Anomaly Awareness Estimation in $b \to s\ell^+\ell^-$ with VAEs

B. Capdevila University of Barcelona, ICCUB

Institut de Ciències del Cosmos
UNIVERSITAT DE BARCELONA
EXCELENCIA MARÍA DE MAEZTU 2020-2023
"la Caixa" Foundation
European Commission

In collaboration with:

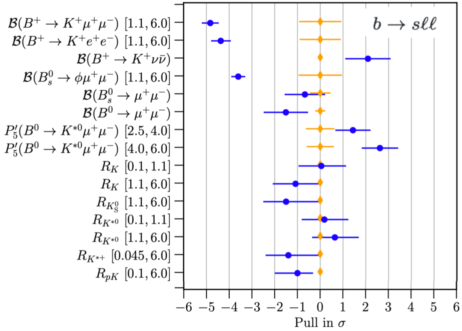
A. Scaffidi; arXiv:241x.xxxx

Anomaly Detection in Low-Energy Observables

- **Anomalies:** Deviations between **experimental data** and **theoretical predictions** under a **null hypothesis** (SM hypothesis)
- Searching for anomalies in **low-energy observables** is critical for **NP exploration**
  - $\Rightarrow$ Observables at **low energies** receive contributions from **higher scales** due to **quantum corrections**
- **Goal:** *Properly* quantify the **statistical significance** of observed anomalies
- Examples of recent anomalies
  - $\Rightarrow$ B anomalies ($b \to s\ell^+\ell^-$, $b \to c\ell\nu$)
  - $\Rightarrow$ $(g-2)_\mu$
  - $\Rightarrow$ V_{cb}, V_{ub} puzzle
  - $\Rightarrow$ ...

Anomaly Detection in $b \to s\ell^+\ell^-$

Data



Stats

p − value

Experimental Data in Phenomenological Analyses

- Experimental data:
  - ⇒ Released from the experiments as a **vector of means** $\mu^{\rm exp}$ and a **covariance matrix** $\Lambda^{\rm exp}$

⇒ **Implicitly assumes a Gaussian distribution** for the experimental measurements

$$p(\boldsymbol{x}^{\rm exp}) = \mathcal{N}(\boldsymbol{x}^{\rm exp}; \boldsymbol{\mu}^{\rm exp}, \boldsymbol{\Lambda}^{\rm exp})$$

	$4.0 < q^2 < 6.0\,\mathrm{GeV}^2/c^4$
P_1	$0.088 \pm 0.235 \pm 0.029$
P_2	$0.105 \pm 0.068 \pm 0.009$
P_3	$-0.090 \pm 0.139 \pm 0.006$
P_4'	$-0.312 \pm 0.115 \pm 0.013$
P_5'	$-0.439 \pm 0.111 \pm 0.036$
P_6'	$-0.293 \pm 0.117 \pm 0.004$
P_8'	$0.166 \pm 0.127 \pm 0.004$

	$F_{\rm L}$	P_1	P_2	P_3	P_4'	P_5'	P_6'	P_8'
$F_{\rm L}$	1.00	0.04	0.05	−0.10	−0.04	−0.14	−0.17	0.14
P_1		1.00	0.06	0.07	−0.06	−0.10	−0.03	0.02
P_2			1.00	−0.02	−0.14	−0.09	−0.03	−0.01
P_3				1.00	−0.01	0.07	0.19	−0.01
P_4'					1.00	0.02	0.04	0.01
P_5'						1.00	0.09	0.00
P_6'							1.00	0.02
P_8'								1.00

LHCb, arXiv:2003.04831

Theoretical Predictions in Phenomenological Analyses

- Theoretical predictions
  - ⇒ For the observables in the analysis $\boldsymbol{x} = (x_1, \ldots, x_n)$, we have functions representing their **theoretical predictions** in terms of several **input parameters** $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_m)$

$$x_i = x_i(\nu_1, \ldots, \nu_m)$$

  - ⇒ The **input parameters** are distributed according to some distribution, usually Gaussian

$$\boldsymbol{\nu} \sim \mathcal{N}(\boldsymbol{\nu}; \boldsymbol{\mu}_\nu, \boldsymbol{\Lambda}_\nu)$$

    where $\boldsymbol{\mu}_\nu$ and $\boldsymbol{\Lambda}_\nu$ means and covariance of the distribution of underlying parameters

