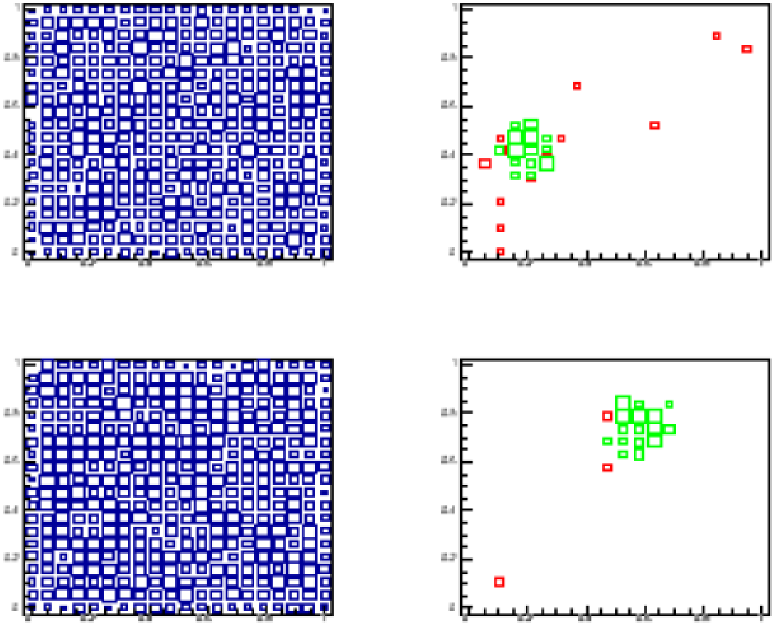
ICNFP, August 27, 2021

RanBox: Anomaly Detection in the Copula Space



Tommaso Dorigo, INFN-Padova

August 27th, 2021

Why should HEP go unsupervised?

Our capability to model known physics processes (SM) is terrific, so supervised and semi-supervised *modi operandi* might be all we need, but...

- We still do not trust our models in extreme regions of the phase space – tails of the distributions, corners of phase space not well lit up by previous studies
- Unfortunately, and obviously, those regions have large overlap with the place where NP can still hide
- In those cases we go data-driven, or just throw our hands up and use parametric models

The above is motivation for making one further step in the unknown: go fully unsupervised

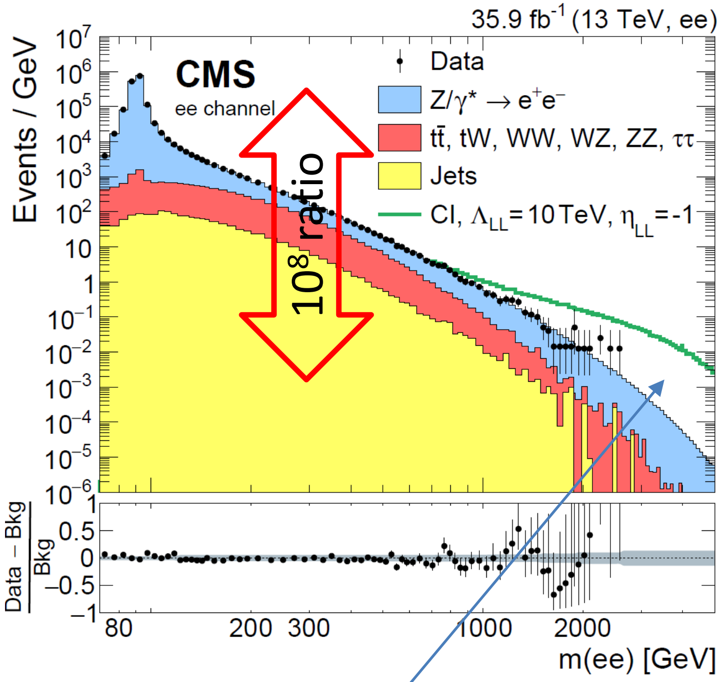
In this talk, I will consider a simple approach to anomaly detection: the search for overdensities in a standardized space

Uneven marginals

Anomaly detection, like many other ML tasks, benefits from suitably defining a metric in feature space

- What we are looking for are overdensities in a complicated, high-D space, populated in a very disuniform and sparsified way by discrete MC data points
  - e.g. think at p_T distributions, or invariant masses: an exponentially falling behaviour is commonplace

If you ask a unsupervised algorithm to locate an overdensity in a space spanned by kinematic variables, it will point at low p_T, low M



Above: an exponentially falling distribution: dielectron mass

New physics (in this case, contact interactions) is always in the tails!

The probability integral transform

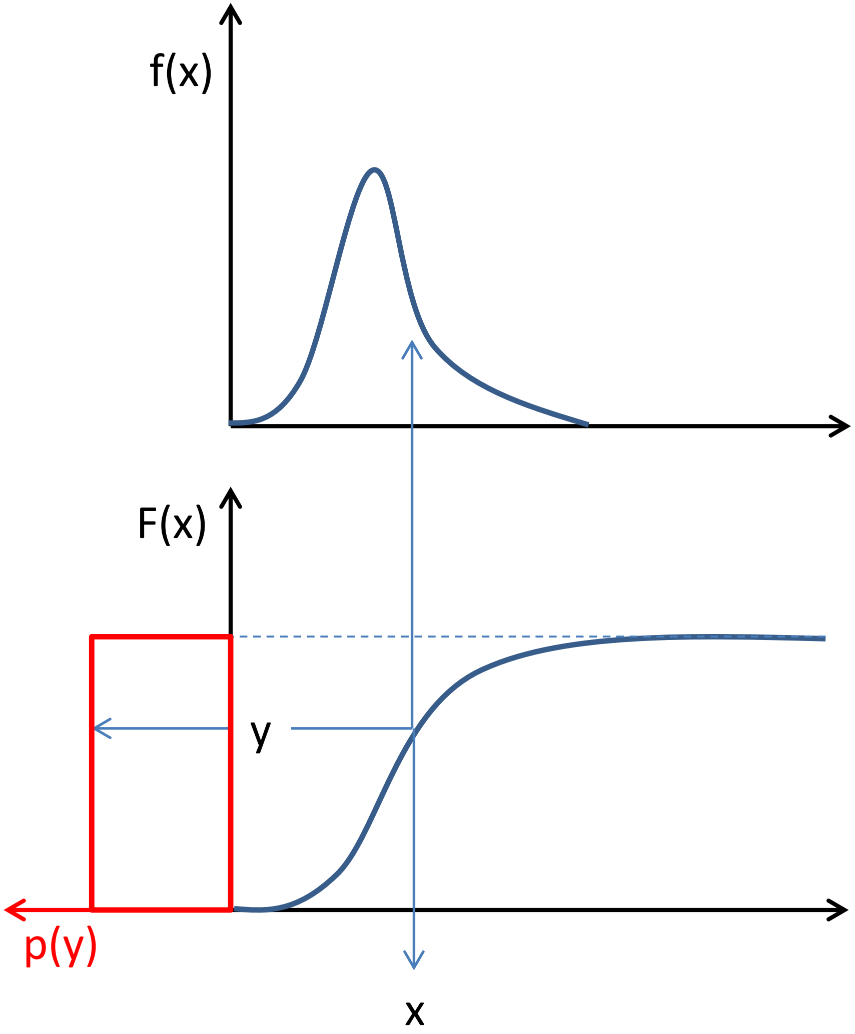
Given a normalized f(x), compute the cumulative distribution function of f:

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

Now,

y = F(x)

is uniformly distributed in [0,1]. There is a 1-to-1 map connecting x and y, so *the transformation is invertible*.

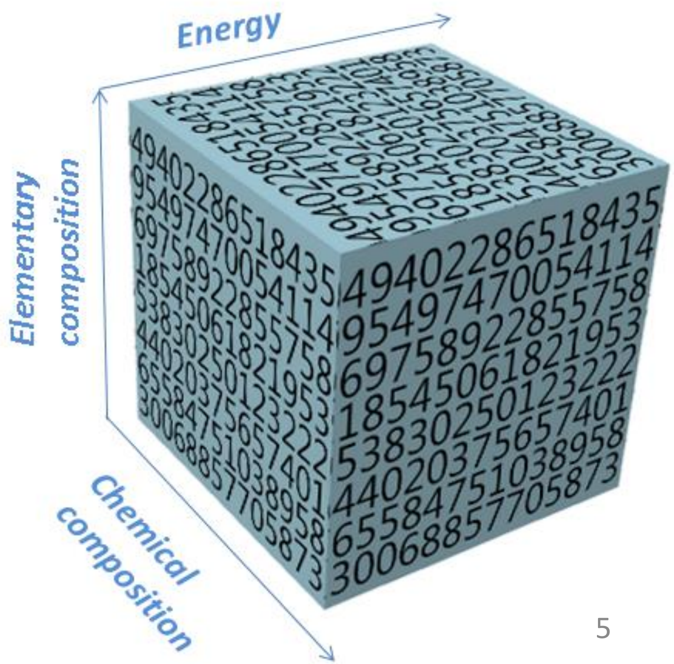


The Copula

In Statistics a useful concept is that of the **copula**: from Sklar's theorem, any multivariate joint distribution is decomposable in univariate marginals and a **uniquely defined copula** which describes the inter-dependence of the features

→ If you apply a **probability integral transform** to each coordinate, such that each feature has a flat marginal, all the information on data structure hides out in the interconnection of the features

