Learning Patterns for Detection with Multiscale Scan Statistics

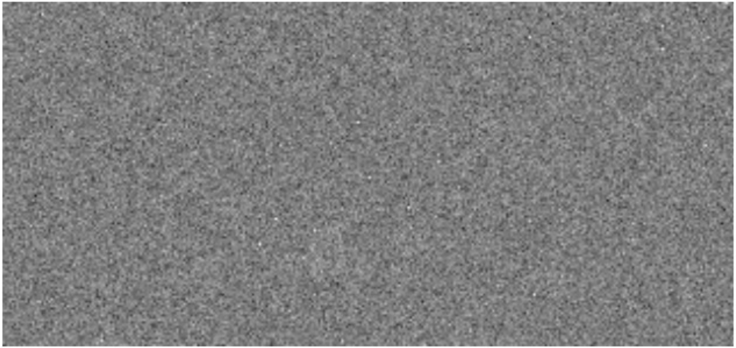
J. Sharpnack[1]

[1]Statistics Department
UC Davis

UC Davis Statistics Seminar 2018

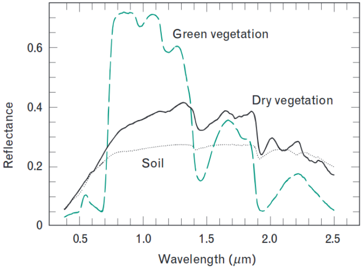
Work supported by NSF DMS-1712996

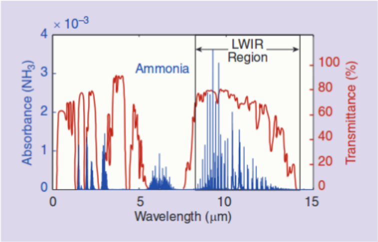
Hyperspectral Gas Detection



Hyperspectral Gas Detection

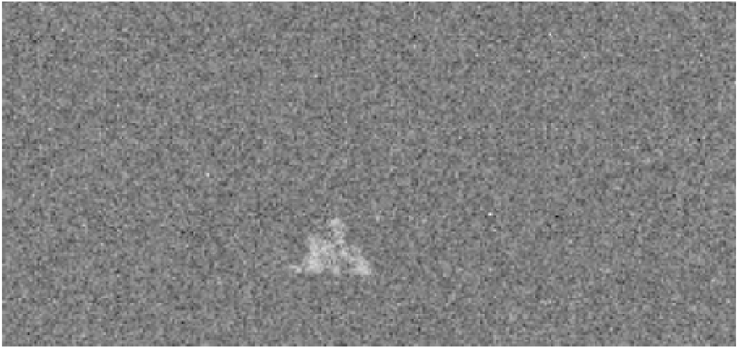
Each pixel of the image is a light spectrum, and we are interested in detecting a single chemical signature.





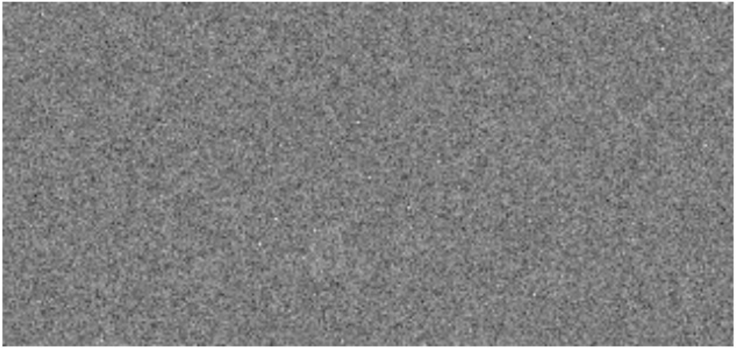
Images from [Manolakis et al. 2014, Manolakis & Shaw 2002]

Hyperspectral Gas Detection



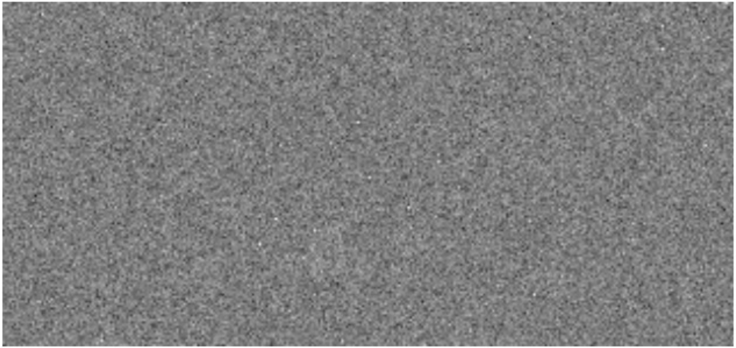
Gas signature level in each pixel. Dataset and chemical signature comparison code from Dimitris G. Manolakis and DTRA.

