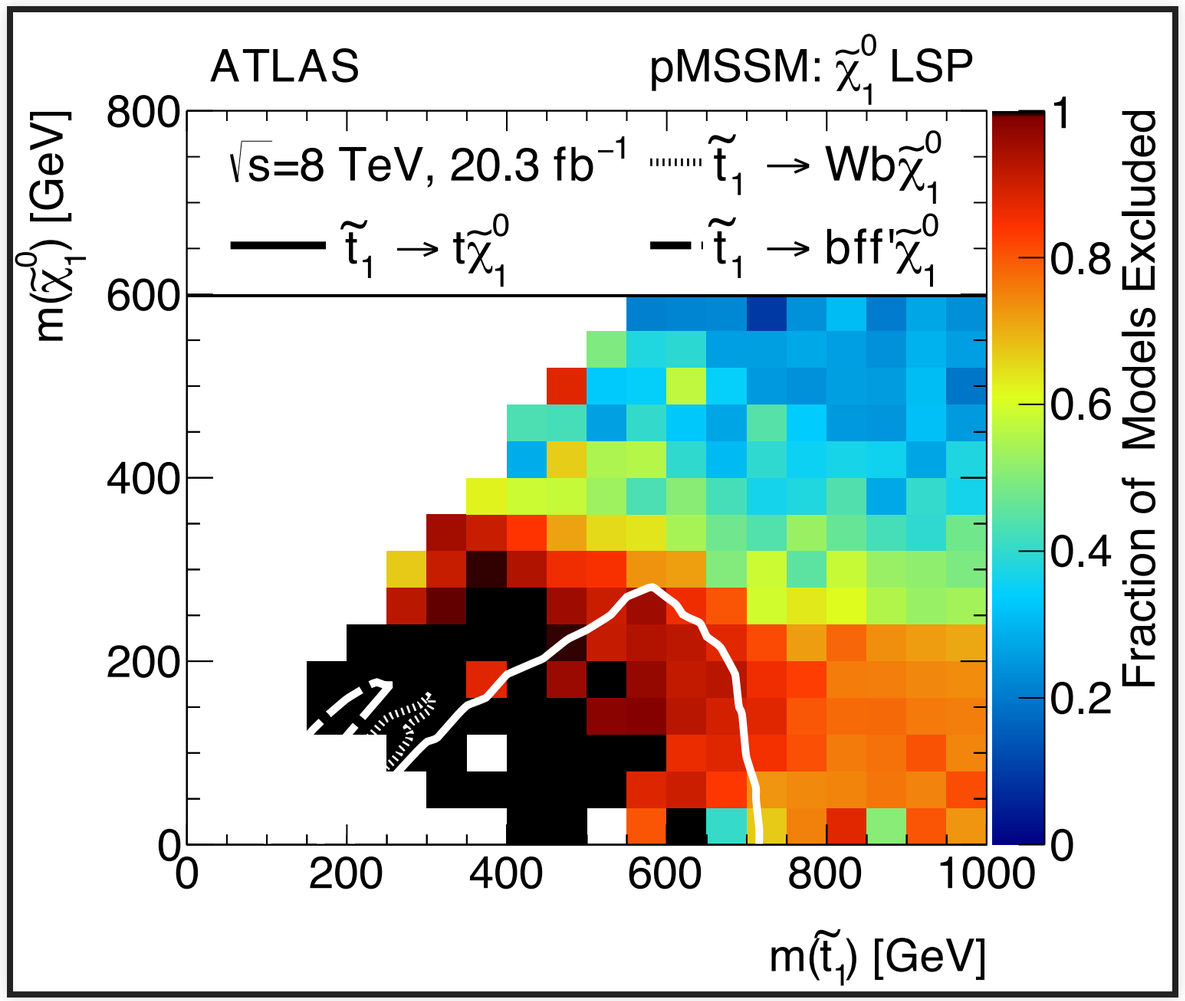
Interpreting non-excluded pMSSM models with clustering