- The data then nicely fit in a $[0,1]^D$ hypercube, of volume V=1
  - If you have N events in total, you expect V*N events in any given subspace of volume V
- Watched from each side, the cube has flat marginals – yet all the information is still there, packed in the correlation of the features
- Once your data fit in a cube, they are easy to represent - No more hassle with dimensionality of the features (you divided units out); discreteness looks less of an issue; boundaries are identical



Correlated variables removal (CVR)

We may reduce the dimensionality by PCA, but an alternative is to discard the M<D variables most correlated with others:

- we order the D(D-1) correlation coefficients
- we find combination of M variables that, once removed, minimize the highest remaining correlation

Coeff.	value
ρ_{34}	0.99
ρ_{12}	0.98
ρ_{13}	0.92
ρ_{26}	0.88
ρ_{38}	0.87
ρ_{35}	0.86
ρ_{28}	0.84
ρ_{68}	0.84
ρ_{14}	0.81
ρ_{46}	0.77
…	…

Remove var. 3

Coeff.	value
ρ_{34}	-
ρ_{12}	0.98
ρ_{13}	-
ρ_{26}	0.88
ρ_{38}	-
ρ_{35}	-
ρ_{28}	0.84
ρ_{68}	0.84
ρ_{14}	0.81
ρ_{46}	0.77
…	…

Remove var. 3,1

Coeff.	value
ρ_{34}	-
ρ_{12}	-
ρ_{13}	-
ρ_{26}	0.88
ρ_{38}	-
ρ_{35}	-
ρ_{28}	0.84
ρ_{68}	0.84
ρ_{14}	-
ρ_{46}	0.77
…	…

Remove var. 3,1,2

Coeff.	value
ρ_{34}	-
ρ_{12}	-
ρ_{13}	-
ρ_{26}	-
ρ_{38}	-
ρ_{35}	-
ρ_{28}	-
ρ_{68}	0.84
ρ_{14}	-
ρ_{46}	0.77
…	…

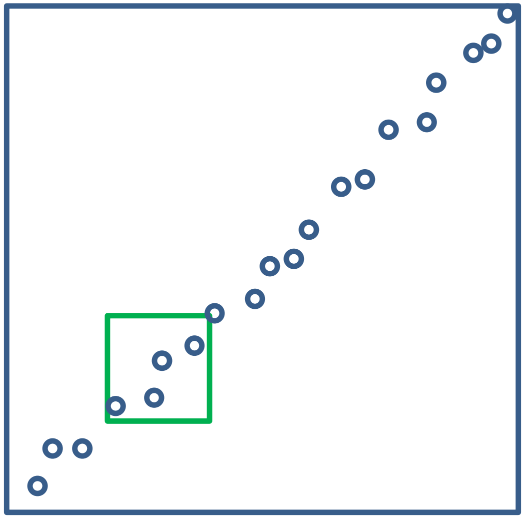
(Why correlations are annoying)

Imagine you put a billion events in a 9-dim space; define a small interval extending by 0.1 in each dimension, e.g [0.2,0.3]^9 → $V_{box} = 10^{-9}$.

If you were to estimate the number of events in such a box, you would do $N_{exp} = N\, V_{box} = 1$ event

Now, what if the space is spanned by the same variable repeated 9 times (=100% correlation)? The box will not contain 1 event, but rather, 100 million! ooops...

This extreme example shows that volume-based estimates of the number of events are heavily affected by correlations.



In 2D: 20 points,
box [0.2,0.4]x[0.2,0.4]
V=0.04 → Expect 0.8, see 4

RanBox

After the integral transform, we have our data in a nice, unit-volume D'-dim cube. Any (large) disuniformity in this space is potentially interesting

Physics creates disuniformities.

- An unsupervised anomaly detection algorithm may spot them and assess their departure from a uniform density hypothesis
  - It may be able to find surprising structures, or only stick to completely predictable ones
  - eventually a human intervention is required to scan and assess the results

Our task is to identify interesting overdense regions in **subspaces of D'**

  - We have the preconception that some, if not most, of the features of NP events will not be "anomalous" → hence we need to **focus on subspaces**

- We may approach this problem by stochastic means:
  - randomly pick a subset of the features
  - scan the resulting space with a D"<D'-dim parallelepiped (a "box"), looking for boundaries which maximize some suitable figure of merit

Sidebands

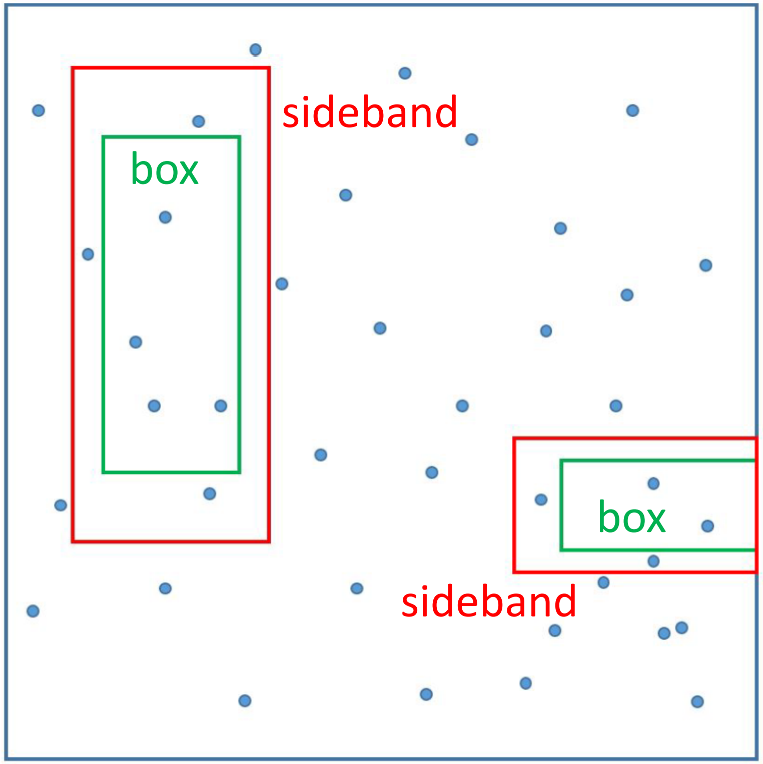
The number of events in the box can be predicted by $N_{exp} = N\ V_{box}$, but a better «local» estimate of density can be obtained by looking within a sideband in D’ dimensions.

If we define

$$\delta_i = 0.5(x^i_{max} - x^i_{min})(2^{1/\mathcal{D}'} - 1),$$
$$x^i_{min,SB} = \max(0, x^i_{min} - \delta_i),$$
$$x^i_{max,SB} = \min(1, x^i_{max} + \delta_i),$$

the «sideband» box has a volume $V_{SB}=2\ V_{box}$, so counting events in the sideband and not contained in the signal box provides a less biased (but higher variance) estimate than the «global» one above) $N_{exp} = N_{SB,not\ in\ box}$

In 2 dimensions (below) sidebands look wide, but in 10D they get quite skinny...



Z_{PL} definition

We need to define a measure of the "pseudo-significance" of observing N_{on} data inside a box, when we see N_{off} outside, with known volume ratio inside/outside.

This is the **so-called "on-off" problem** often encountered in astrophysics searches

A handy definition was proposed by Li and Ma [T. Li, Y. Ma, Astrophysical Journal 272 (1983) 317] in the form of the "Profile Likelihood" Z_{PL} obtained from:

$$\mathcal{L}_{\mathrm{P}} = \frac{(\mu_{\mathrm{s}} + \mu_{\mathrm{b}})^{n_{\mathrm{on}}}}{n_{\mathrm{on}}!} e^{-(\mu_{\mathrm{s}}+\mu_{\mathrm{b}})} \frac{(\tau\mu_{\mathrm{b}})^{n_{\mathrm{off}}}}{n_{\mathrm{off}}!} e^{-\tau\mu_{\mathrm{b}}}$$

(tau is ratio of time spent looking off and on source)

whence one gets

$$\Lambda(\mu_{\mathrm{s}}) = \frac{\mathcal{L}\left(\mu_{\mathrm{s}}, \tilde{\tilde{\mu}}_{\mathrm{b}}(\mu_{\mathrm{s}})\right)}{\mathcal{L}\left(\tilde{\mu}_{\mathrm{s}}, \tilde{\mu}_{\mathrm{b}}\right)}$$

and

$$Z_{\mathrm{PL}} = \sqrt{-2\ln\Lambda\left(\mu_{\mathrm{s}} = 0\right)}$$

- R.D. Cousins, J. Linnemann and J. Tucker examine that and other definitions [Arxiv:0702156]. Bottomline, **Z_{PL} is quite okay for fast calculations**
- Advantage over other definitions: Z_{PL} *can* be computed for any N_{on}, N_{off}

Alternative test statistic: R_{reg}

Z_{PL} is a good approximation to the significance of a counting experiment, but its maximization may chase wide background fluctuations producing large excesses (e.g., 1200 evts when 1000 expected → «6 sigma» or so)

- in a unsupervised search, a more sensitive measure may be the ratio of observed divided by expected density

→ For searches of small, localized signals, use

$$R_{reg} = N_{obs}/(N_{exp}+N_{reg}),$$

e.g., with N_{reg}=1 (to avoid divergence of R)

The full RanBox routine

- Define a dataset you want to search in
- Apply **integral transform** to all features
- (Optional) **do dim. reduction**, select D' features by PCA or CVR
- Loop on subspaces - repeat many times:
  - Choose at random subspace D" of D' of workable dimensionality (e.g. D"=6-10)
  - **Find seed box** in D"
  - Maximize TS by **gradient descent**

