Technion - IIT
Dept. of Electrical Engineering

Signal and Image Processing lab

Anomaly Preserving Redundancy Reduction in High-Dimensional Signals

Oleg Kuybeda

PhD Research under the supervision of:
Prof. David Malah
Dr. Meir Barzohar

Outline

- Signal subspace estimation of noisy linear mixtures
- Drawbacks of using ℓ_2 -norm for anomaly representation
- Proposal of $\ell_{2,\infty}$ -norm as a remedy
- Greedy subspace estimation algorithm
- Anomaly detection algorithm
- Optimal subspace estimation algorithm
- Multispectral filter design for anomaly detection

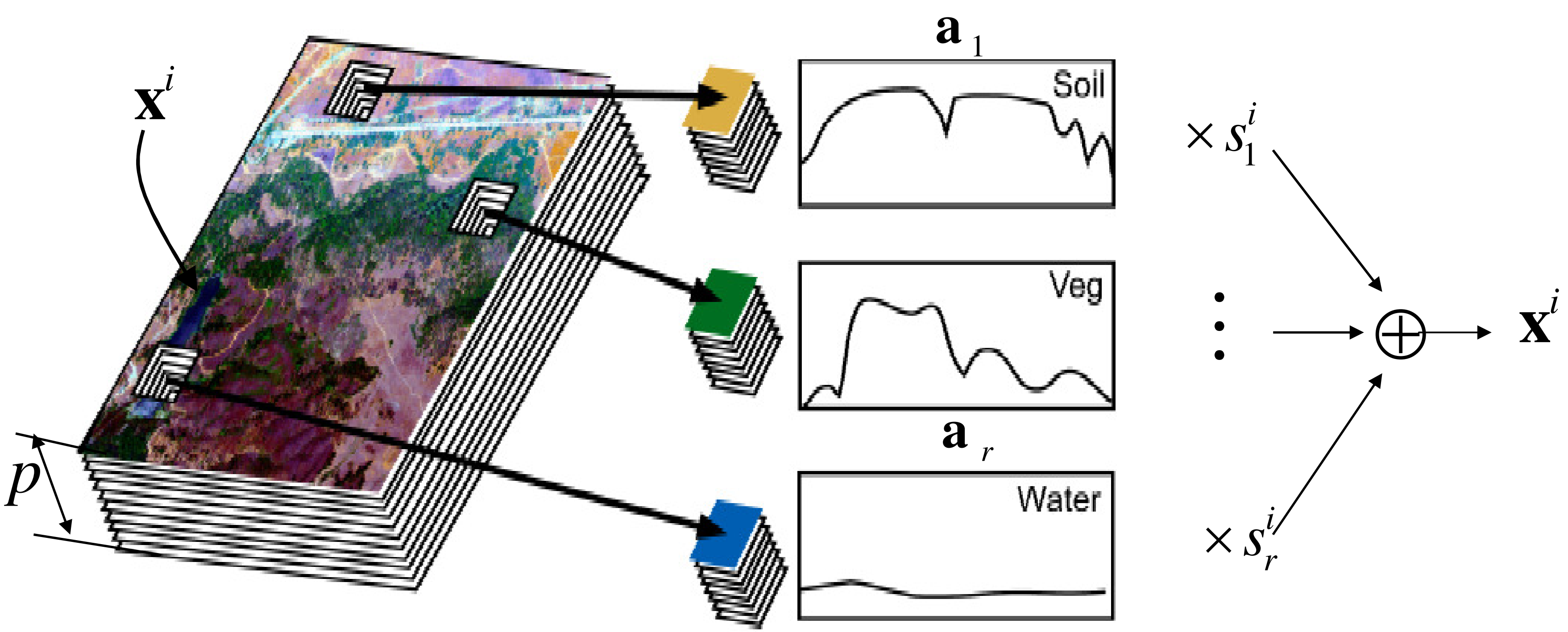
Linear Mixture Model

$$\mathbf{X} = \mathbf{AS} + \mathbf{Z}$$

- $\mathbf{X}$ - pxN matrix of observations
- $\mathbf{A}$ - pxr matrix of endmembers
- $\mathbf{S}$ - rxN matrix of endmember abundances
- $\mathbf{Z}$ - pxN data acquisition/model error

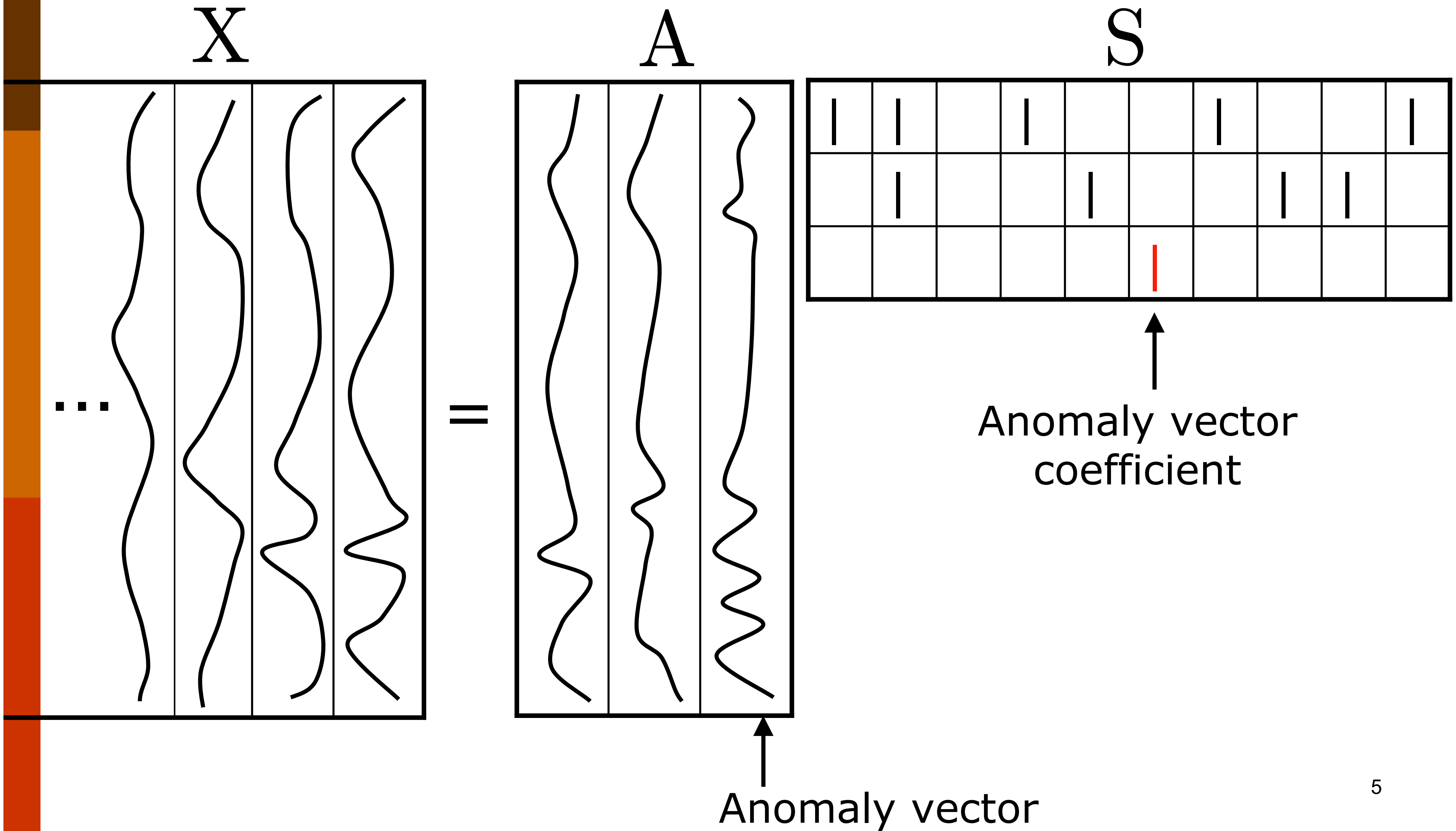
The signal rank is at most r < p !