Walter Hopkins, Taylor Childers, Vangelis Kourlitis, Arindam Fadikar

pMSSM scan meeting

March 9, 2022

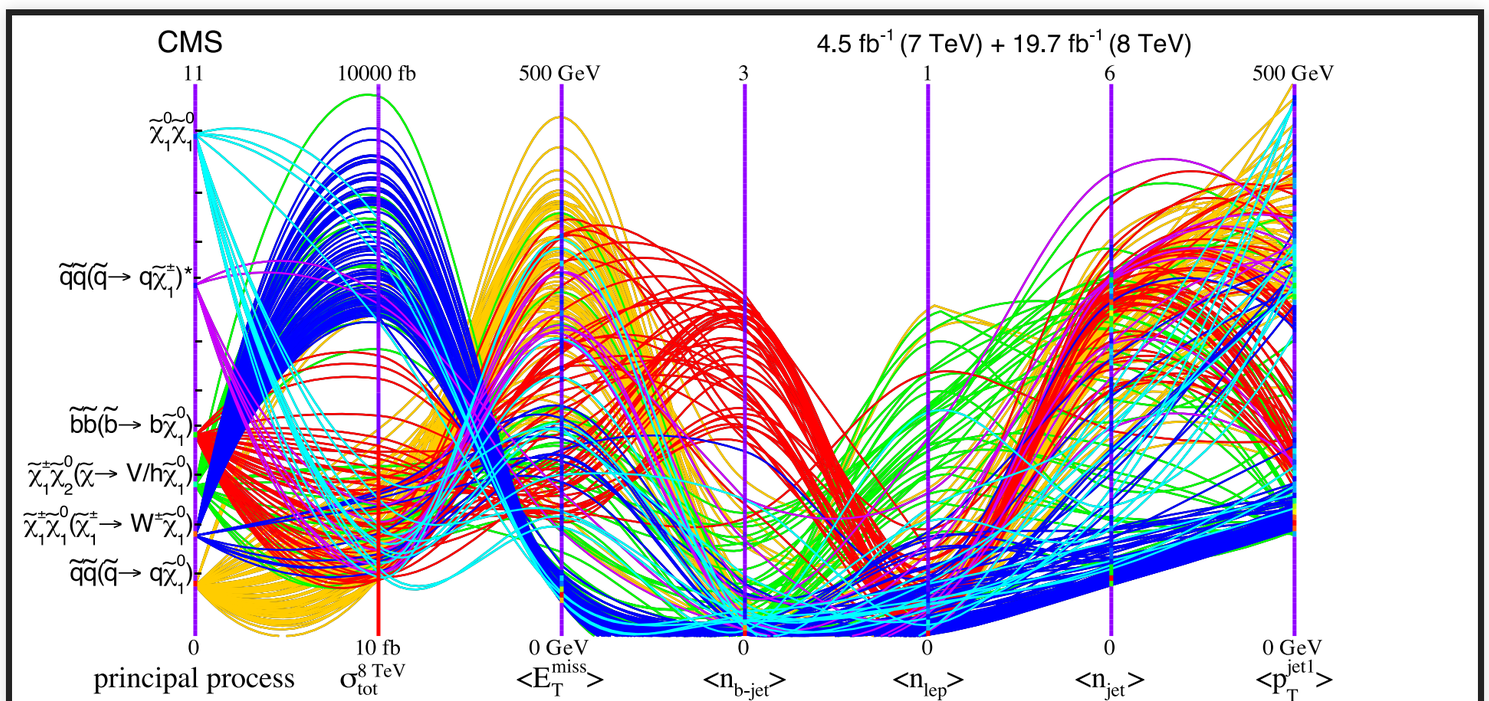
RUN 1 PMSSM SCANS



- ATLAS and CMS sampled 19 parameters of pMSSM to find models that were not excluded.
- Manually inspected models that survived.
  - Some survived due to long cascades and compressed spectra.
  - Cumbersome to really understand what region of observable space the models end up in.

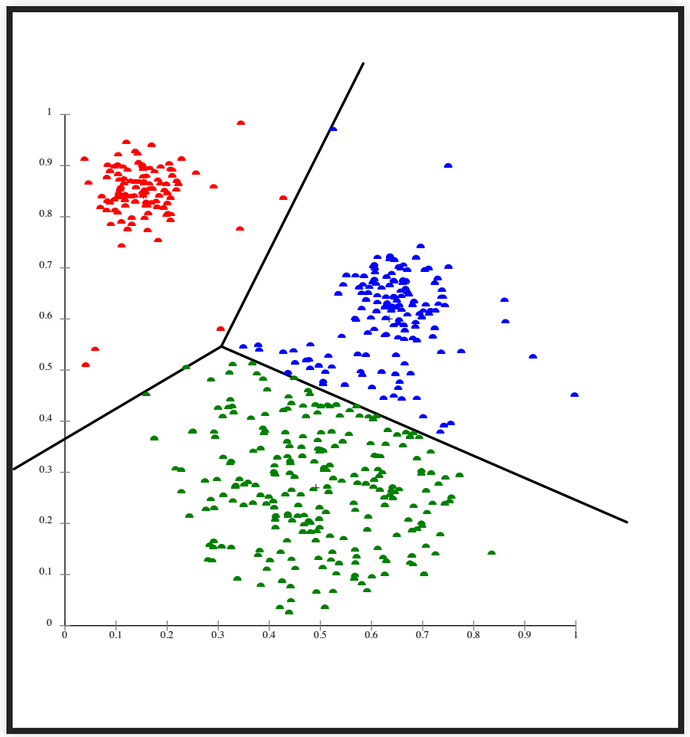
SURVIVING MODELS IN RUN 1

- 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.