Hyperspectral Gas Detection



Decreased SNR by $\frac{1}{5}$th. Can you see it?

Hyperspectral Gas Detection



If classification answers the question, "what am I seeing?" detection answers the question, "do I see anything at all?"

Anomaly detection goals

Goal 1: *Reliably detect anomalous patterns within images beyond what the human eye can see—at the precise information theoretic limit.*

Detection applications

- Contaminant detection in water networks,
- real-time surveillance system,
- radiation monitoring,
- fire detection and other remote sensing applications,
- medical imaging and automated radiology,
- early detection of pathogen outbreaks.

Rectangular multiscale scan statistic

Scan every rectangle (vary location and scale) looking for abnormal concentrations [S, Arias-Castro '16].

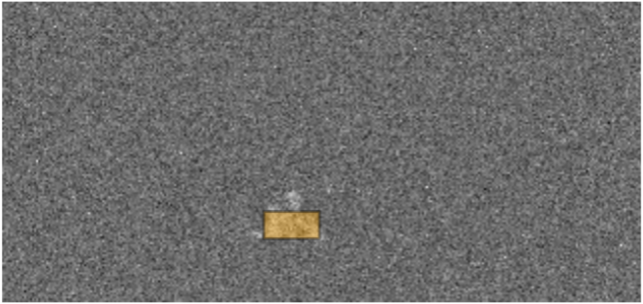


Figure: A rectangle over the active region.

Scan statistic

Represent image over $[-L, L] \times [-L, L]$ as matrix Y and consider scanning rectangle over pixels $[-H_0, H_0] \times [-H_1, H_1]$. Define the pattern to be

$$P_{k,l} = \frac{1}{\sqrt{(2H_0+1)\cdot(2H_1+1)}}$$

then the scan is the following convolution,

$$(P \star Y)_{k,l} = \sum_{k'=-H_0}^{H_0} \sum_{l'=-H_1}^{H_1} Y_{k-k',l-l'} P_{k',l'}, \quad \text{where defined.}$$

The single-scale scan statistic is

$$\hat{s} = \max_{k,l} (P \star Y)_{k,l}.$$

Other patterns

General pattern over the domain $\Omega = [-L, L]^d$ can be hidden in the noisy tensor.

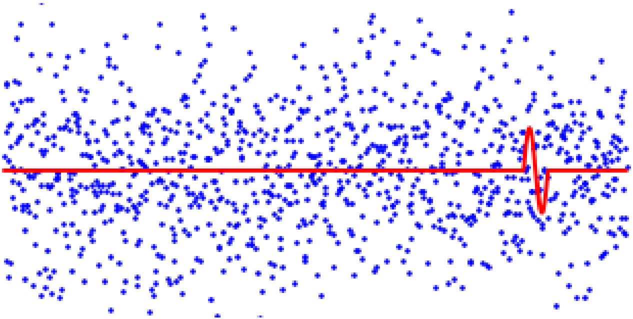
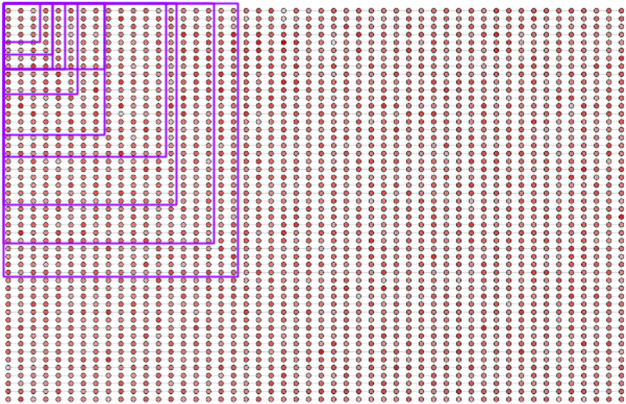


Figure: A simulated time series with an embedded sinusoidal signal with values on the y-axis ($d = 1$).

