Assessing the Decisiveness-Accuracy-Robustness of Probabilistic Forecasts

Kenric P. Nelson
President, Photrek

Photrek

We Enlighten Your Pathway

Through research and development to accelerate innovation in:

- *Machine Intelligence*
- *Blockchain Digital Assets*
- *Complex Systems*

Machine Intelligence

"Dr. Nelson solved a long-standing information fusion problem for us, developing an algorithm based on solid mathematical principals. We implemented this on several programs and successfully demonstrated it in real-world testing. His robust approach has now become our standard solution to this problem across our entire portfolio."

- Adam Art

https://photrek.world

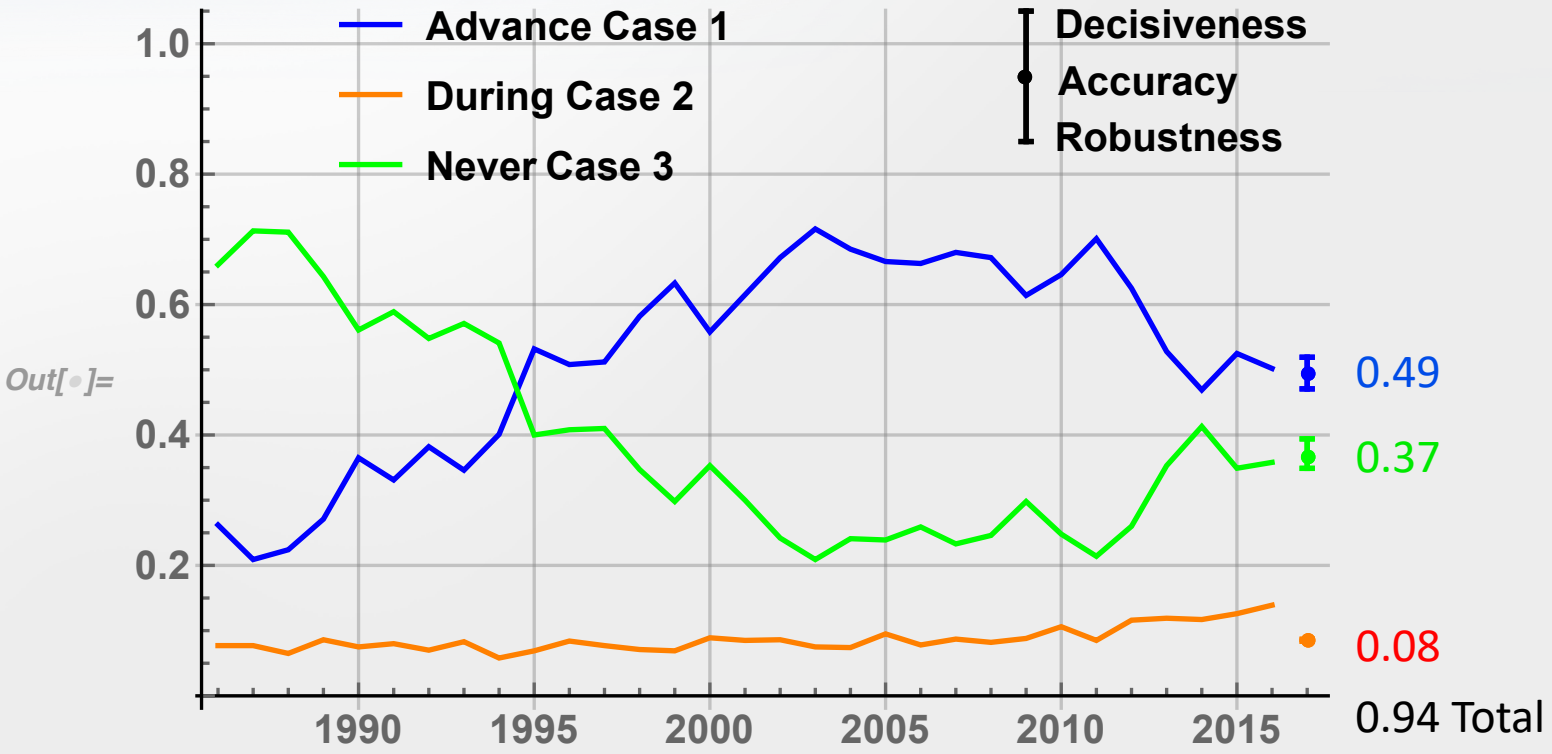
Located at the Impact Hub Boston
50 Milk St, Boston, MA 02109

Introduction & Outline

- A variety of methods have been developed to assess of probabilistic forecasts
  - What's missing is clear, simple calculation of the central tendency
  - Will show that this is achieved using the geometric mean
  - And that variation is measured using the generalized mean

- Show an example from Tornado Forecast Warnings
- Derive methods from Generalized Information Theory
- Step through a simple two-class example
- Example assessment of deep-learning methods
- Discuss application to severe weather forecasts
- Discuss issues of scale and sample size

Tornado Forecast Warnings



- H. E. Brooks and J. Correia, "Long-term performance metrics for National Weather Service Tornado warnings," Weather Forecast., vol. 33, no. 6, 2018
- At the right, the Decisiveness – Accuracy – Robustness means are plotted
- Provides an approach to summarizing probability data with a
  - Central tendancy – Geometric Mean
  - Upper Variation – Arithmetic Mean
  - Lower Variation – -2/3rds Mean

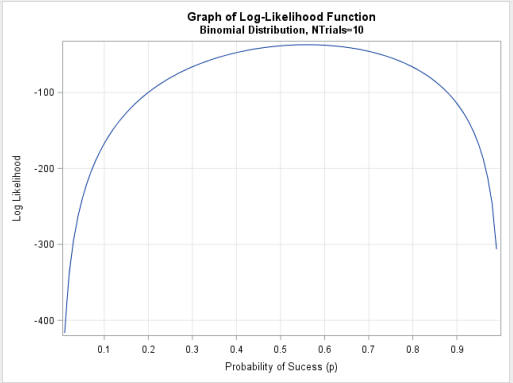
Assessing the performance of a decision-system

- Classification Performance:
  - Did the system make the right decision?
  - Ratios from the confusion matrix form *precision, sensitivity, accuracy, F1-score, etc.*
- Accuracy of Predictions
  - log-likelihood of predicted probability
  - Use of log of the likelihood (probability) grounded in information theory and enables arithmetic average of performance; however the resulting number is abstract
  - Mean-squared average common alternative; however, not well-grounded in probability theory
- Robustness?

Confusion Matrix:

		Actual class	
		Cat	Non-cat
Predicted class	Cat	True Positives	False Positives
	Non-cat	False Negatives	True Negatives

