Clustering for interpreting complex high-energy physics models

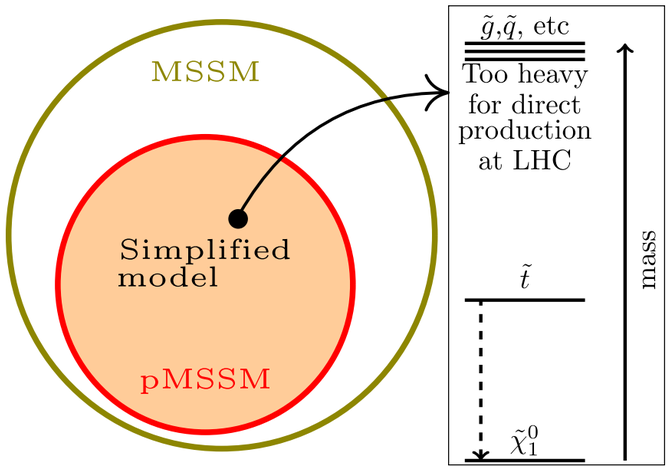
Walter Hopkins, Taylor Childers, Evangelos Kourlitis, Arindam Fadikar

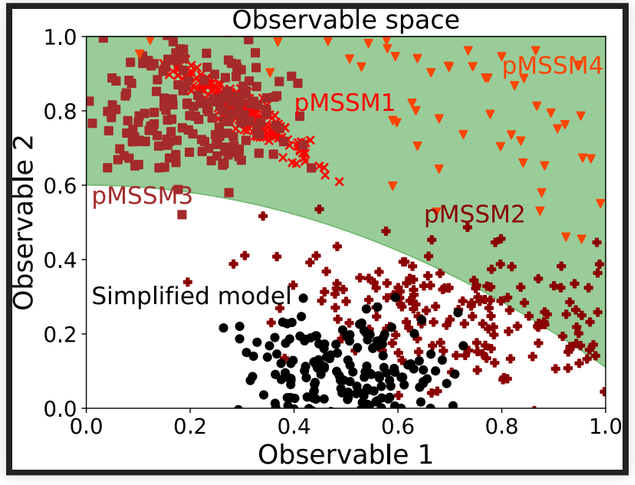
IML meeting

May 11th, 2022

Argonne NATIONAL LABORATORY

NEW PHYSICS SEARCHES AT ATLAS

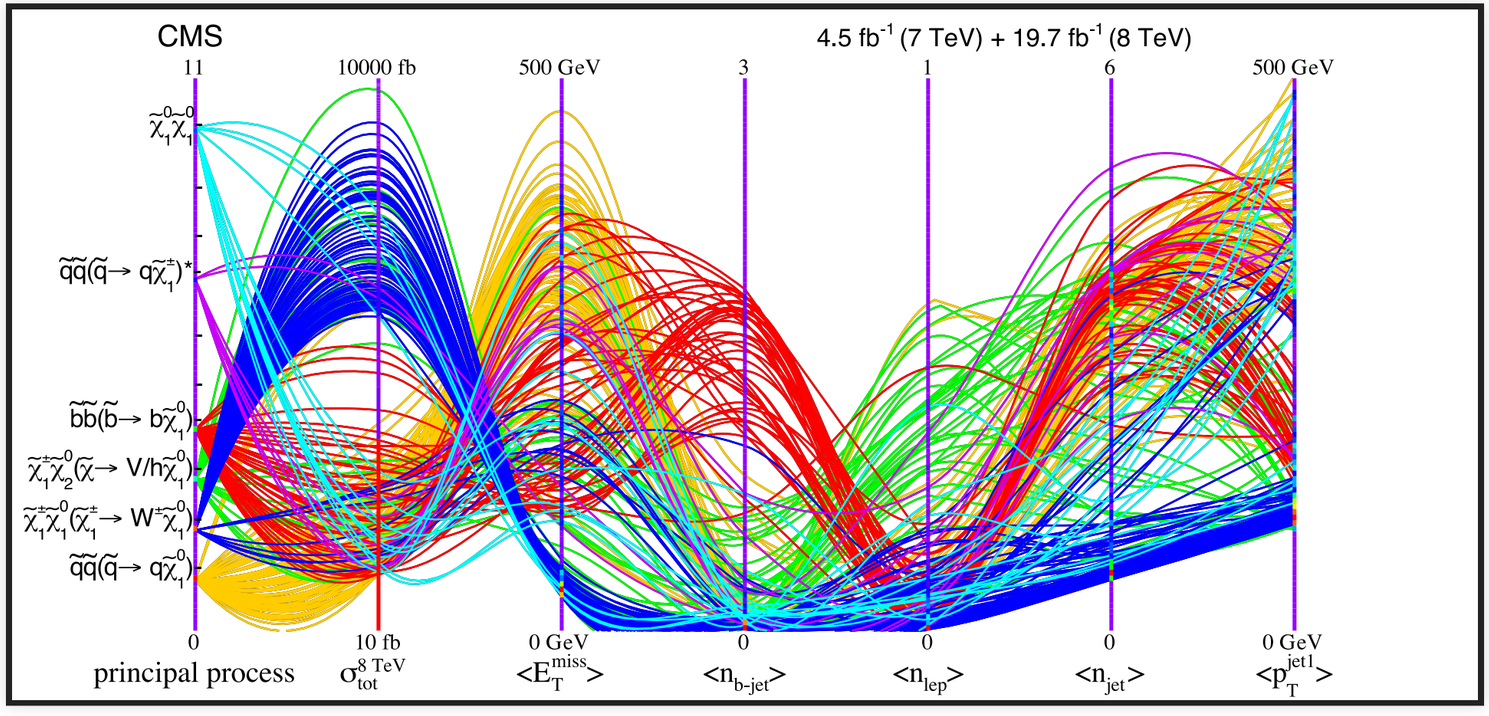




- ATLAS has been searching for physics beyond the Standard Model (SM).
- Most new physics searches use simplified models: only make 2-3 new particle masses accessible at the LHC.
- Simplicity allows us to get a grasp on new signatures that may show up at LHC.
  - But maybe this simplicity is causing us to miss something...
- Broader models exist, e.g. pMSSM with 19 parameters.
  - But how do we know how the models show up in our detector as we scan 19 parameters?
