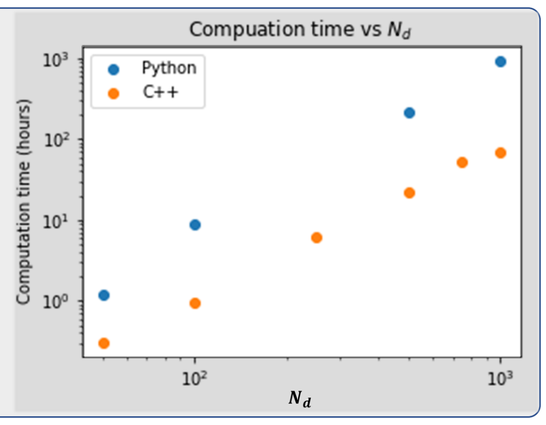
- Smaller χ^2 values mean greater agreement
- Has to be used on binned data since it tests the number of occurrences at a value/range
  - Loss of information due to binning of points causing large variations in the binned data [4].



Reducing the processing time

Python vs C++

- Python is much slower since it uses an interpreter and also determines the data type at run time.
- Python code can easily be converted into Cython code by specifying the data types, which generates C++ optimised code [5].



How are we going to compare the tests?

p-value distributions

The agreement between the data sets is quantified by the test statistics of χ^2 and T. These measure different things and so we want to use them to form a mutual quantity we can compare them on: the distribution of p-values.

The p-value quantifies the significance of any discrepancy between the data and the probability density function (p.d.f. - defines a probability for a discrete random variable). This is defined for a test statistic S as the probability of finding an S-value corresponding to lesser agreement than the observed S-value: $p = \int_S^\infty g_{f_0}(S)dS$

$g_{f_0}(S)$ corresponds to the p.d.f. of the test statistic when both the actual and MC data come from the same parent p.d.f. [6].

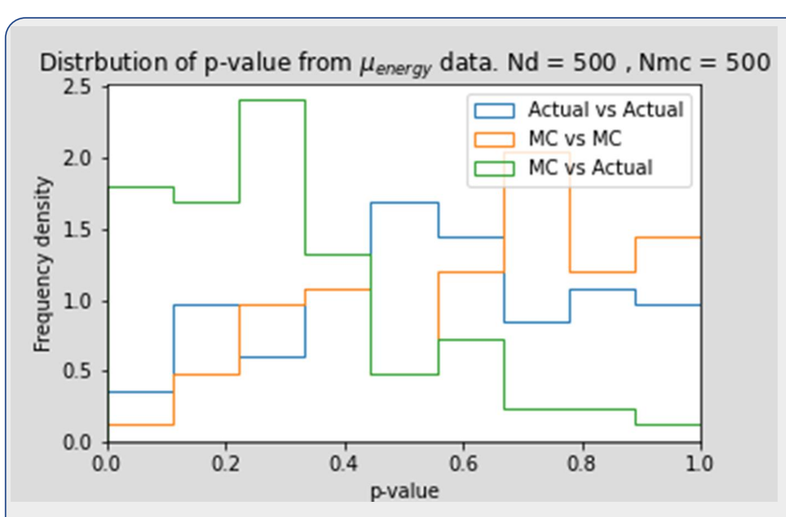
Permutation test [7]

1. Combine both the actual and MC simulated data sets into a pool of size $N_d + N_{mc}$.
2. Randomly draw a sample of size N_d and temporarily label this "actual data", label the remaining "MC simulated" data.
3. Calculate the corresponding T known as T_1 .
4. Repeat this n times to obtain $T_{perm} = \{T_{1,}T_{2,} \dots \dots, T_{N_{perm}}\}$
5. The p-value can then be estimated using:

$$p \approx \left|T_{perm} > T\right| / \left|T_{perm}\right|$$

- The p.d.f. of a χ^2 distribution with n_b degrees of freedom is well documented (and dependent on the data set) meaning the p calculation is straightforward.
- The p.d.f. of the 'point to point dissimilarity function' is dependent on the data, and is not well known meaning p must be estimated using the permutation test.

What are the results of the comparison?



Distributions of p-values were plotted (like the one on the LH side), and the rejection power at 95% confidence interval calculated.

The point to Point Dissimilarity test shown to be superior to the χ^2 test:
This is a very powerful goodness of fit tool. It showed excellent rejection power for $N_d \geq 1{,}000$, and was superior to χ^2 for all N_d tested. The downsides include: not being as easy to understand conceptually as the χ^2 test and slow computation time. However the latter can be greatly reduced by using Cython (or straight C++ code).

Rejection power at 95% confidence interval

N_d	T	χ^2
100	12%	3%
500	83%	33%
1,000	100%	67%

Ideas for further study:

- Analyse other "non binning tests" such as: the distance to nearest neighbour test, the local density test.
- Investigate using different definitions of distance between two points, that is, use values of $p \neq 2$ in the Minkowski distance (generalised distance) [8]: $D(X,Y) = \left(\sum_{i=1}^n |x_i - y_i|^p\right)^{1/p}$.

References: [1] – R.E. Bellman, Adaptive Control Processes, Princeton University Press, Princeton, NJ (1961). [2] – K, *Pearson (1900). "On the criterion […] random sampling". Philosophical Magazine. Series 5.* ***50*** *(302): 157–175. doi:10.1080/14786440009463897.* [3] - C.M. Cuadras and J. Fortiana, Distance-based multivariate two sample tests (2003). [4] - F. Yates, Contingency tables involving small numbers and the χ 2 test, Supplement to the J. Roy. Statistical Society 1, No. 2 (1934) 217-235. [5] – Cython. Website: https://cython.readthedocs.io/en/latest/ (accessed on 11/02/2022). [6] - *Wasserstein RL, Lazar NA (2016). "The ASA's Statement on p-Values: Context, Process, and Purpose". The American Statistician.* ***70*** *(2): 129–133. doi:10.1080/00031305.2016.1154108. [7] - "Invited Articles" . Journal of Modern Applied Statistical Methods.* ***1*** *(2): 202–522. Fall 2011. [8]* - H, Colakog. A generalization of the Minkowski distance and a new definition of the ellipse. Akdeniz University, Vocational School of Technical Sciences, Department of Computer Technologies, Konyaaltı, Antalya. Available at: https://arxiv.org/pdf/1903.09657.pdf. [9] – Cartoon fighters : https://www.alamy.com/stock-photo/boxer-fighter-cartoon.html (accessed on 29/02/2022)