Rate Analysis for Detection of Sparse Mixtures

J. G. Ligo[1], G.V. Moustakides[2] and V. V. Veeravalli[1]

[1]ECE and CSL, University of Illinois at Urbana-Champaign
[2]University of Patras and Rutgers University – New Brunswick

Detecting Sparse Mixtures

- Test between pure noise and sparse signal in noise
- Sparse signal in noise modeled as mixture between noise and signal PDF
- Study trade-off between signal strength, sparsity and sample size
- Applications: Sensor Networks, Disease Outbreak Monitoring, Astrophysics, Bioinformatics, Covert Signaling [2,8-11]
- Initially studied with unit variance Gaussian noise and signal
- Three Questions:
    1. When are there consistent tests?
    2. What are the best rates for consistent tests?
    3. Are there adaptive tests (i.e. unknown signal and noise) with best rate?

Mathematical Model

- Test between:

$$H_{0,n}: X_1, \dots X_n \sim f_{0,n}(x)$$
$$H_{1,n}: X_1, \dots X_n \sim (1-\epsilon_n) f_{0,n}(x) + \epsilon_n f_{1,n}(x)$$

- $\{f_{0,n}(x)\}, \{f_{1,n}(x)\}$ sequence of PDFs; $L_n(x) = \frac{f_{1,n}(x)}{f_{0,n}(x)}$
- $\epsilon_n \to 0, n\,\epsilon_n \to \infty$
- $\mathrm{LLR}(n) = \sum_{i=1}^{n} \log\left(1 - \epsilon_n + \epsilon_n L_n(x_i)\right)$
- Analyze rate and consistency of oracle LRT:
- $\delta_n(x_1, \dots, x_n) = \begin{cases} 1 & \mathrm{LLR}(n) \geq 0 \\ 0 & \mathrm{LLR}(n) < 0 \end{cases}$
- False Alarm: $\mathrm{P}_{\mathrm{FA}}(n) = \mathrm{P}_0[\delta_n = 1]$
- Miss Detection: $\mathrm{P}_{\mathrm{MD}}(n) = \mathrm{P}_1[\delta_n = 0]$
- Consistency: $\mathrm{P}_{\mathrm{FA}}(n), \mathrm{P}_{\mathrm{MD}}(n) \to 0$ as $n \to \infty$
- Prior work [2,3,4,5,6] focuses on consistency and adaptivity, not rate

Acknowledgements

This work was supported in part by the US National Science Foundation under the grants CCF 1514245, DMS 12-22498, through the University of Illinois at Urbana-Champaign, and under the Grant CCF 1513373, through Rutgers University.

Rate Analysis For the LRT

- Rates of consistency: How quickly do error probabilities tend to zero?
- Rate Characterization: $\lim_{n\to\infty} \frac{\log \mathrm{P}_{\mathrm{FA}}(n)}{g_0(n)} = -c.\ \lim_{n\to\infty} \frac{\log \mathrm{P}_{\mathrm{MD}}(n)}{g_1(n)} = -d$

  where $c, d > 0$ and $g_0(n), g_1(n) \to \infty$
- Classic i.i.d. Hypothesis Testing: $g_0(n) = g_1(n) = n$, error exponents (KL divergence); Sublinear $g_0(n), g_1(n)$ in this work
- Rate Characterization for LRT:

Weak Signals: Assume that for all $0 < \gamma < \gamma_0$ where $\gamma_0 \in (0,1)$:

1. $\lim_{n\to\infty} \mathrm{E}_0\left[\frac{(L_n-1)^2}{D_n^2}; L_n \geq 1 + \frac{\gamma}{\epsilon_n}\right] = 0$
2. $\epsilon_n D_n \to 0, n\,\epsilon_n^2 D_n^2 \to \infty$

Where $D_n^2 = \mathrm{E}_0[(L_n - 1)^2] < \infty$ is the χ^2-divergence between $f_{0,n}, f_{1,n}$.