  - Different configurations of the 19 parameters can lead to the same detector signature.

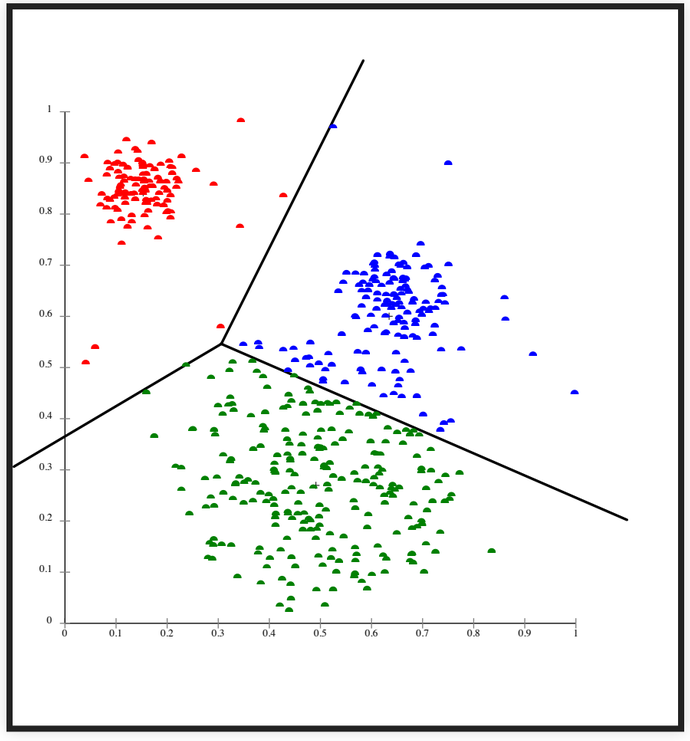
RUN 1 PMSSM SCANS: SURVIVING MODELS

- CMS interpretation of surviving models: averages of observables for surviving models.
- Good idea but still difficult to interpret.
  - Common problem: too many dimensions and models for us to digest.



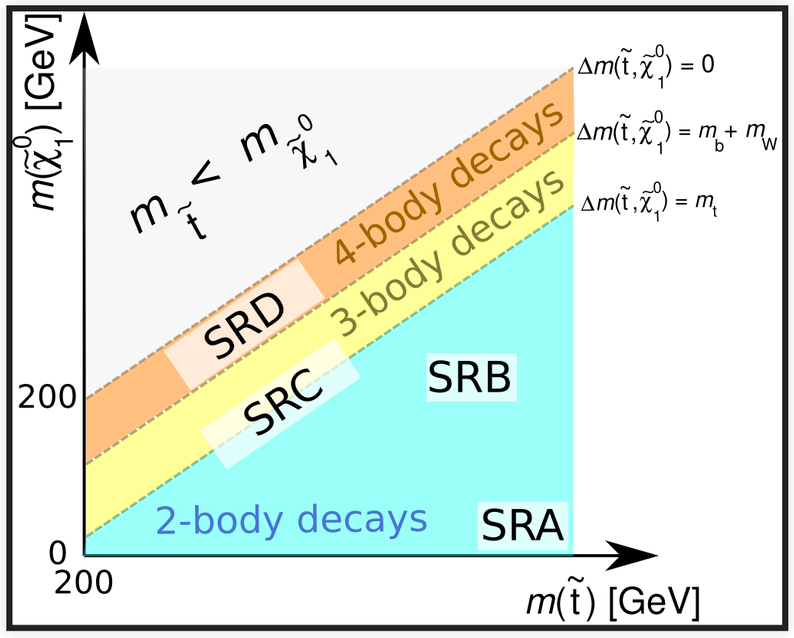
CLUSTERING

- Clustering groups similar data points in a high-dimensional space.
  - Various distance quantities can be used to determine whether points are close.
- Clustering algos are unsupervised learning algorithm: no labels needed.
- Several flavors exist: k-means, hierarchical, density based, etc.
  - For now considering **k-means**.