The procedure produces an ordered list of the boxes of highest TS in the corresponding most promising subspaces

The iterative version: RanBoxIter

The random scan of subspaces may be ineffective when one does not know what is the typical dimensionality of the subspace where features are distinctive of the signal, and/or when the dimensionality of the space is too high

An iterative version of the code does the scan incrementally, starting with all 2D combinations of features, and adding one dim at a time, keeping track of the N (20-50) best boxes at every dimensional step

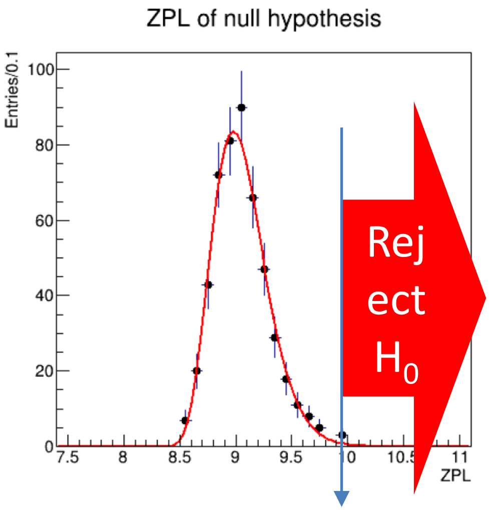
RanBoxIter has quite comparable performance to RanBox, and may be quicker to run in certain situations

Power studies

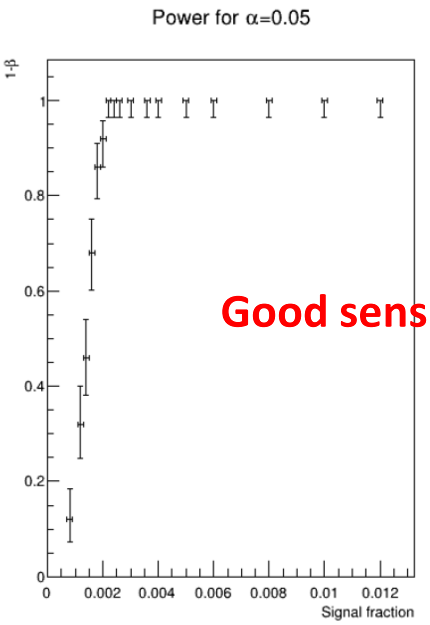
To test the performance of the algorithm, we generate synthetic data with

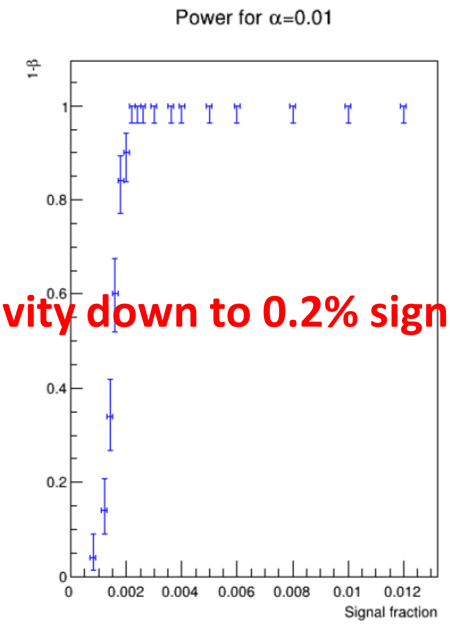
- background: flat in all D features
- signal: multi-variate Gaussian in some features, flat in all others

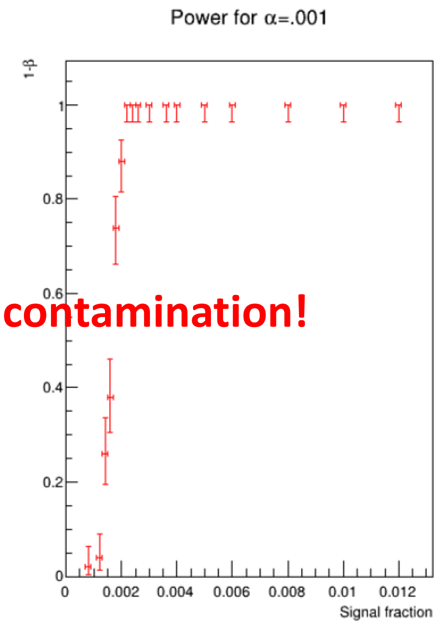
and study whether best box returns signal-rich regions by defining a critical region for the test statistic using background-only trials



Right: power 1-β as a function of the signal fraction, using 5000-event samples, with D=20, and where signal has multivariate Gaussian distr. in 15 features







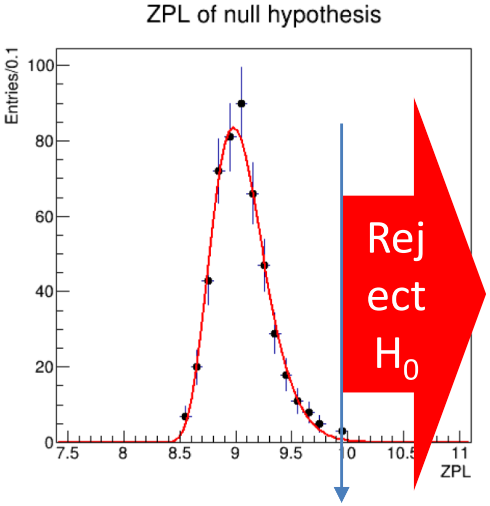
Good sensitivity down to 0.2% signal contamination!

Power studies

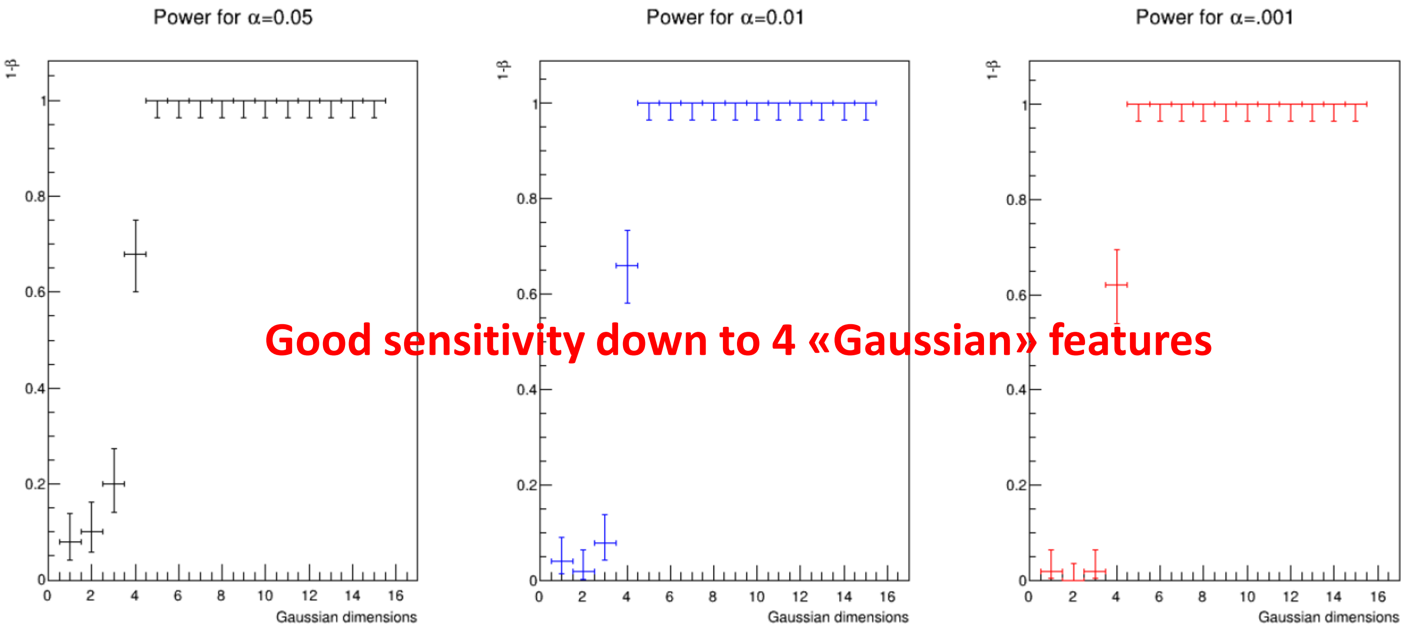
To test the performance of the algorithm, we generate synthetic data with

- background: flat in all D features
- signal: multi-variate Gaussian in some features, flat in all others

and study whether best box returns signal-rich regions by defining a critical region for the test statistic using background-only trials



Right: power 1-β as a function of the number of distinctive features of the signal, using 5000-event samples, with D=20 and a 1% signal contamination



Good sensitivity down to 4 «Gaussian» features

Graphical view in 2D

Here shown is the result of a search in 6-dimensional subspaces. The best box «captures» the small signal by focusing on six features where it has Gaussian distributions

