Optimal Spatial Anomaly Detection: Theory and Applications

Chao Zheng
School of Mathematical Sciences & S3RI
University of Southampton

23 May 2025
YMSC, Tsinghua University

Joint work with

Baiyu Wang (PhD 2022-26, University of Southampton)

Outline

1. Background and preliminaries—Multiple CPD and AD in timeline
2. Spatial Anomaly Detection (SAD)—methodology and theory
   - Set-up and fundamental challenges
   - Penalised regional cost approach
   - Technical assumptions and consistent estimator
   - Minimax optimality in anomaly localisation
   - Extensions
3. Numerical studies
   - A dynamic algorithm for computation
   - Monte Carlo simulations
   - Detect Marine Heatwave

1. Background and preliminaries — multiple CPD/AD in timeline

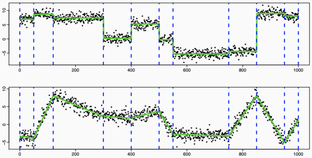
CPD and AD in timeline

Changepoint Detection (CPD) and Anomaly Detection (AD):

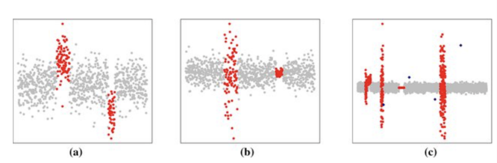
- Data generating process: observations ordered by time index

  $\{x_t\}, t = 1, \dots, n;$ changepoints (anomalies) at $1 < \tau_1 < \cdots < \tau_m < n$

- Common Structure within each segment. Different Structure (different depending on the model) across segments



(a) Change-in-mean and change-in-slope



(b) Anomalies and baseline

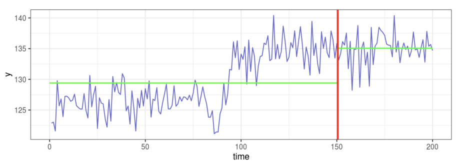
Detect the multiple changepoints

Objectives (in asymptotic view: $\lim_{n\to\infty}$): estimate the number of changepoints, and localise each changepoint within an error-range $\xi = o(n)$:

$$\mathbb{P}(\widehat{m} = m^*) \to 1; \qquad \mathbb{P}\left(\max_{j=1,\dots,m^*} |\widehat{\tau}_j - \tau_j^*| \le \xi\right) \to 1. \tag{1}$$

Detectability relies on SNR:

- minimum signal jump: $\Delta_n = \min_j \text{dist}(\theta_{j+1}, \theta_j)$
- minimum segment length: $\delta_n = \min_j(\tau_{j+1} - \tau_j)$

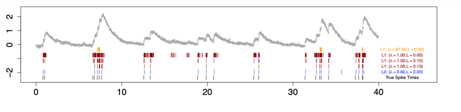


e.g.,for i.i.d Gaussian change-in-mean problem, we require $\delta_n \Delta_n^2 \gtrsim \log n$

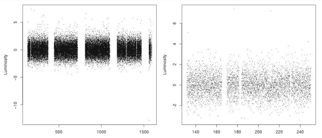
State-of-the-art

Long history since WWII (process/quality control, Duncan, 1952; Page, 1954). Renaissance in recent years.

- Parametric changes $(F_{\tau_{j-1}+1}, \ldots, F_{\tau_j})$ share a common parameter , e.g., change-in-mean (Yao, 1988; Wang et al., 2020), change-in-varaince (Fisch et al., 2022), change-in-slope (Baranowski et al., 2016; Fearnhead et al, 2019)...
- Nonparametric changes (James and Matteson, 2014; Li et al.,2016; Chu and Chen, 2018)....
- High-dimensional data: Roy et al. (2017), Wang and Samworth (2018); Zhang et al. (2022);...
- Time seires/Factor model: Cho and Fryzlewicz (2015); Basu and Michailidis (2015); Barigozzi et al. (2018); Safikhani and Shojaie (2020); McGonigle et al. 2021,...
- Functional data: Hörmann and Kokoszka (2010), Aston and Kirch (2012), Gromenko et al. (2012), ...
- Robust estimation: Fearnhead and Rigaill (2019); Li and Yu (2021);...
- Online setting: Chen et al. (2020); Romano et al. (2021); Padilla et al. (2019)
- ...



