6-7-2018

First evaluation report

Esther Robb
Project Website: e-271.github.io

VIRGINIA TECH

Completed in the first month

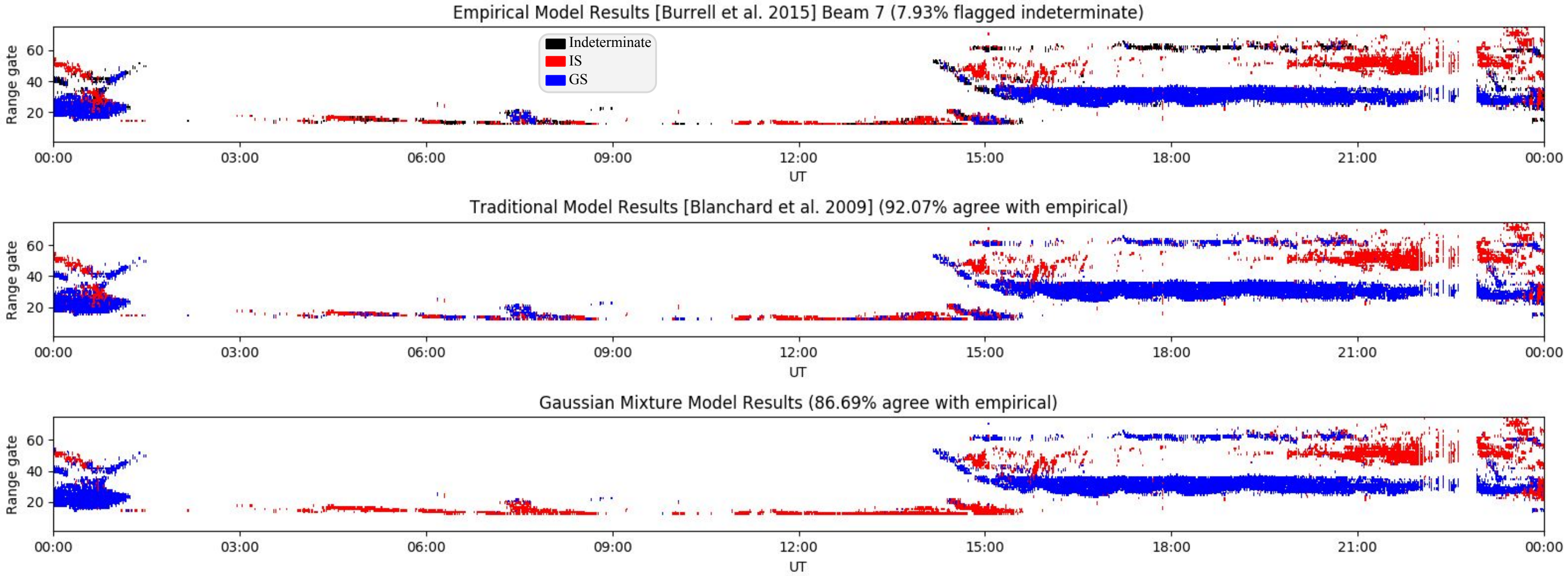
- Set up a project website and GitHub repository
  - e-271.github.io
  - https://github.com/vtsuperdarn/clustering_superdarn_data
- Create poster for presentation at 2018 SuperDARN Workshop
  - https://e-271.github.io/docs/robb_superdarn_clustering.pdf
- Test out high-latitude and mid-latitude radars to compare Gaussian Mixture Model performance
  - Performance is similar on the day we tried (2-7-18)
- Study statistics of the data
  - Some data does not appear Gaussian, but PCA transformation helps (?)
- Select a good model (using BIC and forward-selection)
  - BIC: Found that GMM full covariance is best
  - Forward selection: Preliminary results unclear, more research is needed