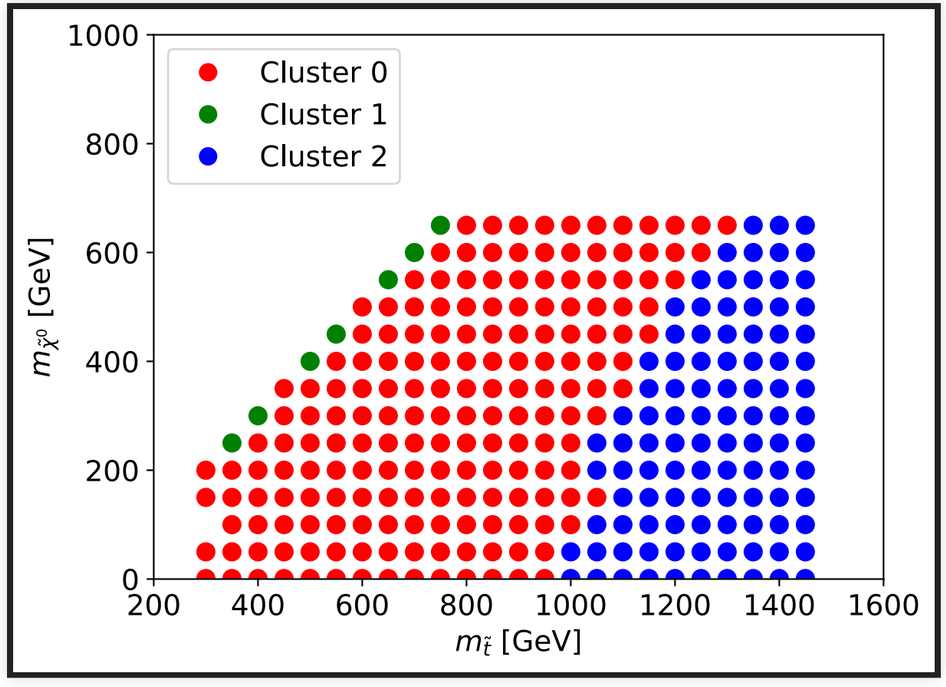
TEST: CLUSTERING SIMPLIFIED MODEL

- Even with simplified models, different regions of theory parameter space result in different observables.
- Example in stop search: compressed stop/LSP vs high $\Delta m(\tilde{t}, \tilde{\chi}_1^0)$.
- **Test clustering on whole stop grid with multiple variables with full event info and averages of quantities.**

