Large-scale statistical inference: detecting sparse and weak effects

Tatjana Pavlenko, KTH Royal Institute of Technology
(joint research with Natalia Stepanova, Carleton University, Canada)

Invited presentation at the virtual workshop
"*Inference problems: algorithms and lower bounds*"
Goethe University Frankfurt
September 3, 2020

Outline

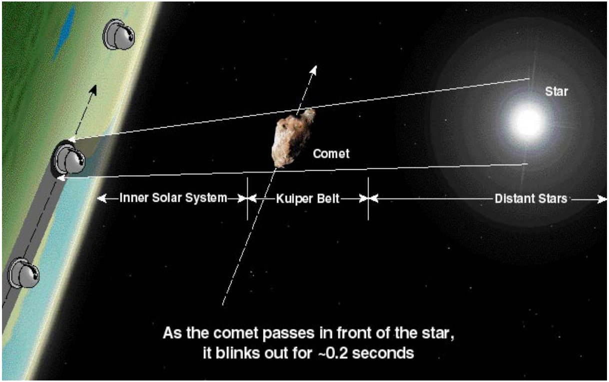
- Motivating examples: signal detection, large-scale testing problems for sparse mixtures and feature selection in high-dimensional classification.
- CsCsHM – the new family of statistics via EFKP upper-class weighted functions for detecting sparse and weak effects.
- Previous work: fundamental phase diagram and Tukey's "higher criticism"(HC) statistics. Motivation for improvements.
- Applications of CsCsHM to feature thresholding in classification.

Detection of sparse signals

Detection of sparse mixtures is an important statistical problem that arises

- in *Signal Processing*: the goal is to *detect* the existence of a signal which only appears in a small fraction of the noisy data
- in *Genomics and Genetics*: the goal of genome-wide association studies (GWAS) is to *identify* positions $j = 1, \ldots, p$ of single-nucleotide polymorphis (SNP) in the genome; a typical value of p is $\approx 10^6$ while sample size is in low thousands
- in *Astrophysics and Cosmology*: The Kuiper Belt containing small solar system bodies which date back to the formation of the solar system. They contain clues to conditions in the early solar nebula. TAOS (Taiwanese-American Occultation Survey generating 10^{11} tests per year where the number of occultations ranges from ten to few thousands. A small objects about 1 km diameter passes in front of a star and the star will *blink out*. Estimation of the rate of these occultation events is one of the primary science goals of TAOS.)

Occultation


Star
Comet
Inner Solar System
Kuiper Belt
Distant Stars
As the comet passes in front of the star,
it blinks out for ~0.2 seconds

Terminology for sparsity and weakness

What we will be talking about:

- Very high-dimensional observations
- High-dimensional underlying parameter vector
- Most parameter entries '*small*' or '*uninteresting*' for the project at hand. Interesting values 'scattered' about in parameter vector.
- Historically ...

What we will *not* be talking about:

- Sparsity of data (paucity low counts)
- Sparsity of matrices (e.g. numerical linear algebra)

A two-group model

- We are interested in making inference based on n units, each represented by a summary statistic X. The cases are either *null* or *non-null*, with non-null units exhibiting interesting patterns of abnormal behavior.
- We do not know the true states of nature but observe a *mixture* of null- and non-null cases.
- To model sparse data we posit a mixture model

$$X_1, \ldots, X_n \overset{iid}{\sim} (1 - \varepsilon_n)F_0 + \varepsilon_n F_1,$$

ε_n the mixing proportion is small, F_0 and F_1 are null and alternative distributions.

Motivating examples

Example 1: One of the earliest work on multi-channel detection in radio-location dates back to Dobrushin (1958).

Let $\{X_i\}_{i=1}^n \overset{iid}{\sim} Ray(\alpha_i)$, with the density $\frac{2x}{\alpha_i}\exp(-\frac{x^2}{\alpha_i})$, $x \geq 0$, representing the random voltages observed on n channels. In the absence of noise, all $\alpha_i = 1$ (nominal value); in the presence of signal, *exactly* one of the α_i's becomes a known value $\alpha > 1$. The goal is to test competing hypotheses

$$H_0: \quad \alpha_i = 1,\ 1 \leq i \leq n,$$

versus n-dependent alternative

$$H_{1,n}: \quad \alpha_i = 1 + (\alpha - 1)\mathbb{1}_{\{i=J\}},\ J \sim U([1,n]).$$

Since the signal only appears ones out of the n, in order to distinguish it from noise, it is necessary for the amplitude α to grow with the sample size, n (in fact, at least logarithmically).

Dobrushin proved log-likelihood ratio convergence to a stable distribution as $n \to \infty$ and obtained sharp asymptotics of the smallest α to achieve the desired false alarm and miss detection levels.

