Thoughts on the Meeting: A Statistician's Perspective

Mikael Kuusela

Department of Statistics and Data Science,
Carnegie Mellon University

PhyStat-ML 2024

Imperial College,
London, UK

September 12, 2024

Outline

In this talk, I'll make some selected remarks on the following topics:

1. Interpretability
2. Model-agnostic searches / anomaly detection
3. Unfolding with ML
4. Generative models
5. Simulation-based inference

Disclaimer: This will not cover everything we've seen during the workshop. Also, this will be more about adding thoughts / context / discussion points and less about reviewing the talks.

Interpretability

What is interpretability and why do we need it?

$\rightarrow$ "Warm, fuzzy feeling that you understand what your NN is doing" (J. Thaler)

Do we always need interpretability?

- Consider some really complex parametric model $p(\boldsymbol{x}|\boldsymbol{\theta})$
- Imagine that you're able to find the MLE $\hat{\boldsymbol{\theta}}$ but it is some really messy function of $\boldsymbol{x}$
- Usually at this point we're happy without starting to ask what features of $\boldsymbol{x}$ the MLE is picking up

Similarly in ML we usually understand from the loss function what the neural network is asymptotically learning (class probability, likelihood ratio, score, conditional mean,...)

- Do we always really need to understand what features the NN is picking up in those cases?

But in certain use cases interpretability seems highly desirable

- For example, in model-agnostic searches / anomaly detection we would like to understand what the NN is seeing in the high-dimensional feature space

Interpretability

How realistic is interpretability?

Notice that interpretability is *really hard* even in classical linear regression

$$y = \sum_{j=1}^{p} \beta_j x_j + \varepsilon$$

There are many ways to assess whether each β_j should be in the model (depends, e.g., on the order in which you consider the β_j's and whether you add or remove them from the model)

But still simple regression models at least feel less "black boxy" than complex neural nets

Suggestion: Additive models

$$y = \sum_{j=1}^{p} f_j(x_j) + \varepsilon,$$

where f_j's are splines, are often recommended as a good compromise between interpretability and model flexibility

$\rightarrow$ Perhaps it would make sense to fit an additive model to the output of a trained NN to interpret the NN?

Interpretability

Notice that in the absence of controlled / randomized experiments, any standard interpretability method merely reflects correlations between the inputs and outputs

But: correlation $\neq$ causation

Causal inference is a subfield of statistics attempting to establish rigorous causal relations between covariates (inputs) and responses (outputs)

- Could these be the ultimate tools for interpretability?
- Perhaps a topic for a future PhyStat seminar?

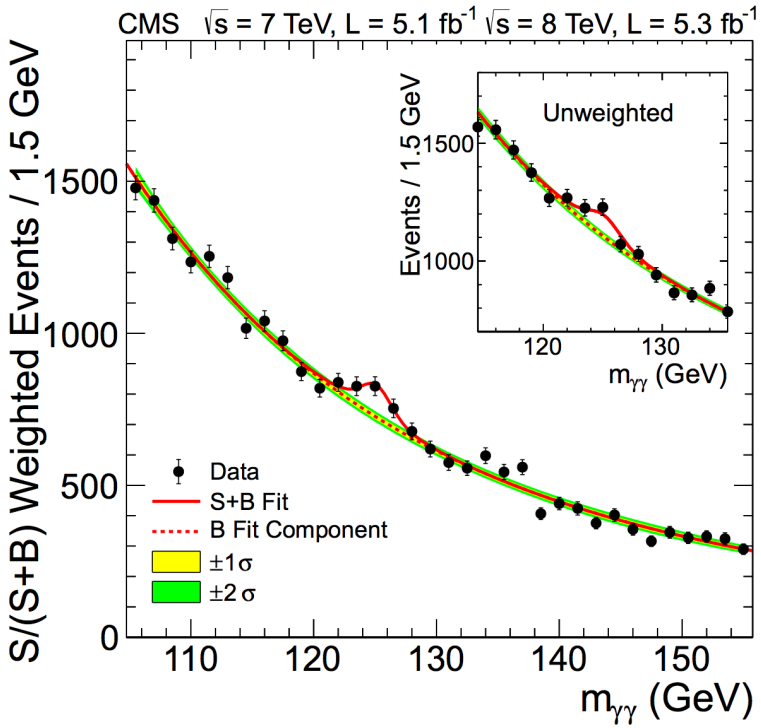
Hypothesis testing for discovery of new physics

Searches of new phenomena at the LHC usually boil down to testing for the presence of a signal distribution over a background of known Standard Model physics:

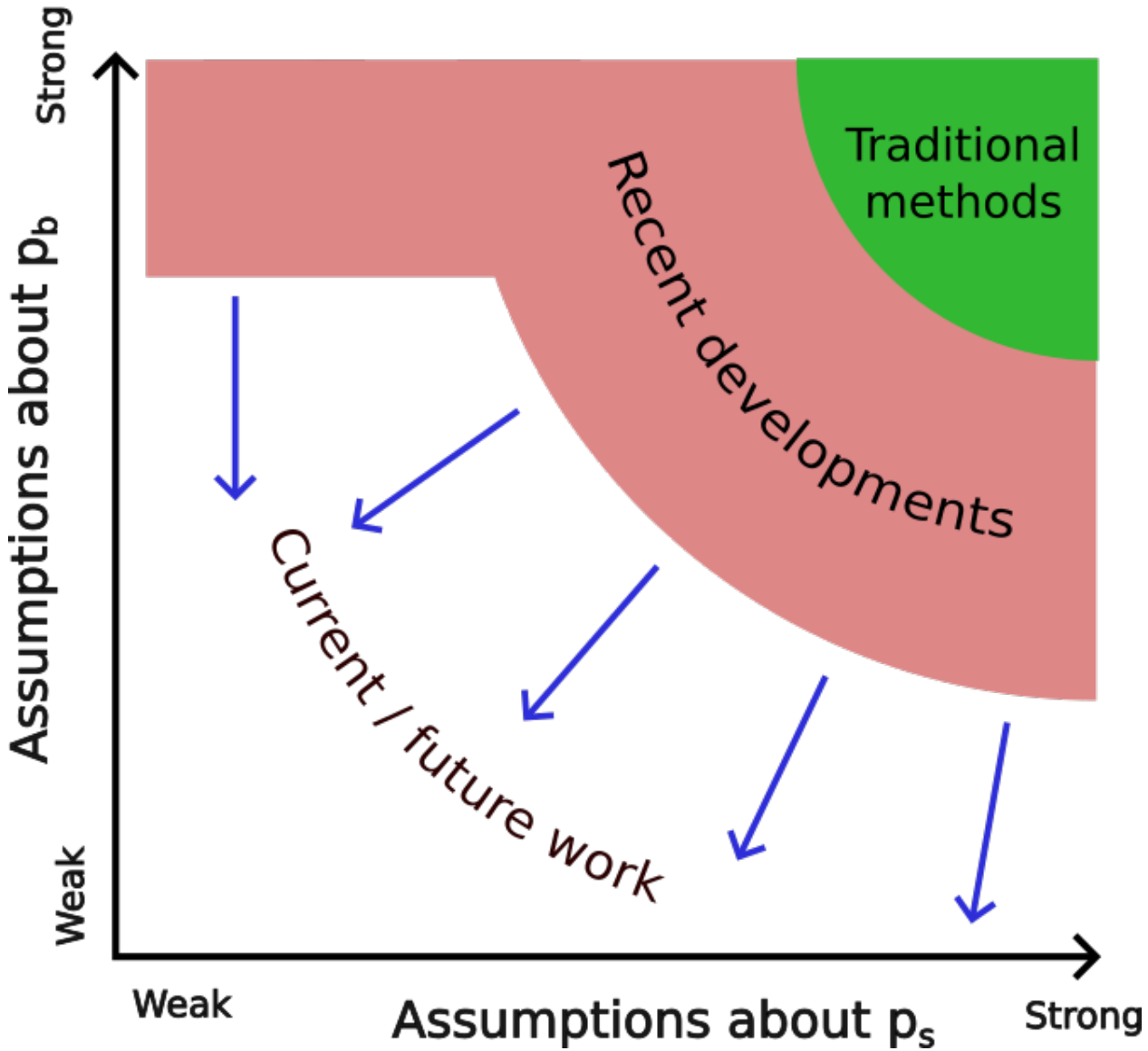
- Known physics: $p_b(z)$
- New signal: $p_s(z)$
- Nature: $q(z) = (1 - \lambda)p_b(z) + \lambda p_s(z)$

Want to test $H_0 : \lambda = 0$ vs. $H_1 : \lambda > 0$

If one rejects H_0 at a high enough significance level, then one might proceed to claim discovery of new physics



Landscape of model-agnostic methods



Related problems in statistics and ML

The model-agnostic search problem is closely related to a number of problems studied in statistics and machine learning

Specifically, in many cases it can be seen as an example of:

1. Two-sample testing (e.g., Kim et al. (2019, 2021)):
   $X_i \overset{\text{iid}}{\sim} p_1$, $Y_i \overset{\text{iid}}{\sim} p_2$, is $p_1 = p_2$?
2. Collective anomaly detection (e.g., Chandola et al. (2009)):
   Is there a *collection* of data points which taken together deviate from the anticipated data?

