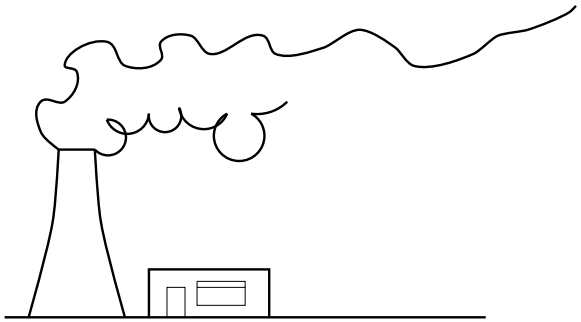
Adventures in High Dimensional Data Analysis: Hyperspectral Gaseous Plume Detection

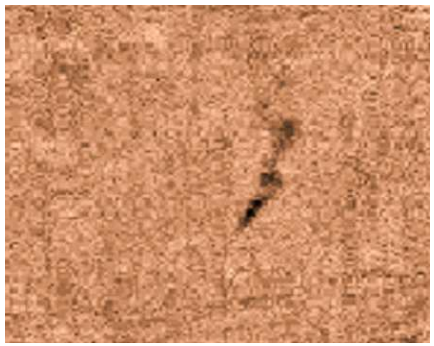
IPAM, UCLA, 18 July 2005



James Theiler*
Space and Remote Sensing Sciences
Los Alamos National Laboratory

***with Bernard R. Foy and Andrew M. Fraser**

- **Goal is to identify, locate, and characterize weak gaseous plumes**
- **Challenge is the spatial and spectral variability of the background (aka clutter)**
  - **0. Baseline: linear signal in gaussian noise**
  - **I. What is the effect of including plume in the clutter?**
  - **II. What if the clutter is not gaussian?**
  - **III. Can machine learning play a role?**



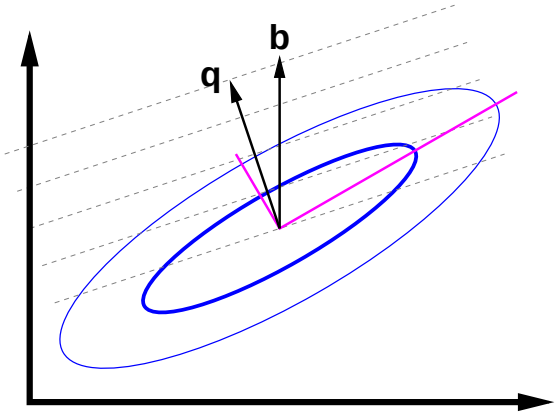
Linear Algebra Formulation

- Signal assumed to superpose linearly with clutter and noise
  - $\mathrm{r} = \text{signal} + \text{clutter} + \text{noise}$
  - signal $= \epsilon \mathrm{b}$ has known signature b, unknown quantity ϵ
  - goal is find filter q to estimate $\epsilon \sim \mathrm{q}^T \mathrm{r}$
- Simple Matched filter: ignore clutter, treat noise as white
  - $\mathrm{q} = \mathrm{b}$
- (Adaptive) Matched Filter: clutter+noise is gaussian
  - Compute covariance from off-plume pixels: $\boldsymbol{K} = \langle (\boldsymbol{r} - \langle \boldsymbol{r} \rangle)(\boldsymbol{r} - \langle \boldsymbol{r} \rangle)^T \rangle$
  - $\mathrm{q} = \mathrm{K}^{-1}\mathrm{b}$
  - some issues in the appropriate regularization of K^{-1}
- Subspace projection: clutter confined to p-dimensional subspace
  - Given "undesired" background clutter spectra: $\mathrm{U} = [\mathrm{u}_1, \mathrm{u}_2, \ldots, \mathrm{u}_p]$
  - Radiance at a pixel: $\mathrm{r} = \epsilon \mathrm{b} + \sum_{i=1}^{p} n_i \mathrm{u}_i + \mathrm{e}$
  - $\mathrm{q} = (\mathrm{I} - \mathrm{U}\mathrm{U}^{\#})\mathrm{b}$, where $\mathrm{U}^{\#}$ is pseudo-inverse

Adaptive matched filter

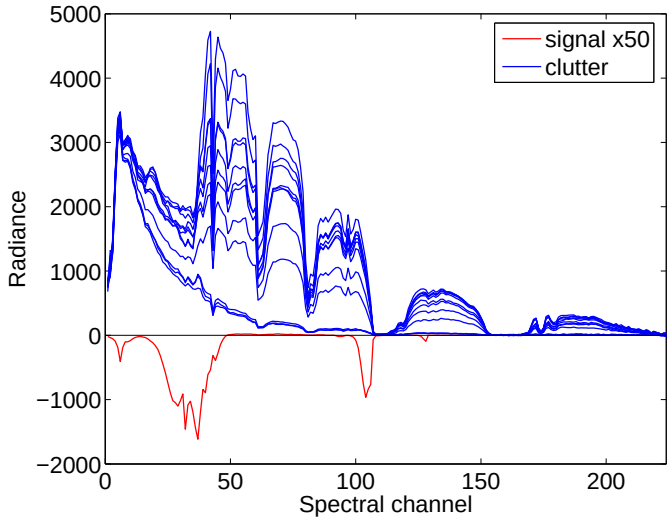
- Pixel radiance: $\mathrm{r} = \epsilon \mathrm{b} + \mathrm{z}$
  - ϵ is *unknown* and generally small; question is whether it is nonzero
  - b is *known* signature of weak signal
  - z is unknown, but
    - statistics given by mean $\mu = \langle \mathrm{z} \rangle$, covariance $\mathrm{K} = \langle (\mathrm{z} - \mu)(\mathrm{z} - \mu)^T \rangle$
- Filter is vector q; scalar $\mathrm{q}^T r$ indicates presence of plume.
  - Signal variance: $(q^T b)^2 = q^T b b^T q$
  - Clutter variance: $\langle (q^T (z - \mu))^2 \rangle = q^T \langle (z - \mu)(z - \mu)^T \rangle q = q^T K q$
- Maximize $\mathrm{q}^T \mathrm{b}$ subject to $\mathrm{q}^T \mathrm{K} \mathrm{q} = 1$:

$$\mathrm{q} = \frac{\mathrm{K}^{-1} \mathrm{b}}{\sqrt{\mathrm{b}^T \mathrm{K}^{-1} \mathrm{b}}}$$



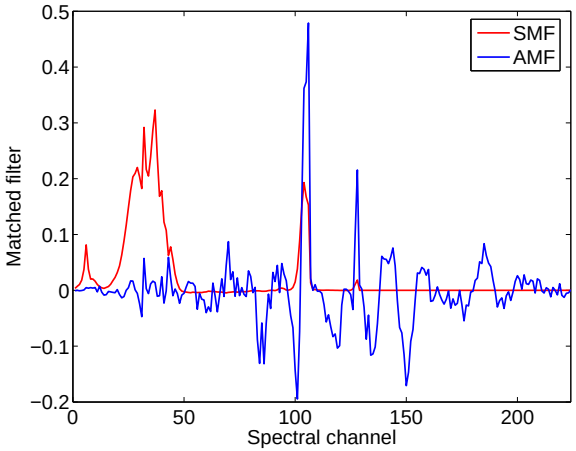
- Informally, a "compromise" between:
  - $\mathrm{q} \sim \mathrm{b}$ in direction of known signature
  - $\mathrm{q} \sim \mathrm{K}^{-1}$ in direction of small eigenvectors
- Note: it is *variances* that are optimized, not detections and false alarms

Illustration: add an artificial plume

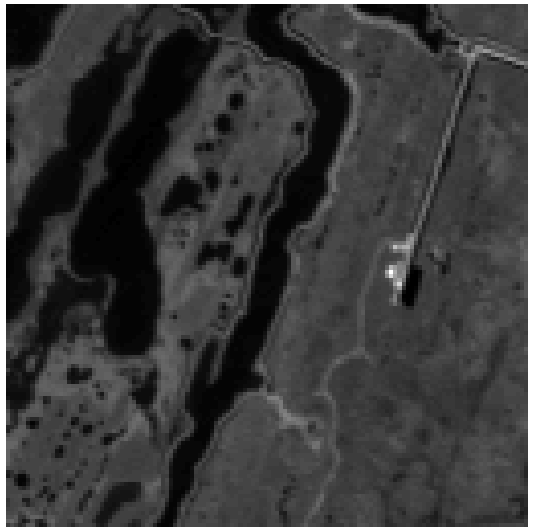


- Start with 224-channel hyperspectral image
- Artificially add a very small plume signal
- Spectrum "based on" Freon 22
- Plume signal can be positive or negative
  - emissive or absorptive

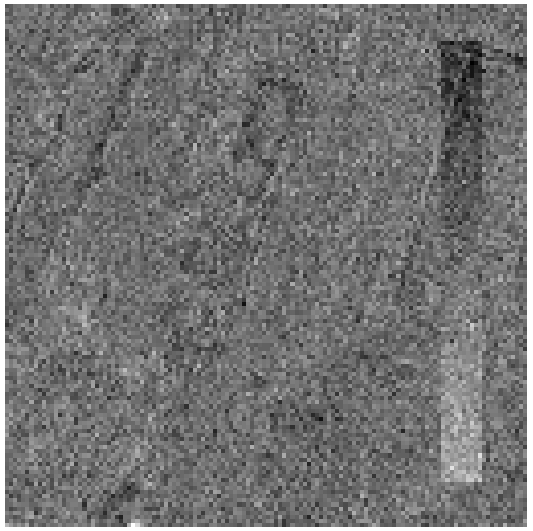
Illustration (cont'd): apply matched filters



matched filters



SMF



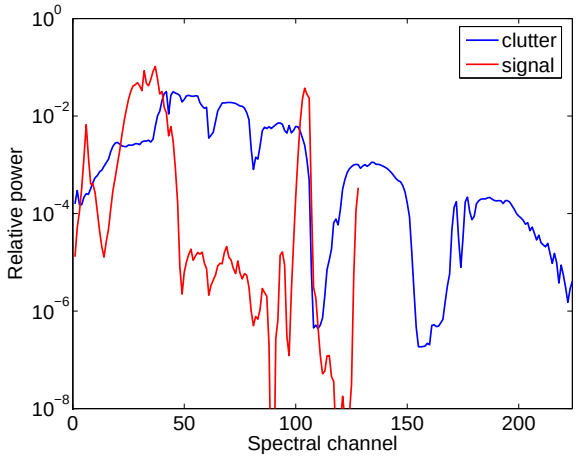
AMF

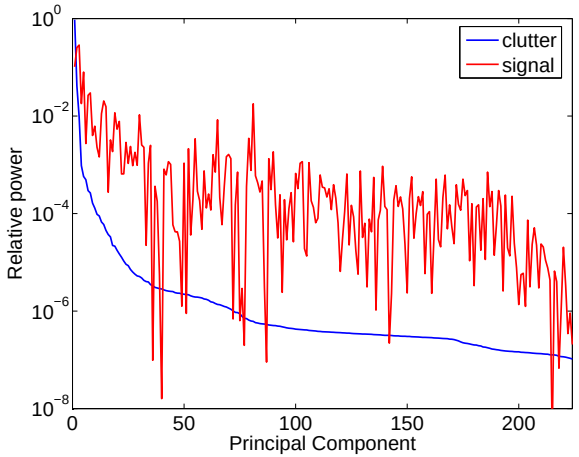
- Simple matched filter (SMF), $q = b$, fails to identify plume
  - background clutter is not supressed
- Adaptive matched filter (AMF), $q = K^{-1}b$, supresses cluttter, finds plume
- Note that AMF uses *all* the spectral channels of the hyperspectral image
  - not just those for which the gas spectrum is nonzero.
- Signal-to-clutter-ratio (SCR): AMF/SMF = 1.6×10^4

The nuggets are in the nullspace

- For plume detection, dimension reduction is the *last* thing you want to do
  - *Keep* the low-variance directions; they are sensitive to small signals
  - In fact, it is a reasonable (but suboptimal*) algorithm to project out the large-variance directions, and keep only the low-variance directions.

    *Optimum supresses large-variance directions with K^{-1}.



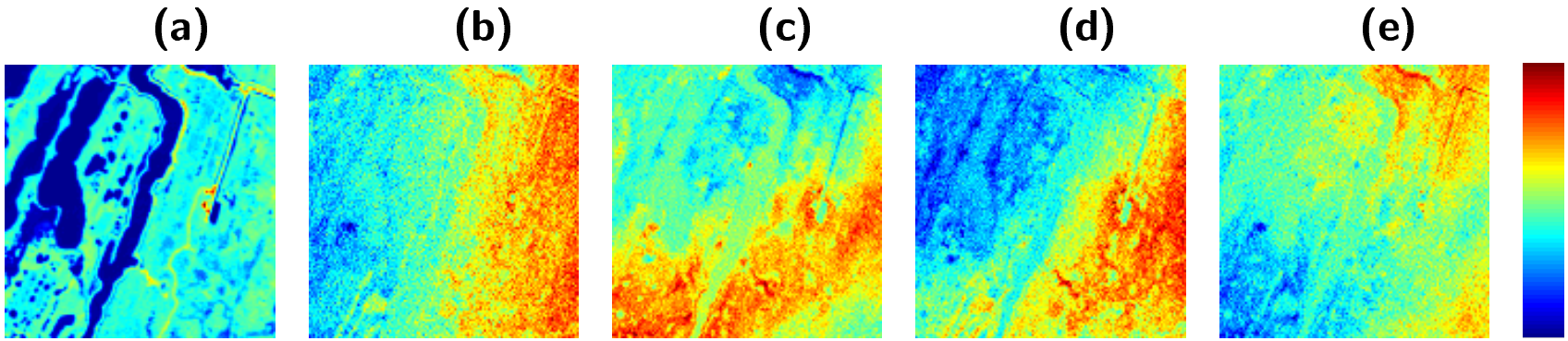


- In band-space, clutter variance is strong in nearly all channels
- In PCA-space, bulk of clutter variance is concentrated in a few channels
  - while signal maintains considerable variance across all channels

I. What is the effect of including plume in the clutter?

see: J. Theiler and B. R. Foy, "Effect of signal contamination in matched-filter detection of the signal on a cluttered background," *IEEE Geoscience and Remote Sensing Letters* (coming soon).