Motivating examples

Example 2: Two-component sparse Gaussian mixture, see e.g. Ingster (1997) and Donoho & Jin (2004) (also Ingster et al (2009), Meinshausen and Rice (2008) and Cai et al (2014))

$$H_0 : \quad X_1, \ldots, X_n \overset{iid}{\sim} N(0,1), vs$$
$$H_{1,n} : \quad X_i \overset{iid}{\sim} (1-\varepsilon_n)N(0,1) + \varepsilon_n N(\mu_n, 1), \quad 1 \leq i \leq n.$$

Example 3: Heteroscedastic two-component Gaussian mixture model. The signal itself varies among the non-null portion of the samples (see e.g. Cai and Jeng (2011)):

$$H_0 : \quad X_1, \ldots, X_n \overset{iid}{\sim} N(0,1), vs$$
$$H_{1,n} : \quad X_i \overset{iid}{\sim} (1-\varepsilon_n)N(0,1) + \varepsilon_n N(\mu_n, \sigma^2), \quad 1 \leq i \leq n.$$

Parametrization: linking sparsity and weakness to n.

$$\mu_n = \sqrt{2r \log n}, \quad \varepsilon_n = n^{-\beta}, \quad 0 < r, \beta < 1.$$

β is the *sparsity index*, usually $\beta \in (1/2, 1)$ and μ_n is the *signal strength* growing with the sample size, i.e $\mu_n \to \infty$ for $n \to \infty$.

Example 4: Beyond normality.

- Detection of *general* sparse mixture.

$$\begin{aligned} H_0: &\quad X_1, \dots, X_n \overset{iid}{\sim} Q_n, vs \\ H_{1,n}: &\quad X_i \overset{iid}{\sim} (1-\varepsilon_n) Q_n + \varepsilon_n G_n, \quad 1 \le i \le n, \end{aligned}$$

  where G_n models the statistical variation of non-null effects.

- Set-based analyses: Detection of significant sets of highly associated variables (gene-sets, functional genomic segments, SNPs aggregated by similar biological functions or genetic factors in GWAS that are sparsely distributed across the genome). See e.g Pavlenko et al (2012), Wu et al (2014), Meinshausen (2015), Bühlmann et al. (2015) and Sun et al. (2017).
- Connection to group-wise feature selection in high-dimensional classification. See Li et al (2015), Moscovich-Eiger et al. (2017), Pavlenko, Stepanova (2018) (ongoing project).

$$\begin{aligned} H_0: &\quad \mathcal{S}_i^2 \overset{iid}{\sim} \chi^2(\cdot, 0), \quad 1 \le i \le b, \text{ vs} \\ H_{1,b}: &\quad \mathcal{S}_i^2 \overset{iid}{\sim} (1-\varepsilon_b)\chi^2(\cdot, 0) + \varepsilon_b \chi^2(\cdot, \omega_i^2), \quad 1 \le i \le b. \end{aligned}$$

In general, we focus on the multiple testing problem where there are many independent null hypothesis H_{0i}, $i = 1, \dots, n$, and we are interested in rejecting the joint null $\cap_{i=1}^n H_{0i}$.

Three questions

in increasing level of difficulty:

Q1: Can we tell if at least one null hypothesis is false? (Is there any signals?)

Q2: What is the proportion of non-null hypotheses? (The theory of estimating ε_n)

Q3: Which null hypotheses are false? (Where is the signal?)

The main objectives of studying the sparse detection problem are two-fold:

- To determine the *detection boundary*, which gives the smallest possible signal strength $r^* = r^*(\beta)$ as a function of the sparsity parameter β, such that reliable detection is possible, in the sense that sum of Type-I and II error probabilities $\to 0$ as $n \to \infty$. The set $\{(r, \beta) : r > \rho(\beta)\}$ is known as the *detectable region*.
- To construct *optimally adaptive* testing procedures which achieve vanishing probability of error simultaneously for all values of (r, β) inside the detectable region, i.e. **without requiring the knowledge of sparsity level and size of the non-null effects**.

Fundamental limits and characterization

Q: *Where the departures from H_0 can be detected reliably?*

- The best possible theoretical detection boundary associated with the testing problem of Example 2 (Ingster (1997)) is

$$\rho(\beta) = \begin{cases} \beta - 1/2, & 1/2 < \beta \leq 3/4, \\ (1 - \sqrt{1-\beta})^2, & 3/4 < \beta < 1. \end{cases}$$

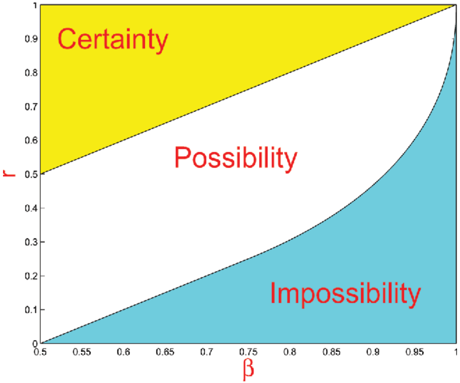
  Results are shown using Neyman-Pearson LRT with **precisely specified** set of parameters $\varepsilon_n = n^{-\beta}$ and $\mu_n = \sqrt{2r \log n}$ calibrated with a pair (r, β): $0 < r < 1$ and $\beta \in (1/2, 1)$.

