Highlights from the Slack Channel of Our BIRS-Systematics Workshop

This document aims to consolidate the key points of the discussions that have taken place on our BIRS-Systematics Slack Channel, starting from the beginning of the workshop on April 23, 2023, until May 12, 2023. Although the Slack Channel will remain open for further discussions, each post will only be visible for 90 days. Recognizing the valuable contributions made by the participants on Slack, the organizers deemed it necessary to create this document as a point of reference for future use.

0. Miscellaneous

0.1 Photos

Various pictures available here and here (Thank you to Phillipp Windischhofer, Purvasha Chakravarti, Richard Lockhart and Lydia Brenner for sharing them).

0.2 On the difference between parametric bootstrap and ``throwing toys''

- **Roger Barlow:** I think there is a difference between 'throwing toy Monte Carlos' and 'the parametric bootstrap'
  The bootstrap involves the use of the dataset to tell you about itself. If you do an experiment, get a dataset, extract some parameters from it, and use those parameters to throw toys and get an ensemble of datasets you can use those to tell you about properties of the experimental dataset - just as in the standard (non-parametric) bootstrap you get an ensemble of datasets by resampling. If you just choose some parameters because you're interesetd in them, perhaps even without doing the experiment, to study coverage or suchlike, this is not a bootstrap.
  It's only the parametric bootstrap if it comes from the Region Parametrique de Bootstrappe. Otherwise it's just sparkling random numbers.

- **Bob Cousins**: Yes, I use parametric bootstrap to mean toys based on parameter values from (profile) likelihood maximization. (Otherwise, that is not using one's bootstraps.) If I said parametric bootstrap in any other context, I did not intend to. I do not recall hearing anyone else use it in another context.

1. Overview on Systematics

1.1. Questions of interest

- **Sara Algeri**: Below are a list of discussion topics/questions proposed by Nick and myself in our respective talks (the full sets of slides are available at https://www.birs.ca/workshops/2023/23w5096/files/).Nick Wardle proposed discussion items
  1-When reporting uncertainties in OPAT is providing covariance enough?
  2- When modelling systematic uncertainties using nuisance parameters, Should we sample many different parameter values to build suitable parameterisation and are

there smarter ways (eg GPs/ML?) to automate this?
3- Are we ok that our fit updates our knowledge of certainty nuisance parameters?
4- Is there a better way to include uncertainties due to model choice than taking difference between (eg) simulations? Or approaches such as inflating uncertainty to cover potential bias / discrete profile method? Sara Algeri proposed discussion items
1- In the context of background mismodelling, can the difference between methods be used to acquire some notion/measure systematic bias?
2- When dealing with nuisance parameters is it at all possible to reach a consensus on what to do when? (e.g., marginalizing or profiling)
3- Can a statistician effectively access published likelihoods?
4- How to check the validity of regularity conditions needed by classical statistics when dealing with complex models?
5- What do we need to robustly bridge the statistics and physics communities?

1.2. Irreducible error vs irreducible background

- **Sara Algeri:** (…) As a follow up on a few discussions I had (both in person and via email) with some of the participants, allow me to clarify that what I call ``irreducible error'' in my slides 3-4 is NOT the same as the ``unresolved background'', but rather, as defined on slide 3**,** such wording refers to the collection of ``unknown unknowns'' which may affect the analysis. I was not aware of the fact that the unresolved background is also referred to as ``irreducible background'' in physics, hence the source of the confusion.

2. Marginalizing vs profiling

2.2. On misspecified priors for the nuisance parameters

- **Tom Loredo**: One or two talks in this session mentioned work on behavior of marginal distributions when the prior for nuisance parameters was "misspecified." I'm struggling to understand what it means for a nuisance parameter prior to be misspecified, unless there is an assumption of some kind of replication structure that is mysteriously being ignored when computing the marginal. Would someone explain what "misspecification" means here, or provide some pointers to studies illustrating the problem where I could see exactly what's being assumed?

- **Anthony Davison**: (…) I was the person making this comment. The paper I referred to can be found at https://www.nuffield.ox.ac.uk/users/cox/cox343.pdf and concerns replicated data (matched pairs of Poisson observations) and then studies the efficiency of modelling them either using random effects (corresponding to marginalisation to estimate the treatment effect) or using a conditional analysis (corresponding to removing the nuisance parameters by profiling in this particular setting). The conclusion (numerical results in Table 1) is that (very broadly) the information gain from the random effects model is rather limited (but of course one would need to check that a physics situation would broadly correspond to the lower part of the table). The comments about robustness correspond to Section 7 of the paper, with numerical work in Table 2 that (Section Eight) suggesting that

inconsistency of the interest parameter can (easily?) arise if the marginalisation distribution is not well-formulated.