Multiscale scan

Scanning with many pattern dimensions, $H_0, \ldots, H_d$, is called a multiscale scan.



How do we compare scan statistics at different scales?

Anomaly detection goals

Goal 2: *General purpose analysis for multiscale scan statistics for a large class of patterns.*

Multiple Tensors

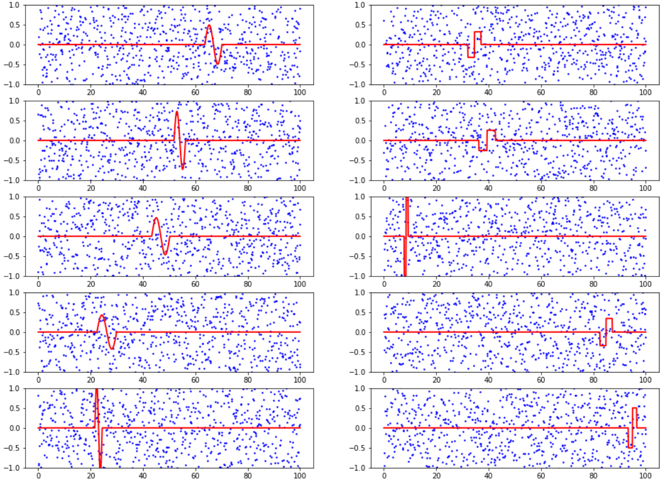


Figure: Multiple tensors, $i = 1, \ldots, n$ $(n = 5)$, with different locations and scales, and two possible patterns (left and right).

Anomaly detection goals

Goal 3: *Leverage database of tensors to simulaneously learn and detect anomalous patterns.*

Prior work

[Naus '65] scan statistics introduced for point cloud data

[Siegmund, Worsley '95] limit distribution of 1-**dimension**al scan

[Glaz and Zhang '04, Kabluchko '11] limit in d-dimensions

[Arias-Castro et al. '05, '11] scan for blob-like **patterns**

[Dumbgen, Spokoiny, '01] **scale adaptive** scan statistic ($d = 1$)

[S, Arias-Castro '16] scale adaptive rectangular scan

[Proksch at al. '17] scale adaptive smooth patterns

This work: learning and detecting **general smooth patterns** in a **database of tensors** with **scale adaptive methods**

Continuous model

- Pattern $f \in \mathcal{F} \subset C^1$ over $[-1,1]^d$, $\|f\|_{L_2} = 1$.
- Data is random measure $\mathrm{d}X^i$ with domain $[-L, L]^d$.
- Scale dilation $f_h := h_\bullet^{-1/2} f(./h)$, $h_\bullet = \prod_j h_j$, $h \in \mathbb{R}^d$
- Null hypothesis: data is just noise ($\mathrm{d}W^i$ is d-dimensional Wiener process)
- Alternative hypothesis: there is a signal f at location t^i, and scale h^i.

$$
\begin{aligned}
H_0 &: \mathrm{d}X^i(\tau) = \mathrm{d}W^i(\tau), i = 1, \dots, n \\
H_1 &: \mathrm{d}X^i(\tau) = \mu f_{h^i}(t^i - \tau)\mathrm{d}\tau + \mathrm{d}W^i(\tau) \\
&\qquad \text{for some } f \in \mathcal{F}, \text{ and } (t^i, h^i) \in \mathcal{D}, i = 1, \dots, n.
\end{aligned}
$$

Continuous multiscale scan statistic

Convolution at scale h,

$$(f_h \star \mathrm{d}X^i)(t) = \int f_h(\tau)\mathrm{d}X^i(t-\tau) = \int \frac{1}{\sqrt{h_\bullet}} f(\tau)\mathrm{d}X^i(t-h\tau),$$

Scale corrected multiscale scan statistic:

$$s(X^i; f) := \max_{h\in\mathcal{H}} v_h \left(\max_{t\in\mathcal{T}_h} (f_h \star \mathrm{d}X^i)(t) - v_h \right). \tag{1}$$

- $h \in \mathcal{H} := \times_j [1, L)$
- $t \in \mathcal{T}_h := \times_j [-(L-h_j), L-h_j]$
- $v_h = \sqrt{2\sum_j \log(L/h_j)}$

Test if the pattern f centered at t and scaled by h is hidden within tensor X^i.