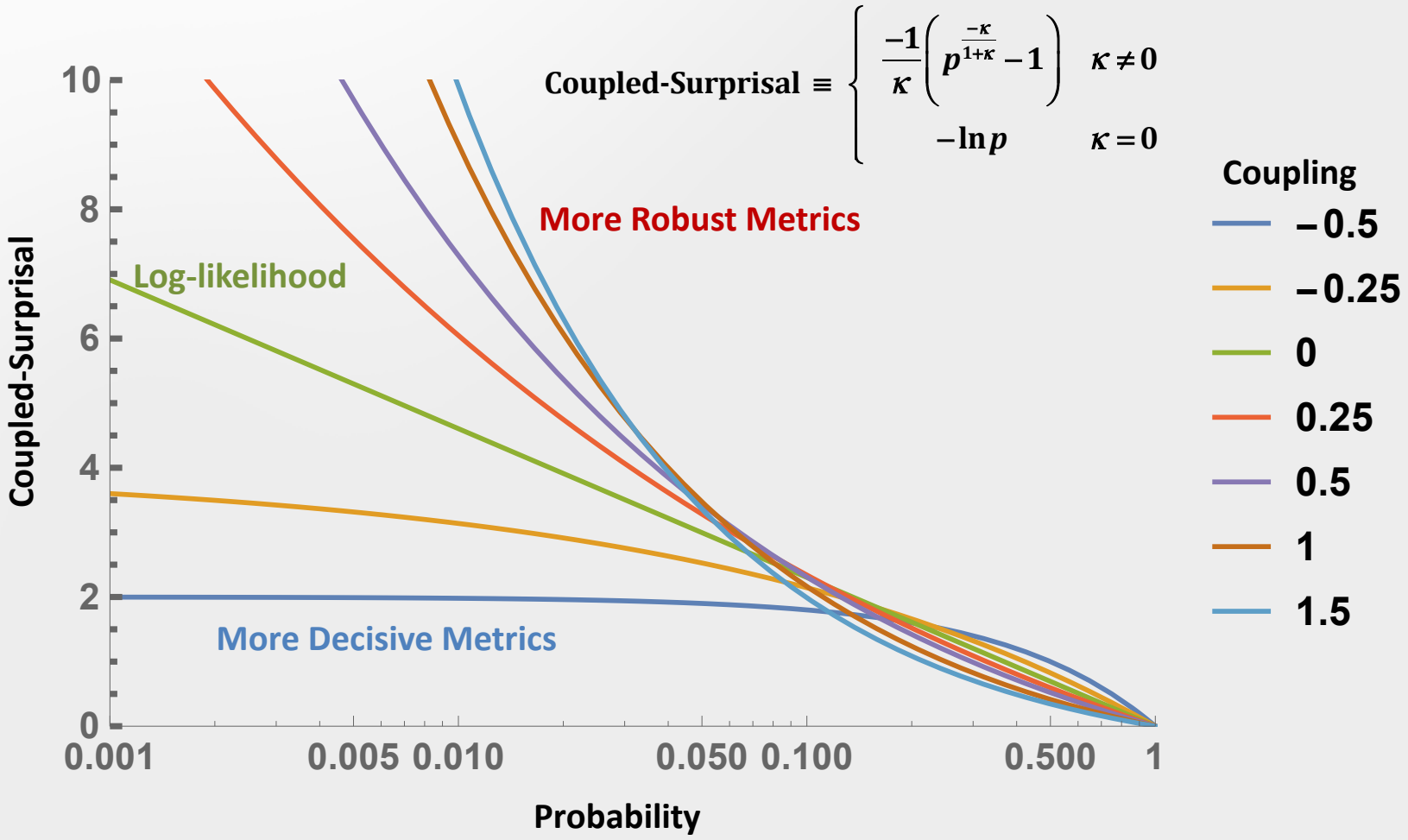
Wikipedia: Confusion Matrix, CatDog



SAS Blog

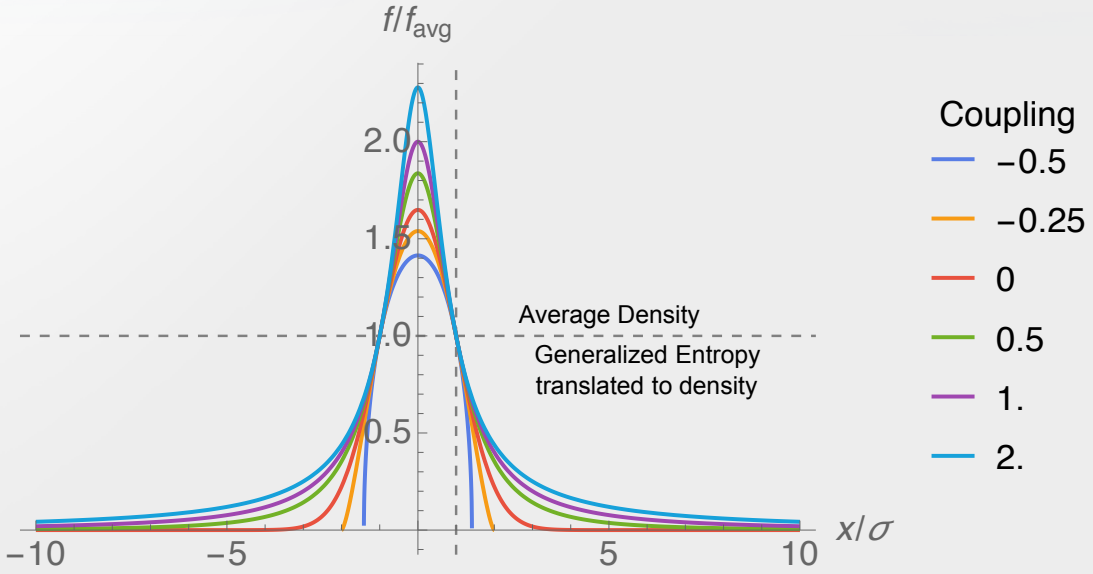
Generalized Log-Likelihood Metrics

- Negative Coupled Logarithm



Average Density: Generalized Mean

- Gaussian central tendency of density is geometric mean and is equal to density at mean plus standard deviation
- Generalizes to family of coupled Gaussian distributions with the generalized mean of density



$$f\left(x;\mu,\sigma,\kappa\right)=\frac{1}{Z\left(\sigma,\kappa\right)}\left(1+\kappa\frac{\left(x-\mu\right)^2}{\sigma^2}\right)_+^{\frac{1+\kappa}{-2\kappa}};\quad \kappa\geq -1,\ \sigma\geq 0,$$

$$f_{\kappa avg}\equiv\left(\int_{x\in X}\left(f\left(x\right)\right)^{1+\frac{2\kappa}{1+\kappa}}dx\right)^{\frac{1+\kappa}{2\kappa}}=f\left(\mu\pm\sigma,\mu,\sigma,\kappa\right)$$

Nonlinear Statistical Coupling

A statistical model for complex systems

"Coupled" Distributions:

$$\left(\exp_\kappa \left(\frac{x}{\sigma} \right)^\alpha \right)^{\frac{-(1+d\kappa)}{\alpha}}$$

Student's t, Pareto & Levy

κ - coupling or tail decay

σ - scale of distribution

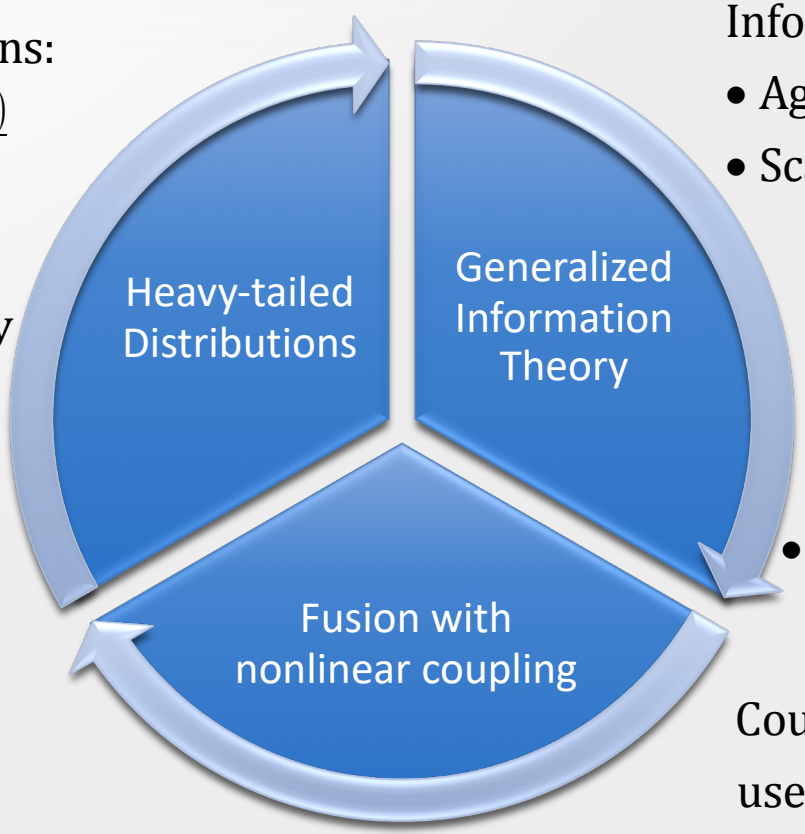
α - polynomial order

d - dimensions

Gen. Exp & Log

$$\exp_\kappa x \equiv \left(1+\kappa x\right)^{\frac{1}{\kappa}}$$

$$\ln_\kappa x \equiv \frac{x^\kappa - 1}{\kappa}$$



Nonlinear Statistical Coupling is the inverse of the **Degree of Freedom**

Information Metric

- Aggregation: Gen. Mean
- Scale: Gen. logarithm

$$\frac{-1}{1+\kappa} \ln_r \left(\sum_i p_i^{1+r} \right)^{\frac{1}{r}}$$

$$r = \frac{\alpha\kappa}{1+\kappa}$$

- r measure of risk tolerance

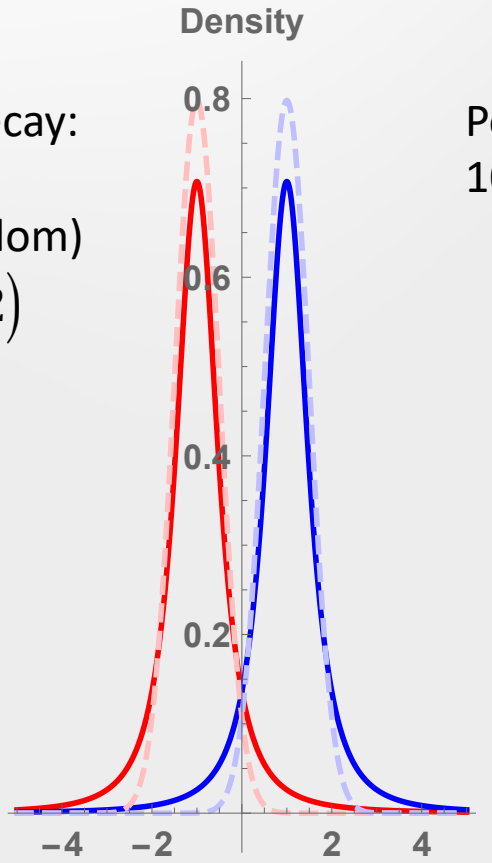
Coupled Bayesian Fusion uses generalized product

$$\prod_i {}_{\otimes_r} p_i \equiv \left(\exp_\kappa \left(\sum_i \ln_\kappa p_i^{\frac{-\alpha}{1+\kappa}} \right) \right)^{\frac{1+d\kappa}{-\alpha}}$$