- Implications
  - ⇒ Even if the distribution of parameters is Gaussian, observables with complex structures **do not distribute normally**
  - ⇒ Except for observables with a **linear dependence** on the underlying parameters
  - ⇒ This means the **likelihood** $p(\boldsymbol{x}|H_i)$ will generally be distributed under a **non-Gaussian distribution**

Gaussian Likelihoods in Phenomenological analyses

- In **conventional frequentist $b \to s\ell^+\ell^-$ analyses**, Gaussian likelihoods are assumed

$$p(\boldsymbol{x}|H_i) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{x}_{H_i}, \boldsymbol{\Lambda}_{H_i}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Lambda}_{H_i}|}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_{H_i})^T \boldsymbol{\Lambda}_{H_i}^{-1} (\boldsymbol{x} - \boldsymbol{x}_{H_i})\right)$$

$\Rightarrow$ $\boldsymbol{\Lambda}_{H_i}$ is the **sum of theoretical and experimental covariances**

$$\boldsymbol{\Lambda}_{H_i} = \boldsymbol{\Lambda}_{H_i}^{\text{th}} + \boldsymbol{\Lambda}^{\text{exp}}$$

$\Rightarrow$ Theory side: Covariance estimated from samples of theoretical predictions

$\Rightarrow$ Experimental side: Covariance read directly from data released by experiments

- Measuring goodness-of-fit

$\Rightarrow$ Use $-\log p(\boldsymbol{x}|H_i)$ as a **statistic** to measure **agreement between data and hypothesis** H_i

$$-\log p(\boldsymbol{x}|H_i) = \frac{\chi^2}{2} + \text{const}$$

where χ^2 is the chi-squared function

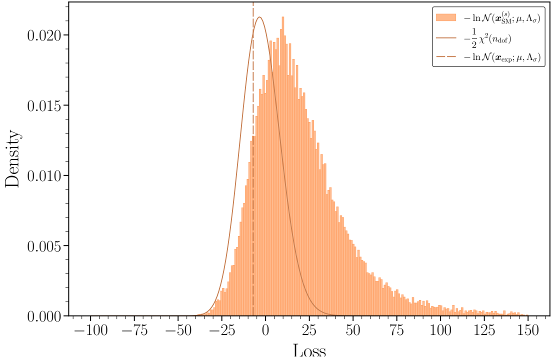
$\Rightarrow$ If theoretical predictions are normally distributed, $-\log p(\boldsymbol{x}|H_i)$ follows a χ^2-distribution with $n_{\text{dof}} = n_{\text{obs}}$ in the analysis

$\Rightarrow$ If **not normally distributed**, $-\log p(\boldsymbol{x}|H_i)$ is **only asymptotically** χ^2 due to the central limit theorem

$\Rightarrow$ Assuming χ^2-distribution when it is not **creates biases** in calculating p-values from $-\log p(\boldsymbol{x}_{\text{exp}}|H_i)$

Distribution of $-\log p(\boldsymbol{x}|H_i)$ vs Asymptotic χ^2

- Distribution of $-\log p(\boldsymbol{x}|H_i)$
  - ⇒ Calculated for the set of observables in the $b \to s\ell\ell$ dataset under the SM hypothesis ($H_0 = \mathrm{SM}$)
  - ⇒ Obtained by calculating $-\log p(\boldsymbol{x}|H_0)$ for each $\boldsymbol{x}$ in a sample $\boldsymbol{x}^s = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^{n_{\text{sample}}})$
  - ⇒ Sample size: $n_{\text{sample}} = 10000$
  - ⇒ Each $\boldsymbol{x}^s$ generated by sampling underlying parameters $\boldsymbol{\nu}$ from $\mathcal{N}(\boldsymbol{\nu}; \boldsymbol{\mu}_\nu, \boldsymbol{\Lambda}_\nu)$
- Comparison
  - ⇒ Distribution of $-\log p(\boldsymbol{x}|H_i)$ vs asymptotic χ^2 distribution with corresponding degrees of freedom
  - ⇒ Difference observed between the two distributions