Hyperspectral Parlor Trick



- Can achieve spatial gradients with spectral processing

  (a) Broadband image (sum of all bands)

  (b) Linear combination of bands chosen to produce horizontal gradient

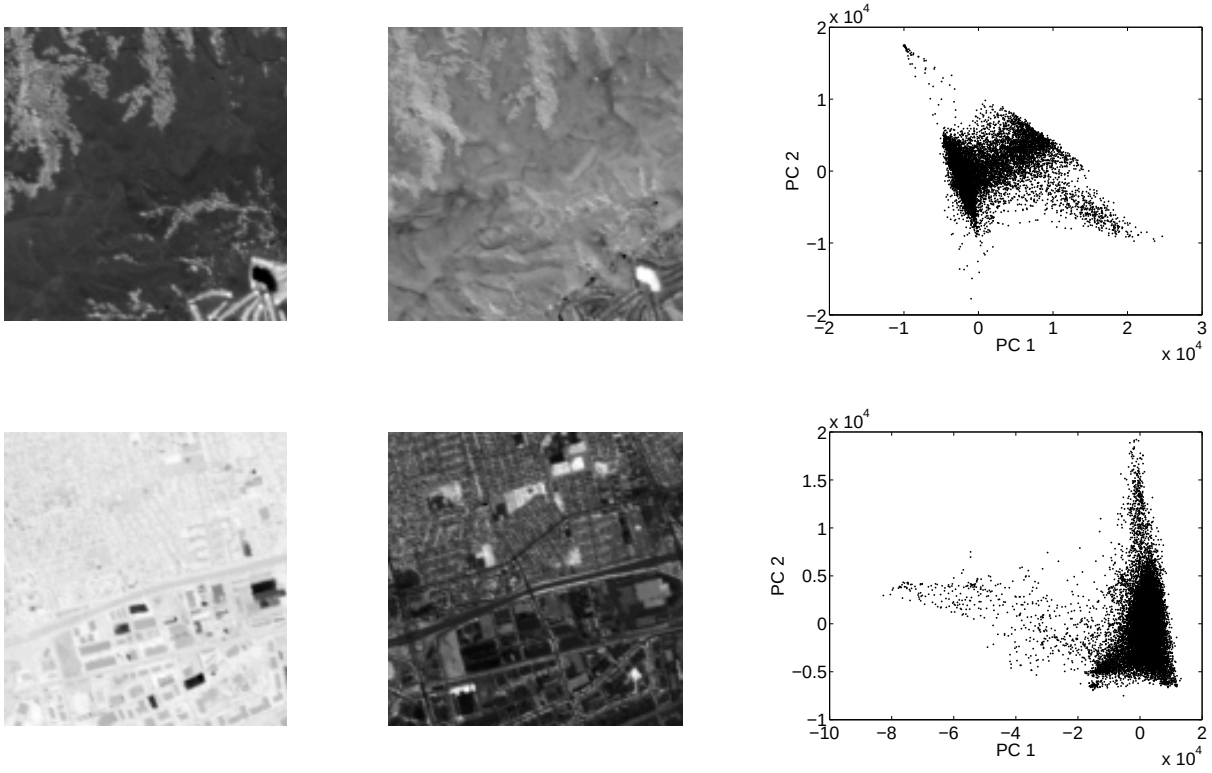
  (c) Different linear combination, produces vertical gradient

  (d,e) Diagonal gradients...

- Can achieve the same effect on random data, just much smaller

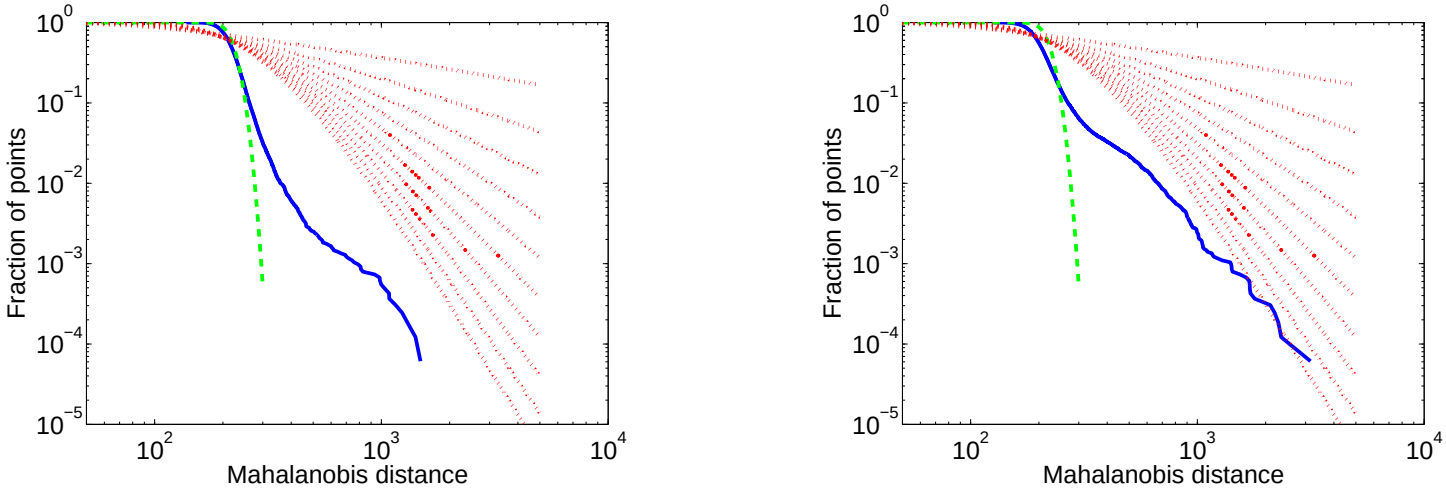
II. What if the clutter is not gaussian?

Nongaussian structure in hyperspectral data



- 128×128 chips of AVIRIS hyperspectral imagery
- Left panels show first and second principal components
- Right panels are scatterplots of two PCs
  - Triangular shapes are indicative of mixed pixels

Distribution of Mahalanobis distances in hyperspectral data



- Mahalanobis distance to centroid: $s^2 = (\mathbf{x} - \mu)^T \mathbf{K}^{-1} (\mathbf{x} - \mu)$
- Plotted is cumulative histogram of s^2 values
  - Hyperspectral data: blue solid lines
  - χ^2_d distribution: green dashed lines
    shows distribution that would be exhibited by gaussian data
  - $F_{d,\nu}$ distribution: red dotted lines
    shows distribution exhibited by multivariate t-distribution
    - $\nu = 1, 2, \ldots, 10$

Elliptically contoured (EC) distributions

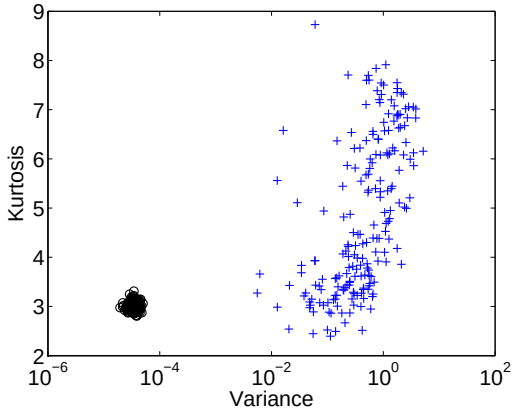
- D. Manolakis, D. Mardin, J. Kerekes, and G. Shaw, "On the statistics of hyperspectral imaging data," *Proc. SPIE* 4381, pp. 308–316, 2001.
  - Observe non-chi-squared distribution of Mahalanobis distances
  - Recommend modelling hyperspectral data with EC distributions
- Distribution described wholly in terms of Mahalanobis distance
- $P(x) \propto |K|^{-1/2} f(s^2)$ for positive scalar $s^2 = x^T K^{-1} x$.
  - e.g., gaussian: $f(s^2) = \exp(-s^2/2)$;
  - e.g., t-distribution: $f(s^2) = (1 + s^2/\nu)^{-(d+\nu)/2}$
    - Fatter tailed for smaller ν
    - gaussian in limit $\nu \to \infty$
- Only one more parameter than gaussian distribution (for t-distribution)
- Can model large dynamic range in eigenvalue spectrum
  - can exploit "thin" directions
- Data can be "whitened": $x' = K^{-1/2} x$

Experiments on the distribution of real hyperspectral data

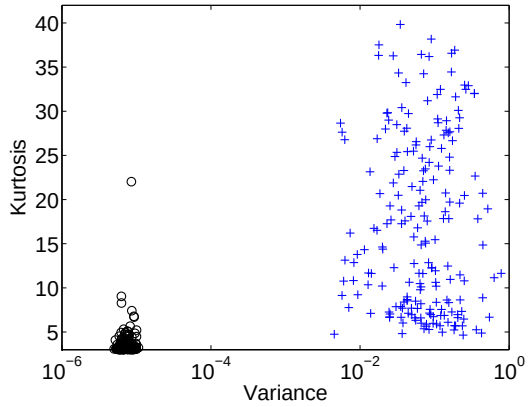
- Kurtosis $\kappa = \langle (z - \langle z \rangle)^4 \rangle / \langle (z - \langle z \rangle)^2 \rangle^2$ is a measure of nongaussian-ness:
  - Applies to a distribution of scalar data,
  - $\kappa = 3$ for gaussian distribution (of any mean/variance)
    - $\kappa > 3$ for leptokurtic; $\kappa < 3$ for platykurtic
- Experiment:
  - Repeat many times:
    - Choose random* direction q
    - Project hyperspectral data in this direction: $z_i = q^T r_i$
    - Compute kurtosis of the scalars $\{z_i\}$
    - Go ahead and compute variance of the scalars $\{z_i\}$
  - Make a scatterplot of kurtosis versus variance
- Interpret results:
  - If kurtosis ~ 3 for all b, then data may be gaussian
  - If kurtosis is consistently larger than ~ 3, then data are leptokurtic
  - If kurtosis varies systematically with variance, then data are not EC

Variance-kurtosis plots for AVIRIS data

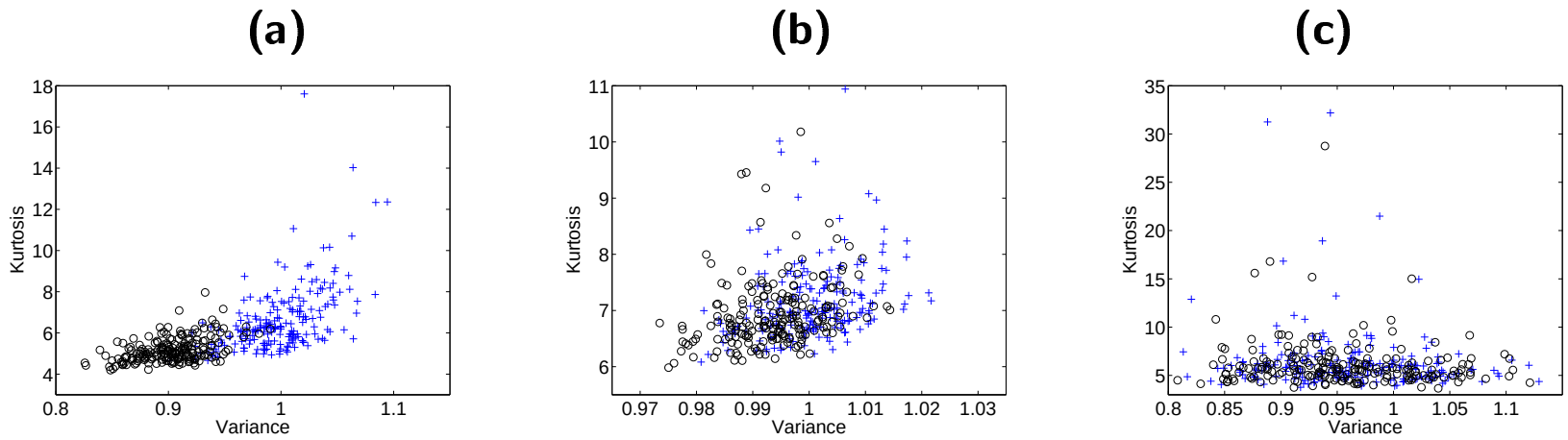
rural



urban



- Kurtosis-variance plots for rural and urban AVIRIS scenes
  - (+) Blue plus symbol indicates isotropic directions: $q = b$
  - (o) Black open circles indicate matched-filter directions: $q = K^{-1}b$
- No surprise that variance is smaller for MF directions
- Urban scene exhibits considerably fatter tails (more kurtosis) than does the rural scene.
- Kurtosis is "more gaussian" in matched-filter directions
  - Maybe matched filters are effective even for data that appears non-gaussian in its dominant large-variance directions



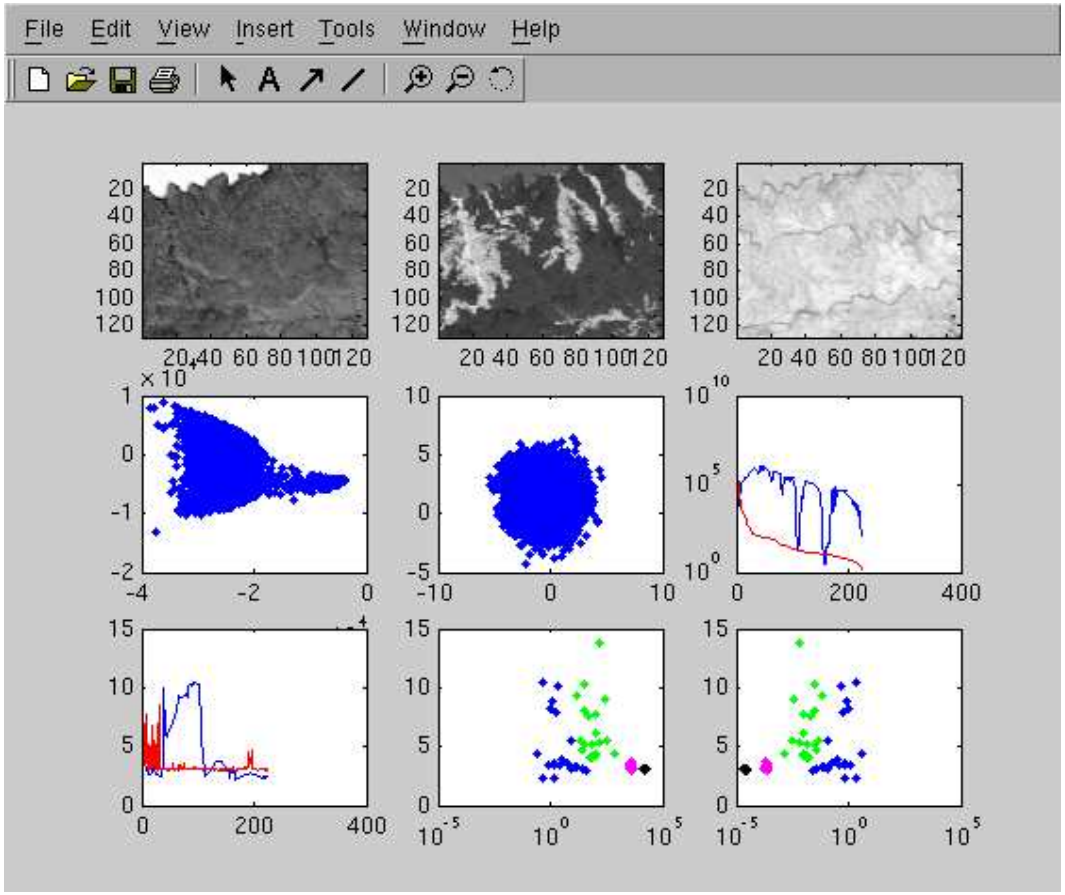
Finite sample effects

- Variance-kurtosis scatterplots for simulated spherical ($K = I$) EC data
- Since EC, there should be no variance-kurtosis correlation
- Since the covariance is spherical, the random (+) and matched-filter (o) directions should exhibit the same statistics.