Linear Mixture Model Example
Hyperspectral images



$$\underbrace{\mathbf{x}^i}_{(p\times 1)} = \underbrace{\mathbf{A}}_{[\mathbf{a}_1 \ \dots \ \mathbf{a}_r]} \cdot \underbrace{\mathbf{s}^i}_{(r\times 1)}$$

Background versus Anomalies



Subspace estimation

ℓ_2 -norm based subspace estimation

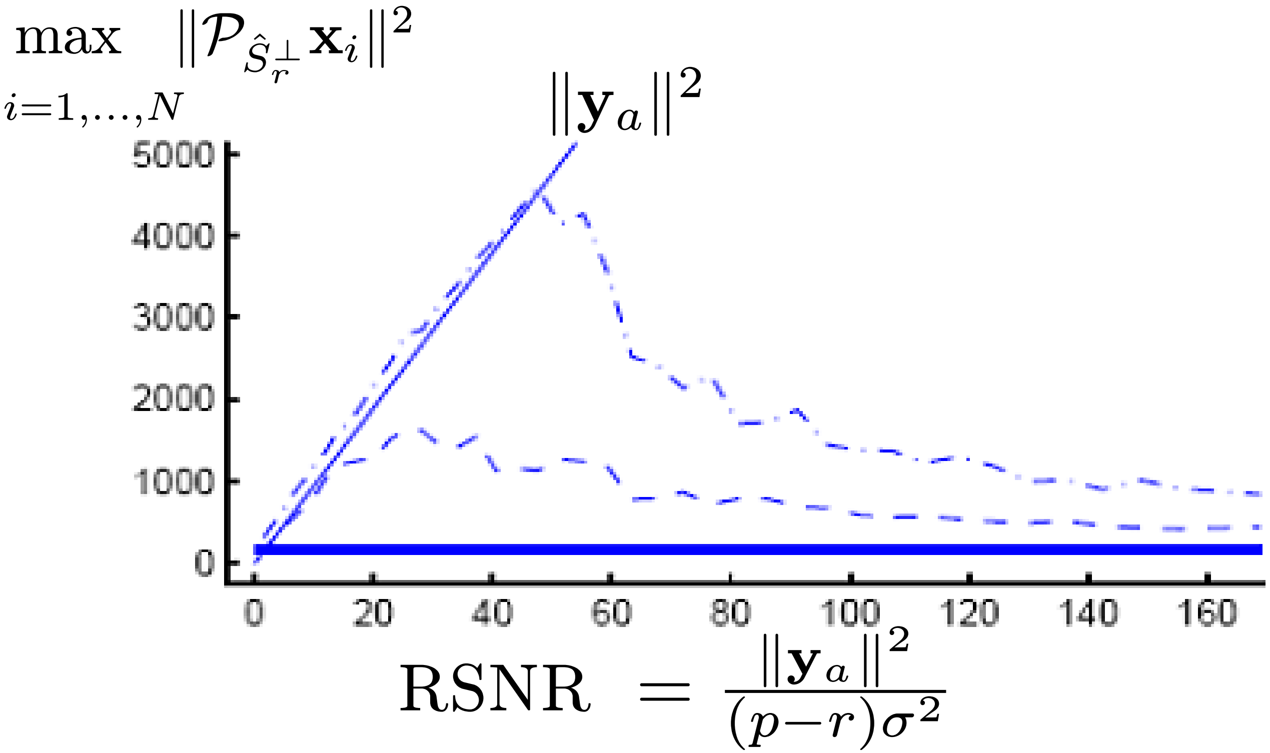
$$\hat{\mathcal{S}}_r = \underset{\mathcal{L}}{\operatorname{argmin}} \; \|\mathcal{P}_{\mathcal{L}^\perp} \mathbf{X}\|_{Fb}^2 \quad \text{s.t.} \quad \operatorname{rank} \mathcal{L} = r,$$

- The signal subspace $\mathcal{L}$ can be estimated via SVD
- Anomaly vector contributions to the ℓ_2 -norm are weaker than contribution of noise
- The resultant subspace is skewed in a way that misses the anomaly vectors with probability close to 1

An example of anomaly misrepresentation as a result of ℓ_2-norm minimization

- Maximum-norm error (dashed and dot-dashed) vs. norm of an anomaly vector (solid)

$$N = 10^5, p = 100, r = r_b + r_a = 5 + 1$$



Subspace estimation
$\ell_{2,\infty}$-norm based subspace estimation

$$\begin{aligned}\hat{\mathcal{S}}_r &= \underset{\mathcal{L}}{\operatorname{argmin}} \max_{i=1,\ldots,N} \|\mathcal{P}_{\mathcal{L}^\perp}\mathbf{x}_i\|^2 = \underset{\mathcal{L}}{\operatorname{argmin}} \|\mathcal{P}_{\mathcal{L}^\perp}\mathbf{X}\|_{2,\infty}^2 \\ & \text{s.t.} \quad \operatorname{rank} \mathcal{L} = r,\end{aligned}$$

- is equivalent to

$$\begin{aligned}\hat{\mathcal{S}}_r &= \underset{\mathcal{L},\gamma}{\operatorname{argmin}}\ \gamma \\ & \text{s.t. } \|\mathcal{P}_{\mathcal{L}^\perp}\mathbf{x}_j\|_2^2 \leq \gamma \qquad \forall j = 1, \ldots, N, \\ & \operatorname{rank} \mathcal{L} = r,\end{aligned}$$

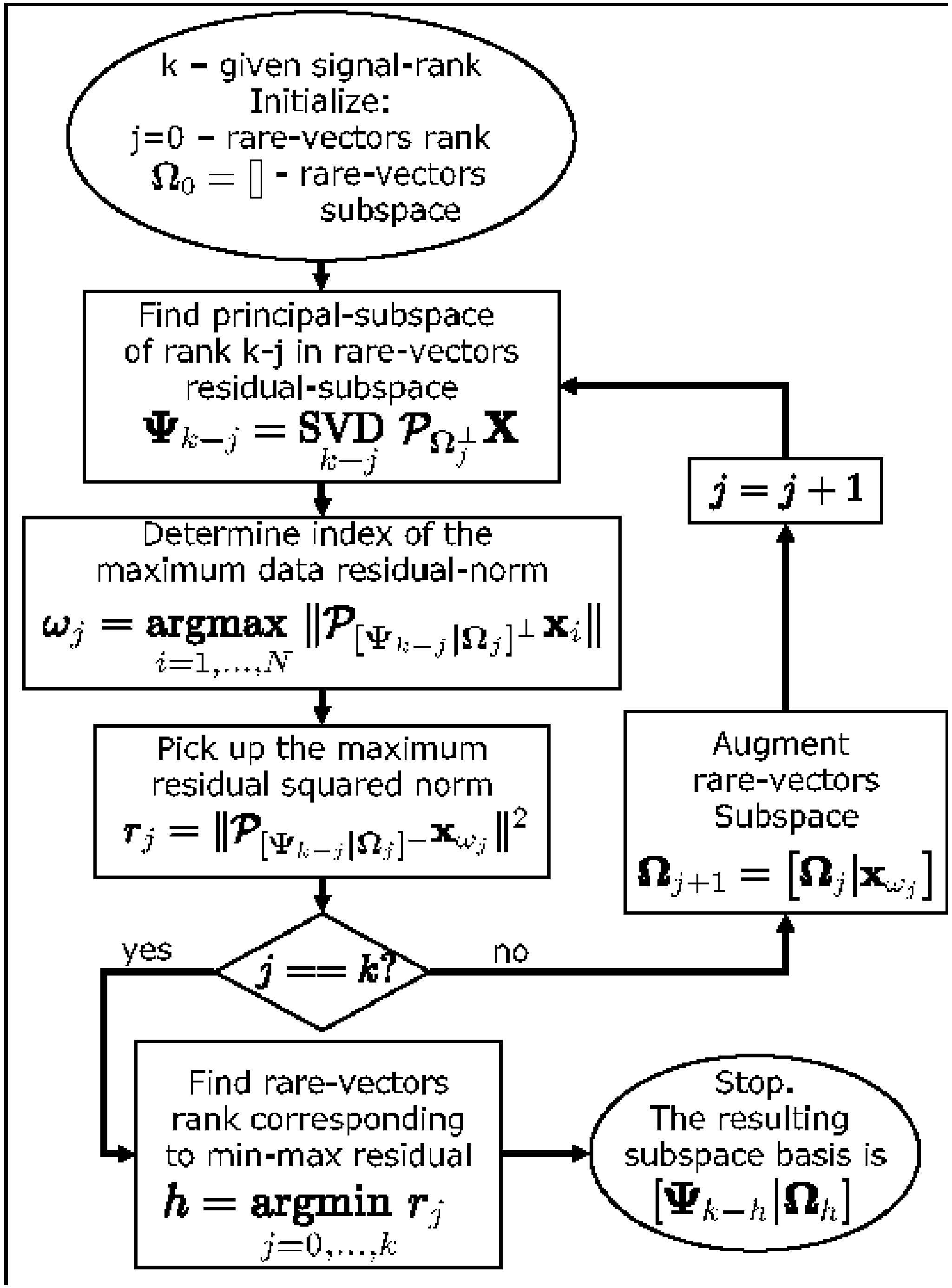
- very hard to optimize due to a large number of constraints and a non-convex constraint

Greedy MX-SVD

- Look for a basis of the form:

$$[\mathbf{\Psi}_{k-h}|\mathbf{\Omega}_h]$$

- $\mathbf{\Omega}_h$ represents anomaly vectors
- $\mathbf{\Psi}_{k-h}$ represents the background