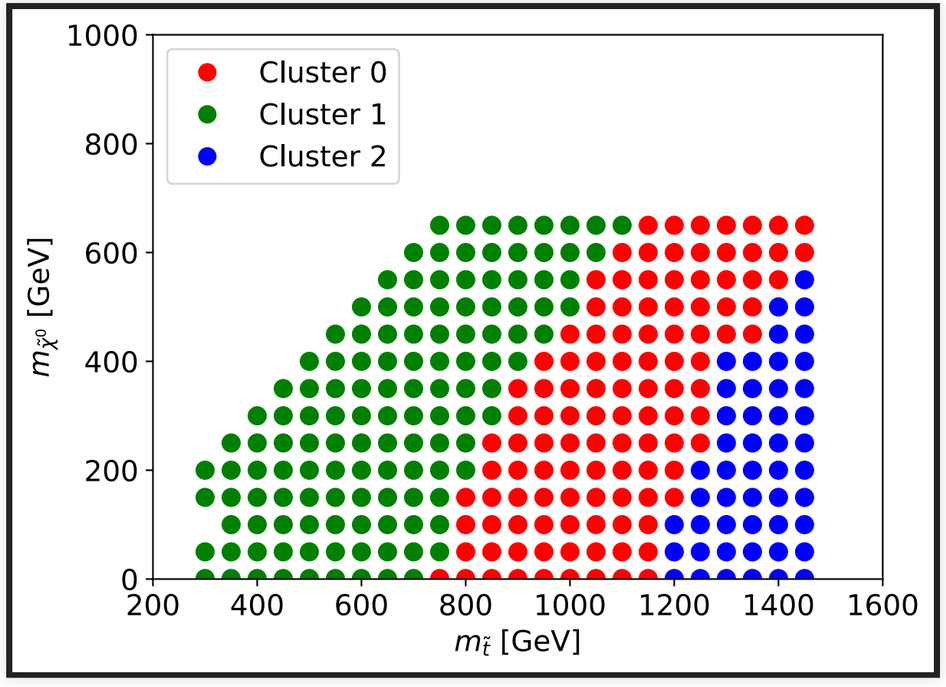


RESULTS: DIVISION OF SIMPLIFIED GRID

Full event info, $k = 3$

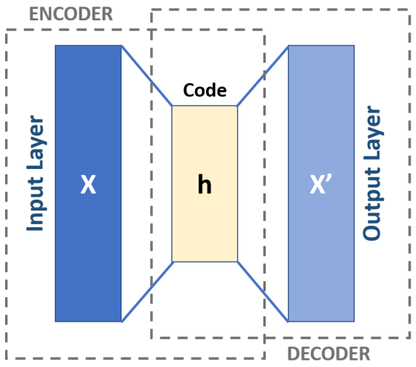


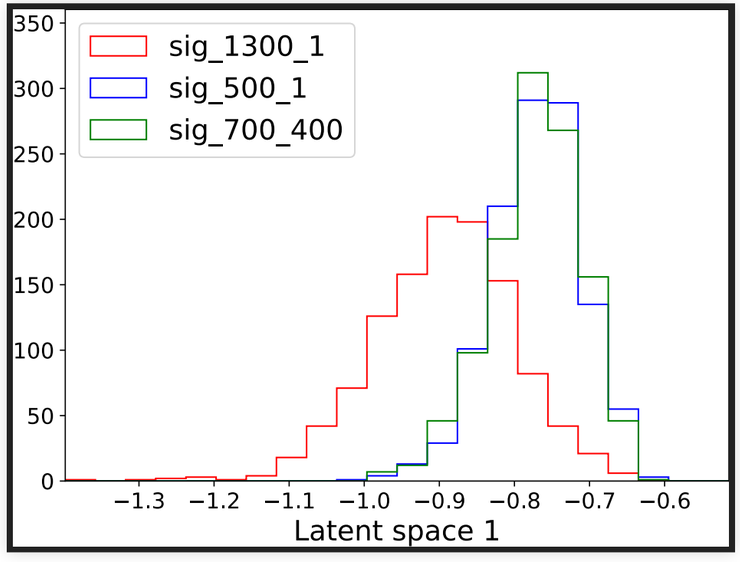
Averages, $k = 3$



REASONABLE GROUPING OF POINTS SIMILAR TO WHAT WAS DONE MANUALLY

R&D: DIMENSIONALITY REDUCTION

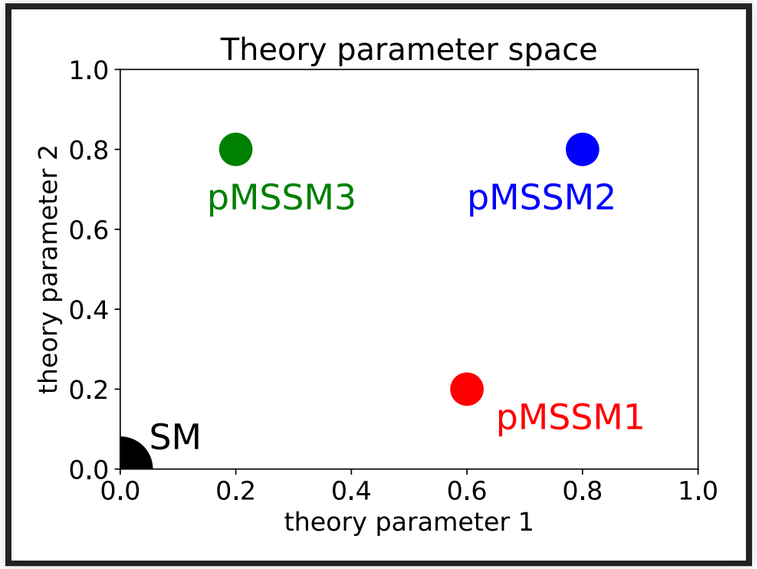


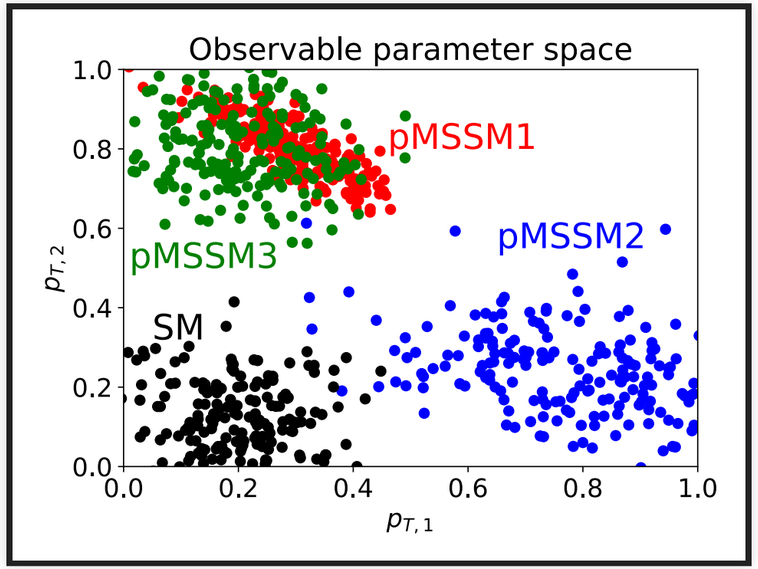


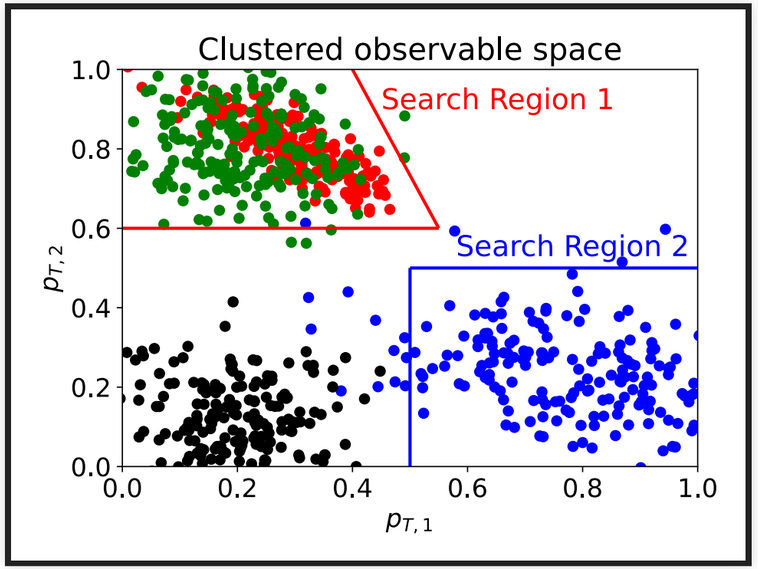
- The data has low-level (particle momenta), correlated features.
- Higher dimensionality reduces power of clustering.
- Would like to have an algorithm construct pertinent features of data set in an unsupervised manner.
- **Autoencoders** can reduce dimensionality of observable space.
- Apply clustering (e.g., k-means) in lower dimension (latent) space.