Red points are events that are discarded because of their value on the shown features (and would pass all other selections)

full copula space

selected box

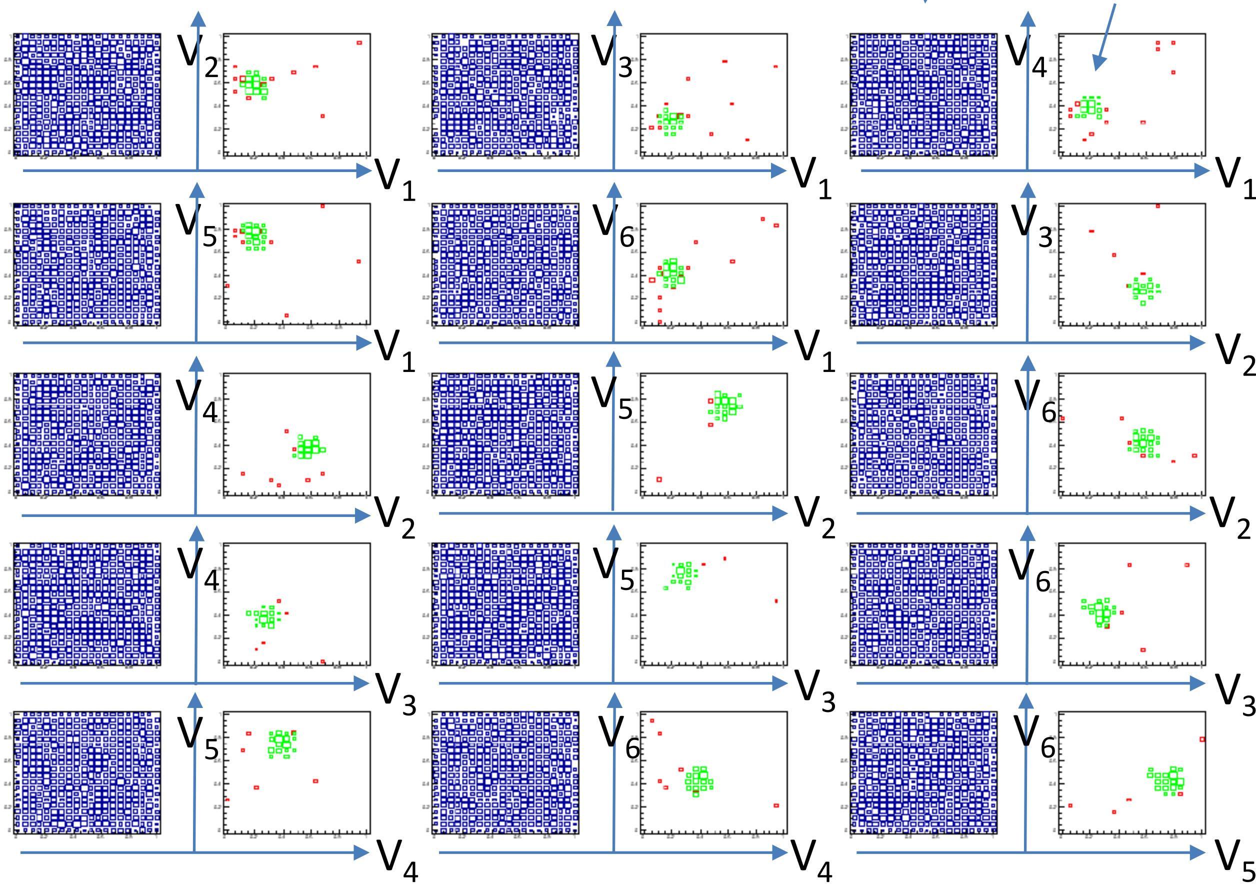
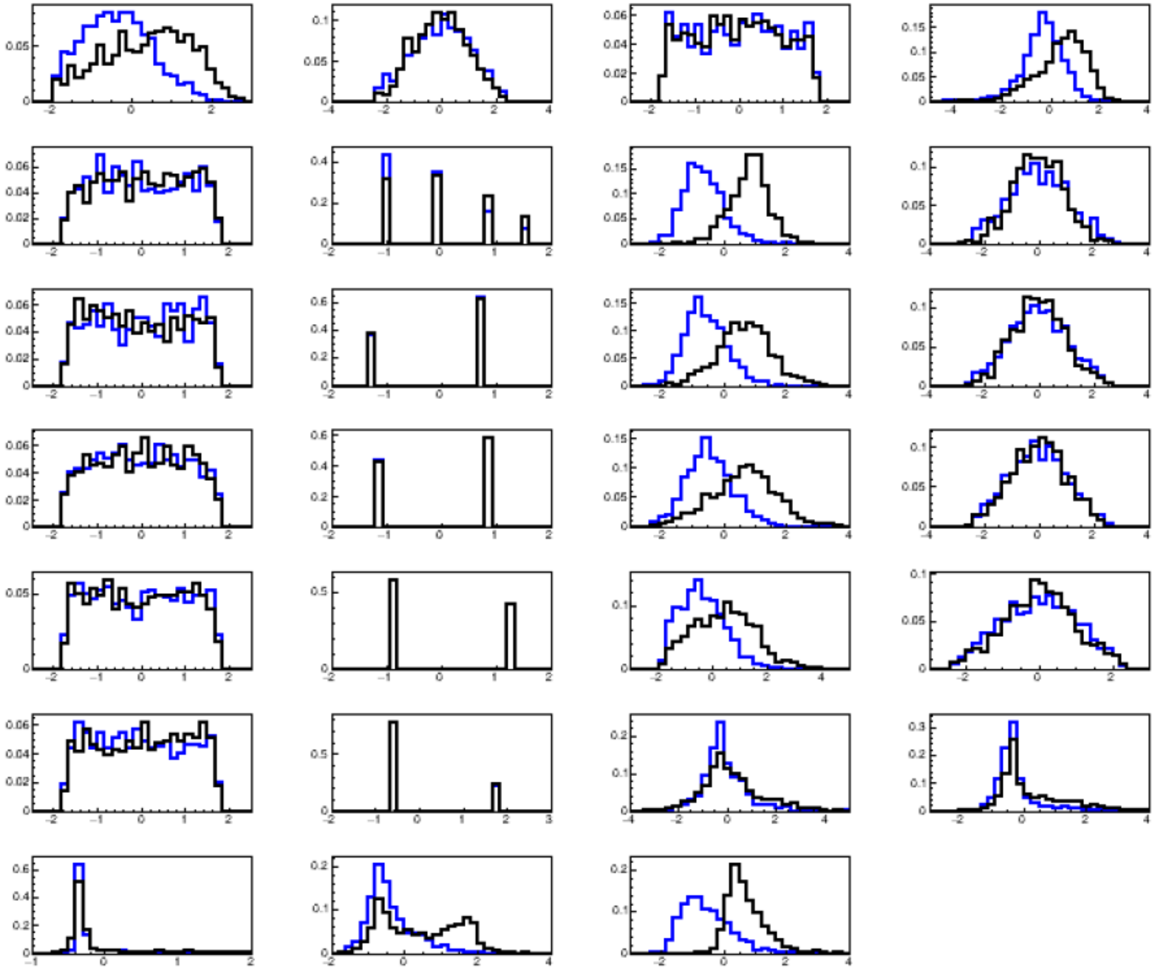


Figure 10: Scatterplots of the six features defining the subspace where RanBox *finds the highest-Z_{PL} box in a run on 5000 synthetic events, 4950 of them generated from a $D = 20$-dimensional uniform distribution and the remaining 50 “signal” events generated with 11 features drawn from a multidimensional Gaussian distribution. The distribution of the totality of the data is shown in blue on the left of each pair of graphs, while the distribution of selected events (in green) is shown in green on the corresponding right graph; in red are events that fail to be included in the highest-Z_{PL} box only because of their value of the shown features. From top to bottom and left to right each pair of graph describes the spaces (v_1, v_2), (v_1, v_3), (v_1, v_4) (first row), (v_1, v_5), (v_1, v_6), (v_2, v_3) (second row), (v_2, v_4), (v_2, v_5), (v_2, v_6) (third row), (v_3, v_4), (v_3, v_5), (v_3, v_6) (fourth row), and (v_4, v_5), (v_4, v_6), (v_5, v_6) (fifth row). See the text for other details.*