MX-SVD vs. SVD
Maximum residual norm distribution

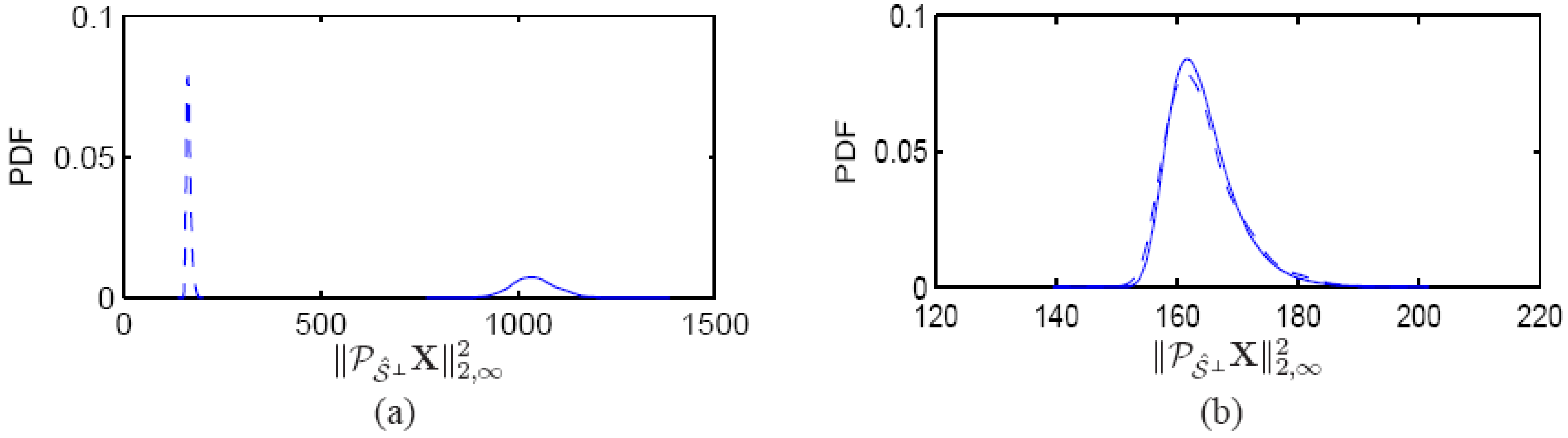


Fig. 4. **The pdfs of $\|\mathcal{P}_{\hat{\mathcal{S}}_k^\perp}\mathbf{X}\|_{2,\infty}^2$, obtained via a Monte-Carlo simulation.** (a) The empirical pdfs of $\|\mathcal{P}_{\hat{\mathcal{S}}_k^\perp}\mathbf{X}\|_{2,\infty}^2$ obtained by MX-SVD (dashed line) and SVD (solid line) for RSNR $= 10$, $\sigma = 1, p = 10^2, N = 10^5$, $k = r_{abund} + r_{rare} = 5 + 3 = 8$ (b) The empirical pdf of $\|\mathcal{P}_{\hat{\mathcal{S}}_k^\perp}\mathbf{X}\|_{2,\infty}^2$ by MX-SVD (dashed-line) versus the exact pdf of $\|\mathcal{P}_{\hat{\mathcal{S}}^\perp}\mathbf{Z}\|_{2,\infty}^2$ (solid line).

Anomaly
misrepresentation
via ℓ_2

Rank estimation via MOCA

Maximum-residual norm = maximum of 2 maxima

Data residual-norms $\{r_i\}$

The group of noise-related residual norms $\{r_{\psi_j}\}$

The group of signal-related residual norms $\{r_{\omega_j}\}$

Max noise-related residual norm $\nu_k = \max\limits_{\psi_j} \{r_{\psi_j}\}$

Max signal-related residual norm $\xi_k = \max\limits_{\omega_j} \{r_{\omega_j}\}$

Max residual norm $y_k = \max\{r_i\} = \max(\nu_k, \xi_k)$

$H_0 : y_k = \nu_k$

$H_1 : y_k = \xi_k$

Rank estimation via MOCA

$P(H_0|y_k)$ and $P(H_1|y_k)$ are functions of $p_{\nu_k}(\cdot)$ and $p_{\xi_k}(\cdot)$

$$p_{\nu_k}(\cdot)$$

- Distribution of the maximum noise-norm:

$$P(\nu_k \le x) = \mathcal{G}\left(a_N\left(x - b_N\right)\right)$$

- Gumbel distribution:

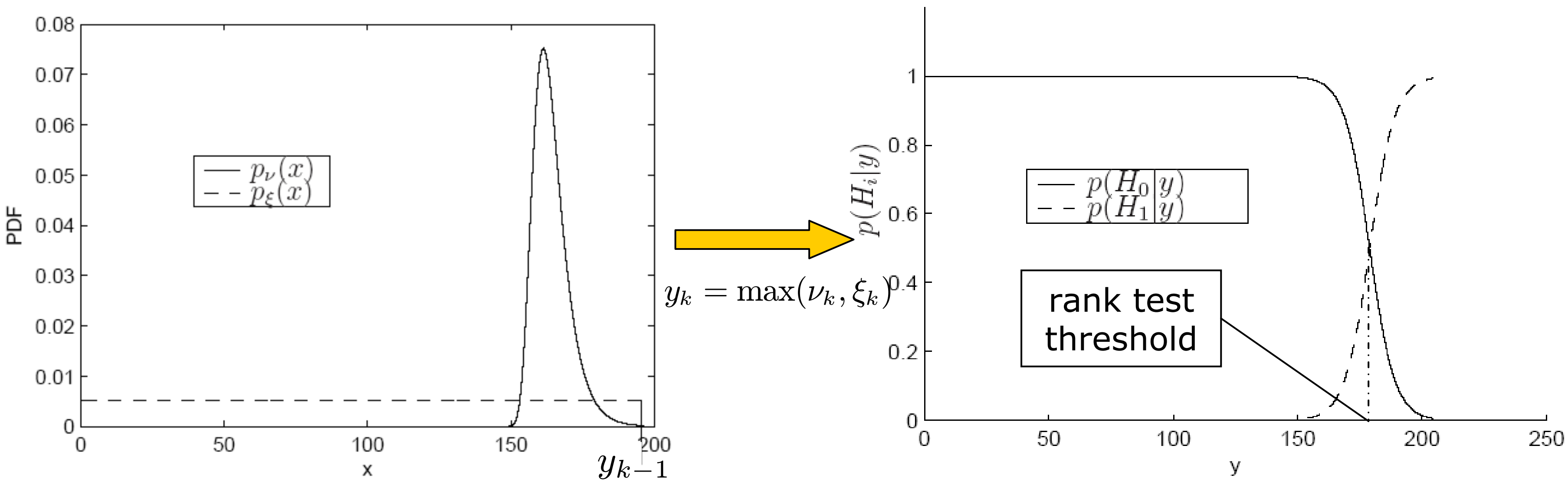
$$\mathcal{G}(x) = e^{-e^{-x}}$$

$$p_{\xi_k}(\cdot)$$

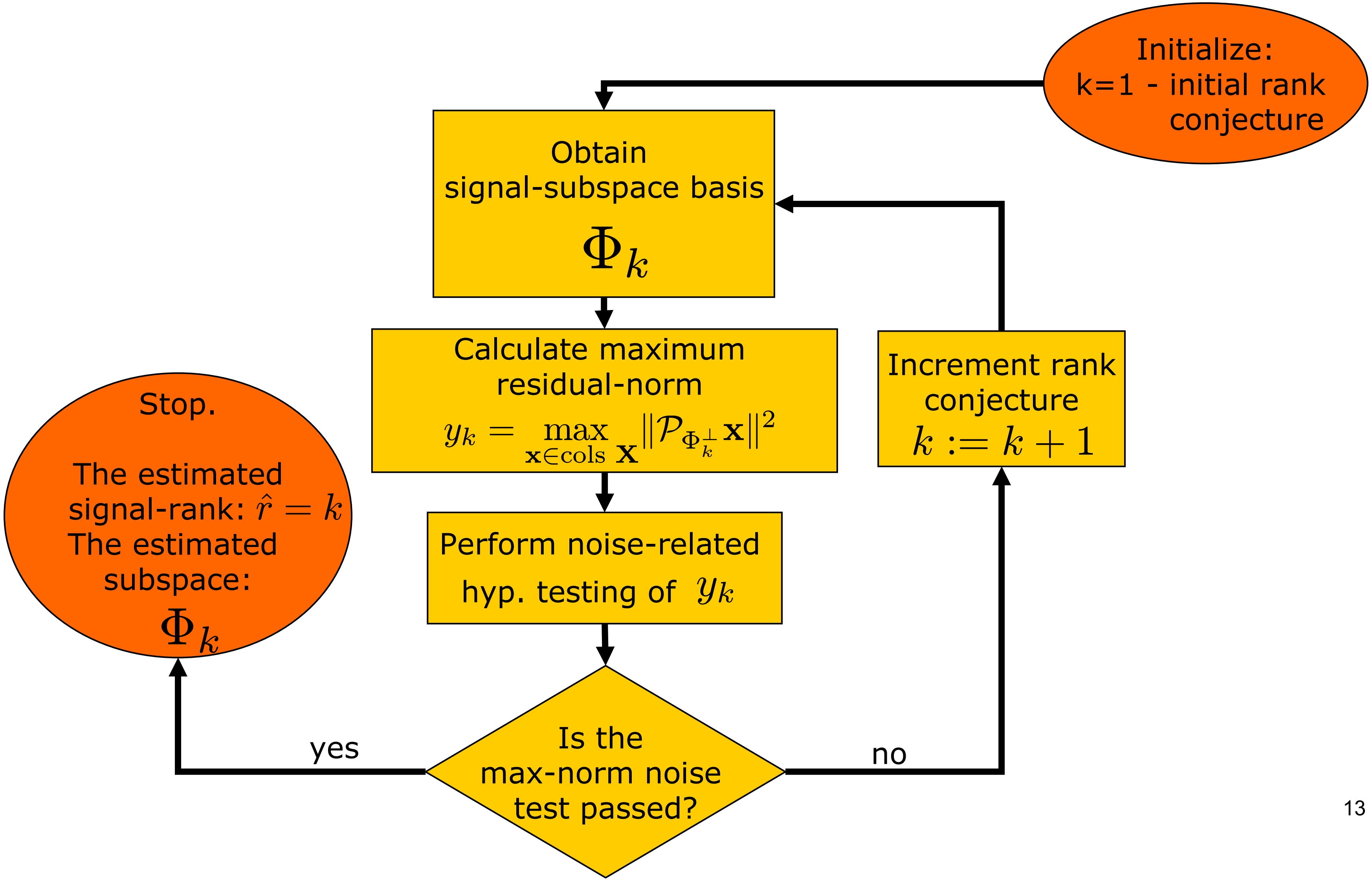
- The maximum signal-norm is assumed to be uniformly-distributed:

$$\xi_k \sim U(0, y_{k-1})$$

- y_{k-1}- is the maxim-norm of data-residuals obtained in the previous MOCA iteration