BACKUP

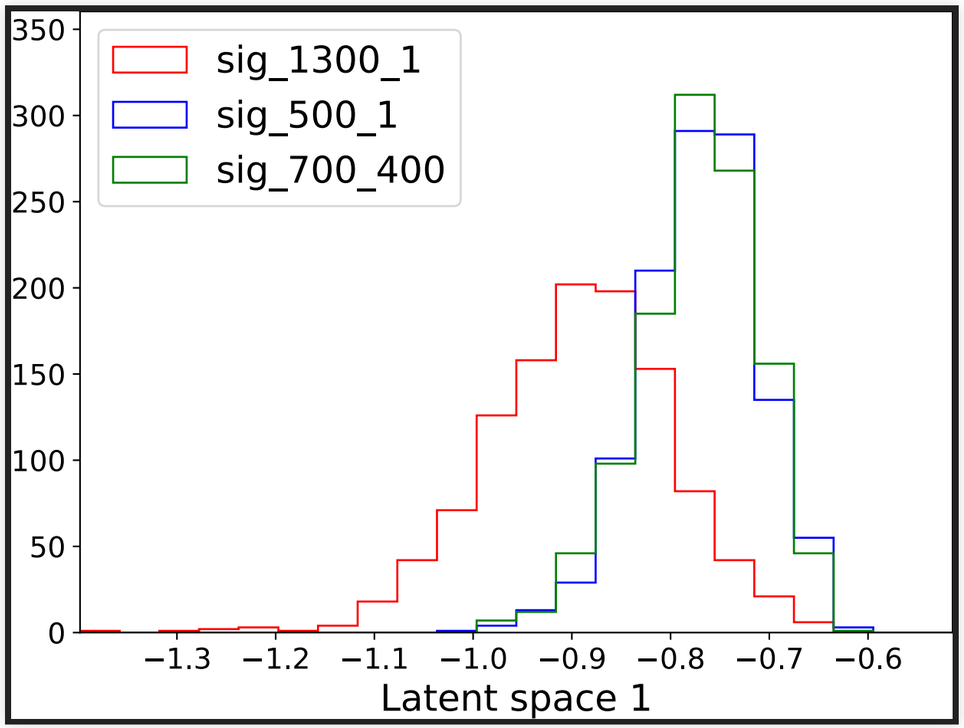
SKETCH OF CLUSTERING WORKFLOW







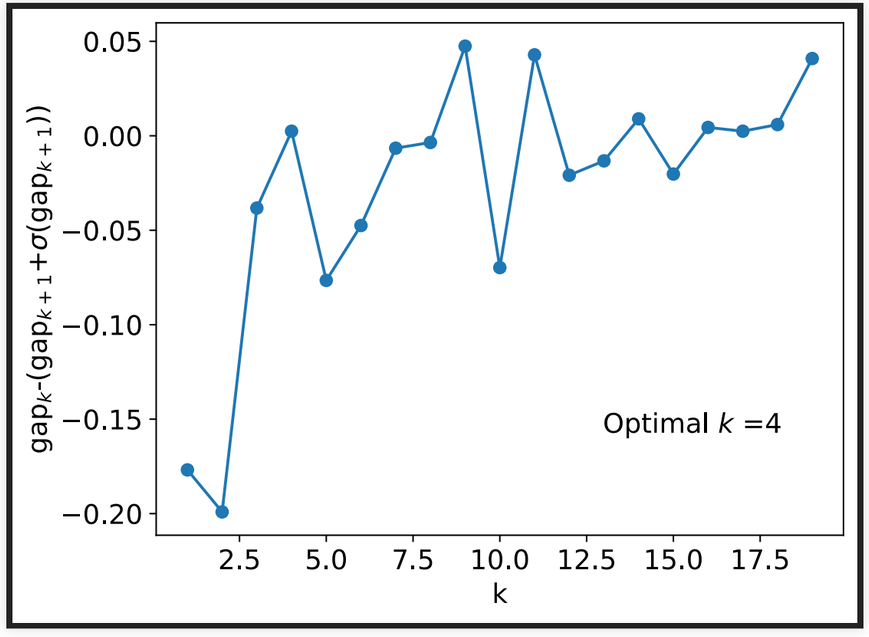
AUTOENCODER PRELIMINARY RESULTS



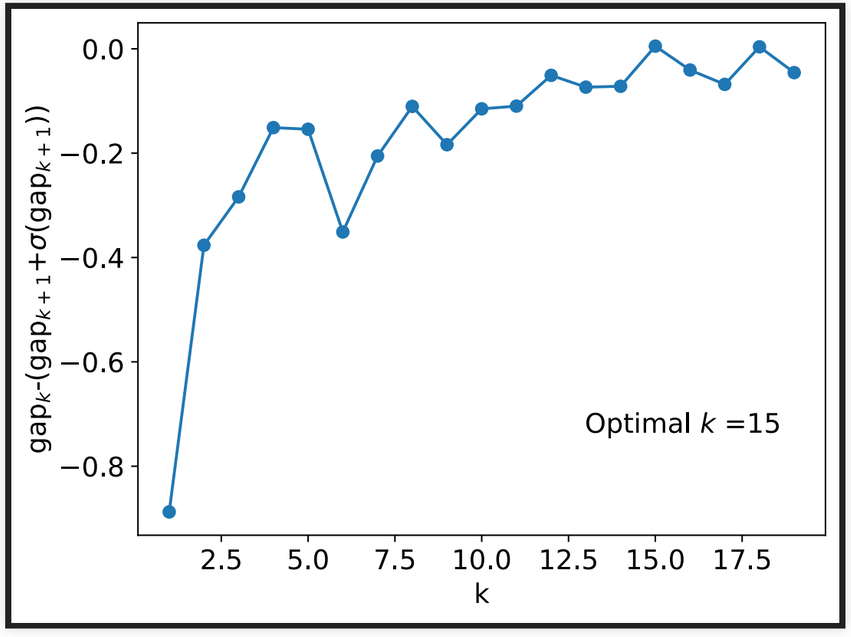
- Considered both a 1D and 2D latent space.
  - However, 2D space was just a line.
- Trained on full grid of samples.