Tests on HEPMASS dataset

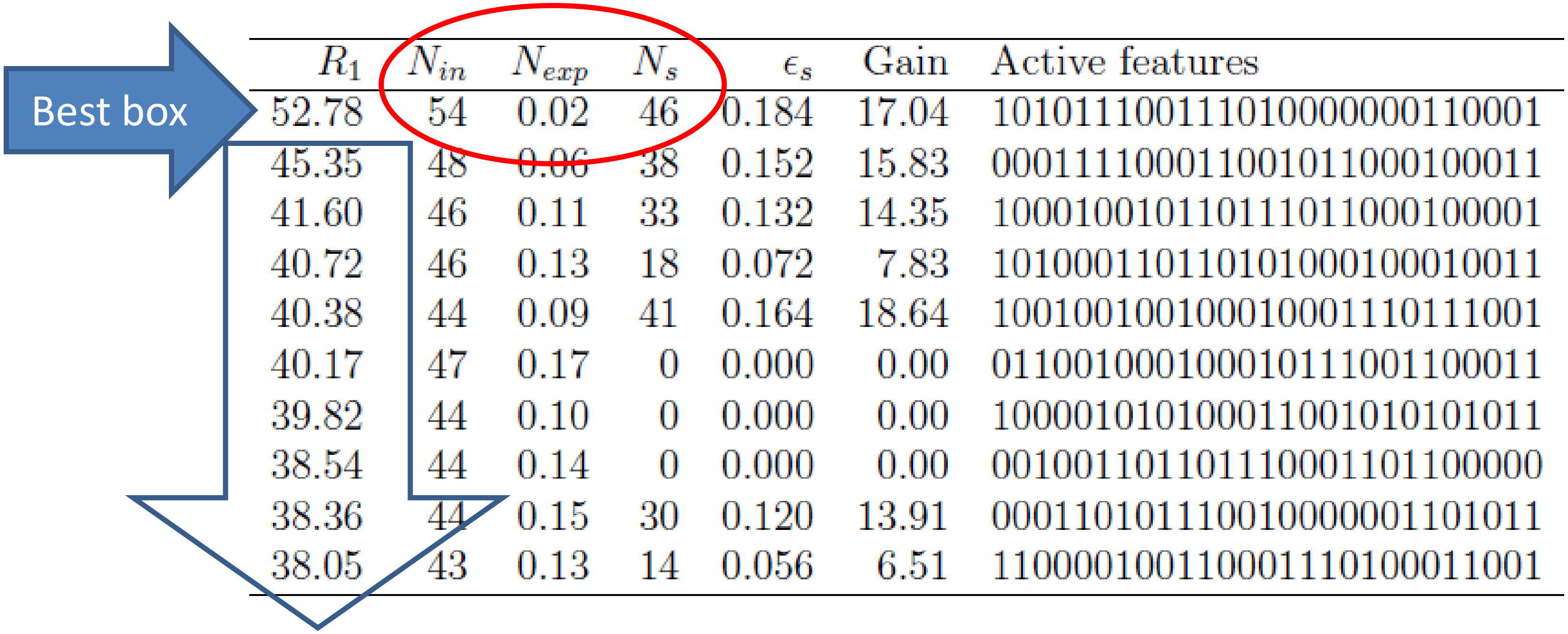
For a more realistic HEP scenario we use the HEPMASS dataset (UCI repository), which contains background (top pair production, blue) and signal (a resonance decaying to ttbar, black), subjected to a fast reconstruction simil-ATLAS.

27 event features are available per event



Results from 10,000 subspace scan

12-dimensional subspaces are scanned by RanBox on 5000 events with a 5% signal contamination (4750 bgr, 250 signal). The algorithm reports the 10 best boxes below:

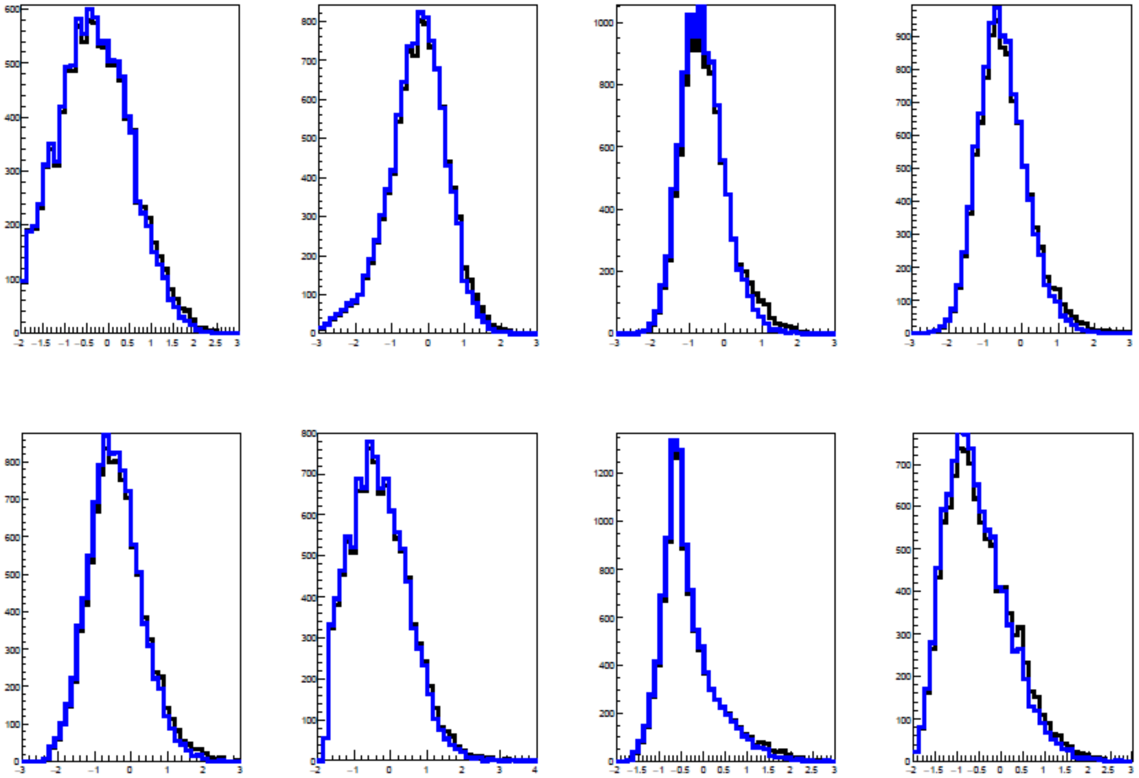


R_1	N_{in}	N_{exp}	N_s	ϵ_s	Gain	Active features
52.78	54	0.02	46	0.184	17.04	101011100111010000000110001
45.35	48	0.06	38	0.152	15.83	000111100011001011000100011
41.60	46	0.11	33	0.132	14.35	100010010110111011000100001
40.72	46	0.13	18	0.072	7.83	101000110110101000100010011
40.38	44	0.09	41	0.164	18.64	100100100100010001110111001
40.17	47	0.17	0	0.000	0.00	011001000100010111001100011
39.82	44	0.10	0	0.000	0.00	100001010100011001010101011
38.54	44	0.14	0	0.000	0.00	001001101101110001101100000
38.36	44	0.15	30	0.120	13.91	000110101110010000001101011
38.05	43	0.13	14	0.056	6.51	110000100110001110100011001

Table 3: Results of an exploratory RanBox *search on the HEPMASS dataset with a 5% signal contamination; data for the 10 most significant boxes are reported. N_s indicates the number of signal events in the search box; ϵ_s is the efficiency of the box selection for the signal component; gain is computed as the increase in the signal fraction of the box over the initial dataset. For other detail see the text.*

5% signal contamination in HEPMASS

A 5% contamination might sound like a large one, but in 1D it would be hard to spot, especially if you don't know what you are looking for, as a unsupervised algorithm

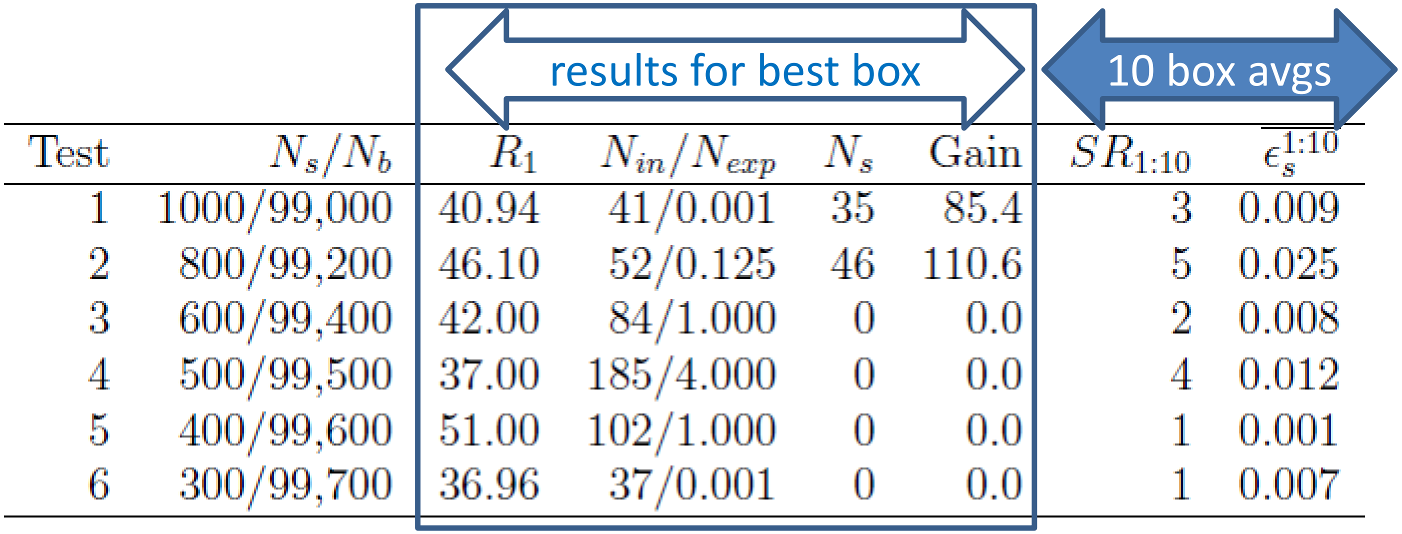


Blue: pure background
Black: 5% signal mixed in

Figure 13: Comparison of the distribution of pure background (blue) and a mixture of 5% signal and background (black) in the most discriminating features in the HEPMASS dataset. Left to right, top to bottom: features 0, 3, 6, 10, 14, 18, 25, and 26.