3. Pragmatic vs full likelihood approaches

3.1. Similarities with cutting feedback and modularization

- **Tom Loredo:** In the Q&A after David's talk on pragmatic Bayes, I asked if it was related to cutting feedback and modularization in the literature on Bayesian graphical models (probabilistic graphical models, hierarchical models, latent variable models, Bayesian networks...). I believe pragmatic Bayes is a special case of this more general idea. I offered to post some links to the literature here in Slack. Well, it's an idea I've looked into on and off since first hearing about it at a SAMSI meeting on emulators for complex models, around 2007. I dug up my scattered notes on the literature from over the years and assembled them into a Markdown document. I'll post a PDF version here, as well as some excerpts to help potential readers decide if this is of interest (if anyone would like the Markdown source, let me know). I should emphasize it's a tool I have not yet used myself, though I heavily use graphical models.

  Jonathan Rougier (statistician working on climate model emulators) on cutting feedback:
  When making probabilistic statements about complex physical systems like climate, it is the end-product that we sign-off on: the probability that global mean temperature in 2100 is two degrees higher than today, for example. How we get there and how we document our journey, in the papers we write and the seminars we give, is an important part of establishing the authority of our assessment. But it is mistaken to think that this authority stands or falls on a simple audit of formal correctness. I'm sure we are all aware of the limitations of probability as a model for reasoning, and to insist on coherence in the development of our inference is rather like treating our climate models as perfect: something we might do as an expedient and temporary place-holder, while we develop a more nuanced approach.

  Berger, Bayarri, and Liu on modularization:
  Bayesian analysis incorporates different sources of information into a single analysis through Bayes theorem. When one or more of the sources of information are suspect (e.g., if the model assumed for the information is viewed as quite possibly being significantly flawed), there can be a concern that Bayes theorem allows this suspect information to overly influence the other sources of information. We consider a variety of situations in which this arises, and give methodological suggestions for dealing with the problem.

  Jacob et al. (including Christian Robert, known to some of you) on modularization:
  In modern applications, statisticians are faced with integrating heterogeneous data modalities relevant for an inference, prediction, or decision problem. In such

circumstances, it is convenient to use a graphical model to represent the statistical dependencies, via a set of connected "modules", each relating to a specific data modality, and drawing on specific domain expertise in their development. In principle, given data, the conventional statistical update then allows for coherent uncertainty quantification and information propagation through and across the modules. However, misspecification of any module can contaminate the estimate and update of others, often in unpredictable ways. In various settings, particularly when certain modules are trusted more than others, practitioners have preferred to avoid learning with the full model in favor of approaches that restrict the information propagation between modules, for example by restricting propagation to only particular directions along the edges of the graph. In this article, we investigate why these modular approaches might be preferable to the full model in misspecified settings. We propose principled criteria to choose between modular and full-model approaches. The question arises in many applied settings, including large stochastic dynamical systems, meta-analysis, epidemiological models, air pollution models, pharmacokinetics-pharmacodynamics, and causal inference with propensity scores.

4. Likelihood-free inference

4.1. On smoothness to ensure validity of interpolation

- **Jim Linnemann:** (…) interpolation assumes smoothness, but Neyman construction for small n enforces jumpiness. How do you get around this conflict?

- **Ann Lee:** The main and calibration branches of our LF2I framework estimate, respectively, the test statistic (such as the LR statistic) and the alpha-level cutoff **as functions of the parameter theta** --- and not of the data/features. It's not immediately clear to me that there would be 'jumpiness' across parameter space (but I may be misunderstanding your question). Nevertheless, the general LF2I framework allows the user to choose a regression method that is appropriate for the problem at hand. We've been using NNs as these can handle high-dimensional inputs and lead to smoother estimates, but one could in principle also use other regression methods that are adapted to discrete parameter spaces, discontinuities etc.