Discrepancy between decisions and probability accuracy

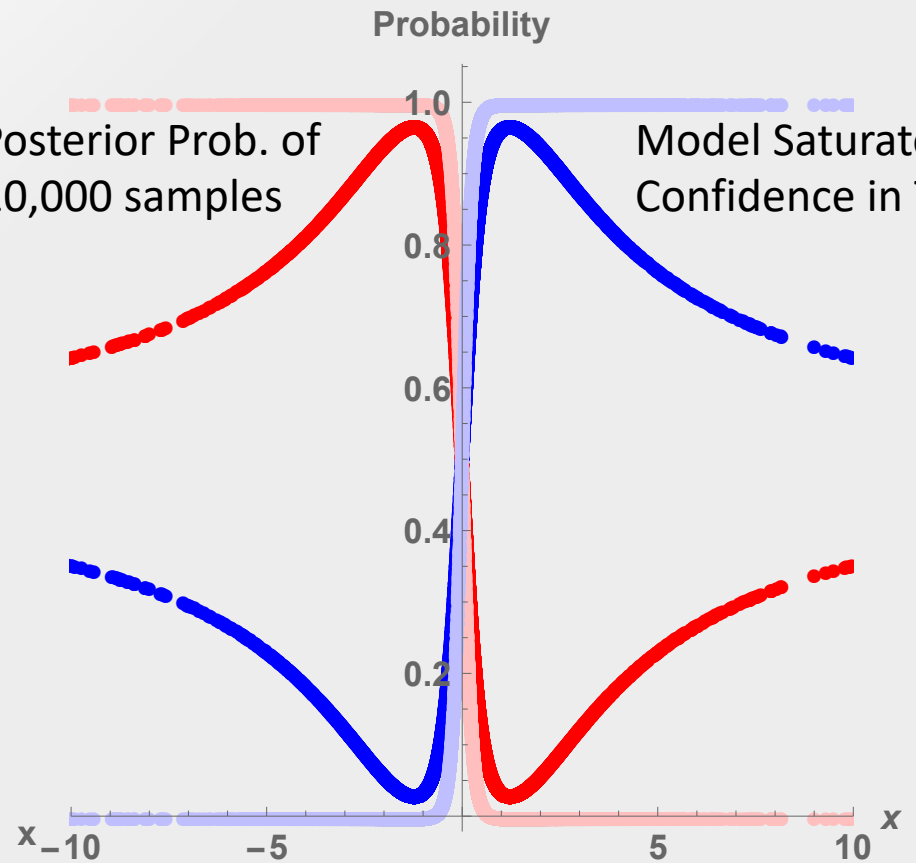
Source (Bold) and Model (Light, Dash) have identical mean and variance
Models are Gaussian, but Source is heavy-tail Coupled Gaussian (or Student's t)

Source Tail Decay:
Coupling
(Deg. of Freedom)
$\kappa = 0.5 \left(\nu = 2 \right)$



Posterior Prob. of
10,000 samples

Model Saturates
Confidence in Tail



Discrepancy between decisions and probability accuracy

Classification Metrics:

- Percent correct decisions, Leakage, Wastage, Precision and Recall

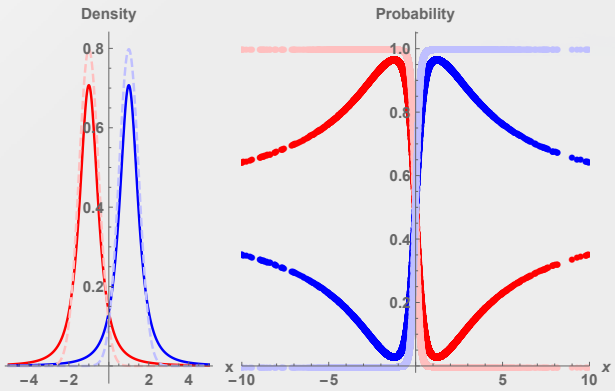
Probability Metric:

- Generalized Mean
- Derives from Gen. Entropy

Decisiveness: Arithmetic Mean of Probabilities

Accuracy: Geometric Mean of Probabilities

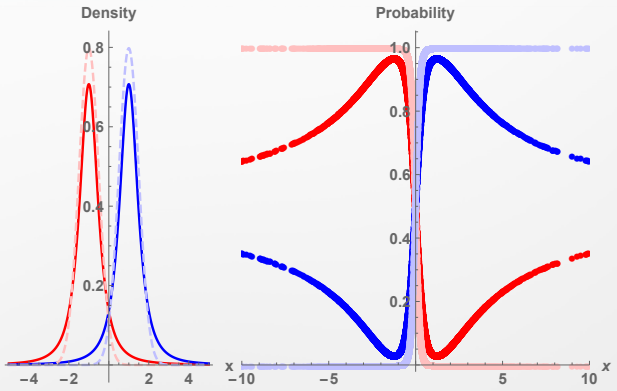
Robustness: -2/3rds Mean of Probabilities



Decision and Probability Performance

Metric	Perfect	Model Gauss
Classification	0.908	0.908
Decisiveness	0.853333	0.895737
Accuracy	0.771746	0.40466
Robustness	0.623693	4.55E-15

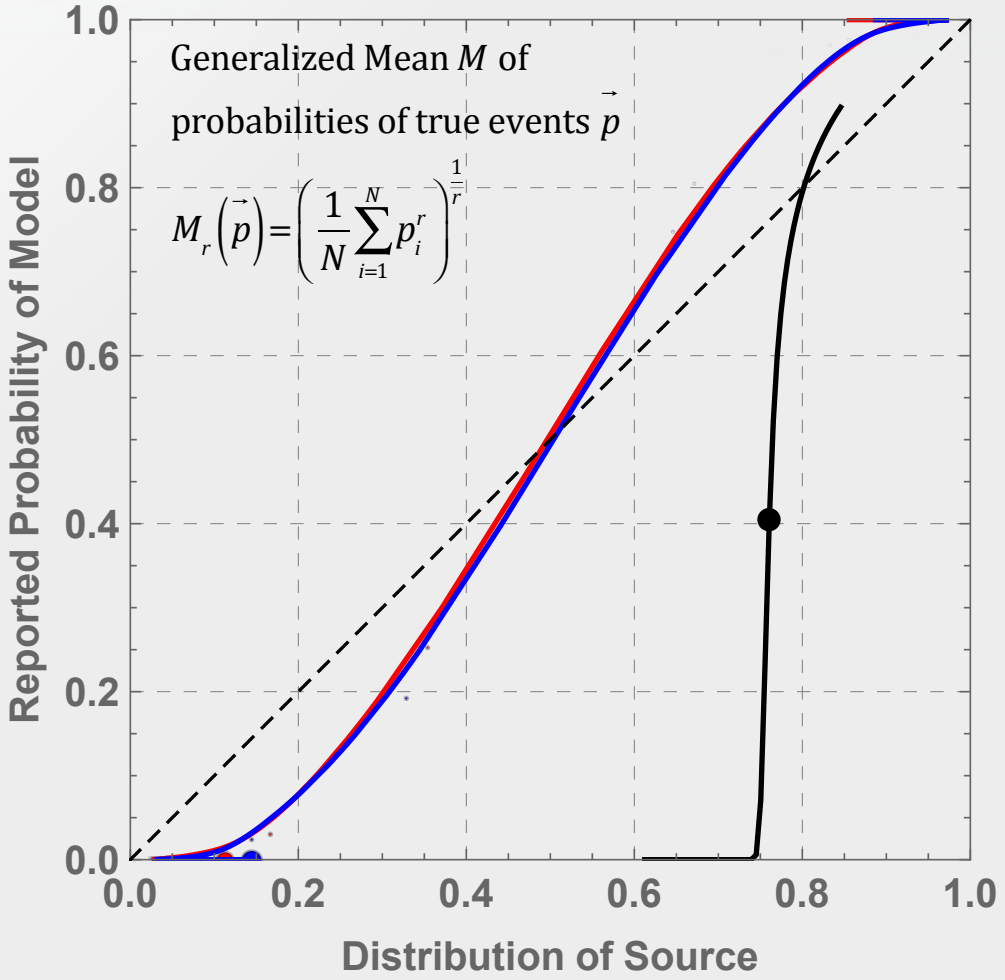
Discrepancy between decisions and probability accuracy



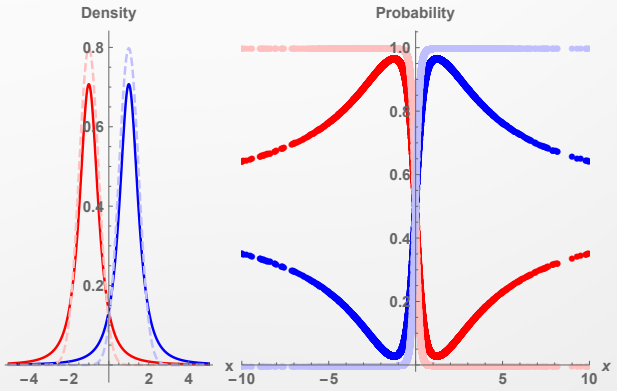
Y - axis: Predicted Probabilities
X – axis: Histogram Measurements

Red & Blue curves are the Probability Calibration curves

Black curve is the generalized mean from -2/3rds to 1
Black dot is the geometric mean