- Couldn't apply clustering but visually inspecting latent space showed expected separation of samples.
- Training was a bit unstable: network sometimes learned the means (due to using MSE loss).

RESULTS OF SCAN OVER k

Full event info



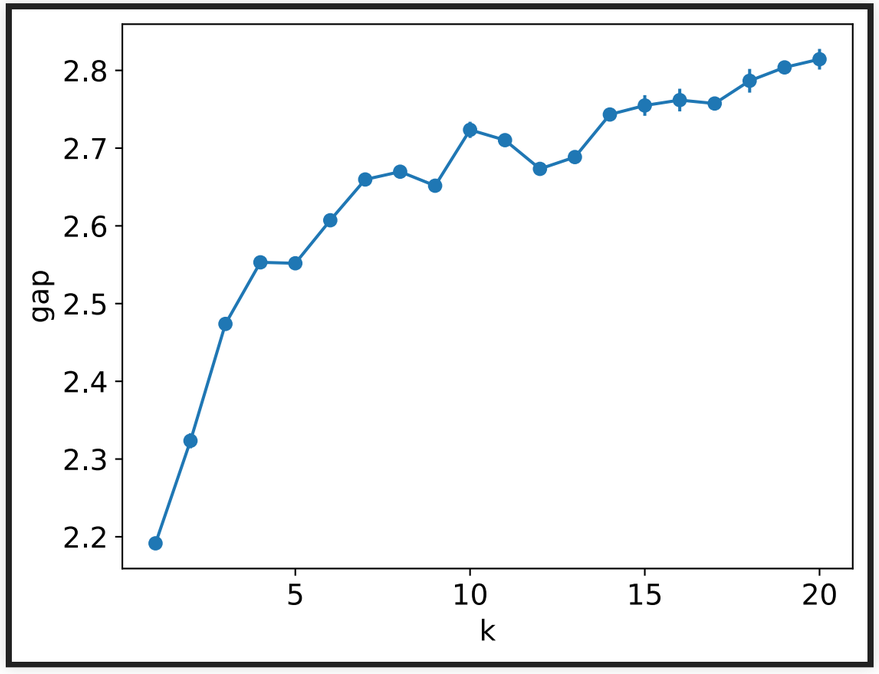
Averages

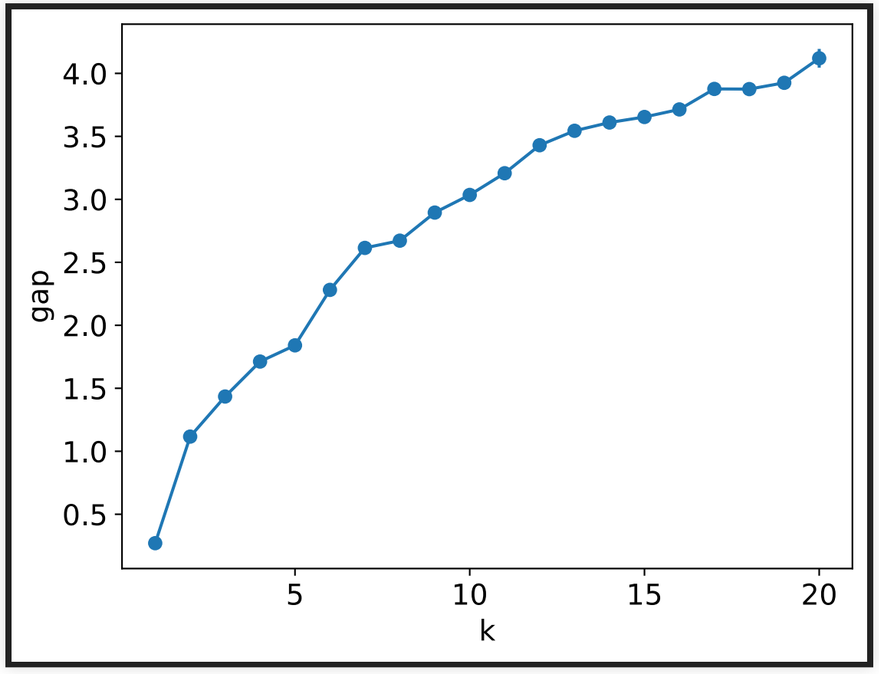


Best k is where gap$_k$-(gap$_{k+1}$-σ_{k+1}) > 0 with σ = error on the gap stat.

