Expressive density models for high energy physics using a custom latent space

Stephen Menary
Department of Physics & Astronomy,
University of Manchester, UK
stephen.menary@manchester.ac.uk

Darren D. Price
Department of Physics & Astronomy,
University of Manchester, UK
darren.price@manchester.ac.uk

Abstract

Testing the compatibility between experimental observations and hypotheses about nature is the heart of the scientific method. This often requires us to model the probability density of possible observations, which is a challenging analytic task when data are multi-dimensional. We show that density models describing multiple observables with (i) hard boundaries and (ii) dependence on external parameters may be created using an auto-regressive Gaussian mixture model. The model is made more expressive using a novel method in which data are projected onto a custom latent space. The method is demonstrated on particle physics data sensitive to anomalies in the electroweak production of a Z boson in association with a dijet system. Such an approach may be applied to similar problems within other fields.

1 Introduction

In high-energy particle physics, we often wish to constrain a set of possible physical models according to their consistency with experimental observations $x \in \mathbb{X}$. To do this we must model the probability density $p(x|\theta)$, where $\theta \in \Theta$ represent parameters of interest or hidden variables such as experimental uncertainties. Analyses may use event selection criteria to enhance the relative contributions of the physical processes under study, then measure high-level observables such as kinematic spectra, for which the likelihood can be approximated by analysing simulations. A challenge of this approach is to ensure that model separability is captured within these low-dimensional summary statistics.

It has recently been demonstrated that machine-learned density models may be constructed which describe such densities (or density ratios) in a high-dimensional observable space $\mathbb{X}$ without the need for binning [1–5]. Provided that model bias can be mitigated and systematic uncertainties properly described, we can then compute $p(x|\theta)$ without the loss of sensitivity caused by discarding data, binning, use of summary statistics or sub-optimal analysis design. Furthermore, it is often possible to sample from density models, providing a compelling alternative to other stochastic generative models such as generative adversarial networks (GANs) [6] and variational auto-encoders (VAEs) [7, 8] for efficiently performing steps in a simulation chain [9, 10].

In this work, we construct a method in which datapoints are projected by function $f : x \mapsto u \in \mathbb{U}$ onto a latent space $\mathbb{U}$ which has the same dimensionality as $\mathbb{X}$. An auto-regressive Gaussian mixture model is used to describe the density $p_\phi(u|\theta)$, where ϕ label the parameters of several neural networks. The use of a latent space has two aims:

1. It guarantees that hard boundaries in physical observables are respected, provided that these boundaries are independent of all other observables.
2. The latent distribution is designed to be well described by a number of overlapping Gaussian modes, each describing a local cluster of density. Variations of θ which induce deformations in observable spectra can then be described by modifying individual Gaussian modes.

Third Workshop on Machine Learning and the Physical Sciences (NeurIPS 2020), Vancouver, Canada.

2 Experimental setup

We consider the electroweak production of a Z boson in association with a dijet system at the Large Hadron Collider (LHC). The kinematic spectra of such events were recently measured by the ATLAS experiment [11]. Exclusion limits were derived for several parameters of the Standard Model (SM) effective field theory (SMEFT) in the Warsaw basis [12, 13], which characterise the presence of any novel physics phenomena in such interactions. We consider the conditional dependence of four kinematic observables capturing distinctive characteristics of these particle interactions, $\mathbb{X} = \{m_{jj},\ m_{ll},\ \Delta\phi\,(j,j)\,,\ \Delta y\,(j,j)\}$, on the parameter $\Theta = \{\tilde{c}_W\}$. Ground truth events are generated using the `Madgraph5` (`MG5`) program at leading order in α_S [14], showered using `Pythia8` [15, 16] and reconstructed using `Rivet` [17]. Neural networks are implemented using `TensorFlow` v1.13.1 interfaced with `Keras` v2.2.4-tf[18, 19]. When training, 50% of the data are used for validation and the early stopping method is used to mitigate overtraining. All data are used when plotting. 400k datapoints are generated in increments of 0.1 on the interval $\tilde{c}_W \in [-0.4, 0.4]$, except at the SM value of $\tilde{c}_W = 0$ for which 1M datapoints are generated. Values of $\tilde{c}_W \in \{-0.3, -0.1, 0.1, 0.3\}$ are not used to train the density model but to test its inductive bias.