Calcium imaging data (Jewell and Witten, 2018)



Kepler light curve data (Fisch et al., 2022)

Approach 1: sequential detection

Single changepoint: carry out a two-sample test on each location $i \in \{1, \ldots, n\}$.

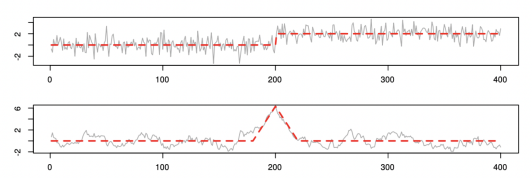


Figure 1: Single changepoint data and test statistics on each location

Multiple changepoints:

Binary Segmentation (Vostrikova, 1981) – detect a first changepoint that split the data into two parts, then repeat the operation on the two resulting sub-data...

Variants: WBS (Fryzlewicz, 2014), NoT (Baranowski et al.,2019), Seeded Binary Segmentation (Kovács, et al. 2020), MOSUM (Eichinger and Kirch, 2018),...

Approach 2: Penalised cost minimisation

Minimize a (additive) segment cost function that measures the fit to data, e..g., least square, negative log-likelihood, Huber/biweight functions,...

$$L(x_{s:t}) = \sum_{i=s}^{t} (x_i - \bar{x}_{s:t})^2.$$

Total cost associated with fitting segmentation $\tau_{1:m}$:

$$\mathbf{L}\left(x_{1:T}; \tau_{1:m}\right) = \sum_{j=1}^{m+1} L\left(x_{\tau_{j-1}+1:\tau_j}\right) \qquad \Rightarrow \text{overfitting}$$

Estimate m^* and $\tau_{1:m^*}$ that minimizes a penalised cost:

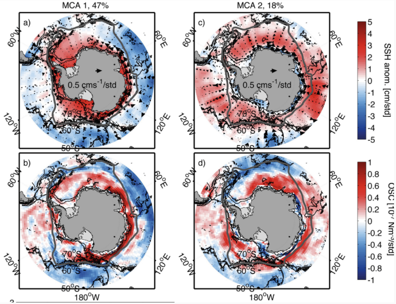
$$\text{argmin}_{(m,\tau_{1:m})} \mathbf{L}\left(x_{1:T}; \tau_{1:m}\right) + \beta m,$$

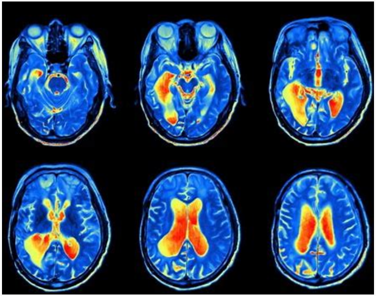
where $\beta = C \log n$ for the change-in-mean problem. (Boysen, et al., 2009).

2. Spatial anomaly detection (SAD) — methodology and theory

Spatial Grid: satellite-derived/geo-spatial data; meteorology; biomedical (fMRI)....

Question: automatically identify/segment anomaly regions in 2D/3D data?





Model set-up: data generating processes

Consider n realisations of $\left\{X(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{S} \subset \mathbb{Z}^2\right\}$, assume

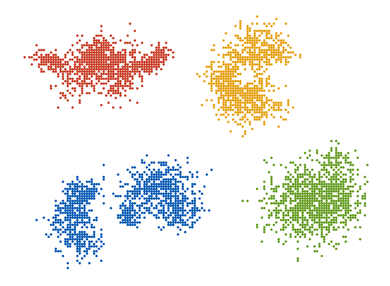
$$X(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \varepsilon(\boldsymbol{s}), \qquad \boldsymbol{s} \in \{1, 2 \ldots, \sqrt{n}\} \times \{1, 2 \ldots, \sqrt{n}\},$$

where $\{\mu(\boldsymbol{s})\}$: deterministic mean signal; $\{\varepsilon(\boldsymbol{s})\}$: random errors.