(a) Naive scheme: same 64×64 data used for estimating K was used for computing variance and kurtosis.

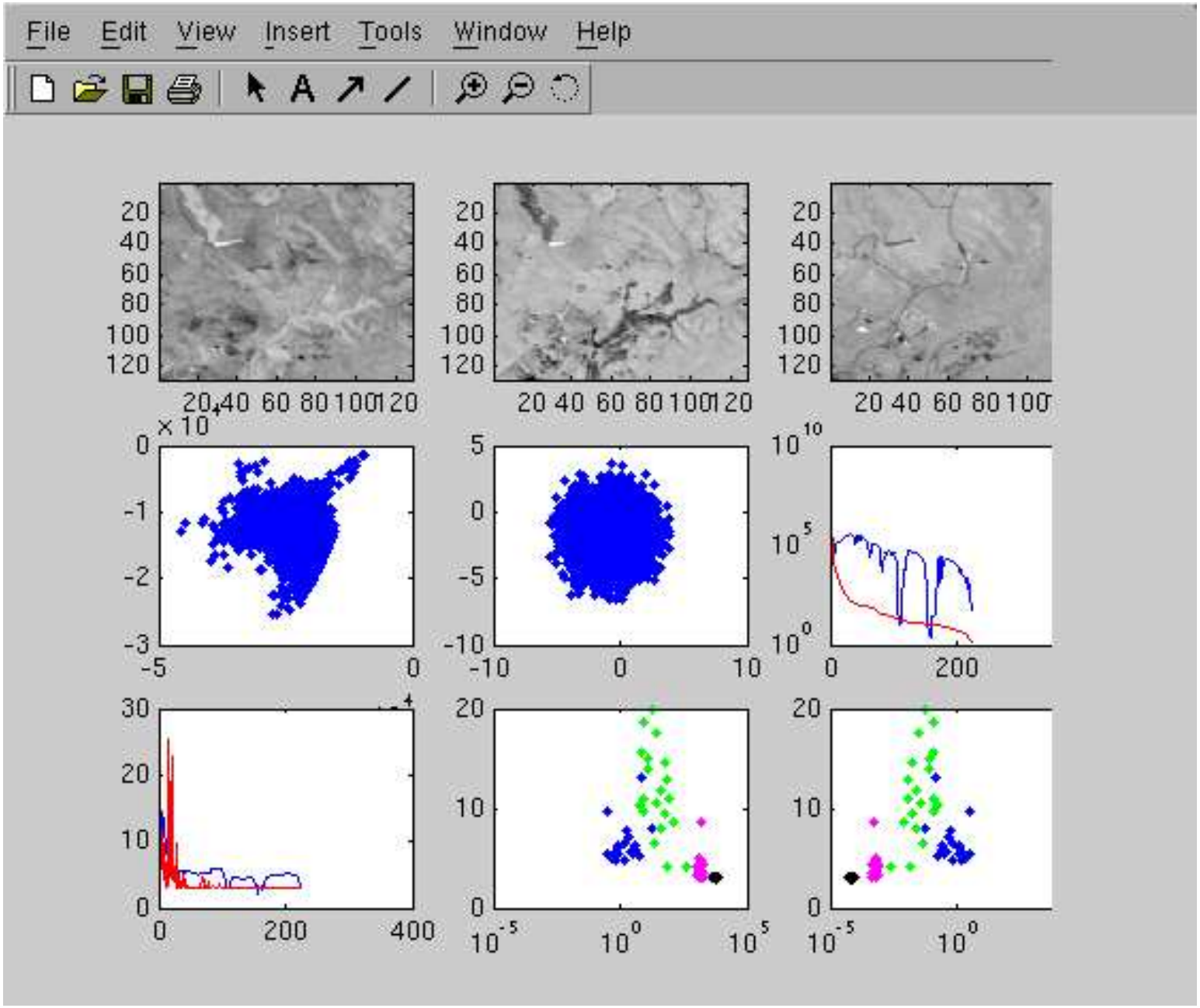
(b) Same naive scheme, but with 256×256 image

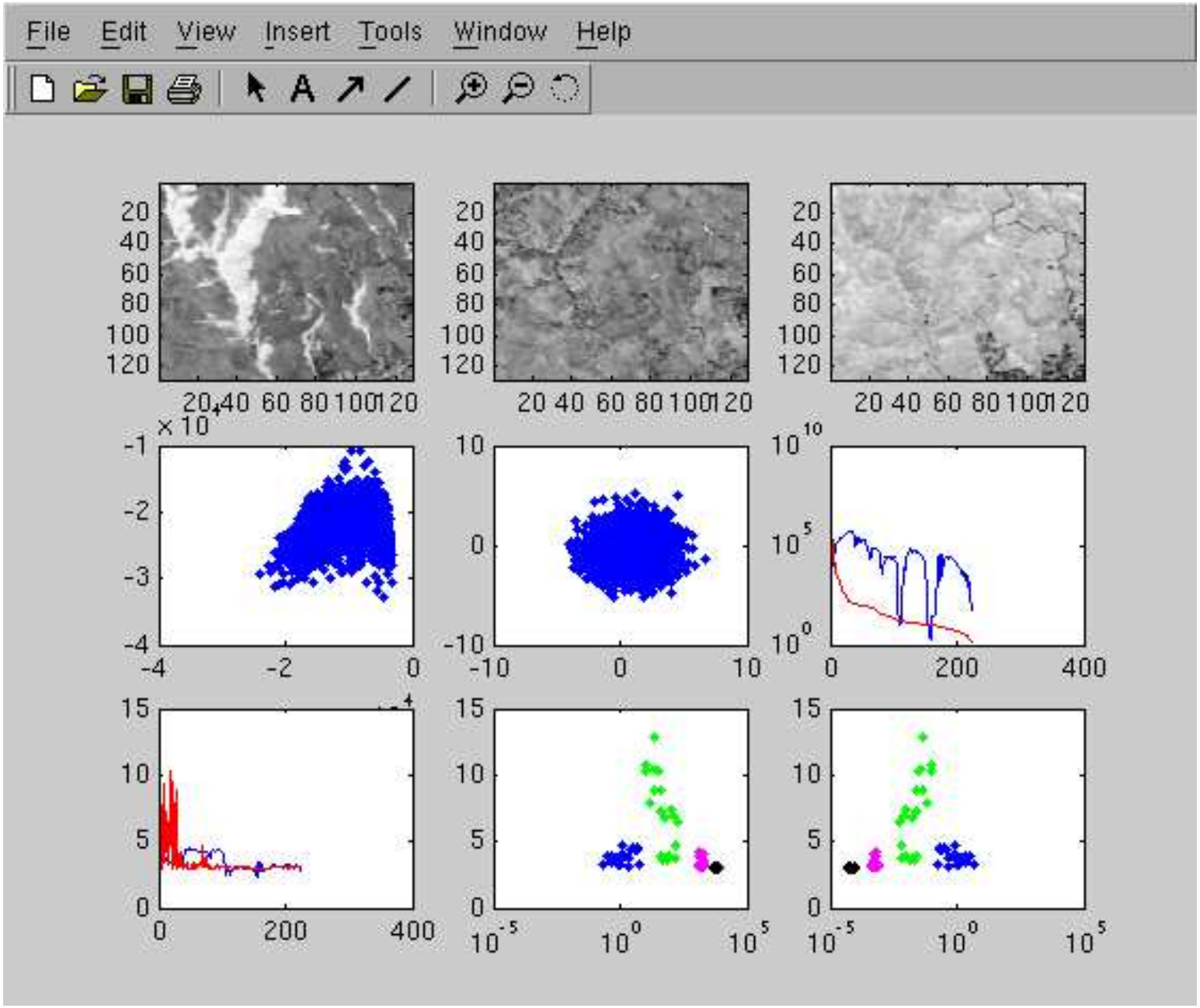
(c) A more careful scheme: 64×64 tile partitioned into three areas: one for estimating K, one for variance, and one for kurtosis.

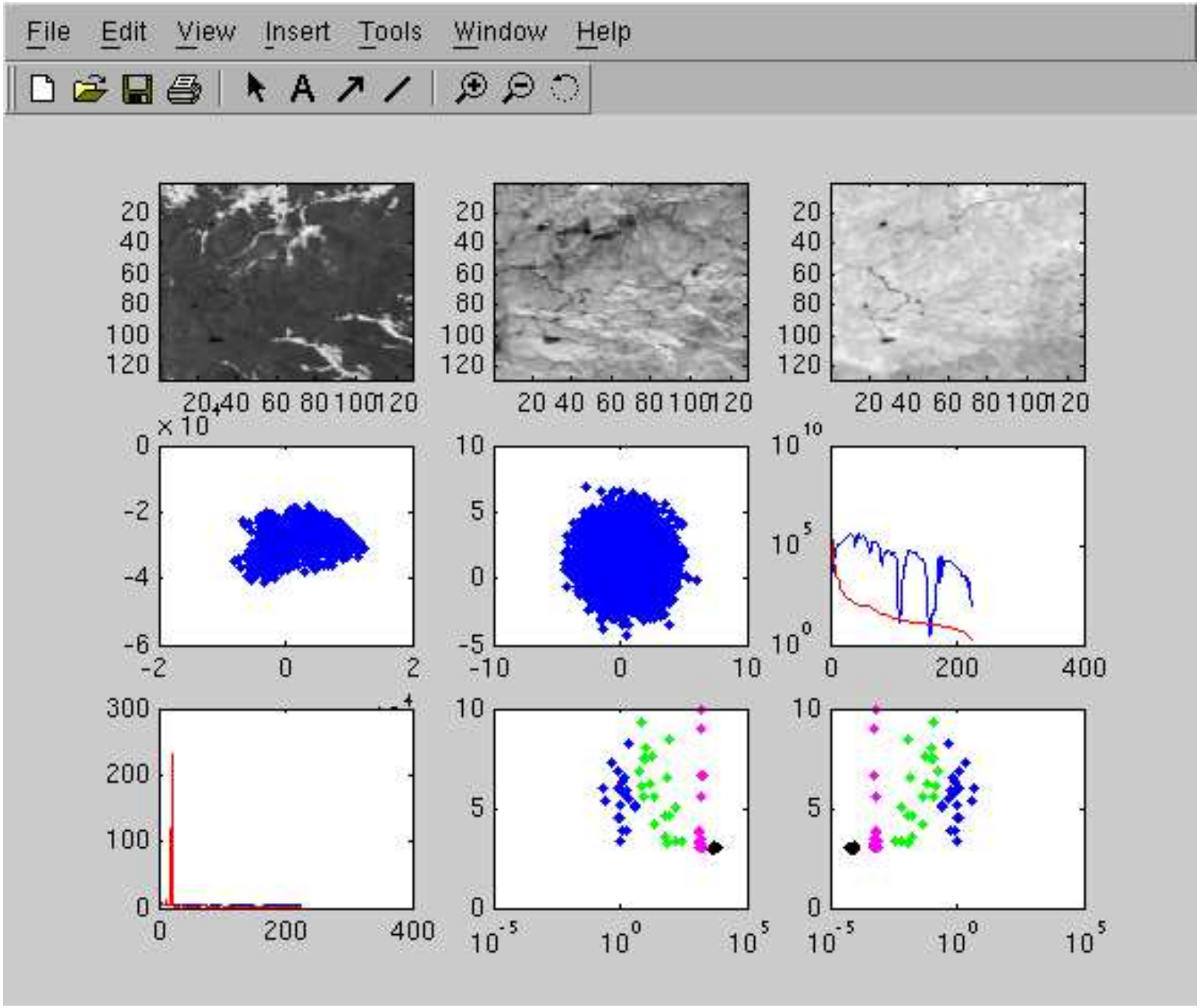


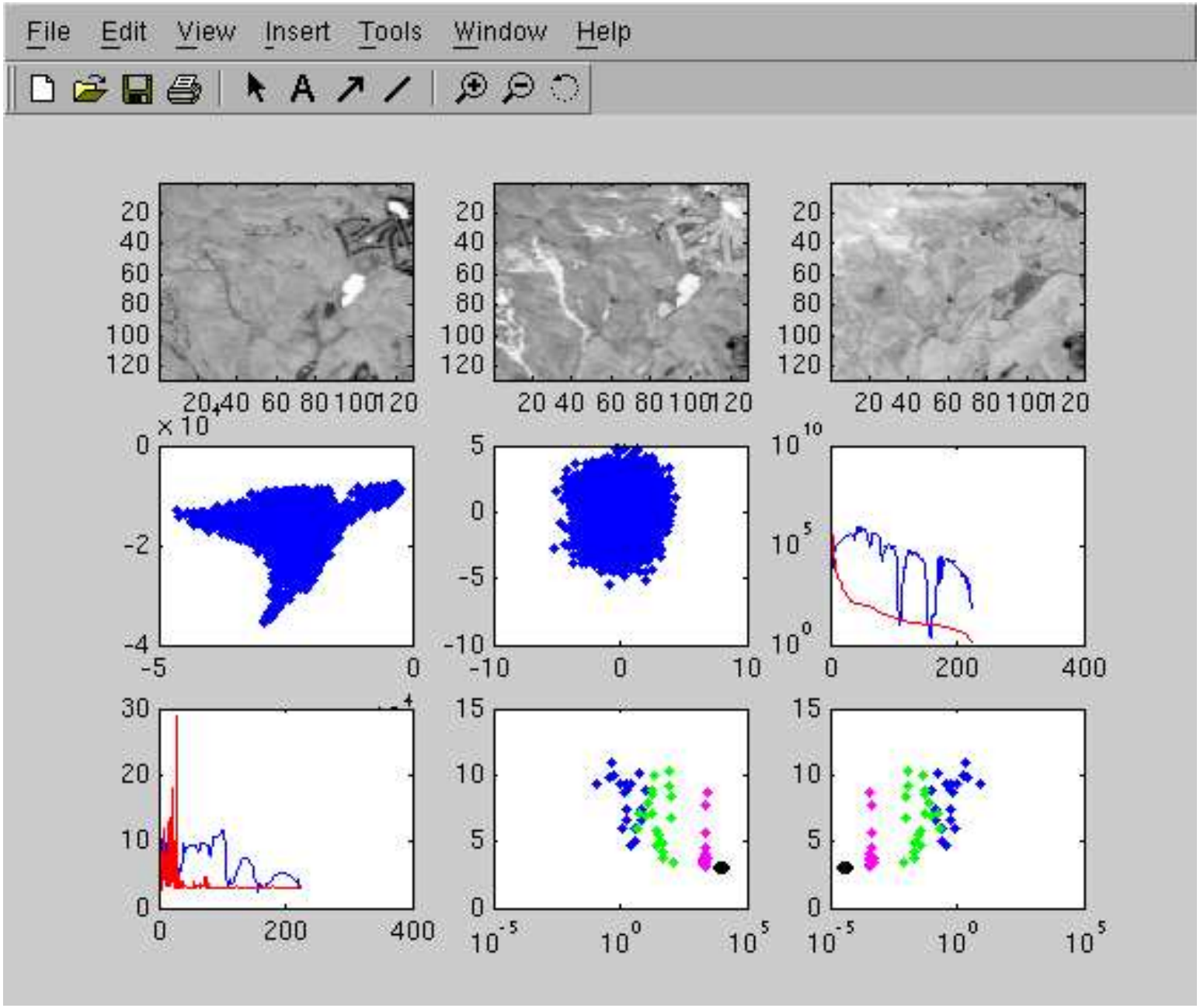
Kurtosis versus Variance in different directions