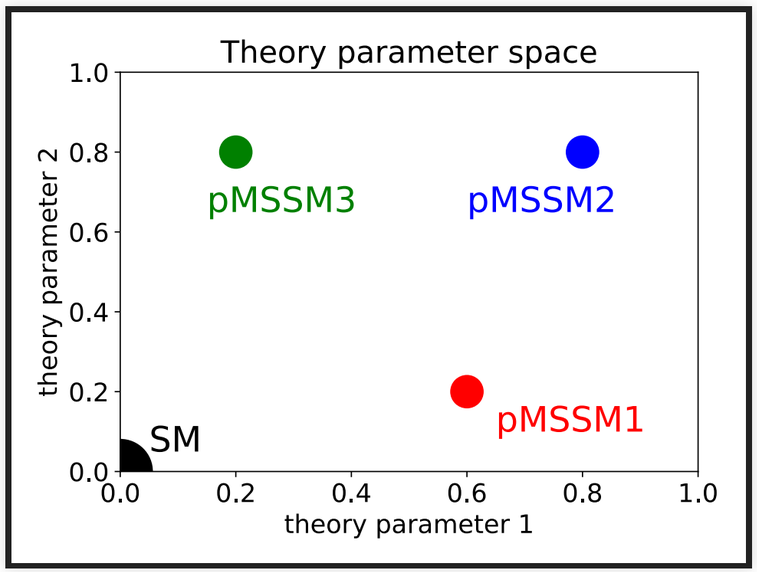


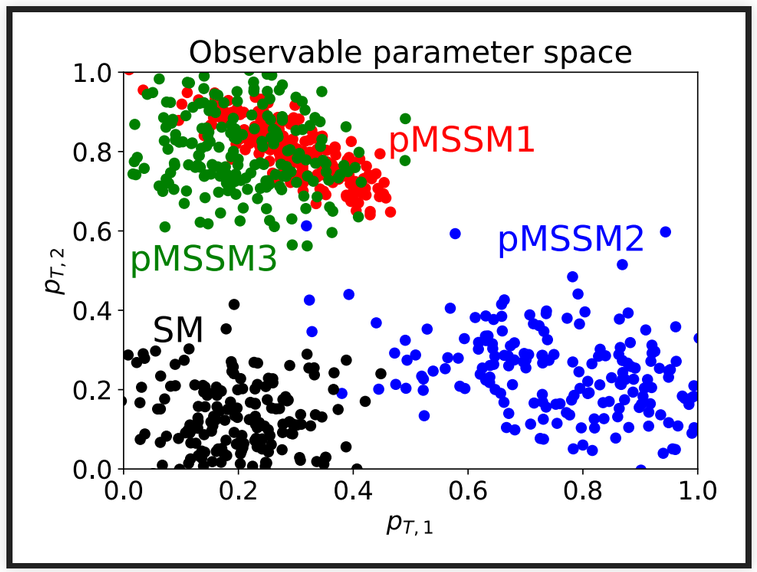
CLUSTERING

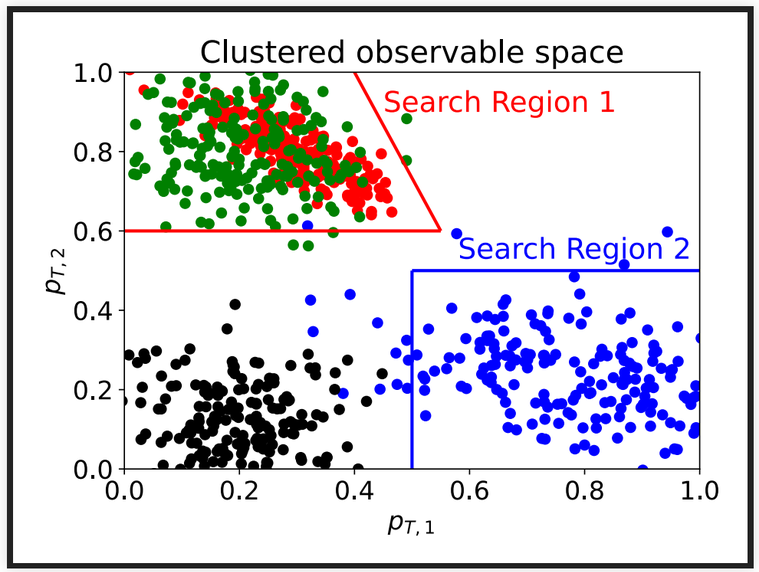
- Clustering groups similar data points in a high-dimensional space.
  - Various distance quantities can be used to determine whether points are close.
- Clustering algos are unsupervised learning algorithm: no labels needed.
- Several flavors exist: k-means, hierarchical, density based, etc.
  - For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



SKETCH OF CLUSTERING WORKFLOW





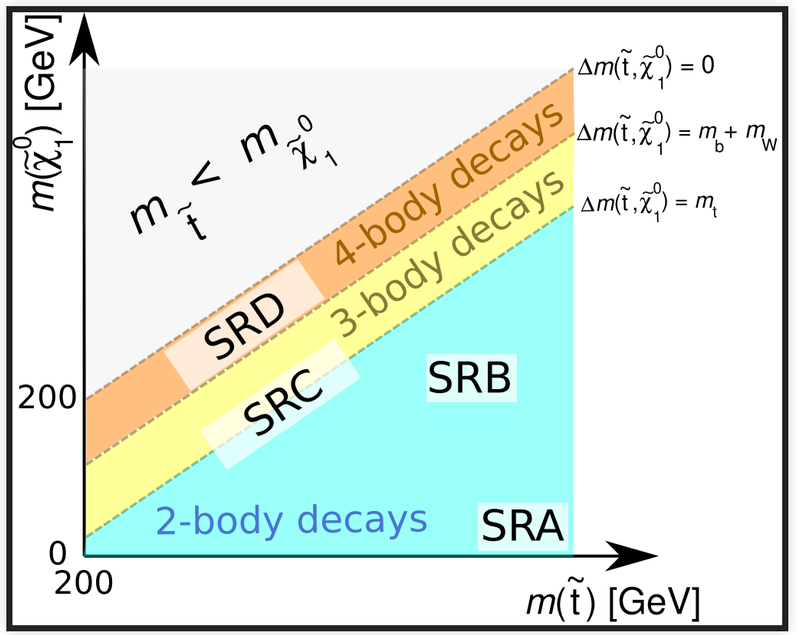


k-MEANS

- Used k-means clustering: minimizes the within-cluster variance.
- k-means doesn't give you the number of clusters (k). You have to specify the number.
  - But we of course don't know k: loop over k and determine which is best.
- Common figure of merit for which k is best: gap statistic.
  - gap statistic compares clusters from data with clusters from uniform pseudodata.
  - For gap stat: k for which gap_k-(gap_{k+1}-σ_{k+1}) > 0 is optimal.
  - More info on gap statistic is here and on an alternative (gap*) is here.
- There is no absolute "right" k.
  - Just as when designing signal regions, one could use more or less bins in, for example, MET with varying but similar significance.
  - Our aim is to identify rough regions in observable space: corresponds to larger bins that we can study further.