Higher-stat runs

When the dataset size is increased, RanBox has higher sensitivity as smaller signal components can emerge from Poisson noise. Below, in 100,000 event mixtures less than 1% contaminations are spotted by the iterative version of the algorithm



results for best box

10 box avgs

Test	N_s/N_b	R_1	N_{in}/N_{exp}	N_s	Gain	$SR_{1:10}$	$\overline{\epsilon_s^{1:10}}$
1	1000/99,000	40.94	41/0.001	35	85.4	3	0.009
2	800/99,200	46.10	52/0.125	46	110.6	5	0.025
3	600/99,400	42.00	84/1.000	0	0.0	2	0.008
4	500/99,500	37.00	185/4.000	0	0.0	4	0.012
5	400/99,600	51.00	102/1.000	0	0.0	1	0.001
6	300/99,700	36.96	37/0.001	0	0.0	1	0.007

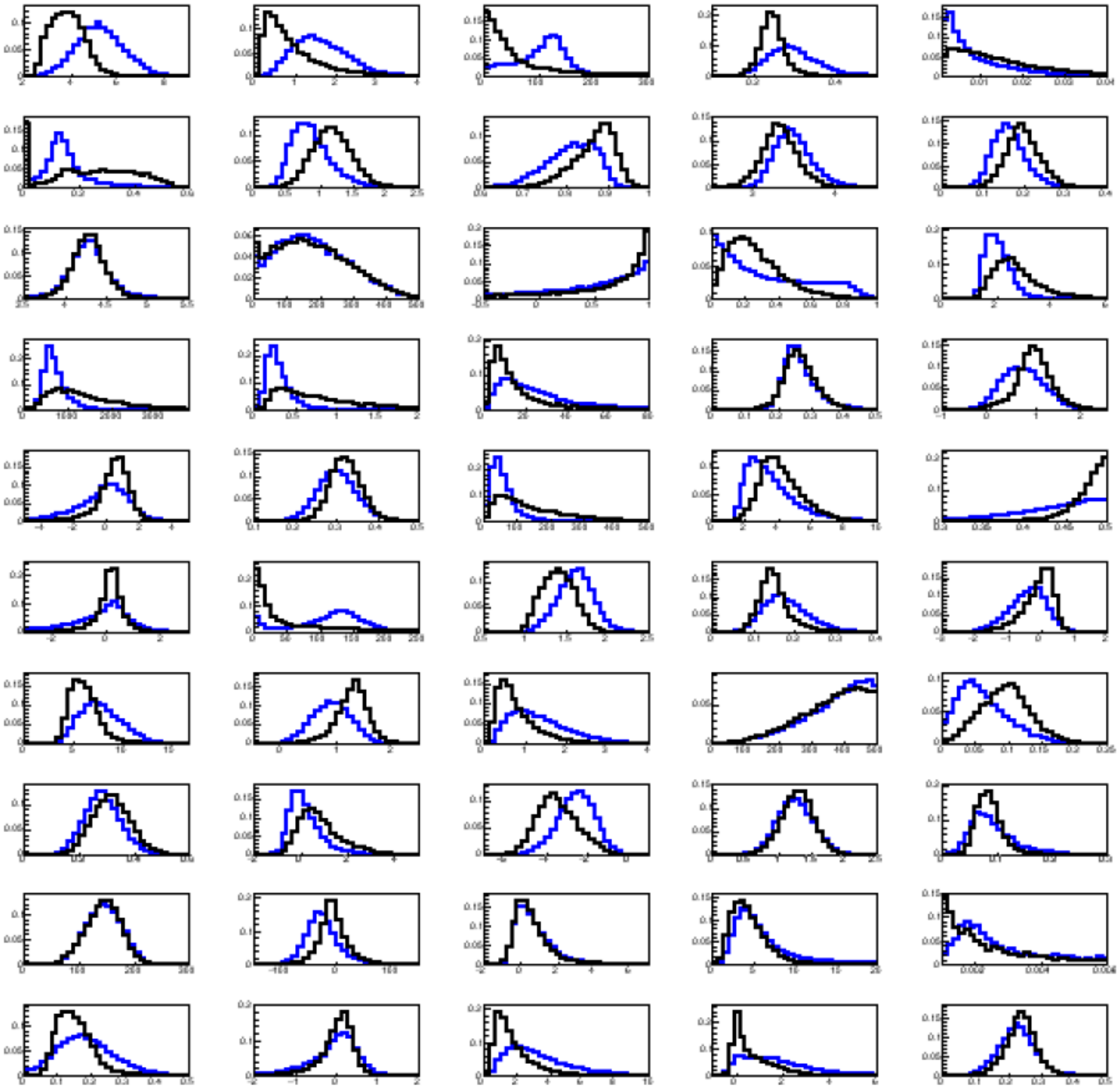
Above, $SR_{1:10}$ indicates the fraction of boxes among the first 10 which include a signal fraction increased by a factor of 6 or larger with respect to the initial sample; also reported is the average signal efficiency of the 10 best boxes.

Electron neutrinos in MiniBooNE

Another famous dataset, also available in UCI repository, is the one of electron neutrino interactions and backgrounds.

It contains 50 features per event, so it lends itself well to studies of dimensionality reduction at preprocessing

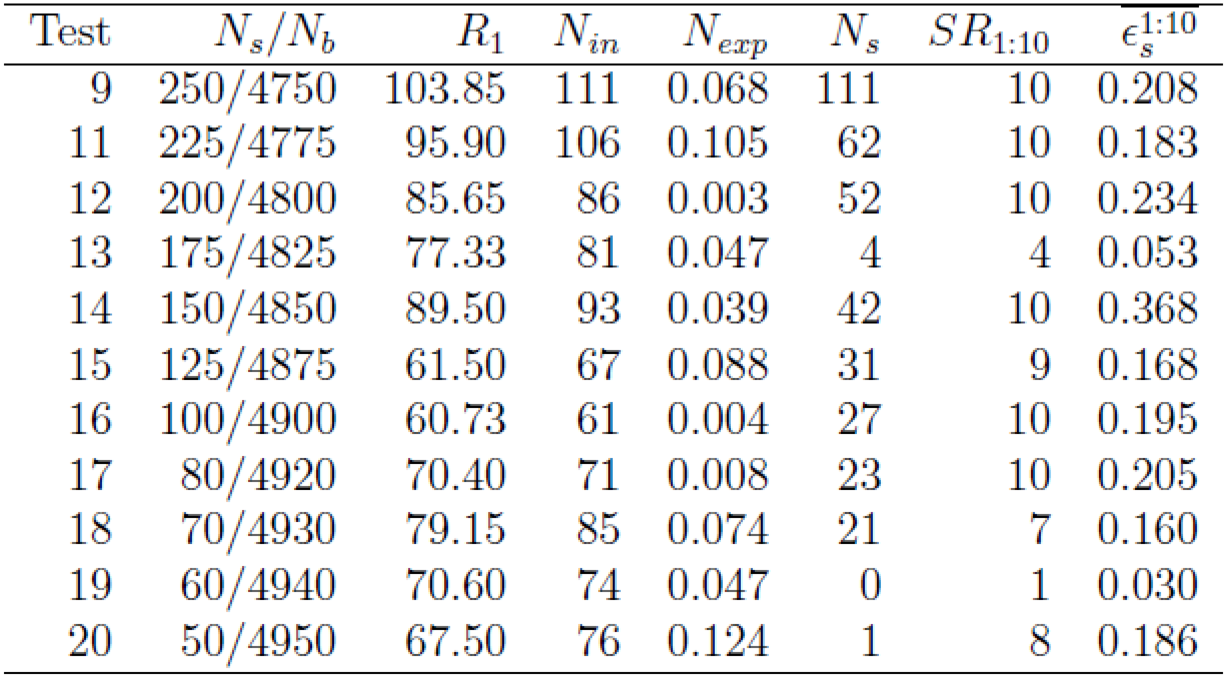
Signal (electron CC interactions) is in black, backgrounds in blue



Results of RanBoxIter scans

D=50 is large → apply dim. reduction. The most effective way is to remove by CVR the 10 variables most correlated with others, bringing the dimensionality to D’=40.

RanBoxIter scans find signal component in a good fraction of the best solutions, down to 1 % signal contaminations.



Test	N_s/N_b	R_1	N_{in}	N_{exp}	N_s	$SR_{1:10}$	$\overline{\epsilon_s^{1:10}}$
9	250/4750	103.85	111	0.068	111	10	0.208
11	225/4775	95.90	106	0.105	62	10	0.183
12	200/4800	85.65	86	0.003	52	10	0.234
13	175/4825	77.33	81	0.047	4	4	0.053
14	150/4850	89.50	93	0.039	42	10	0.368
15	125/4875	61.50	67	0.088	31	9	0.168
16	100/4900	60.73	61	0.004	27	10	0.195
17	80/4920	70.40	71	0.008	23	10	0.205
18	70/4930	79.15	85	0.074	21	7	0.160
19	60/4940	70.60	74	0.047	0	1	0.030
20	50/4950	67.50	76	0.124	1	8	0.186