3 Overview of method using one dimensional example

We first consider a one-dimensional case where $x = \Delta\phi\,(j,j)$ and is restricted to the domain $x \in [-\pi,\ \pi]$. The distribution $p\,(x|\tilde{c}_W = 0)$ is shown in Figure 1 (top left). To project onto the latent space, we construct a response curve between the physical boundaries of x. This is written as $Q_x\,(x) = f \cdot D_x\,(x) + (1-f) \cdot L_x\,(x)$ where $D_x\,(x)$ is the cumulative distribution function of the training data at $\tilde{c}_W = 0$ and $L_x\,(x)$ is a linear function. The hyperparameter f is tuned to ensure that wide regions in x are not collapsed onto narrow regions in u. We then construct a response curve $Q_u\,(u)$ over the latent space, defined as the cumulative distribution function of a target distribution $\tilde{q}_u\,(u)$ given by

$$\tilde{q}_u\,(u) = \frac{1}{1+\exp[\alpha(u-\beta)-\gamma]} \cdot \frac{1}{1+\exp[-\alpha(u+\beta)-\gamma]} \tag{1}$$

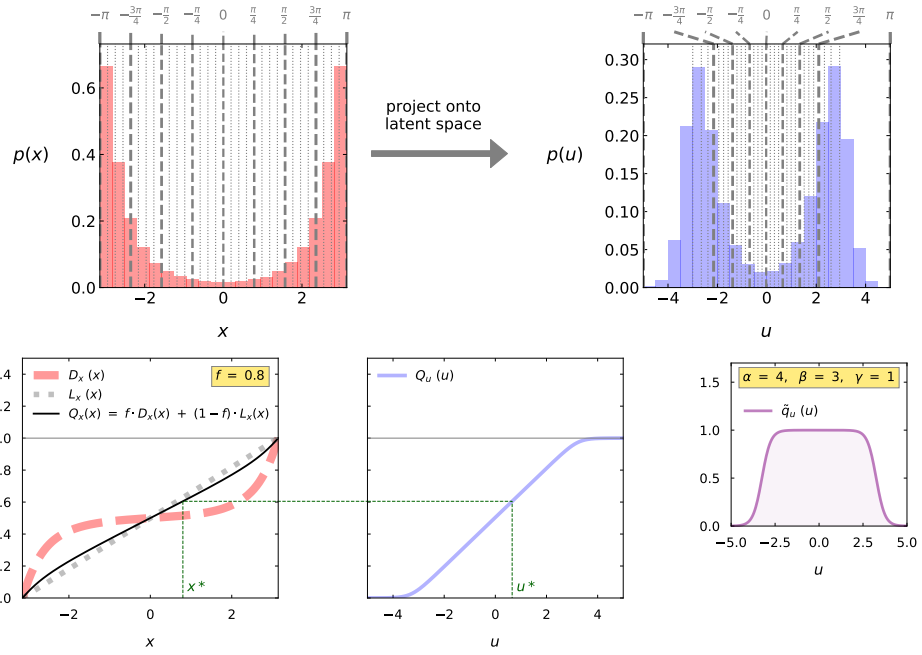

Figure 1: Top left: $p\,(x|\tilde{c}_W = 0)$ with $x = \Delta\phi\,(j,j)$, evaluated using `MG5` events. Top right: distribution over the latent space. Bottom left: response curve over the data space, $Q_x\,(x)$. Bottom middle: response curve over the latent space, $Q_u\,(u)$. Bottom right: target distribution, $\tilde{q}_u\,(u)$.