- When (r, β) satisfy $r > \rho(\beta)$ the LRT has asymptotically full power of detection

$$P_{H_{1,n}} \{\text{Reject } H_0\} \to 1 \quad \text{as } n \to \infty.$$

- In the *detectable region* H_0 and $H_{1,n}$ separate asymptotically. In the interior of *undetectable region* H_0 and $H_{1,n}$ merge asymptotically.
- To construct adaptive optimal procedures, Ingster (2001) and (2002) considered generalized LRT tests over a growing discretized set of (r, β)-pairs and established its optimal asymptotic adaptivity.
- Region of undetectability explains many failures of reproducibility.

Separating boundary on $r - \beta$ plane: existing of phases



A new family of test statistics: initial results

- $X_1, X_2, \ldots$ iid with a continuous CDF F on $\mathbb{R}$, $\mathbb{F}_n(t) = n^{-1} \sum_{i=1}^{n} \mathbb{I}(X_i \leq t)$, $t \in \mathbb{R}$, is the EDF based on $X_1, \ldots, X_n$.

  We are interested in testing the hypothesis of *goodness-of-fit*

  $$H_0 : F = F_0 \quad \text{vs} \quad \text{either} \quad H_1 : F \neq F_0 \quad \text{or} \quad H_1' : F > F_0.$$

- For specific types of alternatives, certain classical goodness-of-fit tests, including the Kolmogorov-Smirnov and Anderson-Darling tests, may benefit significantly from using proper weights, such as *Erdős-Feller-Kolmogorov-Petrovski* (EFKP) upper-class function of a Brownian bridge.

The family of EFKP upper-class functions

Def: Let q be any strictly positive function on $(0,1)$ with the property $q(u)=q(1-u)$ for $u\in(0,1/2)$, which is nondecreasing in a neighborhood of zero and nonincreasing in a neighborhood of one. Such a function is called an *Erdős-Feller-Kolmogorov-Petrovski (EFKP) upper-class function* of a Brownian bridge $\{B(u), 0\le u\le 1\}$, if there exists a constant $0\le b<\infty$ such that

$$\limsup_{u\to 0}|B(u)|/q(u)\overset{\text{a.s.}}{=}b.$$

Def: An EFKP upper-class function q of a Brownian bridge is called a *Chibisov-O'Reilly function* if $b=0$ in (Csörgő et al (1986)). Example of EFKP

upper-class function with $0<b<\infty$

$$q(u)=\sqrt{u(1-u)\log\log(1/(u(1-u)))}.$$

Example of Chibisov-O'Reilly function

$$q(u)=(u(1-u))^{1/2-\nu},\quad 0<\nu<1/2.$$

A new family of test statistics

Back to testing $H_0 : F = F_0$ vs $H_1 : F \neq F_0$ or $H_1^{'} : F > F_0$

$$T_n(q) = \sup_{0<F_0(t)<1} \frac{\sqrt{n}|\mathbb{F}_n(t) - F_0(t)|}{q(F_0(t))}, \; T_n^+(q) = \sup_{0<F_0(t)<1} \frac{\sqrt{n}\,(\mathbb{F}_n(t) - F_0(t))}{q(F_0(t))},$$

q belongs to the family of the EFKP upper-class functions of $\{B(u), 0 \leq u \leq 1\}$.

As $T_n(q)$, in this generality, appeared for the first time in the paper of M. Csörgő, S. Csörgő, Horváth, and Mason (1986), the statistics $T_n(q)$ and $T_n^+(q)$ will be called the *two-sided* and *one-sided Csörgő-Csörgő-Horváth-Mason (CsCsHM) statistics*, respectively.

Under H_0, by defining $U_i = F_0(X_i)$ for each n,

$$T_n(q) \overset{\mathcal{D}}{=} \sup_{0<u<1} \frac{\sqrt{n}|\mathbb{U}_n(u) - u|}{q(u)}, \quad T_n^+(q) \overset{\mathcal{D}}{=} \sup_{0<u<1} \frac{\sqrt{n}(\mathbb{U}_n(u) - u)}{q(u)},$$

$\mathbb{U}_n(u) = n^{-1} \sum_{i=1}^n \mathbb{I}(U_i \leq u)$ with $U_1, \ldots, U_n$ being iid $U(0,1)$ r.v. The corresponding order statistics are denoted by $U_{(1)} < \ldots < U_{(n)}$.

Asymptotic theory

The utility of the test depends on whether one can work out the distribution theory of the test statistic.

Fact 1 (Theorem 4.2.3 in Csörgő et. al (1986): *The sequence of r. v.'s* $\sup_{0<u<1} \sqrt{n}|\mathbb{U}_n(u) - u|/q(u)$ *converges in distribution to a nondegenerate random variable if and only if q is an EFKP upper-class function. The latter nondegenerate random variable must be the random variable* $\sup_{0<u<1} |B(u)|/q(u)$.