VIRGINIA TECH

Demo: Project setup and running scripts

Comparing mid-latitude and high-latitude

Saskatoon 2-7-18

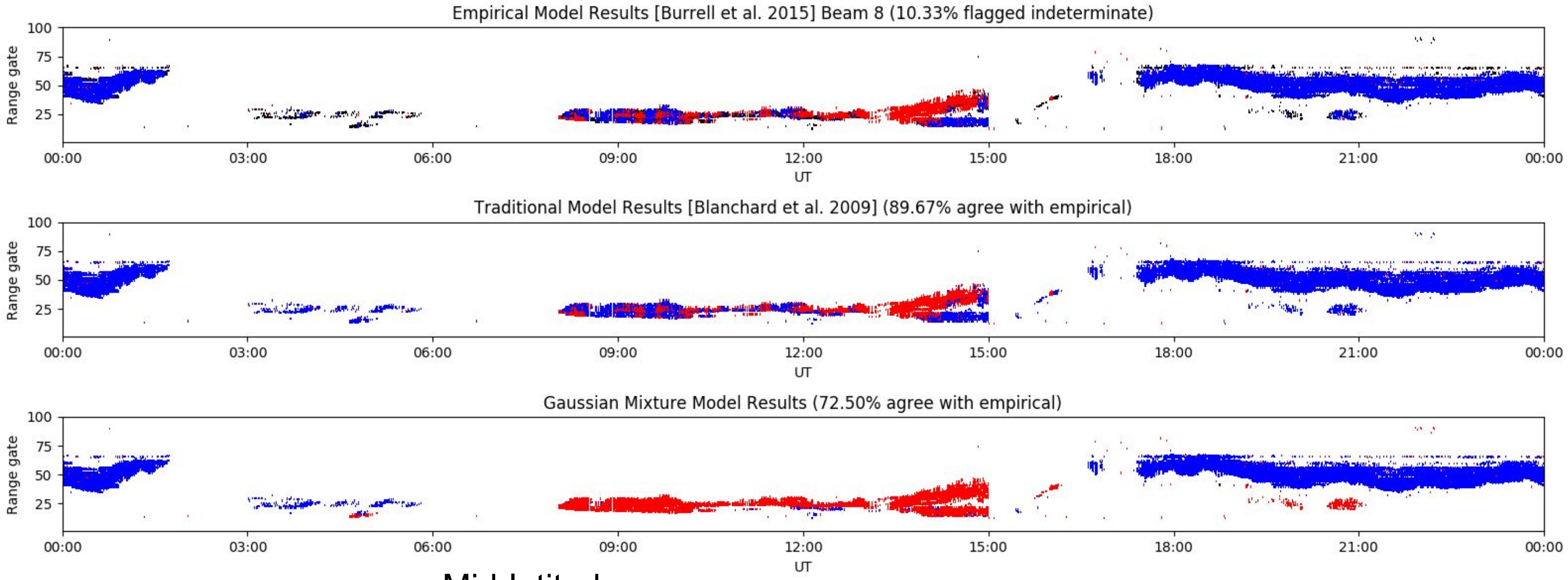


High-latitude

- GMM is doing a good job
- Outliers get clustered together and misclassified
- Need to find a way to better capture outliers