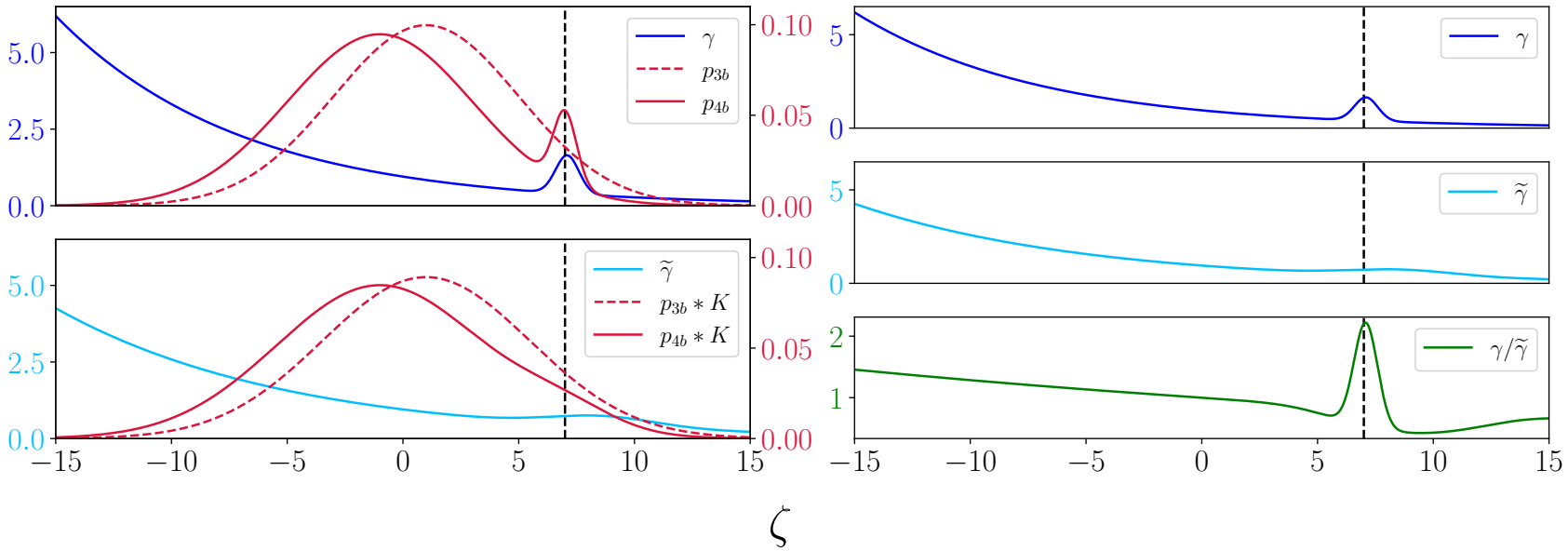
Notice that

model-agnostic search $\neq$ outlier detection

Each signal event is typically indistinguishable from the background on its own; it is the *collection* of many signal events that defines the excess

How to be model-agnostic for both signal and background?

New idea in arXiv:2409.06960: apply an event-level low-pass filter (smearing) and neural density ratios to find data-driven signal regions in a model-agnostic way in a high-dimensional feature space



To define the signal region, look for large values in $\gamma/\widetilde{\gamma}$, where

$$\gamma(\zeta) = \frac{p_{4b}(\zeta)}{p_{3b}(\zeta)}, \quad \widetilde{\gamma}(\zeta) = \frac{(p_{4b} * K)(\zeta)}{(p_{3b} * K)(\zeta)}$$

We learn γ by training Z_{3b} vs. Z_{4b} and $\widetilde{\gamma}$ by training $Z_{3b} + \mathcal{E}$ vs. $Z_{4b} + \mathcal{E}$

Unfolding with machine learning

A major development in the past few years: using machine learning to help solve the unfolding problem

Two main approaches:

1. OmniFold (Andreassen et al., 2020): iteratively reweight particle-level MC events using classifier-based density ratios
2. Generative unfolding (Bellagente et al., 2020; Backes et al., 2024): Train a generative model to sample from $p(X = t|Y = s)$; iterate to reduce dependence on $p^{\mathrm{MC}}(X = t)$

Benefits of ML-based unfolding:

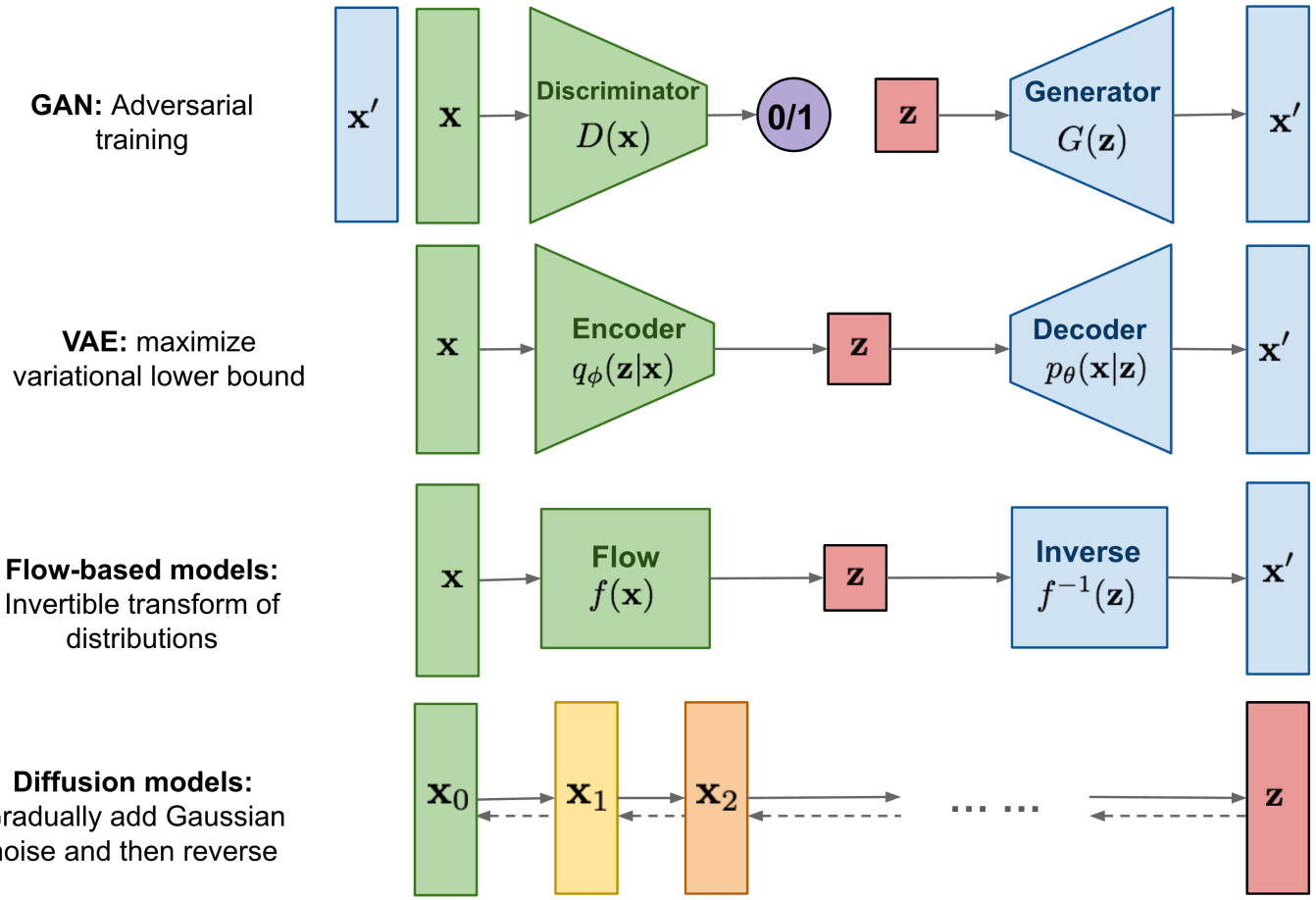
- Does not rely on binning
- Provides event-level unfolded results
- Can handle (moderately) high-dimensional phase spaces
- Does not need a separate estimate of the response kernel $k(s, t) = p(Y = s|X = t)$

Open question: What kind of regularization do these methods impose on the unfolded solution?

Conjecture: The regularization is implicitly in the structure and training of the neural networks

Generative models

The basic idea in all ML generative models is to train a neural model to map from a latent variable $\boldsymbol{z}$ generated from some tractable distribution (usually $N(0, \boldsymbol{I})$) to the data distribution



(Source: `https://lilianweng.github.io/posts/2021-07-11-diffusion-models/`)

Generative models

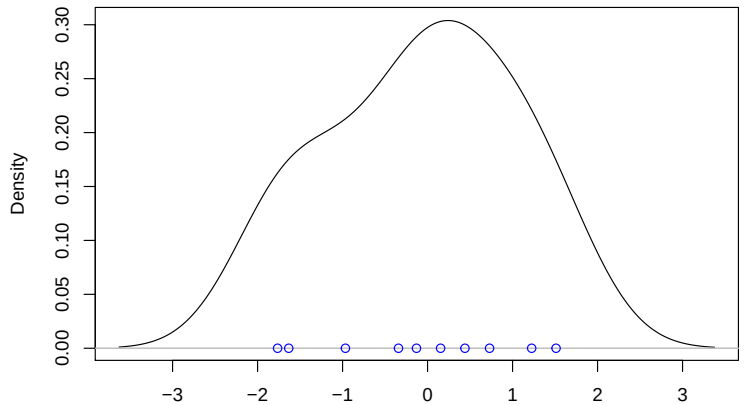
My take on the active (and converging?) discussion around generative models:

Q: If I train a generative model with n data points and use it to sample m data points with $m > n$, have I gained anything? ("GANplification")

A: A cautious "yes". The information content is still that of n data points but the generative model has a built-in inductive bias. To the extent that the inductive bias is a close enough reflection of reality, the larger sample can be more useful than the smaller sample.