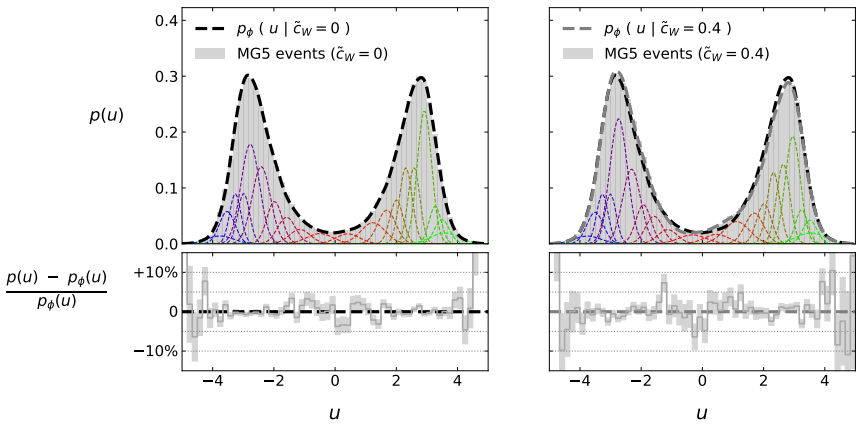

Figure 2: Gaussian mixture model over the latent space for the one-dimensional example of $x = \Delta\phi\,(j,j)$. We show the comparison with MG5 events when $\tilde{c}_W = 0$ (left) and $\tilde{c}_W = 0.4$ (right).

The mapping function is defined as $f\,(x) = Q_u^{-1}\,(Q_x\,(x))$, and its derivation is shown visually in Figure 1 (bottom). Parameter values of $f = 0.8$, $\alpha = 4$, $\beta = 3$ and $\gamma = 1$ are chosen. Figure 1 (top right) shows the distribution of $p\,(u|\tilde{c}_W = 0)$ using MG5 events. We compute $Q_u\,(u)$ as a piecewise-linear function over the interval $u \in [-5,\ 5]$, however this can be extended arbitrarily far in u so that all sampled points $u^* \in \mathbb{U}$ are mapped onto the physically allowed domain of $\mathbb{X}$. The functions f derived using $\tilde{c}_W = 0$ are applied to all values of $\tilde{c}_W$, so that variations in observable spectra become parameterisable deformations of $p\,(u|\tilde{c}_W)$.

The function $\tilde{q}_u$ is designed to have smooth tails and be significantly flatter than a Gaussian distribution. This encourages the resulting $p\,(u|\tilde{c}_W)$ to be well described by a mixture of many narrow Gaussian modes, instead of being dominated by any single mode. This is demonstrated in Figure 2, which shows the trained distributions for two values of $\tilde{c}_W$ in the $\Delta\phi\,(j,j)$ example, using $N_G = 20$. Systematic mismodelling is below 5% except in the sparsely populated tails of the distribution. A key observation is that both positive and negative deformations to the spectrum can be modelled as modifications to the amplitudes, means and widths of the Gaussian modes local to the deformation. This provides a mechanism for expressing the conditional dependence on external parameters, as well as the co-dependence between observables, as follows.

When modelling d observables, we construct an auto-regressive probability density

$$p_\phi\,(u|\tilde{c}_W) = \prod_{i=1}^{d} p_{\phi,i}\,(u_i|u_{<i};\ \tilde{c}_W) \tag{2}$$

where i label observables and $u_{<i}$ is the list of all prior latent observables. Note that the model is dependent on the observable ordering. Each conditional distribution is modelled as a sum of N_G Gaussian distributions $\mathcal{N}$ according to