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.**

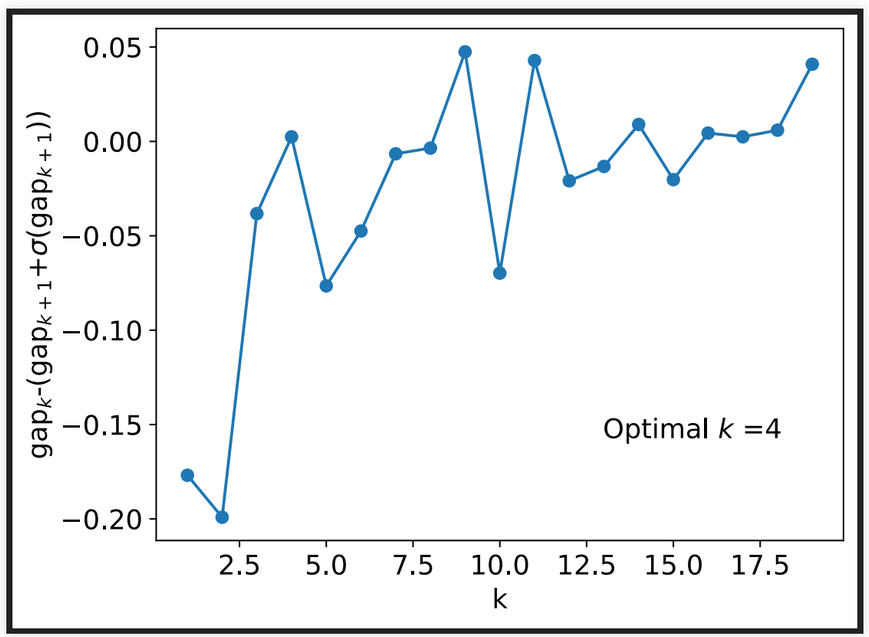


SET UP

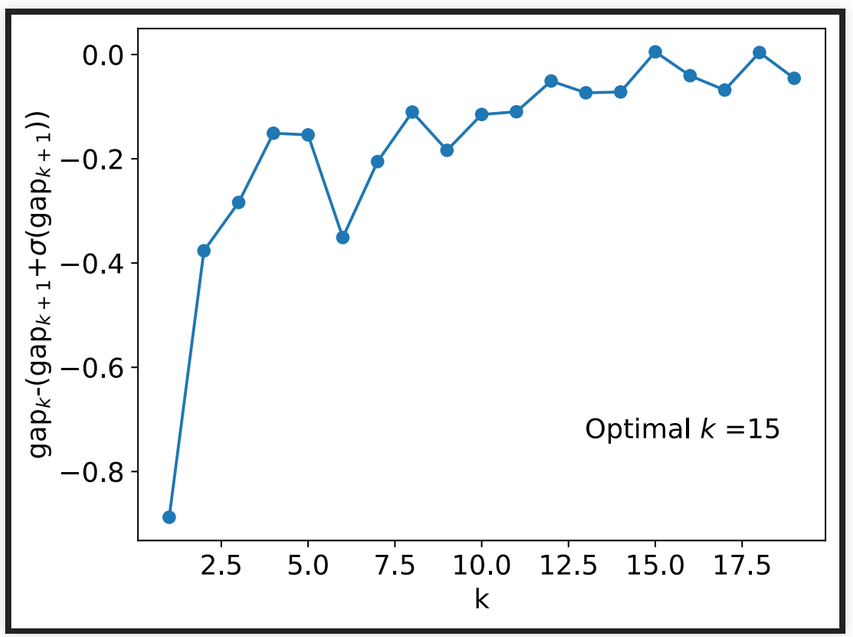
- Used signal grid similar to what was used in ATLAS tt+$E_{\mathrm{T}}^{\mathrm{miss}}$ search (but reproduced the samples with MG+Pythia+Delphes due to holes in grid).
- No preselection applied and used simple input variables: MET, HT, leading four jet pTs, jet multiplicity
- Inputs were scaled to 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: GAP_k-(GAP_{k+1}-σ_{k+1})