Comparing mid-latitude and high-latitude

CVW 2-7-18



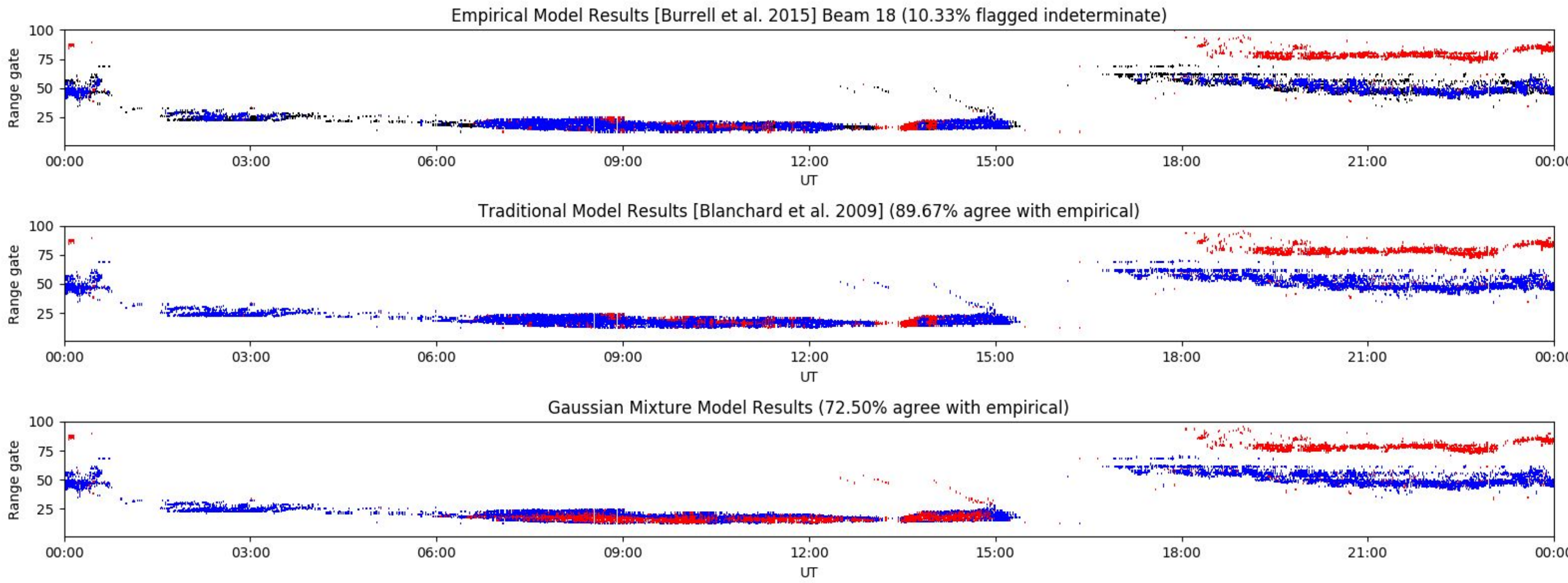
Mid-latitude

- GMM is doing a good job
- Outliers clustered together and misclassified
- Need to find a way to better capture outliers