Fact 2 (Lemma 4.2.2 in Csörgő et. al (1986): *Whenever q is an EFKP upper-class function, then for each* $-\infty < x < \infty$ *and any Brownian bridge B*

$$P\left(\sup_{1/n \leq u \leq 1-1/n} |B(u)|/q(u) \leq x\right) \to P\left(\sup_{0<u<1} |B(u)|/q(u) \leq x\right), \quad n \to \infty.$$

By Facts 1 and 2, if H_0 is true, then as $n \to \infty$

$$\begin{aligned} T_n(q) &= \sup_{0<F_0(t)<1} \frac{\sqrt{n}|\mathbb{F}_n(t) - F_0(t)|}{q(F_0(t))} \xrightarrow{\mathcal{D}} \sup_{0<u<1} |B(u)|/q(u), \\ T_n^+(q) &= \sup_{0<F_0(t)<1} \frac{\sqrt{n}(\mathbb{F}_n(t) - F_0(t))}{q(F_0(t))} \xrightarrow{\mathcal{D}} \sup_{0<u<1} B(u)/q(u). \end{aligned}$$

Test procedure based on CsCsHM: advantages

In general, for $I = (a, b)$, $0 \leq a < b \leq 1$ and q an EFKP upper-class function, the statistics

$$T_n(q, I) = \sup_{a<F_0(t)<b} \frac{\sqrt{n}|\mathbb{F}_n(t) - F_0(t)|}{q(F_0(t))}, \quad T_n^+(q, I) = \sup_{a<F_0(t)<b} \frac{\sqrt{n}(\mathbb{F}_n(t) - F_0(t))}{q(F_0(t))}$$

have the same null distributions as the respective **uniform empirical processes in weighted sup-norms**:

$$\sup_{u \in I} |\mathbb{U}_n(u) - u| / q(u), \quad \sup_{u \in I} (\mathbb{U}_n(u) - u) / q(u).$$

The convergence results suggest the following test procedures of asymptotic level α:

- Set $T_n(q) := \sup_{0<u<1} |B(u)| / q(u), \quad T_n^+(q) := \sup_{0<u<1} B(u) / q(u).$
- Reject H_0 in favor of H_1 when $T_n(q) > t_\alpha(q)$, the critical point $t_\alpha(q)$ is chosen to have $P(T(q) \geq t_\alpha(q)) = \alpha$; and reject H_0 in favour of H_1' whenever $T_n^+(q) > t_\alpha^+(q)$, where $t_\alpha^+(q)$ is determined by $P(T^+(q) \geq t_\alpha^+(q)) = \alpha$.
- An effective algorithm for tabulating the CDFs of $T_n(q)$ and $T^+(q)$ is given in Stepanova, Pavlenko (2018).

Motivation for the use of $T_n(q, I)$ and $T_n^+(q, I)$

Tukey's *higher criticism* test statistics (using probability integral transform for each n, see Donoho, Jin (2004), see e.g. Jager and Wellner (2007), Cai et al (2014))

$$\mathrm{HC}_n = \sup_{0<u<\alpha_0} \frac{\sqrt{n}(\mathbb{U}_n(u) - u)}{\sqrt{u(1-u)}}, \quad 0 < \alpha_0 < 1.$$

Construction: The test statistic HC_n^+ is derived from the random variable

$$\max_{0<\alpha\leq\alpha_0} \frac{\sqrt{n}\,(M_n/n - \alpha)}{\sqrt{\alpha(1-\alpha)}},$$

where M_n is the number of hypotheses among H_{0i}, $i = 1, \ldots, n$, that are rejected at level α, which measures the maximum deviation of the observed proportion of rejections from what one would expect it to be purely by chance as the Type I error level changes from zero to α_0 (see DasGupta (2008)). The parameter α_0 defines a range of significance levels in multiple-comparison testing and therefore is a number like 0.1 or 0.2.

The convergence properties of the statistic HC_n are largely determined by the behaviour of $\sqrt{n}(\mathbb{U}_n(u) - u)/\sqrt{u(1-u)}$ in the vicinity of zero and one: the latter inflates when u is close to zero and one.

Motivation for the use of $T_n(q, I)$ and $T_n^+(q, I)$ (cont.)

To overcome this problem, Donoho and Jin (2004), (2008) suggested to truncate the range

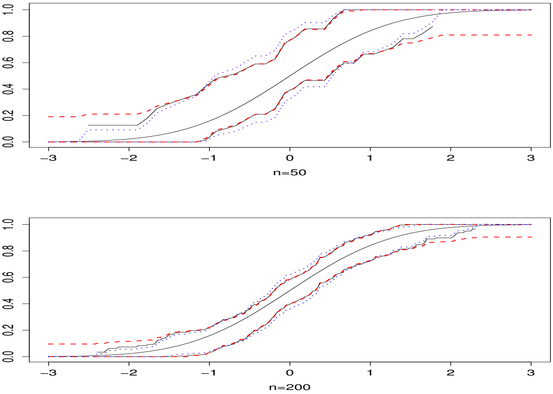
$$\mathrm{HC}_n^+ = \sup_{U_{(1)} < u < U_{([\alpha_0 n])}} \frac{\sqrt{n}(\mathbb{U}_n(u) - u)}{\sqrt{u(1-u)}}, \quad 0 < \alpha_0 < 1$$

which does not eliminate the problem!

