Visa Koivunen
Dept of Signal Processing and Acoustics
School of Electrical Engineering
Aalto University
Finland

Abdelhak M. Zoubir
Signal Processing Group
Institute of Telecommunications
Technische Universität Darmstadt
Germany

The support of Dr.-Ing. Michael Muma and Dipl.-Ing. Michael Fauss is acknowledged
Overview of the Tutorial

- Introduction and Basic Concepts of Robustness
- Robustifying the Regression Model
- Robust Methods for Complex Valued Multichannel Data
- Robust Methods for Dependent Data
- Recent Trends in Robust Signal Processing
  - Advances in Robust Detection
  - Robust, Scalable and Fast Bootstrap Method for Analyzing Large Scale Data
- Additional Material
  - High-Dimensional Data
  - Regularized M-Estimation
Overview of the Tutorial

Remarks to the content of this talk:

- We do not deal with outlier removal.
- We consider robustness in the sense of "unwanted" effects with respect to a nominal distribution (usually the Gaussian). Of course, in the case of natural heavy-tailed phenomena the corresponding models should be used.
Introduction and Basic Concepts of Robustness
Classical Theoretical Approach to Estimation

- strong and precise assumptions
- e.g., estimators, detectors or filters optimal under nominal distribution

Optimality only achieved when assumptions hold exactly.
Real World Data in Signal Processing

- in many cases Gaussian assumption well justified
- measurement campaigns confirmed impulsive (heavy-tailed) noise, e.g. in
  - outdoor and indoor mobile communication channels
  - radar and sonar systems
  - biomedical sensor (array) measurements, e.g. magnetic resonance imaging (MRI)
- outliers in the measurements, e.g. in
  - geolocation position estimation and tracking (NLOS)
  - short-term load forecasting
  - motion artifacts for portable medical devices

Performance of optimal procedures may deteriorate significantly, even for minor departures from the assumed model.
Introduction

Motivation

Real World Data Example 1: Intracranial Pressure (ICP)

photo: www.esa.int
ten hour excerpt of a typical ICP measurement
Introduction

Motivation

Real World Data Example 1: Intracranial Pressure (ICP)

photo: www.esa.int
ten hour excerpt of a typical ICP measurement

artifacts $\rightarrow$ robust ICP forecasting $\rightarrow$ early interventions for patients with traumatic brain injuries
Real World Data Example 2: Electricity Consumption

- first corrected from weather influence by regression model
- residual series encompasses (daily, weekly, yearly) seasonalities, modeled as seasonal autoregressive integrated moving average (SARIMA)

- one-week difference in consumption in France at 7:00 a.m. in 2005
- outliers due to special days and events (marked in red)
Aim 1: Near Optimality [Huber-09]:

‘The procedure should behave “reasonably well” at the assumed model.’
Aims of Robust Methods
Approximate Quotes by Huber and Hampel

Aim 1: Near Optimality [Huber-09]:

‘The procedure should behave “reasonably well” at the assumed model.’

Aim 2: Qualitative Robustness [Hampel-85]:

‘The effect of an erroneous observation, even if it takes an arbitrary value, should not have a large impact on the method.’
Aims of Robust Methods
Approximate Quotes by Huber and Hampel

Aim 1: Near Optimality [Huber-09]:

‘The procedure should behave “reasonably well” at the assumed model.’

Aim 2: Qualitative Robustness [Hampel-85]:

‘The effect of an erroneous observation, even if it takes an arbitrary value, should not have a large impact on the method.’

Aim 3: Quantitative Robustness [Huber-09]:

‘Somewhat larger deviations from the model should not cause a catastrophe.’
Aims of Robust Methods
Approximate Quotes by Huber and Hampel

Aim 1: Near Optimality [Huber-09]:
‘The procedure should behave “reasonably well” at the assumed model.’

Aim 2: Qualitative Robustness [Hampel-85]:
‘The effect of an erroneous observation, even if it takes an arbitrary value, should not have a large impact on the method.’

Aim 3: Quantitative Robustness [Huber-09]:
‘Somewhat larger deviations from the model should not cause a catastrophe.’

How can I quantify, if my estimator fulfills these aims?
Measuring Robustness
Intuitive Definitions

**Measure 1: Relative Efficiency**

- increase in the variance \( (\sigma^2) \) of the estimator *at the assumed model* compared to the optimal method

optimal estimator of the mean of a zero-mean Gaussian random variable
Measure 1: Relative Efficiency

- increase in the variance ($\sigma^2$) of the estimator at the assumed model compared to the optimal method

robust estimators of the mean of a zero-mean Gaussian random variable
Measure 2: Influence Function (IF)

- bias impact of an infinitesimal contamination on the estimator, standardized by the fraction of contamination

bounded and continuous IF $\rightarrow$ qualitative robustness
Measure 3: Maximum Bias Curve (MBC)

- maximum possible bias of an estimator plotted over the fraction of outliers
Measure 3: Maximum Bias Curve (MBC)
▶ maximum possible bias of an estimator plotted over the fraction of outliers

Measure 4: Breakdown Point (BP)
▶ maximal fraction of outliers that an estimator can handle (0 ≤ BP ≤ 50%)
Robustifying the Regression Model
Robustifying the Regression Model

Linear Regression Model

Many Engineering Problems: Written as Linear Regression Model

\[ Y = X\theta + E \]  \hspace{1cm} (1)

- \( X \) and \( E \) are i.i.d., \( X \) is of dimension \( N \times p \), \( \theta \) is of dimension \( p \times 1 \)
- \( Y \) is of dimension \( N \times 1 \), \( Y(n) = y_n \)
- \( x_n^T \) is the \( n^{th} \) \( p \)-valued row of the matrix \( X \)

Applications Require Robustness, e.g.