5. Model Selection

5.1. Score tests

- **Tom Loredo:** In his model selection talk, Chad briefly described score tests. For those unfamiliar with them, my talk for the 2010 Banff PhyStat meeting included a quick-and-dirty 2-slide description of the connection of score tests to maximum likelihood ratio (MLR) testing for small departures from the null. See slides 61 & 62 here on the Banff 2010 site: https://www.birs.ca/workshops/2010/10w5068/files/loredo.pdf. The interesting

thing about a score test is that it approximates a MLR test for a small change in a parameter from its null *without having to find the MLE under the alternative to the null*. I learned about score tests from John Rice and Peter Bickel, who introduced them to astronomers at an earlier SCMA meeting (ca. 2006). They still are seldom used in astronomy, however.

6. Background and signal shapes

6.1. On the spurious signal approach

- **Lydia Brenner:** (…) the description of how spurious signal is done is written in section 5.5 of https://arxiv.org/pdf/1207.7214.pdf where they call it the description of the bias estimation (later renamed to spurious signal), and this is the official Atlas reference for the spurious signal method.
  A slightly more specific description of the method can be found here; https://cds.cern.ch/record/2743717/files/ATL-PHYS-PUB-2020-028.pdf in section 3.1

- **Sara Algeri:** (…) the ``spurios signal method'' essentially the same as the ``safe-guard'' method described in this paper: https://iopscience.iop.org/article/10.1088/1475-7516/2017/05/013 (...) describes the method quite nicely from a statistical point of view so it many be more accessible to statisticians

- **Lydia Brenner:** The thing that can still be discussed is if/how this method can be adapted if you know where your signal is expected to be, since all these descriptions are on how to safe-guard against modelled background uncertainties in the case where you don't know where the signal lies

7. Template morphing

7.1. Additional references

- **Lydia Brenner:** If you want to know more you can read about the effective Lagrangian morphing here; http://cds.cern.ch/record/2066980/files/ATL-PHYS-PUB-2015-047.pdf or in my thesis in chapter 5 and 6 here; https://cds.cern.ch/record/2292147/files/CERN-THESIS-2017-220.pdf
  if you want to read how this works for EFT models you can read more here; https://arxiv.org/abs/2202.13612