Signal-subspace and rank by MOCA Flowchart



Anomaly Detection Approaches: Matched Subspace Detector (MSD)

- Define two hypotheses:
$$\begin{aligned} H_0 &: \quad \mathbf{x}_i \sim \mathcal{N}[\mathbf{B}\mathbf{b}_i, \sigma^2 \mathbf{I}] \\ H_1 &: \quad \mathbf{x}_i \sim \mathcal{N}[\mathbf{B}\mathbf{b}_i + \mathbf{T}\theta_i, \sigma^2 \mathbf{I}] \end{aligned}$$
  - $\mathbf{B}$ background subspace basis
  - $\mathbf{T}$ anomaly subspace basis
- Generalized Log-Likelihood Ratio Test (GLRT)
$$L(\mathbf{x}) = \frac{1}{\sigma^2} \mathbf{x}^T \mathcal{P}_{\mathbf{B}^\perp \mathbf{T}} \mathbf{x} \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$
$\mathcal{P}_{\mathbf{B}^\perp \mathbf{T}}$ is a projection onto $(range\ \mathbf{B})^\perp \bigcap range\ \mathbf{T}$
- Drawbacks
  - The Background and Anomaly subspaces and their ranks are not known

Anomaly Detection Approaches: Gaussian Mixture Model (GMM)

- Gaussian mixture

$$p(\mathbf{x}) = \sum_{c=1}^{C} \alpha_c N(\mathbf{x}|\mu_c, \mathbf{\Gamma}_c)$$

- Hypotheses:

$$\begin{aligned} H_0 &: \quad \mathbf{x} \sim p(\mathbf{x}) \\ H_1 &: \quad \mathbf{x} - \mathbf{a} \sim p(\mathbf{x}) \end{aligned}$$

- GLRT - Reed Xiaoli (RX)

$$L(\mathbf{x}) = -\log p(\mathbf{x}) \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$

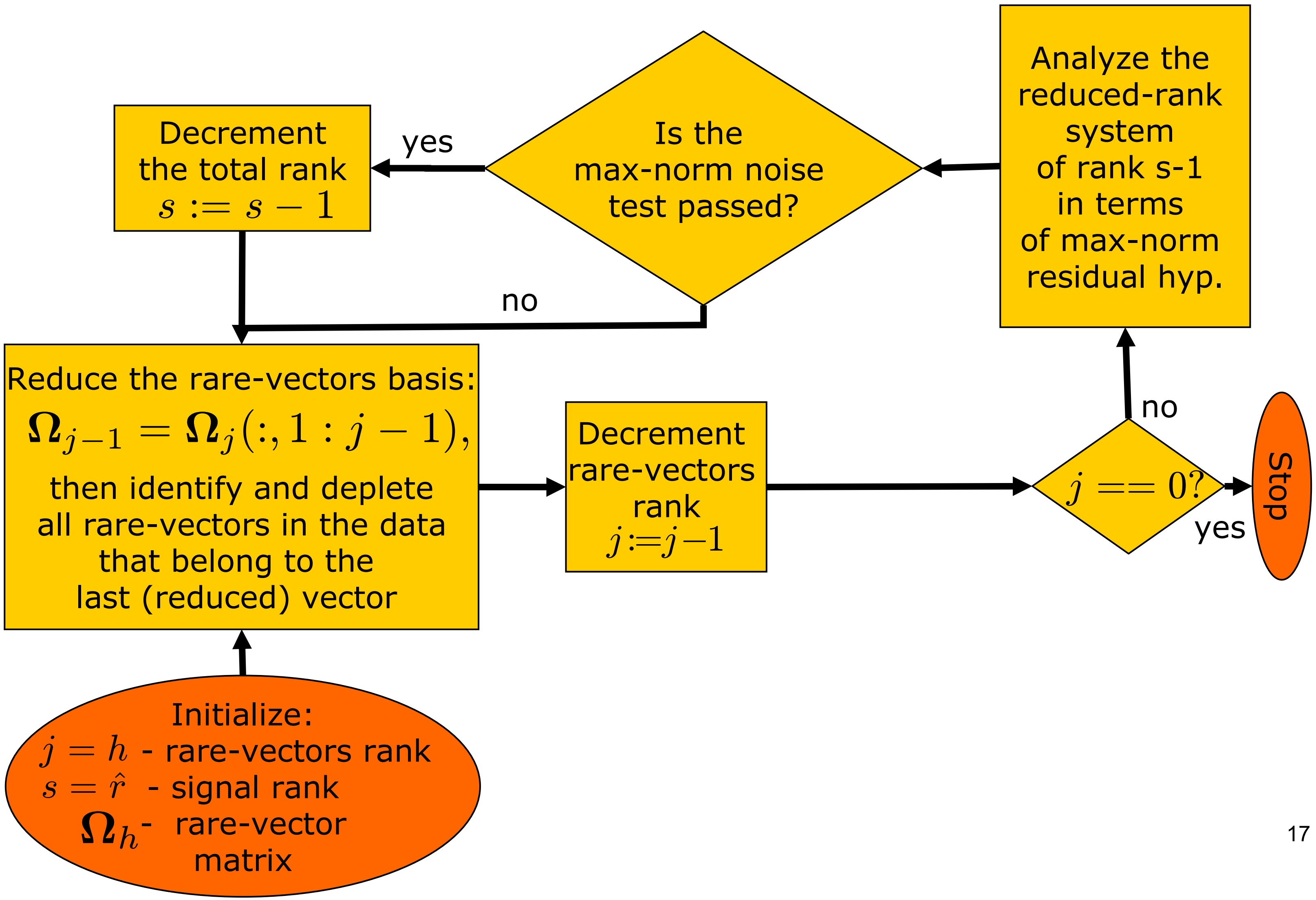
- Drawbacks
  - The number of Gaussians is not known
  - Initial-Condition dependent (prone to local minima)
  - Doesn't fit well the Hyperspectral data

MOCA versus MSD

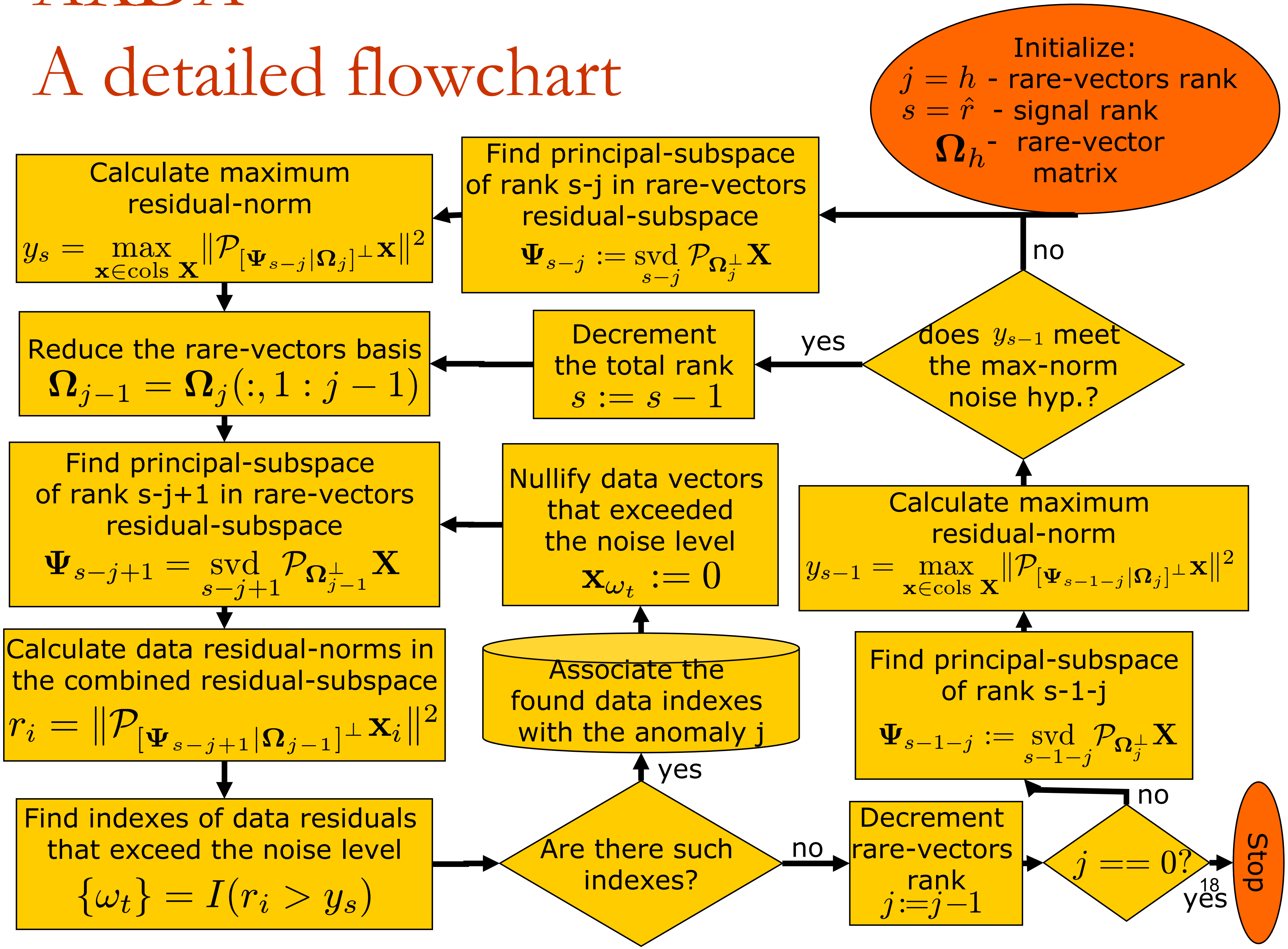
$$\left[\mathbf{\Psi}_{k-h}\middle|\mathbf{\Omega}_h\right]$$

