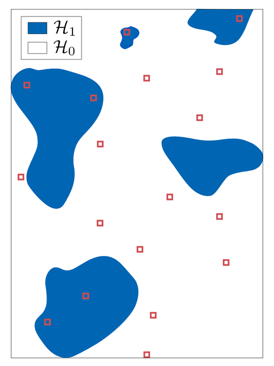
Improving Inference for Spatial Signals by Contextual False Discovery Rates

Martin Gölz[†], Abdelhak M. Zoubir[†] and Visa Koivunen[*]

Motivation



- Spatial signal/phenomenon: physical process that varies smoothly as a function of location, occur in
  - e.g. RADAR, wireless, meteorology, environmental monitoring, ...
- Monitored by large-scale sensor networks (IoT)
  - congested wireless spectrum, battery powered
- Fundamental problem: Detection of spatial regions associated with interesting, different or anomalous behavior under strict error control
- Previously [1] proposed: multiple hypothesis testing (MHT) approach with false discovery rate (FDR) control

Contribution

- Use contextual lfdr's (clfdr's) to incorporate a spatially varying empirical Bayes prior into the MHT approach from [1]
  $\rightarrow$ significant improve in detection power
- Two methods to estimate the empirical Bayes prior from the data
  1. sensor lfdr smoothing (SLS)
  2. screened null sensor smoothing (SNS)
- A generally applicable criterion for prior selection

System Model

- Discrete grid of Q points, sensors placed at $N \leq Q$ grid points
- $H_q = \mathrm{H}_0$ ($H_q = \mathrm{H}_1$): signal in nominal (any deviating) state at $q \in [Q]$
  - $\mathcal{H}_0 = \{q \in [Q] : H_q = \mathrm{H}_0\}$: The null region
  - $\mathcal{H}_1 = \{q \in [Q] : H_q = \mathrm{H}_1\}$: The alternative region

$$\mathrm{FDR} = \mathrm{E}\left[\frac{\text{\#false positives}}{\text{\#positivs}}\right] = \mathrm{E}\left[\frac{\sum_{q \in \mathcal{H}_0} \mathbb{1}\{\hat{H}_q = \mathrm{H}_1\}}{\max\left(\sum_{q=1}^{Q} \mathbb{1}\{\hat{H}_q = \mathrm{H}_1\}, 1\right)}\right]$$

- $\mathcal{P}^N = \{p_1, \dots, p_N\}$: set of sensor p-values at fusion center (FC), realizations of random variable P
- $P \sim f_P(p) = \pi_0 \cdot f_{P|\mathrm{H}_0}(p) + (1 - \pi_0) \cdot f_{P|\mathrm{H}_1}(p)$
  - π_0 : fraction of sensors where H_0 holds
  - $f_{P|\mathrm{H}_h}(p)$: PDF of p-values from sensors $n \in \mathcal{H}_X, h \in [1, 2]$

Spatial Inference with contextual local false discovery rates

At the sensors

- local observations
- computation of soft decision statistics

At the fusion center

- estimation of clfdr's at sensor locations
- clfdr interpolation in between sensors
- find region $\hat{\mathcal{H}}_1$ by grid point decisions

- Contextual local false discovery rates (clfdr's) at sensors $n \in [N]$

$$\mathrm{clfdr}_n(p) = \frac{\pi_{n,0} \cdot f_{P|\mathrm{H}_0}(p)}{\pi_{n,0} \cdot f_{P|\mathrm{H}_0}(p) + (1 - \pi_{n,0}) \cdot f_{P|\mathrm{H}_1}(p)}$$

✓ $f_{P|\mathrm{H}_0}(p) = \mathcal{U}[0, 1]$: known
✓ $f_{P|\mathrm{H}_1}(p)$: estimate using `lfdr-sMoM` [1]
! $\pi_{n,0}$: the local prior of $H_n = \mathrm{H}_0$

- Estimate $\mathrm{clfdr}_q, q \in [Q] \setminus [N]$ by radial basis function interpolation with thin plate splines
- Estimate regions associated with H_1 s.t. $\mathrm{FDR} \leq \alpha$ by

$$\hat{\mathcal{H}}_1 = \underset{\mathcal{H} \subseteq [N]}{\mathrm{argmax}} \left\{ |\mathcal{H}| \middle| \frac{1}{|\mathcal{H}|} \cdot \sum_{q \in \mathcal{H}} \mathrm{clfdr}_q \leq \alpha \right\}$$

Learning the empirical Bayes local null prior

1. For $K(\cdot) \in \mathcal{K}$
   2. For $b \in \mathcal{B}$
      3. Compute $\pi_{n,0}(K(\cdot); b) \forall n \in [N]$ using
         - Eq. (1): sensor lfdr smoothing (SLS) or
         - Eq. (2): screened null sensor smoothing (SNS)
4. Determine $(K^*(\cdot), b^*) = \theta^* = \underset{\theta = (K(\cdot) \in \mathcal{K}, b \in \mathcal{B})}{\mathrm{argmax}} \; c(\theta)$

$$\hat{\pi}_{n,0}^{\mathrm{SLS}}(K(\cdot); b) = \frac{\sum_{m=1, m \neq n}^{N} K(d_{n,m}; b) \cdot \hat{\mathrm{lfdr}}(p_m)}{\sum_{m=1, m \neq n}^{N} K(d_{n,m}; b)} \quad (1)$$

$$\hat{\pi}_{n,0}^{\mathrm{SNS}}(K(\cdot); b) = \frac{\sum_{m \in \mathcal{N}_S, m \neq n} K(d_{n,m}; b)}{(1 - \tau) \sum_{m=1, m \neq n}^{N} K(d_{n,m}; b)} \quad (2)$$