Q: How to quantify the uncertainty of these generative models?

A: Perhaps best viewed as a combination of aleatoric and epistemic uncertainty? BNNs suggested as a potential way to account for both.



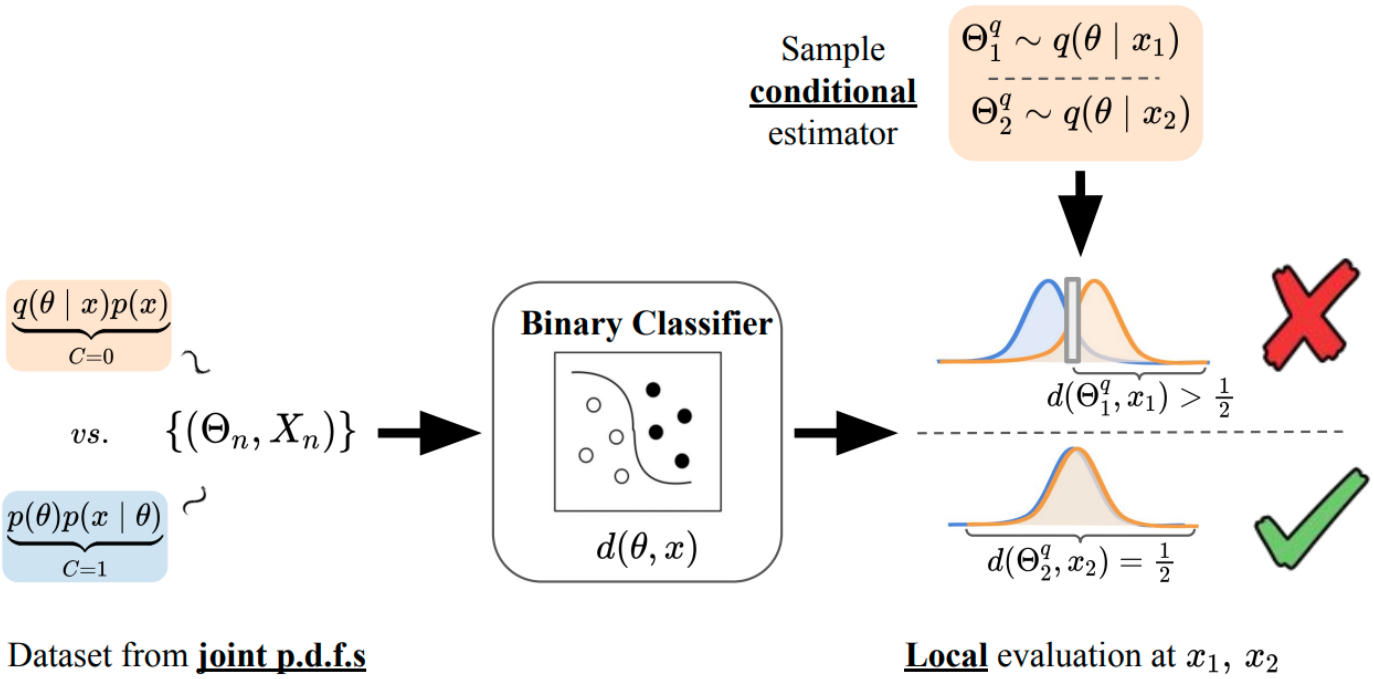
Kernel density estimator: not a neural generative model, but conceptually useful

Generative models

Q: How to validate a generative model?

A: Train a classifier to separate the generated data from the training data. Just keep in mind that this is a high-dimensional two-sample test and there does not exist a test that has high power for every possible departure from the null.

For validation of conditional generative models, see ▸ talk by J. Linhart at PhyStat-SBI:



Simulation-based inference

Simulation-based inference (SBI) makes inferences about a parameter of interest $\boldsymbol{\theta}$ given data $\boldsymbol{x}$ from a parametric model

$$\boldsymbol{x} \sim F_{\boldsymbol{\theta}}$$

when $F_{\boldsymbol{\theta}}$ is only available as a simulator

Ingredients:

- Sample of parameters: $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \sim p(\boldsymbol{\theta})$
- Corresponding simulations from the model: $\boldsymbol{x}_i \sim F_{\boldsymbol{\theta}_i},\ i = 1, \ldots, n$
- Observed data: $\boldsymbol{x}_{\text{obs}}$

Task: Infer $\boldsymbol{\theta}$ that generated $\boldsymbol{x}_{\text{obs}}$ (i.e., produce point estimates, confidence sets, credible sets, posteriors, hypothesis tests, etc.)