VIRGINIA TECH

Comparing mid-latitude and high-latitude

CVW 2-7-18

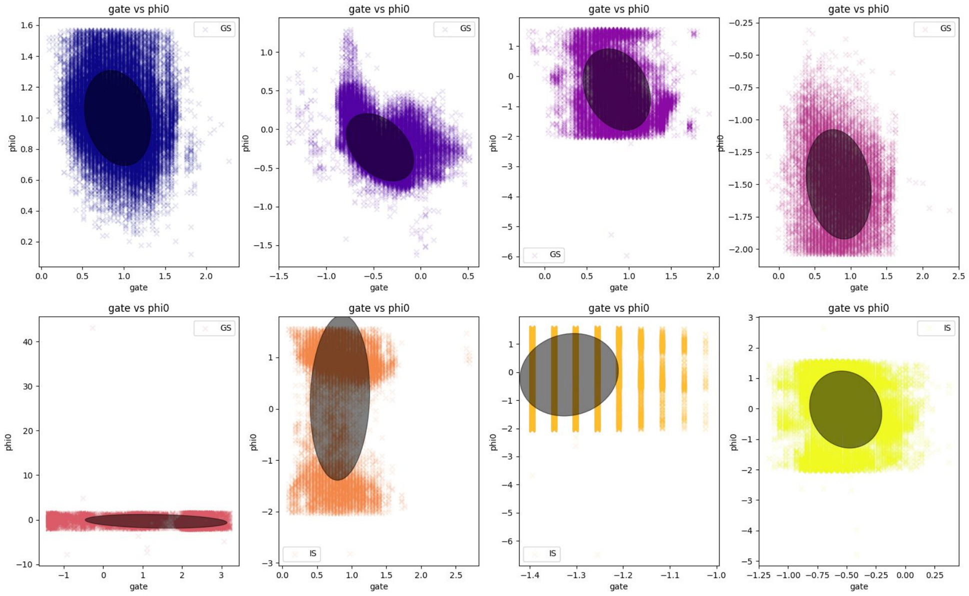


Mid-latitude

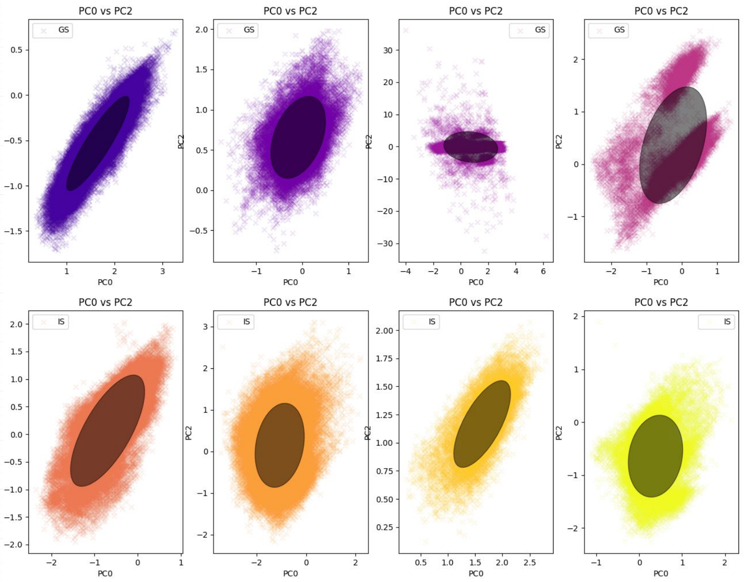
- GMM is doing a bad job (so is empirical model)
  - Likely low-velocity 0.5 hop IS
- Threshold should be adjusted

VIRGINIA TECH

Studying the dataset



PCA



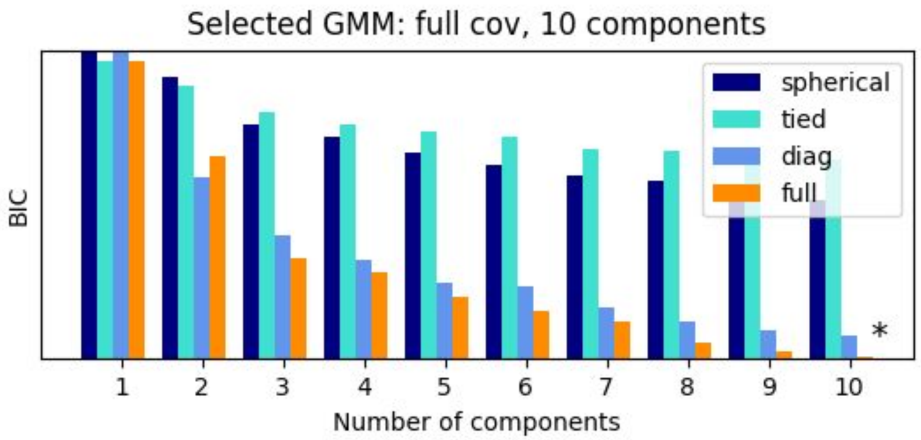
- Some features don't look Gaussian
  - Phi0 (above), beam, power