- MSD
  - $\mathbf{B}$ background subspace basis
  - $\mathbf{T}$ anomaly subspace basis
- MOCA
  - $\mathbf{\Omega}_h$ represents anomaly subspace
  - $\mathbf{\Psi}_{k-h}$ complements $\mathbf{\Omega}_h$ to represent background

Anomaly Extraction and Discrimination Algorithm (AXDA)
A simplified outline

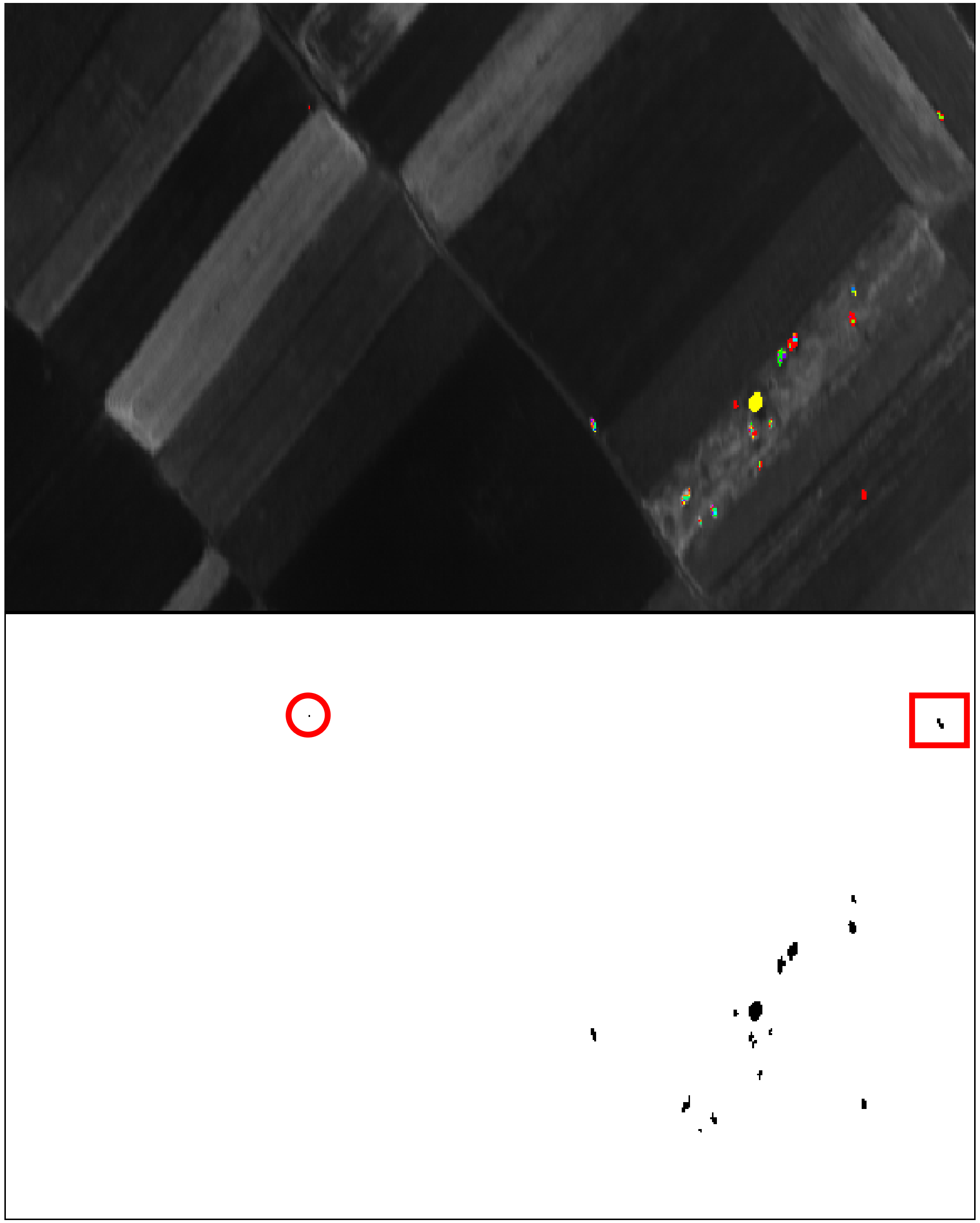


AXDA
A detailed flowchart



Results of MOCA/AXDA applied on real data

r = 19, h = 11, r_b = 14, # anomaly spectra = 11



- r – the estimated signal-subspace rank
- h – the estimated number of rare-vectors
- r_b – the obtained background rank
- # anomaly spectra – the obtained number of anomaly spectra in the scene
- ○ - a single-pixel man-made anomaly
- □ - a vegetation-related anomaly

r = 12, h = 6, r_b = 9, # anomaly spectra = 6

$r = 9, h = 0, r_b = 9$, # anomaly spectra = 0

No anomalies detected!

ROC curves

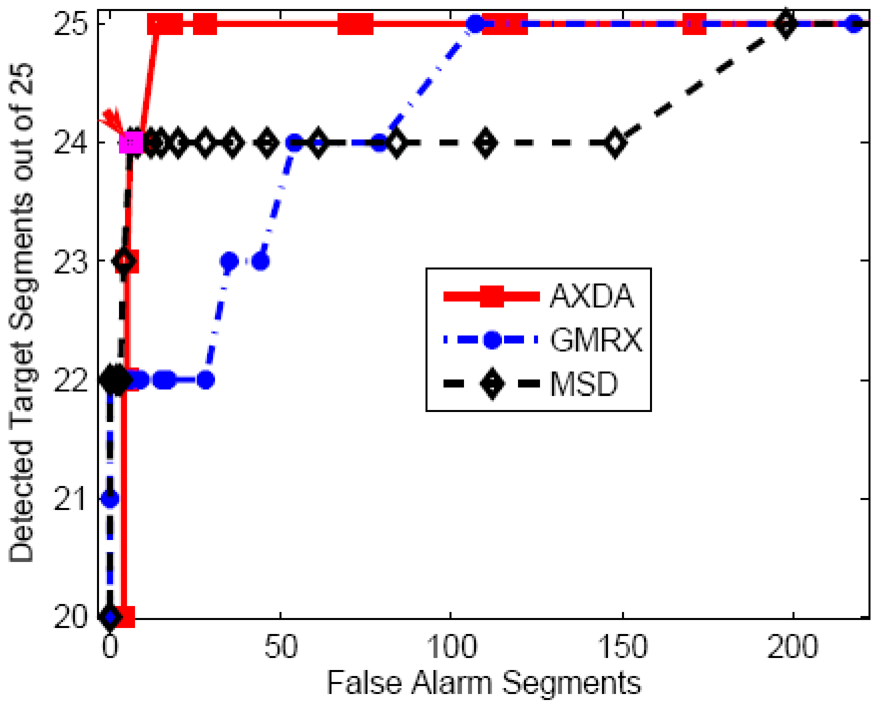


Fig. 6. **ROC curves corresponding to GMRX, MSD and AXDA.** The nominal operating point of AXDA is marked in magenta color and is pointed out by the arrow. This point corresponds to 24 detected anomalies and 6 false alarm segments.