Full event info



Averages

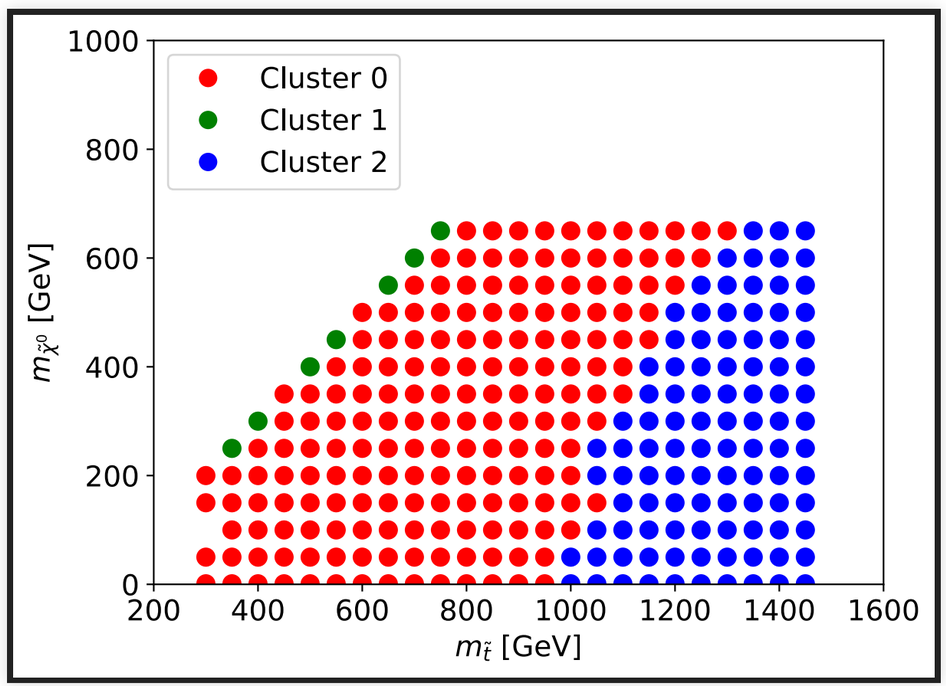


Best k is where gap_k-(gap_{k+1}-σ_{k+1}) > 0 with σ = error on the gap stat.

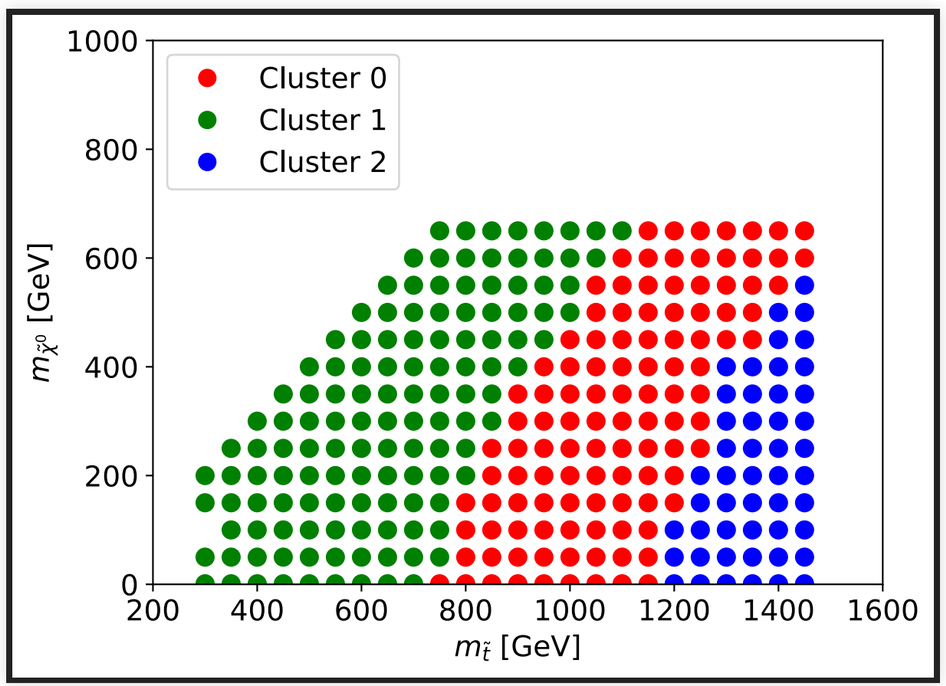
This occurs at $k = 4$ ($k = 15$) for event (average) info but gap_k-(gap_{k+1}-σ_{k+1}) doesn't actually change much after $k \sim 3$.

RESULTS: DIVISION OF SIMPLIFIED GRID

Full event info, $k = 3$



Averages, $k = 3$



Reasonable grouping of points similar to what was done manually

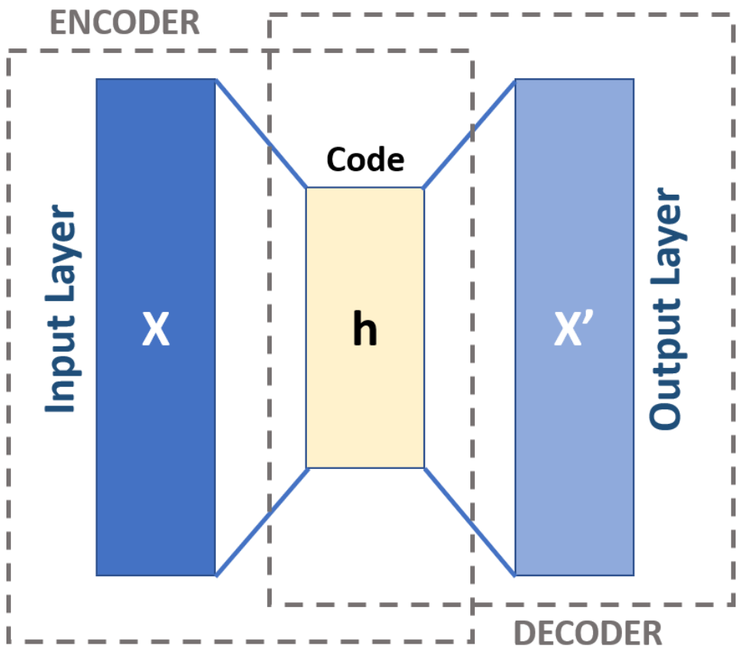
TECHNICAL NOTES: SCALING

- Using minibatch k-means: cluster on subset of data and update as you go through full dataset.
  - No problems with loading all the data to memory (a challenge with the pMSSM ntuples).
- Converting ntuples to scikit-learn friendly format needs to be scaled. Currently load full data set into memory which is not feasible for pMSSM.
- Preprocessing needs to work in batches. Will look for a way to do this.
  - Current 0-1 scaling is easy to implement but other methods more robust to outliers might require work.
- Calculating average (for input to k-means) could be done with ROOT quite easily.