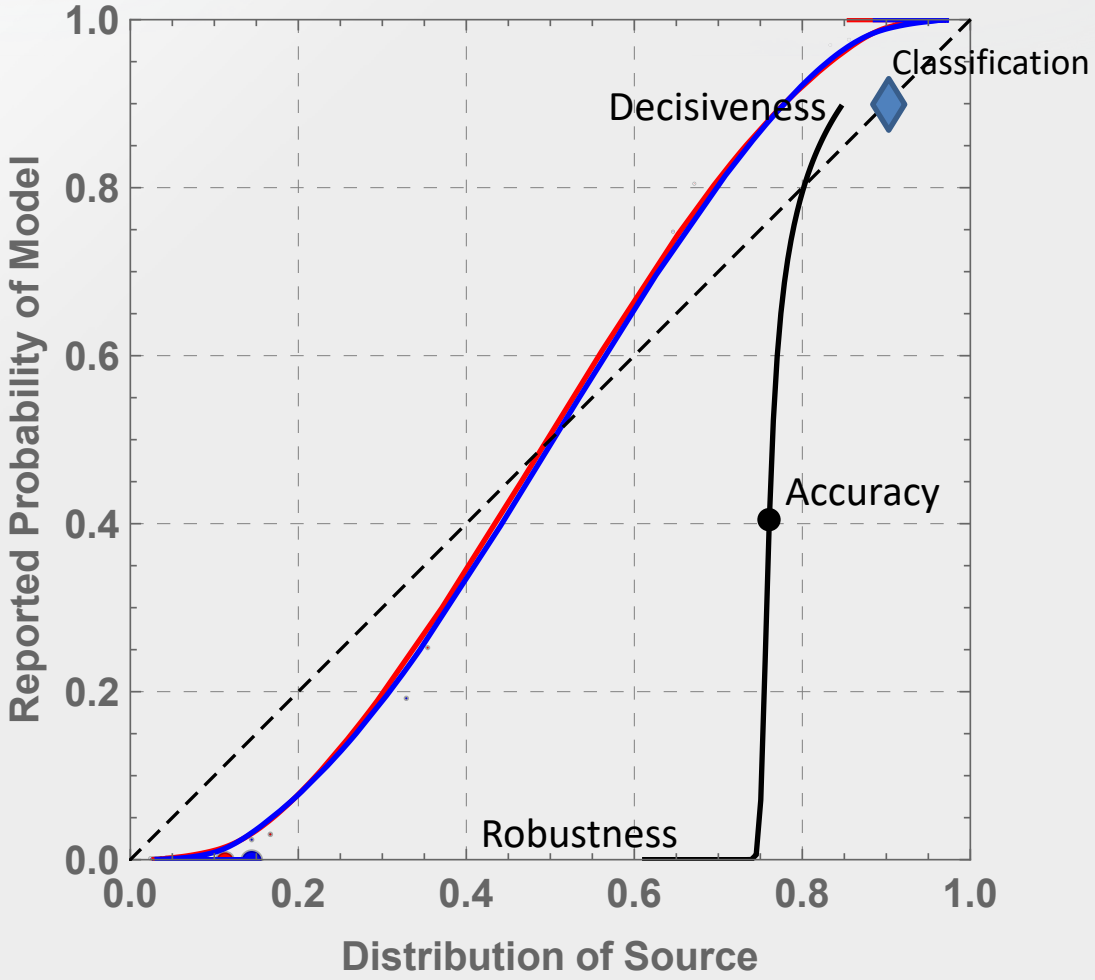


Discrepancy between decisions and probability accuracy

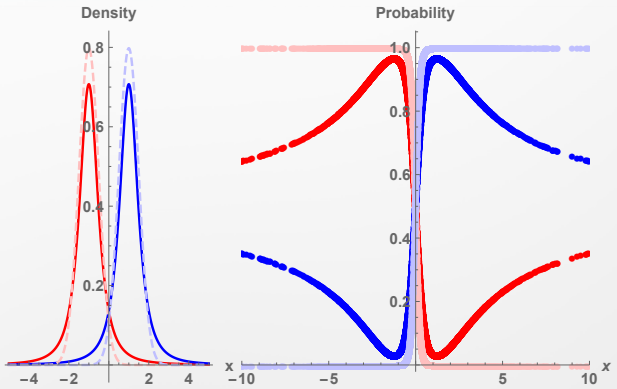


Decisiveness – Arithmetic Mean
Accuracy – Geometric Mean
Robustness - -2/3rds Mean

Classification ≈ Decisiveness



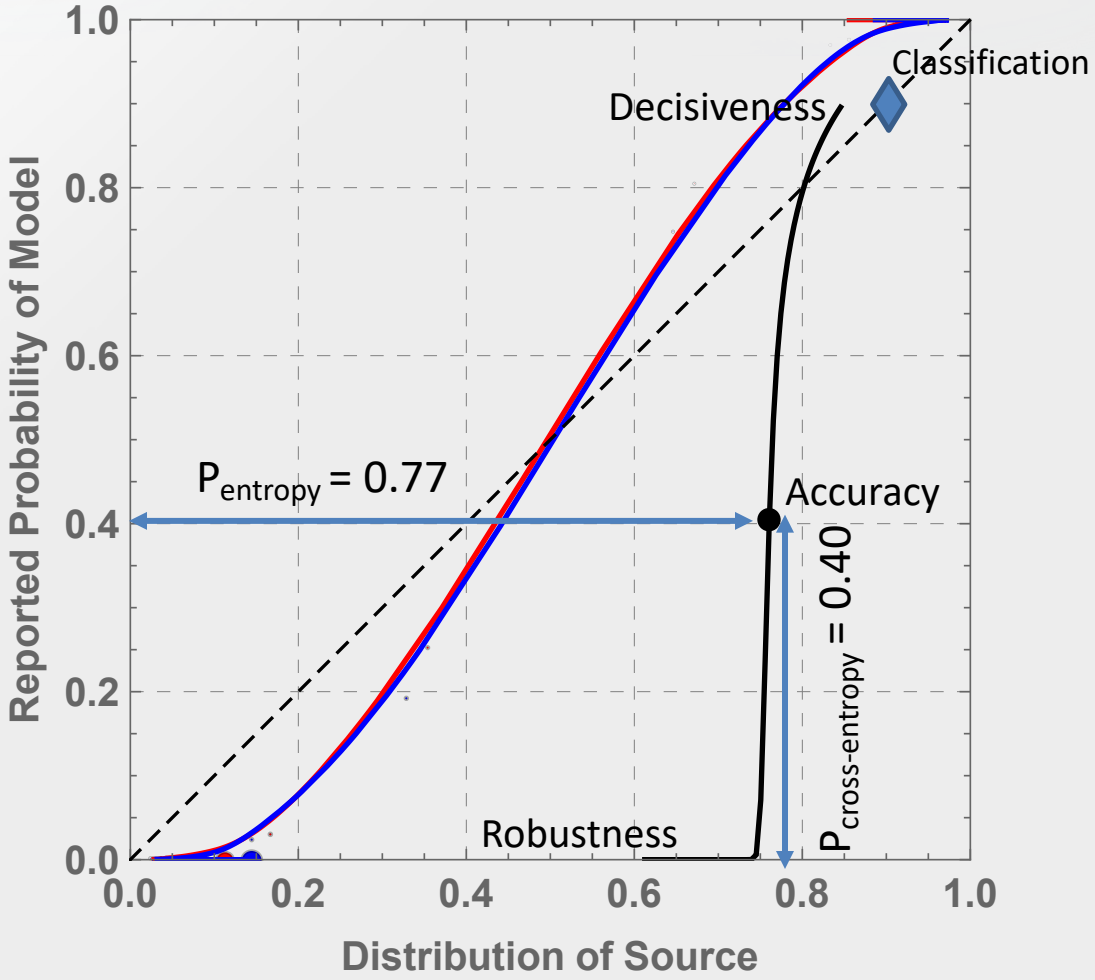
Discrepancy between decisions and probability accuracy