(For now we just consider the simplest 2D data)

Mean Anomaly Regions: $\mathcal{S}$ is partitioned into: one baseline region R_0 and m anomaly regions $R_1, , \ldots, R_m$, on which:

$$\mu(\boldsymbol{s}) = \mu_j, \ \forall \boldsymbol{s} \in R_j \quad \text{and} \quad \mu_j \neq \mu_0, \qquad j = 1, \ldots, m.$$



Fundamental challenges of SAD

Quantity of interest: estimate $(m; R_{1:m}; \mu_{1:m})$

Comparing to CPD/AD in timeline:

- spatial regions are not specify by "just two boundary points"
- arbitrary shape (non-convexity, non-connective components, holes,...)

From a high-level perspective: lack of total ordering in higher dimensions — no spatial past/future.

To name a few consequent challenges:

Methodology: Sequential detection methods, e.g., based on Binary Segmentation and Moving Window, are no longer applicable.

Theory: Number of candidate anomaly regions exploded from $\binom{n}{2}$ to 2^n

Computation: Linear time pruning algorithms, such as PELT and FPOP, do not work.

Similar in spirit to clustering (k-means)/image segmentation...

Penalised cost approach (and problem?)

Similarly, consider $\{R_1, ..., R_m\}$ and $R_0 = \mathcal{S} \setminus \bigcup_{j=1}^{m} R_j$. Regional cost (least square) and L_0 penalty can be used to measure its fit to the data:

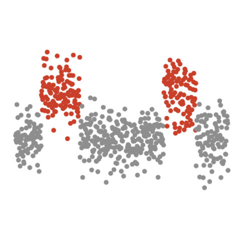
$$L(R_0) + \sum_{j=1}^{m} L(R_j) + \beta m,$$

Penalised cost approach (and problem?)

Similarly, consider $\{R_1, ..., R_m\}$ and $R_0 = \mathcal{S} \setminus \bigcup_{j=1}^{m} R_j$. Regional cost (least square) and L_0 penalty can be used to measure its fit to the data:

$$L(R_0) + \sum_{j=1}^{m} L(R_j) + \beta m,$$

However the above cost cannot specify distant spatial anomaly regions !!!



Penalised cost approach (and problem?)

Similarly, consider $\{R_1, ..., R_m\}$ and $R_0 = \mathcal{S} \setminus \bigcup_{j=1}^m R_j$. Regional cost (least square) and L_0 penalty can be used to measure its fit to the data:

$$L(R_0) + \sum_{j=1}^{m} L(R_j) + \beta m,$$

However the above cost cannot specify distant spatial anomaly regions !!!

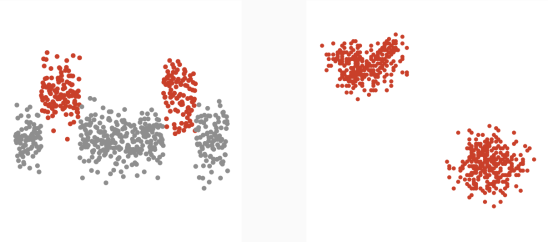


Figure 3: anomaly regions having similar similar mean signals in timeline and in 2-D

Regional penalty: size of minimum convex hull

Therefore we add a second penalty on the intrinsic area of anomalies:

$$\mathcal{C}(m, R_{1:m}) := \mathbf{L}(R_{1:m}) + \beta m + \lambda \sum_{j=1}^{m} \left|\mathrm{Co}(R_j)\right|, \tag{2}$$

where $\mathrm{Co}(\cdot)$ is the minimum convex hull, $|\cdot|$ is the set cardinality; β and λ are penalty parameters.

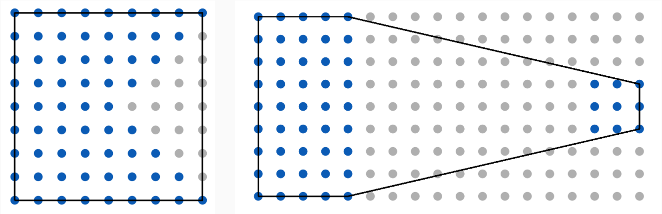


Figure 4: minimum convex hull illustrations: the smallest convex set that covers the region (a polygon with no interior angles $> 180°$)

Next we will show the minimiser of (2) is a valid estimate for SAD.

Technical assumptions

(**Sub-Gaussianity**) $\{\varepsilon(\boldsymbol{s})\}$ are i.i.d centered subG(σ^2).