In the sup-norm scenario, when the process $\sqrt{n}(\mathbb{U}_n(u) - u)$ is normalized by $\sqrt{u(1-u)}$, **all the action takes place in the tails but, unfortunately, near infinity**. This, together with the fact that, under H_0, the statistic HC_n tends to ∞ in probability, as well as almost surely (see Ch. 16 in Shorack Wellner (1986)), motivated us to search for a better weighed analog of the statistic HC_n^+, for which the **action is shifted somewhat to the middle, while properly regulated on the tails**, and whose limit distribution is sensitive to α_0.

We propose CsCsHM (the weighted KS test statistic) $T_n^+(q, I)$ with $I = (0, \alpha_0)$ and $q(u) = \sqrt{u(1-u) \log\log(1/(u(1-u)))}$ as a competitor HC_n^+. Unlike the test procedures based on HC_n^+ and its modifications with the suprema over $(1/n, \alpha_0)$ and $(U_{(1)}, U_{([\alpha_0 n])})$, in order to perform well, the test procedure based on $T_n^+(q, I)$ do not require a very large sample size of$n = 10^6$ and works well for $n = 10^2$, see Stepanova Pavlenko (2018).

Convergence properties of CsCsHM



Figur: Confidence bands for simulated data. The solid line is the true CDF. The solid lines above and below the middle line are a 95% CsCsHM confidence band. The red dashed lines are a 95% Kolmogorov-Smirnov confidence band.

CsCsHM for optimal detection of sparse heterogeneous mixtures

- Detection of sparse mixtures is a special case of a goodness-of-fit testing problem. Extensive studies after publications of Ingster(1997, 1999), (see also Donoho, Jin (2009), Cai et. al (2011) and Cai et al (2014)).
- We focus on (back to Example 2)

$$H_0 : X_1, \dots, X_n \overset{iid}{\sim} N(0, 1),$$

i.e., F_0 in the goodness-of-fit testing is the standard normal CDF, vs

$$H_{1,n} : X_1, \dots, X_n \overset{iid}{\sim} (1 - \varepsilon_n) N(0, 1) + \varepsilon_n N(\mu_n, 1),$$

where $\varepsilon_n = n^{-\beta}$ for some unknown $\beta \in (1/2, 1)$ and $\mu_n = \sqrt{2r \log n}$ with $0 < r < 1$.

- The choice of the non-zero μ_n makes the problem very hard but yet solvable! If $\xi_1, \xi_2, \dots$ are iid standard normal r.v.'s, then

$$P\left(\max_{1 \le i \le n} |\xi_i| \ge \sqrt{2 \log n}\right) \to 0, \quad n \to \infty.$$

Attainment of the optimal detection boundary

To apply the previously developed theory to the problem of testing H_0 vs $H_{1,n}$, we need to transform X_i's to $Y_i = 1 - \Phi(X_i)$. Let $\mathcal{G}(u)$ denote a common CDF of the Y_i's taking values in $[0, 1]$. Then the problem of testing H_0 versus $H_{1,n}$ transforms to testing

$$\mathcal{H}_0 : \mathcal{G}(u) = F_0(u), \quad \text{the uniform } U(0,1) \text{ CDF}$$

against a sequence of upper-tailed alternatives

$$\mathcal{H}_{1,n} : \mathcal{G}(u) = F_0(u) + \varepsilon_n \left((1-u) - \Phi\left(\Phi^{-1}(1-u) - \mu_n \right) \right) > F_0(u).$$

The test statistic takes the form

$$\mathrm{T}_n^+(q) = \sup_{0<u<1} \frac{\sqrt{n}(\mathbb{G}_n(u) - u)}{q(u)},$$

where $\mathbb{G}_n(u) = n^{-1} \sum_{i=1}^{n} \mathbb{I}(Y_i \leq u)$ is the EDF based on the transforms variables Y_i's.

Attainment of the optimal detection boundary (cont.)

Thm (see S., P. (2018)) *For (r, β) satisfying $r > \rho(\beta)$, the test based on $T_n^+(q)$ is size and power consistent for testing $\mathcal{H}_0$ vs $\mathcal{H}_{1,n}$.* For an EFKP upper class function $q(\cdot)$, consider the test of asymptotic level α that rejects $\mathcal{H}_0$ when $\mathrm{T}_n^+(q) > t_\alpha^+(q)$. Then, for every alternative $\mathcal{H}_{1,n}$, with r exceeding $\rho(\beta)$, the asymptotic level α test based on $\mathrm{T}_n^+(q)$ has a full power,

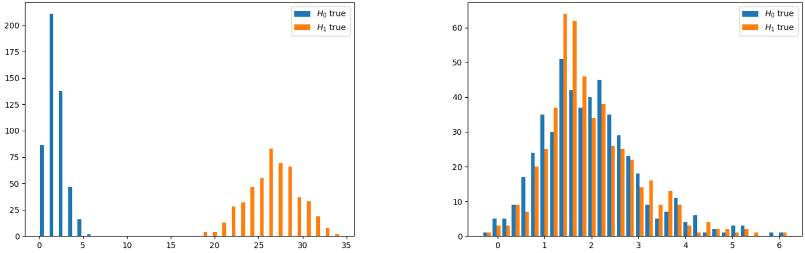
$$P_{\mathcal{H}_{1,n}}(\mathrm{T}_n^+(q) > t_\alpha^+(q)) \to 1 \quad \text{as} \quad n \to \infty.$$

If $r > \rho(\beta)$, then asymptotically the test procedure based on $\mathrm{T}_n^+(q)$ *distinguishes* between $\mathcal{H}_0$ and $\mathcal{H}_{1,n}$

Opt. Everywhere in (r, β) plane - where LRT is successful – the CsCsHM testing with $T_n^+(q)$ also completely separates $\mathcal{H}_0$ and $\mathcal{H}_{1,n}$.