Then, $\lim_{n\to\infty} \frac{\log \mathrm{P}_{\mathrm{FA}}(n)}{n\epsilon_n^2 D_n^2} = \lim_{n\to\infty} \frac{\log \mathrm{P}_{\mathrm{MD}}(n)}{n\epsilon_n^2 D_n^2} = -\frac{1}{8}$

Strong Signals: If for all M sufficiently large, $\lim_{n\to\infty} \mathrm{E}_0\left[L_n; L_n \geq 1 + \frac{M}{\epsilon_n}\right] = 1$,

then $\lim_{n\to\infty} \frac{\log \mathrm{P}_{\mathrm{FA}}(n)}{n\epsilon_n} \leq -1,\ \lim_{n\to\infty} \frac{\log \mathrm{P}_{\mathrm{MD}}(n)}{n\epsilon_n} = -1$

- Weak signals characterized by χ^2-divergence
    - Proof: Chernoff (UB), Mod. Cramer's Theorem w. n-dependent tilting (LB)
- Strong signal rate is independent of divergence (beyond condition)
    - Proof: Chernoff (UB), Universal $\liminf_{n\to\infty} \frac{\log \mathrm{P}_{\mathrm{MD}}(n)}{n\epsilon_n} \geq -1$ (typicality LB)

Gaussian Location Model (GLM)

- $f_{0,n} = \mathcal{N}(0,1),\ f_{1,n} = \mathcal{N}(\mu_n, 1)$ and $\epsilon_n = n^{-\beta}, \beta \in (0,1)$
- $\{(\epsilon_n, \mu_n)\}$ conditions for consistency [2,3,4,5]
- Adaptive testing (without rate guarantees) [2,3,4,5,6] (Different tests have very different power [4,7])

Conditions for consistency [2,3,4,5]:

1. (Dense) If $0 < \beta < \frac{1}{2}, \mu_{crit,n} = n^{\beta - \frac{1}{2}}$
2. (Moderately Sparse) If $\frac{1}{2} < \beta < \frac{3}{4}, \mu_{crit,n} = \sqrt{2(\beta - \frac{1}{2}) \log n}$
3. (Very Sparse) If $\frac{3}{4} < \beta < 1, \mu_{crit,n} = \sqrt{2\left(1 - \sqrt{1-\beta}\right)^2 \log n}$

If $\mu_n < \mu_{crit,n}, P_{FA} + P_{MD} \to 1$ (Detection is impossible)

