Less is more: *sparse* kernel methods with dictionary learning

Expressive, *regularized* and *interpretable* models for **statistical anomaly detection**

Gaia Grosso[1,2,3], Demba Ba[1,2], Phil Harris[1,3]

[1]NSF Institute for Artificial Intelligence and Fundamental Interaction (IAIFI)

[2]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA [3]MIT Laboratory for Nuclear Science, Cambridge, MA

HARVARD UNIVERSITY

Massachusetts Institute of Technology

EUROPEAN AI FOR FUNDAMENTAL PHYSICS CONFERENCE
EuCAIFCon 2024

Physics

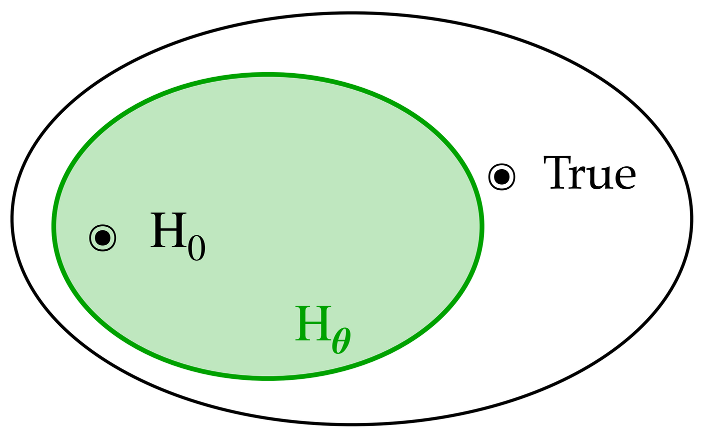
GOAL

Signal-agnostic statistical detection of new physical processes

Maximum-likelihood-ratio goodness of fit test [1-3]:

$$t(\mathcal{D}) = 2\max_{\boldsymbol{\theta}} \log \frac{\mathcal{L}(\mathcal{D}|\mathrm{H}_{\boldsymbol{\theta}})}{\mathcal{L}(\mathcal{D}|\mathrm{H}_0)} \qquad \boxed{n(x|\mathrm{H}_{\boldsymbol{\theta}}) = n(x|\mathrm{H}_0)\,\exp[f_{\boldsymbol{\theta}}(x)]}$$

$$= -2\min_{\boldsymbol{\theta}} L_{\mathrm{LR}}[f_{\boldsymbol{\theta}}]$$

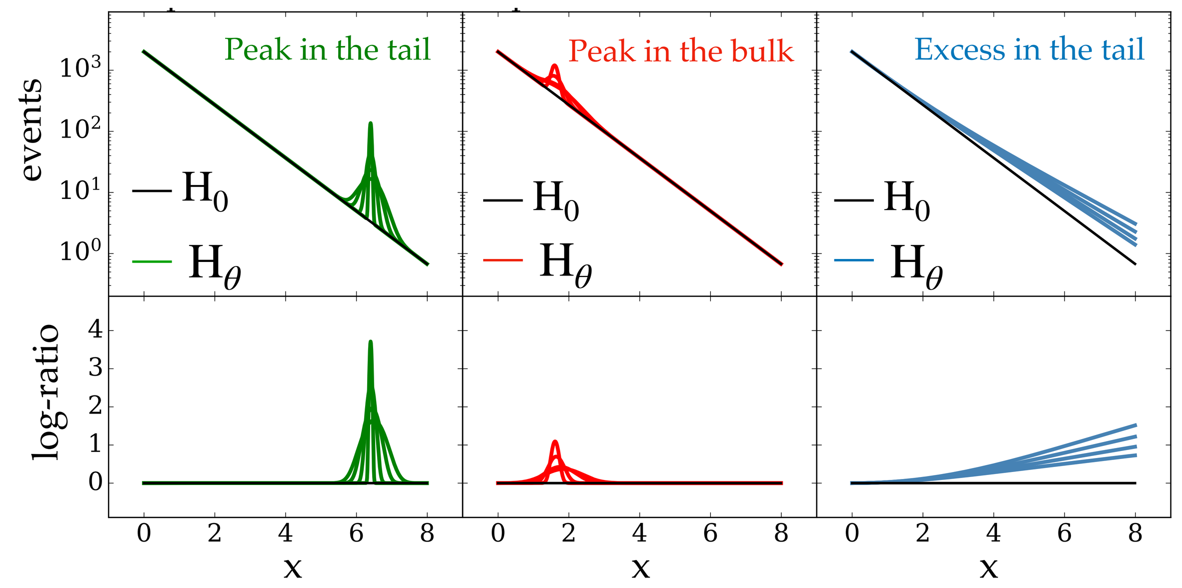


Loss function:

$$\boxed{L_{\mathrm{LR}}[f_{\boldsymbol{\theta}}] = \sum_{x\in\mathcal{R}} w_0(x)\,(\exp[f_{\boldsymbol{\theta}}(x)]-1) - \sum_{x\in\mathcal{D}} f_{\boldsymbol{\theta}}(x)}$$

PROBLEM

How to design $f_\theta(x)$ to capture *rare* and *unexpected* subtle perturbations on top of the known physics?