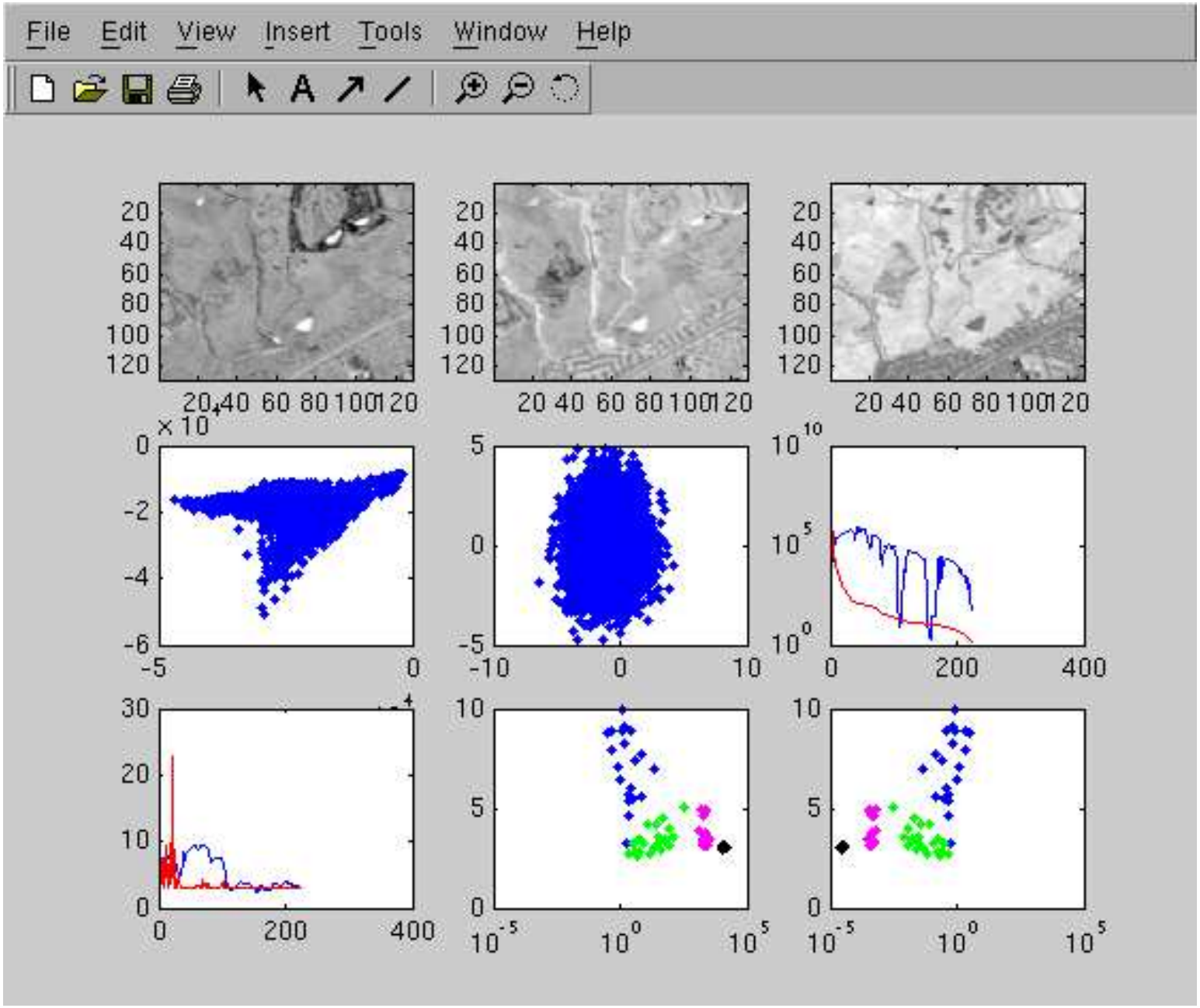
- Top three panels are first three principal component images
- Middle left panels are scatter plots with first two PCs, last two PCs
- Next two panels are variance, then kurtosis, of bands (blue) and PCs (red)
- Bottom right panels are kurtosis vs SCR and vs variance
  - Black: CMF, Magenta: OBS k=20, Green: OBS k=2, Blue: SMF


File Edit View Insert Tools Window Help
20 40 60 80 100 120
20 40 60 80 100 120
20 40 60 80 100 120
20 40 60 80 100 120
20 40 60 80 100 120
20 40 60 80 100
× 10
0 -1 -2 -3
-5 0
5 0 -5 -10
-10 0 10
10^10 10^5 10^0
0 200
30 20 10 0
0 200 400
20 10 0
10^-5 10^0 10^5
20 10 0
10^-5 10^0


File Edit View Insert Tools Window Help
20 40 60 80 100 120
20 40 60 80 100 120
× 10^-4
-1 -2 -3 -4
-4 -2 0
10 0 -10
-10 0 10
10^10 10^5 10^0
0 200 400
15 10 5 0
0 200 400
10^-5 10^0 10^5


File Edit View Insert Tools Window Help
20 40 60 80 100 120
20 40 60 80 100 120
20 40 60 80 100 120
20 40 60 80 100 120
20 40 60 80 100 120
20 40 60 80 100 120
× 10
0 -2 -4 -6
-2 0 2
10 5 0 -5
-10 0 10
10^10 10^5 10^0
0 200 400
300 200 100 0
0 200 400
10 5 0
10^-5 10^0 10^5
10 5 0
10^-5 10^0 10^5

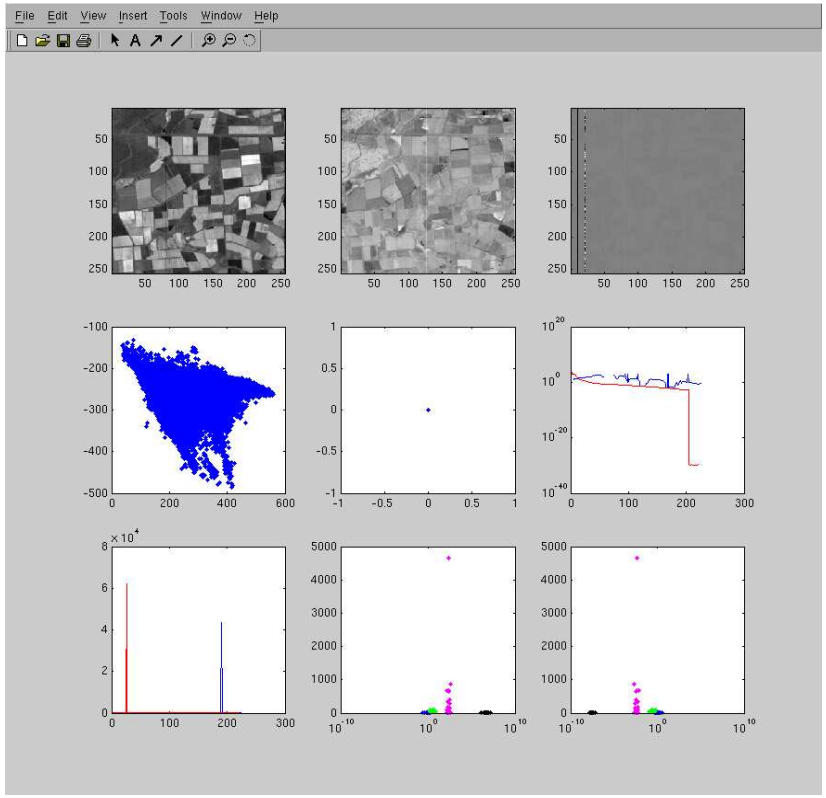

File Edit View Insert Tools Window Help
20 40 60 80 100 120
20 40 60 80 100 120
20 40 60 80 100 120
20 40 60 80 100 120
20 40 60 80 100 120
20 40 60 80 100 120
× 10
0
-2
-4
-5
0
5
0
-5
-10
-10
0
10
10^{10}
10^{5}
10^{0}
0
200
400
30
20
10
0
0
200
400
15
10
5
0
10^{-5}
10^{0}
10^{5}
15
10
5
0
10^{-5}
10^{0}
10^{5}


File Edit View Insert Tools Window Help
20 40 60 80 100 120
20 40 60 80 100 120
20 40 60 80 100 120
20 40 60 80 100 120
20 40 60 80 100 120
20 40 60 80 100 120
× 10
0 -2 -4 -6
-5 0
5 0 -5
-10 0 10
10^10 10^5 10^0
0 200 400
30 20 10 0
0 200 400
10 5 0
10^-5 10^0 10^5
10 5 0
10^-5 10^0 10^5

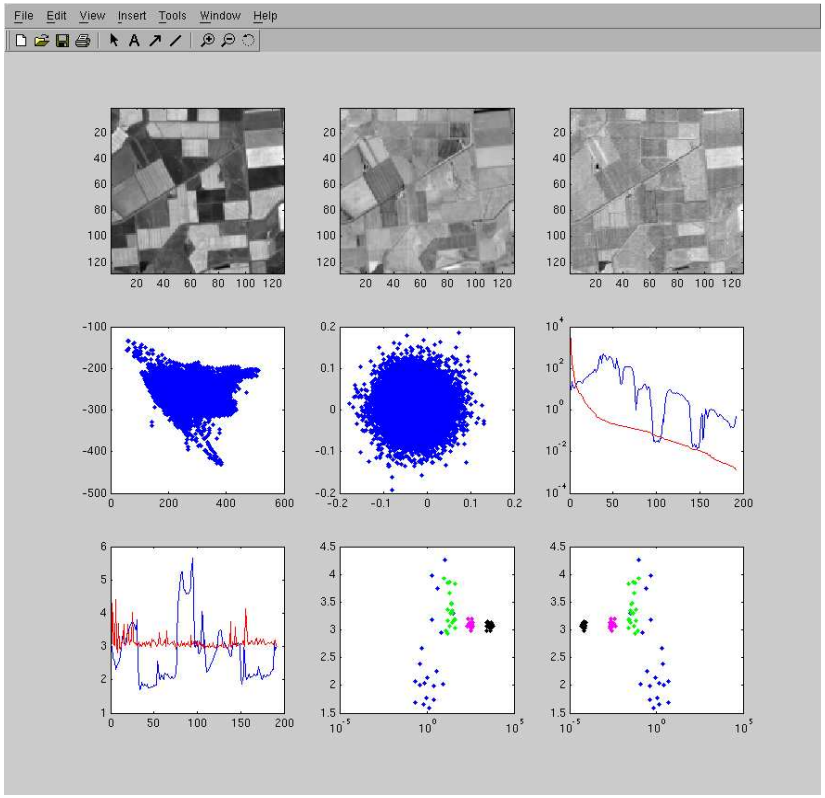
Kurtosis versus Variance (preliminary observations)

- Although all these plots are from the same image – just different tiles – there is a lot of variation from one tile to another.
- Scatterplot of first and second PCs strikingly nongaussian;
  - Generally more "triangular" than "elliptical": indicative of linear mixing.
  - Interestingly, triangles have curved sides, often non-convex.
- Scatterplot of higher PCs generally look a lot more gaussian
- Comparing kurtosis vs band number (blue) to kurtosis vs PC number (red)
  - Neither shows *consistently* larger kurtosis, but...
  - kurtosis for PCs sometimes *much larger* than kurtosis for bands
- Comparing random directions to matched-filter directions
  - Kurtosis of the random b directions (blue dots) is usually larger, sometimes much larger, than the gaussian value of 3.
  - Kurtosis of the matched-filter $q = \hat{K}^{-1}b$ directions (black dots) is almost always near the gaussian value of 3.
- Kurtosis also seems to be near 3 for the OBS with k=20 (magenta dots)

Kurtosis: effect of image artifacts



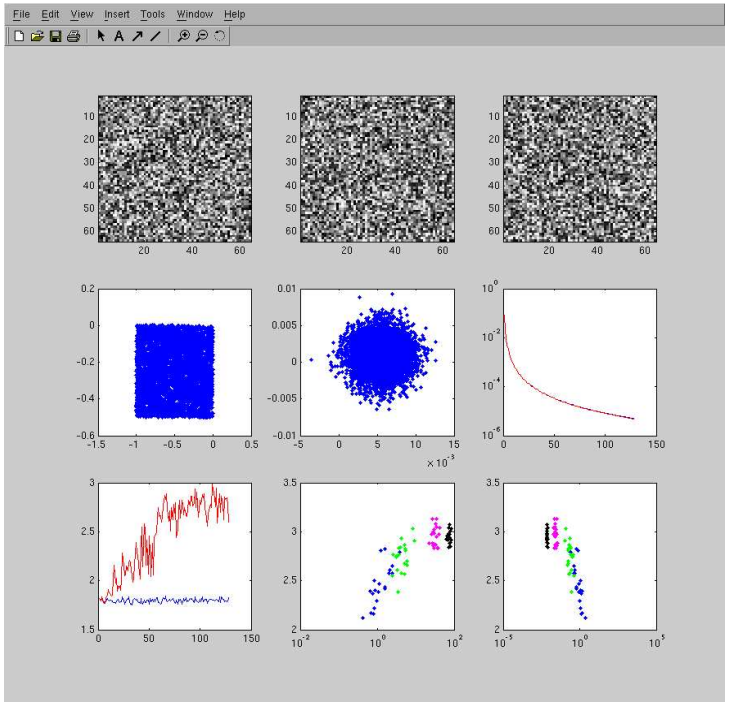
original data, full datacube



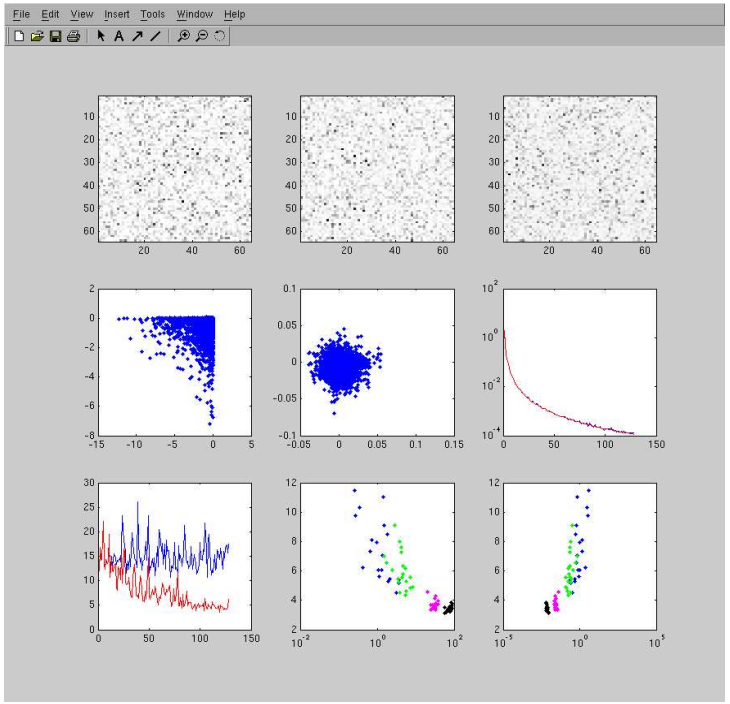
with some artifacts removed

- Original Hyperion imagery included a few bad channels
- Kurtosis effects can be reduced by fixing the artifacts
  - reduced, but not eliminated, as evidenced by simulated data

