A Gaussian Mixture Model Approach to Forecast Verification

A Parametric, Feature-Based Method

Valliappa Lakshmanan[1,2] John Kain[2]

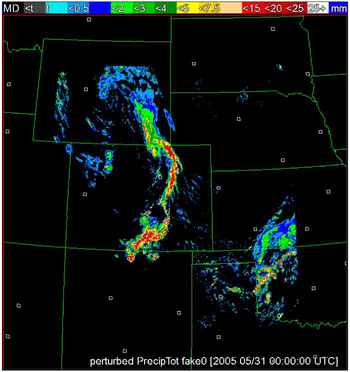
[1]Cooperative Institute of Mesoscale Meteorological Studies
University of Oklahoma

[2]Radar Research and Development Division
National Severe Storms Laboratory

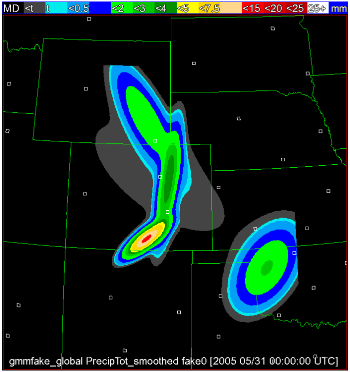
Intercomparison Workshop, Boulder, Aug. 2009

What is a GMM?

Intuitively: find an optimal way to place Gaussian functions at various points in the image such that the sum of these Gaussians mimics the input gridded field.



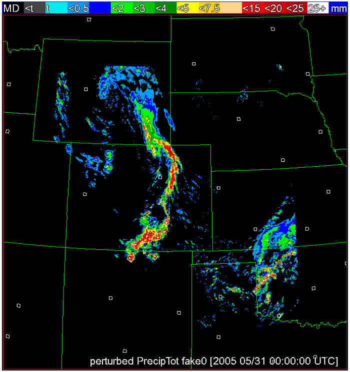
Original



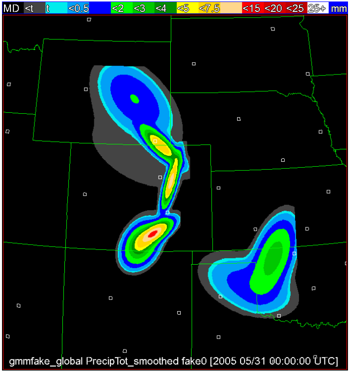
5 Gaussians

Number of Gaussians

The accuracy of fit gets better as you increase number of Gaussians.

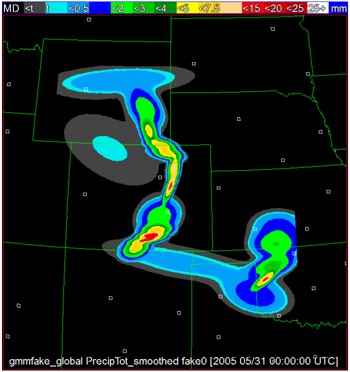


Original

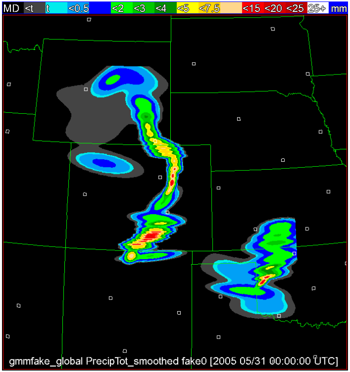


10 Gaussians

Diminishing Returns



20 Gaussians



50 Gaussians

At some point, the benefits of a parametric model are lost and you might as well just use the pixel values.

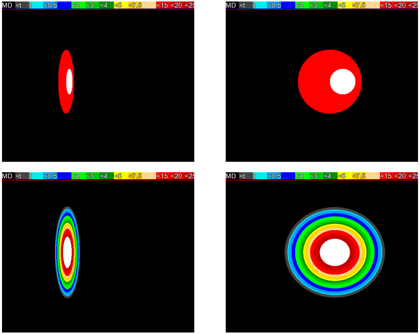
Forecast Verification with GMM

The GMM captures the ”key” features in an image (subject to Gaussian approximations). Compare the Gaussians' parameters to gain insight into how two images differ:

Parameter	Meaning	Implication
μ_x, μ_y	Center point	Translation error
$\sigma_x^2, \sigma_y^2, \sigma_{xy}$	Variance, covariance	Rotation, aspect ratio, size
π_k	Amplitude of Gaussian	Intensity error