(**Signal strength**) (i) For any $\eta > 0$, there exists $C_\eta > 0$, such that

$$\frac{\Delta^2}{\sigma^2} \geq C_\eta \cdot \frac{\log^{1+\eta} n}{\sqrt{n}}.$$

(ii) There exists $0 < C_R < 1$, s.t. $\delta \geq C_R \cdot n$.

(iii) There exist $0 < d_A < d_B < 1$, s.t.

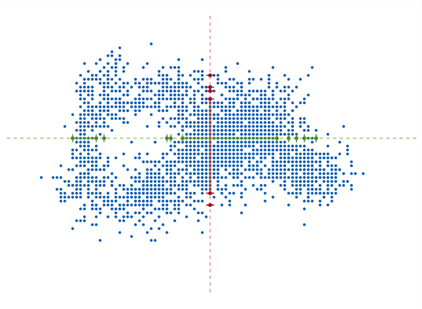
$$\max_j r_{R_j^*} \leq d_A\sqrt{n}, \quad \text{and} \quad \min_{i,j} \text{dist}(R_i^*, R_j^*) \geq d_B\sqrt{n}.$$

Geometric assumptions: region smoothness

We allow arbitrary shape of anomaly regions (nonconvexity, non-connectivity, holes,...)

We take the minimisation over a subset of regions satisfying:

Smoothness condition (1): At any fixed x or y coordinate, each anomaly region can be divided into at most K consecutive segments for some $K < \infty$.



Smoothness condition (2): For all true anomaly regions $R^*_{1:m^*}$, $\text{Co}(R^*_j)$ have at most H vertices for some $H < \infty$.

Detection consistency

Define the measure of symmetric regional difference

$$H(R, R') := |R \setminus R'| + |R' \setminus R| = |R \cup R'| - |R \cap R'|$$

(similar to $|\tau - \tau'|$ in timeline)

Theorem 1: Let $\beta = C_\beta \sqrt{n} \log n$ and $\lambda = C_\lambda \log n / \sqrt{n}$. There exists constant $c_\gamma > 0$ that only depends on K and σ^2, such that

$$\widehat{m} = m^* \quad \text{and} \quad H\left(\widehat{R}_i, R_i^*\right) \leq \frac{C_\epsilon \sigma^2}{\Delta^2} \sqrt{n} \log n, \;\; i = 1, \ldots, m$$

holds with prob. at least $1 - 2 \exp\left(-c_\gamma \sqrt{n} \log n\right)$

Remark: Detection rate $\sqrt{n} \log n$ is different from that in time line case ($\log n$).

Minimax optimality

Minimax theory provides the optimal localisation error that can be achieved among a class of models. For univariate timeline change-in-mean problem: (Wang and Samworth, 2018; Wang et al. 2020; Verzelen et al. 2020):

- minimax lower bound of the localization error $O(1)$, when $\delta\Delta^2/\sigma^2 \geq \log n$.
- L_0 penalisition and WBS achieve the optimal rate up to a logarithm factor

Minimax optimality

Minimax theory provides the optimal localisation error that can be achieved among a class of models. For univariate timeline change-in-mean problem: (Wang and Samworth, 2018; Wang et al. 2020; Verzelen et al. 2020):

- minimax lower bound of the localization error $O(1)$, when $\delta\Delta^2/\sigma^2 \geq \log n$.
- L_0 penalisition and WBS achieve the optimal rate up to a logarithm factor

Whilst we derive the following result for SAD:

Theorem 2: Under previous model and technical assumptions, we have the minimax lower bound:

$$\inf_{\widehat{R}} \sup_{P\in\mathcal{P}} \mathbb{E}_P\left[H\left(\widehat{R}, R^*(P)\right)\right] \geq \begin{cases} \frac{n}{64}, & \text{if } \delta\Delta^2/\sigma^2 < \sqrt{n}\log n \\ \frac{\sigma^2}{\Delta^2}\frac{\sqrt{n}}{2}, & \text{if } \delta\Delta^2|/\sigma^2 \geq \sqrt{n}\log n, \end{cases}$$

Extension to spatial dependent data

Previously we assume $\{\varepsilon(\boldsymbol{s})\}$ are i.i.d. Now consider the spatial dependence for any $R \in \mathbb{R}_K$ satisfying:

$$\mathbb{E}\left[\exp\left(\lambda \sum_{\boldsymbol{s}\in R} \varepsilon(\boldsymbol{s})\right)\right] \leq \exp(\lambda^2 \sigma^2 |R|^{\phi}), \quad 1 \leq \phi < \frac{3}{2}.$$

Theorem 3 With the same penalised cost function and choice of penalty parameters as in the independent case, we have

$$\widehat{m} = m^* \quad \text{and} \quad H\left(\widehat{R}_i, R_i^*\right) \leq \frac{C_\epsilon \sigma^2}{\Delta^2} n^{\phi - \frac{1}{2}} \log n, \;\; i = 1, \ldots, m$$

holds with probability one as $n \to \infty$.