Estimating Likelihoods

- Understanding $p(\boldsymbol{x}|H_i)$
  - ⇒ Goal: Determine the likelihood $p(\boldsymbol{x}|H_i) = p(\boldsymbol{x})$ for enhanced statistical rigor in hypothesis testing
  - ⇒ Known: Distribution of underlying inputs $p(\boldsymbol{\nu})$, which informs us about the prior probability of the model parameters
  - ⇒ Computable: $p(\boldsymbol{x}|\boldsymbol{\nu})$, achievable by simulating observables $\boldsymbol{x}$ using sampled parameters $\boldsymbol{\nu}$ from their known distributions

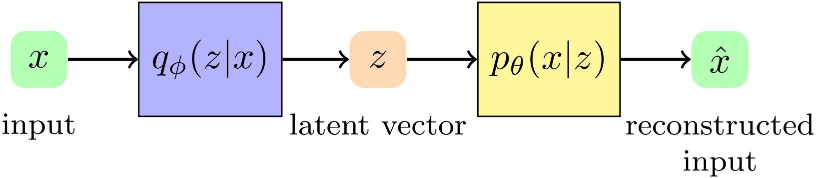
- Obtaining $p(\boldsymbol{x})$ as a marginal likelihood

$$p(\boldsymbol{x}) = \int d\boldsymbol{\nu}\, p(\boldsymbol{x}|\boldsymbol{\nu})p(\boldsymbol{\nu})$$

  - ⇒ In most real-life applications $\boldsymbol{\nu}$ is usually high-dimensional
  - ⇒ Challenge: a direct computation the $d\boldsymbol{\nu}$ integral is generally computationally prohibitive
  - ⇒ Hence, the likelihood $p(\boldsymbol{x})$ is typically **intractable**

- Estimation using Variational Autoencoders (VAEs)
  - ⇒ VAEs provide a feasible approach to approximate $p(\boldsymbol{x})$ with **arbitrary precision**

Introducing Variational Autoencoders



- VAE framework
  - ⇒ **Model**: Pairs a probabilistic decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ with a probabilistic encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x})$
  - ⇒ **Latent variable $\boldsymbol{z}$** approximates underlying parameters $\boldsymbol{\nu}$
  - ⇒ VAEs **do not** map inputs to a deterministic latent variable, but to a **probability space** $p(\boldsymbol{z})$
  - ⇒ θ: parameters of the decoder
  - ⇒ ϕ: parameters of the decoder

The Variational Lower Bound

- The Variational Lower Bound (ELBO) relates to two joint probability density functions: p_θ and q_ϕ

$$\mathcal{L}(\theta, \phi; \boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right]$$

  - ⇒ $p_\theta(\boldsymbol{x}, \boldsymbol{z})$: joint distribution of $\boldsymbol{x}$ and $\boldsymbol{z}$
  - ⇒ $q_\phi(\boldsymbol{z}|\boldsymbol{x})$: approximate encoder posterior
  - ⇒ Simplifies to

$$\mathcal{L}(\theta, \phi; \boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] - D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}))$$

- Includes Kullback-Leibler divergence (KL-div) which is a distance in distribution space

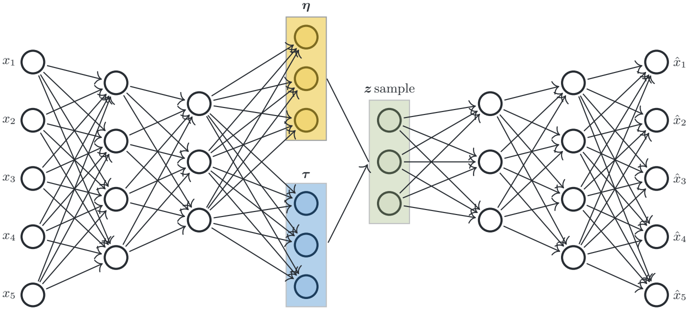
$$D_{KL}(q_\phi(w)||p_\theta(w)) = \mathbb{E}_{q_\phi(w)}\left[\log \frac{q_\phi(w)}{p_\theta(w)}\right]$$