Example Verification: Geometric

Geometric dataset from [Gilleland et al., 2009]. Chose 3 Gaussians to demonstrate that exact number is not critical.



geom000 geom003

Example Verification: Geometric

		μ_x	μ_y	σ_x^2	σ_{xy}	σ_y^2	π_k
0	Original	249	203	1720	4	128	49734
		249	203	1667	4	127	49734
		250	203	1668	9	127	49737
1	50 pts. right	249	**253**	1694	0	129	49731
		250	254	1682	4	121	49741
		250	253	1679	4	131	49732
2	200 pts. right	249	**404**	1612	4	126	49739
		250	403	1682	4	127	49735
		250	403	1760	0	129	49731
3	125 pts. right too big	250	**339**	1696	9	**2110**	**167034**
		249	340	1696	13	2048	167018
		250	341	1647	4	2021	167032

Example Verification: Geometric

		μ_x	μ_y	σ_x^2	σ_{xy}	σ_y^2	π_k
4	125 pts.	249	**341**	**104**	1	**2046**	49736
	right	249	340	101	1	2027	49729
	turned	250	339	105	2	2120	49740
5	125 pts	249	**355**	1678	17	**8271**	**323126**
	right	250	356	1688	34	8203	323125
	huge	250	356	1668	16	8265	323121

Yes, but ...

Three questions:

1. How does this compare to Method X ...?
2. How do you fit a Gaussian Mixture Model to data?
3. How well does it do on real (not synthetic) data?

Advantages of GMM approach

1. Splits, merges happen automatically if needed for optimal fit.
2. The Gaussian is a parametric function: a highly compressed view of the information in the data
3. Number of Gaussians used a good measure of the scale at which the image is being represented.
4. Transformations of Gaussians correspond to easily identifiable changes in their parameters.

How the concept compares to ...

Technique	Similarity	Difference
Filtering	Naturally incorporates scale	Compare model parameters not pixel values
Object	Compare ”features”	No thresholds, split/merge problems, etc.

How the concept compares to ...

Wavelets	Wavelets provide multiresolution i.e. *images* at different scales; GMM provides *objects* at different scales
Field deformation	Optical flow approaches are non-parametric; you get an answer at each grid point.

Why consider a GMM approach?

Multi-scale like filtering-based methods
Transformation-detecting like object-based methods
Simple to implement
Mathematically **elegant**

How to fit a GMM

1. Initialize the GMM.
2. Carry out Expectation-Minimization (EM) algorithm to iteratively ”tune” the GMM.
3. Store the parameters of each Gaussian component of the GMM.

The GMM

The GMM is defined as a weighted sum of K two-dimensional Gaussians:

$$G(x, y) = \sum_{k=1}^{K} \pi_k f_k(x, y) \quad (1)$$

$$f(x, y) = \frac{1}{2\pi\sqrt{|\Sigma_{xy}|}} e^{-((x-\mu_x)(y-\mu_y))\Sigma_{xy}^{-1}((x-\mu_x)(y-\mu_y))^T/2} \quad (2)$$

Σ_{xy} is:

$$\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} \quad (3)$$

The Expecation-Minimization (EM) method

Given a set of points x_i, y_i, it is possible to fit these points to a GMM, $G(x, y)$, by following an iterative method known as the expectation-minimization (EM) method.
Assume that an initial choice of parameters $\mu_{x_k}, \mu_{y_k}, \Sigma_{xy_k}$ exists for each of the K components.

$$P(x_i, y_i|\theta) = \sum_{k=1}^{K} \pi_k f_k(x_i, y_i|\mu_{x_k}, \mu_{y_k}, \Sigma_{xy_k}) \quad (4)$$

E-M method

E-step:

$$P(k|x_i, y_i, \theta) = \frac{\pi_k f_k(x_i, y_i|\mu_{x_k}, \mu_{y_k}, \Sigma_{xy_k})}{P(x_i, y_i|\theta)} \tag{5}$$

M-step:

$$\mu_x = E(x) = \frac{\sum_{i=1}^{N} (P_k(x_i, y_i) x_i)}{\sum_{i=1}^{N} P_k(x_i, y_i)} \tag{6}$$

$$\begin{pmatrix} E((x-\mu_x)^2) & E((x-\mu_x)(y-\mu_y)) \\ E((x-\mu_x)(y-\mu_y)) & E((y-\mu_y)^2) \end{pmatrix} \tag{7}$$

$$\pi_k = \frac{1}{N} \sum_{i=1}^{N} P_k(x_i, y_i) \tag{8}$$

The EM method is problematic

The EM process has to be bootstrapped with some initial guess at a GMM.
The EM process will start at that point and slowly climb towards the local maximum in likelihood space.
Only promises a local maximum

Initializing the GMM

A significant amount of spatial coherence in weather images that we can take advantage of to place the initial mixture components:

1. Group pixels into contiguous regions
2. Randomize the pixels within a region (to remove order-dependence)
3. Arrange regions in order (maybe in order of top-left point, or of centroid, or K-means cluster centroids)
4. Break pixel list into K equal parts
5. If a pixel falls into the kth group, the initial weight is one for the kth component and zero for all other components.