$$p_{\phi,i}\,(u_i|u_{<i},\tilde{c}_W) = \sum_{g=1}^{N_G} f_{\phi,g,i}\,(u_{<i},\tilde{c}_W)\cdot\mathcal{N}\,(u_i;\ \mu_{\phi,g,i}\,(u_{<i},\tilde{c}_W)\,;\ \log\sigma_{\phi,g,i}\,(u_{<i},\tilde{c}_W)) \tag{3}$$

where $f_{\phi,g,i}$, $\mu_{\phi,g,i}$ and $\log\sigma_{\phi,g,i}$ are respectively the amplitude, mean and log-width of the g^{th} Gaussian subject to $\sum_{g=1}^{N_G} f_{\phi,g,i} = 1\ \forall\ i$. These are modelled using neural networks which capture the dependence on prior observables and external parameters.

4 Extending to four dimensions

To extend the model to all four observables, hyperparameter values of $f = 0.2,\ 0.8,\ 0.2$ and 0.8 are chosen for m_{jj}, m_{ll}, $\Delta y\,(j,j)$ and $\Delta\phi\,(j,j)$ respectively. We sample the trained density model 500k times by randomly drawing $u_0^* \sim p_{\phi,0}\,(u_0|\tilde{c}_W)$, $u_1^* \sim p_{\phi,1}\,(u_1|u_0^*;\tilde{c}_W)$ and so on until datapoints u in four dimensions are constructed. These are transformed back onto data space using $x^* = f^{-1}\,(u^*)$.

Figure 3 compares the one-dimensional marginal distributions using the density model (red) and MG5 events (black) for $\tilde{c}_W = 0$. Some mismodelling is observed up to the level of 5 %, most notably in the $\Delta\phi\,(j,j)$ distribution. Greater mismodelling is observed in the tail of $\Delta y\,(j,j)$. Figure 4 compares the two-dimensional marginal distributions for $\tilde{c}_W = 0$. This demonstrates that the model has captured the nontrivial high-dimensional correlations between the observables.

Figure 5 shows how the one-dimensional marginal distributions evolve as $\tilde{c}_W$ is scanned between $[-0.4, 0.4]$. These are presented as a ratio with respect to the $\tilde{c}_W = 0$ case. Whilst the data do not show any clear dependence as a function of m_{jj}, m_{ll} or $\Delta y\,(j,j)$, the $\Delta\phi\,(j,j)$ distribution is seen to oscillate in a nontrivial way as a function of $\tilde{c}_W$. The density model has captured this dependence, demonstrating that it can be used to perform a hypothesis test on an unbinned dataset.

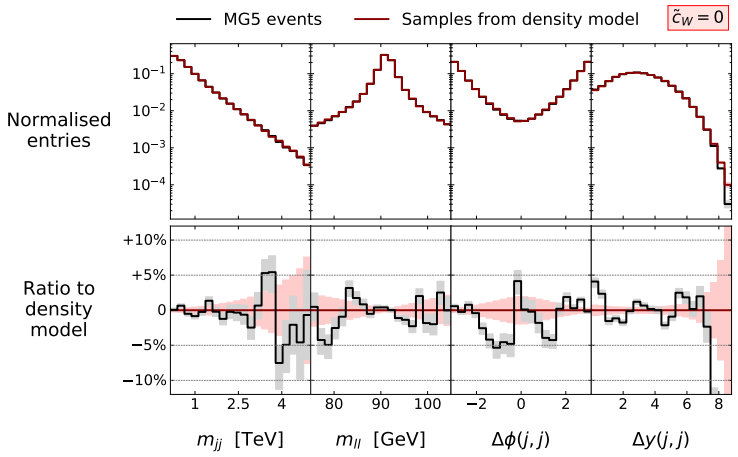

Figure 3: Marginal distributions of events sampled using the density model (red) compared with those generated using MG5 (black) for a value of $\tilde{c}_W = 0$. Shaded areas show sampling uncertainties.