Synthetic random data



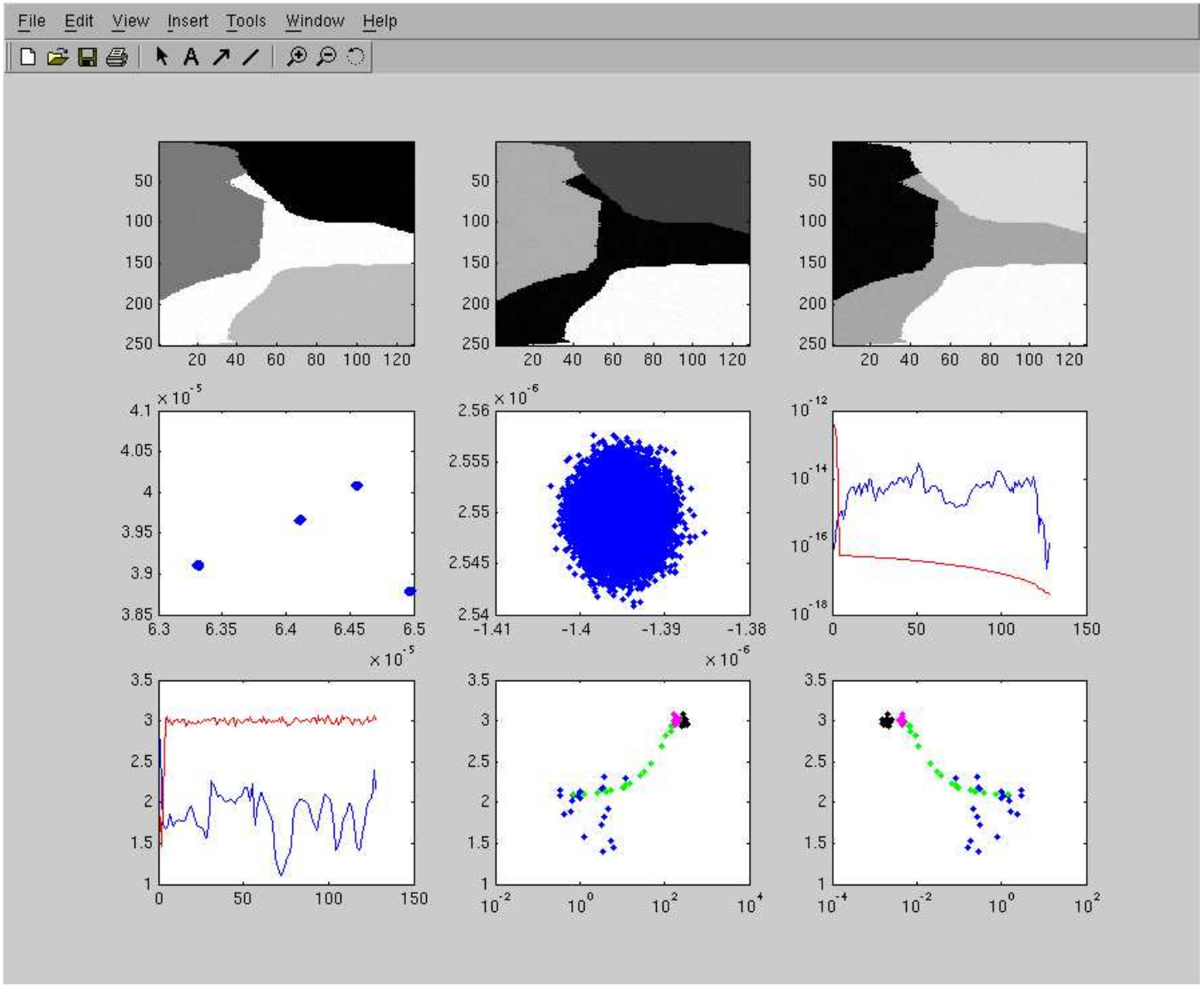
uniform



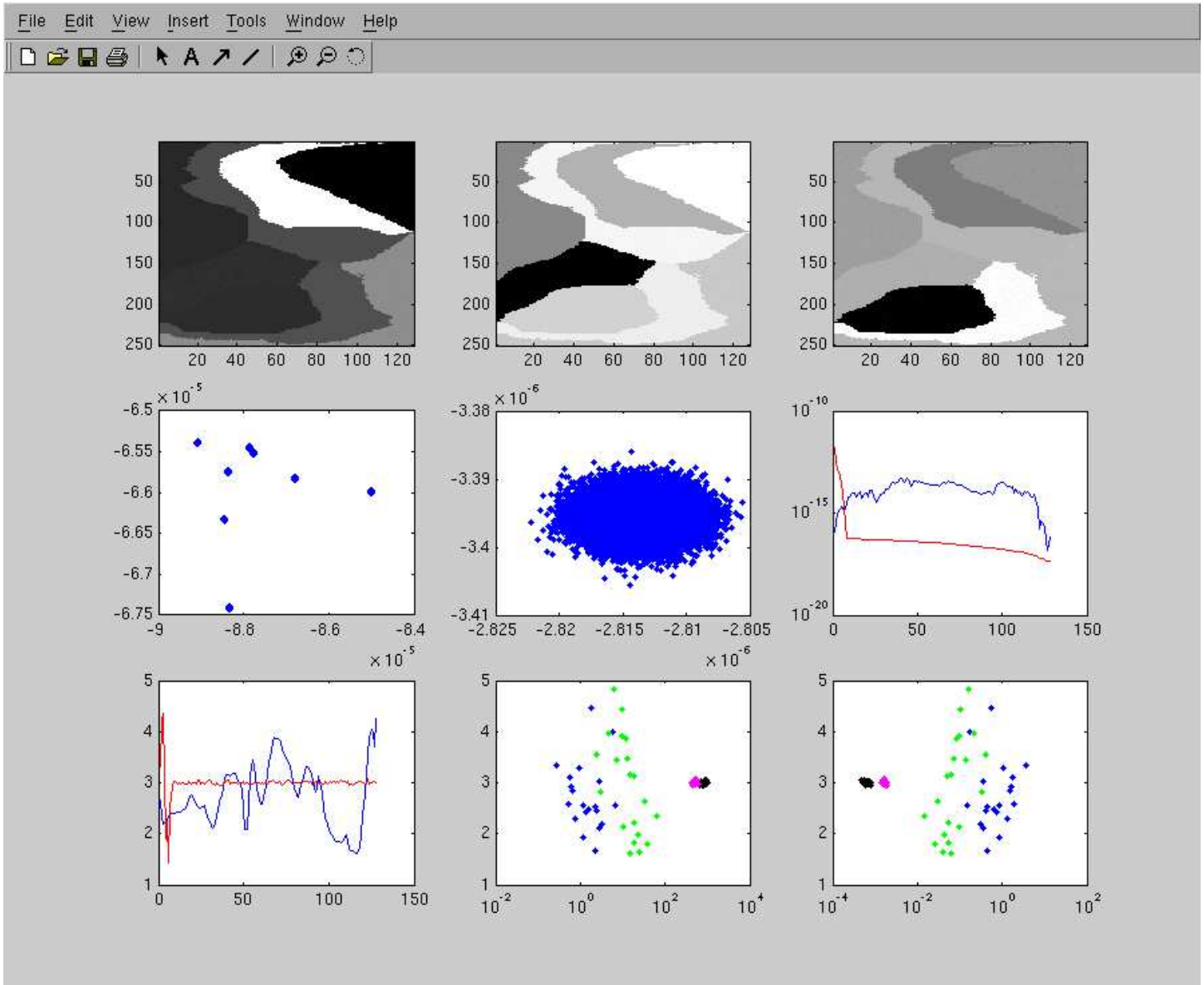
squared gaussian

- Uniform is platykurtic ($\kappa < 3$); squared gaussian is leptokurtic ($\kappa > 3$)
- Yet, for both distributions:
  - Kurtosis for high PCs tends toward gaussian ($\kappa = 3$)
  - Kurtosis for low-variance directions tends toward gaussian
- Kurtosis-variance effect not particular to hyperspectral data

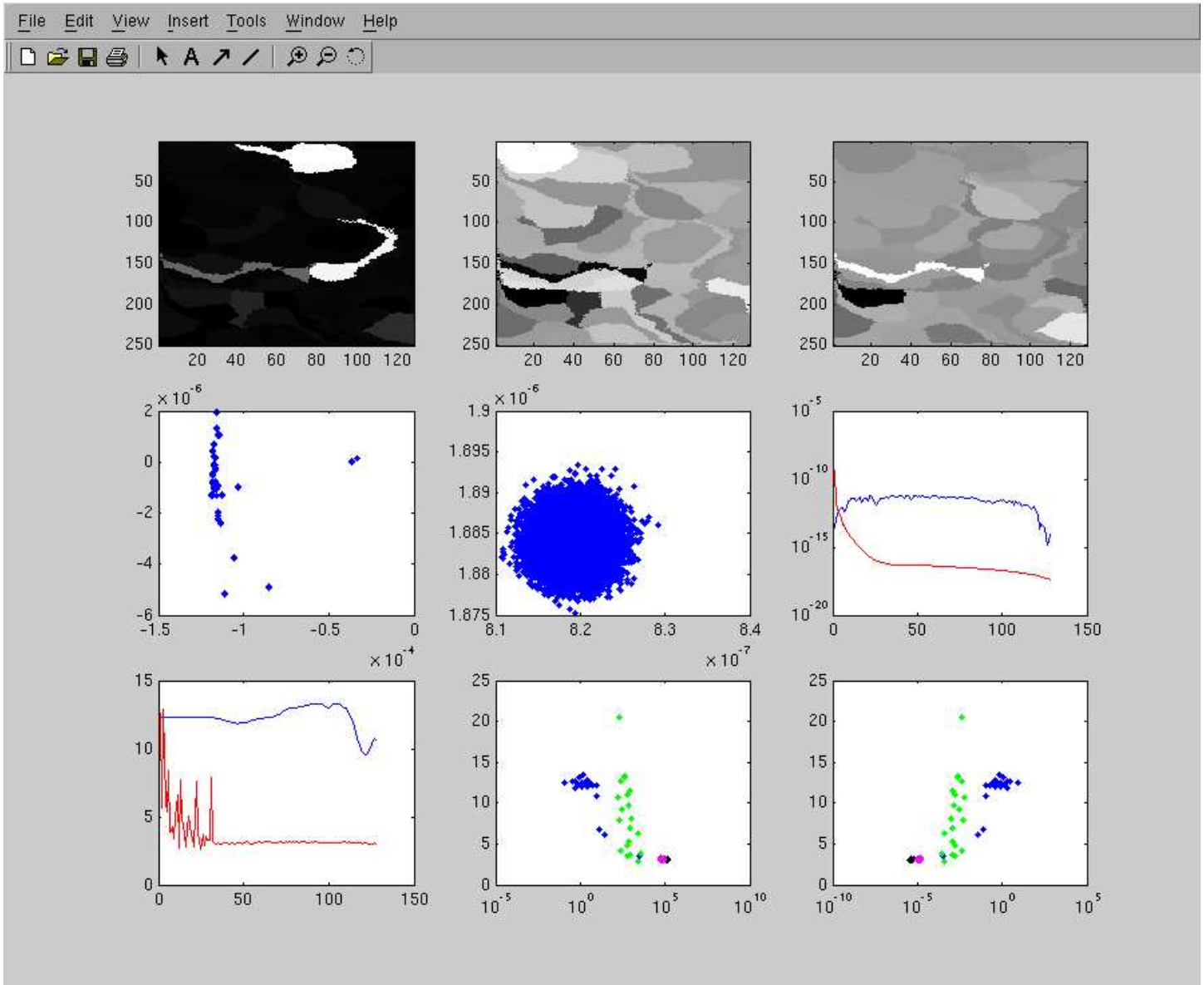
Synthetic data: 4 components



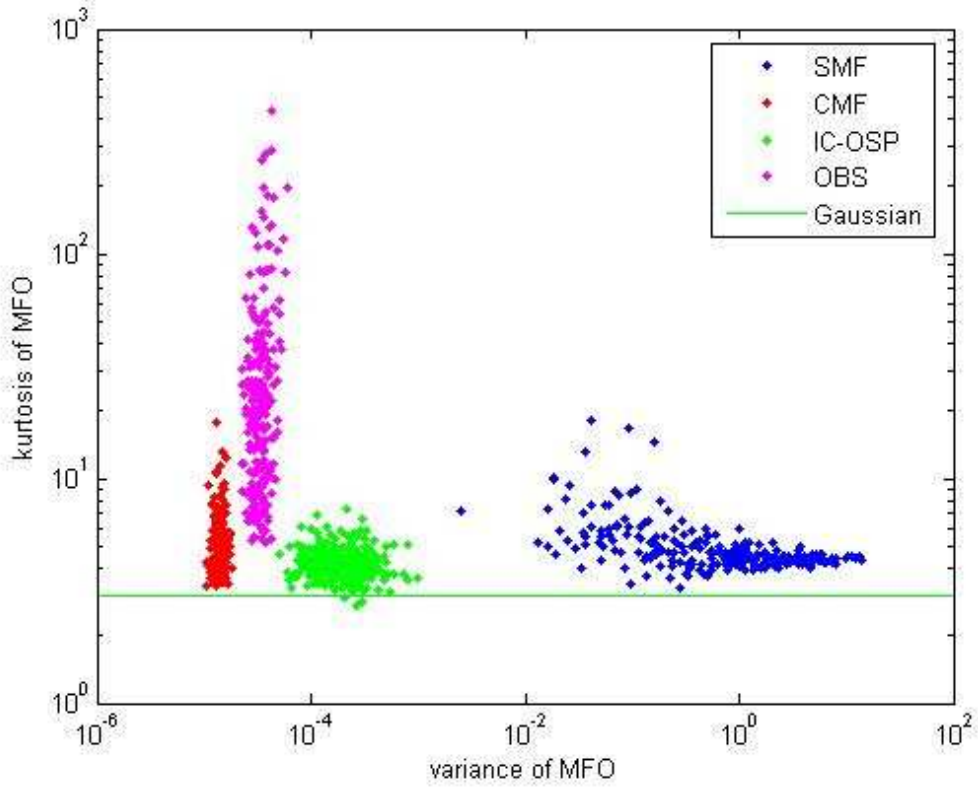
Synthetic data: 8 components



Synthetic data: 40 components



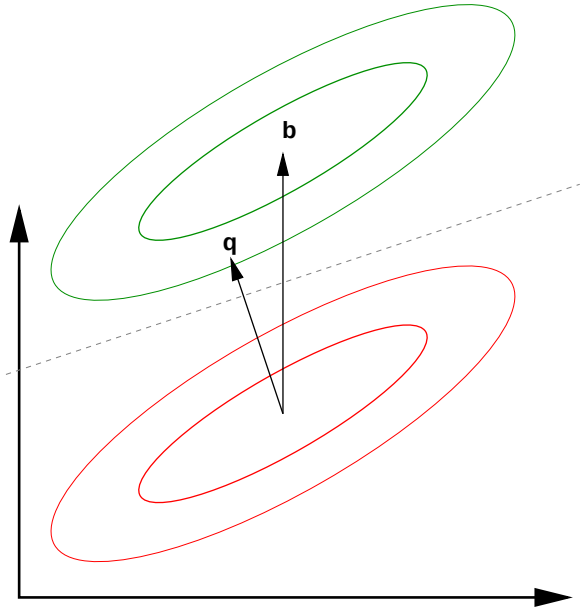
Experiment with Independent Components Analysis