Optimal $\ell_{2,\infty}$ minimization

$$\hat{\mathcal{S}}_r \quad = \quad \underset{\mathcal{L}}{\operatorname{argmin}} \; \|\mathcal{P}_{\mathcal{L}^\perp}\mathbf{X}\|_{2,\infty}^2 \equiv \underset{[\mathbf{W}]}{\operatorname{argmin}} \; \|\mathbf{W}^\top\mathbf{X}\|_{2,\infty}^2$$

$$\text{s.t.} \qquad \operatorname{rank} \mathcal{L} = r \qquad \text{s.t.} \qquad [\mathbf{W}] \in G_{p,p-r}$$

- $[\mathbf{W}]$ - equivalence class of $p \times p - r$ orthogonal matrices whose columns span the same subspace in $\mathbb{R}^p$ as $\mathbf{W}$
- $G_{p,p-r}$ - Grassmann manifold, the set of all $p - r$ -dimensional subspaces in $\mathbb{R}^p$

Line search on Grassmannian

- Continuous choice of subspaces on geodesics
- $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ - a descend direction of a function $F([\mathbf{W}])$
- The corresponding geodesic is:

$$\mathbf{W}(t) = (\mathbf{W}\mathbf{V} \;\; \mathbf{U}) \begin{pmatrix} \cos(t\mathbf{\Sigma}) \\ \sin(t\mathbf{\Sigma}) \end{pmatrix} \mathbf{V}^\top$$

- $t\mathbf{\Sigma}$ traverse *Principal Angles* between column spaces $[\mathbf{W}]$ and $[\mathbf{W}(t)]$
- Geodesic distance $d([\mathbf{W}(t)], [\mathbf{W}]) = t\sqrt{\mathrm{tr}(\mathbf{\Sigma}^2)}$
- Line search $\underset{t}{\operatorname{argmin}} F([\mathbf{W}(t)])$ (Armijo rule)

Gradient on Grassmannian

$$\nabla F = F_{\mathbf{W}} - \mathbf{W}\mathbf{W}^\top F_{\mathbf{W}}$$

- $F_{\mathbf{W}}$ - is the matrix of partial derivatives of $F([\mathbf{W}])$ with respect to the elements of $\mathbf{W}$
- ∇F is the projection of $F_{\mathbf{W}}$ onto the tangent space at $\mathbf{W}$
- Gradient of $\|\mathbf{W}^\top \mathbf{X}\|_{2,\infty}^2$:

single *maximum vector*

$$F_{\mathbf{W}} = \mathbf{x}_j \mathbf{x}_j^\top \mathbf{W}$$

multiple maximum vectors $\{\mathbf{x}_j\}\ j \in J$

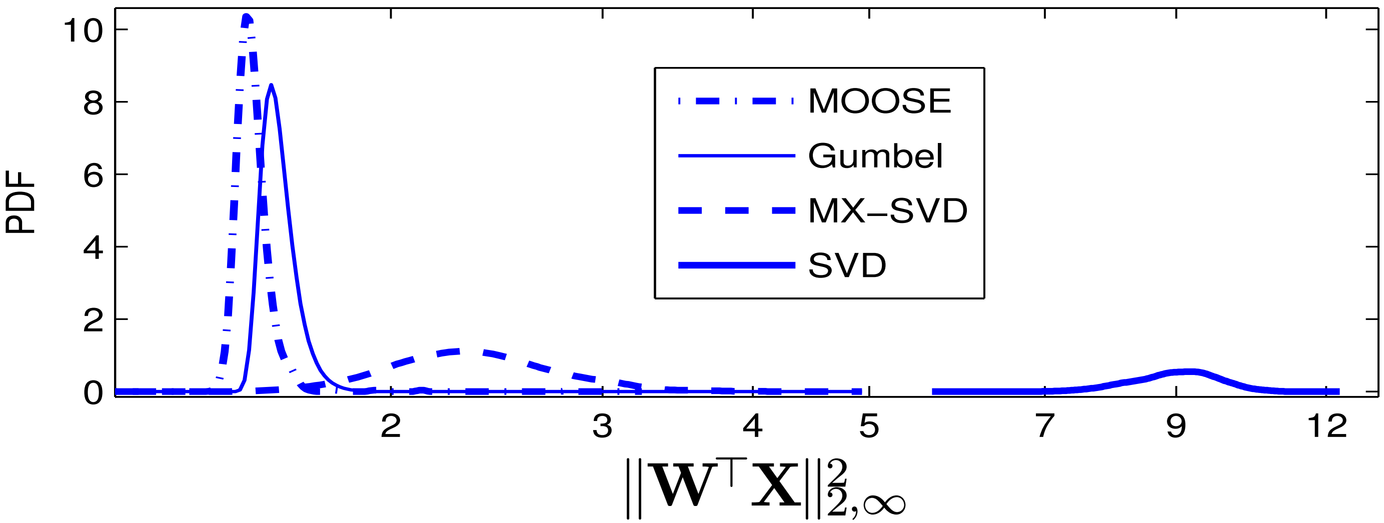
$$\begin{aligned} F_{\mathbf{W}} / \|F_{\mathbf{W}}\|_2 \quad &= \quad \max_{\mathbf{G}} \min_{j \in J} \langle \mathbf{G}, \mathbf{x}_j \mathbf{x}_j^\top \mathbf{W} \rangle \\ \text{s.t.} \quad & \langle \mathbf{G}, \mathbf{x}_j \mathbf{x}_j^\top \mathbf{W} \rangle > 0 \ \ \forall j \in J \\ & \langle \mathbf{G}, \mathbf{G} \rangle = 1 \end{aligned}$$