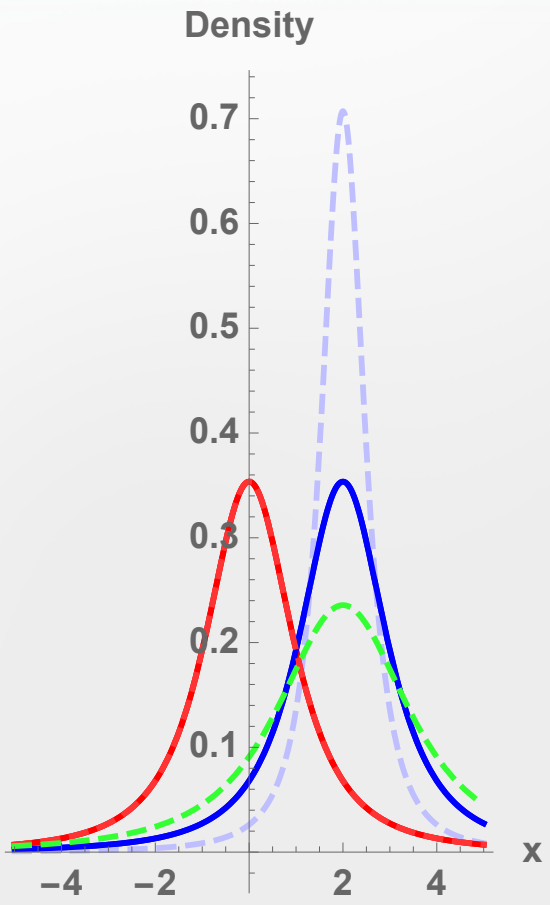
Cross-Entropy = Entropy + Divergence

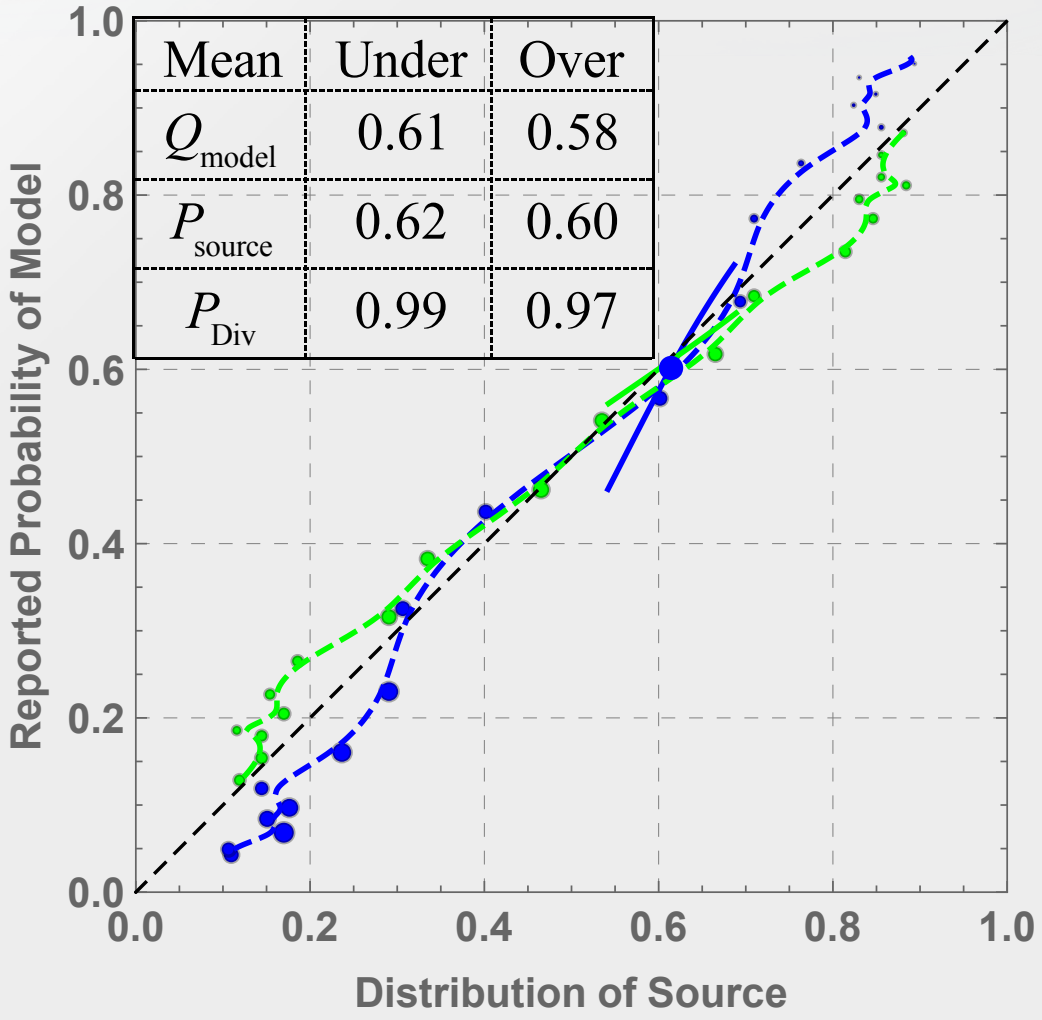
$P_{cross\text{-}entropy} = P_{entropy} \bullet P_{divergence}$

$$P_{divergence} = \frac{P_{cross-entropy}}{P_{entropy}} = \frac{0.40}{0.77} = 0.52$$



Under & Over Confident Standard Deviation





Ensemble Probability of Vorticity

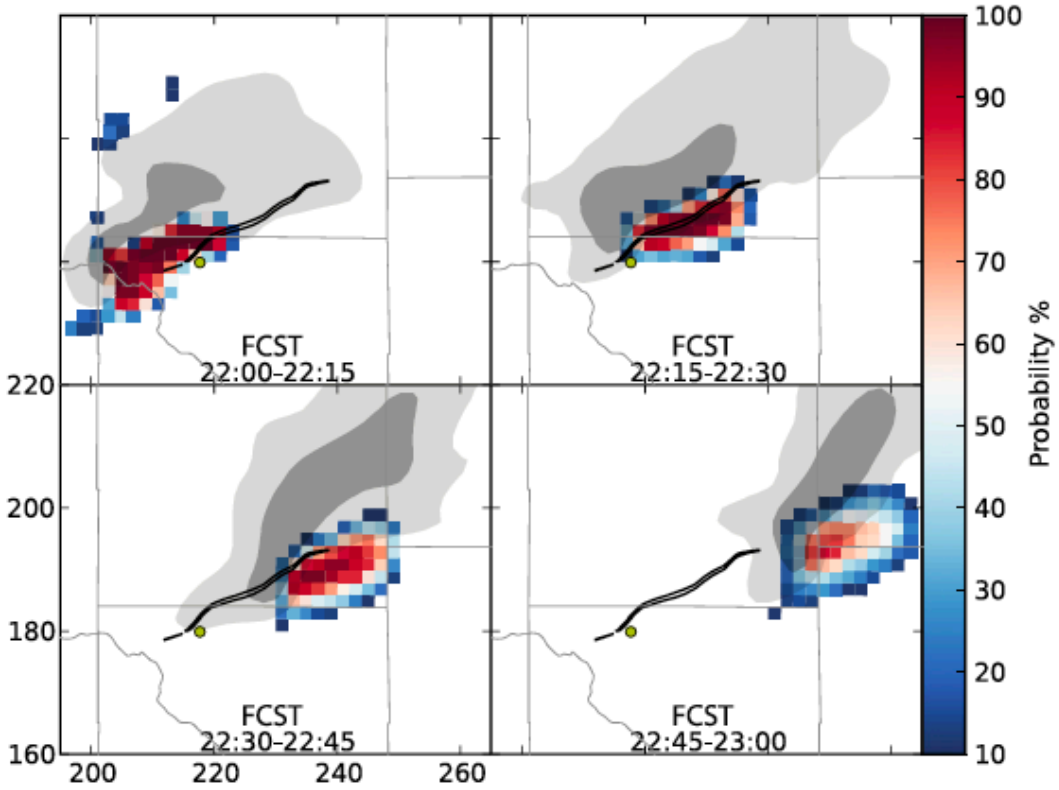


Fig. 4. Ensemble probability of vorticity exceeding 0.00375 s^{-1} at ~1 km AGL for the 8 May 2003 forecast experiment for four 15-min time windows starting at 2200 UTC 8 May 2003. Simulated radar reflectivity regions greater than or equal to 30 dBZ and 50 dBZ are shaded in light and dark gray, respectively, for ensemble member 7 at the beginning of each time interval for each panel. Overlaid in each panel is the observed tornado damage track (black outline), the location of Moore, Oklahoma (yellow dot), and county borders (thin black lines). The time interval (UTC) of each 15-min period is indicated in each panel.

Incorporating Sample Size into Assessment

- Error bars are calculated using the standard deviation divided by the number of samples
- Some function of the probabilities and the sample size is need to transform the variation bounds into a meaningful errorbar
- However, it's not necessarily division by *n.* Would need to analyze variation for typical distributions
- Beta distribution is commonly used to model the variation in probabilities

Conclusion

- Central tendency of a set of probabilistic forecasts is measured by the **Geometric Mean** of the probabilities for the observed events.

- Variation in probabilistic forecasts can measured by the generalized mean
- The generalized mean is the aggregation function associated with generalized entropy
  - Geometric Mean: associated with Shannon Entropy
  - Arithmetic Mean: associated with decision performance
  - -2/3 ^{rds} Mean: conjugate to arithmetic & associated with robustness; i.e. optimizing this smooths predictions
- Incorporation of sample size would provide error bars on the probabilistic forecast assessment

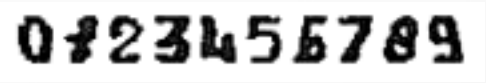
References

- Tutorial on methods
  - K. P. Nelson, "Reduced Perplexity: A simplified perspective on assessing probabilistic forecasts," in *Info-Metrics Volume*, M. Chen, J. M. Dunn, A. Golan, and A. Ullah, Eds. Oxford University Press, 2020. http://arxiv.org/abs/1603.08830
- Information Theoretic Proofs
  - K. P. Nelson, S. R. Umarov, and M. A. Kon, "On the average uncertainty for systems with nonlinear coupling," *Phys. A Stat. Mech. its Appl.*, vol. 468, pp. 30–43, 2017.
  - K. P. Nelson, M. A. Kon, and S. R. Umarov, "Use of the geometric mean as a statistic for the scale of the coupled Gaussian distributions," *Physica A*, vol. 515, pp. 248–257, 2019.
- Applications
  - S. Cao, J. Li, K. P. Nelson, and M. A. Kon, "Coupled VAE: Improved Accuracy and Robustness of a Variational Autoencoder," *arXiv:1906.00536[cs.LG]*, Jun. 2019.
  - K. P. Nelson, B. J. Scannell, and H. Landau, "A Risk Profile for Information Fusion Algorithms," *Entropy*, vol. 13, no. 8, pp. 1518–1532, 2011.
  - K. P. Nelson, M. Barbu, and B. J. Scannell, "Probabilistic graphs using coupled random variables," in *SPIE Sensing Technology & Applications*, 2014, p. 911903.