Intensity?

Recall that the GMM was defined so as to sum to 1, and that the EM method optimized the likelihood of the parameters given the *positions* of the pixels (and not the intensity).
Two minor changes:

1. The total intensity associated with all the pixels in the image is used to scale the GMM
2. More intensive locations are repeated several (m) times:

$$m = 1 + round(\frac{CDF(I_{xy})}{freq(I_{mode})}) \forall I_{xy} < I_{mode} \tag{9}$$

Number of components?

Traditional way to estimate number of components:

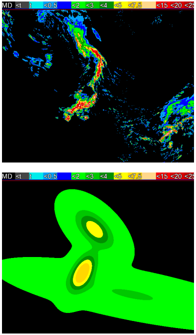
$$BIC = 2l(\theta) - 6Klog(N) \tag{10}$$

$$l(\theta) = \sum_{i=1}^{N} log(P(x_i, y_i)) \tag{11}$$

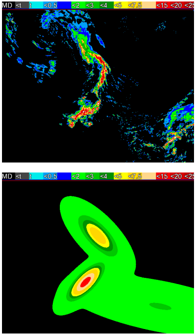
Doesn't work: hundreds of components.
We just used 3.

Perturbed dataset

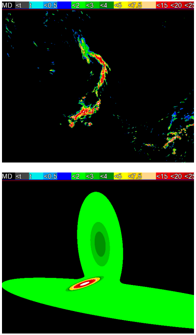
2km WRF from CAPS perturbed (See [Gilleland et al., 2009]).



fake000



fake003



fake007

GMM parameters for perturbed cases

	μ_x	μ_y	σ_x^2	σ_{xy}	σ_y^2	π_k
Original	176	289	1305	743	1328	26437
	309	252	1272	482	665	26437
	379	407	1456	3919	20490	26437
3 pts. right	**181**	**292**	1306	743	1328	26437
5 pts. down	314	255	1270	490	675	26437
	384	410	1456	3918	20424	26437
6 pts. right	**186**	**295**	1307	744	1329	26437
10 pts. down	319	258	1269	496	675	26437
	389	414	1472	3928	20348	26437
12 pts. right	**195**	**299**	1206	840	1133	27101
20 pts. down	340	261	774	578	767	34201
	416	495	1051	1900	10252	17843

GMM parameters for perturbed cases

Description	μ_x	μ_y	σ_x^2	σ_{xy}	σ_y^2	π_k
24 pts. right	**212**	**311**	1059	813	1111	26527
40 pts. down	354	276	1239	802	837	33773
	432	483	1347	3110	13743	17566
48 pts. right	**250**	**335**	968	801	1121	25113
80 pts. down	387	304	1772	1052	934	33256
	452	447	1405	4659	20003	15666
12 pts. right	**192**	**298**	1096	859	**1198**	**33338**
20 pts. down	335	263	1178	773	829	42294
times 1.5	412	483	1264	2538	12634	22304
12pts. right	**222**	**306**	**2355**	194	**459**	**17815**
20 pts. down	345	258	79	162	486	20620
minus 2 mm	409	431	755	2884	20770	15932

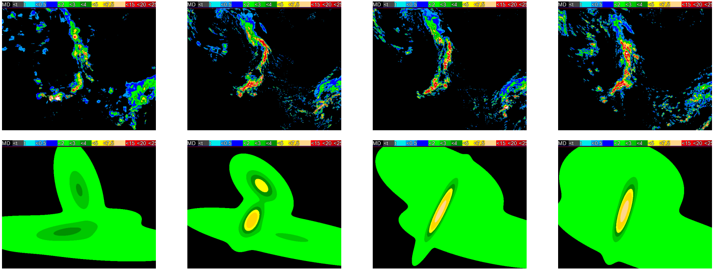
June 1, 2005

Three quite different systems: Plains, NW, SE 3-member GMM fit does not capture these three events. Instead, NW ignored.

As pointed out by [Wernli et al., 2009], it would be advantageous to carry out this analysis on smaller domains where only one type of of meteorological system predominates.