Since $T_n^+(q)$ does not require the knowledge of β and r we call such a test procedure *optimally adaptive*.

Optimal adaptivity of $\mathrm{T}_n^+(q)$: the non-asymptotic context



Figur: Simulated CsCsHM values. Histograms for $T_n^+(q)$ under H_0 and under the alternative $H_{1,n}$ obtained by $n = 10^4$. Left: parametrization, $\beta = 0.55$, $r = 0.9$ (region of success). Right: parametrization, $\beta = 0.9$, $r = 0.3$ (region of failure). See more on non-asymptotic results in Stattin (2017).

Applications: CsCsHM detection threshold in classification

LDA: a quick tour

- Supervised setting: given is n labeled training samples $\{(\mathbf{X}_i, \mathcal{C}_i)\}_{i=1}^n$,
  - $\mathbf{X}_i \sim N(\mathcal{C}_i \boldsymbol{\mu}, \boldsymbol{\Sigma})$ are i.i.d. feature vectors in R^p
  - $\mathcal{C}_i \in \{-1, +1\}$ are class variables
- Goal: given a fresh feature vector $\mathbf{X}$, predict the associated class variable $\mathcal{C}$.
- Fisher linear discriminant: $\mathcal{L}(\mathbf{X}) = \sum_{j=1}^p \omega_j x_j$
- Feature weights $\boldsymbol{\omega} \propto \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ are estimated using $\{(\mathbf{X}_i, \mathcal{C}_i)\}_{i=1}^n$
- Decision rule: Allocate $\mathbf{x}$ to $\mathcal{C} \pm 1$ according to $\mathcal{L}(\mathbf{x}) \gtrless 0$
- Widely used in a huge amount of empirical statistical research
- Optimal weights are asymptotically approachable with $n \gg p$

New data types and modern challenges

Examples: Tumor classification by gene expression data

Data name	Source	n, total sample size	p, features
Colon	Alon et al. (99)	62 $(n_1 = 22, n_2 = 40)$	$p = 2000$
Prostate	Singh et al.(02)	102 $(n_1 = 50, n_2 = 52)$	$p = 12600$
Breast	Pawitan et al. (05)	159 $(n_1 = 128, n_2 = 31)$	$p = 6573$

- **Problem:** Too few observations to estimate Σ^{-1} if $p \gg n$.

  Solution: Regularization of Σ, graph-based technique to learn the structure underlying Σ^{-1} (concentration matrix, $\Omega = \Sigma^{-1}$).

- **Problem:** Many features, most useless, a few useful. Sparse model, rare and weak model.

  Solution: DLDA (for today $\text{diag}(\Sigma^{-1})$) & Feature selection with CsCsHM-thresholding.

Our strategy: *"Use a method that does well in sparse problems, since no procedure does well in dense problems"*, Friedman et al. (2004).

Feature selection by thresholding

1: Obtain a vector $\boldsymbol{z}$ quantifying the feature separation strength, $\boldsymbol{z}$-scores, within a given classification problem: $\boldsymbol{z} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathcal{C}_i \mathbf{x}_i$

2: Define the jth feature weight as *hard threshold function*

$$\omega_j(\psi) = \mathrm{sing}(z_j) \mathbb{1}_{|z_j| > \psi}$$

3: LDA with feature thresholding:

$$\mathcal{L}^*(\mathbf{x}; \psi) = \sum_{j=1}^{p} \omega_j(\psi) x_j \gtrless 0$$

Crucial question: How to determine the **best** threshold ψ?

CsCsHM detection threshold

z_j: z-score as the test statistics for testing the null hypothesis

$$H_{0,j} : j\text{th feature variable is not informative}$$

1. Transform z-scores to p-values: $\pi_j = P(|N(0,1)| > |z_j|)$
2. Arrange: $\pi_{(1)} < \pi_{(2)} < \cdots < \pi_{(p)}$
3. Define the CsCsHM objective function

$$\mathrm{T}^{+}_{p,q}(j;\pi_{(j)}) = \frac{\sqrt{p}(j/p - \pi_{(j)})}{q(j/p)}$$

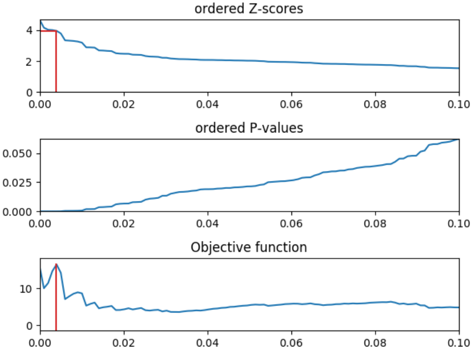
4. Fix $\alpha_0 \in [0, 1/2)$. Obtain the maximizing index of $\mathrm{T}^{+}_{p,q}(j;\pi_{(j)})$:

$$j^* = \arg \max_{1 \le j \le [\alpha_0 p]} \mathrm{T}^{+}_{p,q}(j;\pi_{(j)})$$

5. Set the CsCsHM threshold as (new ingredient):

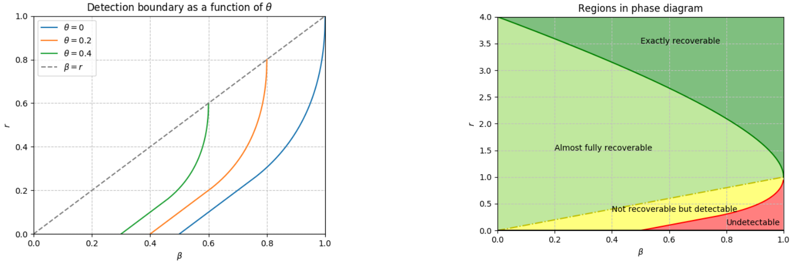
$$\psi^* = |z|_{(j^*)} \text{ corresonding to maximizing, } j^*$$

Illustration of CsCsHM thresholding



Figur: Parametrization: $p = 10^3$, $\beta = 0.6$, $r = 0.7$, $\alpha = 0.5$ (only fragment to $\alpha_0 = 0.1$ is shown). The component score maximizing the objective function $\mathrm{T}^{+}_{p,q}(j; \pi_{(j)})$ located at the red line.

Phase space for classification problem

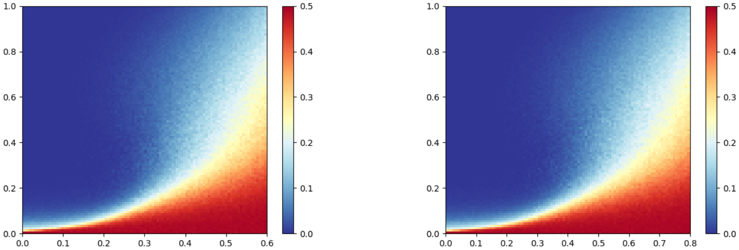


Figur: Left: Classification boundaries and optimality for ARW models. Calibration using p as a driving index:

$$\varepsilon_p = p^{-\beta},\ \mu = \sqrt{2r\log p},\ n_p = p^{\theta},\ \theta \in (0,1)$$
$$\rho_\theta(\beta) = (1-\theta)\rho(\beta/(1-\theta)),\ 0 < \beta < (1-\theta).$$

Right: Phase diagram (*regular growth*: $n_p = p^{\theta}$), $0 < \beta < 1$ and $0 < r < 4$. In the dark green region, it is not only possible to construct successful classifiers, but is also possible to separate useful features from useless ones. Feature variables assumed to be independent (Naive approach).

Phase transition in non-asymptotic context



Figur: Empirical behavior of finite sample size of $\mathrm{T}^{+}_{p,q}(j;\pi_{(j)})$ thresholding with $\alpha_0 = 0.1$. Heat map of $\widehat{\mathrm{Err}}$ with $\theta = 0.4$, $p = 10^3$ (left) and $\theta = 0.2$ $p = 10^4$ (right), showing the transition zone near the detection boundary. The grid is set to (100×100). The simulations are performed on resources provided by the *Swedish National Infrastructure for Computing* (SNIC) at Tegner PDC.

Analysis of classification accuracy of CsCsHM-thresholding

Tabell: Mean misclassification rates and mean number of variables selected with std for the *Leukemia* data set, with data size of (73×7129)

Method	Misclassification rate	N. variables selected
HC^+	0.024 (± 0.00771)	151.4 (± 4.35)
$\mathrm{T}^+_{p,q}$	0.026 (± 0.00284)	68.4 (± 3.09)

Tabell: Mean misclassification rates and mean number of variables selected with standard deviation for the *Colon caner* data set, with a size of (62×2000)

Method	Misclassification rate	N. variables selected
HC^+	0.1058 (± 0.01024)	87.6 (± 12.30)
$\mathrm{T}^+_{p,q}$	0.1075 (± 0.01176)	35.8 (± 4.78)

Conceptual advantages of CsCsHM

- is the data-driven nonparametric statistics
- performs optimally under ARW – *without* needing to know the underlying ARW parameters (*optimal adaptivity*)
- is simple (compare e.g. with internally very complex SVM classifier) and extremely fast, does not require tuning, cross-validation, ...
- is competitive in performance accuracy (outperforms out-of-the box classifiers such as SVM, BagBost)
- See e.g result reported in Donoho and Jin (2008):
  - *Colon cancer*, SVM: $\widehat{\mathrm{Err}} = 0.1505$, BagBost: $\widehat{\mathrm{Err}} = 0.1610$
  - *Leukemia*, SVM: $\widehat{\mathrm{Err}} = 0.0189$, BagBost: $\widehat{\mathrm{Err}} = 0.0408$

Further challenges: beyond normality

Threshold choice for optimal detection of informative sets of features (gene-sets or *pathways* whose expressions jointly change significantly between given conditions, i.e. *tumor/normal*).