This occurs at $k = 4$ ($k = 15$) for event (average) info but gap$_k$-(gap$_{k+1}$-σ_{k+1}) doesn't actually change much after $k \sim 3$.

GAP STATISTIC





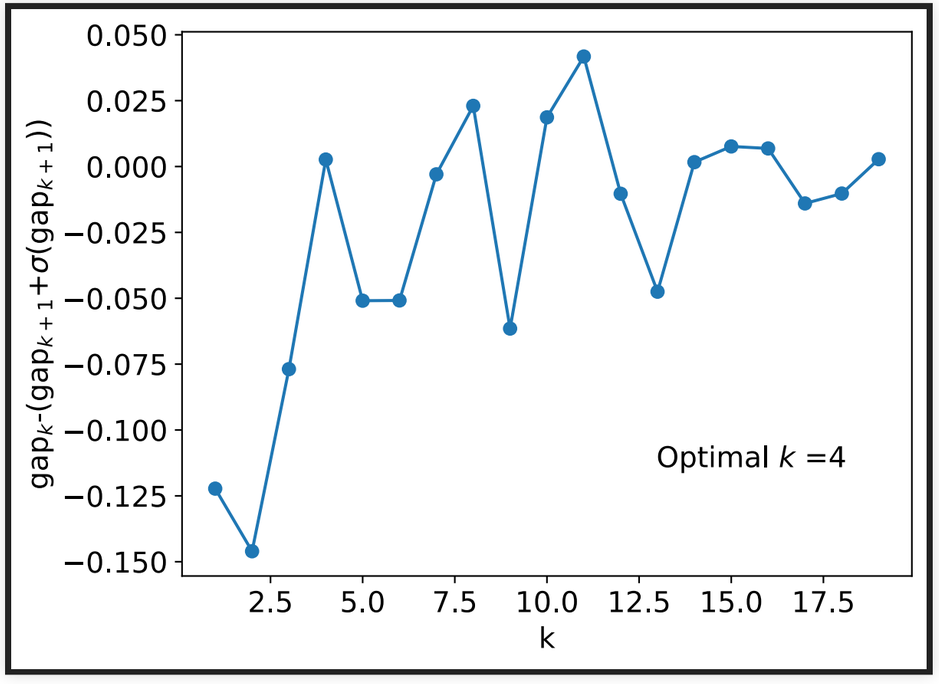
Best k is where gap(k)-[gap(k+1)+σ(k+1)] > 0 with σ = error on the gap stat.

SET UP

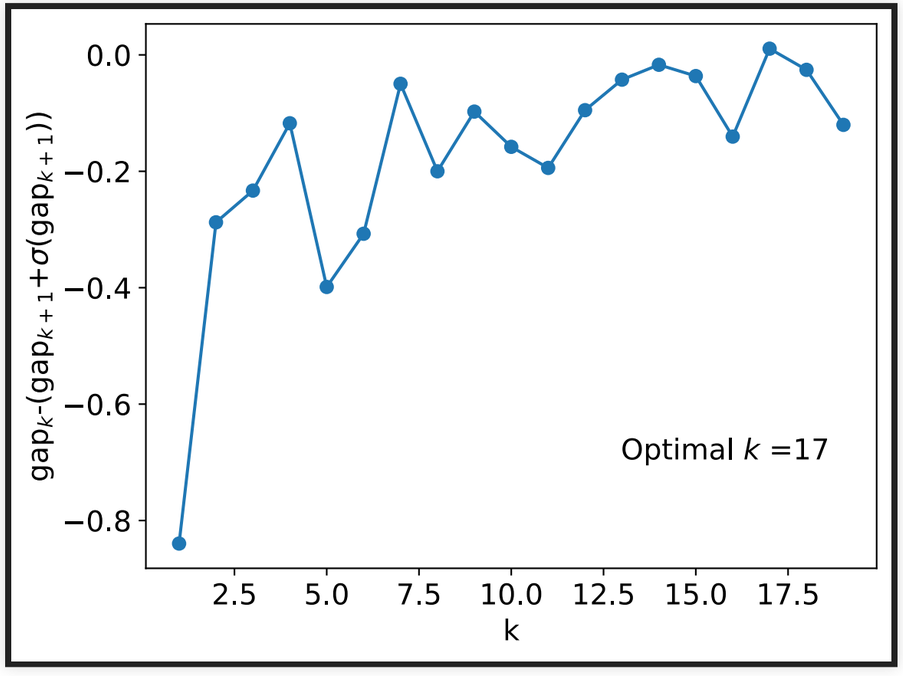
- Used signal grid similar to what was used in ATLAS tt+$E_{\mathrm{T}}^{\mathrm{miss}}$ search (but reproduced the samples with MG+Pythia+Delphes due to holes in grid).
- No preselection applied and used simple input variables: MET, HT, leading four jet pTs, jet multiplicity
- Inputs were scaled to have range from 0 to 1 for all samples together (individual signal points could have different ranges).
  - Leaving the shapes as-is.
- Clustered using all events.
- Clustered using averages → thus you have one value per variable per signal grid point.
  - Helps makes this more easily scalable if we have 10^4 to 10^5 models.
- Best clusters chosen to show results:
  - Best means cluster with highest signal efficiency for that model.
  - This can result in clusters that are never the best... not a problem we just ignore these.

