Hypothesis testing with low-degree polynomials in the Morris class of exponential families

Dmitriy Kunisky (New York University)

Motivation

How to assess the asymptotic **computational** difficulty of **hypothesis testing** in high dimension?

We want to know when we can distinguish with high probability between:

- $\boldsymbol{Y} \sim \mathbb{Q}_n$ (null, unstructured, symmetric)
- $\boldsymbol{Y} \sim \mathbb{P}_n$ (planted, containing a signal)

for $\mathbb{Q}_n, \mathbb{P}_n$ over $\mathbb{R}^{N=N(n)}$ for $n \to \infty$, N growing with n.

What are useful **summaries** of difficulty and how can we compute them?

Likelihood Ratio

Ignoring computational constraints, a good test should be:

$$\begin{aligned} &\text{maximize } \mathbb{E}_{\mathbb{P}_n} p(\boldsymbol{Y}) \\ &\text{subject to } \mathbb{E}_{\mathbb{Q}_n} p(\boldsymbol{Y})^2 \le 1 \end{aligned}$$

Rewriting in $L^2(\mathbb{Q}_n)$,

$$\begin{aligned} &\text{maximize } \left\langle p, \frac{d\mathbb{P}_n}{d\mathbb{Q}_n} \right\rangle \text{ in } L^2(\mathbb{Q}_n) \\ &\text{subject to } \|p\|^2 \le 1 \text{ in } L^2(\mathbb{Q}_n) \end{aligned}$$

whose optimizer is the (normalized) **likelihood ratio**:

$$p^\star(\boldsymbol{Y}) = \underbrace{\frac{d\mathbb{P}_n}{d\mathbb{Q}_n}(\boldsymbol{Y})}_{L_n(\boldsymbol{Y})} \Big/ \underbrace{\left\| \frac{d\mathbb{P}_n}{d\mathbb{Q}_n} \right\|}_{\text{objective value}}$$

Neyman-Pearson Lemma (1933)

Thresholded likelihood ratio is an optimal test for any fixed tolerance of Type I and II errors.

Le Cam's Second Moment Method (1960s)

If $\|L_n\|$ is bounded as $n \to \infty$, no test (efficient or not) can distinguish.

Low-Degree Method

To incorporate computational constraints, take {low-degree polynomials} $\approx$ {efficient algorithms}:

$$\begin{aligned} &\text{maximize } \mathbb{E}_{\mathbb{P}_n} p(\boldsymbol{Y}) \\ &\text{subject to } \mathbb{E}_{\mathbb{Q}_n} p(\boldsymbol{Y})^2 \le 1, \\ &\qquad\qquad p(\boldsymbol{Y}) \in \mathbb{R}[\boldsymbol{Y}]_{\le D} \end{aligned}$$

By the same token as above, optimizer is the (normalized) **low-degree likelihood ratio**

$$p^\star(\boldsymbol{Y}) = \underbrace{\mathcal{P}^{\le D}\frac{d\mathbb{P}_n}{d\mathbb{Q}_n}(\boldsymbol{Y})}_{L_n^{\le D}(\boldsymbol{Y})} \Big/ \underbrace{\left\| \mathcal{P}^{\le D}\frac{d\mathbb{P}_n}{d\mathbb{Q}_n} \right\|}_{\text{objective value}}$$

for $\mathcal{P}^{\le D}$ the orthogonal projector to $\mathbb{R}[\boldsymbol{Y}]_{\le D}$.

Low-Degree Conjecture (2017-2018)

If $\|L_n^{\le D}\|$ is bounded as $n \to \infty$, no test can distinguish in time $e^{\widetilde{O}(D(n))}$.

But when can we actually carry out this computation?

Morris Class of Exponential Families

Any probability measure ρ_0 on $\mathbb{R}$ generates a *natural exponential family (NEF)*,

$$d\rho_\theta(x) \propto e^{\theta x} d\rho_0(x).$$

We can also reparametrize by the mean: $\mu = \mu(\theta) = \mathbb{E}_{x \sim \rho_\theta}[x] \rightsquigarrow \widetilde{\rho}_\mu = \rho_\theta$.

The variance is a function of the mean,

$$V(\mu) = \mathsf{Var}_{x \sim \widetilde{\rho}_\mu}[x],$$

and **simple families have simple variance functions**, in particular low-degree polynomial variance functions. Morris (1982-1983) fully characterized the NEFs with **quadratic** variance function (NEF-QVFs).