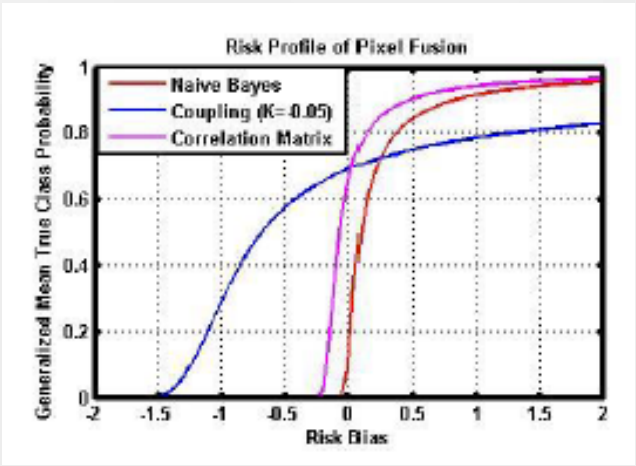
Back-ups

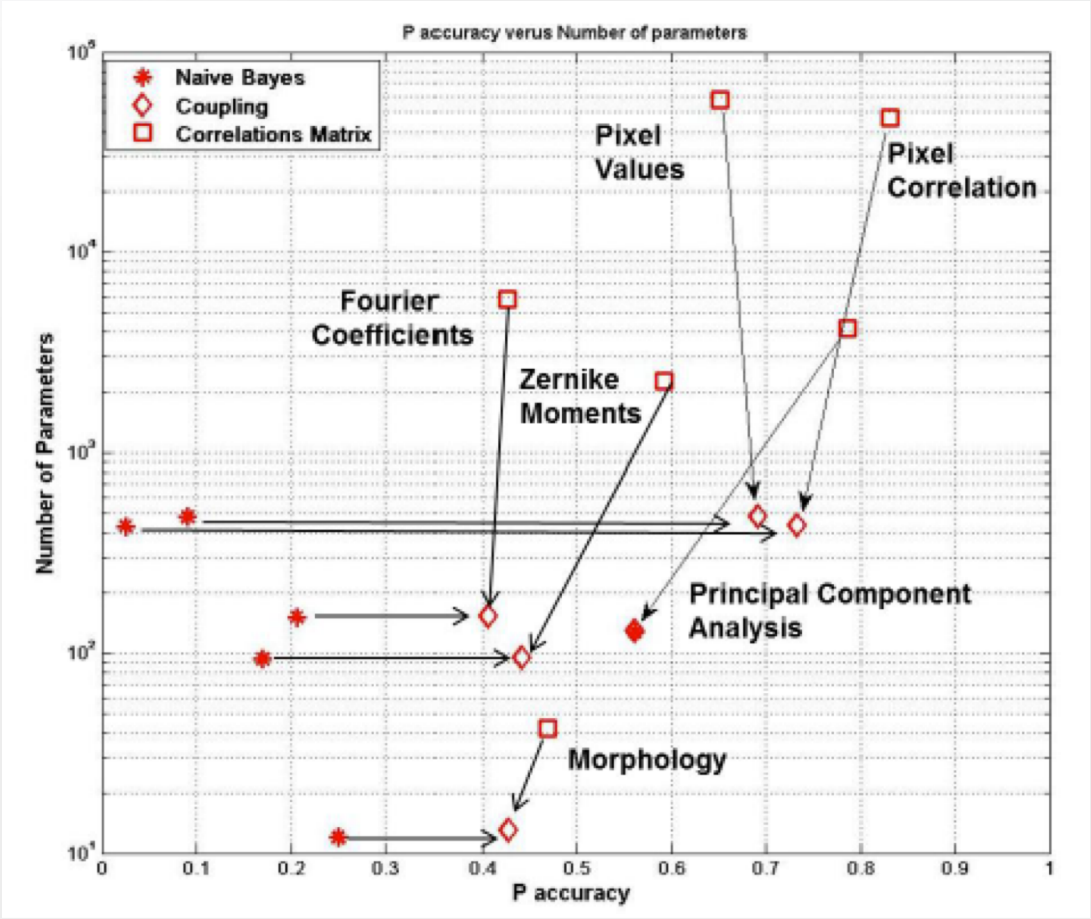
Improving Probability Accuracy & Robustness

UC Davis Machine Learning Repository
Multiple Features Data Set



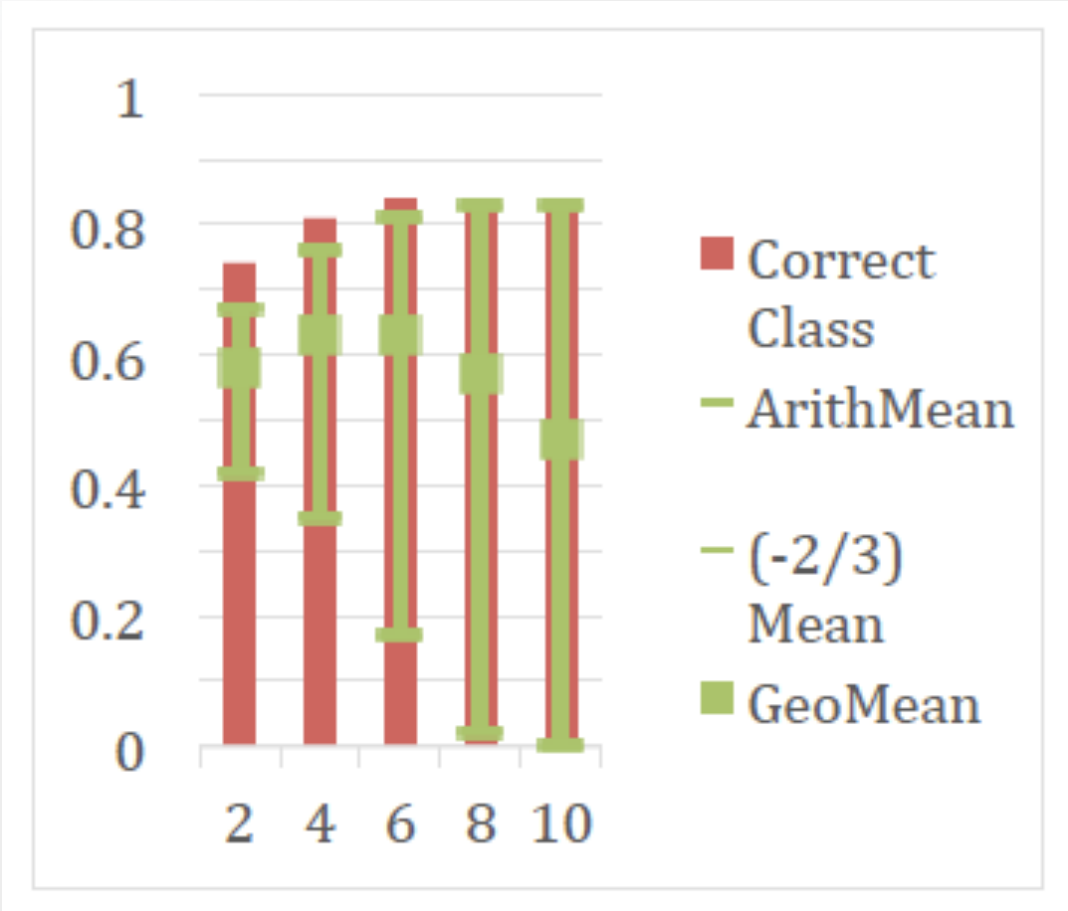
Features Modeled w Gaussians
76 Fourier; 216 correlations;
64 PCA; 240 Pixels;
27 Zernike; 6 morphology





