Topical Lectures Statistics –

Exercises Set 3

Wouter Verkerke (Dec 2022)

Exercise 16 – The important of accurate modeling

Statistical inference is only valid in the limit that the probability model that is used can accurately describe the data. While this seems trivially obvious, plenty of problems can arise in practice

Here we investigate a common area of problems – model with tail regimes that assign very low probabilities to events occurring in these regions

Copy file `ex16.C`. This macro performs the following steps

- Construct a narrow Gaussian probability model: the observable range spans about 20σ of the Gaussian.
- Fit sample of 100 points, drawn from the Gaussian to the model

Run the macro and observe its behavior. *Is the fit unbiased*?

- Uncomment the second code block. In this section, a single outlier event is added 'by hand' to the dataset and the fit is repeated. *How large is the fit bias due to this single event?*
- What happens if you move the outlier event to x=9? Can you explain why only the width parameter is biased and not the mean parameter?

Outlier events like the one manually added here can occur in real data if detectors or reconstruction techniques do not behave according their assumed specifications. To avoid strong biases due to (exceedingly) rare outlier points, it is prudent to include a term in models that can absorb such unexpected events.

- A prudent solution for narrow signals like the Gaussian model studied here is to add uniform background to the Gaussian signal that can model a small fraction of events that do not conform to the expected behavior
- Construct a model that is the sum of the original Gaussian plus a Uniform background model, where the fraction of Gaussian signal is floating and initialized to a value close to 1 (but not exactly one). Consult e.g. the model of `ex04.C, ex05.C` on the syntax to construct such model. Run the fit and visualize the result in the 3rd plot panel.
- What is the loss in precision on the estimate on the mean and sigma of the Gaussian model compared to the original fit without outlier and without outlier absorption term?

Exercise 18 – Fourier Convolution models

In this final macro we return to unbinned likelihood models and explore convolution models based on discrete FFT transformation.

Copy file `ex18_convolution.C`. This macro performs the following steps

- Construct a Landau physics model and a Gaussian resolution model, then construct a FFT-based convolution of these two
- Generate 1000 events and fit the model to this data
- Plot the fitted convolution model on the data, and also make separate plots of the fitted shapes of the physics truth model (the Landau) and the resolution model

Questions and explorations

- Run the macro. Change number of events to different counts (e.g. 100k events) and observe how little CPU time is needed to perform the convolution calculations.

- Change samples size back to 1000 and hen change the value of the mean of the Landau from 30 to 80. Rerun the macro. Do you see any signs of cyclical spillover?

- Now uncomment the line that sets the overflow buffer fraction to zero and rerun. Do you see signs of cyclical spillover?

- Try to calculate some other convolutions numerically

  For example a convolution of a Gaussian with Gaussian should be another Gaussian where the width/mean is the sum of the two input Gaussians width/mean (note that you may need to fix one of the Gaussians width parameters as only the sum of the width can be constrained by the fit).

  Next try a convolution of an exponential distribution with Gaussian and see how well the fit can resolve the Gaussian smearing versus the exponential slope of the convoluted distribution. To successfully fit this convolution you should increase the spillover buffer size of the FFT convolution to 1.0 (as the exponential takes on large values close to the boundaries).