Learning Patterns for Detection with Multiscale Scan Statistics

J. Sharpnack[1]

[1]Statistics Department
UC Davis

ITA 2018

Outline

Outline

Anomaly detection

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

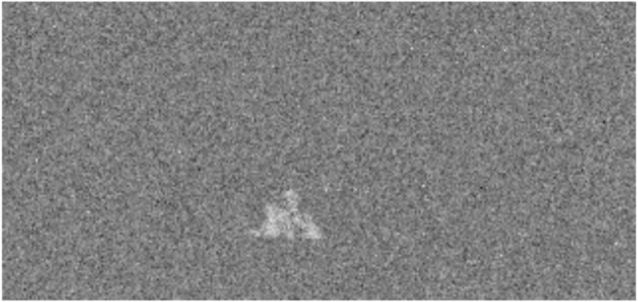


Figure: An image with an anomalous region of contaminant.

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].

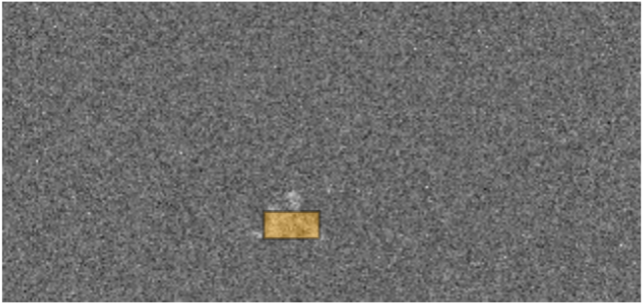


Figure: A rectangle over the active region.

Other patterns

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

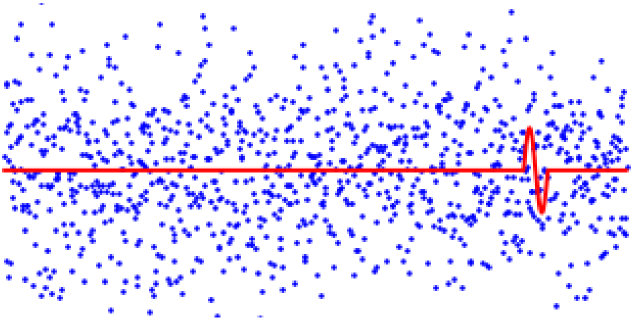


Figure: A simulated time series with an embedded sinusoidal signal with values on the y-axis ($d = 1$).

Outline

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**

Outline

Simple scan statistic

For an image, $Y_{k,l} : k, l = -L, \dots, L$ we can convolve a pattern $P_{k,l} : k, l = -H, \dots, H$ with the image,

$$(P \star Y)_{k,l} = \sum_{k',l'=-H}^{H} Y_{k-k',l-l'} P_{k',l'}, \quad k, l = -L + H, \dots, L - H.$$

Then the simple scan statistic is $\max_{k,l} (P \star Y)_{k,l}$.

- For an arbitrary P, would like to scale both dimensions, so that $H \leftarrow H'_j$ in dimension j
- For general functions f over Ω need to rasterize/iterpolate
- Cumbersome and unenlightening analysis, so we model the problem as continuous

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.

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

Outline

Continuous multiscale scan statistic

New 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:

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

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

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 $f \in \mathcal{F}$?

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

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

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

Outline

Epsilon-net architecture

Natural notion of distance (shift operator S_t):

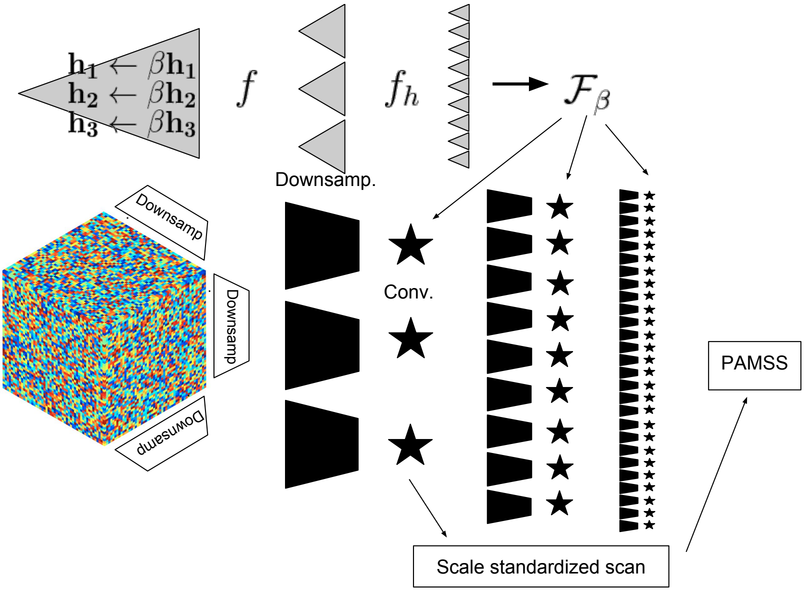
$$\nu_f((t,h),(t',h')) := \|S_t f_h - S_{t'} f_{h'}\|_{L_2}$$