- wireless localisation [Hammes et al., 2009], [Guvenc et al., 2009]
- ultrasonic systems [Prieto et al., 2009]
- computer vision [Stewart, 1999], [Ye et al., 2003]
- electric power systems [Mili et al., 2002]
- automated detection of defects [Thomas et al., 2007]
- electroencephalographic (EEG) signal analysis [Bénar et al., 2007]
Robustifying the Regression Model

Effects of Outliers

Different Outliers in the Regression Model

1. **Gaussian ML estimator**

   \[
   \arg\min_{\theta} \sum_{n=1}^{N} (y_n - x_n^T \theta)^2
   \]

   \(\Theta\) not robust against any type of outliers
Robustifying the Regression Model

M-estimation

Key Idea:
▶ replace negative log-likelihood function by $\rho(x)$

Example: Huber’s Robust M-estimator

$$
\rho(x) = \begin{cases} 
\frac{1}{2} x^2 & |x| \leq c_{\text{Hub}} \\
-\frac{1}{2} c_{\text{Hub}}^2 & |x| > c_{\text{Hub}} 
\end{cases}
$$

$$
\psi(x) = \begin{cases} 
x & |x| \leq c_{\text{Hub}} \\
c_{\text{Hub}} \text{sign}(x) & |x| > c_{\text{Hub}} 
\end{cases}
$$

(a) $\rho$-function

(b) score function $\psi$
Robustifying the Regression Model

Effects of Outliers

Different Outliers in the Regression Model

2. M-estimators:

\[ \arg\min_{\theta} \sum_{n=1}^{N} \rho \left( \frac{y_n - x_n^T \theta}{\hat{\sigma}_{E,\text{rob}}} \right) \]

⊕ robust against vertical outliers, easy to compute
Robustifying the Regression Model

Effects of Outliers

Different Outliers in the Regression Model

- M-estimators:

\[
\sum_{n=1}^{N} \psi \left( \frac{y_n - x_n^T \theta}{\hat{\sigma}_{E,rob}} \right) x_n = 0
\]

\( \oplus \) BP=0 for outliers in \( x_n \) (leverage points)
3. **S-estimator**: [Rousseeuw et al., 1984], [Salibian-Barrera et al., 2006] minimizes robust scale of the residuals

\[ \hat{\theta}_S = \arg\min_{\theta} \hat{\sigma}(r(\theta)) \]

- \( \hat{\sigma}(\cdot) \) is an M-estimate of scale defined by

\[ \frac{1}{N} \sum_{n=1}^{N} \rho \left( \frac{r_n(\theta)}{\hat{\sigma}} \right) = \delta \]  

- \( r(\theta) = (r_1(\theta), \ldots, r_N(\theta))^T \), with \( r_n(\theta) = y_n - x_n^T \theta \)

- combination of high efficiency and BP not possible
- fast algorithm [Salibian-Barrera et al., 2006]
4. **LTS-estimator**: [Rousseeuw and Leroy, 1987], [Rousseeuw and Van Driessen, 2006]

   LTS stands for "Least Trimmed Squares"
   finds an estimate based on a subset $N/2 < h < N$ of data points, whose least-squares fit has smallest sum of squared residuals

   ⊕ combination of high efficiency and BP not possible
   ⊗ fast algorithm [Rousseeuw and Van Driessen, 2006]
Robustifying the Regression Model

Robust and Efficient Estimators

5. \( \tau \)-estimator: [Yohai, 1988], [Salibian-Barrera et al., 2006]

minimizes robust and efficient \( \tau \)-scale of the residuals

\[
\hat{\theta}_\tau = \arg\min_{\theta} \hat{\sigma}_\tau(r(\theta))
\]

\( \hat{\sigma}_\tau(\cdot) \) is a \( \tau \)-estimate of scale defined by

\[
\hat{\sigma}_\tau(r(\theta))) = \hat{\sigma}_1(r(\theta)) \sqrt{ \frac{1}{N} \sum_{n=1}^N \rho_2 \left( \frac{r_n(\theta)}{\hat{\sigma}_1(r(\theta))} \right) }
\]

\( r(\theta) = (r_1(\theta), \ldots, r_N(\theta))^T \), with \( r_n(\theta) = y_n - x_n^T \theta \)

\( \hat{\sigma}_1(\cdot) \) satisfies Eq. (2) with highly robust \( \rho_1(\cdot) \), while \( \rho_2(\cdot) \) is highly efficient

\( \Box \) highly robust and efficient: BP=0.5 and Eff=0.95

\( \Box \) fast algorithm [Salibian-Barrera et al., 2006]
6. **MM-estimator**: [Yohai, 1987], [Salibian-Barrera et al., 2006]

**Step 1.** Compute an initial consistent high BP S-estimate: $\hat{\theta}_S$.

**Step 2.** Compute the high BP M-scale of the residuals of Step 1: $\hat{\sigma}_1(r(\hat{\theta}_S))$.

**Step 3.** Compute an efficient M-estimate of regression, using an iterative procedure starting at $\hat{\theta}_S$:

$$
\arg\min_\theta \sum_{n=1}^{N} \rho_2 \left( \frac{y_n - x_n^T \theta}{\hat{\sigma}_1(r(\hat{\theta}_S))} \right)
$$

⊕ highly robust and efficient: BP=0.5 and Eff=0.95
⊕ fast algorithm [Salibian-Barrera et al., 2006]
Robustifying the Regression Model
Application Example: Localisation

Application Example: Localisation of a Mobile User Equipment

- localise wireless transmitter device using different base stations
- important task in many civilian and military applications
- urban scenario: Line-of-Sight (LOS) and Non-Line-of-Sight (NLOS) propagation → severe degradation of position estimates
Robustifying the Regression Model
Application Example: Localisation

Localisation Based on Time of Arrival (TOA) Measurements:

[Gustafsson et al., 2005]

non-linear measurement equation at each base station (BS)

\[ y_n = h(\theta) + v_n, \quad n = 1, \ldots, N \]

- \( h = \sqrt{(x - x_{BS,m})^2 + (y - y_{BS,m})^2} \) distance from the UE to the \( m \)-th BS, where \( m = 1, \ldots, M \)
- \( \theta = (x, y)^T \) position of the UE
Robustifying the Regression Model

Application Example: Localisation

Localisation Based on Time of Arrival (TOA) Measurements:
[Gustafsson et al., 2005]

Linearisation of $h(\cdot) \rightarrow$ linear regression model of Eq. (1):

$$Y = X\theta + E$$
Robustifying the Regression Model

Application Example: Localisation

Localisation Based on Time of Arrival (TOA) Measurements:
[Gustafsson et al., 2005]

NLOS propagation → outliers in TOA measurements

possible model

\[ f_E(e) = (1 - \varepsilon)f_{LOS}(0, \sigma^2_{LOS}) + \varepsilon f_{NLOS}(\mu_{NLOS}, \sigma^2_{NLOS}) \]
Simulation Setup:

- NLOS effects modeled as i.i.d. random variables with probability density
  \[ f_V(v) = (1 - \varepsilon)f_G(v; 0, \sigma_{LOS}^2) + \varepsilon h_{NLOS}(v) \]

- \( \sigma_{LOS} = 150 \text{m} \), \( \varepsilon = 0.4 \), \( h_{NLOS}(v) \) is the exponential density
- \( M = 10 \) base stations (located at blue dots)
- position of UE: randomly generated with uniform distribution in the green area
- \( N = 5 \) measurements available at each base station.
Robustifying the Regression Model
Application Example: Localisation

Simulation Setup:

- 4 estimators with computational complexity increasing with decreasing order:
  1. ML under Gaussian distribution (LSE)
  2. Huber’s M-estimator (tuned for 95% efficiency)
  3. S-estimator
  4. MM-estimator

- Mean Circular Positioning Error (MCPE) using 10,000 Monte-Carlo runs:

\[
MCPE = \frac{1}{10^4} \sum_{i=1}^{10^4} \sqrt{(\hat{x}(i) - x(i))^2 + (\hat{y}(i) - y(i))^2}.
\]

displayed w.r.t. scale of \( h_{NLOS}(v) \), Gaussian case: \( \sigma_{NLOS} = 0m \)
Robustifying the Regression Model
Application Example: Localisation

Simulation Results:

- M-Estimator outperforms ML for all $\sigma_{NLOS} > 0m$
- S-Estimator better than M-Estimator for $\sigma_{NLOS} > 750m$
- MM-Estimator outperforms all competitors, gives stable performance over all $\sigma_{NLOS}$ (simultaneously robust and efficient)
Robust Methods for Complex Valued Multichannel Data
Complex Valued Multichannel Data

Outline

- Complex-valued Signal Models
- Multivariate Statistics: distribution models,
- Uncertainty in noise modeling
- Robustness Theory: quantitative and qualitative robustness
- Robust statistical methods
- Array processing examples
Many key signal processing applications use multiple sensors to acquire complex multi-channel data.

Typical examples are MIMO communication systems, phased array, surveillance and radar systems.

Many biomedical measurement systems such as MEG and EEG, and environmental monitoring produce multi-channel data.

Mobile terminals are equipped with 10+ radios and other sensors.

The measurements are vector-valued and often realizations of random signals and noise are described by a multivariate probability distribution function.

The most commonly used multivariate statistical model is the multivariate Gaussian distribution, which is fully described by first and second order statistics, i.e., the mean vector and the covariance (and pseudocovariance) matrix.
One may attempt to derive robust signal processing techniques for complex multi-channel data by simply applying robust methods developed for scalar signals to each signal component independently.

This approach is unsuitable as it may lead to unexpected results. For example, in the case of location estimation, the resulting estimate is not necessarily within the convex hull of the data.

Moreover, none of the data vector components may be outlying alone but the multivariate observations they form may be far away from the majority of the data.

Such outliers would not be detected if robust signal processing is performed on individual single components.

The vector components in multi-channel data may be correlated and have different variances which makes detecting outliers more difficult.

Data may be non-circular ($E[zz^T] \neq 0$)
Typical multivariate estimation problems are estimation of location and scatter parameter.

Robust estimation of the multivariate location parameter is needed in many multi-channel filtering problems such as noise attenuation and outlier (impulsive noise) removal in color images, multi-channel biomedical measurements, multimodal imaging as well as sensor array signal processing.

Location estimation in real-valued case is well covered in statistics literature.

Robust estimation of scatter (covariance) matrices is a key to numerous signal processing tasks of optimal multi-channel estimation and filtering, such as direction finding, beamforming, interference cancellation, spatial multiplexing, and signal separation.

We focus on scatter estimation in this part.
Affine equivariant estimators are highly desirable since it facilitates analyzing estimators even if components have different scales and are correlated. By denoting an estimator by a functional $\mathbf{T}(\mathbf{z})$, affine equivariance means in multivariate location estimation that

$$\mathbf{T}(B\mathbf{z} + \mathbf{b}) = B\mathbf{T}(\mathbf{z}) + \mathbf{b}$$

and in scatter (covariance) estimation that

$$\mathbf{T}(B\mathbf{z} + \mathbf{b}) = B\mathbf{T}(\mathbf{z})B^H,$$

where $\mathbf{B}$ is a $K \times K$ full rank transformation matrix and $\mathbf{b}$ is a $K \times 1$ vector and the superscript $^H$ denotes Hermitian operation.
Indoor measurements campaigns show the impulsive nature of noise and interference due to, e.g., microwave ovens and devices with electromechanical switches.

Noise and interference in biomedical sensor array measurements (brain activity, MRI) is often non-Gaussian.

Heavy-tailed noise and clutter models in radar, e.g. [Gini and Greco, 2001]

Robust multichannel algorithms are especially interesting because the efficiency of the robust estimators tends to grow as a function of array size. The efficiencies are often very close to optimal in Gaussian case already with relatively small sensor arrays (6-8 elements).

Common noise models include symmetric $\alpha$-stable ($\text{S}_\alpha\text{S}$) noise and $\epsilon$-contaminated noise.
Covariance matrix of multichannel complex data $\mathbf{R} = E[\mathbf{zz}^H]$ is used for example in

- MVDR and classical beamformers, spatial smoothing use array covariance matrix or its inverse
- Multiantenna CMOE and MMSE receivers
- Eigenvalues and eigenvectors of covariance matrix (MUSIC, ESPRIT, MDL), eigenbeamforming and spatial multiplexing in MIMO systems.
- High-resolution methods, signal and noise subspaces (MUSIC, ESPRIT,...), projection matrices (Stochastic and Deterministic ML, Weighted subspace fitting, noise statistics estimation), jammer and interference cancellation
For example, in array processing commonly used conventional sample covariance matrix

\[ \hat{R} = \frac{1}{N} \sum_{i=1}^{N} z_i z_i^H \]

computed from spatial sensor array snapshots \( z_1, \ldots, z_N \) leads to optimal estimates in Gaussian noise.

However, it has poor performance even for small departures from the nominal Gaussian noise model.

Consequently, the statistics (such as eigenvalues, eigenvectors, projection matrices) based on it are unreliable and far from optimal.