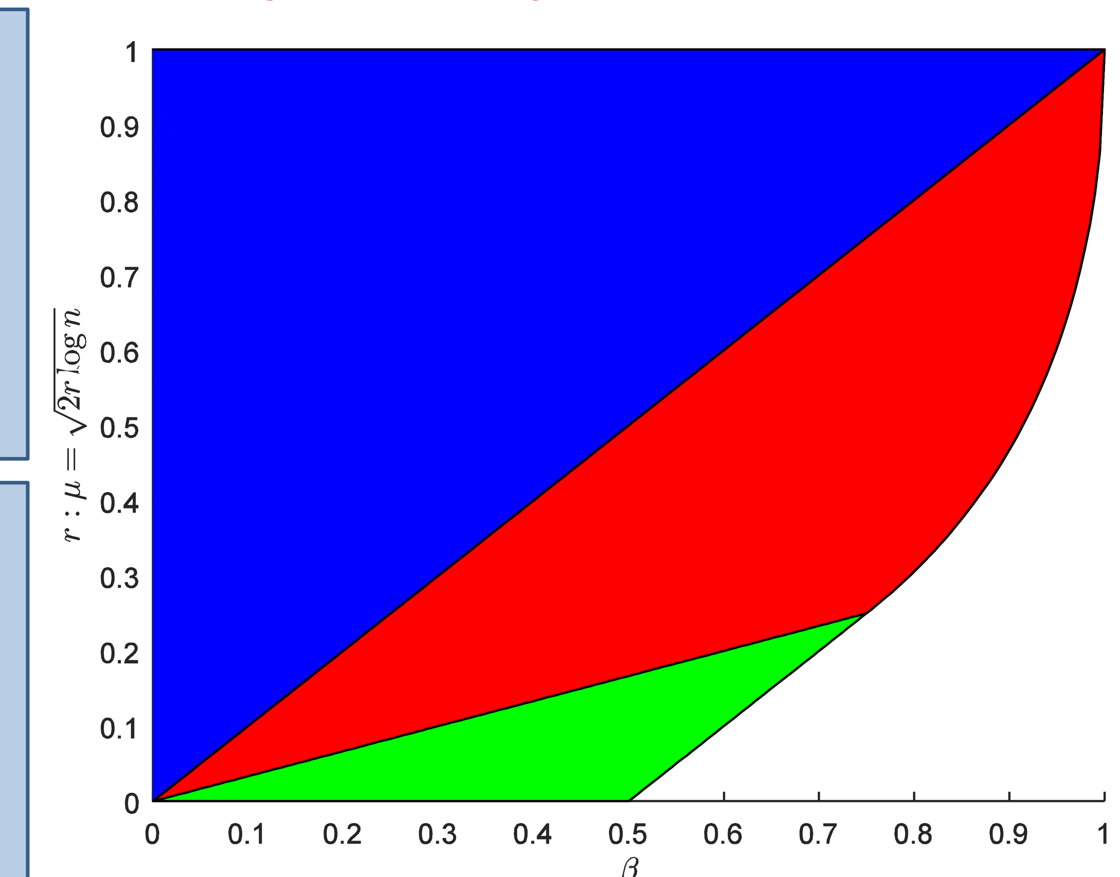
Rate Characterization for GLM:

Weak Signals (Green): If $\mu_{crit,n} < \mu_n < \sqrt{\frac{2\beta}{3} \log n}$,

$\lim_{n\to\infty} \frac{\log \mathrm{P}_{\mathrm{FA}}(n)}{n\epsilon_n^2 (e^{\mu_n^2} - 1)} = \lim_{n\to\infty} \frac{\log \mathrm{P}_{\mathrm{MD}}(n)}{n\epsilon_n^2 (e^{\mu_n^2} - 1)} = -\frac{1}{8}$

Strong Signals (Blue): If $\mu_n > \sqrt{2\beta \log n}$, $\lim_{n\to\infty} \frac{\log \mathrm{P}_{\mathrm{MD}}(n)}{n\epsilon_n} = -1.$

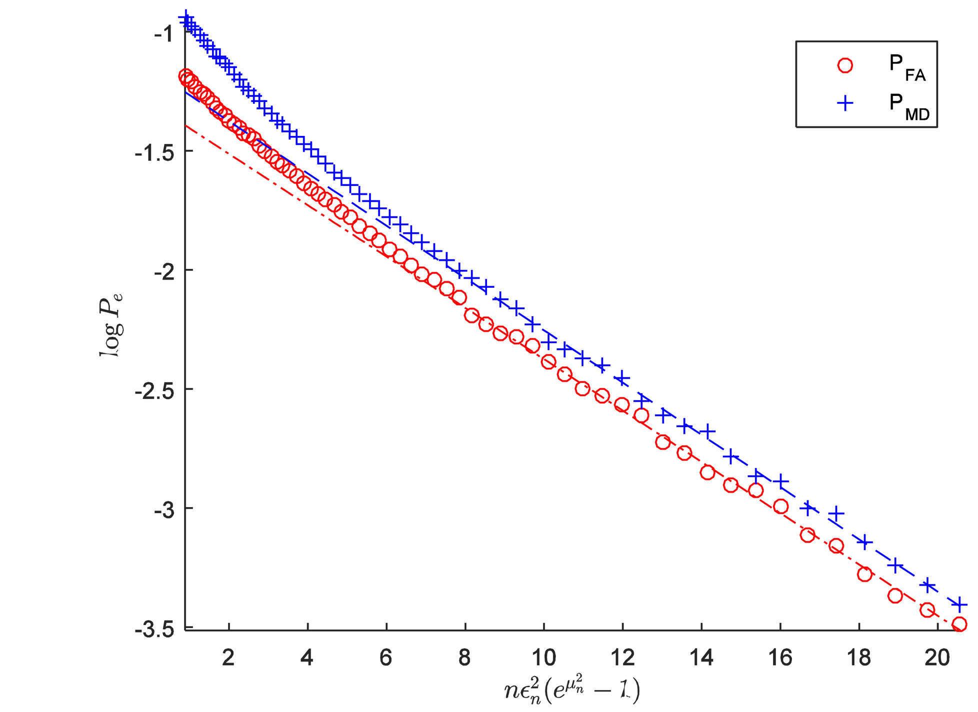
Furthermore, if $\frac{n\epsilon_n}{\mu_n^2} \to \infty$, $\lim_{n\to\infty} \frac{\log \mathrm{P}_{\mathrm{FA}}(n)}{n\epsilon_n} = -1$