Key insight: Machine learning enables us to do this with very high-dimensional $\boldsymbol{x}$

Which test statistic to use?

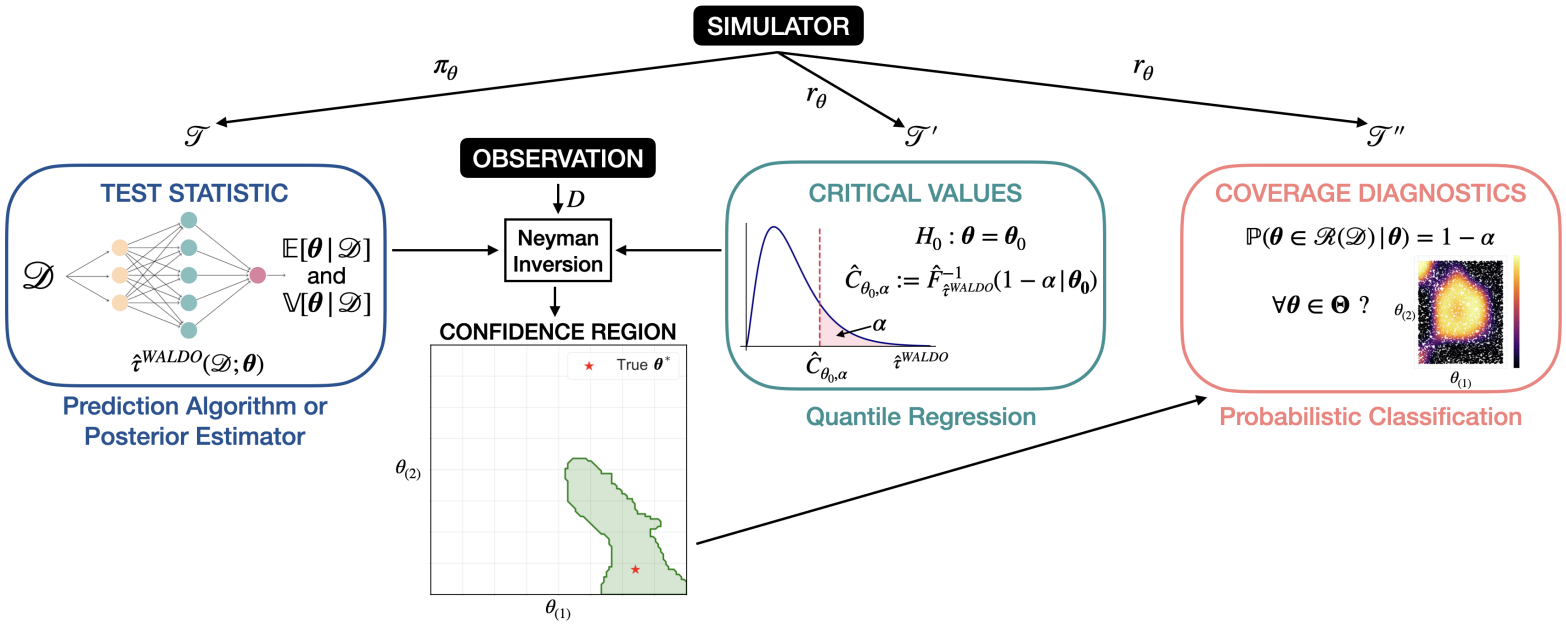
"Likelihood ratio is not guaranteed to be optimal for simple vs. composite tests. Are there benefits from considering some other test statistic?"

WALDO (Masserano et al., 2023) uses the following test statistic:

$$\tau^{\text{WALDO}}(\mathcal{D}; \boldsymbol{\theta}_0) = (\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}] - \boldsymbol{\theta}_0)^T \mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]^{-1} (\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}] - \boldsymbol{\theta}_0)$$

Benefits:

- Many SBI methods already produce $\mathbb{E}[\boldsymbol{\theta}|\mathcal{D}]$ and $\mathbb{V}[\boldsymbol{\theta}|\mathcal{D}]$
- Can take advantage of prior information without losing frequentist validity



SBI for spatial statistics

"Is there a more general notion of SBI?" $\rightarrow$ **Yes!**

In recent years, there has been an explosion of interest in SBI / neural inference / amortized inference for purely statistical models especially in spatial statistics:

- Neural prediction for spatial models (Gerber and Nychka, 2021; Lenzi et al., 2023; Sainsbury-Dale et al., 2024)
- Neural likelihood for spatial models (Walchessen et al., 2024)
- Neural prediction with censored observations (Richards et al., 2023)
- Neural prediction with irregularly spaced observations (Sainsbury-Dale et al., 2023)

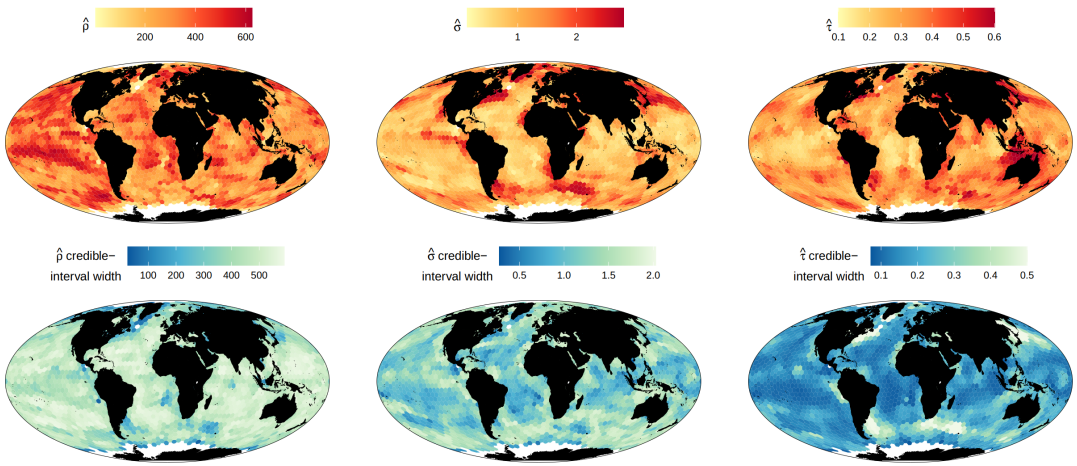


Figure: Neural prediction for satellite sea surface temperature using locally stationary Gaussian processes (Sainsbury-Dale et al., 2023)

References I

A. Andreassen, P. T. Komiske, E. M. Metodiev, B. Nachman, and J. Thaler. OmniFold: A method to simultaneously unfold all observables. *Physical Review Letters*, 124: 182001, 2020. doi: 10.1103/PhysRevLett.124.182001.

M. Backes, A. Butter, M. Dunford, and B. Malaescu. An unfolding method based on conditional invertible neural networks (cINN) using iterative training. *SciPost Phys. Core*, 7:007, 2024.

M. Bellagente, A. Butter, G. Kasieczka, T. Plehn, A. Rousselot, R. Winterhalder, L. Ardizzone, and U. Köthe. Invertible networks or partons to detector and back again. *SciPost Phys.*, 9:074, 2020. doi: 10.21468/SciPostPhys.9.5.074.

V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41:15:1–15:58, 2009.

F. Gerber and D. Nychka. Fast Covariance Parameter Estimation of Spatial Gaussian Process Models using Neural Networks. *Stat*, 10(1):e382, 2021. doi: https://doi.org/10.1002/sta4.382.

I. Kim, A. B. Lee, J. Lei, et al. Global and local two-sample tests via regression. *Electronic Journal of Statistics*, 13(2):5253–5305, 2019.

I. Kim, A. Ramdas, A. Singh, and L. Wasserman. Classification accuracy as a proxy for two-sample testing. *The Annals of Statistics*, 49(1):411 – 434, 2021.

References II

M. Kuusela and M. L. Stein. Locally stationary spatio-temporal interpolation of Argo profiling float data. *Proceedings of the Royal Society A*, 474:20180400, 2018.

A. Lenzi, J. Bessac, J. Rudi, and M. L. Stein. Neural Networks for Parameter Estimation in Intractable Models. *Computational Statistics & Data Analysis*, 185: 107762, 2023. doi: https://doi.org/10.1016/j.csda.2023.107762.

L. Masserano, T. Dorigo, R. Izbicki, M. Kuusela, and A. Lee. Simulator-based inference with WALDO: Confidence regions by leveraging prediction algorithms and posterior estimators for inverse problems. In F. Ruiz, J. Dy, and J.-W. van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 2960–2974. PMLR, 25–27 Apr 2023.

J. Richards, M. Sainsbury-Dale, A. Zammit-Mangion, and R. Huser. Neural Bayes estimators for censored inference with peaks-over-threshold models. Preprint arXiv:2306.15642 [stat.ME], 2023.

M. Sainsbury-Dale, J. Richards, A. Zammit-Mangion, and R. Huser. Neural Bayes estimators for irregular spatial data using graph neural networks. Preprint arXiv:2310.02600 [stat.ME], 2023.