We consider \( M \)-estimation, nonparametric methods and ML methods for heavy-tailed distributions as robust alternatives.

Pseudocovariance \( P = E[zz^T] \) needs to be taken into account, see [Schreier and Scharf, 2010] or [Ollila et al., 2012].
**DEFINITION** random vector $z = x + jy \in \mathbb{C}^k$ is said to have a complex elliptically symmetric (CES) distribution with parameters

$$\mu = \mu_1 + j\mu_2 \in \mathbb{C}^k \quad \text{and} \quad \Sigma = \Sigma_1 + j\Sigma_2 \in \mathcal{H}(k)$$

if and only if

$$z_R = \begin{pmatrix} x \\ y \end{pmatrix}$$

has real elliptically symmetric distribution with parameters

$$\mu_R = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \in \mathbb{R}^{2k} \quad \text{and} \quad \Sigma_R = \frac{1}{2} \begin{pmatrix} \Sigma_1 & -\Sigma_2 \\ \Sigma_2 & \Sigma_1 \end{pmatrix} \in \text{PDS}(2k)$$
If pdf of $z_R$ is

$$f_R(z_R; \mu_R, \Sigma_R) = |\Sigma_R|^{-1/2} g \left( (z_R - \mu_R)^T \Sigma_R^{-1} (z_R - \mu_R) \right),$$

for some nonnegative function $g$, then the pdf of $z$ is

$$f_C(z; \mu, \Sigma) = 2^k |\Sigma|^{-1} g \left( 2(z - \mu)^H \Sigma^{-1} (z - \mu) \right).$$

**Example 1:** Complex $k$-variate normal: $g(t) = (2\pi)^{-k} \exp(-t/2)$

**Example 2:** Complex $k$-variate $t$-distribution:

$$g(t) \propto c (1 + t/\nu)^{-(2k+\nu)/2},$$

where $\nu$ is the number of degrees of freedom.

Note: $\nu = 1$ corresponds to complex multivariate Cauchy distribution.
Location $\mu$ is the symmetry center of the distr., equal to $E_{\mu,\Sigma}(z)$ (when exists). Symmetry: $z - \mu \overset{d}{=} \exp(i\theta)(z - \mu) \ \forall \theta \in \mathbb{R}$.

Scatter matrix $\Sigma$ is proportional to covariance matrix of $z = x + jy$ (when exists). Specifically, $\text{Var}(z) = E_{0,1}(|z_1|^2)\Sigma$,

$$\text{Var}(z) = \text{Var}(y) \quad \text{and} \quad \text{Cov}(x, y) = -\text{Cov}(y, x).$$
Several authors have proposed robust DOA and channel parameter estimation and detection methods for non-Gaussian noise environments, see e.g. [Williams and Johnson, 1993], [Tsakalides and Nikias, 1996], [Swami and Sadler, 2002], [Arce et al. 2001], [Visuri, Koivunen and Oja, 2001], [Kozick and Sadler, 2000], [Gini and Greco, 2001]

The following topics are addressed:

- Concept of $M$-estimation of scatter matrix for data in complex vector space $\mathbb{C}^k$.
- Stochastic ML-estimators for heavy-tailed models.
- Influence functionals for eigenvalues and eigenvectors
- Subspace DOA estimation and MDL number of signals estimation.
- Nonparametric scatter matrix estimates
We consider robust statistical procedures for estimating parameters in the face of noise model uncertainty. The concepts of quantitative and qualitative robustness are presented to formally establish robustness. This is achieved by using the concepts of breakdown point and influence function. Nonparametric and approximately parametric robust methods for array processing applications are presented. Maximum likelihood estimators for heavy-tailed models (e.g. complex multivariate t-distribution) lead to robust estimators that are also statistically efficient (close to optimal).
The concept of breakdown point is typically used for quantitative robustness.

The breakdown point is defined by the smallest fraction of outliers that can cause the estimator to break down, i.e., bias may go to $\infty$.

In the context of covariance estimation this could mean the smallest number $k$ observations out of $N$ observations that could make the condition number very large. The (finite sample) breakdown point for a scatter matrix estimator may be defined as follows

$$
\varepsilon^* = \min_{1 \leq k \leq N} \left\{ \frac{k}{N} \left| \sup_{Z_{N,k}} D\left( \hat{\Sigma}(Z_N), \hat{\Sigma}(Z_{N,k}) \right) = \infty \right. \right\},
$$

where $Z_{N,k}$ is the corrupted sample obtained by replacing $k$ of observations of $Z_N$ with arbitrary values and the supremum is taken over all possible data sets $Z_{N,k}$.
$D(,)$ above is a measure of dissimilarity of two positive definite Hermitian matrices and defined as:

$$D(A, B) = \max\{ |\lambda_1(A) - \lambda_1(B)|, |\lambda_M(A)^{-1} - \lambda_M(B)^{-1}| \},$$

with $\lambda_1(A) \geq \ldots \geq \lambda_M(A)$ being the ordered eigenvalues of the matrix $A \in \mathcal{H}(M)$.

- The definition is given using eigenvalues but similar concepts for eigenvectors has been constructed.
- Breakdown point is always below 50% and it often depends on the dimension.
A robust estimator should have a bounded and continuous IF. Once it is derived, it may be evaluated for different distributions.

IF is essentially the first derivative of the functional version of an estimator. If we denote the point mass at $z$ by $\Delta_z$ and consider the contaminated distribution $F_\varepsilon = (1 - \varepsilon)F + \varepsilon\Delta_z$ then the influence function (IF) of a functional $T$ at $F$ is

$$\text{IF}(z; T, F) = \lim_{\varepsilon \downarrow 0} \frac{T(F_\varepsilon) - T(F)}{\varepsilon} = \left. \frac{\partial}{\partial \varepsilon} T(F_\varepsilon) \right|_{\varepsilon = 0}. \quad (3)$$
In several applications the eigenvectors and eigenvalues of $\Sigma$ are the statistics of interest. They are found from the eigenvalue decomposition $\Sigma = \Gamma \Lambda \Gamma^H$.