pdf of $\ell_{2,\infty}$- norm of residuals Monte-Carlo simulations

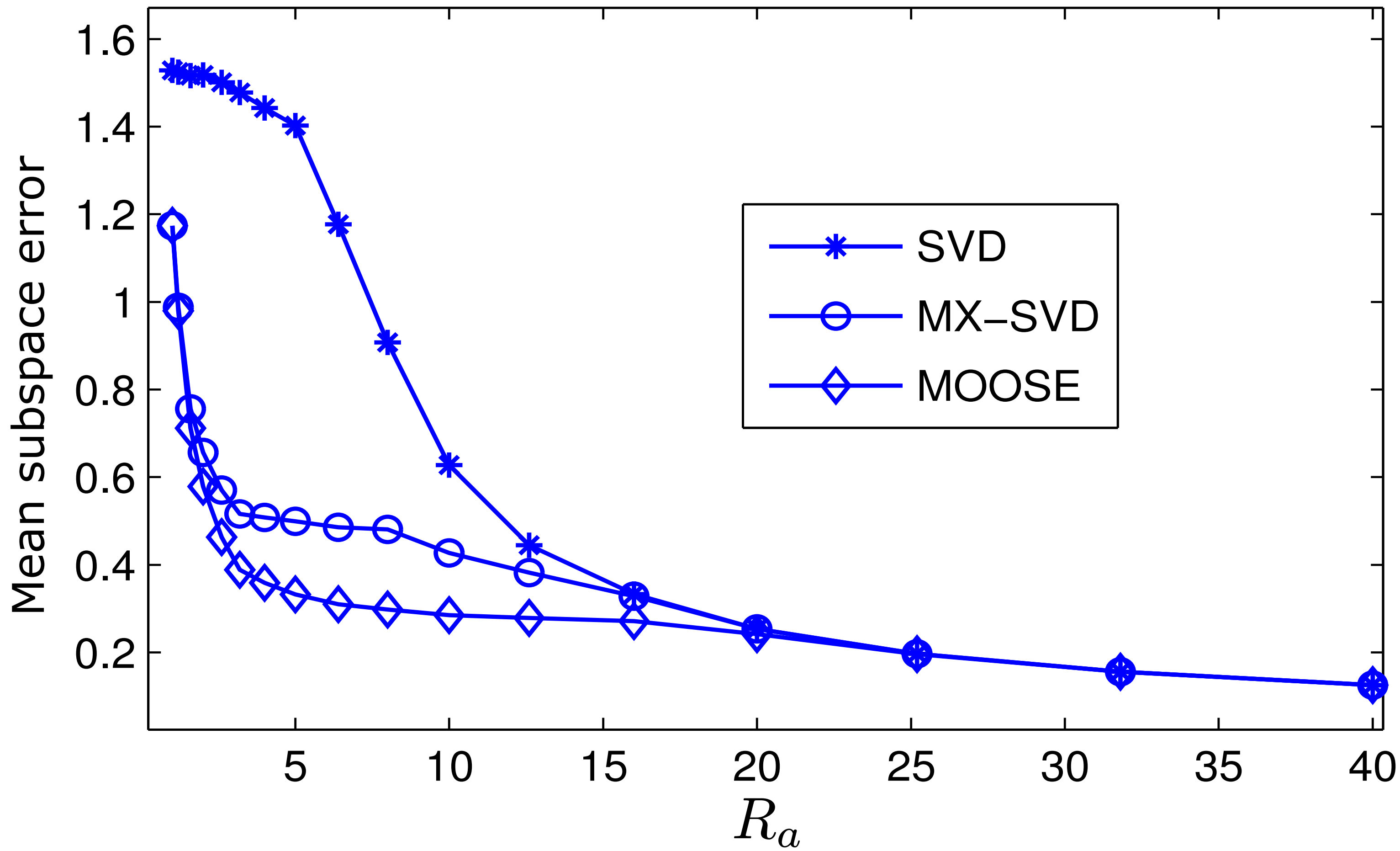
number of anomalies $N_a = 10$

anomaly subspace rank $r_a = 5$

anomaly loading ratio $R_a \triangleq N_a / r_a = 2$



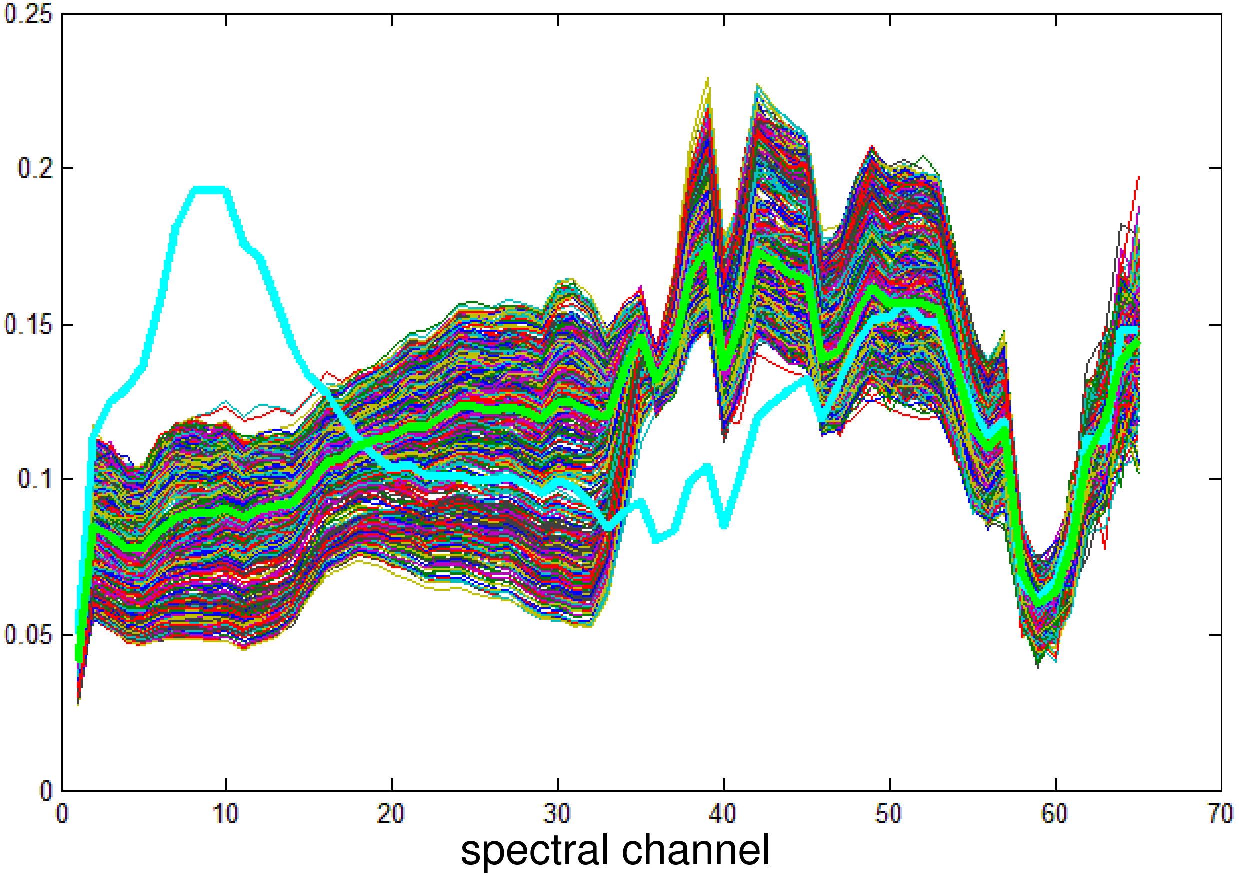
Mean estimated subspace error $\angle\{\hat{s}, s\}$



Designing Multispectral Filters for Anomaly Detection in Remote Sensing

Hyperspectral Anomalies

- Anomaly vs. Background spectra scatter



Mean background spectrum

Anomaly spectrum

65 channels

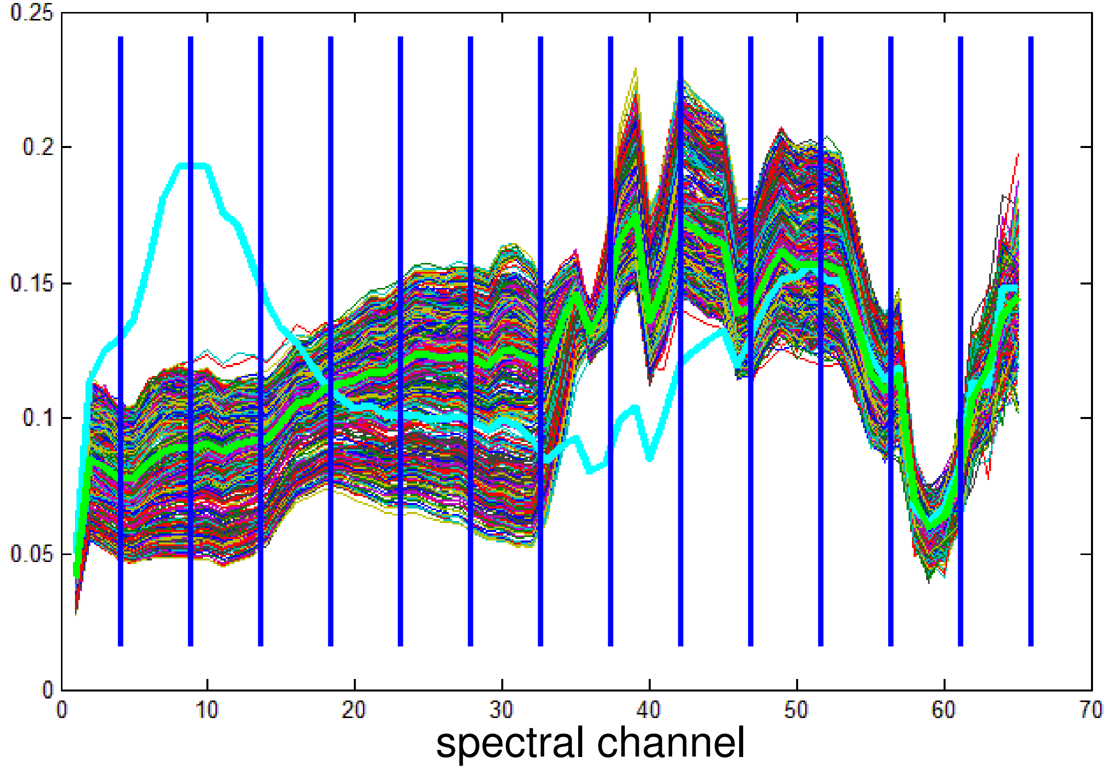
400nm – 1000nm

Hyperspectral Imager Drawbacks

- Expensive
- Heavy
- High power consumption
- Fragile

From Hyperspectral to Multispectral

- Given hyperspectral images, partition spectra into a number of bands K



- Which partition is better for detecting anomalies for a given K?

Problem Statement

- Determine a vector of *K* breakpoints $\mathbf{b}_K = \{b_1, \ldots, b_K\}$
- Corresponding to $K - 1$ contiguous intervals $I_k = [b_k, b_k + 1)$
- Producing a set of constants at each pixel *j* $\mu_{k,j} = \frac{1}{|I_k|} \sum_{i \in I_k} x_{i,j}$
- Such that a cost function is minimized $J(\mathbf{b}_K, \mathbf{X})$

Fast Hyperspectral Feature Reduction (FFR) [1]

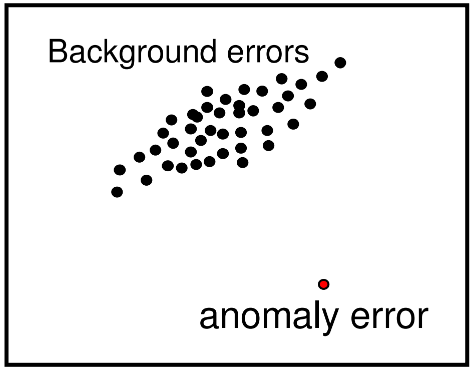
- Error vector at interval *k* in pixel *j*
  $$\mathbf{e}_{k,j} \triangleq \{(x_{i,j} - \mu_{k,j}) : i \in I_k\}$$
- Squared error at pixel *j*
  $$e_j^2 \triangleq \sum_{k=1}^{K-1} \|\mathbf{e}_{k,j}\|^2$$
- Sum of squared error cost
  $$J(\mathbf{b}_K, \mathbf{X}) \triangleq \sum_{j=1}^{N} e_j^2$$

- *Is not sensitive enough to anomaly contributions*
- *Partition is governed by the background process*
- *Anomalies are misrepresented*

The proposed Maximum of Mahalanobis Norms (MXMN)

- Error vector at interval *k* in pixel *j* $\mathbf{e}_{k,j} \triangleq \{(x_{i,j} - \mu_{k,j}) : \ i \in I_k\}$
- Error covariance matrix in an interval *k* $\mathbf{\Sigma}_k$
- Mahalanobis norm of error *j* in interval *k* $G(\mathbf{e}_{k,j}) = \sqrt{\mathbf{e}_{k,j}^\top \mathbf{\Sigma}_k^{-1} \mathbf{e}_{k,j}}$
- Potential Anomaly Loss (PAL) measure $D_k = \max_{j=1}^{N} G(\mathbf{e}_{j,k})$
- Maximum of PAL measures cost $J(\mathbf{b}_K, \mathbf{X}) = \max_{k=1}^{K} D_k$

- *Allows capturing misrepresented anomaly contributions*
- *Eliminates heavy tails of errors pdf*
- *The more "Gaussian" interval is, the coarser is the partition in it*