PCA transformation makes features look more Gaussian

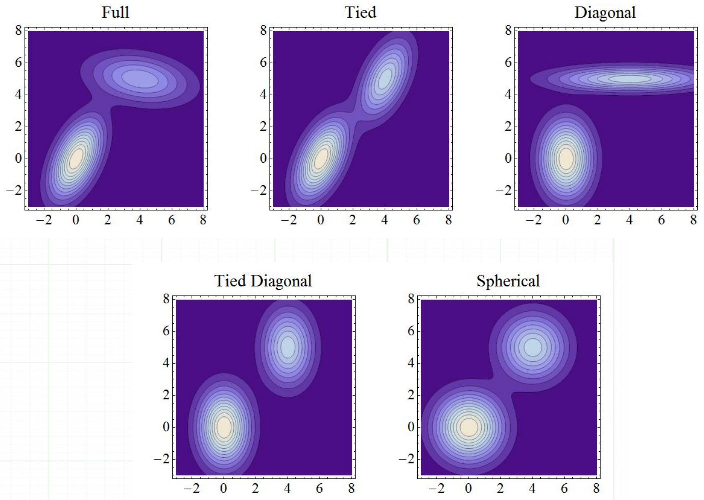
- PCA does an axis transformation - so our data looks more Gaussian after axis change
- PCA tries to get rid of 'noise' by dropping lowest-variance axis - assuming 'signal' has high variance 'noise' has low variance

Selecting GMM covariance type (BIC)

BIC for different covariances, # clusters



GMM Covariance types



Running BIC with different covariance types found that full covariance is best

VIRGINIA TECH

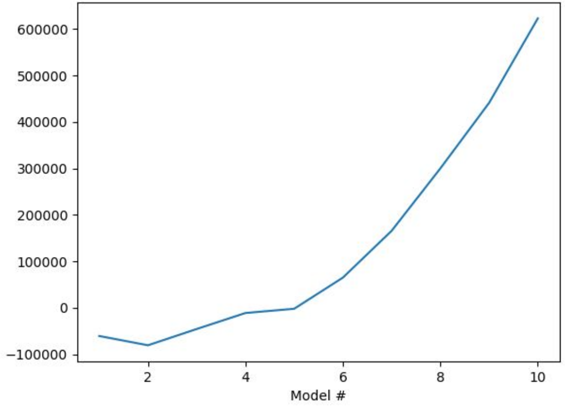
Next Steps

- Study a few days worth of data and use human expert analysis as ‘ground truth’ to:
    - Compare different GMM models
    - Adjust the threshold
        - Test out Ribiero method, test an adjustment of the traditional method
    - Test results of removing non-Gaussian features
        - Beam and power are low importance on all tests, sometimes phi0
    - Test transformations to capture edge behavior

Thank you!

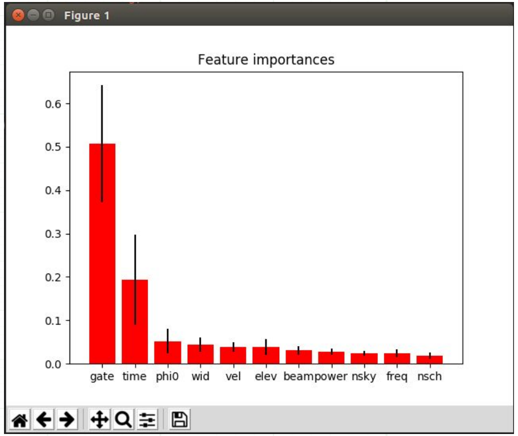
Selecting features/clusters (forward selection)