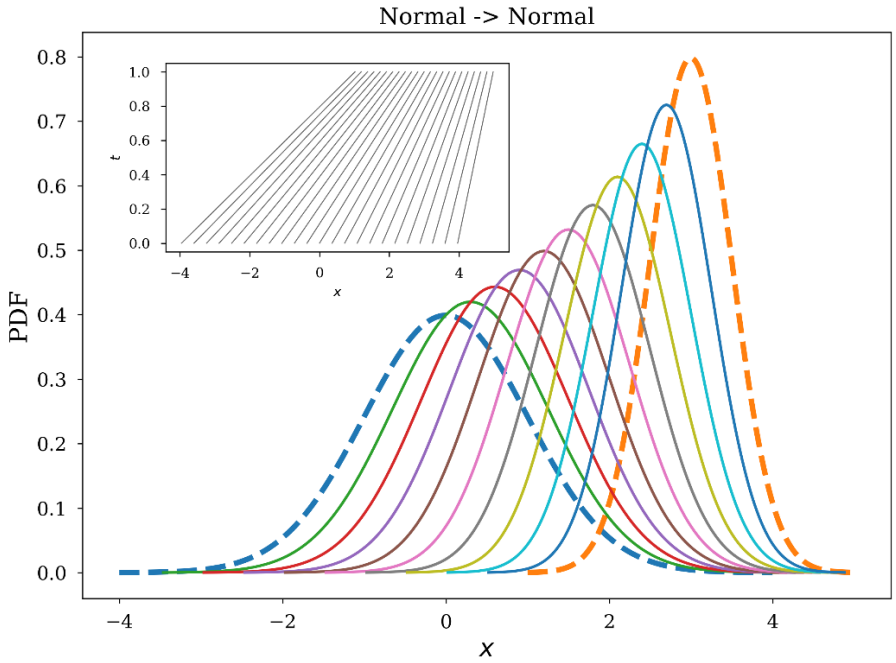
8. Optimal transport

8.1. Geodesic interpolations via Optimal transport

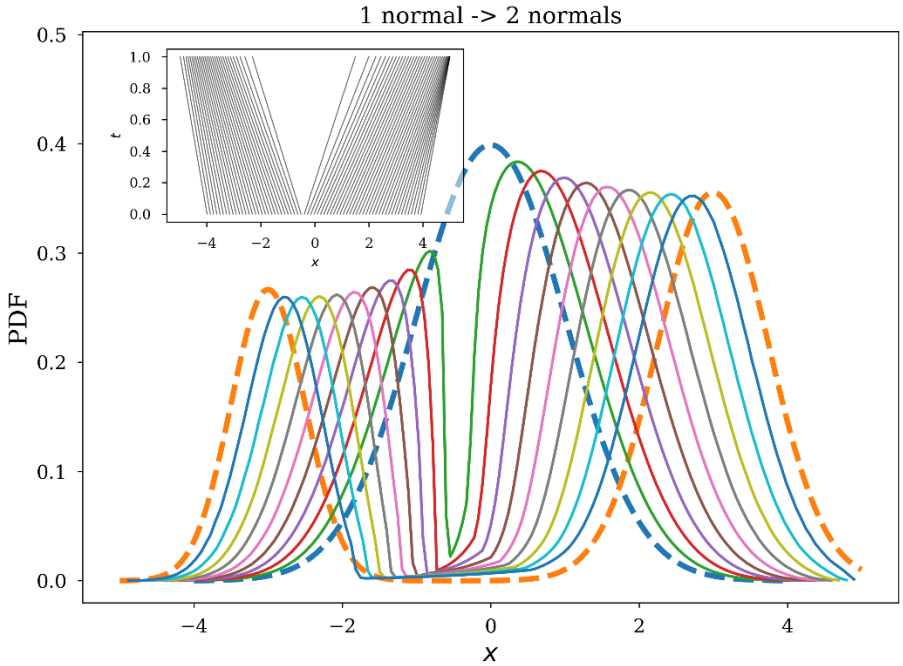
- **Tom Loredo:** Banffers, I am quite far from an expert on optimal transport. But I thought it would be worth sharing some examples of 1D density geodesic interpolations via OT that I did as an experiment about two years ago. In yesterday's discussion after the OT talk, there was some verbal description of things that could happen; some of the plots I'll show here give some concrete examples of the behavior that was surmised and discussed.There are quite a few papers and blogs with OT tutorials that show some appealing interpolations. I haven't come across many (well, any) showing unappealing behavior. I'll show some here. The behavior makes sense once you think about it, but it illustrates that OT may not always interpolate the way one would like.Incidentally, regarding OT tutorials, there are excellent long-form tutorials out there. But for a quick start, the best entry I've come across is the note titled "Optimal Transport and Wasserstein Distance" that our own Larry Wasserman wrote for one of his CMU courses (Larry is attending remotely, but not currently in Slack). You'll find it here: 36-708 Statistical Machine Learning, Spring 2018.

  Background: One of my research groups wanted to do quick-and-dirty interpolation of sequences of galaxy spectral energy distributions (SEDs). Roughly speaking, a galaxy SED is composed of a smooth continuum with lines superposed on it (mainly emission lines). The most important factor influencing the SED is the age of the galaxy. The continuum peaks in the blue end of the spectrum for young galaxies (dominated by young stellar populations dominated by hot stars), and in the red end of the spectrum for old galaxies (dominated by longer-lived smaller and cooler stars). The line features are more prominent in the bluer SEDs than in the redder SEDs. They are atomic spectral lines, so of course their positions don't change across the sequence, just their amplitudes. This is a rather crude overview of the properties of galaxy SEDs, but it should suffice for these purposes.To test some algorithms, we hoped to build a simple tool that would take a dozen or so SEDs in a sequence (from a sequence of simulations, or a human-defined sequence from observations) and interpolate smoothly between them. Interpolation via OT seemed like a good candidate. Note that this is a simpler problem than was addressed in the talk—we have high S/N data that basically serve as given density functions; we're not trying to interpolate a family of densities from *samples* from two members spanning the family.