Learning patterns

Given a dataset of images $X^i, i = 1, \ldots, n$ then can we also learn the pattern $f \in \mathcal{F}$?

$$S_n(X; \mathcal{F}) := \max_{f \in \mathcal{F}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s(X^i; f) \tag{PAMSS}$$

The pattern adapted multiscale scan statistic (PAMSS) averages the MSS for each tensor.

Smoothness conditions on $\mathcal{F}$ are required: bounded variation (TVC) or average Hölder condition (AHC).

Smoothness assumptions

Assumption TVC: Define the isotropic total variation,

$$\|f\|_{\mathrm{TV}} := \int_{\Omega} \|\nabla f(u)\|_2 \mathrm{d}u,$$

function are of bounded variation,

$$\exists \gamma_1 > 0 \text{ s.t. } \forall f \in \mathcal{F}, \quad \|f\|_{\mathrm{TV}} \leq \gamma_1. \tag{TVC}$$

or **Assumption AHC:** Define the Hölder functional,

$$A_{t,s}(f) := \int_{\Omega_L} |f(t-z) - f(s-z)|^2 \, \mathrm{d}z.$$

functions have bounded average Hölder condition,

$$\exists 0 < \gamma_2 \leq 1 \text{ s.t. } \forall f \in \mathcal{F}, \quad A_{t,s}(f) \leq c_A \|t-s\|_2^{2\gamma_2}. \tag{AHC}$$

Type 1 error control

We can simulate from the null distribution to obtain a significance level for

$$s(X^i; f) := \max_{h \in \mathcal{H}} v_h \left(\max_{t \in \mathcal{T}_h} (f_h \star \mathrm{d}X^i)(t) - v_h \right). \tag{2}$$

[Dumbgen & Spokoiny '01] showed that under H_0 for L large enough

$$s(X^i, f) = O_{\mathbb{P}} (\log \log L)$$

for functions satisfying (TVC) in 1D.

> *We need a more precise control to analyze PAMSS (tail bound)—$s(X^i, f)$ is subexponential random variable.*

SubGaussian process

Definition

We say that a random field, $\{Z(\iota)\}_{\iota \in \mathcal{I}}$, is a (zero mean) standard subGaussian process if there exists a constant $u_0 > 0$ such that

$$\mathbb{P}\{|Z(\iota_0) - Z(\iota_1)| \geq u\} \leq 2\exp\left(-\frac{u^2}{2\nu(\iota_0, \iota_1)}\right), \tag{3}$$

$$\mathbb{P}\{Z(\iota_0) \geq u\} \leq \exp\left(-\frac{u^2}{2}\right), \tag{4}$$

for any $\iota_0, \iota_1 \in \mathcal{I}$, $u > u_0$, and $\nu(\iota_0, \iota_1) = \sqrt{\mathbb{E}(Z(\iota_0) - Z(\iota_1))^2}$, is the canonical distance.

Under H_0, $\{(f_h \star \mathrm{d}X^i)(t) : (t, h) \in \mathcal{D}\}$ is a subGaussian random field with canonical distance

$$\nu_f((h_0, t_0), (h_1, t_1)) := \|f_{h_0}(t_0 - .) - f_{h_1}(t_1 - .)\|_{L_2}.$$

Dudley's chaining

Define the **covering number** of metric space $(\mathcal{I}, \nu)$, $\mathcal{N}(\mathcal{I}, \nu, \epsilon)$, to be the number of balls of ν-radius ϵ that is required to cover $\mathcal{I}$.

