Surrogate datasets

for sharing experimental information with the theory community

Ulrik Egede, **Riley Henderson**

WARWICK THE UNIVERSITY OF WARWICK

MONASH University

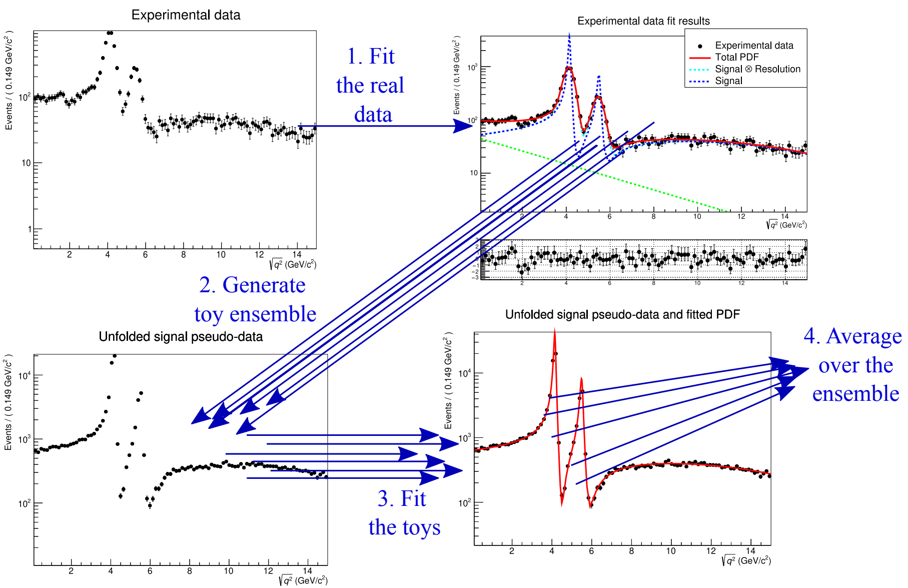
Motivation

- Some experimental analyses, particularly in flavour physics, can be quite complex.
- They may involve fitting large multidimensional datasets with complicated models to determine large numbers of free parameters.
- Often the model also contains many nuisance parameters which describe critical experimental effects, but are not of interest to the wider community.
  - One example is our (soon to be published) LHCb $B^0 \to K^{*0}\mu^+\mu^-$ amplitude analysis which involves fitting for 150 parameters, of which around 60 are nuisance parameters that we do not publish.
- Somehow how the results need to be communicated in a clear, correct, and useful way.
  - This can be difficult to achieve by simply publishing the numbers in a paper.

Surrogate data approach

- One possibility is to provide *unfolded signal only toy datasets* that we call *surrogates*, which essentially represent the real data but without the experimental complications such as resolution and background.
- An ensemble of surrogates would be generated from the covariance matrix of the full model used in the experimental analysis.
    - Uncertainties and correlations accounting for all nuisance parameters would thus be encoded in the ensemble.
- A theorist could then fit back these surrogates with whatever model they choose, neglecting experimental effects, as long as they average the results over the ensemble.

Surrogate data approach



Key points to demonstrate

- For this to be useful, it is necessary to demonstrate that the approach really works and to try out some potential use cases, *e.g.* to show that:
  1. The results of fitting back the ensemble of surrogates with the same model reproduces the central values and covariance matrix used to generate them.
  2. Alternative models can be fit to the surrogates such that the averaged results also reproduce the results of fitting the original data with that alternative model.

Key points to demonstrate

- For this to be useful, it is necessary to demonstrate that the approach really works and to try out some potential use cases, *e.g.* to show that:
  1. **The results of fitting back the ensemble of surrogates with the same model reproduces the central values and covariance matrix used to generate them.**
  2. Alternative models can be fit to the surrogates such that the averaged results also reproduce the results of fitting the original data with that alternative model.
- I have done a **proof of concept** study to demonstrate **point 1** quite clearly using a simple example model.
- Unfortunately I haven't quite made it around to demonstrating point 2 yet, but have a few ideas worth exploring.

