Classification of SuperDARN backscatter using machine learning algorithms

E. A. Robb[1], X. Shi[1], S. Chakraborty[1], J. M. Ruohoniemi[1], J. B. H. Baker[1], and A. G. Burrell[2]

[1]*Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, USA*

[2] *Department of Physics, University of Texas at Dallas, Dallas, USA*

Introduction

Previous methods of classifying backscatter:

- Traditional model [Blanchard et. al. 2009]
  - $v < 33.1$ m/s + 0.139w – (0.00133 s/m)w^2
  - Biased to classify low-velocity ionospheric scatter (IS) as ground scatter (GS)
  - *Problem*: Classifies data point-by-point, does not consider clusters of data
  - *Problem*: Bias in low-velocity IS
- Depth-first search clustering model [Ribiero et. al. 2011]
  - Do depth-first search to cluster data in time/space
  - Classify cluster as GS or IS based on velocity ratio
  - Reduces bias so that more low-velocity IS is correctly classified
  - *Problem*: Uses boxcar filtering, which reduces spatial and time resolution
- Empirical model [Burrell et. al. 2015]
  - Based on elevation angle and virtual height model
  - *Problem*: Requires reliable elevation angle measurements, which are often unavailable
  - *Problem*: Only considers spatial variations along a single beam and time

Methodology

Goals:

- Preserve data resolution
- Get clean GS and IS data
- Cluster similar data the way an expert human would

Solution:

- Apply a Gaussian Mixture Model (GMM)

Gaussian Mixtures:

- Composed of several Gaussian distributions added together (Figure 1)
- Each distribution is called a component

GMM:

- Learn a mean and covariance matrix for each component, based on how well it fits the data
- Number of components is defined by the programmer
- Components are called 'clusters'
- Unsupervised learning - no knowledge of ground truth.
- Figure 2 - example output of GMM on 2 features

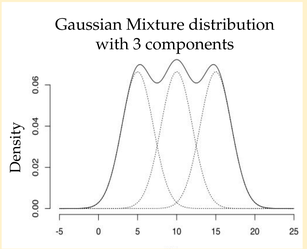

Figure 1: A Gaussian Mixture with 3 components.

Figure 2: Example GMM results on 2 features with 3 clusters.

Methodology (continued)

Our Method:

- Use 7 features: velocity, spectral width, power, phi0, time, range gate, beam
  - Determined best features by training a decision tree on the data using empirical method as 'ground truth' (Figure 3)
- Create 6-30 clusters on 1 day of SuperDARN data
- Classify each cluster as IS or GS based on:
  - |median vel| > 15 m/s
    - Threshold was determined by trial and error

Parameters to Tune:

- Covariance type
- Number of clusters
- Velocity threshold

Evaluating Model Quality:

- Use Bayes Information Criterion (BIC), a statistical measure of inverse 'goodness of fit' (less is better)
  - BIC includes a penalty term to avoid overfitting
- 5-10 clusters works for 1-2 features - more is needed for all 7

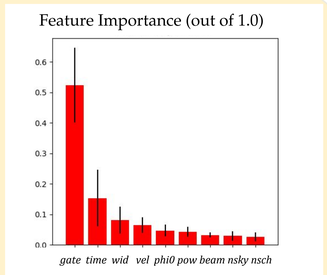

Figure 3: Features ordered by importance using a decision tree.

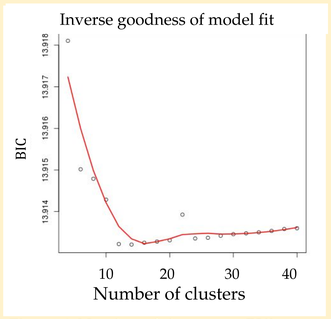

Figure 4: Number of clusters vs. BIC score for model fit to spectral width and velocity.

Results

GMM has Low Bias:

- Applied GMM and plotted |velocity| of scatter classified as IS and GS (Figure 5)
  - Curves are smooth with a small bump in IS around 30 m/s
- *Comparison*: Traditional method shows strong bias to miscategorize low-velocity IS (Figure 6, dashed line)
- *Comparison*: Ribiero depth-first search method shows low bias, performs similarly to GMM (Figure 6, solid line)
  - Bump at 15 m/s caused by the high / low velocity threshold

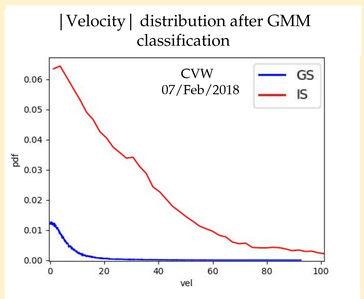

Figure 5: Velocity distribution of scatter classified as GS and IS by GMM.

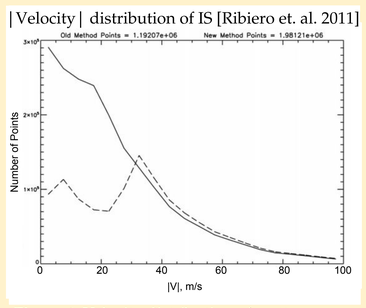

Figure 6: Velocity distribution of scatter classified as IS by the traditional method (dashed line) and the Ribiero depth-first search method (solid line).