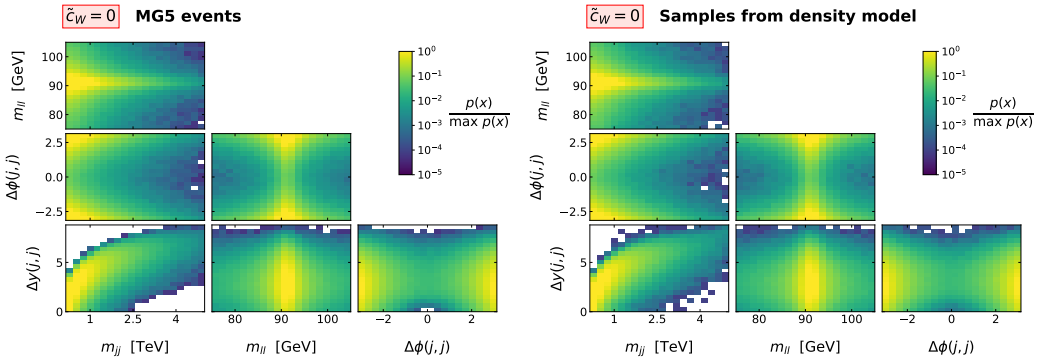

Figure 4: Two dimensional marginal distributions of events sampled using the density model (red) compared with those generated using MG5 (black) for a value of $\tilde{c}_W = 0$.

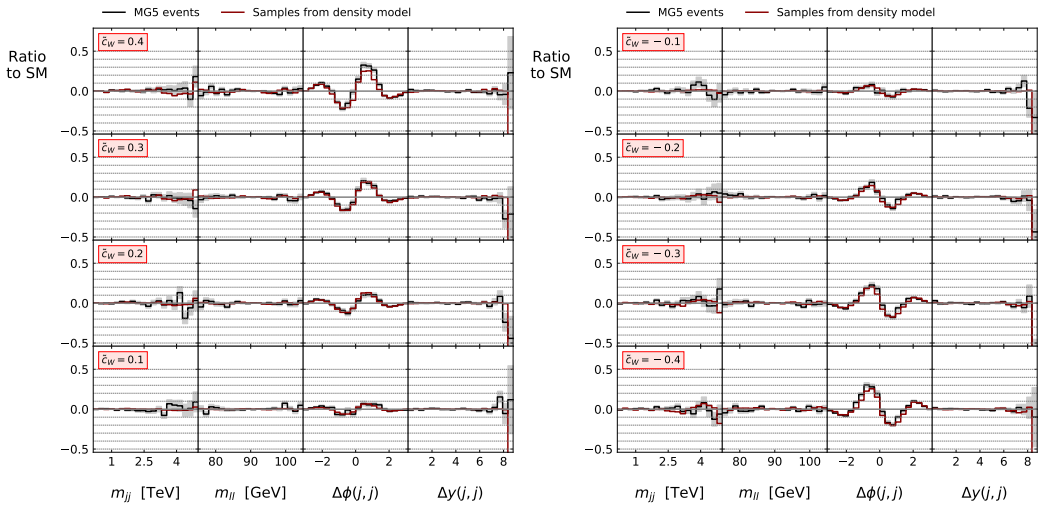

Figure 5: Demonstration of how the marginal distributions evolve with $\tilde{c}_W$, presented as a ratio with respect to the SM at $\tilde{c}_W = 0$. The dependence is well captured by the density model.

Broader Impact

The techniques described in this work are demonstrated for hypothesis testing within the field of high energy physics, including but not limited to parameter estimation, scientific discovery, and the setting of constraints on theoretical models. However, we anticipate that they may be applied to probabilistic modelling tasks in any domain for which high dimensional observable distributions are expected to be deformed by external parameter variations. We note that automisation of the method to general scenarios would require careful validation beyond the scope of what has been tested in this work.