- Implications and optimisation objective
  - ⇒ Log-likelihood relation

$$\log p_\theta(\boldsymbol{x}) \geq \mathcal{L}(\theta, \phi; \boldsymbol{x})$$

  - ⇒ Objective function to train VAEs
    - ⇒ Maximize ELBO to approximate the true log-likelihood.
    - ⇒ Equivalent to minimising the negative ELBO (-ELBO)

Deep Learning Implementation of Variational Autoencoders



- Implementation and parametrisation
  - ⇒ **Neural Networks as parametrisers**
  - ⇒ $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ and $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ parameterised using deep neural networks

$$\theta = \{W_{l_1}, \ldots, W_{l_L}, b_{l_1}, \ldots, b_{l_L}\}$$
$$\phi = \{V_{l_1}, \ldots, V_{l_L}, c_{l_1}, \ldots, c_{l_L}\}$$

  where W_l (V_l), b_l (c_l) are the weights and biases of the encoder (decoder) network
  - ⇒ This setup enables the modeling of complex, non-linear relationships between observed data and latent variables

Preparing Training Data with Theoretical and Experimental Inputs

- Generating Theoretical Predictions
  - ⇒ Start by sampling the distribution of underlying inputs under hypothesis H_0, $p(\boldsymbol{\nu}|H_0)$
  - ⇒ Compute the vector of observables $\boldsymbol{x}^s$ for these values to obtain a sample of theoretical predictions: $\boldsymbol{x}^s = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^{n_{\text{sample}}})$
- Incorporating Experimental Uncertainties
  - ⇒ Smear the samples with experimental uncertainties to simulate realistic observational data

$$\boldsymbol{x}'^s = \boldsymbol{x}^s + \boldsymbol{L}_{\boldsymbol{\Lambda}^{\text{exp}}} \boldsymbol{w}$$

  - ⇒ $\boldsymbol{L}_{\boldsymbol{\Lambda}^{\text{exp}}}$: Cholesky decomposition of the experimental covariance matrix $\boldsymbol{\Lambda}^{\text{exp}}$
  - ⇒ $\boldsymbol{w} \sim \mathcal{N}(w; \mathbf{0}, \mathbf{1})$: Normal noise vector simulating experimental noise

Training the Variational Autoencoder with Experimental Uncertainties

- Training the VAE
  - ⇒ Divide the smeared dataset into **training and testing datasets**
  - ⇒ Use the training dataset to **optimise the parameters** of the VAE, **minimising** the -ELBO

$$\mathcal{L}(\theta, \phi; \boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] - \beta D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}))$$

  - ⇒ Notice the parameter β in the ELBO. It allows us to test the **generative properties** of the model
  - ⇒ Objective: **approximate the full log-likelihood distribution** of the observables under H_0 as closely as possible

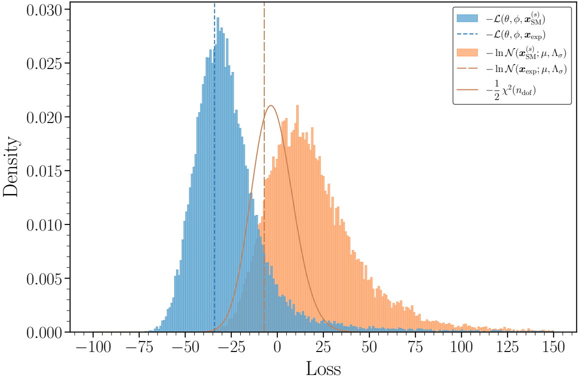
- Model Assumptions
  - ⇒ Assume distributions for **model simplicity**:

$$\begin{aligned}
p_\theta(\boldsymbol{z}) &= \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{1}) \\
p_\theta(\boldsymbol{x}|\boldsymbol{z}) &= \mathcal{N}(\boldsymbol{x}; \hat{\boldsymbol{x}}, \boldsymbol{\Lambda}_{\hat{\boldsymbol{x}}}) \text{ with } \boldsymbol{\Lambda}_{\hat{\boldsymbol{x}}} = \mathrm{diag}(\sigma^2_{\hat{\boldsymbol{x}}}) \\
q_\phi(\boldsymbol{z}|\boldsymbol{x}) &= \mathcal{N}(\boldsymbol{z}; \boldsymbol{\eta}, \boldsymbol{\Lambda}_{\boldsymbol{\tau}}) \text{ with } \boldsymbol{\Lambda}_{\boldsymbol{\tau}} = \mathrm{diag}(\tau^2)
\end{aligned}$$