Remark 1: Theorem 1 immediately follows when $\phi = 1$.

Remark 2: Stronger dependence (larger ϕ) leads to inferior detection rate.

Extension to general dimension

Consider $\boldsymbol{s} \in \mathcal{S} \subset \mathbb{R}^d$, and each dimension can be of different size (hyper-rectangular):

$$\boldsymbol{s} \in \{1, 2, \dots, n_1\} \times \{1, 2, \dots, n_2\} \times \{1, 2, \dots, n_d\}$$

Define $n_{max} = \max\{n_1, \dots, n_d\}$ is the maximum size of any dimension.

Theorem 4: Let $\beta = C_\beta \frac{n}{n_{max}} \log n$ and $\lambda = C_\lambda \frac{1}{n_{max}} \log n$. There exists constant $c_\gamma > 0$, such that

$$\widehat{m} = m^* \quad \text{and} \quad H\left(\widehat{R}_i, R_i^*\right) \leq \frac{C_\epsilon \sigma^2}{\Delta^2} \frac{n}{n_{max}} \log n, \;\; i = 1, \dots, m$$

holds with prob. at least $1 - 2\exp\left(-c_\gamma \frac{n}{n_{max}} \log n\right)$

Remark 1: When $d = 1, n_{max} = n$ and $d = 2, n_{max} = \sqrt{n}$, we have the results for timeline and 2D cases.

Remark 2: when $n_{max} = n^{1/d}$ (hyper-cube), the localisation rate is $n^{\frac{d-1}{d}} \log n$.

3. Numerical studies

Computational challenge

We need to solve the following optimisation under smoothness constraints

$$\min_{m, R_1, \ldots, R_m \in \mathcal{R}} L(R_0) + \sum_{j=1}^{m} L(R_j) + \beta m + \lambda \sum_{j=1}^{m} \left| \mathrm{Co}(R_j) \right|,$$

$\Rightarrow$ highly non-convex and complex (NP hard).

Computational challenge

We need to solve the following optimisation under smoothness constraints

$$\min_{m, R_1, \ldots, R_m \in \mathcal{R}} L(R_0) + \sum_{j=1}^{m} L(R_j) + \beta m + \lambda \sum_{j=1}^{m} \left| \mathrm{Co}(R_j) \right|,$$

$\Rightarrow$ highly non-convex and complex (NP hard).

Dynamic programming inspired by 1-D k-mean clustering algorithm:

- w.l.o.g. suppose $\mu_0^* = 0$ and all the $\mu_j^* > 0$
- Sort the data in ascending order
- Divide the ordered data into consecutive segments
- Recursively update the optimal cost matrix $C(i, j)$, $i = 1, \ldots, n$ (number of data) and $j = 1, \ldots, m$ (number of segments). For calculation of $C(i, j)$, we just need to go back to $C(i - k, t)$ for some fixed k and $1 \leq t \leq j$.

Algorithm: CRS

Consider a rearrangement of $\left\{Y(\boldsymbol{s})-\mu_0\right\}_{\boldsymbol{s}\in\mathcal{S}}$, with $|Y_1|\geq|Y_2|\geq ...\geq|Y_n|$, as a rearrangement of $\left\{Y(\boldsymbol{s})-\mu_0\right\}_{\boldsymbol{s}\in\mathcal{S}}$. Solve the following problem to find baseline cut-off:

$$\min_{1\leq N\leq n}\left[L\left(R_0^N\right)+\min_m\left\{\min_{R_{1:m}^N}\left(\sum_{j=1}^m L\left(R_j^N\right)+\lambda\sum_{j=1}^m\left|\text{Co}\left(R_j^N\right)\right|\right)+\beta m\right\}\right],$$