Higher order GMM fits do capture all these systems. We chose to use only a 3rd order fit so as to keep the analysis of member parameters tractable.

The GMM



Observed 2CAPS 4NCAR 4NCEP

GMM Parameters

	μ_x	μ_y	σ_x^2	σ_{xy}	σ_y^2	π_k
Obs.	193	301	3546	841	936	22136
	350	264	684	1218	7508	22616
	383	309	921	2032	22181	20061
2CAPS	176	289	1305	743	1328	26437
	309	252	1272	482	665	26437
	379	407	1456	3919	20490	26437
4NCAR	159	260	3134	2344	7636	16464
	277	264	3369	1607	932	39139
	379	461	1729	2840	14879	21068
4NCEP	168	247	3518	747	6888	23002
	278	258	3153	906	484	43675
	405	416	3920	6740	24879	20010

First Gaussian Component

1st Gaussian component corresponds to the Northern Great Plains.

	μ_x	μ_y	σ_x^2	σ_{xy}	σ_y^2	π_k
Observed	193	301	3546	841	936	22136
2CAPS	176	289	1305	743	1328	26437
4NCAR	159	260	3134	2344	7636	16464
4NCEP	168	247	3518	747	6888	23002

All 3 forecasts displaced to the north and west; 2CAPS is the least displaced.
4NCAR underestimates the precip; 2CAPS overestimates; 4NCEP gets it correct.
2CAPS gets shape wrong; 4NCAR, 4NCEP get north-south extent correct but overestimate east-west extent.

Second Gaussian Component

Corresponds to Southern Great Plains into Texas.

	μ_x	μ_y	σ_x^2	σ_{xy}	σ_y^2	π_k
Obs	350	264	684	1218	7508	22616
2CAPS	309	252	1272	482	665	26437
4NCAR	277	264	3369	1607	932	39139
4NCEP	278	258	3153	906	484	43675

All forecasts displaced to north; with 2CAPS again the best.
Shape: 4NCEP, 4NCAR overly vertical; 2CAPS also wrong orientation, but better.
Intensity: 2CAPS is closest; 4NCAR, 4NCEP significant overestimates

Third Gaussian Component

Corresponds to Southeastern US

	μ_x	μ_y	σ_x^2	σ_{xy}	σ_y^2	π_k
Obs	383	309	921	2032	22181	20061
2CAPS	379	407	1456	3919	20490	26437
4NCAR	379	461	1729	2840	14879	21068
4NCEP	405	416	3920	6740	24879	20010

All get intensity and orientation correct but displaced to the east; 4NCEP also displaced to north.
4NCEP too large in north-south direction: precipitation even in correct in aggregate is spread over too large an area.

Areas for further exploration

We presented a GMM approach to model verification, not a full-fledged verification technique. Ideally, a verification technique is fully automated and objective.

1. *Number of components* An information criterion more amenable to the forecast verification problem needs to be developed.
2. *Scale* The correspondence of the number of components to the scale at which model verification is carried is very inexact.
3. *Automated analysis* In order to use more than 3 components, rules to identify and analyze GMM parameter changes automatically have to be developed.

Areas for further exploration

1. *Association or Deformation?* Matching Gaussian components across images may become hard with more than 3 components. An alternative approach: start the E-M on the forecast field with the GMM that corresponds to the observed field and observe how the GMM components get deformed.
2. *Initialization of EM* Exploration into other algorithms for initializing the EM process may prove beneficial.
3. *Low intensity regions* should not be ignored. Perhaps break up large spatial areas into smaller areas and then fit GMMs to them . . .

Acknowledgements

Funding for this research was provided under NOAA-OU Cooperative Agreement NA17RJ1227.
The GMM fitting technique described in this paper has been implemented within the Warning Decision Support System Integrated Information (WDSSII; [Lakshmanan et al., 2007]) as part of the w2smooth process. It is available for download at www.wdssii.org.

References

- Gilleland, E., Ahijevych, D., Brown, B., Casati, B., and Ebert, E. (2009).
  Intercomparison of spatial forecast verification methods.
  Weather and Forecasting, 0(1):DOI: 10.1175/2009WAF2222269.1.
- Lakshmanan, V., Smith, T., Stumpf, G. J., and Hondl, K. (2007).
  The warning decision support system – integrated information.
  Weather and Forecasting, 22(3):596–612.
- Wernli, H., Hofmann, C., and Zimmer, M. (2009).
  Spatial forecast verification methods intercomparison project – application of the SAL technique.
  Weather and Forecasting, 0(2):DOI: 10.1175/2009WAF2222271.1.