TECHNICAL NOTES: PREPROCESSING

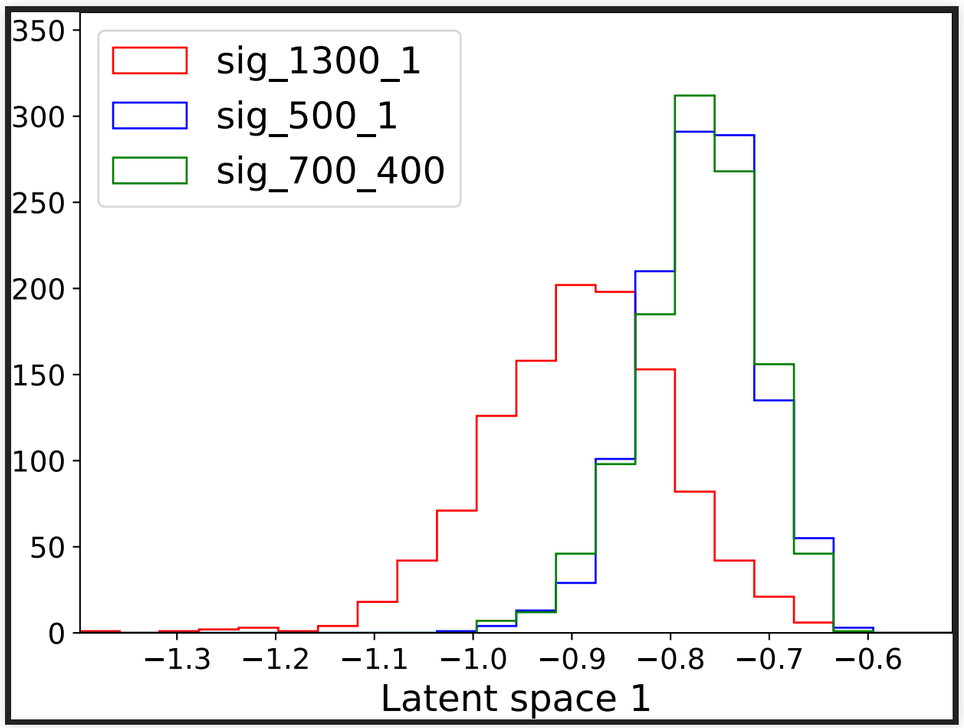
- Currently scaling dimensions to have range 0-1.
  - Not robust against outliers (do we really care about $\mathcal{O}(10)$ outliers in signal?)
  - Considering quantile scaler (make data uniform in all dimensions).
- k-means is not optimal when mixing input types (continuous vs n_bjets, n_leptons, etc. which is more like categorical data). Maybe try k-mediods.
  - Not really a problem when using averages since discrete variables become more continuous.

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.

AUTOENCODER PRELIMINARY RESULTS



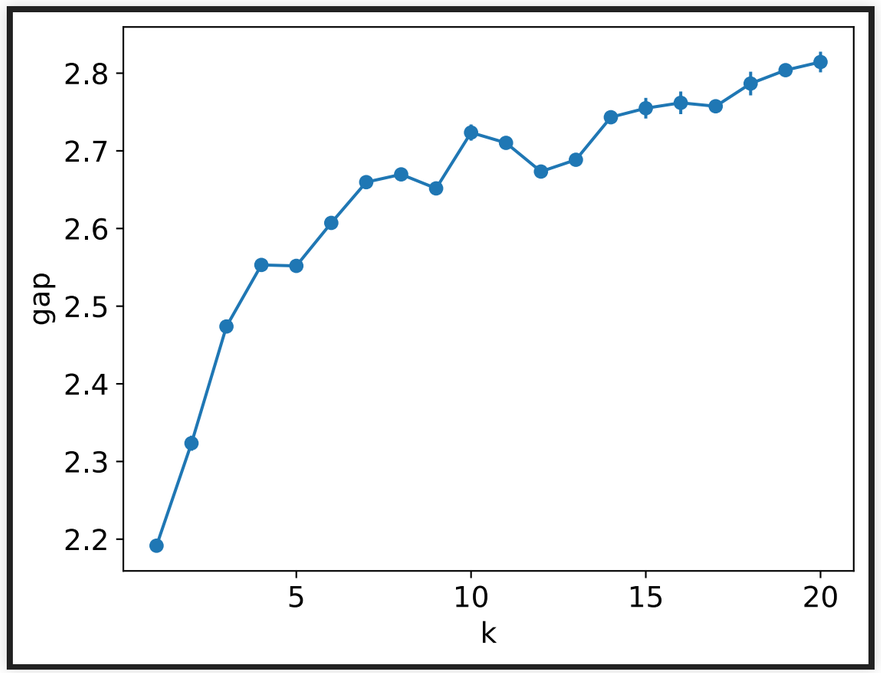
- Considered both a 1D and 2D latent space.
  - However, 2D space was just a line.
- Trained on full grid of samples.
- Couldn't apply clustering but visually inspecting latent space showed expected separation of samples.
- Training was a bit unstable: network sometimes learned the means (due to using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