CVW, 2/7/18, beam 12



- Used 50 as max # of clusters

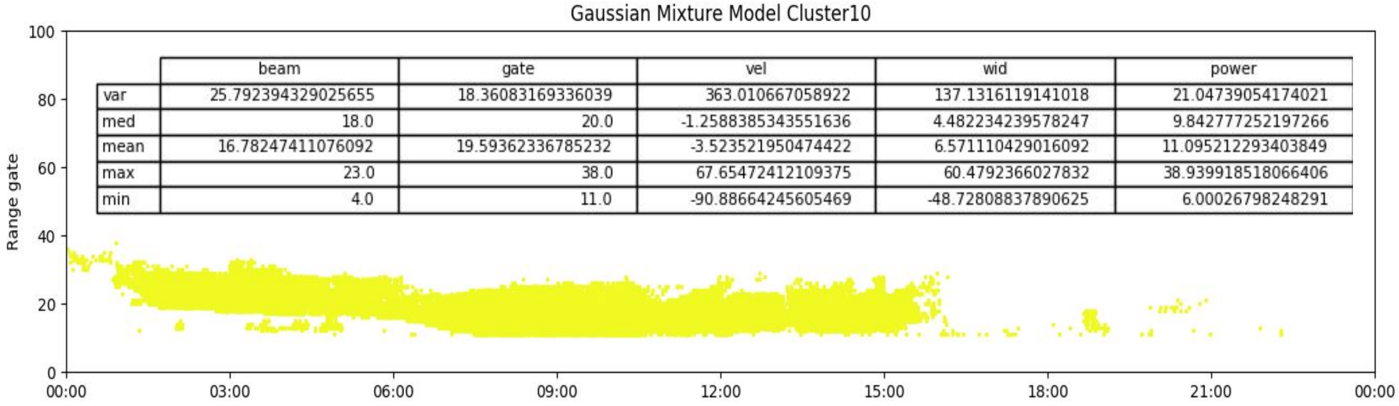
Model #	Features	# Clusters	BIC
1	Freq	9	-60,929
2	Freq, time	20	-80,932
3	Freq, time, nsky	35	-45,856
4	Freq, time, nsky, phi0	50	-11,452
5	Freq, time, nsky, phi0, elev	45	-2,317
6	Freq, time, nsky, phi0, elev, nsch	50	64,805
7	Freq, time, nsky, phi0, elev, nsch, gate	50	165,587
8	Freq, time, nsky, phi0, elev, nsch, gate, power	25	299,981
9	Freq, time, nsky, phi0, elev, nsch, gate, power, wid	25	441,036
10	Freq, time, nsky, phi0, elev, nsch, gate, power, wid, vel	20	623,056



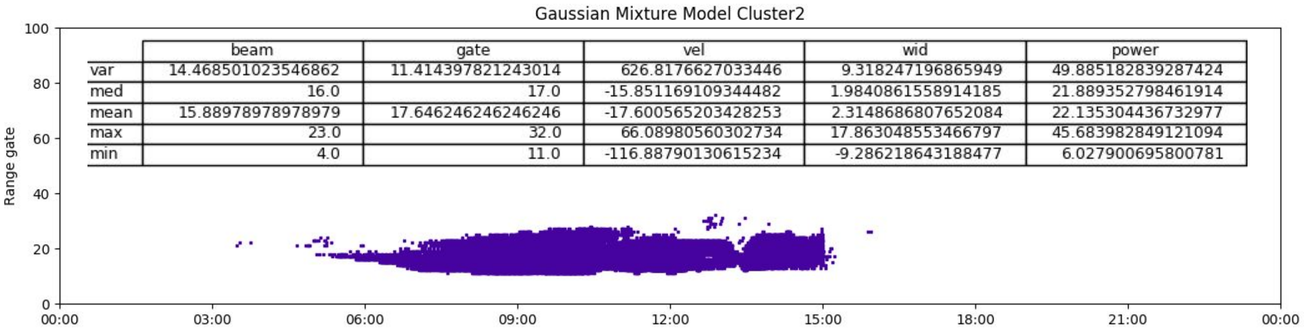
The old feature selection method

Comparing mid-latitude and high-latitude

CVW 2-7-18



	beam	gate	vel	wid	power
var	25.792394329025655	18.36083169336039	363.010667058922	137.1316119141018	21.04739054174021
med	18.0	20.0	-1.2588385343551636	4.482234239578247	9.842777252197266
mean	16.78247411076092	19.59362336785232	-3.523521950474422	6.571110429016092	11.095212293403849
max	23.0	38.0	67.65472412109375	60.4792366027832	38.939918518066406
min	4.0	11.0	-90.88664245605469	-48.72808837890625	6.00026798248291