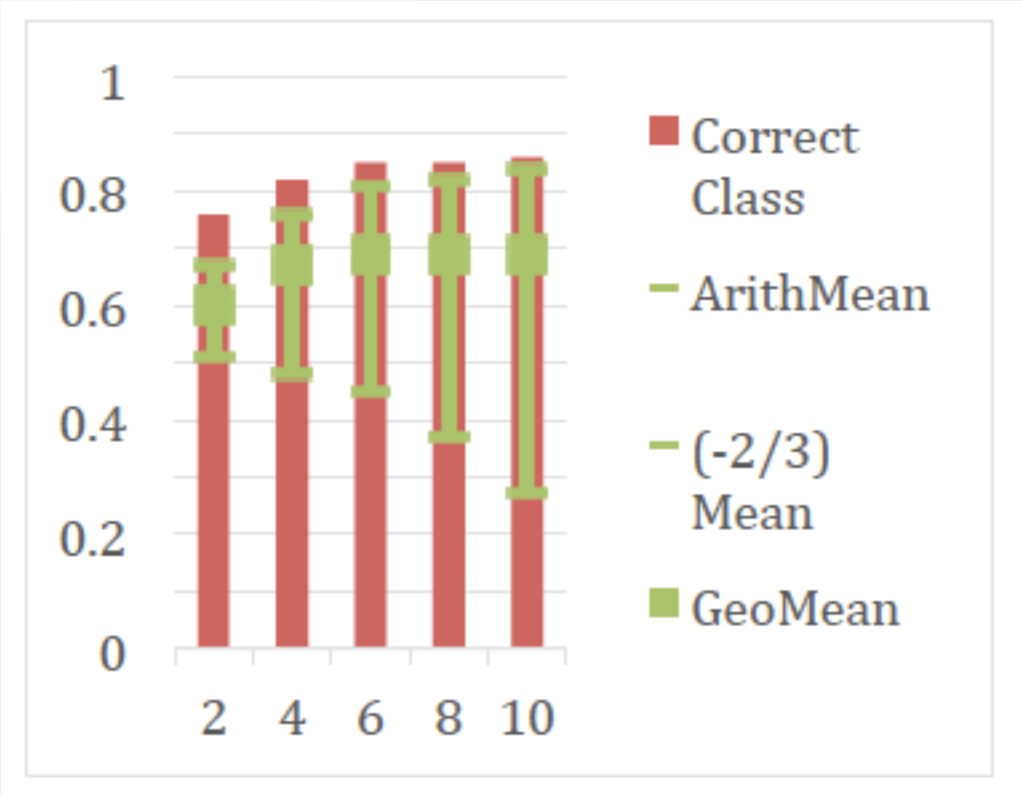
More features can improve decision and decrease accuracy of probabilities

- Source: 10 dim. Gaussian
- Model: 2 to 10 dim. Gaussian
- As features modeled increases
  - Better decisions
  - P_{acc} peaks at 6
  - P_{robust} decreases



Model with slower decaying tail preserves accuracy with higher dimensional features

- Source: 10 dim. Gaussian
- Model: 2 to 10 dim. Coupled-Gaussian $r = -0.65$
- As features modeled increases
  - Better decisions
  - Better P_{acc}
  - P_{robust} decreases modestly

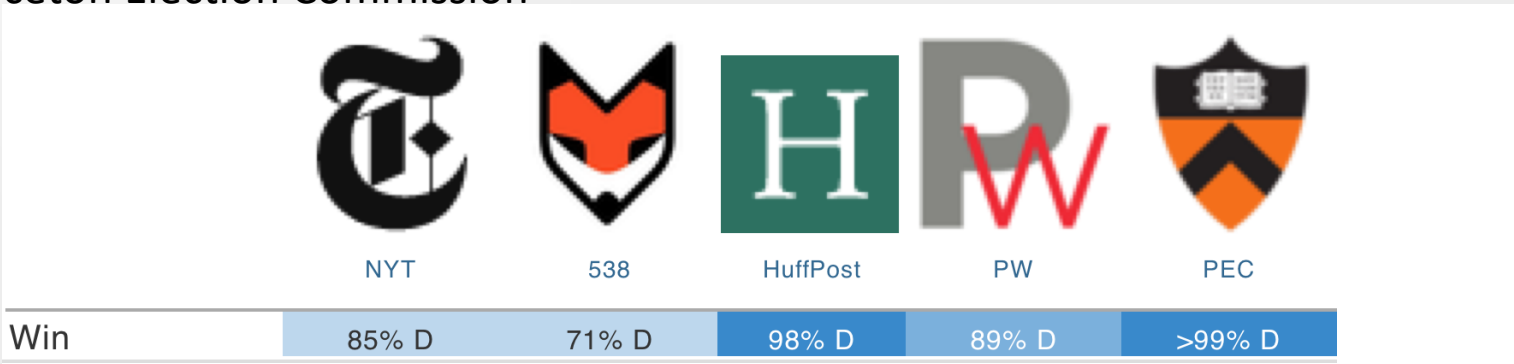


High dimensional models require tail decay slower than source

2016 Presidential Election Forecasts

- On Nov. 8, 2016 the New York Times published a table of forecasters state-by-state predictions. 5 of the forecasters provided numerical probabilities for each state or electoral district.
- New York Times, FiveThirtyEight, Huffington Post, PredictWise, and Princeton Election Commission

Forecasts

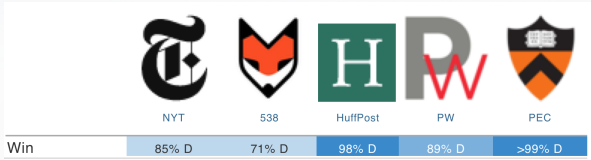


% Correct Electoral Votes

	NYT	538	HuffPost	PW	PEC
Clinton	100%	100%	100%	100%	100%
Trump	69%	68%	69%	69%	68%

Source: NYT, Nov. 8, 2016

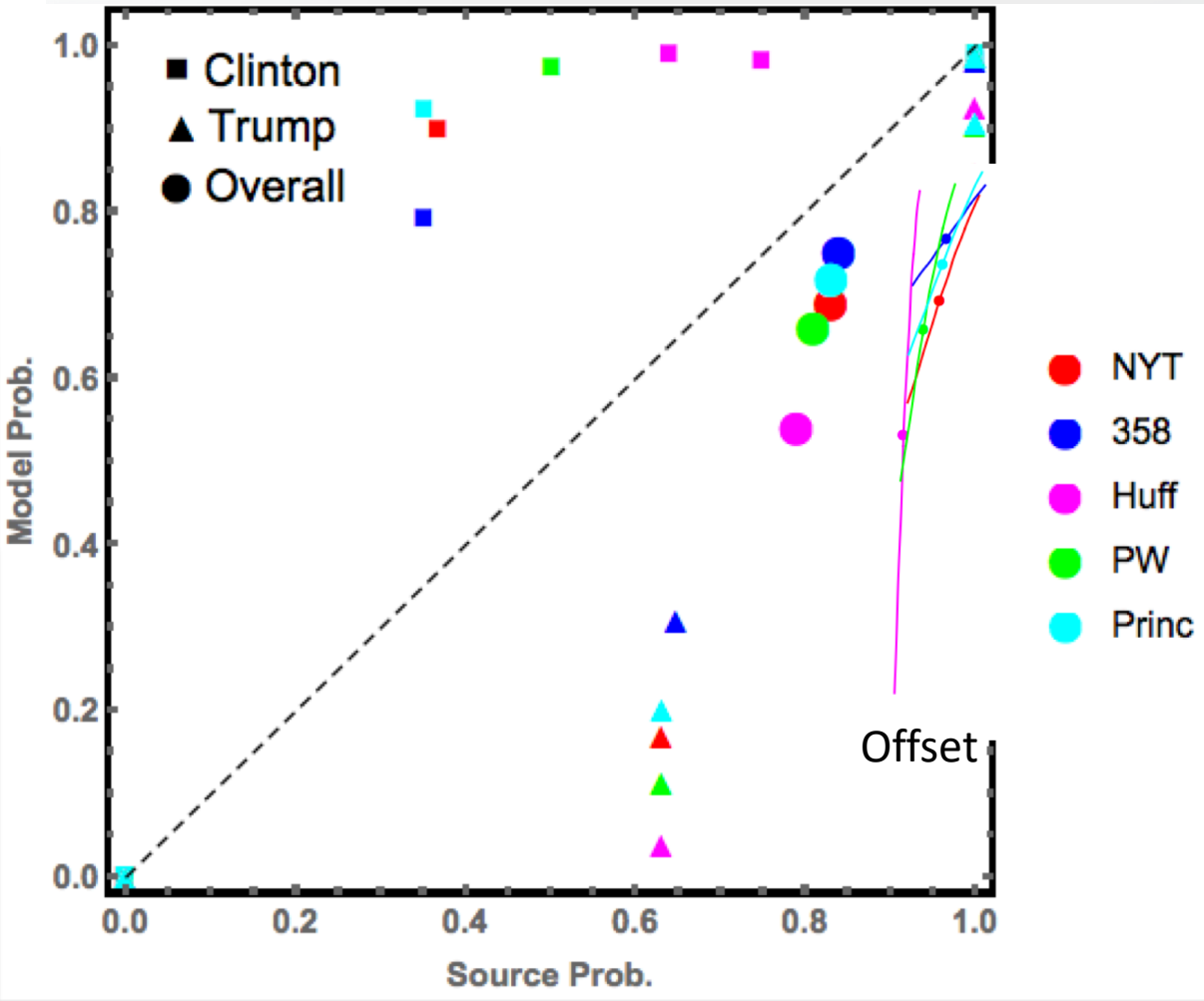
Performance of 2016 Presidential Forecasts



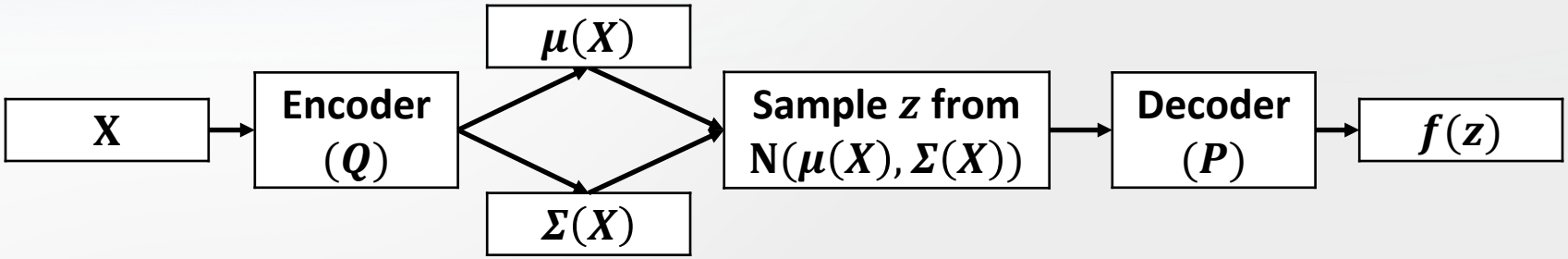
- 56 Electoral Groups in 4 bins of 14
- Weighted by number of electoral votes

Model = Source * Calib.