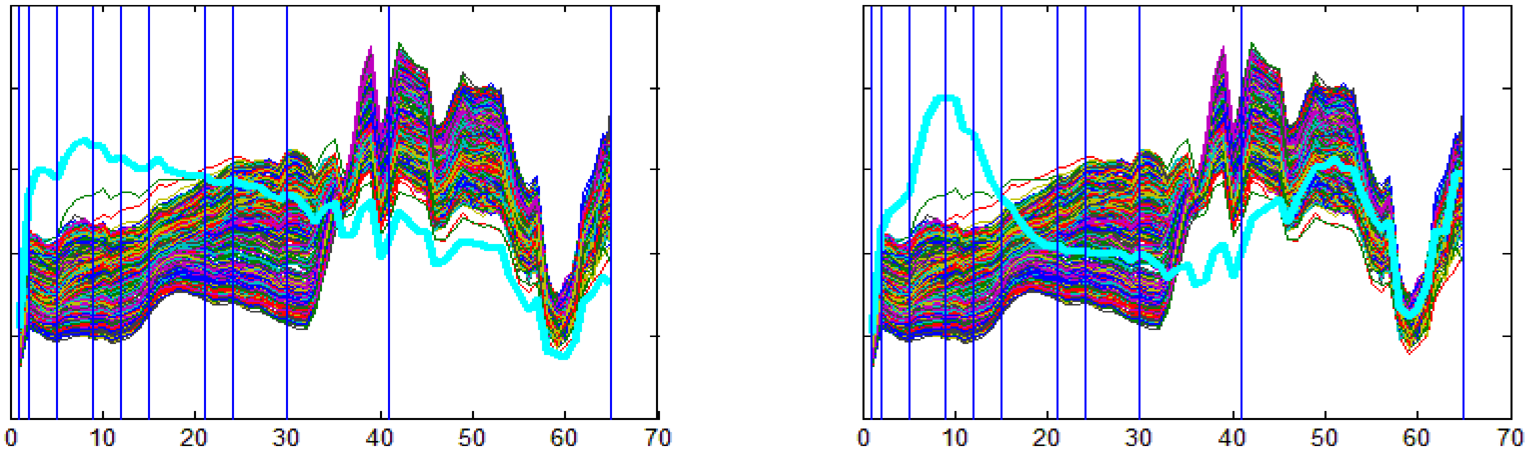


MXMN Training

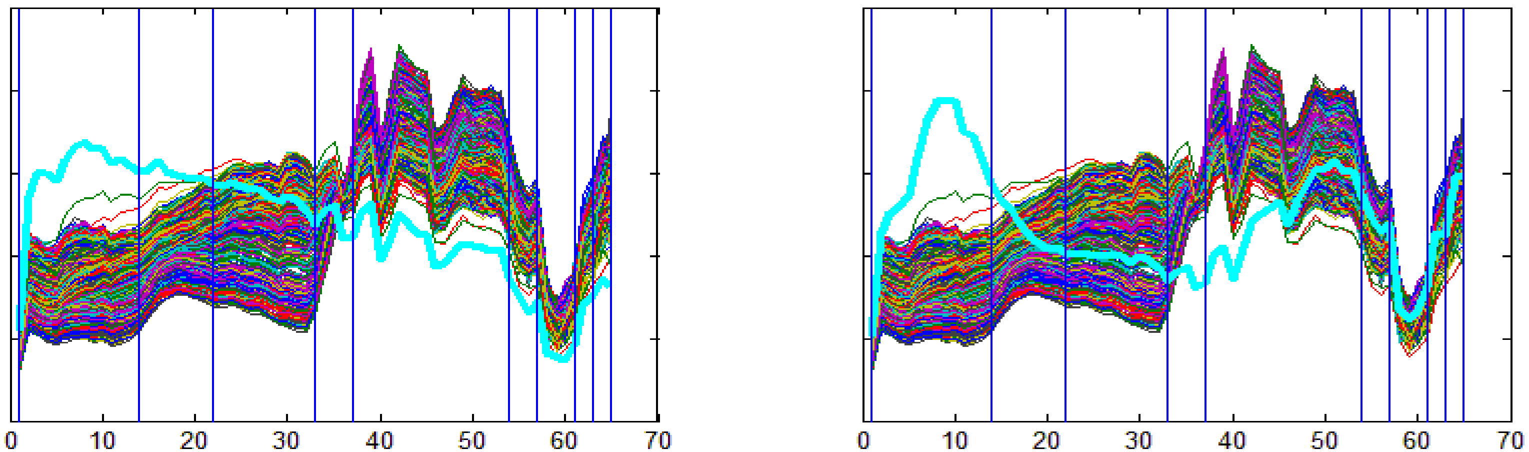
- MXMN produces a coarser partition in “more Gaussian” channels
- We attribute these channels to background clutter
- Usually, background clutter channels are not good for anomaly mining
- Moreover, they may mask anomalies during detection
- As a matter of fact, all background clutter related information is found in images without anomalies
- By design, MXMN is able to identify and disregard background clutter channels by a coarse partitioning
- A reasonable question is whether exploiting background clutter statistics alone suffice for obtaining a good partition for anomaly detection purposes
- In our simulations we examined the hard case by training MXMN on an image that does not contain known anomalies

Partition breakpoints: MXMN vs. FFR
MXMN trained on image without anomalies

Partition breakpoints by MXMN



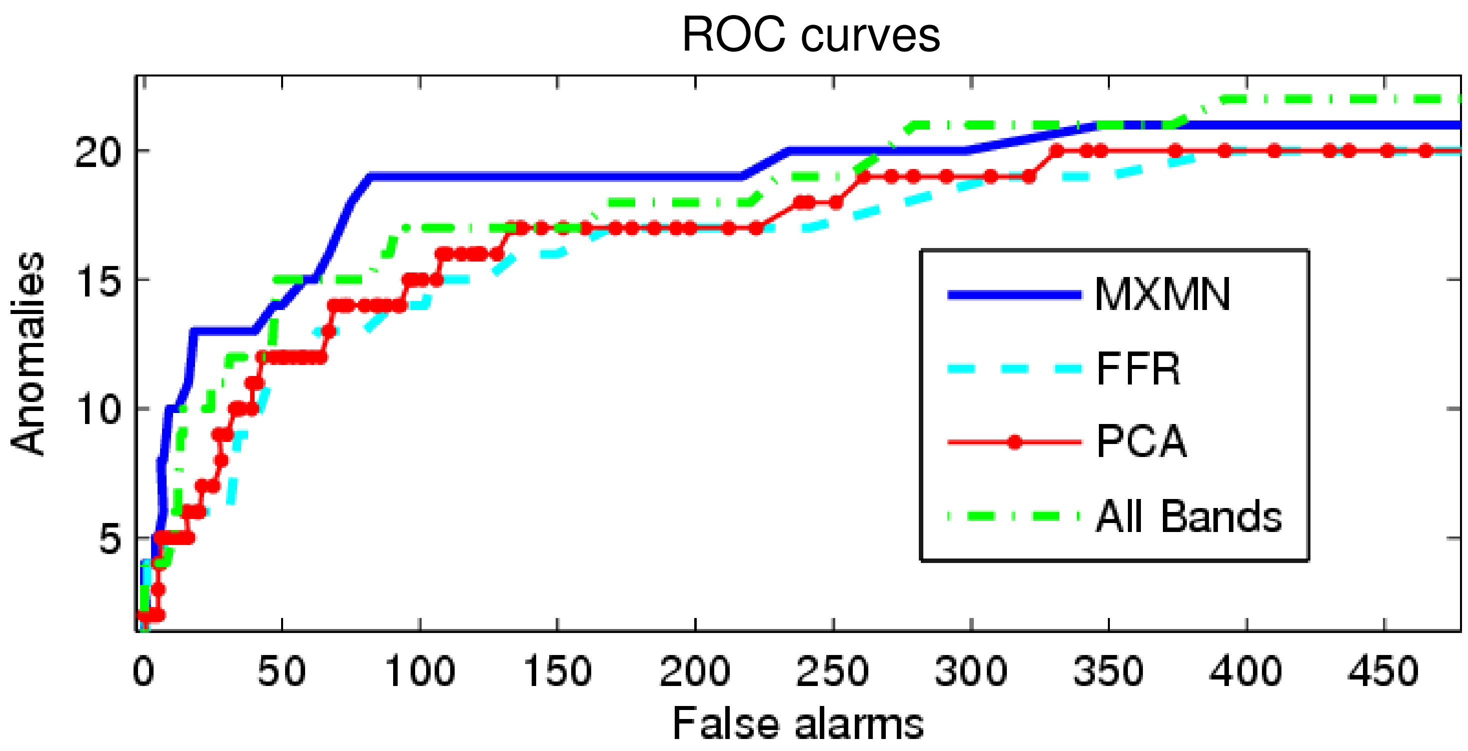
Partition breakpoints by FFR



Number of Multispectral Filters 10 | Number of Hyperspectral Channels 65

ROC curves by applying RX

- Number of hyperspectral channels 65
- Number of multispectral channels 10
- Number of hyperspectral images 6 (300 x 400 x 65)



Summary

- MOCA – a greedy algorithm for anomaly preserving signal subspace estimation based on $\ell_{2,\infty}$ - norm minimization
- The signal-subspace rank is estimated by applying *Extreme Value Theory* results to the $\ell_{2,\infty}$ - norm of residuals
- The structure of MOCA is used for developing *Anomaly Extraction and Discrimination Algorithm* - AXDA
- $\ell_{2,\infty}$ - optimal anomaly preserving subspace estimation on a *Grassmann Manifold*
- The principle of maximum error norm minimization is used for multispectral filter design tuned to anomaly detection algorithms

Future directions

- Develop an anomaly detection algorithm that is based on $\ell_{2,\infty}$ -optimal subspace
- Develop a technique for a robust anomaly discrimination/classification
- Determination of multispectral dimensionality, which is optimal in terms of anomaly detection algorithm performance

Thank you!