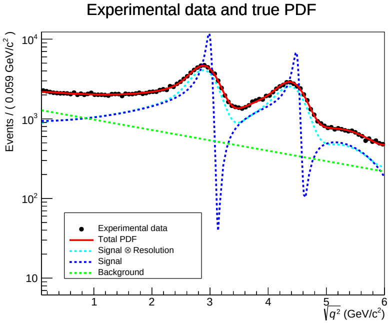
Example: Two *relativistic Breit-Wigner* resonances interfering with a non-resonant component, all convolved with a Gaussian resolution model and added on an exponential background shape.

e.g. $X_i \to X_f \ell^+ \ell^-$ dilepton spectrum:

$$\frac{d\Gamma}{dq^2} \propto |\mathcal{A}_{\text{total}}(q^2)|^2,$$

$$\mathcal{A}_{\text{total}} = \mathcal{A}_{\text{NR}} + \sum_j \eta_j e^{i\delta_j} \mathcal{A}_{\text{BW},j},$$

$$\mathcal{A}_{\text{NR}}(q^2) = \sum_i \mathcal{C}_i F_i(q^2).$$



1. Perform an amplitude analysis to measure the magnitudes and phases of the resonances relative to the non-resonant component.

Proof-of-concept - real data proxy

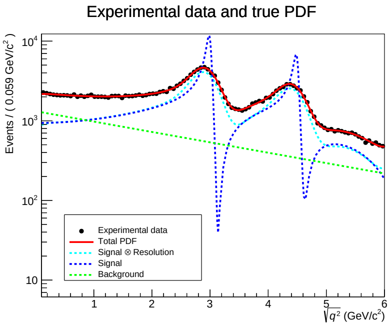
Example: Two *relativistic Breit-Wigner* resonances interfering with a non-resonant component, all convolved with a Gaussian resolution model and added on an exponential background shape.

e.g. $X_i \to X_f \ell^+ \ell^-$ dilepton spectrum:

$$\frac{d\Gamma}{dq^2} \propto |\mathcal{A}_{\text{total}}(q^2)|^2,$$

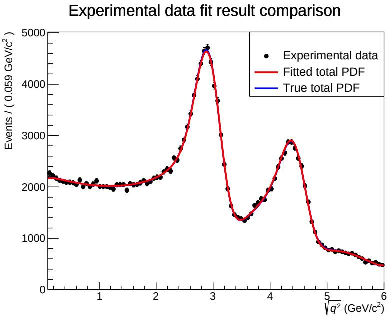
$$\mathcal{A}_{\text{total}} = \mathcal{A}_{\text{NR}} + \sum_j \eta_j e^{i\delta_j} \mathcal{A}_{\text{BW},j},$$

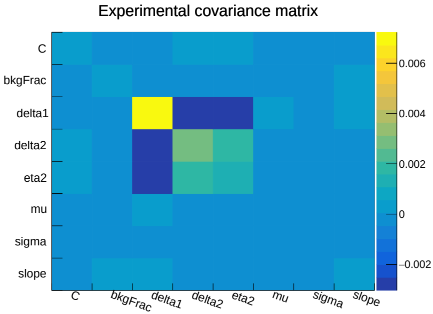
$$\mathcal{A}_{\text{NR}}(q^2) = \sum_i \mathcal{C}_i F_i(q^2).$$



1. **Generate a dataset from the full true model to act as a proxy for the "real data".**
   - This a sample of 200k events.