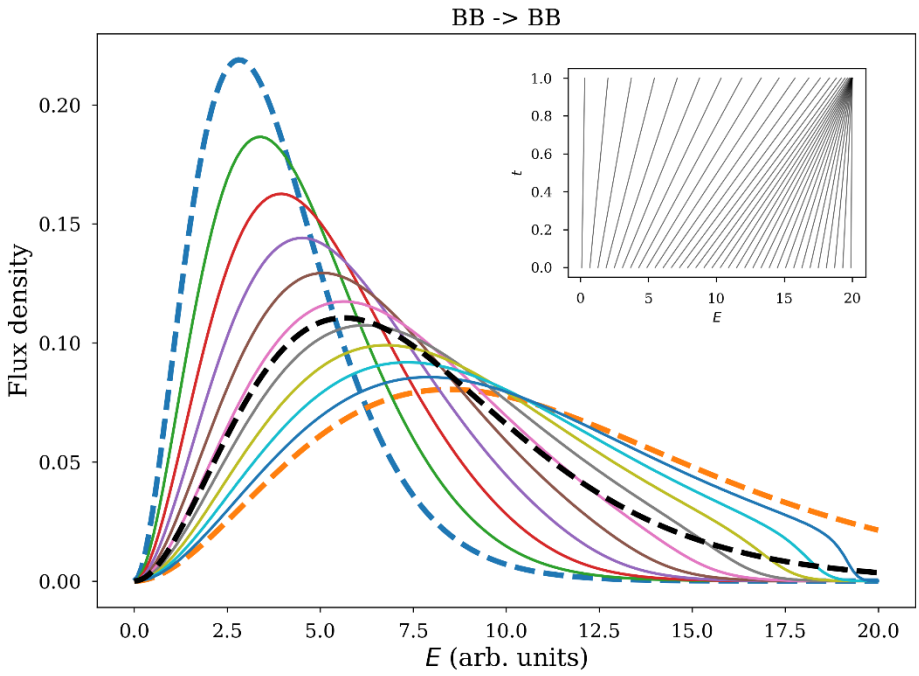
  Two normals: As initial tests of our OT interpolation code, we duplicated the usual kinds of examples. Here is interpolation from a wide normal distribution to a narrow one. The two endpoint distributions are plotted as thick dashed curves; the solid thin curves are interpolants. The inset shows the coordinate map that takes you from the wide normal (bottom) to the narrow one (top), as a function of the [0,1] interpolation scalar, t (ordinate). All is well.



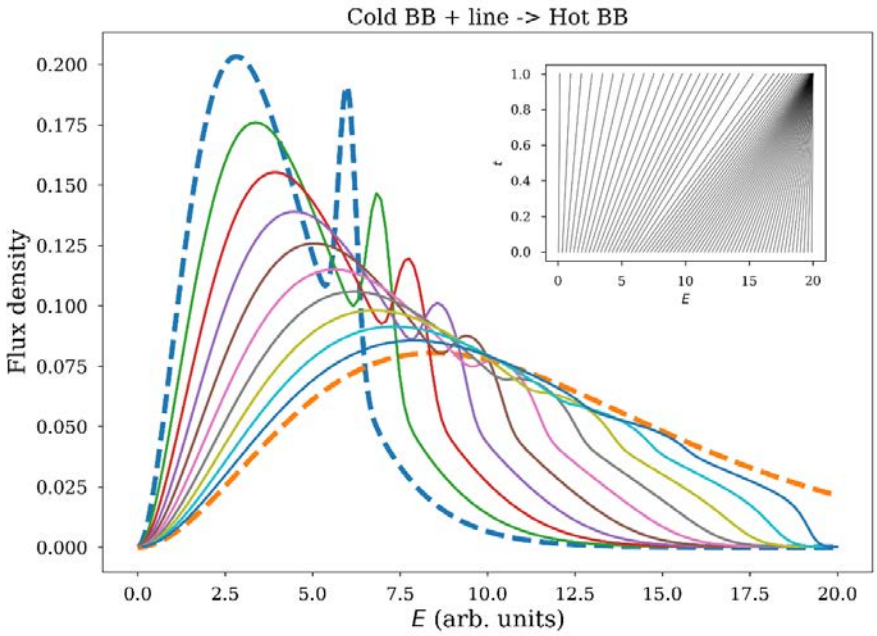
One normal to two: Here's another mostly appealing example: interpolating between a single normal and a mixture of two normals straddling the original one. It is not splitting the original normal into two evolving nearly-normal clumps; especially on the left, you can see that the shape of the bump is very asymmetric at first. Still, there's no reason to expect OT to "know" it should be splitting one normal into two, so it's hard to complain about what it's doing here.