For fixed N, we propose the following Circular Region Segmentation:

Algorithm 1 Circular Region Segmentation (CRS)

Require: $(Y_{1:N},\, m,\, \xi)$, $\mathcal{N}=\{1,...,N\}$

1: **for** $k=1,\dots,m$ **do**
2: Pick $i=\min\mathcal{N}$;
3: Calculate $\widetilde{R}_k^N=\mathcal{S}_{\mathcal{N}}\cap\mathcal{B}\left(\boldsymbol{s}_i,\sqrt{\frac{n}{m\pi}}\right)$
4: Update $\mathcal{N}=\mathcal{N}\setminus\{j:\boldsymbol{s}_j\in\widetilde{R}_k^N\}$
5: **if** $|\widetilde{R}_k^N|\geq\xi$ **then**
6: $k=k+1$

Ensure: $\widetilde{R}_{1:m}^N$

Algorithm (PLS-SAD)

Algorithm 2 Penalised Least Square for Spatial Anomaly Detection (PLS-SAD)

Require: $(Y_{1:n}, \beta, \lambda, \xi)$,

for $N = 1, ..., n$ **do**

  for $m = 1, ..., N$ **do**

    $R^N_{1:m} = \mathbf{CRS}(Y_{1:N}, m, \xi)$;

    Calculate

      $R^N_0 = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\} \setminus \cup_{j=1}^m R^N_j$

      $C(m, N) = L(R^N_0) + \sum_{j=i}^m L(R^N_j) + \lambda \sum_{j=1}^m \left|\mathrm{Co}(R^N_j)\right| + \beta m$

$(\tilde{m}, \tilde{N}) = \arg\min_{1\leq m\leq n, 1\leq N\leq n} C(m, N)$

Ensure: $(\tilde{m}, \widetilde{R^{\tilde{N}}_{1:\tilde{m}}})$

It provide an approximate solution to (2) with computational cost $O(n^4)$.

Simulation: 2-D experiments

We construct the following 3 different anomaly settings:

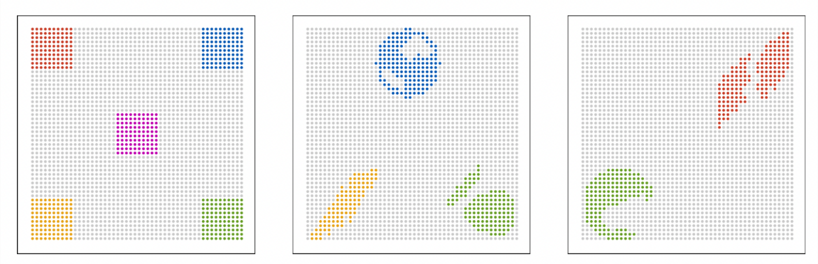


Figure 5: Setting 1-3 for illustration. We examined different sample size $n \in \{400, 1225, 2500\}$ and different scale of anomaly regions.

Define the symmetric error for two sets of anomalies:

$$\text{Error}\left(R^*_{1:m}, \widehat{R}_{1:\widehat{m}}\right) = \frac{\text{Error}\left(R^*_{1:m} \setminus \widehat{R}_{1:\widehat{m}}\right) + \text{Error}\left(\widehat{R}_{1:\widehat{m}} \setminus R^*_{1:m}\right)}{\sum_{j=1}^{m^*} \left|R^*_j\right|}$$

Simulation: 2-D i.i.d data

Detection performance with different δ and Δ under three different settings

		$\Delta = 1$		$\Delta = 2$		$\Delta = 3$	
		N	$\text{Error}(R^*, \widehat{R})$	N	$\text{Error}(R^*, \widehat{R})$	N	$\text{Error}(R^*, \widehat{R})$
Setting 1	$\lvert R^*_{1:5}\rvert = 9$	13	105.36%	45	34.36%	89	11.56%
	$\lvert R^*_{1:5}\rvert = 16$	43	63.30%	83	18.43%	98	8.35%
	$\lvert R^*_{1:5}\rvert = 25$	73	46.52%	92	16.59%	99	8.08%
Setting 2	$\lvert R^*_{1:3}\rvert = 9,9,11$	18	180.93%	34	67.21%	83	23.31%
	$\lvert R^*_{1:3}\rvert = 14,14,15$	22	122.50%	63	40.63%	95	14.14%
	$\lvert R^*_{1:3}\rvert = 18,19,23$	47	88.43%	78	27.53%	96	13.43%
Setting 3	$\lvert R^*_{1:2}\rvert = 8,10$	9	387.72%	14	243.78%	22	138.33%
	$\lvert R^*_{1:2}\rvert = 13,13$	14	276.23%	31	145.65%	54	65.62%
	$\lvert R^*_{1:2}\rvert = 19,23$	25	177.56%	54	90.29%	87	38.20%