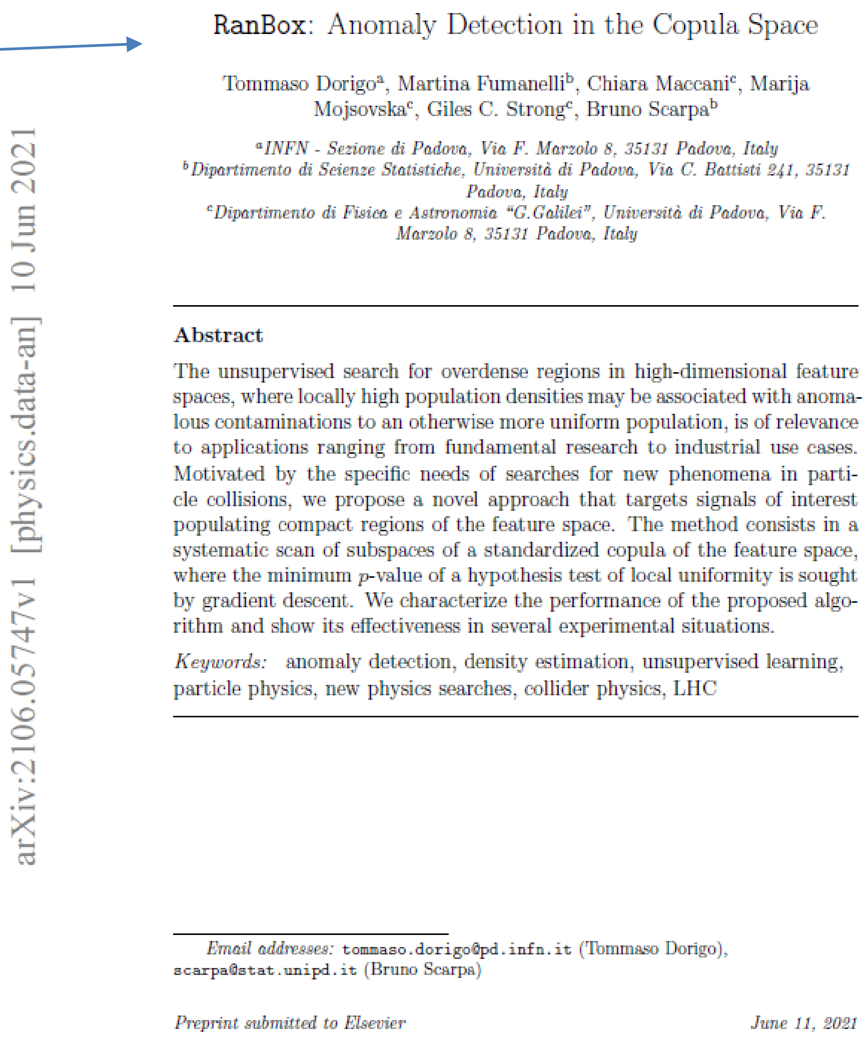
Conclusions and future work

A paper is out (and sent to Comput. Phys. Comms.)

The algorithm is ready to take on real HEP data! Meanwhile, we will

- Improve search of minimum (GD implementation is artisanal)
- Calibrate Z with Bonferroni correction for Hypothesis test applications
- **Investigate semi-supervised approach** (partial MC input for background)

- Release python and root macro versions of code



arXiv:2106.05747v1 [physics.data-an] 10 Jun 2021

RanBox: Anomaly Detection in the Copula Space

Tommaso Dorigo[a], Martina Fumanelli[b], Chiara Maccani[c], Marija Mojsovska[c], Giles C. Strong[c], Bruno Scarpa[b]

[a]INFN - Sezione di Padova, Via F. Marzolo 8, 35131 Padova, Italy
[b]Dipartimento di Scienze Statistiche, Università di Padova, Via C. Battisti 241, 35131 Padova, Italy
[c]Dipartimento di Fisica e Astronomia "G.Galilei", Università di Padova, Via F. Marzolo 8, 35131 Padova, Italy

Abstract

The unsupervised search for overdense regions in high-dimensional feature spaces, where locally high population densities may be associated with anomalous contaminations to an otherwise more uniform population, is of relevance to applications ranging from fundamental research to industrial use cases. Motivated by the specific needs of searches for new phenomena in particle collisions, we propose a novel approach that targets signals of interest populating compact regions of the feature space. The method consists in a systematic scan of subspaces of a standardized copula of the feature space, where the minimum p-value of a hypothesis test of local uniformity is sought by gradient descent. We characterize the performance of the proposed algorithm and show its effectiveness in several experimental situations.

Keywords: anomaly detection, density estimation, unsupervised learning, particle physics, new physics searches, collider physics, LHC

Email addresses: tommaso.dorigo@pd.infn.it (Tommaso Dorigo), scarpa@stat.unipd.it (Bruno Scarpa)

Preprint submitted to Elsevier *June 11, 2021*

Thank you for your interest !

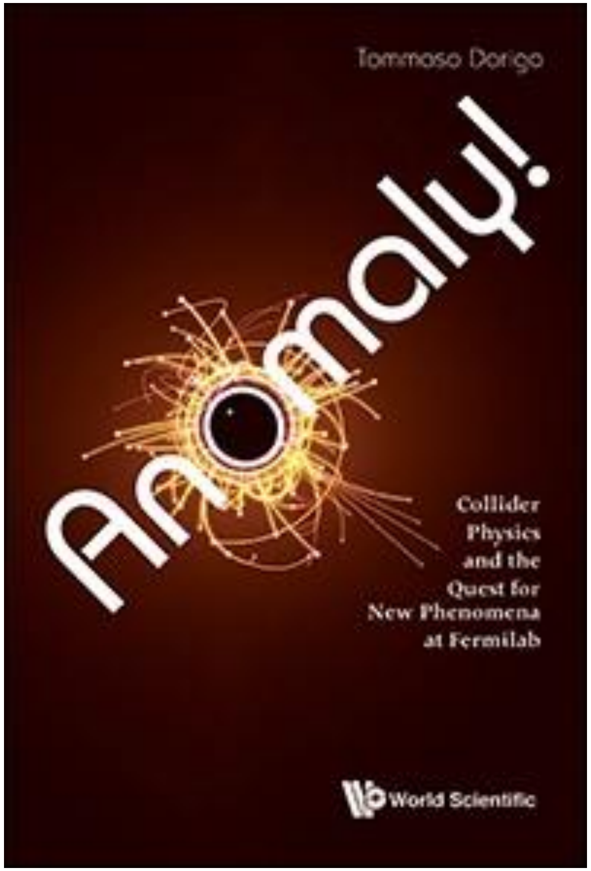
Backup

Anomaly!

A sideline:

Your speaker is not a true expert in anomaly detection...

Which is a great opportunity to have a unbiased view of the problem!


Tommaso Dorigo
Anomaly!
Collider Physics and the Quest for New Phenomena at Fermilab
World Scientific

OTOH I wrote a book on anomalies in collider data

- An account of the complex sociology of a particle physics experiment and the controversies around anomalous findings
  - Ed Witten: "*a fascinating look*"
  - Gordy Kane: *"charming and irreverent"*
  - Peter Woit: *"clear-eyed view... compelling"*
  - Sean Carroll: *"entertaining and provocative"*
  - GF Giudice*: "a captivating narrative"*
  - Times HE: *"a guilty pleasure"*
  - Physics World: *"engaging and insightful"*

In case you are interested:

- More info at World Scientific site,

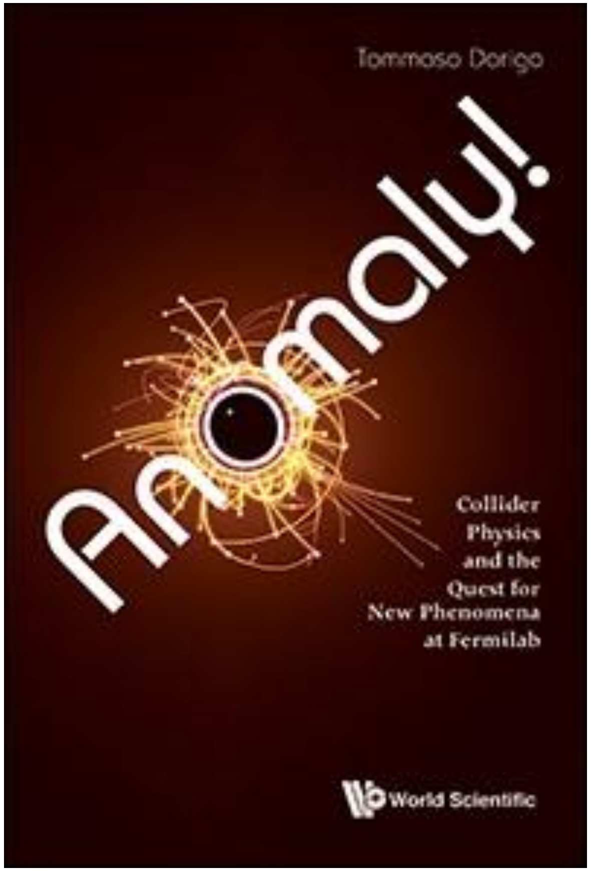
https://www.worldscientific.com/worldscibooks/10.1142/q0032

- Get a copy at https://tinyurl.com/y69h7jpm

(or at the CERN bookstore)

- Or email me for a free pdf if your financial means don’t suffice

Anomaly!


Tommaso Dorigo
Anomaly!
Collider Physics and the Quest for New Phenomena at Fermilab
World Scientific

A sideline:

Your speaker is not a true expert in anomaly detection...

Which is a great opportunity to have a unbiased view of the problem!

OTOH I wrote a book on anomalies in collider data

- An account of the complex sociology of a particle physics experiment and the controversies around anomalous findings

shameless self-promotion censored

In case you are interested:

- More info at World Scientific site,

https://www.worldscientific.com/worldscibooks/10.1142/q0032

- Get a copy at https://tinyurl.com/y69h7jpm

(or at the CERN bookstore)

- Or email me for a free pdf if your financial means don’t suffice

One more (optional) step: PCA

In the typical feature space of data passing some trigger, or even preselected with basic cuts ("*two photons plus anything*", "*four or more jets*", etc.) one finds **large correlations** between informative features

- the basic reason is that the momenta of observed objects are related by physical conservation laws: **matrix element + selection biases** do it

How can we even hope to get sensitive to those extra hidden correlations between some of the features (and **we do not know which**) that a signal may have added to the mix?