The functional representations of the eigenvalues $\hat{\lambda}_i$ and eigenvectors $\hat{\gamma}_i$, $i = 1, \ldots, M$ of $\hat{\Sigma}$ are obtained from the eigenvalue decomposition of the scatter functional, $C(F) = G(F)L(F)G(F)^H$, where $G(F) = (g_i(F) \cdots g_M(F))$ and $L(F) = \text{diag}\{l_1(F), \ldots, l_M(F)\}$, $g_i(F) \in \mathbb{C}^M$ and $l_i(F) \in \mathbb{R}$ are the eigenvector and eigenvalue functionals of $C(F)$.

Since $C(F) = c\Sigma$ ($c$ is constant, proportionality), it immediately follows that $g_i(F) = \gamma_i$ and $l_i(F) = c\lambda_i$. 
As an example, the influence function for eigenvector functional $g_i(F)$ of $C(F)$ corresponding to a simple eigenvalue $\lambda_i$ is given by:

$$IF(z; g_i, F) = \frac{\alpha(r)}{c} \sum_{j=1, j\neq i}^{M} \frac{\sqrt{\lambda_j \lambda_i}}{\lambda_i - \lambda_j} u_j u_i^\ast \gamma_j,$$

where $\gamma_1, \ldots, \gamma_M$ denotes the eigenvectors of the scatter matrix $\Sigma$.

For sample covariance matrix, for example

$$\alpha(r) = r^2$$

and

$$\beta(r) = r^2 / M - E_l(|Z_1|^2),$$

i.e. they are quadratic in $r$ and hence unbounded and non-robust.
\|IF(z; g_1, F)\| for the sample covariance estimator (left) and $t_1 M$-estimator (right) at bivariate complex normal distribution ($\lambda_1 = 1$, $\lambda_2 = 0.6$).
Different robust covariance matrix estimators are presented using basic array signal processing model and direction finding application.

First, basic model for high resolution parameter estimation using Uniform Linear Array (ULA) model is presented.

Robust covariance matrix estimators based on $M$-estimation, ML approach and nonparametrics are described.

Direction of arrival estimation examples using ULA are provided.
A uniform linear array of $M$ sensors receiving plane waves from $K$ far-field point sources.
Low rank signal model of $K$ incoherent source signals impinging to an array of $M$ sensors ($K < M$), see, for example [Krim and Viberg, 1996]. The received signal vector $z(n)$ is an $M \times 1$ complex vector given by

$$z(n) = As(n) + w(n)$$

(4)

where $A$ is an $M \times K$ matrix such that $A = [a(\theta_1), a(\theta_2), \ldots, a(\theta_K)]$ with $a(\theta_k)$ being the $M \times 1$ array steering vectors, $s(n) = [s_1(n), s_2(n), \ldots, s_K(n)]^T$ is the signal vector and $w(n)$ is the $M \times 1$ circular complex noise vector.

The covariance matrix of $z(n)$ (if it exists) is

$$R = E[z(n)z^H(n)] = AR_sA^H + \sigma^2 I$$
DEFINITION The $M$-estimator of covariance based upon a sample $\mathbf{z}_1, \ldots, \mathbf{z}_N$ in $\mathbb{C}^K$ is the positive definite hermitian (PDH) $K \times K$ matrix $\hat{\Sigma}$ which solves

$$
\Sigma = \frac{1}{N} \sum_{i=1}^{N} u\left(\mathbf{z}_i^H \Sigma^{-1} \mathbf{z}_i\right) \mathbf{z}_i \mathbf{z}_i^H,
$$

where $u : [0, \infty) \rightarrow \mathbb{R}$ is a weighting function.

- A robust weighting function $u(\cdot)$ is descending to zero, i.e. highly deviating observation $\mathbf{z}_i$ with large $\|\Sigma^{-1/2} \mathbf{z}_i\|^2 = \mathbf{z}_i^H \Sigma^{-1} \mathbf{z}_i$ is given less weight.
We call $t_\nu M$-estimator the $M$-estimator obtained with the weight function

$$u(s) = u_\nu (s) = \frac{2K + \nu}{\nu + 2s},$$

where $\nu > 0$ (and $K$ is the dimension of the data). This $M$-estimator is also the ML-estimator if the data comes from complex $K$-variate $t_\nu$-distribution.

- Complex $t_\nu$-distribution belongs to a wider class of Complex Elliptically Symmetric (CES) distributions.

- Random vector from complex $t_\nu$-distribution has pdf

$$f_\nu (z) \propto |\Sigma|^{-1} \left\{ 1 + 2z\Sigma^{-1}z/\nu \right\}^{-(2K+\nu)/2}$$

- $t_\nu$-distribution with $\nu = 1$ is called multivariate complex Cauchy distribution which is a heavy tailed alternative for the multivariate Gaussian distribution.
COMPUTATION: The $t_\nu M$-estimate can be calculated by a convenient iterative algorithm: Given an initial estimate, a $K \times K$ PDH matrix $\Sigma_0$, the iterations

$$
\Sigma_{m+1} = \frac{1}{N} \sum_{i=1}^{N} u_\nu(z_i^H \Sigma_m^{-1} z_i) z_i z_i^H
$$

converge to the unique solution $\hat{\Sigma}$ under mild reg. cond. on data.

In practical implementation of the proposed algorithm we used as termination point

$$
\varepsilon_{m+1} = \| I - \Sigma_m^{-1} \Sigma_{m+1} \| < \varepsilon,
$$

where $\| \cdot \|$ is some matrix norm and $\varepsilon$ is some predetermined tolerance level, e.g. $\varepsilon = 0.001$.

The computational complexity is $O(N_{\text{iter}} NK^2)$, where $N_{\text{iter}}$ is the number of iteration.
The complex analogue of $M$-estimator by [Tyler, 1987] can be constructed with the weight function $u(s) = K/s$ ($K$ is the dimension of the data), that is, Complex Tyler’s $M$-estimator $\hat{\Sigma}$ satisfy

$$\hat{\Sigma} = \frac{K}{N} \sum_{i=1}^{N} \frac{z_i z_i^H}{z_i^H \hat{\Sigma}^{-1} z_i}.$$ 

We need to exclude all $z_i = 0$. 
Examples of weight functions. The lack of robustness of sample covariance matrix is obvious.
Root-MUSIC signal zeros (for 15 simulated samples) in the Cauchy noise environment. Noise subspace is estimated using the sample cov. matrix (left picture) and $\mathcal{M}$-estimator with Cauchy weights (right picture). Setup: 4-element ULA, two Gaussian (20dB) signals at $15^\circ$ and $25^\circ$. 
In the classical stochastic ML approach the array output vector $z$ is modeled as complex Gaussian with cov. matrix $R = A(\theta)\Omega A(\theta)^H + \sigma^2 I$, where $\Omega$ is the signal cov. matrix and $\sigma^2$ is the noise variance.