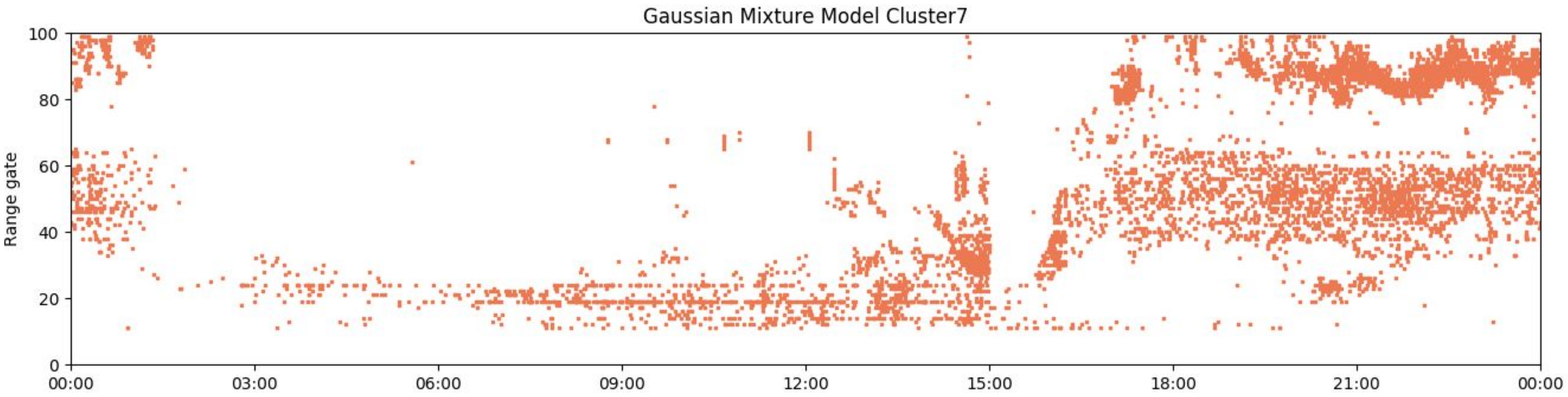


	beam	gate	vel	wid	power
var	14.468501023546862	11.414397821243014	626.8176627033446	9.318247196865949	49.885182839287424
med	16.0	17.0	-15.851169109344482	1.9840861558914185	21.889352798461914
mean	15.88978978978979	17.646246246246246	-17.600565203428253	2.3148686807652084	22.135304436732977
max	23.0	32.0	66.08980560302734	17.863048553466797	45.683982849121094
min	4.0	11.0	-116.88790130615234	-9.286218643188477	6.027900695800781