Morris' Classification of NEF-QVFs (1982)

All natural exponential families with quadratic variance function are one of the following or an affine transformation thereof:

Name	$d\rho_0(x)$	Support	$V(\mu)$
Gaussian (variance $\sigma^2 > 0$)	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}x^2)dx$	$\mathbb{R}$	σ^2
Poisson	$\frac{1}{e}\frac{1}{x!}$	$\mathbb{Z}_{\ge 0}$	μ
Gamma (shape $\alpha > 0$)	$\frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x} dx$	$(0, +\infty)$	$\frac{1}{\alpha}\mu^2$
Binomial (m trials)	$\frac{1}{2^m}\binom{m}{x}$	$\{0, \dots, m\}$	$-\frac{1}{m}\mu^2 + \mu$
Negative Binomial (m successes)	$\frac{1}{2^{m+x}}\binom{x+m-1}{x}$	$\mathbb{Z}_{\ge 0}$	$\frac{1}{m}\mu^2 + \mu$
Hyperbolic Secant (shape $r > 0$)	$\frac{1}{2}\mathsf{sech}(\pi x/2)$ (e.g.)	$\mathbb{R}$	$\frac{1}{r}\mu^2 + r$

Computing $\|L_n^{\le D}\|$

Prior work: **Gaussian models**. Using orthogonal (Hermite) polynomials, can treat the setup:

- Under $\mathbb{Q}_n$, $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.
- Under $\mathbb{P}_n$, draw $\boldsymbol{x} \sim \mathcal{P}$ a *signal* and take $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{x}, \boldsymbol{I})$.

This yields an **overlap formula**:

$$\|L_n^{\le D}\|^2 = \mathop{\mathbb{E}}_{\boldsymbol{x}^1, \boldsymbol{x}^2} \exp^{\le D}(\langle \boldsymbol{x}^1, \boldsymbol{x}^2 \rangle)$$

Here $\boldsymbol{x}^i \sim \mathcal{P}$ are i.i.d. copies, and $\exp^{\le D}(t) = \sum_{k=0}^D t^k/k!$ is the truncated Taylor series. Such a formula is **very convenient to work with!** It reduces the high-dimensional expectation to a scalar one.

Theorem 1: Generalization to NEF-QVFs

Suppose $\mathbb{Q}_n$ and $\mathbb{P}_n$ are products of distributions in an NEF-QVF, with

- $\mathbb{Q}_n$ has coordinates with means μ_i, variances $\sigma_i^2 = V(\mu_i)$,
- $\mathbb{P}_n$ has coordinates with means x_i for $\boldsymbol{x} \sim \mathcal{P}$.

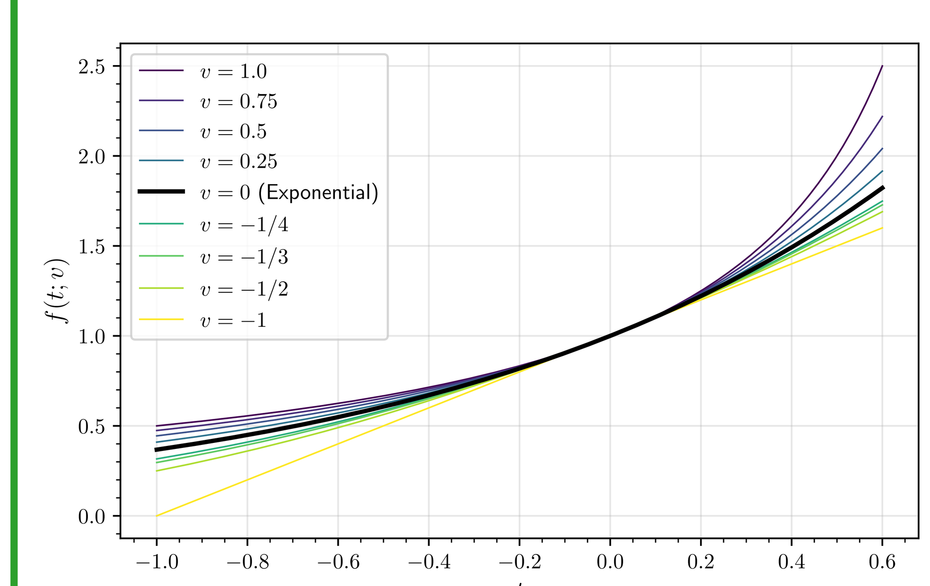
Suppose $V(\mu) = v_2\mu^2 + v_1\mu + v_0$. Then,

$$r(\boldsymbol{x}^1, \boldsymbol{x}^2) := \sum_{i=1}^N \frac{x_i^1 - \mu_i}{\sigma_i} \cdot \frac{x_i^2 - \mu_i}{\sigma_i}$$

$$f(t) := (1 - v_2 t)^{-1/v_2},$$

$$\|L_n^{\le D}\|^2 \le \mathop{\mathbb{E}}_{\boldsymbol{x}^1, \boldsymbol{x}^2} \left[f^{\le D}(r(\boldsymbol{x}^1, \boldsymbol{x}^2)) \right].$$

Examples of the $f(t) = f(t; v_2)$:



Channel Monotonicity

From the picture: $f(t; v_2)$ is monotone in v_2. In fact...