- $K(\cdot), \mathcal{K}$: kernel function, set of candidates $\rightarrow$ found automatically
- $b, \mathcal{B}$: bandwidth parameter, grid $\rightarrow$ found automatically
- $d_{m,n}$: Euclidean distance between sensors $n, m \in [N]$
- $\mathcal{N}_S = \{n \in [N] : p_n \geq \tau\}$: screened sensor set with threshold τ
  - contains only sensors where $H_n = \mathrm{H}_0$ very likely
- $c(\theta) = \sum_{n=1}^{N} w_n l_n(\hat{\pi}_{n,0}(\theta))$
  - w_n: pre-defined weights, $\sum_{n=1}^{N} w_n = 1$
  - $l_n(\hat{\pi}_{n,0}(\theta))$: likelihood function for sensor $n \in [N]$

Simulation results

- Identification of areas with occupied radio frequency spectrum
  - ScA: 2 sources, suburban line-of-sight (LOS) environment, low transmission power
  - ScB: 8 sources, suburban LOS environment, low transmission power
  - ScC: 1 source, urban non-LOS (NLOS) environment, high transmission power
- Observation area: 100×100 grid points
- Performance measures averaged over 200 independent repetitions
- Methods: FDRS (from the literature), `lfdr-sMoM` (previously proposed), **`clfdr-sMoM-SNS`**, **`clfdr-sMoM-SLS`** (both proposed in this work)

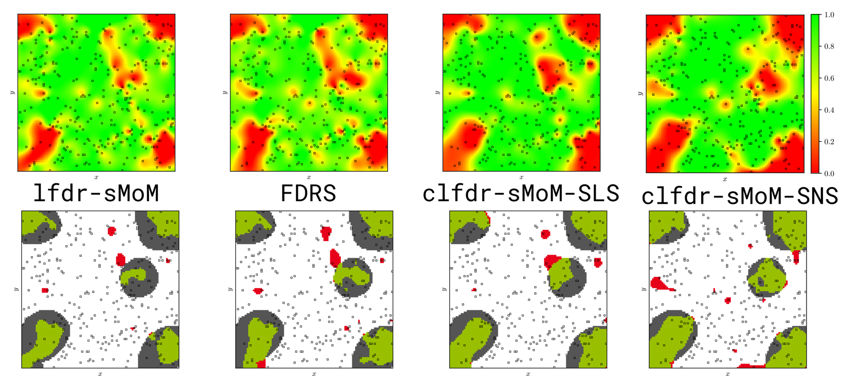
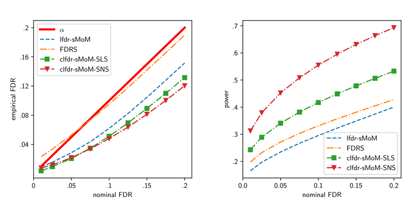


Fig. 1: ScB with $N = 300$ sensors. Estimated (c)lfdr's (top) detection patterns with $\alpha = 0.1$ (bottom). Green, red, gray: true, false, missed discoveries.



empirical vs. nominal FDR — power vs. nominal FDR

Fig. 2: Sc. C with $N = 1000$ sensors. The FDR is controlled at all levels except for FDRS at small α. The proposed **`clfdr-sMoM-SLS`** and **`clfdr-sMoM-SNS`** provide the largest detection power.

Table 1: $\alpha = 0.1$. Strict FDR control: `lfdr-sMoM`, **`clfdr-sMoM-SLS`**. **`clfdr-sMoM-SNS`**: problems in ScC, N very small, but overall provides highest power. FDRS breaks down in ScA

		$N = 300$		$N = 1000$		$N = 3000$	
		FDR	Power	FDR	Power	FDR	Power
ScA	`lfdr-sMoM`	.039	.117	.023	.138	.028	.163
	FDRS	.415	.239	.324	.294	.264	.377
	`clfdr-sMoM-SLS`	.031	.211	.022	.260	.026	.283
	`clfdr-sMoM-SNS`	.109	<u>.224</u>	.034	<u>.320</u>	.013	<u>.404</u>
ScB	`lfdr-sMoM`	.070	.263	.050	.286	.052	.302
	FDRS	.082	.257	.057	.312	.047	.378
	`clfdr-sMoM-SLS`	.052	<u>.375</u>	.038	.410	.038	.431
	`clfdr-sMoM-SNS`	.156	.494	.068	<u>.560</u>	.030	<u>.604</u>
ScC	`lfdr-sMoM`	.079	.283	.062	.295	.060	.297
	FDRS	.149	.303	.095	.331	.074	.388
	`clfdr-sMoM-SLS`	.063	.406	.051	.417	.055	.418
	`clfdr-sMoM-SNS`	.088	<u>.518</u>	.048	<u>.555</u>	.028	<u>.577</u>

Conclusion

- Exploitation of spatial smoothness by spatially varying empirical Bayes prior considerably increases detection power while FDR is controlled
  - strict control for any network size with `clfdr-sMoM-SNS`
  - largest power gain with `clfdr-sMoM-SLS`
- Empirical Bayes prior can be learned from the data in autonomous fashion using one of the proposed methods $\rightarrow$ no parameter tuning required
- Future research directions: robustification against wrongly reporting sensors, sequentially arriving local summary statistics, ...

[†] Signal Processing Group, Technische Universität Darmstadt, Germany
[*] Department of Signal Processing and Acoustics, Aalto University, Finland

The work of M. Gölz was supported by the German Research Foundation (DFG) under grant ZO 215/17-2.

Reference: [1] M. Gölz, A.M. Zoubir and V. Koivunen, "Multiple Hypothesis Testing Framework for Spatial Signals". Submitted to IEEE Trans. Signal Inf. Process. Netw., preprint available online. arXiv: 2108.12314.