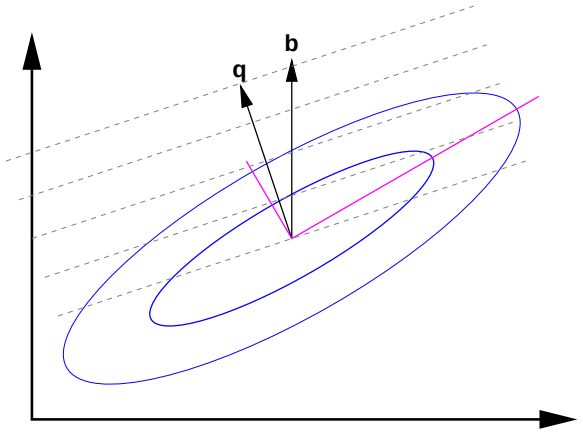


- Project out the first 20 ICA bands instead of first 20 PCA bands
  - higher variance (PCA is more efficient way to reduce variance), but
  - lower kurtosis
- Speculation: might lead to a scheme for a "more gaussian matched filter"

III. Can machine learning play a role?

Binary classification: the Fisher discriminant





- Two classes with a common covariance
  - Means μ_1 and μ_{-1}; Covariance K.
  - Let $b = \mu_1 - \mu_{-1}$
- Between-class scatter matrix $S_B = bb^T$
- Within-class scatter matrix $S_W = K$
- Maximize Rayleigh quotient:

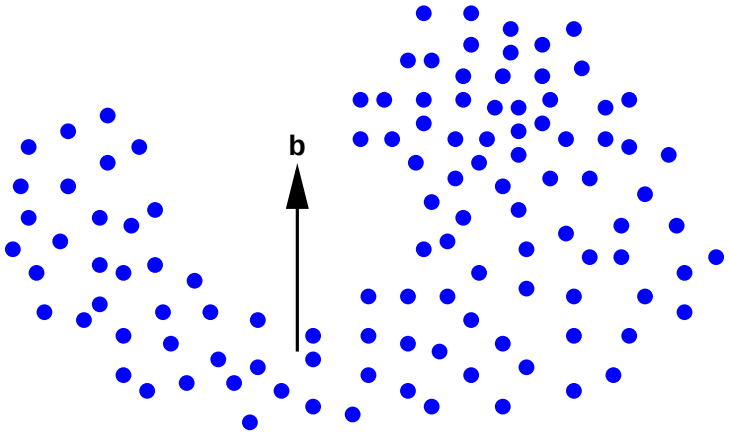
$$J(q) = \frac{q^T S_B q}{q^T S_W q} = \frac{q^T b b^T q}{q^T K q}$$

- Solution given by eigenvectors of $S_W^{-1} S_B$
  - There is only one with nonzero eigenvalue
  - Fisher discriminant: $q = \mathrm{K}^{-1} b$
- Say, this looks a lot like the matched filter!

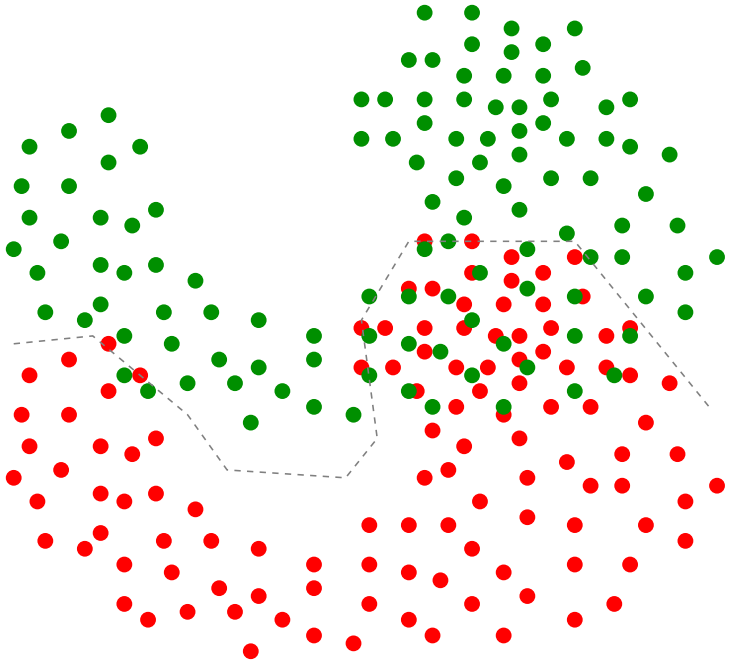
Weak Signal Identification *vs.* Classification

- For gaussian clutter, the two tasks are equivalent
- Main idea: Convert the weak signal problem to the classification problem, and then use something besides the Fisher discriminant
  - Something that does not assume gaussian clutter
  - Something that is not so sensitive to high dimensionality
- We have addressed these points previously, but in piecemeal fashion
  - Funk *et al.*: clutter as a mixture of gaussians
  - Villeneuve *et al.*: regularization of gaussian using principal components with a cutoff/saturated eigenvalue
- What does ML provide?
  - ML never assumes anything is gaussian
  - ML can be well-behaved in high-dimensional situations
  - ML already knows how to do classification

Convert matched filter problem to classification problem



- Image clutter given by blue dots
- Vector b in direction of desired signature



- Convert into classification problem
  - Red or Green?

Hypothesis testing approach, when $P(z)$ is known

- **Simple Hypothesis test**
  - Measured $r = \epsilon b + z$, with $z \sim P(z)$
    - Null hypothesis $H_o : \epsilon = 0$: $r \sim P(r)$
    - Alternative $H_1 : \epsilon = \epsilon_o$: $r \sim P(r - \epsilon_o b)$
  - Two-class classification problem
    - Discriminant function: iff $\mathcal{D}(r) > \gamma$; then take H_1
    - False alarm rate: $\alpha = \int 1_{\{\mathcal{D}(r) > \gamma\}} P(r) dr$
    - Detection rate: $\int 1_{\{\mathcal{D}(r) > \gamma\}} P(r - \epsilon_o b) dr$
    - Bayes optimal solution given by likelihood ratio:

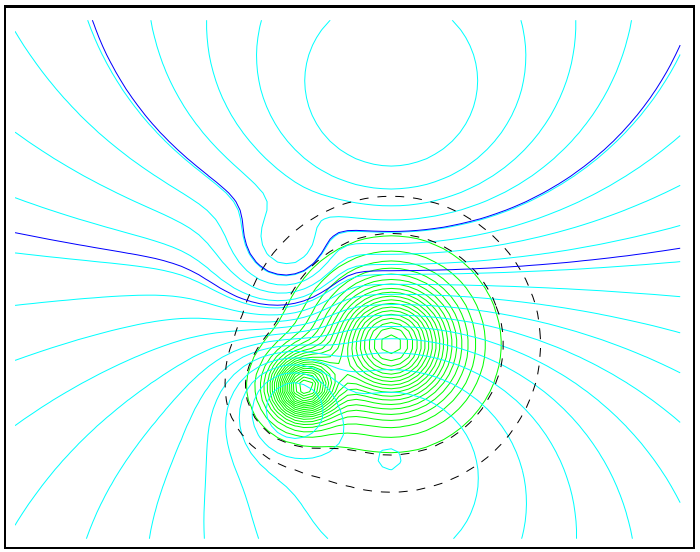
$$\mathcal{D}(r) = \frac{P(r - \epsilon_o b)}{P(r)}$$

    - Threshold γ adjusted to trade false alarms vs detections

- **Composite Hypothesis test:** $H_1 : \epsilon \neq 0$
  - One-sided variant: $H_1 : \epsilon > 0$

Simple Hypothesis test

- Null and alternative hypotheses both precisely known
- Measured $r = \epsilon b + z$, with $z \sim P(z)$
  - Null $H_o : \epsilon = 0$, so $r \sim P(r)$
  - Alternative $H_1 : \epsilon = \epsilon_o$, so $r \sim P(r - \epsilon_o b)$



- Green contours show distribution under the null hypothesis: $P(r)$
- Black dashed contours enclose 95% and 99% of the integrated probability in $P(r)$
- Cyan contours are of the likelihood ratio
- Blue contours correspond to 5% and 1% false alarm rates

- Figure is based on mixture of two EC t-distributions with $\nu = 15$

Generalized Likelihood Ratio Test (GLRT)

- Replace all nuisance parameters with their maximum likelihood estimates

$$\mathcal{D}(\boldsymbol{r}) = \frac{\max_{\theta \in \boldsymbol{\Theta_1}} \mathcal{P}_{\boldsymbol{H_1}}(\theta; \boldsymbol{r})}{\max_{\theta \in \boldsymbol{\Theta_o}} \mathcal{P}_{\boldsymbol{H_o}}(\theta; \boldsymbol{r})}$$

- GLRT introduced for Gaussian distributions by Kelly (1986).
  - $\boldsymbol{H_o}$: Estimate centroid $\hat{\mu}$ and covariance $\hat{\boldsymbol{K}}$ from plume-free data
  - $\boldsymbol{H_1}$: Estimate $\hat{\mu}$, $\hat{\boldsymbol{K}}$ and $\hat{\epsilon}$ from all data

$$\mathcal{D}(\boldsymbol{r}) = \frac{|\boldsymbol{b}^T \hat{\boldsymbol{K}}^{-1}(\boldsymbol{r} - \hat{\mu})|^2}{\boldsymbol{b}^T \hat{\boldsymbol{K}}^{-1} \boldsymbol{b}} \times \frac{\boldsymbol{1}}{\boldsymbol{1} + \frac{\boldsymbol{1}}{\boldsymbol{N}} (\boldsymbol{r} - \hat{\mu})^T \hat{\boldsymbol{K}}^{-1} (\boldsymbol{r} - \hat{\mu})}$$

  - Adaptive matched filter with $\boldsymbol{O(1/N)}$ correction... which, in practice, is usually neglected

- In case $\boldsymbol{P(z)}$ is known, then ϵ is the only nuisance parameter

$$\mathcal{D}(\boldsymbol{r}) = \frac{\max_{\epsilon} \boldsymbol{P}(\boldsymbol{r} - \epsilon \boldsymbol{b})}{\boldsymbol{P}(\boldsymbol{r})}$$

- For Gaussian $\boldsymbol{P(z)}$, GLRT provides *uniformly most powerful* detector

Kelly, E. J. (1986) "An adaptive detection algorithm," *IEEE Trans. Aerospace and Electronic Systems* 22, pp. 115–127.

Bayesian Likelihood Ratio Test ... a.k.a. Bayes Factor

- Requires a prior* on plume strength: $P_\epsilon(\epsilon)$

$$\mathcal{D}(\boldsymbol{r}) = \frac{\mathcal{P}_{H_1}(\boldsymbol{r})}{\mathcal{P}_{H_o}(\boldsymbol{r})} = \frac{\int P_\epsilon(\epsilon)\, P(\boldsymbol{r} - \epsilon \boldsymbol{b})\, d\epsilon}{P(\boldsymbol{r})}$$

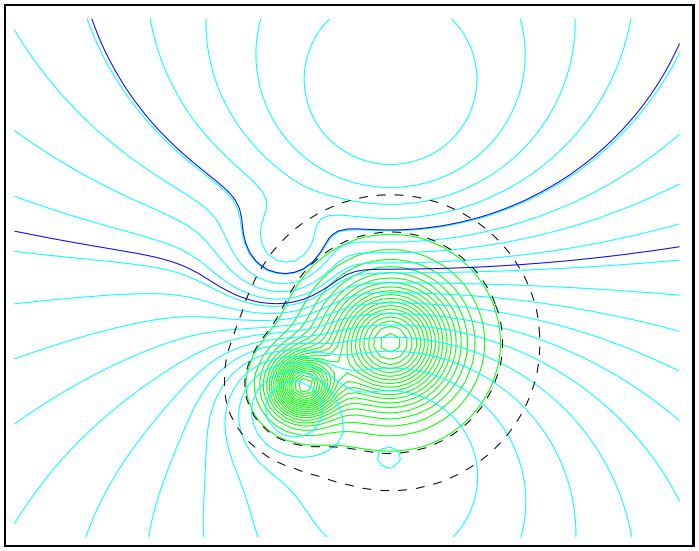
  - Ratio of two *bona fide* probability distribution functions
  - Contours of $\mathcal{D}(\boldsymbol{r})$ align with $\boldsymbol{q}^T\boldsymbol{r}$ for gaussian $P(\boldsymbol{r})$, *regardless of prior*
- Some Special cases
  - Fixed alternative: $P_\epsilon(\epsilon) = \delta(\epsilon - \epsilon_o)$
    - Weak plume: $\epsilon_o \to 0$ (leads to the "derivative method")
    - Strong plume: $\epsilon_o \to \infty$
  - Two sided: $P_\epsilon(\epsilon) = (1/2)\delta(\epsilon - \epsilon_o) + (1/2)\delta(\epsilon + \epsilon_o)$
  - Bounded uniform prior: $P_\epsilon(\epsilon) = (1/2\epsilon_o)\, 1_{\{-\epsilon_o \le \epsilon \le \epsilon_o\}}$
    - One-sided: $P_\epsilon(\epsilon) = (1/\epsilon_o)\, 1_{\{0 \le \epsilon \le \epsilon_o\}}$
  - Flat (improper) prior: $P_\epsilon(\epsilon) = 1$
    - Numerator is just a projection along the b direction
    - One-sided: $P_\epsilon(\epsilon) = 1_{\{\epsilon > 0\}}$

*ack!