Problems:

- Quantifying the separation strength of a subset of variables within a given classification problem.
- Too many subsets identified at the structure learning stage, a few informative/weak signals.

Our strategy:

- Squared Mahalanobis norm for the location shift and its distribution in high-dimensions. Relation to misclassification probability.
- Modeling of sparse and weak data in classification setting.
- Linking the optimal choice of threshold to multiple testing procedures.

Detection problem: Whether any sets are informative?

- Distributional properties. For $p_0 \ll p$

$$\hat{\delta}_j^2 = \frac{n_1 n_2}{n_1 + n_2}(\hat{\boldsymbol{\mu}}_{1,[j]} - \hat{\boldsymbol{\mu}}_{2,[j]})'\hat{\boldsymbol{\Omega}}_{[j]}(\hat{\boldsymbol{\mu}}_{1,[j]} - \hat{\boldsymbol{\mu}}_{2,[j]}).$$

$$\mathcal{S}_i^2 = \frac{n_1 + n_2 - p_0 - 1}{(n_1 + n_2 - 2)p_0}\hat{\delta}_j^2 \in \mathcal{F}\left(\cdot; p_0, n_1 + n_2 - p_0 - 1, \frac{n_1 n_2}{n_1 + n_2}\delta_j^2\right).$$

- In (n, p)-asymptotic, uniformly over j

$$\mathcal{S}_j^2 \to \chi^2(\cdot; p_0, \gamma_j^2) \text{ in distribution.}$$

- The goal to investigate the properties of CsCsHM for testing and thresholding

$$\begin{aligned} H_0 &: \mathcal{S}_j^2 \in \chi^2(\cdot; p_0, 0), \\ &\quad \text{vs } p\text{-dependent alternatives} \\ H_{1,b} &: \mathcal{S}_j^2 \in (1 - \varepsilon_p)\chi^2(\cdot; p_0, 0) + \varepsilon_p \chi^2(\cdot; p_0, \gamma^2), \end{aligned}$$

Current work

- When the separation strength may be unequal

  Our strategy: Replace the single χ^2 distribution by a mixture of non-central χ^2s so that the alternative hypothesis becomes

  $$H_{1,b}: \quad \mathcal{S}_i^2 \text{ are iid from } \quad (1-\beta_p)\chi^2(\cdot,0) + \beta \int \chi^2(\cdot,\gamma^2)dG_n(\gamma^2)$$

  where $\chi^2(\cdot,u)$ is the density of $\chi^2(\cdot,\gamma^2)$ and $G_n(\gamma^2)$ is some distribution function of the separation strength.
- How to (efficiently) estimate proportion of informative blocks, β in SWBM?
- Exact classification boundary separating detectable and undetectable regions for χ^2 family.
- Relationship between ideal block-thresholding, CsCsHM-thresholding and misclassification probability.

Some references

- Cai T., Jeng, X. J, & Jin, J. (2011) Optimal detection of heterogeneous and heteroscedastic mixtures. J.R. Statist. Soc B, 73,pp. 629-662.
- Cai T. and Wu Y. (2014) Optimal detection of sparse miixtures against a given null alternative. IEEE Transactions on information theory, 60(4), pp. 2217-2232.
- Csörgő, M., Csörgő, S., Horváth, L., & Mason, D. (1986) Weighted empirical and quantile processes. *Ann. Probab.*, 14, pp. 31–85.
- Donoho, D. and Jin, J. (2009). Feature selection by Higher criticism thresholding achieves the optimal phase diagram. Phil. Trans. R. Soc. Lond. A, 367, pp. 4449-4470.
- Dobrushin R. (1958). A statistical problem arising in the theory of detection of signals in the presence of noise in a multi-channel system and leading to stable distribution laws. Theory Probab. Appl., no. 2, pp. 161–173.
- Ingster Y. (2001) Adaptive detection of a signal of growing dimension. I," Math. Methods Statist., vol. 10, no. 4, pp. 395–421.
- Ingster Y. (2002) Adaptive detection of a signal of growing dimension. II," Math. Methods Statist., vol. 11, no. 1, pp. 37–68.

Some references

- Jager L. and Wellner J. (2007). Goodness-of-fit tests via phi-divergences. Ann. Statist. 35, 2018-2053.
- Meinshausen N. and Rice J. (2006) Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses.
- Pavlenko T Björkström A. and Tillander A. (2012) Covariance structure approximation via gLasso in high-dimensional supervised classification. J. of Applied Stat. 8, 1643-1666.
- Shorack G. and Wellner J. (2009) Empirical Processes With Applications to Statistics. Philadelphia SIAM.
- Stepanova N Pavlenko T. (2018) Goodness-of-fit tests based on sup-functionals of weighted empirical processes. Theor. Probab. Appl. (TVP) SIAM, 63(2), 358-388.
- Stepanova N. Pavlenko T. Stattin O. (2018) Variable selection via CsCsHM thresholding. Working paper.