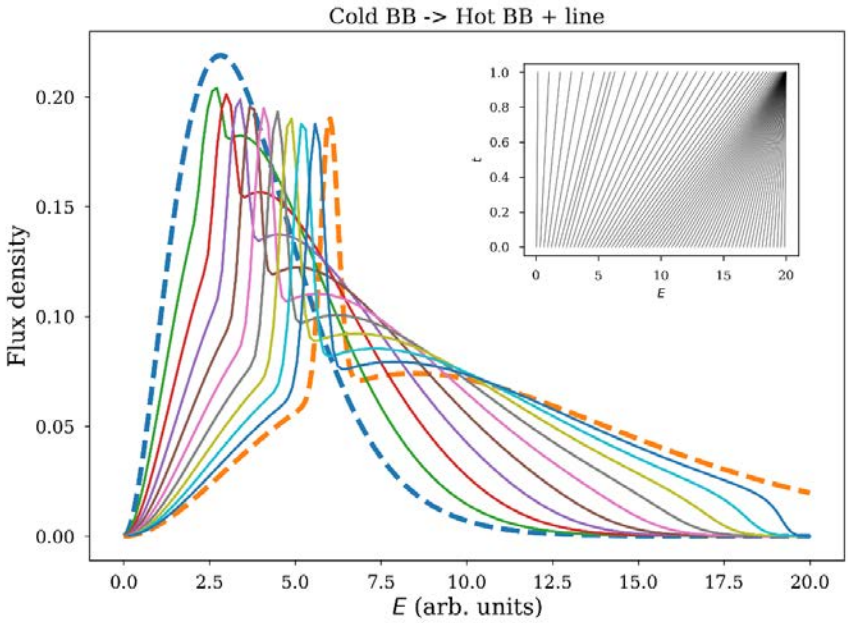
Two BBs: Next, we tried some examples meant to *very* crudely capture some elements of SED interpolation. We used black body (BB) spectra (Planck distributions) for continuum models, and superposed a single Gaussian line on some of them. First, we tried interpolating between two BBs with *no* lines. In this plot, the middle thick dashed curve is a BB spectrum with an intermediate temperature, so we could see how much the interpolants departed from the BB form. There's some weird behavior happening at the high energy (E) end, but it is mostly doing what we'd hope for in terms of qualitative behavior. (The weird tail behavior may be an artifact of the finite support chopping off a not-quite-negligible part of the hotter spectrum.)Incidentally, this two-BB example raises an issue that may deserve some attention. If we wanted to use the interpolants for inference of the temperature, how should we relate the interpolation coordinate, t, to the temperature? It seems to me that that's a nontrivial problem.



Line to no line: Next, we took a cold BB with a line and mapped it to a hot BB *without* a line (not like galaxy SEDs, where the hot ones have lines, but interesting anyway). It would have been nice if OT somehow just damped the line away while keeping it in the same position. Instead it interpolates a series of spectra with line-like features that shift and spread to higher energy. When you think a bit about what it's "told" to do—shift the mass around as little as possible to redistribute it—it makes sense. To keep the line in place, it would somehow have to move mass from the left of the line to the right of it. That's not "optimal" in terms of the default cost function (Wasserstein). It just slides all of the mass toward the right, taking the line with it.

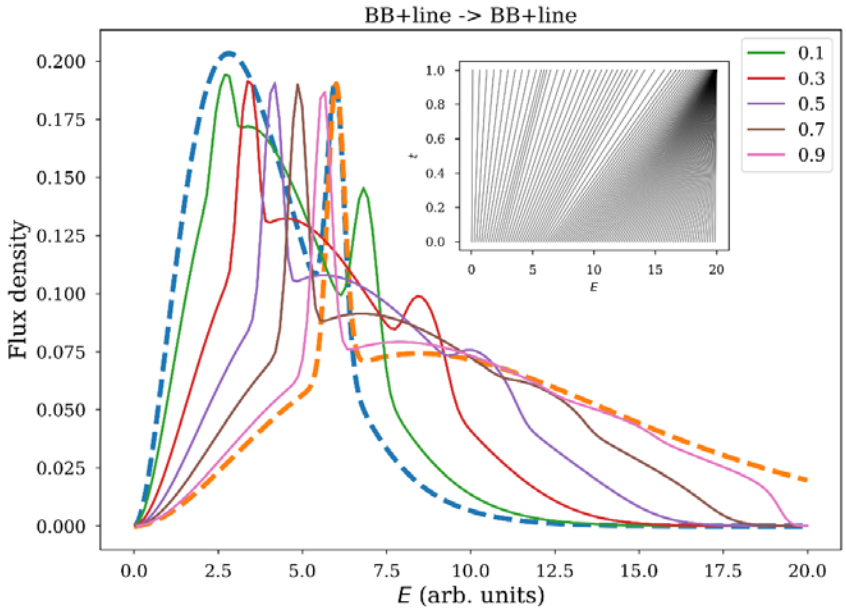


No line to line: Next is something crudely more SED-like, a cold BB with no line mapped to a hot BB with a line. The algorithm builds up the line from the left rather than from "underneath," which makes sense given the way the OT optimization problem is set up. Alas, that's nothing like how SEDs behave.