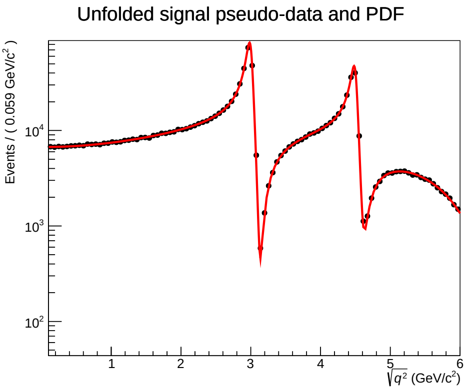
2. **Fit the "real data" proxy with the same model to obtain a set of experimental fit results.**





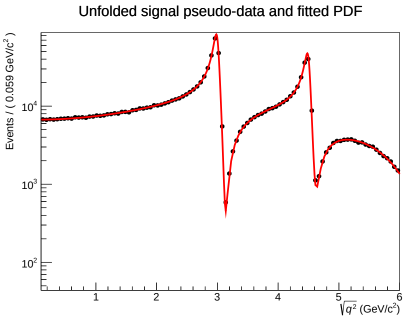
- Here we are measuring the relative phases, δ_i, and magnitudes, C and η_i, of the three signal components, plus the experimental nuisance parameters, *i.e.* background slope and fraction, width of resolution.

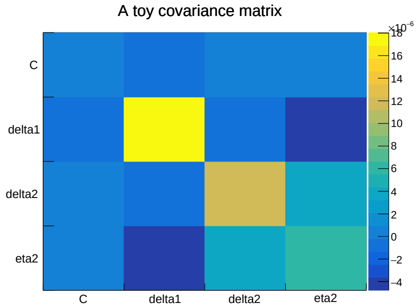
3. **Generate an ensemble of signal only toy datasets by fluctuating best fit values according to the covariance matrix.**



- An ensemble of 350 toys was generated with the background and resolution removed.
- Each toy has 1M events in this example.

4. **Fit back each toy in the ensemble using only the signal-only model.**

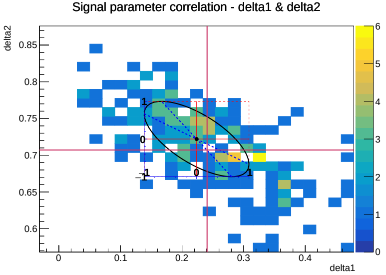


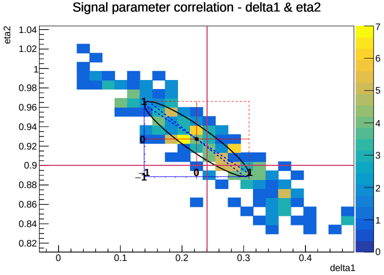


- N.b. the individual covariance matrices from these fits **do not** capture the experimental uncertainties.
- In fact, the toys should be large enough that their statistics do not contribute an additional significant source of uncertainty in the end.

Proof-of-concept - check results

5. **Analyse the results over the ensemble.**
   - Parameter central values, uncertainties, and correlations obtained by *averaging over the toy ensemble*.
   - Should all agree with those of the original fit to the real data.

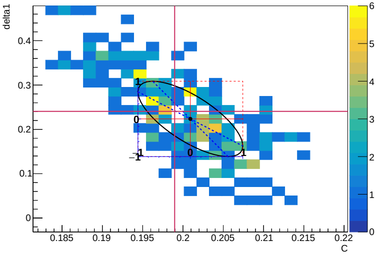




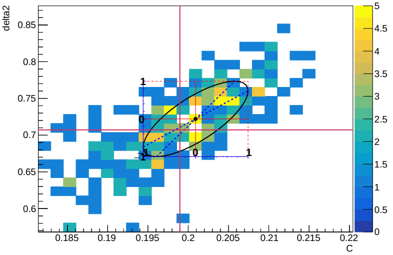
- Ellipses show correlation between fit parameters in the fit to the "real data".
- Histogram shows the distribution of fit results from the surrogate ensemble.
- Central values, errors, and correlations from experimental results are reproduced.