The signal parameters are then found by solving

$$\{\hat{\theta}, \hat{\Omega}, \hat{\sigma}^2\} = \arg\min_{\theta, \Omega, \sigma^2} \{\log[\det(R)] + Tr[R^{-1}\hat{R}]\},$$

where $\hat{R} = \frac{1}{N} \sum_{i=1}^{N} z_i z_i^H$. 

To obtain robust estimates of the DOAs, we use the complex $t_{\nu}$-distribution as an array output model distribution and find robust DOA estimate $\hat{\theta}$ by solving

$$\{\hat{\theta}, \hat{\Omega}, \hat{\sigma^2}\} = \arg\min_{\theta,\Omega,\sigma^2} \left\{ \log[\det(R)] + \frac{2K + \nu}{2N} \sum_{i=1}^{N} \log\left(1 + 2z_i^H R^{-1} z_i / \nu \right) \right\}.$$ 

where $\hat{\sigma^2}$ is the estimate of noise scatter (variance not necessarily defined).
The MUSIC method is noise-subspace DOA estimation technique. We now compare the quality of the MUSIC pseudospectrum

\[
V(\theta) = \frac{1}{a^H(\theta)\hat{\Gamma}_N\hat{\Gamma}_N^H a(\theta)}.
\]

associated with noise subspace estimate $\hat{\Gamma}_N$ based on different covariance matrix estimators by simulations. Here $a(\theta)$ is the array steering vector for the Uniform Linear Array (ULA).

Simulation setup:

- ULA, two uncorrelated signals at $\theta_1 = -2^\circ$ and $\theta_2 = 2^\circ$ with SNR 20dB (same dispersion parameter for complex Cauchy $(S\alpha S)$ distribution)
- $N = 300$ snapshots, $M = 8$ element array with $\lambda/2$ element spacing.
- data sets $z_1, \ldots, z_N$ generated from complex Gaussian and Cauchy distribution.
Complex Valued Multichannel Data
MUSIC pseudospectrum, Gaussian case

Figure: a) sample cov. b) $t_1 M$-, c) Huber’s $M$- and d) Cov. matrix via ML-estimator.
Figure: a) sample cov. b) $t_1 M$-, c) Huber’s $M$- and d) Cov. matrix via ML-estimator.
Complex Valued Multichannel Data

Efficiencies for complex Gaussian and t(5)-distr.

The efficiencies of the covariance matrix \( E[zz^H] \) estimators corresponding to the limiting distribution of standardized array covariance matrix estimators at the complex normal and complex \( t \)-distribution with 5 deg. of freedom. \( k = 2, 5, 10 \) refers to array size.

<table>
<thead>
<tr>
<th></th>
<th>Normal distr.</th>
<th></th>
<th>t(5)-distr.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( k = 2 )</td>
<td>( k = 5 )</td>
<td>( k = 10 )</td>
</tr>
<tr>
<td>SampleCM ( \tau )</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Huber ( \tau )</td>
<td>1.041</td>
<td>1.014</td>
<td>1.006</td>
</tr>
<tr>
<td>Cau ( \tau )</td>
<td>1.301</td>
<td>1.161</td>
<td>1.090</td>
</tr>
</tbody>
</table>
Conclusions

- Robustness and Complex Multivariate signals were considered.
- The concepts of qualitative and quantitative robustness were described.
- Influence functions for eigenvalues and eigenvectors were found.
- Robust procedures for covariance matrix estimation were presented (M-estimation, ML estimation).
- DoA estimation using ULA was used as an example.
Robust Methods for Dependent Data
Robustness for Dependent Data

- research on robustness for non-i.i.d. case limited
- first contributions in the mid 80’s, mainly theoretical
- not much progress for a long time
- recently, research significantly increased
- novel estimators proposed and applied to engineering problems, e.g.
  - robust discovery of periodically expressed genes [Liang et al., 2009]
  - robust electricity consumption forecasting [Chakhchoukh et al., 2010]
  - robust motion artifact cancellation in ECG signals [Strasser et al., 2012]
Dependent Data

Introduction

Difficulty Introduced by the Correlation:

- different definitions of robustness measures from the i.i.d. case
- several types of outliers: additive, innovative, patchy, isolated, ...
- estimators need to be adapted to deal with correlation

Most Popular Model for Dependent Data:

- as in classical signal processing for dependent data: AutoRegressive Moving Average (ARMA)

→ **Focus in This Tutorial**: robust methods for ARMA models
  
  Note: robust model selection is not considered (see references)
Example: Electricity Consumption Forecasting [Chakhchoukh et al., 2010]

- consumption data first corrected from weather influence (temperature, nebulosity) by regression model whose parameters must be estimated robustly
- residual series encompasses (daily, weekly, yearly) seasonalities, modeled as Seasonal Autoregressive Integrated Moving Average (SARIMA)

- one-week difference in consumption in France at 7:00 a.m. in 2005
- outliers due to special days and events (marked in red)
Example: Electricity Consumption Forecasting [Chakhchoukh et al., 2010]

- evaluation of the one-day ahead forecasting quality

Mean Absolute Percentage (Prediction) Error
- lead times are in half-hourly intervals from (0:00, 0:30, ..., 23:30)
- sophisticated robust forecasts: robust double exponential smoothing (DEXR) and Filtered $\tau$ (explained later) outperform the cleaned maximum likelihood (ML) which is used by the majority of electricity companies.
Consider an ARMA($p,q$) Process:

$$X_n = Z_n - \sum_{k=1}^{p} a_k X_{n-k} + \sum_{k=1}^{q} b_k Z_{n-k}, \quad n \in \mathbb{Z}$$

where $Z_n$ are independent and identically distributed (i.i.d.) Gaussian random variables.

**Classically Assumed:**

- observations given by $y_n = x_n$
- in this case, we call $Y_n$ a “clean ARMA process”

→ In the following: different outlier models and their effects on parameter estimation
Dependent Data
Which Types of Outliers Do We Face in ARMA Models?