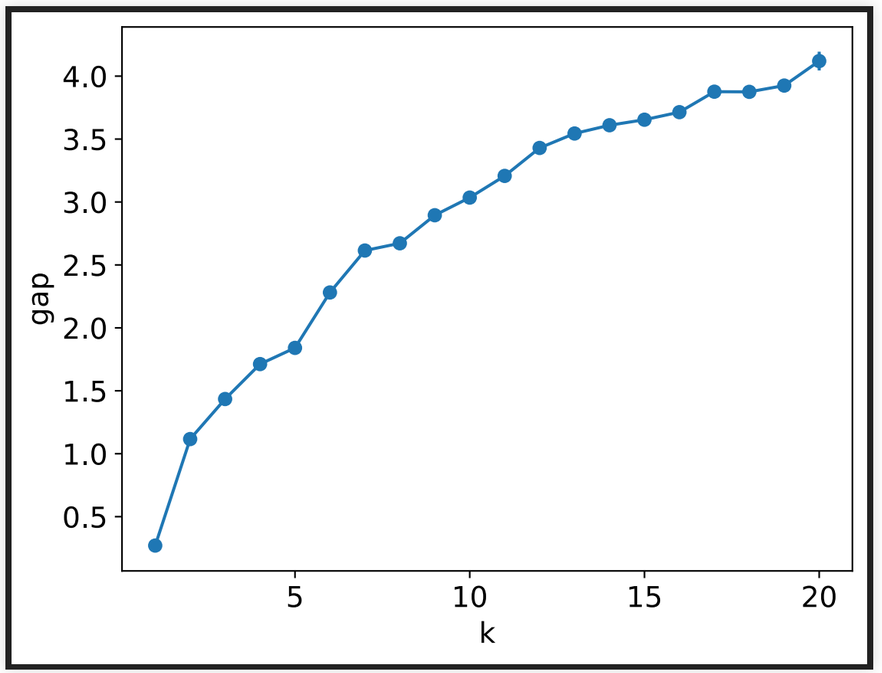
CONCLUSION/NEXT STEPS

- k-means with gap yields reasonable results for a simplified model grid.
  - Both with full event info and averages (averages makes processing much easier).
  - The division of the stop grid is sensible and does not have to "optimal"
- Depending on the scale and output format of the Snowmass pMSSM scan, some technical work may be needed.
  - Ntuple conversion to scikit-learn friendly format. Chunking into memory.
  - Preprocessing (scaling from 0-1 or quantile scaling) in chunks.
    - 0-1 scaling is trivial to scale.

BACKUP

GAP STATISTIC

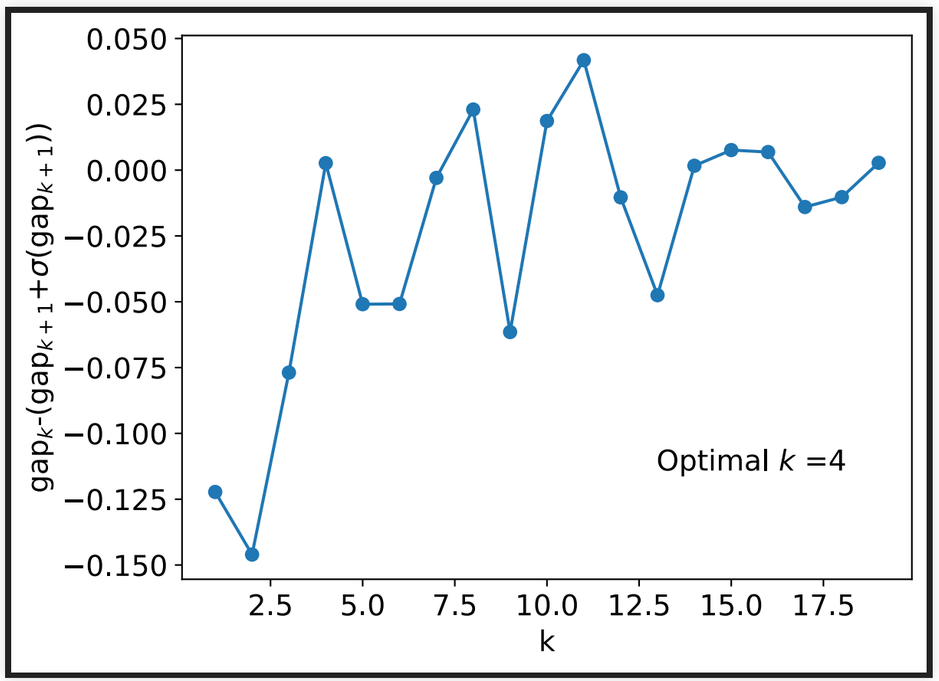




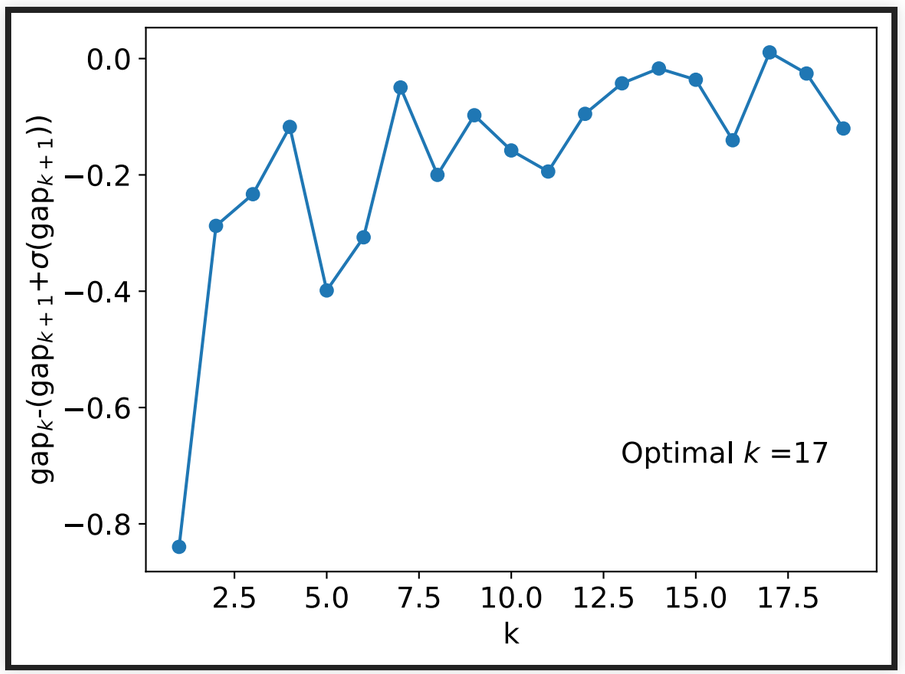
Best k is where gap(k)-[gap(k+1)+σ(k+1)] > 0 with σ = error on the gap stat.