Figure 6: Result over 100 Monte Carlo simulations with $n = 400$. N is the number of times we correctly estimate m^*.

Simulation: 2-D i.i.d data

Heatmap of Setting 1 under different combinations of δ and Δ.

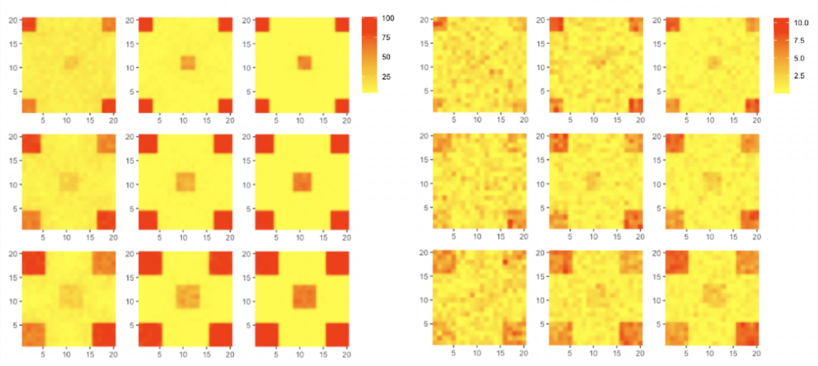


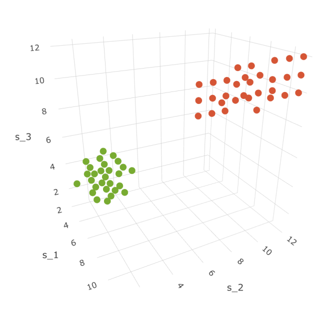
Figure 7: Left: correct detection over 100 Monte Carlo simulations. Right: data (with noise) in one realisations

2-D dependent data: The random error $\varepsilon(\boldsymbol{s})$ have covariance $\exp(-\varphi \cdot \mathrm{dist}(\boldsymbol{s}, \boldsymbol{s}'))$.
Sample size $n = 2500$, $|R^*_{1:3}| = 71, 72, 85$.

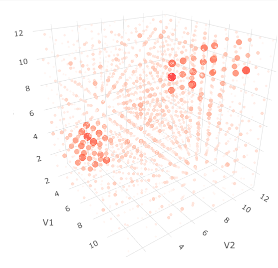
	$\Delta = 1$		$\Delta = 2$		$\Delta = 3$	
	N	$\mathrm{Error}(R^*, \widehat{R})$	N	$\mathrm{Error}(R^*, \widehat{R})$	N	$\mathrm{Error}(R^*, \widehat{R})$
$\varphi = 0$	17	74.60%	32	31.87%	42	17.14%
$\varphi = 0.5$	29	74.18%	54	31.58%	91	17.01%
$\varphi = 3$	38	73.82%	81	31.34%	100	16.85%

Simulation: 3-D i.i.d data

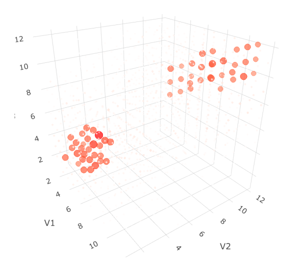
Sample size $n = 1728\,(12 \times 12 \times 12)$, $|R^*_{1:2}| = 28, 31$.