Theorem 2: Truncated Monotonicity

The truncated Taylor series $f^{\le D}(t; v_2)$ are monotone in v_2.

Recalling that, from the low-degree conjecture, an expectation of $f^{\le D}(\cdot; v_2)$ governs the cost of testing, we find **channel monotonicity**:

With a fixed distribution of z-scores of the planted model signal ($\boldsymbol{x} \sim \mathcal{P}$) with respect to the null model,

Bernoulli $\ge$ Binomial $\ge$ Gaussian $=$ Poisson $\ge$ Exponential

harder ⟷ easier

Exploration Before Inference

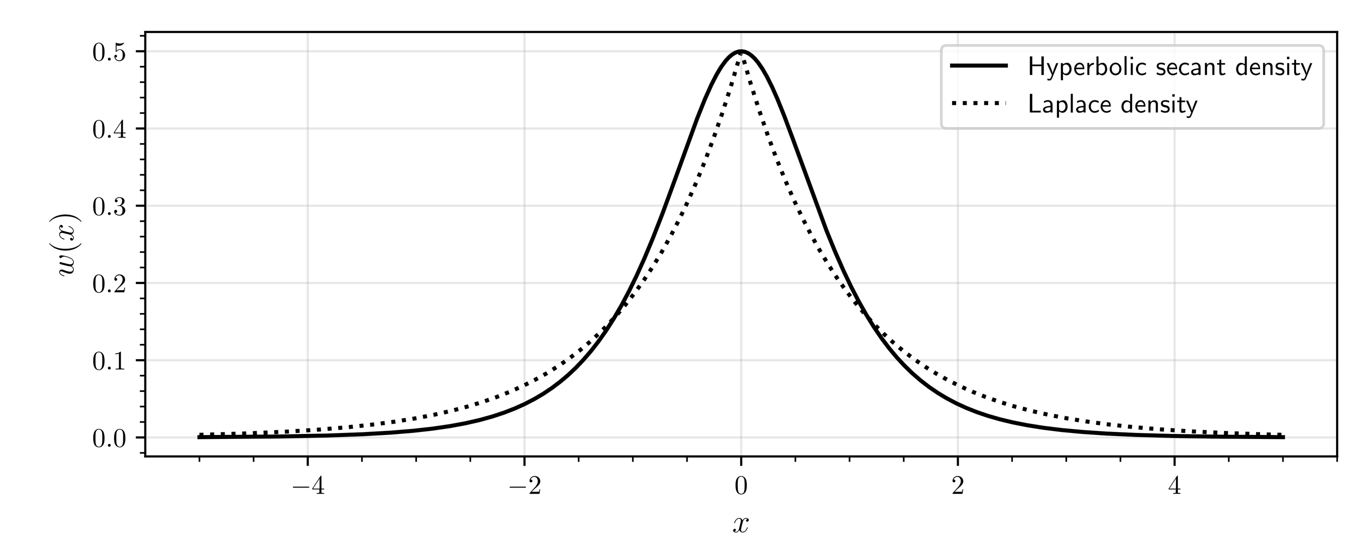
Apply the same tools to a **non-Gaussian spiked matrix model.** In Gaussian spiked matrix models, an optimal testing statistic is

$$\begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix} \rightsquigarrow \lambda_{\max}\left(\begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix} \right)$$

In general, we need F a transcendental function tailored to the noise law:

$$\begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix} \rightsquigarrow \lambda_{\max}\left(\begin{bmatrix} F(\cdot) & F(\cdot) & F(\cdot) \\ F(\cdot) & F(\cdot) & F(\cdot) \\ F(\cdot) & F(\cdot) & F(\cdot) \end{bmatrix} \right)$$

We consider the specific **hyperbolic secant** noise distribution:



Need high **entrywise degree** (approximating F) to test $\rightsquigarrow$ we can confuse low-degree polynomials by a **mixture** with heavy-tailed noise.

General prediction: Low-degree polynomials have trouble "exploring the data" (e.g., kernel density estimation, adaptive preprocessing, calibration).

Some References

Low-degree method and conjecture: *The power of sum-of-squares for detecting hidden structures* (Hopkins, Kothari, Potechin, Raghavendra, Schramm, Steurer 2017); *Efficient Bayesian estimation from few samples: community detection and related problems* (Hopkins, Steurer 2018); *Statistical inference and the sum-of-squares method* (Hopkins 2018).

NEF-QVFs: *Natural exponential families with quadratic variance functions* (Morris 1982).

Spiked matrix models: *Optimality and sub-optimality of PCA I: Spiked random matrix models* (Perry, Wein, Bandeira, Moitra 2018); *Adapting to unknown noise distribution in matrix denoising* (Montanari, Ruan, Yan 2018); *Estimating rank-one spikes from heavy-tailed noise via self-avoiding walks* (Ding, Hopkins, Steurer 2020).