Results (continued)

GMM vs. Empirical / Traditional Models:

- Often performs better than the empirical and traditional models (Figures 7 and 8, see black arrows)
  - GMM classifies scatter by clusters in time and space, which is an advantage
- Works on high-latitude (Figure 7) and mid-latitude (Figure 8) radars
- Works without elevation angle or phi0
- Preserves data resolution, no filtering

Saskatoon, Feb. 7 2018 (High-latitude)

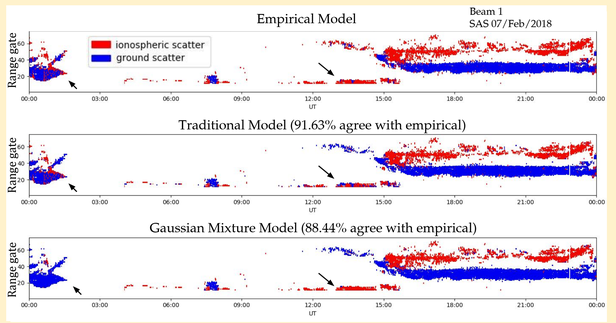

Figure 7: Model comparison between the empirical model, traditional model, and GMM on data from Saskatoon (high-latitude).
Note: we exclude data from range gate <= 10.

Christmas Valley West, Feb. 7 2018 (Mid-latitude)

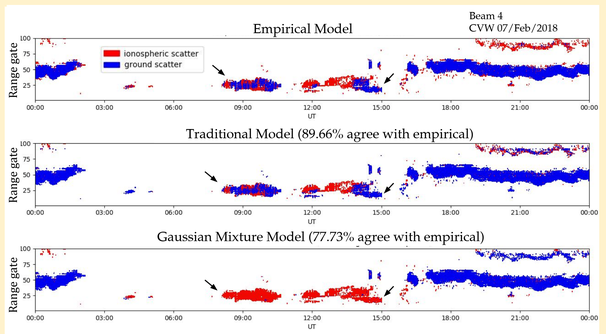

Figure 8: Model comparison between the empirical model, traditional model, and GMM on data from Christmas Valley West (mid-latitude).

GMM Output Distributions (Mid-latitude):

- Velocity appears Gaussian (Figure 9), reasonable to fit using Gaussian model
- Range gate appears to be a mixture of Gaussians (Figure 10), reasonable to fit using Gaussian Mixture (3 - 4 components)
- Range gate graph shows GMM is misclassifying some IS as GS (Figure 10, blue arrows)
  - This IS may be low-velocity
  - We need to adjust velocity threshold, or let the majority cluster at that range gate dictate the classification

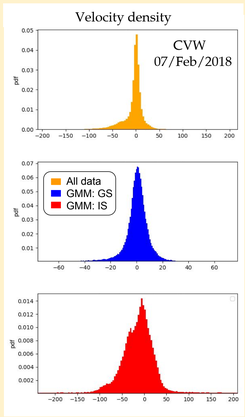

Figure 9: Velocity density for the full set of scatter combined, compared to scatter classified by GMM.

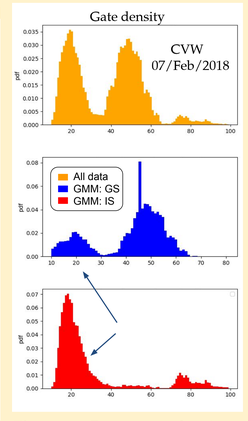

Figure 10: Range gate density for the full set of scatter combined, compared to scatter classified by GMM.

Conclusions

- Gaussian Mixture Model provides an alternative to:
  - The depth-first search model - GMM maintains high data resolution and works at mid-latitudes
  - The empirical model - GMM can work without elevation angle
  - The traditional model - GMM preserves GS
- Future research steps:
  - Statistically validate the results in a more rigorous way
  - Show that GMM's model can accurately capture the distribution of all features and does not exclude outliers
  - Work towards a module that can be distributed to the community to aid in scatter identification

References

- Blanchard, G. T., S. Sundeen, and K. B. Baker (2009), Probabilistic identification of high-frequency radar backscatter from the ground and ionosphere based on spectral characteristics, Radio Sci., 44, RS5012, doi: 10.1029/2009RS004141.
- Burrell, A. G., S. E. Milan, G. W. Perry, T. K. Yeoman, and M. Lester (2015), Automatically determining the origin direction and propagation mode of high-frequency radar backscatter, Radio Sci., 50, 1225–1245, doi:10.1002/2015RS005808.
- Ribeiro, A. J., J. M. Ruohoniemi, J. B. H. Baker, L. B. N. Clausen, S. de Larquier, and R. A. Greenwald (2011), A new approach for identifying ionospheric backscatter in midlatitude SuperDARN HF radar observations, Radio Sci., 46, RS4011, doi:10.1029/2011RS004676.

This project is sponsored by Google Summer of Code 2018.

Contact: *erobb@vt.edu, xueling7@vt.edu*

Code: *github.com/vtsuperdarn/clustering_superdarn_data*