Theorem (Dudley's entropy (tail) bound)

$$\mathbb{P}\left\{\sup_{\eta\in\mathcal{I}} Z(\eta) > u \cdot c \cdot \mathbf{D} + C\right\} \leq Ce^{-\frac{u^2}{2}},$$

such that

$$\mathbf{D} = \int_0^\infty \sqrt{2\log\mathcal{N}(\mathcal{I}, \nu, \epsilon)}\mathrm{d}\epsilon$$

for universal constants c, C. Specifically, if

$$\mathcal{N}(\mathcal{I}, \nu, \epsilon) \leq \Gamma\epsilon^{-\rho}.$$

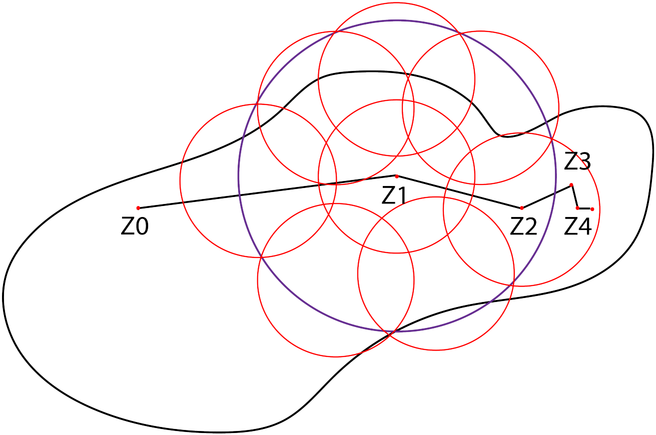
then (as $\Gamma \to \infty$)

$$\mathbf{D} = \sqrt{2\log\Gamma} + o(1).$$

Dudley's chaining

Find a sequence (the chain) $\{\eta_k\}_{k=0}^{\infty}$ such that $\lim_{k\to\infty} \eta_k = \eta$,

$$\sup_{\eta\in\mathcal{I}} Z(\eta) = \sup_{\eta_0,\eta_1,\dots} Z(\eta_0) + (Z(\eta_1) - Z(\eta_0)) + (Z(\eta_2) - Z(\eta_1)) + \dots$$



Dudley's chaining

Find a sequence (the chain) $\{\eta_k\}_{k=0}^{\infty}$ such that $\lim_{k\to\infty} \eta_k = \eta$,

$$\sup_{\eta\in\mathcal{I}} Z(\eta) = \sup_{\eta_0,\eta_1,\dots} Z(\eta_0) + (Z(\eta_1) - Z(\eta_0)) + (Z(\eta_2) - Z(\eta_1)) + \dots$$

- $|Z(\eta_{i+1}) - Z(\eta_i)|$ is prop. to $\nu(\eta_{i+1}, \eta_i)$
- Choose covering layers such that $\mathcal{N}$ grows like 2^{2^k}
- Variance of the bound is dominated by $Z(\eta_0)$

Chaining standardized suprema

Theorem

Let $Z(\eta)$ be a standard subGaussian process over an index set $\mathcal{I}$. Suppose that metric $(\mathcal{I}, d_Z)$ has covering number, $\mathcal{N}$ s.t.,

$$\mathcal{N}(\mathcal{I}, d_Z, \epsilon) \leq \Gamma \epsilon^{-\rho}. \tag{5}$$

Then there exists an $\Gamma_0 > 0$ such that for any $\Gamma \geq \Gamma_0$, the following supremum is bounded in probability,

$$\mathbb{P}\left\{\sqrt{c_0 \log \Gamma}\left(\sup_{\eta \in \mathcal{I}} Z(\eta) - \sqrt{2 \log \Gamma}\right) - a_0 \log \log \Gamma > u\right\} \leq e^{-u}, \tag{6}$$

for $u > u_0$ where u_0, c_0, a_0 are constant depending on ρ (but not on Γ). In words, the supremum of such a subGaussian process is subexponential with location and rate parameter, $(2 \log \Gamma)^{1/2}$ (omitting the $\log \log$ term).

Proof sketch for chaining bound

For iid normals, $\{z_i\}_{i=1}^N$, from union bound

$$\mathbb{P}\left\{\max_i z_i > \sqrt{2\log N + u^2}\right\} \leq e^{-\frac{u^2}{2}}.$$

Generic chaining: $\sqrt{2\log N + u^2} \leq u + \sqrt{2\log N}/(2u)$
Our chaining: $\sqrt{2\log N + u^2} \leq \sqrt{2\log N} + u^2/(2\sqrt{2\log N})$

$$\mathbb{P}\left\{2\sqrt{2\log N}\left(\max_i z_i - \sqrt{2\log N}\right) > u\right\} \leq e^{-u}.$$

Modify chain so that

(1) start the chain at a deeper level (N large enough)

(2) make the covers grow slowly $\mathcal{N} \leq a^{a^k}$ for $a \to 1$.