GAP STATISTIC: DIFF

Full event info, best $k = 7$



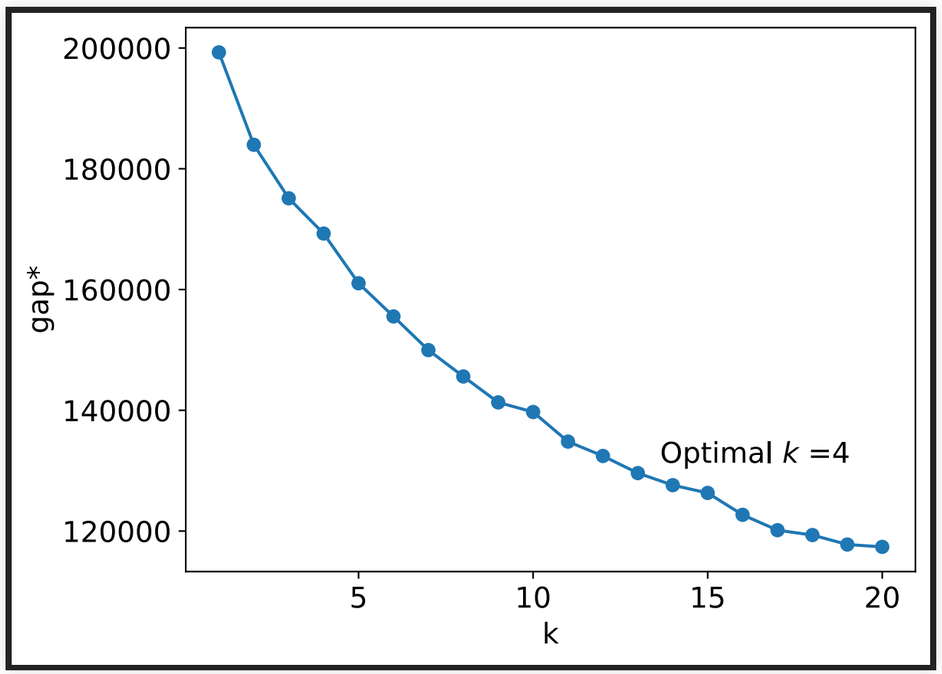
Averages, best $k = 8$



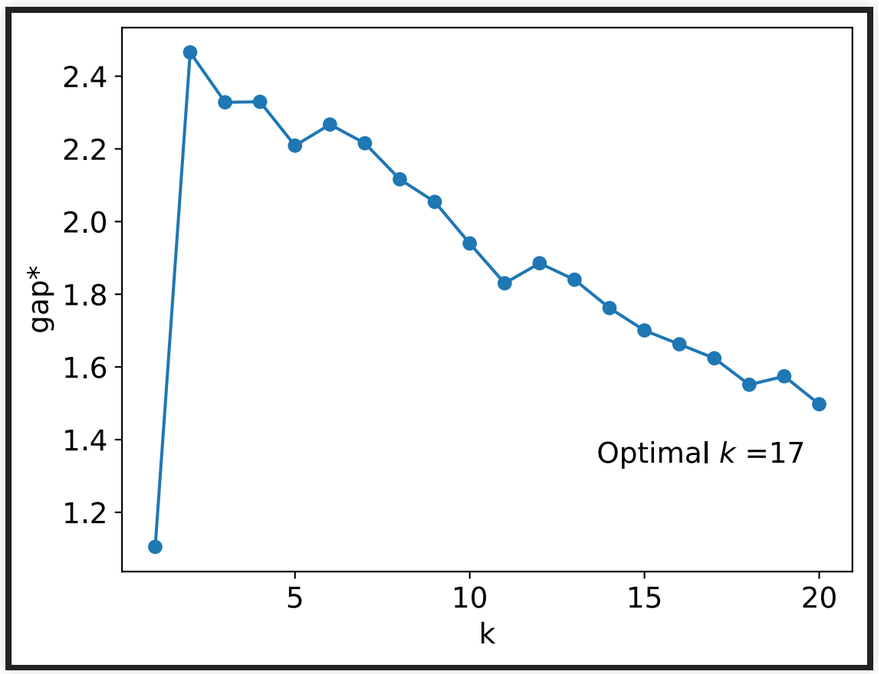
Best k is where gap(k)-[gap(k+1)+σ(k+1)] > 0 with σ = error on the gap stat.

GAP* STATISTIC

Full event info, best $k = 1$



Averages, best $k = 3$



Best k is at max gap*.