References III

M. Sainsbury-Dale, A. Zammit-Mangion, and R. Huser. Likelihood-free parameter estimation with neural Bayes estimators. *The American Statistician*, 78(1):1–14, 2024. doi: 10.1080/00031305.2023.2249522.

J. Walchessen, A. Lenzi, and M. Kuusela. Neural likelihood surfaces for spatial processes with computationally intensive or intractable likelihoods. *Spatial Statistics*, 62: 100848, 2024.

Backup

SBI for purely statistical models

Purely statistical models ("spherical cows") vs. mechanistic simulators:

Domain	Purely statistical model	Mechanistic simulator
Oceanography	Gaussian process regression of irregularly sampled observations	Data assimilation with general circulation models
Epidemiology	ARIMA time series models	Compartmental models
Finance	Stochastic volatility models	??
Particle physics	??	MC generators

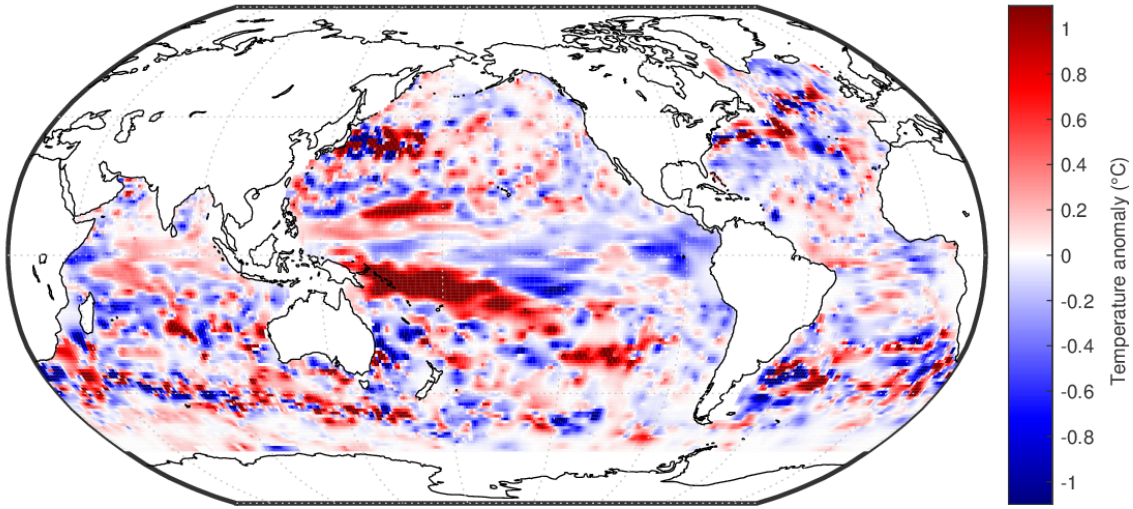


Figure: Spatio-temporal interpolation of subsurface ocean temperature anomalies using moving window-based locally stationary Gaussian processes (Kuusela and Stein, 2018)

The statistical fundamentals have not changed

While SBI has enabled inference in many previously intractable settings, it is important to keep in mind that it cannot circumvent fundamental limitations of statistics:

- Cramer–Rao lower bound: $\mathrm{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$, where $I(\theta)$ is the Fisher information
- Uniformly most powerful tests do not exist in general
- Sufficient statistics only exist in exponential families
- Goodness-of-fit tests place power on specific alternatives
- ...

Two notions of coverage

When validating SBI techniques (or inferential procedures more generally), a common desideratum is the coverage of an interval estimator $[\underline{\theta}(\boldsymbol{x}), \overline{\theta}(\boldsymbol{x})]$ of θ (for simplicity, let's take θ to be scalar here)

It's good to keep in mind that there are two different notions of coverage that often get mixed up:

Marginal coverage: $\mathbb{P}_{\boldsymbol{x},\theta}(\theta \in [\underline{\theta}(\boldsymbol{x}), \overline{\theta}(\boldsymbol{x})]) = 1 - \alpha$, where both θ and $\boldsymbol{x}$ are random inside the probability statement

Conditional coverage: $\mathbb{P}_{\boldsymbol{x}|\theta}(\theta \in [\underline{\theta}(\boldsymbol{x}), \overline{\theta}(\boldsymbol{x})]) = 1 - \alpha$, for all θ, where $\boldsymbol{x}$ is random but θ is fixed inside the probability statement

Even though these look similar, these are fundamentally different notions of what we mean by "uncertainty"

- Marginal coverage is easier to achieve but is a weaker notion (in fact, conditional coverage implies marginal coverage but the reverse is not true)
- Marginal coverage only makes sense if it is sensible to think of θ as being random