  - ⇒ These **assumptions do not imply a Gaussian likelihood** but approximate relationships within the VAE structure

Statistical Analysis Using Trained VAE for $b \to s\ell^+\ell^-$

- Analysing the test dataset
  - $\Rightarrow$ Compute the -ELBO distribution using the test dataset, approximating the full -log-likelihood under hypothesis H_0.
  - $\Rightarrow$ Evaluate -ELBO for the experimental data to compute the p-value
- Preliminary results
  - $\Rightarrow$ Performed for the $b \to s\ell\ell$ dataset with promising preliminary outcomes

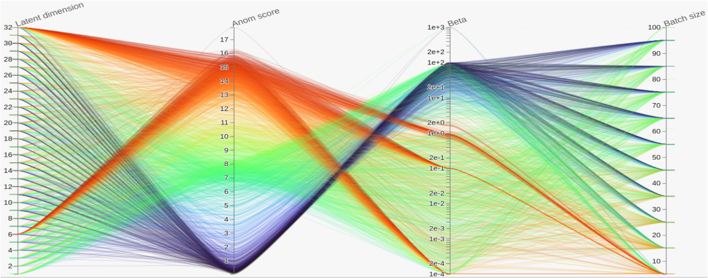


Refining Model Tuning and Validity in VAE Training

- Challenges in Model Tuning
  - ⇒ How do we determine the optimal dimensionality for the DNNs of the encoder and decoder, or the correct value of β?
  - ⇒ Could these choices bias the p-value?
  - ⇒ The choice of neural networks' architecture and β significantly affects the model's performance and the fidelity of the statistical results
- Testing and Validating Model Parameters
  - ⇒ Ongoing research and empirical testing are essential to optimise these parameters while minimising biases
- Strategies for Validation and Hyperparameter Optimisation
  - ⇒ Employ validation techniques to ensure model outputs are stable and reliable across various parameter configurations
  - ⇒ Use synthetic datasets to evaluate the impact of hyperparameter adjustments on model performance

Optimising Hyperparameters

- Generating artificially anomalous data
  - ⇒ Used to train the VAE across different configurations to optimise anomaly detection
- Tuning β in the ELBO
  - ⇒ Exploring the impact of β on anomaly detection and VAE's generative accuracy
- Pre-experimental blind analysis
  - ⇒ Ensures that the final measurement of experimental data is unbiased



Outlook and Continued Research

- Deepening understanding of VAE parameters
  - $\Rightarrow$ Exploring how different parameters influence the anomaly score and VAE's generative properties
- Ongoing hyperparameter optimisation
  - $\Rightarrow$ Continuously refining the model to enhance its predictive accuracy and anomaly detection
- Addressing sparse covariance matrices
  - $\Rightarrow$ Using random matrix theory techniques to sample observables and covariance matrices at the same time (LKJ distribution, Wishart distribution)
  - $\Rightarrow$ Will allow us to quantify the uncertainty attached to many unnatural zeros in the experimental covariance matrix
- Expanding application scope
  - $\Rightarrow$ Applying methodologies to SMEFT fits beyond $b \to s\ell\ell$

Thank You!

Backup Slides

Dual-Branch VAE Architecture

- $\Rightarrow$ We need the VAE to estimate $p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \hat{\boldsymbol{x}}, \boldsymbol{\Lambda}_{\hat{\boldsymbol{x}}})$
- $\Rightarrow$ We need two output branches: one for $\hat{\boldsymbol{x}}$ and one for $\boldsymbol{\Lambda}_{\hat{\boldsymbol{x}}}$