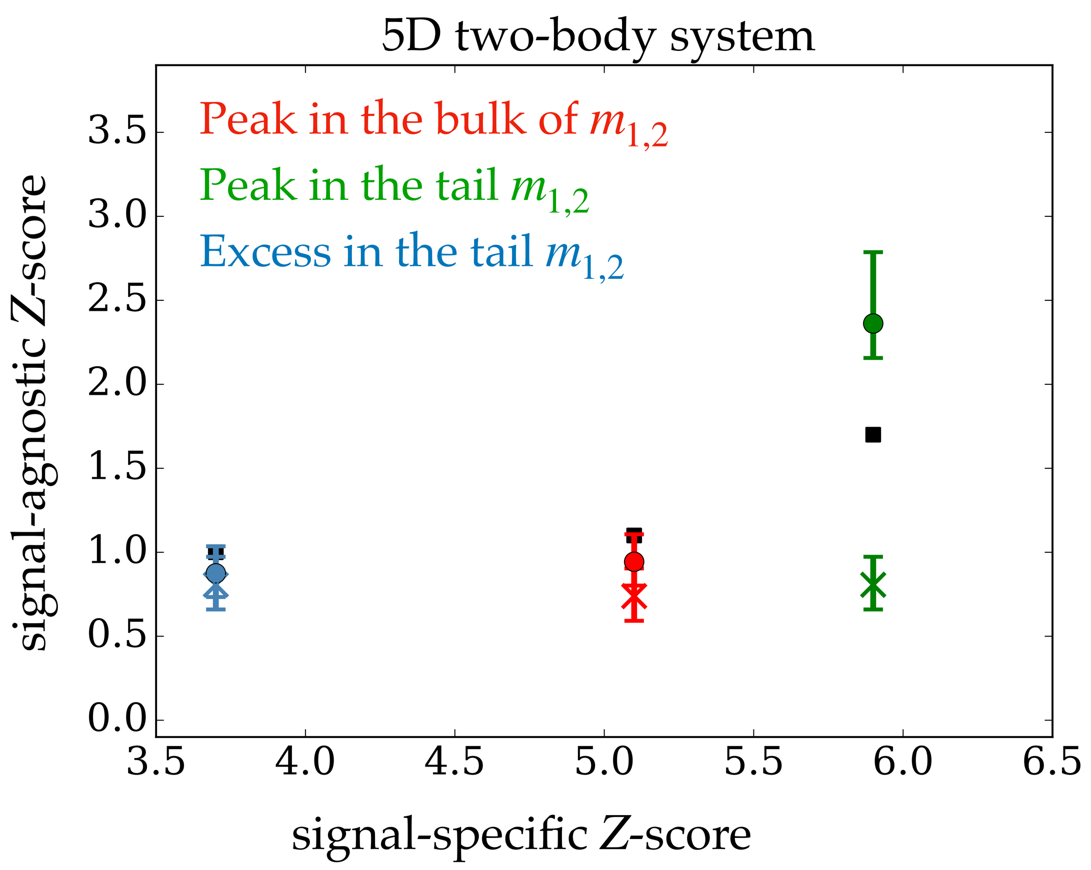
RESULTS

	Model	#par		time	Ref.
■	NN	96		~ 4.5 h	[1]
✕	GK*	10k	(M=10k, random int. $\{\mu_i\}$)	~ 40 s	[2]
○	**SGK***	**600**	**(M=100, learnable $\{\mu_i\}$)**	~ 1.5 h	**NEW!**

*$\sigma = q_{50\%}$: **median of pair-wise distance between points**

more with less

Same or improved sensitivity to signal benchmarks!



IMPLICATIONS

Resource efficient representation of anomalies

→ Interpretability

→ Data compression?

AI

SOLUTION

Sparse linear combination of Gaussian Kernels (SGK)

$$f_{\boldsymbol{\mu},\boldsymbol{w}}(x) = \sum_{i=1}^{\mathrm{M}} w_i\, k(x;\,\mu_i,\sigma_i)$$