Proof-of-concept - check results

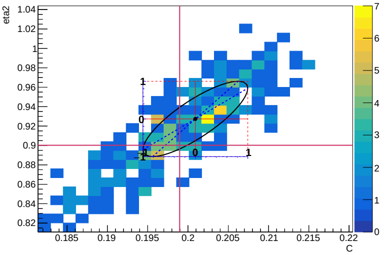
Signal parameter correlation - C & delta1



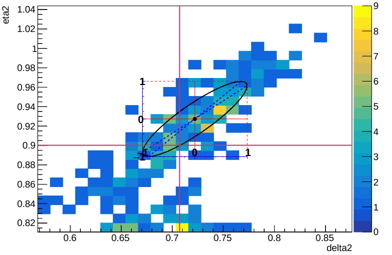
Signal parameter correlation - C & delta2



Signal parameter correlation - C & eta2



Signal parameter correlation - delta2 & eta2



Key points to demonstrate

- For this to be useful, it is necessary to demonstrate that the approach really works and to try out some potential use case, *e.g.* to show that:
  1. The results of fitting back the ensemble of surrogates reproduces the central values and covariance matrix used to generate them.
  2. **Alternative models can be fit to the surrogates such that the averaged results also reproduce the results of fitting the original data with that alternative model.**

Key points to demonstrate

- For this to be useful, it is necessary to demonstrate that the approach really works and to try out some potential use case, *e.g.* to show that:
  1. The results of fitting back the ensemble of surrogates reproduces the central values and covariance matrix used to generate them.
  2. **Alternative models can be fit to the surrogates such that the averaged results also reproduce the results of fitting the original data with that alternative model.**

Below are some ideas we have considered for showing how this can be interesting:

1. Fitting for interfering resonance magnitudes and phases often results in multiple solutions.
   - It would be good to show that the multiple solutions persist in the surrogates, *e.g.* fit back starting near expected symmetry points.

Key points to demonstrate

- For this to be useful, it is necessary to demonstrate that the approach really works and to try out some potential use case, *e.g.* to show that:
  1. The results of fitting back the ensemble of surrogates reproduces the central values and covariance matrix used to generate them.
  2. **Alternative models can be fit to the surrogates such that the averaged results also reproduce the results of fitting the original data with that alternative model.**

Below are some ideas we have considered for showing how this can be interesting:

1. Fitting for interfering resonance magnitudes and phases often results in multiple solutions.
   - It would be good to show that the multiple solutions persist in the surrogates, *e.g.* fit back starting near expected symmetry points.
2. Fitting for the non-resonant amplitude $\mathcal{A}_{\text{NR}}(q^2) = \mathcal{C}F(q^2)$ uses "theory input" to fix/constrain the "form factor". Alternative input will modify the "Wilson coefficient".
   - Fit back surrogates with alternative "form factor" model, $F(q^2)$, and see if the change in $\mathcal{C}$ is recovered.

Additional considerations

1. There is *model dependence* baked into the surrogates.
   - To consider an extreme example, one could never search for a hypothetical third resonant contribution in the example given earlier — it simply wasn't generated in the surrogates.
   - The best fit to the surrogates will always be given by the model that was used to generate them.

Additional considerations

1. There is *model dependence* baked into the surrogates.
    - To consider an extreme example, one could never search for a hypothetical third resonant contribution in the example given earlier — it simply wasn't generated in the surrogates.
    - The best fit to the surrogates will always be given by the model that was used to generate them.
2. With that said, if one can somehow show that the description of background and experimental effects is accurate in the full model AND that the full model gives a good fit to the data, then it follows that the signal must be well described by the model too.
    - In that sense, it makes sense to use the surrogate data approach to test the compatibility of alternative signal models with the data and investigate how parameters of interest might change.
    - This would amount to reattributing features of the data to different physics parameters, *e.g.* WCs vs FFs or non-local interference.

Summary

- We are proposing a method of sharing complicated experimental results with the theory community that we call the surrogate data approach.
- It involves generating ensembles of toy datasets with experimental nuisance effects removed, *i.e.* unfolded signal only toys.
- Experimental uncertainties and correlations are encoded in the ensemble by fluctuating model parameters according the covariance matrix.
- The basics of the approach have been validated with simple example models.
- I have highlighted a few potential limitations of the approach.
- Some additional demonstrations of possible use cases are still on the to do list.