Cold to hot with fixed line: Finally, we wondered if having a line in both spectra would sort of "lock" the line in the interpolations. So we tried a cold BB with a line mapped to a hot BB with the same line (shown with just 5 interpolants here, because so much is going on). Instead of the line staying

in place, its mass gets shifted to the right to fill up the blue (hi-E) end of the spectrum. The line gets rebuilt from the left by moving mass from the cold BB peak into the line location. So instead of the interpolation maintaining the line in place (one might hope by refilling it from the left as quickly as it emptied it to the right), it creates interpolants that go from having a line to having two lines and then back to having the original line again. This one was more surprising than the others, though in retrospect it makes sense, esp. once one sees the maps—e.g., you can see that clump of lines starting near x=2 just steadily shifting mass from the blue peak to recreate the line, while the original line is used to fill up the hi-E tail. It's also a good illustration of a remark that either Philipp or Tudor made: "no mass can stay in place." So even though there's a line in a fixed location, OT interpolation can't just leave the corresponding mass there. The final line is built up of different "pieces" of the SED than were in the original line. That's the only way OT can try to maintain a steady feature, and evidently it's hard to balance the flow to keep a feature in place.



Hopefully having some examples with unappealing (but retrospectively sensible) behavior will help build insight into what OT does. It would be interesting to explore how the behavior can be influenced by altering the cost function. I have to say the complexity of behavior exhibited in these 1D examples makes me skeptical that I could build sound intuition about what OT interpolation is doing in many dimensions. But I suppose I should be skeptical of my intuition about any high-dimensional phenomenon!

9. Data Challenge

9.1. On Banff Challenge 2

Thomas Junk: I have added the now 13-year-old Banff Challenge 2 problem statement and data files to a subdirectory of the Banff Challenge 3 directory on my Google Drive.https://drive.google.com/drive/folders/1i2yDyiQo7wQOw0hGv2guwSPwAgIuCfdoI have the answer key files on my laptop (backed up of course) and can provide them if people want to score their own answers. My old scoring programs ran using PAW. I still have access to a computer that can run PAW but possibly not forever.

The writeup describing submissions is part of the Phystat 2011 proceedings: http://cds.cern.ch/record/1306523

10. Errors on errors

10.1. Notation and quantities involved

- **Nick Wardle:** (…) On slide 40, for w* how are M and E[w] calculated for the simple setup used to study the asymptotic properties shown in the slides? I guess in that case M=1 (right?) but what does E[w] look like and how was that (practically) determined?

- **Enzo Canonero:** (…) the expectation value can be calculated analytically to the order (n^-1) using a method known as the "Lawley formula." This formula is applicable to a broad range of statistical models, including our Gamma-Variance model. If an analytical computation is not feasible, one can utilize Monte Carlo (MC) simulations to estimate the expectation value, which generally requires fewer iterations compared to sampling the distribution of the statistic in question. For instance, when applying a "Bartlett correction" to a goodness-of-fit statistic with the aim of catching a 4-sigma effect, at least one million events have to be generated. However, estimating an expectation value typically necessitates a significantly lower number of iterations.If you want to know some more details about these methods and how we apply them to our gamma variance model you can give a look to the paper Glen Alessandra and I just pushed on the arXiv: https://arxiv.org/abs/2304.10574.

- **Nick Wardle:** thanks for the answer, the MC approach sounds very reasonable indeed. I think Bob made the comment that we rarely use those test-statistics with better asymptotic properties at the LHC so I was interested to see how easily it can be implemented generally, seems like it could be very straightforward. Thanks again!

10.2. Handling outliers

- **Olaf Behnke:** (…) here is another point for discussion following what I said orally in the discussion: I would find it interesting to see how the proposed (by Glen and Enzo) treatment of outliers in data combination compares to the method of M-estimates as discussed in the BIRS talk from Volker Blobel, see section 6 of https://www.birs.ca/workshops/2006/06w5054/files/volker-blobel.pdf In the method of M-estimates measurements are down weighted if they are more than some not so small distance away from the weighted average, e.g more than 4 sigma.