→ We may **rotate the data such that we get uncorrelated features**, by doing a Principal Component Analysis (PCA)

RanBox optionally does PCA before fitting data in a cube, but this transformation may reduce the sensitivity (as «low variance» components for the background may still be the most distinguishable feature of signal)

Principal Component Analysis

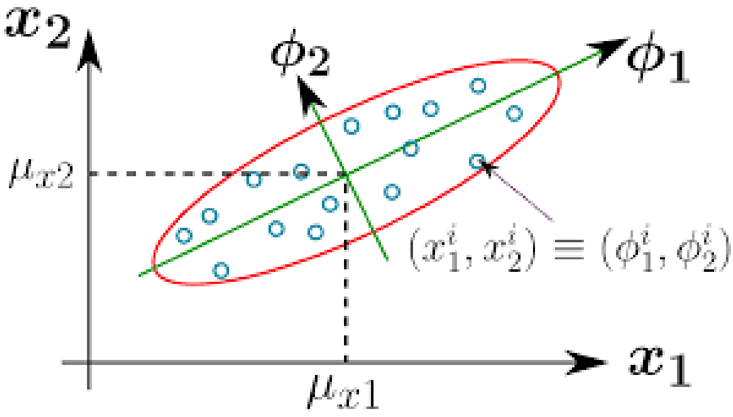
PCA applies an orthogonal transformation to a D-dimensional space, obtaining a new D'<=D –dimensional space where each of the new features (aptly called "principal components") are **linearly uncorrelated**

- The first PC is the one that accounts for the maximum amount of variance in the data; each successive PC accounts for most of the residual variance, while living in a subspace orthogonal to the previous one(s)

What PCA does is a hyperellipsoid fit to the data.
The ellipsoid is narrow along uninformative directions

One may thus only consider the first D'<D principal components, killing two birds with one stone:

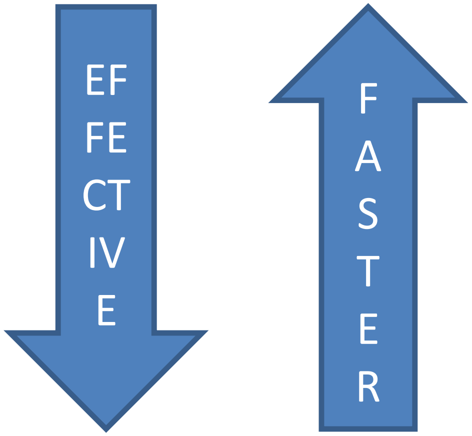
1) dimensionality reduction
2) the new variables are uncorrelated, so they simplify the problem of finding pockets of overdensity in the copula



Other nuts and bolts

The search for the highest-significance box is complicated, because data are sparse. There are a huge number of local maxima where a maximization routine can get stuck.

- One thus needs a smart seeding to inform the choice of initial box boundaries, easing the convergence task
- Three strategies considered:
  - **random initial box** boundaries
    - fastest but, er, random
  - window around region of **maximum kernel density**
    - slow, but good unless high-D sparsity makes it ineffective
  - area containing **cluster of close points**
    - best for high-D search, slowest


EFFECTIVE
FASTER

RanBox parameters

- (If PCA) Number of PC retained as "restricted feature space" where the search is made, **D’**
- (If CVR) Number of features retained, **D’**
- Number of variables **D’’** of subspaces searched for best box
- Number of subspaces scanned
  - if large (>1/3) compared with combinatorial D’!/[(D’’!)(D’-D’’)!], a full scan is instead made
- Box initialization method

  (0 – random, 1:4 – various clustering choices, 5 – kernel density)
- Mode "real data": amount of data / kind of data to read / fraction of signal and background
- Mode "toy": number of Gaussian dimensions, fraction of non-flat data, correlation among features
- Choice of test statistic (Z_{PL}/R_1)
- Choice of density estimation method (full volume/sideband)
- Max number of GD loops in minimization (100 – typical convergence takes 40-50 steps)

Maximization: gradient descent

Being such a snob SOB, I do not rely on prepackaged minimization routines, and wrote my own gradient descent routine

- "The problem has distinct features that make it amenable to specialized solutions"

Not interesting to discuss details here, except that indeed:

- discreteness and sparsity of data in multi-D create large number of minima, noisy convergence to good ones
  - useful to schedule periodic variation of learning rate
- we are **not interested in generalization**, as we want to minimize on data at hand
  - stochastic GD useless

- Acceleration not really crucial but it does improve speed

Bonferroni correction?

Of course, we are looking for excesses of events in a gazillion places → saying that Z is overestimated is a understatement

- The ordering by Z_{PL} that RanBox performs does imply that what we are looking for are interesting regions of phase space, not a quantitative assessment of the "significance" of an excess
- However, the actual value of max(Z_{PL}) might be useful, to inform on how interesting the region is. One then needs to calibrate it.

One may define a procedure to determine a Bonferroni correction (aka a way to account for the look-elsewhere effect)

- This is **work in progress**. It will require:
  - finding a proper way to create a space with as much information as the one examined
  - finding a way to give toy data a "similar" correlation structure

- Ultimately, an absolute evaluation of a trial-factor-corrected Z value remains unnecessary for the algorithm to be meaningful

Seeding recipe

1) Random box: choose two numbers in [0,1] per each dimension → initial volume is V=$(1/3)^N$

2) Kernel density: Substitute each data point with a (wide) multi-D Gaussian density, compute sum of densities, get

$$\arg max_{x_j} \sum_{i=1}^{N} G(x_j - x_i)$$

then pick initial box boundaries as [x_j-k, x_j+k]
- k=0.2 reasonable choice for the cases explored so far

3) Find box boundaries with clustering routine
→next slide

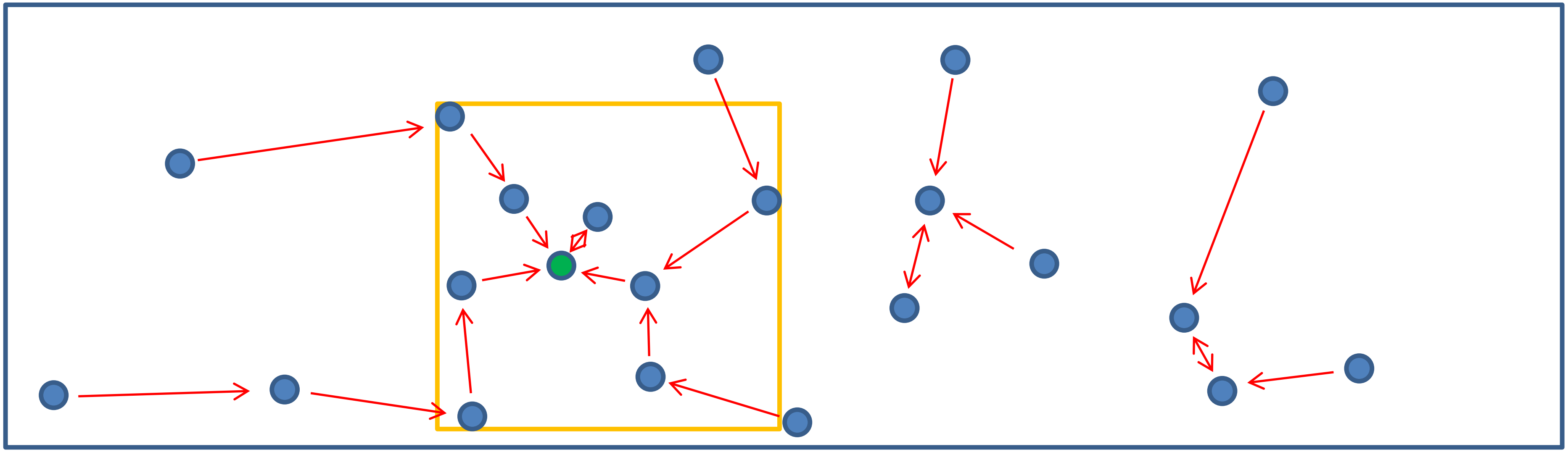
Clustering for seeding

We are searching for a local overdensity, so it looks reasonable to seed the box search with the point where several data events are close in space.

Routine:

1) for each data event i, find list of other data events $j_1 \ldots j_N$ that have i as closest (relatives); **identify event i*** which has maximum number of relatives
2) for each relative of i* $j_1 \ldots j_N$, find list of data events that have j_k as closest (relatives of j_k)
3) Draw **minimum box in multi-D** which includes all second-order relatives of i*

Many variants investigated – some perform similarly to this one



Variants of metric in cluster definition

The clustering described in previous slide works by finding "closest" points

- Given that one wants high density and maybe only a fraction of the subspace features are significant to isolate a signal, several possible variants come to mind to define the distance of x and y in D"-dimensional subspace

1) Euclidean, but only consider the half of coordinates where points are closest
1b) Same, but use all D" dimensions
2) Linear sum of the smallest half of absolute values of coordinate differences
2b) Same, but use all D" dimensions
3) Minimum volume of box containing x,y points in half of subspace dimensions
3b) Same, but using all D" dimensions
4) Largest among the D" absolute values of x-y coordinate differences

Although the choice is certainly application-specific, in my studied use cases I found that consistently 3 > 1 > 3b > 1b > 2 > 2b > 4

→ **use recipe 3**