The model is potentially biased by (i) any assumptions and methods used to simulate the experimentally-derived training data, (ii) the inductive bias imposed by the neural network architectures when interpolating between training points and (iii) systematic mis-modelling. As is common for traditional analyses within the field, such sources of bias must be carefully understood and captured within final exclusion limits, and any remaining model dependence made clear. Whilst the modelling and profiling of systematic nuisance parameters is possible in principle, we have not demonstrated this in the current work.

Applied to the physical sciences, the work benefits those who wish to set the most stringent limits on nature using their finite data. We hope that the continued development of trainable likelihood models, for which this work contributes, will enable model testing using multiple datasets, allowing more stringent limits to be set, or revealing inconsistencies or biases in the data not previously considered. Applied to the physical sciences, the impact of unmitigated model bias would be the drawing of false conclusions about nature. Applied within other fields, the impact of mis-modelling is defined by the domain being considered.

Acknowledgments and Disclosure of Funding

Darren Price is supported by a Turing Fellowship from the Alan Turing Institute, London, UK and by the University of Manchester. Stephen Menary is supported through a grant from the Alan Turing Institute.

References

[1] Johann Brehmer, Kyle Cranmer, Gilles Louppe, and Juan Pavez. A Guide to Constraining Effective Field Theories with Machine Learning. *Phys. Rev. D*, 98(5):052004, 2018.

[2] Johann Brehmer, Felix Kling, Irina Espejo, and Kyle Cranmer. MadMiner: Machine learning-based inference for particle physics. *Comput. Softw. Big Sci.*, 4(1):3, 2020.

[3] Johann Brehmer, Gilles Louppe, Juan Pavez, and Kyle Cranmer. Mining gold from implicit models to improve likelihood-free inference. *Proceedings of the National Academy of Sciences*, 117(10):5242–5249, 2020.

[4] Kyle Cranmer, Juan Pavez, and Gilles Louppe. Approximating Likelihood Ratios with Calibrated Discriminative Classifiers. `https://arxiv.org/abs/1506.02169`, 6 2015.

[5] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. `https://arxiv.org/abs/1705.07057`, 2018.

[6] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. `https://arxiv.org/abs/1406.2661`, 2014.

[7] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. `https://arxiv.org/abs/1312.6114`, 2014.

[8] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[9] Anja Butter and Tilman Plehn. Generative Networks for LHC events. `https://arxiv.org/abs/2008.08558`, 8 2020.

[10] Aishik Ghosh, for the ATLAS Collaboration. Deep generative models for fast shower simulation in ATLAS. Technical Report ATL-SOFT-PROC-2019-007, CERN, Geneva, Jun 2019.

[11] ATLAS Collaboration. Differential cross-section measurements for the electroweak production of dijets in association with a Z boson in proton-proton collisions at ATLAS. `https://arxiv.org/abs/2006.15458`, 2020.

[12] Ilaria Brivio, Yun Jiang, and Michael Trott. The SMEFTsim package, theory and tools. *Journal of High Energy Physics*, 2017(12), Dec 2017.

[13] B. Grzadkowski, M. Iskrzyński, M. Misiak, and J. Rosiek. Dimension-six terms in the standard model lagrangian. *Journal of High Energy Physics*, 2010(10), Oct 2010.

[14] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *Journal of High Energy Physics*, 07:079, 2014.

[15] Torbjörn Sjöstrand, Stefan Ask, Jesper R. Christiansen, Richard Corke, Nishita Desai, Philip Ilten, Stephen Mrenna, Stefan Prestel, Christine O. Rasmussen, and Peter Z. Skands. An introduction to PYTHIA 8.2. *Comput. Phys. Commun.*, 191:159–177, 2015.

[16] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. A Brief Introduction to PYTHIA 8.1. *Comput. Phys. Commun.*, 178:852–867, 2008.

[17] Christian Bierlich et al. Robust Independent Validation of Experiment and Theory: Rivet version 3. *SciPost Phys.*, 8:026, 2020.

[18] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[19] François Chollet et al. Keras. `https://keras.io`, 2015.