	Model	Source	Calib.
358	**0.75**	0.84	0.90
Princ	**0.72**	0.83	0.87
NYT	**0.69**	0.83	0.83
PW	**0.66**	0.81	0.81
Huff	**0.54**	0.79	0.68



Example Decision System: Variational Autoencoder



Input MNIST Images

Generated Image

Input image has a log-likelihood of -87,
but what does this mean and can it be improved?

D. P. Kingma & M. Welling, "Auto-Encoding Variational Bayes," in *International Conference on Learning Representations,* 2014.

Improved Variational Autoencoder

The cost functions are modified to penalize outliers in the tails

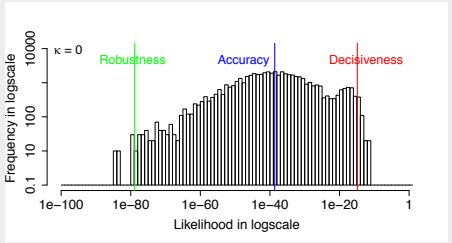
$$logp(x) \rightarrow \frac{1}{2} log_{\kappa}\left(p(x)^{\frac{-2}{1+\kappa}}\right) = \frac{1}{2\kappa}\left(p(x)^{\frac{-2\kappa}{1+\kappa}} - 1\right)$$

$$D_{KL}(q(x)||p(x)) \rightarrow \int \frac{q^{1+\frac{2\kappa}{1+\kappa}}(x)}{\int q^{1+\frac{2\kappa}{1+\kappa}}(x)dx} \cdot \frac{1}{2\kappa}\left(p^{-\frac{2\kappa}{1+\kappa}}(x) - q^{-\frac{2\kappa}{1+\kappa}}(x)\right)dx$$

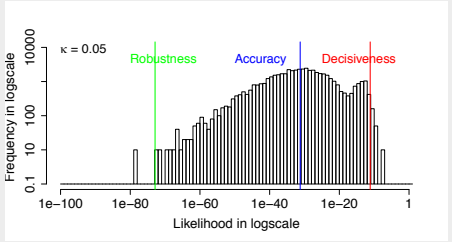
Log-likelihood of -87
Corresponds to
P = exp(-87) = 10^{-38}
Compared with
$2^{784} = 10^{236}$ possible images

Improvement in the
robustness & accuracy is
8 – 9 orders of magnitude
P_robustness: $10^{-79} \rightarrow 10^{-71}$
P_accuracy: $10^{-39} \rightarrow 10^{-30}$

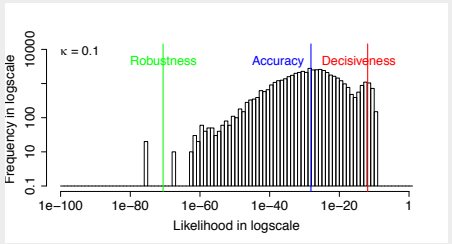
$\boldsymbol{\kappa}$ = 0.00



$\boldsymbol{\kappa}$ = 0.05

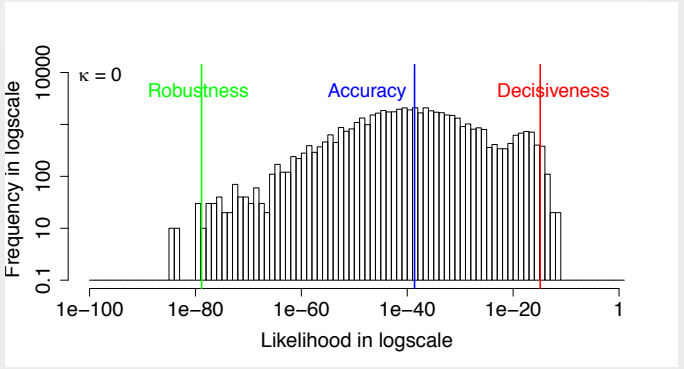


$\boldsymbol{\kappa}$ = 0.10

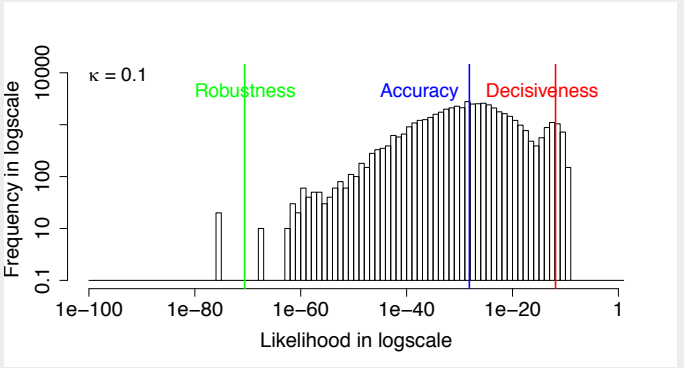


Improved training of Variational Autoencoder

$\kappa = 0.0$



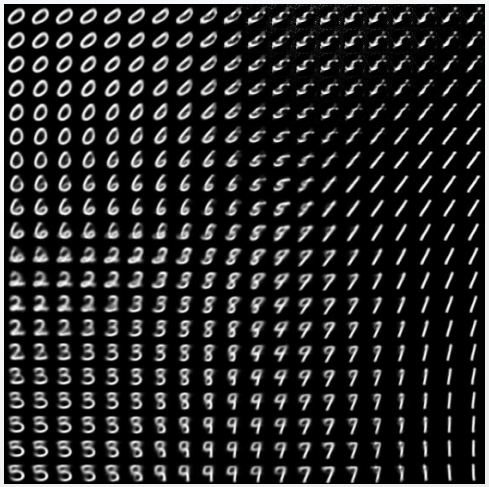
$\kappa = 0.1$

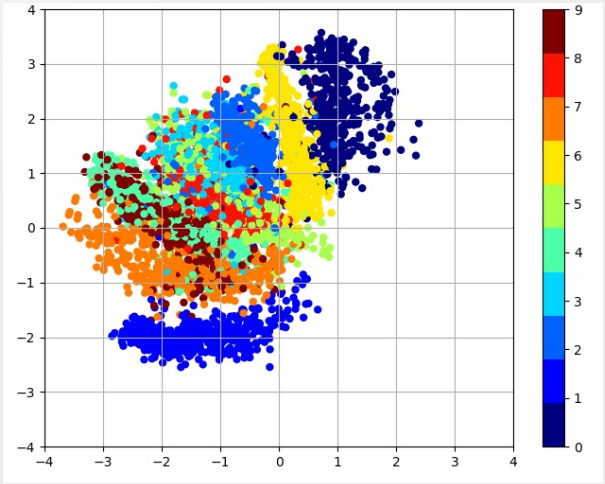


Further improvement planned by using a coupled Gaussian model

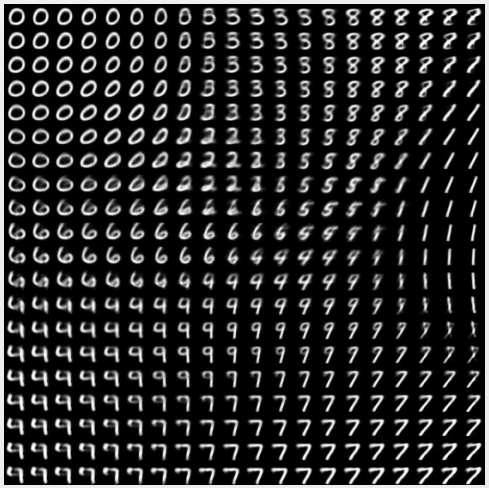
Visualization of a 2-D latent variable

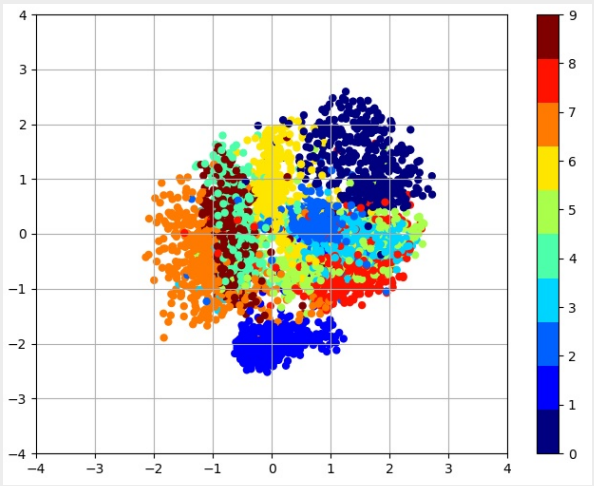
κ = 0.00





κ = 0.75





Coupled entropy loss function forms tighter clusters,
but does not improve separation of clusters