VIRGINIA TECH™

Comparing mid-latitude and high-latitude

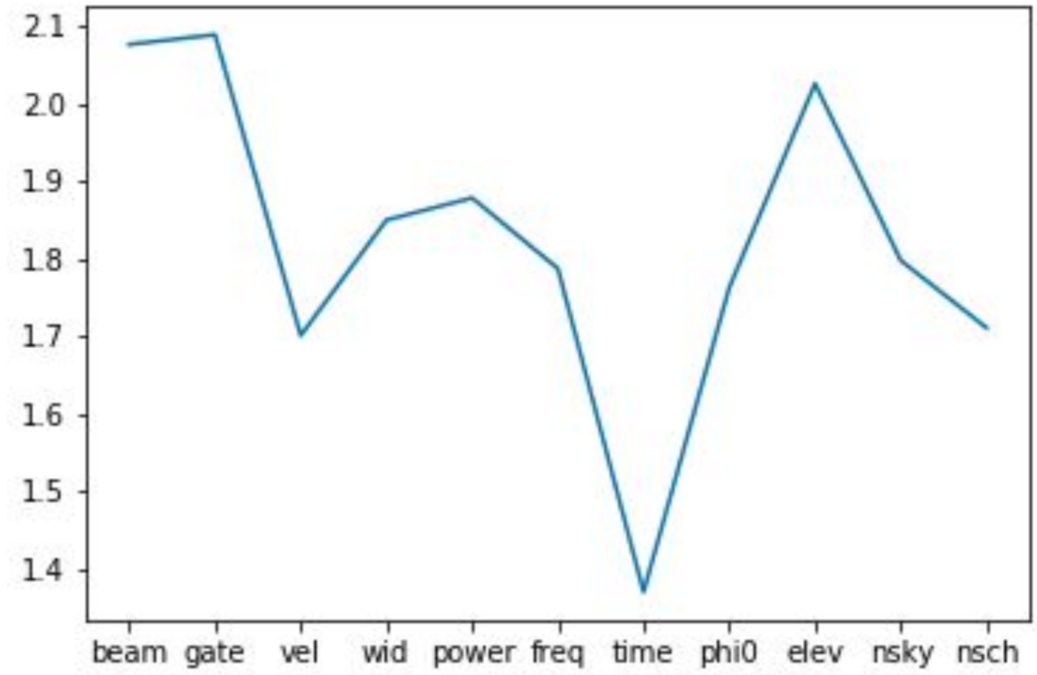
CVW 2-7-18



	beam	gate	vel	wid	power	phi0	time
var	36.8554986255 14166	720.8709707119998	139546.84563549238	5366.414710805141	24.344025027322395	4.170622114823706	0.06321287378376932
med	7.0	53.0	-0.7506966590881348	27.274415969848633	10.879316806793213	-0.3799201250076294	736732.8154141551
mean	10.067556060175987	57.09977292080613	4.997717326808907	43.79629185254508	12.06709424413269	-0.158927113260468	736732.7212298575
max	23.0	99.0	3669.947021484375	798.5252075195312	37.452022552490234	75.33321380615234	736732.9999608797
min	4.0	11.0	-3456.23388671875	-512.9129638671875	6.000430583953857	-12.399154663085938	736732.0001034491

- High-variance data gets grouped into 1 cluster
- Ways to solve:
  - Data transformation
  - Covariance matrix that limits shape of clusters

Selecting features/clusters (PCA)

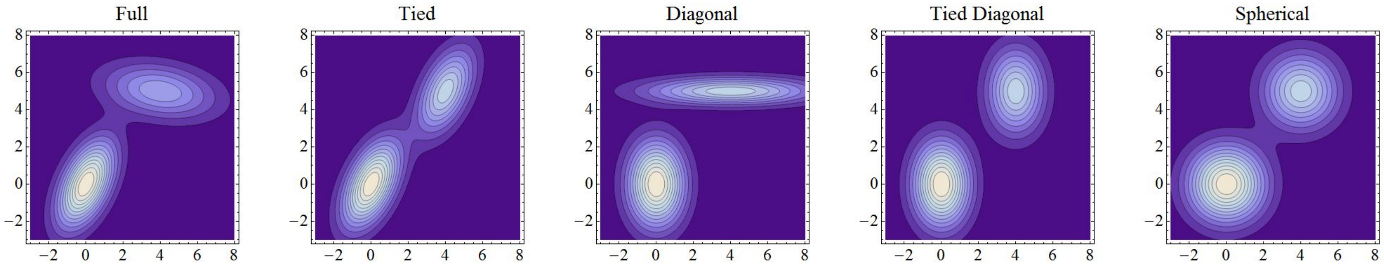


CVW, 2/7/18, all beams

Importance order:

1. Gate
2. Beam
3. Elev
4. Power
5. Width
6. Nsky
7. Freq
8. Phi0
9. Nsch
10. Vel
11. Time

Covariance matrices



- **Full** means the components may independently adopt any position and shape.
- **Tied** means they have the same shape, but the shape may be anything.
- **Diagonal** means the contour axes are oriented along the coordinate axes, but otherwise the eccentricities may vary between components.
- **Tied Diagonal** is a "tied" situation where the contour axes are oriented along the coordinate axes. (I have added this because initially it was how I misinterpreted "diagonal.")
- **Spherical** is a "diagonal" situation with circular contours (spherical in higher dimensions, whence the name).

https://stats.stackexchange.com/questions/326671/different-covariance-types-for-gaussian-mixture-models

VIRGINIA TECH