Different Outliers in an ARMA\((p,q)\) Process:

\[
X_n = Z_n - \sum_{k=1}^{p} a_k X_{n-k} + \sum_{k=1}^{q} b_k Z_{n-k}, \quad n \in \mathbb{Z}
\]

1. **Additive Outliers:**

The process \(Y_n\) is given by

\[
Y_n = X_n + \xi_n \varepsilon V_n
\]

\(\xi_n \varepsilon = 1\) with probability \(\varepsilon\) and \(\xi_n \varepsilon = 0\) with probability \((1 - \varepsilon)\).

\(X_n\) and \(V_n\) are independent processes \(\forall n \in \mathbb{Z}\).
Different Outliers in an ARMA\((p, q)\) Process:

\[
X_n = Z_n - \sum_{k=1}^{p} a_k X_{n-k} + \sum_{k=1}^{q} b_k Z_{n-k}, \quad n \in \mathbb{Z}
\]

2. Replacement Outliers:

The process \(Y_n\) is given by

\[
Y_n = (1 - \xi^\varepsilon_n) X_n + \xi^\varepsilon_n W_n
\]

\(\xi^\varepsilon_n = 1\) with probability \(\varepsilon\) and \(\xi^\varepsilon_n = 0\) with probability \((1 - \varepsilon)\).

\(X_n\) and \(W_n\) are independent \(\forall n \in \mathbb{Z}\).

Usually, no knowledge is assumed about the distribution of \(W_n\).
Dependent Data
Which Types of Outliers Do We Face in ARMA Models?

Different Outliers in an ARMA($p,q$) Process:

\[ X_n = Z_n - \sum_{k=1}^{p} a_k X_{n-k} + \sum_{k=1}^{q} b_k Z_{n-k}, \quad n \in \mathbb{Z} \]

3. **Innovation Outliers:** \( Z_n \) contains outliers, e.g.
   - because distribution of \( Z_n \) is heavy tailed
   - \( Z_n \) is contaminated by additive outliers:
     \[ Z_n^e = Z_n + \xi_n^e V_n \]
   - \( Z_n \) is contaminated by replacement outliers:
     \[ Z_n^e = (1 - \xi_n^e) Z_n + \xi_n^e W_n \]
Example: Outliers in an AR(1) process

**Example: AR(1) Process:**

- $X_n = Z_n + 0.5X_{n-1}, \quad n = 1, \ldots, 250$
  
  $Z_n$ are zero mean i.i.d. Gaussian random variables with $\sigma_Z = 1$

- observations given by $y_{100} = x_{100} + 10, \ y_{150} = x_{150} + 10, \ y_n = x_n, \text{otherwise}$

→ two additive outliers
Dependent Data
Example: Outliers in an AR(1) process

Representation of the AR(1) Process With Two Additive Outliers as a Regression:

→ four outliers in regression representation: two vertical outliers and two 'bad' leverage points
→ regression M-estimates non-robust
→ for $p > 1$ or ARMA, even highly robust i.i.d. regression estimators break down (reasons shown later)
Dependent Data

Example: Outliers in an AR(1) process

Now Consider Two Innovations Outliers:

\[ z_{100} = z_{100} + 10, \quad z_{150} = z_{150} + 10 \]

and the corresponding regression

\[ x_n = 0.5x_{n-1} \]

→ two vertical outliers, but several 'good' leverage points, which lie on the tangent given by \( x_n = 0.5x_{n-1} \).

→ regression M-estimates robust

→ even classical estimators give good results, since the 'good leverage points' compensate for the vertical outliers
For Dependent Data, Outliers Can Inhibit Temporal Structure:

- isolated outliers:
  \( \xi_n, n \in \mathbb{Z} \) takes value 1 such that at least one non-outlying observation is between two outliers (e.g. follows an independent Bernoulli-distribution)

- patchy outliers:
  \( \xi_n, n \in \mathbb{Z} \) takes value 1 for \( k \leq N/2 \) subsequent values (patches)
  - switching between LOS/NLOS environments in geolocation position estimation
  - loose connection in sensor creates patchy impulsive noise
  - motion artifacts in patient movement for biomedical measurements

- seasonal outliers, e.g.
  - in average rainfall data
Example: Electrocardiogram (ECG) Measurement Where Arm Movement Creates Additive Patchy Outlier

- cleaned signal = measured signal - estimated artifacts

- artifacts estimated with method by [Strasser et al., 2012]
- recorded at Signal Processing Group Laboratory TU Darmstadt
  (Matlab code freely available at http://www.spg.tu-darmstadt.de/res/dl)
Dependent Data

Robustness Measures for Dependent Data

Difficulty with Robustness Theory for Dependent Data:

- extension of theoretical robustness measures from i.i.d. case is not straightforward and difficult to evaluate
- robustness measures depend on the type of outliers

In the Following:
brief summary of the most important differences for the breakdown point, influence function and maximum bias curve.
1. **Breakdown Point (BP)**

   - parameter space generally bounded
   - effect of outliers more complicated than in i.i.d. case

   e.g. for an AR(1) model with additive outliers, breakdown occurs when the contamination drives the parameter estimate to 0

---

**General Intuitive Definition of the BP** [Genton and Lucas, 2003]

Given some percentage of contamination $\epsilon$, breakdown occurs when increasing $\epsilon$ does not further enlarge the range of values taken by the estimate over the contamination neighborhood.
2. **Influence Function (IF)**

- two definitions exist [Künsch, 1984], [Martin and Yohai, 1986], which are mathematically related
- influence functions depends on the type of outliers, e.g. are different for patchy and isolated outliers
- correlation makes computation of joint probability distributions (necessary for derivation of IF’s) nearly impossible, except for the AR(1) and MA(1) models
3. **Maximum Bias Curve**
   - same definition holds as in the i.i.d. case
   - MBC depends on the type of outliers
   - **However**: more complex and harder to compute, e.g. for an AR(1) process joint distribution of \((Y_n, Y_{n-1})\) needs to be considered.
   - **In Practice**: MBC is generally obtained using Monte Carlo simulations

→ **Next**: Some robust estimators for ARMA\((p,q)\) Models
Robust Methods for Dependent Data
Some Estimators for ARMA\((p,q)\) Models

Cleaned Maximum Likelihood Estimator (MLE):