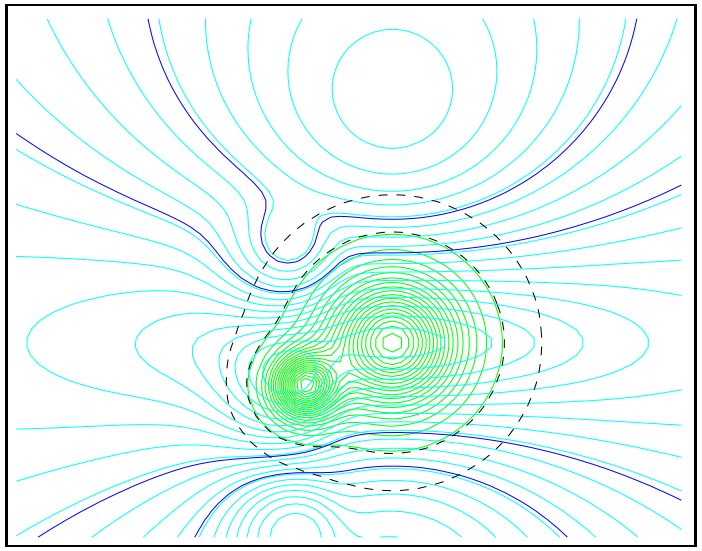
Simple example of a Composite Hypothesis test

- **Alternative hypothesis is two-sided**
  - $H_1 : \epsilon = \epsilon_o$ **or** $\epsilon = -\epsilon_o$
  - **Bayes approach gives equal prior weight to both cases**

$$\mathcal{P}_{H_1}(r) = \frac{1}{2} P(r - \epsilon_o b) + \frac{1}{2} P(r + \epsilon_o b)$$

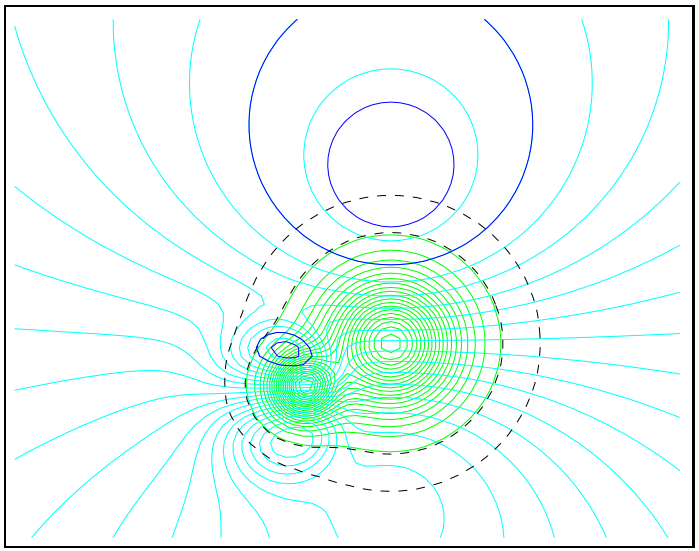


one-sided: $\epsilon = \epsilon_o$

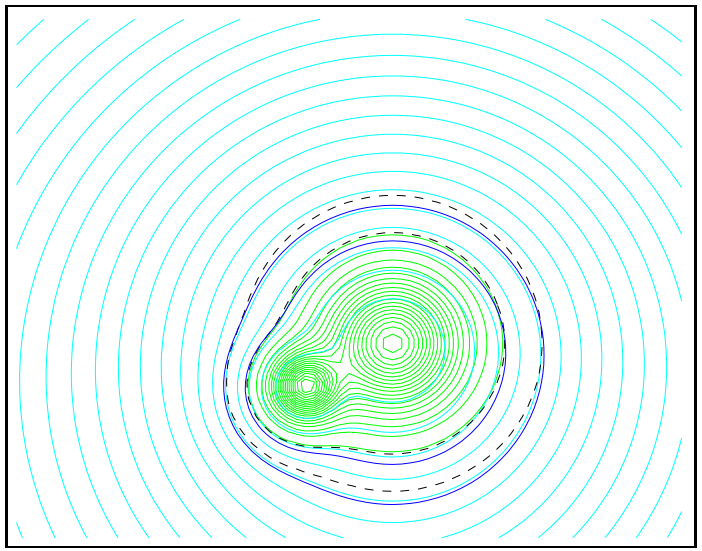


two-sided: $\epsilon = \pm\epsilon_o$

Bayes factor: special limits



$\epsilon \to 0$



$\epsilon \to \infty$

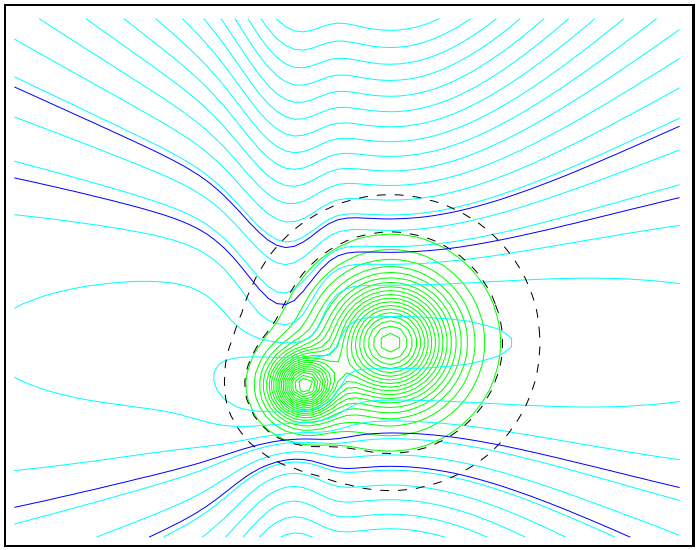
- In weak plume limit, $\epsilon \to 0$ suggests $\boldsymbol{P}(\boldsymbol{r} - \epsilon \boldsymbol{b}) \approx \boldsymbol{P}(\boldsymbol{r}) - \epsilon(\boldsymbol{b} \cdot \nabla)\boldsymbol{P}(\boldsymbol{r})$
  - Likelihood ratio: $\mathcal{D}(\boldsymbol{r}) = \dfrac{\boldsymbol{P}(\boldsymbol{r} - \epsilon \boldsymbol{b})}{\boldsymbol{P}(\boldsymbol{r})} \approx 1 - \epsilon \dfrac{\boldsymbol{b} \cdot \nabla \boldsymbol{P}(\boldsymbol{r})}{\boldsymbol{P}(\boldsymbol{r})}$
  - Contours of constant $\mathcal{D}(\boldsymbol{r})$ will be contours of constant $\mathcal{D}'(\boldsymbol{r})$:

$$\mathcal{D}'(\boldsymbol{r}) \equiv \lim_{\epsilon \to 0} \frac{1 - \mathcal{D}(\boldsymbol{r})}{\epsilon} = \frac{\boldsymbol{b} \cdot \nabla \boldsymbol{P}(\boldsymbol{r})}{\boldsymbol{P}(\boldsymbol{r})} = \boldsymbol{b} \cdot \nabla \log \boldsymbol{P}(\boldsymbol{r})$$

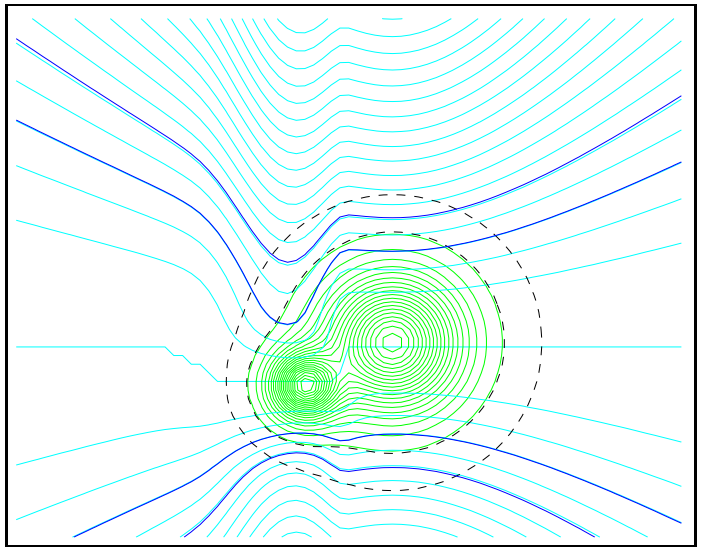
- Strong plume limit, $\epsilon \to \infty$
  - approaches anomaly detection for fat-tailed distributions

Bayes factor vs GLRT

- **Bayes factor with uniform prior:** $\mathcal{D}(\boldsymbol{r}) = \dfrac{\int \boldsymbol{P}(\boldsymbol{r} - \epsilon \boldsymbol{b})\, \mathrm{d}\epsilon}{\boldsymbol{P}(\boldsymbol{r})}$
- **Generalized likelihood ratio:** $\mathcal{D}(\boldsymbol{r}) = \dfrac{\max_\epsilon \boldsymbol{P}(\boldsymbol{r} - \epsilon \boldsymbol{b})}{\boldsymbol{P}(\boldsymbol{r})}$



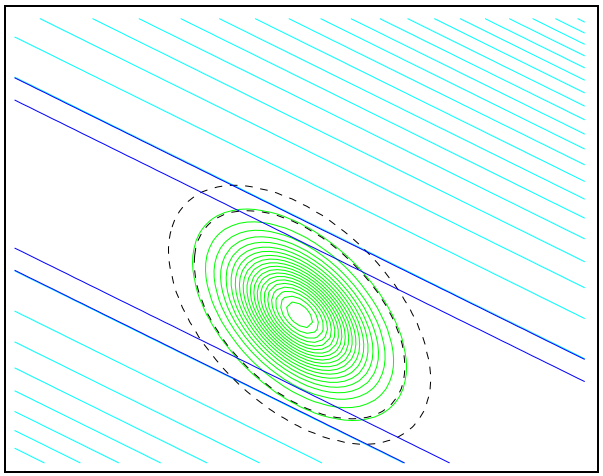
Bayes factor



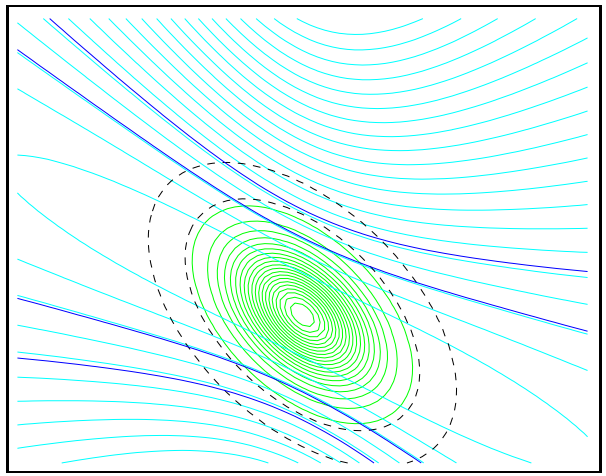
GLRT

- **Methods give different, though generally "similar" contours**
- **Bayes factor seems to give "smoother" contours**

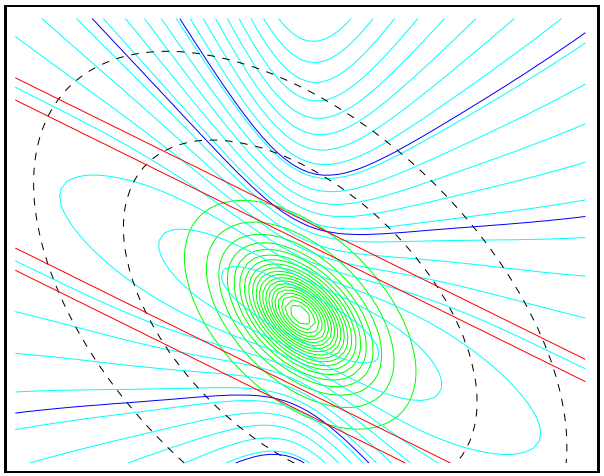
Detection contours for single EC distribution



$\nu \to \infty$



$\nu = 15$



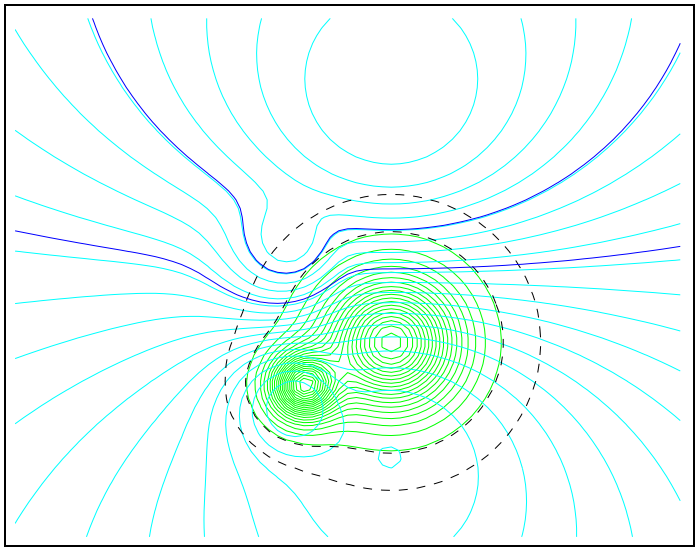
$\nu = 3$

- Multivariate t-distribution
  - with varying ν, but same μ and K for all distributions
- Contours for $\nu \to \infty$ (gaussian) limit are straight lines
- Contours for fat-tailed distributions "bend away"
- Using linear matched filter on EC data leads to:
  - more false alarms (this can be calibrated out)
  - less efficient detection

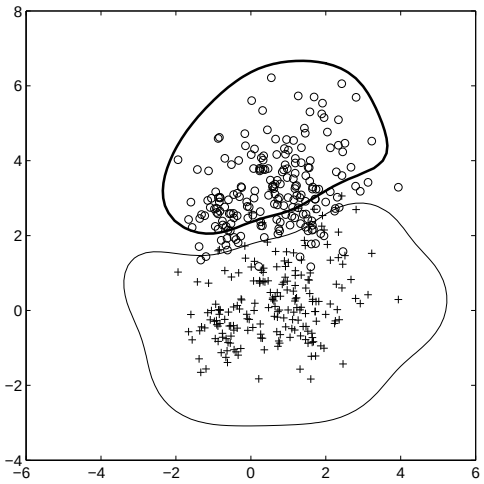
Plume finding and two-class classification when ϵ_o known