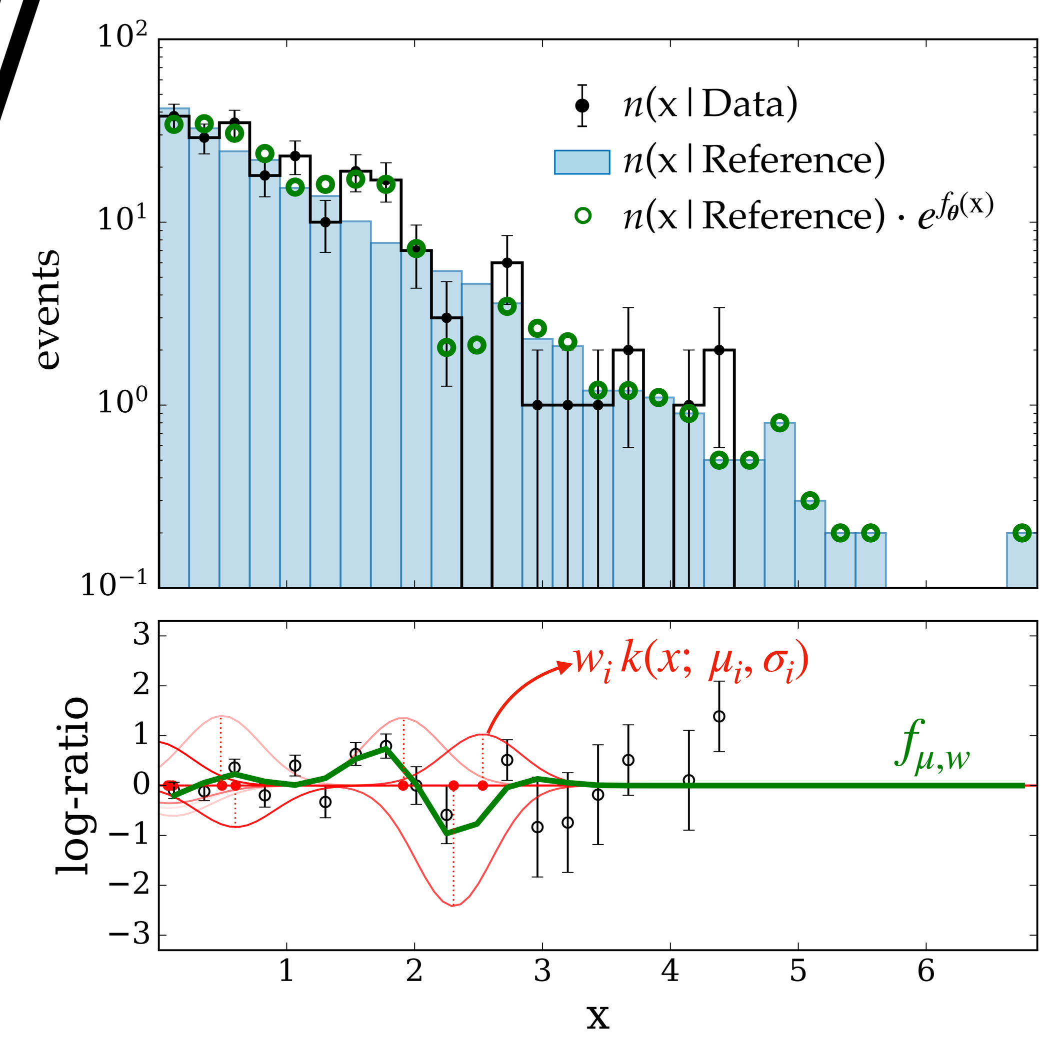
Local interpretability
Active kernels highlight anomalous regions

$$k(x;\,\mu_i,\sigma_i) = \mathrm{A}\exp\left[-\frac{||x-\mu_i||^2}{2\sigma_i^2}\right]$$

Sparse model (M ≪ N)
competition between data points to attract the kernels

Adaptive model (learnable μ)
directing *attention* to anomalous features

Smooth model ($\sigma^2 = \sigma^2_{\mathrm{exp}} + \sigma^2_X$)
Physics constraints (e.g. experimental resolution).
What is the scale of New Physics?



Loss function

$$L = L_{\mathrm{LR}} - \lambda_{\mathrm{H}} L_{\mathrm{H}} + \lambda_w L_w$$

Alternate training over $\{\mu_i\}$ and $\{w_i\}$

Likelihood-ratio loss

$$L_{\mathrm{LR}} = \sum_{x\in R} w_R(x)\,(\exp[f_{\boldsymbol{\mu},\boldsymbol{w}}(x)]-1) - \sum_{x\in D} f_{\boldsymbol{\mu},\boldsymbol{w}}(x)$$

Entropy regularization on $\{\mu_i\}$
Uniform distributed kernels prior

$$L_{\mathrm{H}} = -\sum_{j=1}^{\mathrm{M}} p(\mu_j)\,\log p(\mu_j)\,,\quad p(\mu) = \frac{1}{\mathrm{M}}\sum_{i=1}^{\mathrm{M}} k(\mu;\,\mu_i,\sigma_i)$$

L2 regularization on $\{w_i\}$
Smoothness constraint

$$L_w = \frac{1}{\mathrm{M}}\sum_{j=1}^{\mathrm{M}} w_j^2$$

References:

[1] "Learning multivariate new physics" Eur. Phys. J. C 81, 89 (2021)

[2] "Learning new physics efficiently with nonparametric methods" Eur. Phys. J. C, 82(10)

[3] "Goodness of fit by Neyman-Pearson testing" arXiv:2305.14137

[4] "ProtoNN: Compressed and Accurate kNN for Resource-scarce Devices" Int. Conf. on machine learning. PMLR, 2017