Definition

An ϵ-net is a subset of scale and locations $\mathcal{D}_{\text{net}}$ such that for any t, h there is an element $t', h' \in \mathcal{D}_{\text{net}}$ such that

$$\nu_f((t,h),(t',h')) \leq \epsilon.$$

Epsilon-net architecture



Outline

Chaining standardized suprema

Theorem

Let $Z(\eta)$ be a standard subGaussian process over an index set $\mathcal{I}$. Suppose that the metric space $(\mathcal{I}, d_Z)$ has

$$\mathcal{N}(\mathcal{I}, d_Z, \epsilon) \leq \Gamma \epsilon^{-\gamma}. \tag{2}$$

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

$$\mathbb{P}\left\{ \sqrt{c_0 \log \Gamma} \left(\sup_{\eta \in \mathcal{I}} Z(\eta) - \sqrt{2 \log \Gamma} \right) - a_0 \log \log \Gamma > u \right\} \leq e^{-u}, \tag{3}$$

for $u > u_0$ where u_0, c_0, a_0 are constant depending on γ (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 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: $\sqrt{2\log N + u^2} \leq \sqrt{2\log N} + u^2/(2\sqrt{2\log N})$

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

Chain is a sequence of partitions $\mathcal{A}_k$ of $\mathcal{I}$, we do

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

(2) make the partitions be smaller $|\mathcal{A}_k| \leq a^{a^k}$ for $a \to 1$.

Outline

Main Theorem

Theorem

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

$$F_n(\delta) := \begin{cases} \sqrt{K \log\left(\frac{|\mathcal{F}|}{\delta}\right)}, & \log|\mathcal{F}| \leq \frac{n}{K} + \log\delta \\ \frac{K}{\sqrt{n}} \log\left(\frac{|\mathcal{F}|}{\delta}\right), & \log|\mathcal{F}| > \frac{n}{K} + \log\delta \end{cases} \tag{4}$$

then for some constant K,

$$\mathbb{P}\left\{S_n(X, \mathcal{F}) > F_n(\delta) \cdot \log\log L\right\} \leq \delta. \tag{5}$$

Asymptotic Distinguishability

Define $V_n = \sum_i v_{h^i}^2$ and $W_n = \sum_i v_{h^i}$, then

$$\frac{\mu W_n - V_n}{\sqrt{n}} - F_n(\delta) \log\log L = \omega\left(\sqrt{\frac{V_n}{n}}\right)$$

is sufficient for distinguishing H_0 from H_1. For $n = 1 = |\mathcal{F}|$,

$$\mu - v_{h^1} - \frac{K}{v_{h^1}} \log\frac{1}{\delta} \cdot \log\log L = \omega(1),$$

which matches known conditions (up to constants).

Summary

- The pattern adapted multiscale scan statistic can be implemented with a deep convolutional architecture
- We proved a refined concentration result for the supremum of subGaussian processes
- We controlled the error probabilities for the PAMSS
- Future work
  - Theory for infinite function classes (outer loop chaining)
  - SGD with soft-max activations (active)
  - Goodness-of-fit tests for convolutional nets and deep autoencoders

Thanks!

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 $\ell \in \{0, \ldots, \lfloor \log_2 L \rfloor\}^d$, and $\mathcal{H}_2(\ell) = \times_j [2^{\ell_j}, 2^{\ell_j+1}]$. Then

$$\mathbb{P}\left\{ c_1 \cdot \max_{h \in \mathcal{H}_\ell, t \in \mathcal{T}_h} v_h \left((f_h \star \mathrm{d}X^i)(t) - v_h \right) - a_1 > u \right\} \leq e^{-u} \qquad (6)$$

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.