Main Theorem

Theorem

Let $\mathcal{F}$ be finite and assume that either all functions in $\mathcal{F}$ satisfy either (TVC) or (AHC). Let

$$F_n(\delta) := \begin{cases} \sqrt{K \log\left(\frac{|\mathcal{F}|}{\delta}\right)}, & \log|\mathcal{F}| \leq \frac{n}{K} + \log\delta \\ \frac{K}{\sqrt{n}} \log\left(\frac{|\mathcal{F}|}{\delta}\right), & \log|\mathcal{F}| > \frac{n}{K} + \log\delta \end{cases} \tag{7}$$

then for some constant K, under H_0,

$$\mathbb{P}\left\{S_n(X, \mathcal{F}) > F_n(\delta) \cdot \log\log L\right\} \leq \delta. \tag{8}$$

Proof of main theorem

Lemma

Then, under the above conditions, there is a constant C depending on d alone such that

1. *Suppose that (TVC) holds for the class $\mathcal{F}$, then*

$$\nu_f((t,h),(t',h'))^2 \leq C\gamma_1 \left(\left\| \frac{t-t'}{h} \right\|_2^2 + \left(\sqrt{\frac{h'_\bullet}{h_\bullet}} - 1 \right)^2 \right).$$

2. *[Proksch et al. '16] Suppose that (AHC) holds for the class $\mathcal{F}$, then*

$$\nu_f((t,h),(t',h'))^2 \leq C \left(\left\| \frac{t_j - t'_j}{h_j} \right\|_{2\gamma_2}^{2\gamma_2} + \left\| \frac{h_j - h'_j}{\sqrt{h_j h'_j}} \right\|_{2\gamma_2}^{2\gamma_2} \right).$$

Proof of main theorem

Lemma

Suppose that $f \in \mathcal{F}$ satisfies either (TVC) or (AHC). Let $\ell \in \{0, \dots, \lfloor \log_2 L \rfloor\}^d$, and $\mathcal{H}_2(\ell) = \times_j [2^{\ell_j}, 2^{\ell_j+1}]$. Then

$$\mathbb{P}\left\{ c_1 \cdot \max_{h \in \mathcal{H}_\ell, t \in \mathcal{T}_h} v_h \left((f_h \star \mathrm{d}X^i)(t) - v_h \right) - a_1 \log \log L > u \right\} \leq e^{-u}$$

for constants $a_1, c_1 > 0$ depending on γ, d only.

With the union bound over ℓ,

$$\mathbb{P}\left\{ c_2 \cdot \frac{s_n(X^i, f)}{\log \log L} - a_2 > u \right\} \leq e^{-u},$$

then use subexponential Bernstein inequality.

Asymptotic Distinguishability

Corollary

Suppose that $\log|\mathcal{F}| = o(n)$*, and recall that under the alternative hypothesis,* H_1*,* X^i *has an embedded pattern* f *at scale* h^i *and* $v_{h^i}^2 = \sum_j \log(L/h_j^i)$*, and the noise is a standard Wiener process. Suppose also that* $h_j^i \leq L^c$ *for some* $0 \leq c < 1$ *for all* i, j*, then the PAMSS is asymptotically powerful (has diminishing probability of type 1 and type 2 error) if*

$$\mu - \sqrt{2} \cdot \frac{\sum_{i=1}^{n} v_{h^i}^2}{\sum_{i=1}^{n} v_{h^i}} \to \infty. \tag{9}$$

We take this result to mean that as long as the function class, $|\mathcal{F}|$*, does not grow exponentially in* n*, we achieve asymptotic power under the same conditions as if* $|\mathcal{F}| = 1$*.*

Summary

- Proposed the Pattern Adapted Multiscale Scan Statistic for learning anomalous patterns
- We proved a refined concentration result for the supremum of subGaussian processes
- We controlled the error probabilities for the PAMSS showing that it can learn the function from a class for free as long as $n \gg \log |\mathcal{F}|$

For a link to this paper: http://jsharpna.github.io

Thanks!