10.3. On the inputs provided by the experiments

- **Bogdan Malaescu:** Hello. I wanted to follow-up on one of the discussions earlier, concerning to the inputs that can be / are provided by the experiments. The ATLAS jet energy scale uncertainties are publishes since many years together with the uncertainties on the uncertainties and on the correlations (see e.g. Sec. 13.7 and Fig. 41 of 1406.0076).
- Similarly, the jet cross-sections are published with uncertainties on uncertainties and on correlations, both statistical and systematic. Several uncertainty and correlation models are published in order to provide the information on the precision with which this information is known (instead of publishing a single set of uncertainties, as it is done for many other measurements in the literature). (see Sec. 10.3 and Appendix of 1706.03192, as well as the corresponding HEPdata entry) A bootstrap procedure is also used to evaluate a statistical uncertainty on the evaluation/propagation of the systematic uncertainties, although the dominant effect generally comes from systematic uncertainties on uncertainties (discussed in the paper and provided in HEPdata).
- I think such approach should be adopted in other areas. I try to promote this in the area of the hadronic contribution to the muon g-2 theoretical prediction, but there are certainly many other studies where this is relevant.

10.4. Additional reference

- **Enzo Canonero:** This is the Bayesian “errors-on-errors” paper I was talking about: https://arxiv.org/abs/hep-ex/9910036

- **Tom Loredo:** Some current work in astrostatistics by Massimillano Bonamente, a high-energy astrophysicist at U. Alabama, Huntsville, may be of interest in this discussion: Systematic errors in the maximum likelihood regression of Poisson count data: introducing the overdispersed chi-square distribution - NASA/ADS.

11. Systematics in neutrino analyses

11.1. Additional references

- **Bogdan Malaescu:** For the discussion on the migration effects for binned fits you may want to also have a look at this document: https://cds.cern.ch/record/2839912/files/ATL-PHYS-PUB-2022-046.pdf

12. Machine learning for reducing systematics

12.1. On Cherenkov tank array optimization program

- **Thomas Junk:** The Cherenkov tank array optimization program looks like it might benefit from taking advantage of symmetry. Some existing arrays are hexagonal perhaps only for convenience. Azimuthal symmetry sounds like it might be imposable without loss of sensitivity. A square array may be sub-optimal.

13. Semi-supervised classifiers

13.1 Additional references

- **Sara Algeri:** Here is the paper I was talking about where it is shown that when simulating GOF test statistics in presence of nuisance parameters using the non-parametric bootstrap a bias correction is needed. It also shows that the problem does not occur when using the parametric bootstrap http://repository.ias.ac.in/71910/1/116_PUB.pdf

- **Ann Lee:** Here are the papers I mentioned by Ilmun Kim et al on a related regression test (prob classification instead of binary classification): See Section 2.2 of https://arxiv.org/abs/1905.11505 for a short description of the regression test. *Theorem 2 says that if the chosen regression ('prob classification') estimator has a small MISE, the power of the test is large over a wide region of the alternative hypothesis. What this means in practice is that we should choose a regression method that predicts the "class membership" Y well.*
(It may be a NN classifier or some other prediction method)Here's our 2017 applied astronomy paper (to compare distributions of galaxy images) https://academic.oup.com/mnras/article/471/3/3273/3979021?login=falseFor the stats theory paper, see https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-13/issue-2[…]ocal-two-sample-tests-via-regression/10.1214/19-EJS1648.full

13. Systematics in flavor physics

13.1 Additional references

- **Pueh Leng Tan:** The Yellin optimal interval method was quite popular in the dark matter direct detection community a while back when the background was not well modelled. But I just realised that this probably worked for us because the dark matter

spectrum is not as narrow as yours. And asymptotic assumption usually doesn't hold for for us but we toy MC out the test statistic distribution, but that can be computationally expensive depending on how complicated your likelihood is.https://arxiv.org/abs/physics/0203002

14. Final thoughts on the meetings (Summary slide available here)

14.1 Additional references

- **Olaf behnke:** Mikael's point 5) on Model Discrepancy: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9868.00294 Article from Kennedy & O'Hagan

- **Lydia Brenner:** Nonlinear Regression Analysis and Its Applications by Bates and Watts

- **Alessandra Brazzale:** ARBrazzale_PhD-thesis and here's one example of the plots