- LB for $\mathrm{P}_{\mathrm{FA}}(\mathrm{n})$ via LB on $\mathrm{LLR}(n)$ + Modified Cramer's Theorem

Numerical Results (GLM)

- $\epsilon_n = n^{-0.6}, \mu_n = \sqrt{2(0.19)\log n}$ (Sparse, Weak)
- Best fit slope with $n \geq 100000$: -0.108 (FA, MD)
- Theoretical prediction: $-\frac{1}{8}$



Conclusions and Future Work

- Analyzed rates for oracle LRT for general sparse mixtures
- Sublinear rate with non-KL divergence
- Future Work: Design adaptive tests with oracle rate for GLM and other mixture models

References

1. J.G. Ligo, G. V. Moustakides, V. V. Veeravalli, "Rate Analysis for Detection of Sparse Mixtures", ICASSP 2016 (Extended version: arXiv:1509.07566 [cs.IT])
2. D. Donoho, J. Jin, "Higher criticism for detecting sparse heterogeneous mixtures", Ann. Statist., vol. 32, no. 3, pp. 962–994, 06 2004.
3. T. T. Cai, X. J. Jeng, and J. Jin, "Optimal detection of heterogeneous and heteroscedastic mixtures," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 73, no. 5, pp. 629–662, 2011.
4. E. Arias-Castro and M. Wang, "Distribution-free tests for sparse heterogeneous mixtures," arXiv preprint arXiv:1308.0346 [math.ST], 2013.
5. Y. Ingster and I. A. Suslina, Nonparametric goodness-of-fit testing under Gaussian models. Springer Science & Business Media, 2003, vol. 169
6. T. T. Cai and Y. Wu, "Optimal detection of sparse mixtures against a given null distribution," IEEE Trans. Info. Theory, vol. 60, no. 4, pp. 2217–2232, 2014.
7. G. Walther, "The average likelihood ratio for large-scale multiple testing and detecting sparse mixtures," arXiv preprint arXiv:1111.0328 [stat.ME], 2011.
8. E. Mossel and S. Roch, "Distance-based species tree estimation: information-theoretic trade-off between number of loci and sequence length under the coalescent," arXiv preprint arXiv:1504.05289 [math.PR], 2015.
9. J. J. Goeman and P. Buhlmann, "Analyzing gene expression data in terms of gene sets: methodological issues," Bioinformatics, vol. 23, no. 8, pp. 980–987, 2007.
10. L. Cayon, J. Jin, and A. Treaster, "Higher criticism statistic: detecting and identifying non-gaussianity in the wmap first-year data," Monthly Notices of the Royal Astronomical Society, vol. 362, no. 3, pp. 826–832, 2005.
11. V. Saligrama and M. Zhao. "Local anomaly detection." International Conference on Artificial Intelligence and Statistics, pp. 969-983. 2012.