GAP STATISTIC: DIFF

Full event info, best $k = 7$



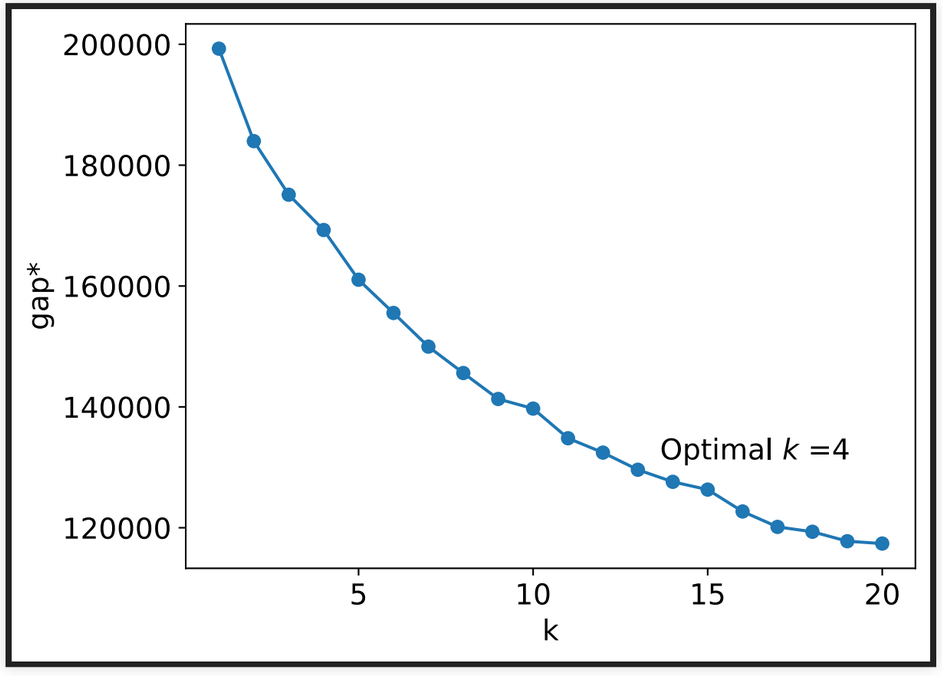
Averages, best $k = 8$



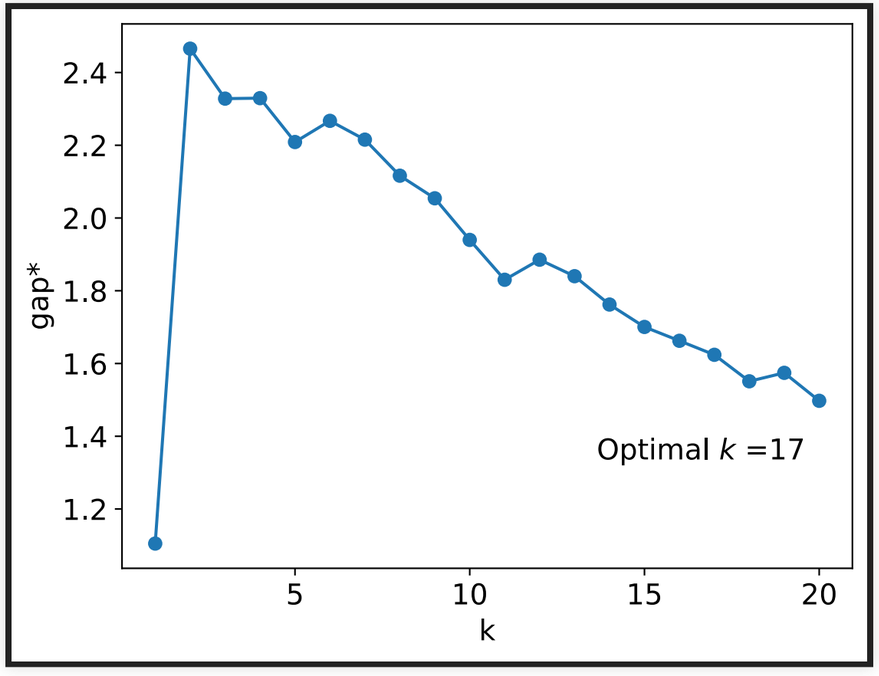
Best k is where gap(k)-[gap(k+1)+σ(k+1)] > 0 with σ = error on the gap stat.

GAP* STATISTIC

Full event info, best $k = 1$



Averages, best $k = 3$



Best k is at max gap*.