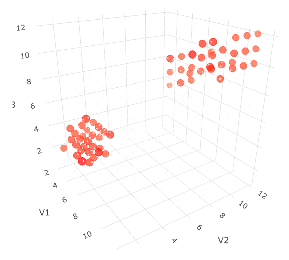
$\Delta = 1$		$\Delta = 2$		$\Delta = 3$	
N	$\mathrm{Error}(R^*, \widehat{R})$	N	$\mathrm{Error}(R^*, \widehat{R})$	N	$\mathrm{Error}(R^*, \widehat{R})$
24	136.78%	75	85.35%	97	40.52%



(a) $\Delta = 1$



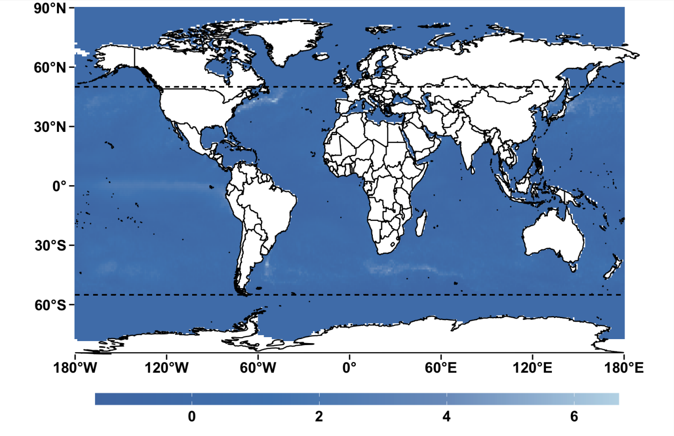
(b) $\Delta = 2$



(c) $\Delta = 3$

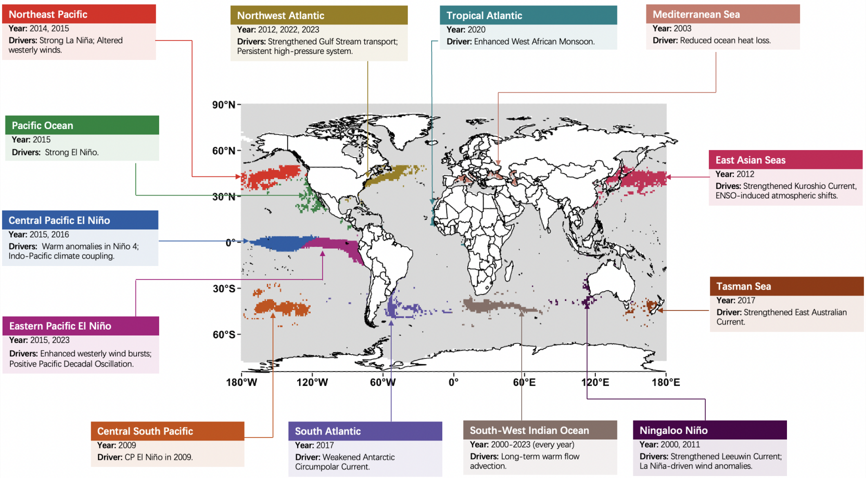
Figure 8: Result of estimated anomaly regions under different values of Δ.

Marine Heatwaves (MHVs) and Sea Surface Temperature (SST)



- Level-4 SST data from ESA CCI Programme
- Daily mean SST since 2000 with grid resolution of 0.5° (720×360)
- Linear yearly detrending, then taking maximum monthly average

Detecting MHVs with PLS-SAD



Summary and Discussions

Summary of the project

- Formulate the problem of SAD
- Give a valid penalised cost with both penalty on the number and intrinsic area of anomaly regions (via minimum convex hull)
- Establish the consistency theory and prove the detection rate is minimax optimal
- Propose a dynamic programming algorithm for computation

Summary of the project

- Formulate the problem of SAD
- Give a valid penalised cost with both penalty on the number and intrinsic area of anomaly regions (via minimum convex hull)
- Establish the consistency theory and prove the detection rate is minimax optimal
- Propose a dynamic programming algorithm for computation

Follow-up works:

- Flexible spatially dependent data — ongoing
- Application to oceanographic sciences — ongoing
- Faster/better algorithm – to be done

Summary of the project

- Formulate the problem of SAD
- Give a valid penalised cost with both penalty on the number and intrinsic area of anomaly regions (via minimum convex hull)
- Establish the consistency theory and prove the detection rate is minimax optimal
- Propose a dynamic programming algorithm for computation

Follow-up works:

- Flexible spatially dependent data — ongoing
- Application to oceanographic sciences — ongoing
- Faster/better algorithm – to be done

Future works:

- All kinds of spatial anomalies/changepoints