- **Plume finding problem: simple hypothesis test**
  - **Measured $r = \epsilon b + z$, with $z \sim P(z)$**
    - **Null hypothesis $H_o : \epsilon = 0$, $r \sim P(r)$**
    - **Alternative $H_1 : \epsilon = \epsilon_o$, $r \sim P(r - \epsilon_o b)$**
  - **Bayes optimal solution given by likelihood ratio: $\mathcal{D}(r) = \dfrac{P(r - \epsilon_o b)}{P(r)}$**
- **Two-class classification problem: simple hypothesis test**
  - **Measure r, choose between two hypotheses**
    - **$H_1 : r \sim P_1(r)$**
    - **$H_2 : r \sim P_2(r)$**
  - **Bayes optimal solution given by likelihood ratio: $\mathcal{D}(r) = \dfrac{P_1(r)}{P_2(r)}$**
- **In the machine learning paradigm,**
  - **Underlying distributions P_1, P_2 are unknown**
  - **Independently sampled data is available for both distributions**
  - **The ML solution avoids estimating $P_{1,2}(r)$ as an intermediate step**

ML solution when ϵ_o is known, but $P(z)$ is unknown



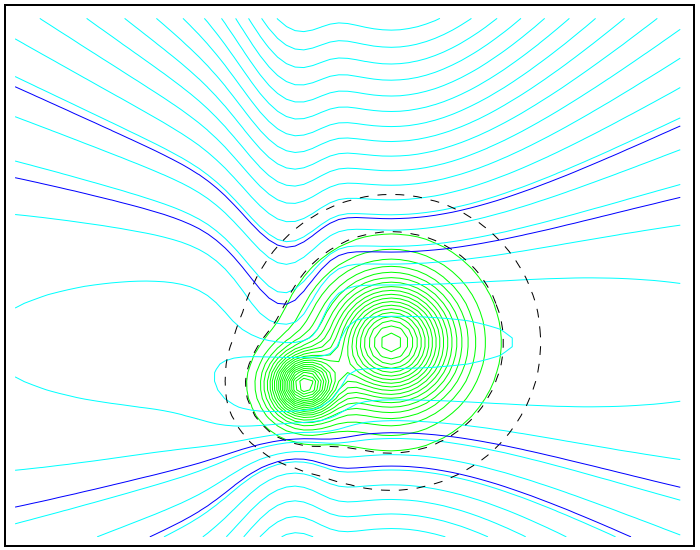
distribution $P(z)$ known



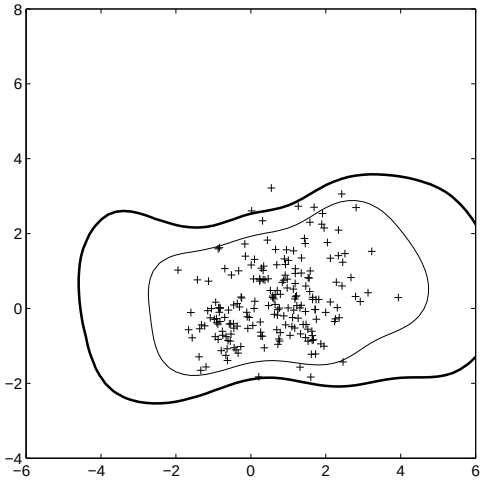
samples $\{z_1, \ldots, z_N\}$ available

- Data for null class: $\{z_1, \ldots, z_N\}$
- Generate artificial data for plume class: $\{z_1 + \epsilon_o b, \ldots, z_N + \epsilon_o b\}$
  - Note that this data is *not* sampled independently of the null class
  - Need new theorems?!
- Use your favorite binary classifier (SVM) to estimate $D(r)$
  - Want Neyman-Pearson boundaries (fixed false alarm rates)
  - If you use Fisher discriminant, you'll get Adaptive Matched Filter

ML solution when ϵ_o and $P(z)$ are both unknown



distribution $P(z)$ known



data available

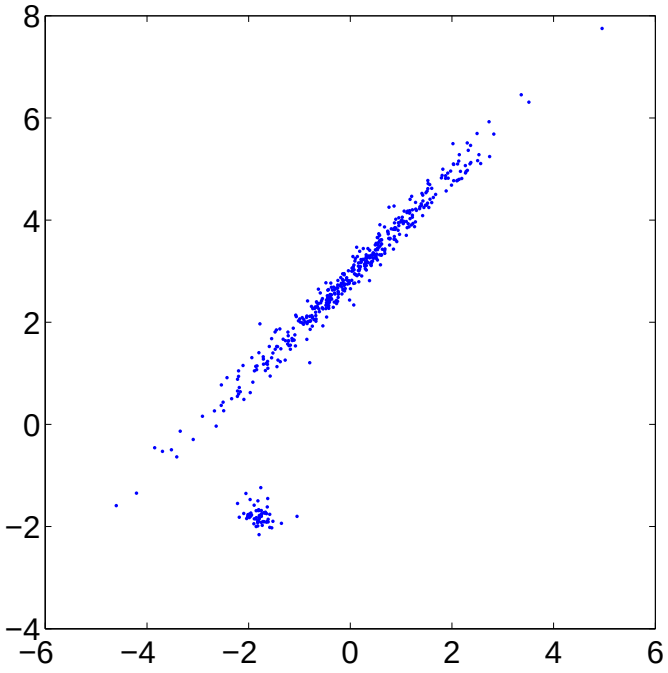
- In Bayes factor formulation, we have a ratio of true distributions: collect samples from both distributions; use binary classification to discriminate.
  - Distribution for null class is unknown: $P(z)$
  - Samples from null class are available data: $\{z_1, \ldots, z_N\}$
  - Distribution for plume class is also unknown, but related to $P(z)$:

$$\int P_\epsilon(\epsilon)\, P(z - \epsilon b)\, d\epsilon$$

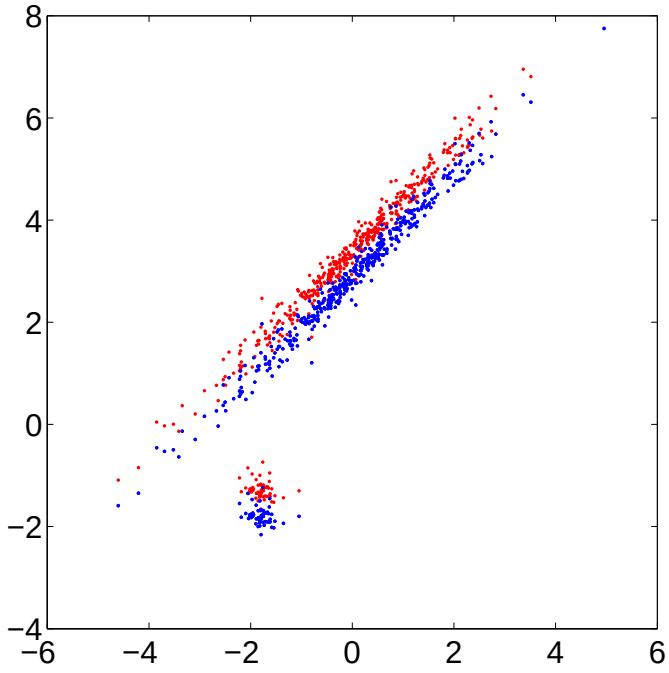
  - Resample data for plume class: $\{z_1 + \epsilon_1, \ldots, z_N + \epsilon_N\}$ with $\epsilon_i \sim P_\epsilon(\epsilon)$

Toy example: non-gaussian clutter

clutter (z)



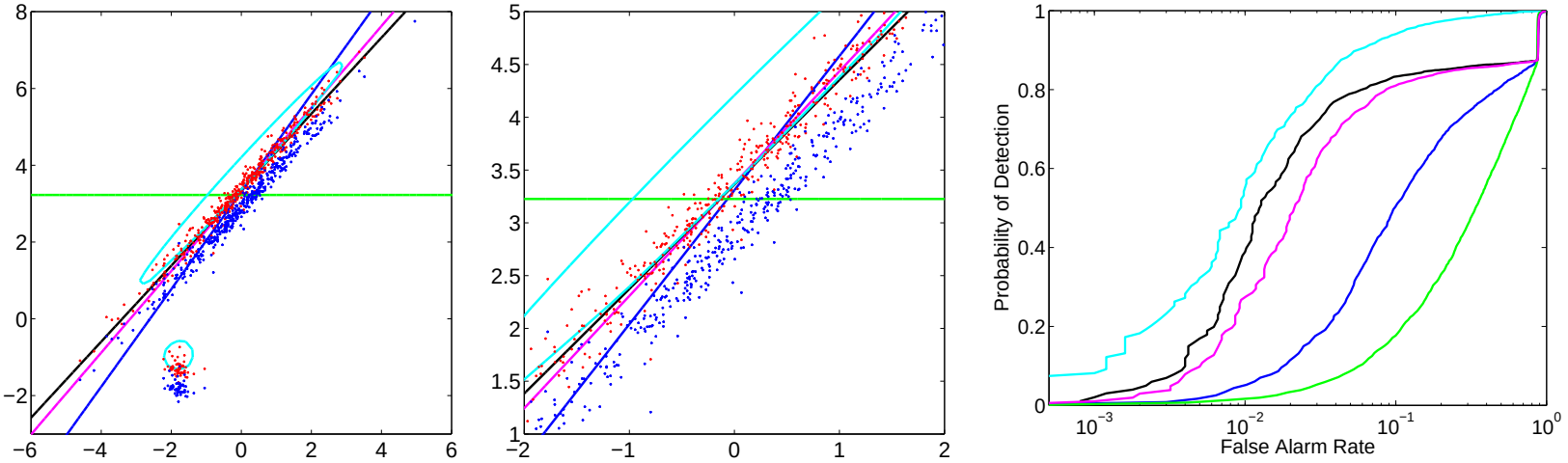
clutter+plume (z, $z + \epsilon b$)



- Two components
  - Main component is thin near-gaussian distribution
    - thin: variance ratio ≈ 200
    - near-gaussian: EC with $\nu = 4$
  - Small fraction of points are in secondary lobe
- Binary classification: separate "red" from "blue"

Toy example: detection contours

detection contours ($P_d = 0.5$ contours (inset) ROC curves (out-of-sample)

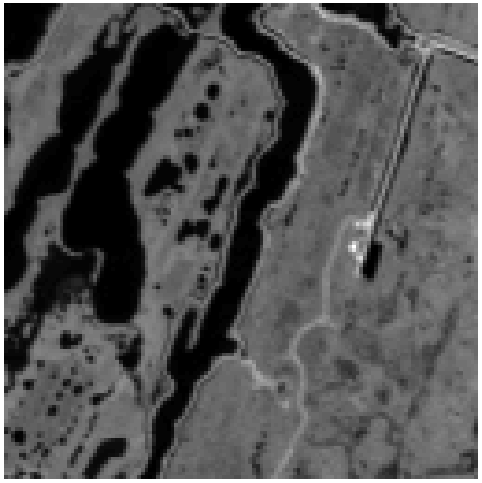


- Cyan: likelihood ratio (best nonlinear detector)
- Green: simple matched filter, $q = b$ (fails utterly for thin distributions)
- Blue: adaptive matched filter, $q = K^{-1}b$ (is sensitive to the lobe)
- Magenta: linear SVM (is somewhat more robust to the lobe)
- Black: linear detector (computed by ignoring the lobe)

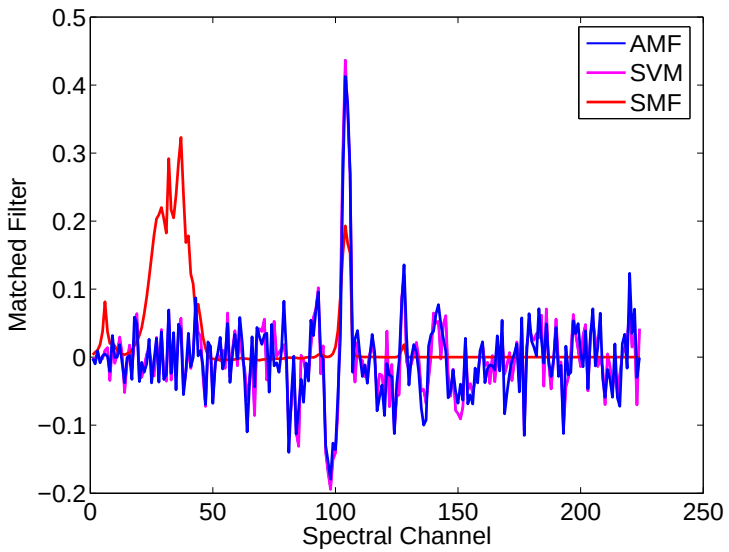
■ Lesson: Linear detection may be adequate

. . . if you can find the right linear detector

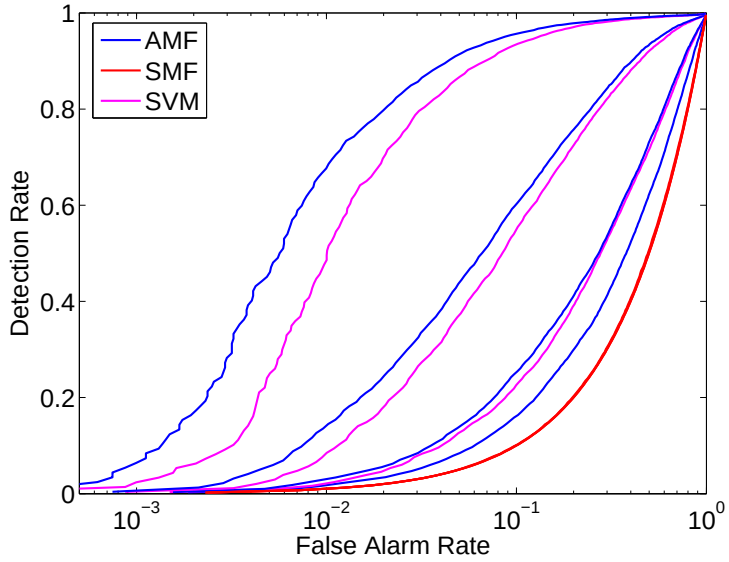
Real data (fake plume)



- Is it possible to find a linear filter that beats $q = K^{-1}b$ (AMF) . . . on real hyperspectral data.
- Test: out of sample, with different plume strengths



matched filter



ROC curves

Are there any questions?

- How to choose the prior* $P_\epsilon(\epsilon)$?
  - And how sensitive is the result to that choice?
- What assurances that the resampling doesn't introduce biases?
- Can the resampling be *built-in* to algorithm?
  - *e.g.*, avoid having the resampled points be support vectors?
- Efficient algorithm when two classes have a high degree of overlap?
- How to deal with (or maybe how to exploit) the huge dynamic range in eigenvalues of the covariance K
- Can a purely data-driven model succeed in such a high-dimensional space?
  - Is this a job for semi-parametric modeling?
  - *e.g.*, mixtures of EC distributions?
- Can ML provide distribution-independent/non-Bayesian algorithms
  - *e.g.*, based on GLRT
- Can new algorithms beat the old matched filter?

*ack!