- most popular method among practicing engineers
- Idea: first reject all samples with value beyond three times the known or estimated standard deviation (3–\(\sigma\) rejection rule), then apply a maximum likelihood approach that handles missing data
- simple in terms of computation and implementation
- when estimated non-robustly inflated \(\hat{\sigma}\) leads to masking effect: large outliers mask small ones
- even when \(\sigma\) is estimated robustly, ignoring the correlation structure in the data leads to false decisions in the outlier rejection and to imprecise parameter estimates

→ How about adapting robust estimates designed for i.i.d.?
Example: ARMA(2,1) With Additive Outliers:

- $a = (-0.39, -0.3)^T$, $b = 0.9$, additive outliers equally spaced with amplitude 10

- black: innovations process $Z_n$, red: additive outlier positions, blue: residuals using ARMA model fit

Robust estimates designed for i.i.d. case can only be used for ARMA($p$, $q$) models, if they include a mechanism that prevents the ‘smearing’ of a single outlier onto multiple residuals!
Two possibilities exist to prevent “smearing”:

1. estimates that involve a robust filtering operation
2. Bounded Influence Propagation (BIP) estimates

In both cases the propagation of the outlier’s influence is bounded.
→ robust i.i.d. regression estimates (e.g. S, τ, MM, can be adapted)

In the following:
How to compute these estimates?
State-Space Representation of AR(p) Process:

- state equation:
  \[ X_n = AX_{n-1} + Z_n \]

- non-observable \( p \)-dimensional state vector: \( X_n = [X_n, X_{n-1}, \cdots, X_{n-p+1}]^T \)

- innovations: \( Z_n = [Z_n, 0, \cdots, 0]^T \)

- state transition matrix:
  \[
  A = \begin{bmatrix}
    a_1 & \cdots & a_{p-1} & a_p \\
    1 & \cdots & 0 & 0 \\
    \vdots & \ddots & \vdots & \vdots \\
    0 & \cdots & 1 & 0
  \end{bmatrix}
  \]

- measurement equation:
  \[ Y_n = X_n + V_n \]

- \( V_n \) and \( Z_n \) are independently distributed
Robust Filtering

Principle of Robust Filter

- Robust Initial State
  - find robust estimates $\hat{X}_{0|0}$ and $\hat{\Sigma}_{0|0}$

- Prediction Step
  - predict the next state vector $\hat{X}_{n|n-1}$
  - compute the prediction error $\hat{V}_n$
  - compute the prediction error covariance matrix $\hat{\Sigma}_{n|n-1}$

- Robust Correction Step
  - compute the robust state est. $\hat{X}_{n|n}$
  - compute the robust correction error covariance matrix $\hat{\Sigma}_{n|n}$

$x_n = AX_{n-1} + Z_n$
$y_n = CX_n + V_n$

Time step $n$

Time step $n + 1$
Approximate Conditional Mean (ACM) Filter [Masreliez 1975], [Martin 1982]:
computes robustly filtered estimate

\[
\hat{X}_{n|n} = A\hat{X}_{n-1|n-1} + \frac{\hat{\Sigma}_{1,n}}{\hat{\sigma}_n} \psi \left( \frac{Y_n - \hat{Y}_{n|n-1}}{\hat{\sigma}_n} \right)
\]

which is an approximation of \(E[X_n|Y_1, Y_2, \ldots, Y_n]\)

- \(\psi(\cdot)\) is an odd, bounded and continuous score function [Maronna et al., 2006]
- \(\hat{\Sigma}_{1,n}\) is the first column of \(\hat{\Sigma}_n\) (prediction error covariance matrix), which is computed recursively. \(\hat{\sigma}^2_n\) is the first element of \(\hat{\Sigma}_{1,n}\)
- \(\hat{Y}_{n|n-1}\) is the robust one step ahead predictor of \(Y_n\) based on \(\{Y_1, \ldots, Y_{n-1}\}\)

\[
\hat{Y}_{n|n-1} = (A\hat{X}_{n-1|n-1})_1
\]

- For a detailed description of the algorithm, see e.g. [Spangl & Dutter, 2007]
Robust Methods for Dependent Data
Robust Filtering of an AR(\(p\)) Process

Example: Filter-Cleaned Process AR(2)

- \(a = (0.5, 0.3)^T\) 10 % additive outliers equally spaced with amplitude 10

- (left plot) black: clean process, red: process contaminated by additive outliers
- (right plot) black: clean process, red: robustly filtered (cleaned) process

\[\hat{Y}_{n|n-1} = (A\hat{X}_{n-1|n-1})_1\]
Example: Filter-Cleaned Process AR(2)

- filtered residual at time $n$: $\tilde{Z}_n = Y_n - [a_1 \ a_2 \ldots \ a_p]^T \hat{X}_{n-1|n-1}$

- black: innovations process $Z_n$, blue: filtered residuals (effect of outlier bounded to outlier position)

- use filtered residuals for robust parameter estimation
Robust Methods for Dependent Data
Bounded Influence Propagation

Bounded Innovation Propagation Autoregressive Moving Average Model (BIP-ARMA) [Muler et al., 2009]

\[ X_n = Z_n - \sum_{k=1}^{p} a_k X_{n-k} + \sum_{k=1}^{r} \left( a_k Z_{n-k} + (b_k - a_k)\sigma \psi \left( \frac{Z_{n-k}}{\sigma} \right) \right) \]

- auxiliary model to robustly estimate ARMA models
- ARMA models are included by setting \( \psi(x) = x \)
- \( r = \max(p, q) \), if \( r > p \), \( a_{p+1} = \ldots = a_r = 0 \), while if \( r > q \), \( b_{q+1} = \ldots = b_r = 0 \)
- \( \sigma \) is a robust M-scale of \( Z_n \)
- \( \psi(x) \) is an odd, bounded and continuous function
Robust Methods for Dependent Data
Bounded Influence Propagation Model

Example: ARMA(2,1) With Additive Outliers:

- $a = (-0.39, -0.3)^T$, $b = 0.9$, additive outliers equally spaced with amplitude 10

black: true innovations process $Z_n$, red: additive outlier positions, left plot blue: residuals from ARMA, right plot blue: residuals from BIP-ARMA

→ one additive outlier contaminates less samples (ideally only one) for the BIP-ARMA model
Relation of BIP-ARMA Estimates to Robust Filter-Cleaner Based Estimates

Essentially: BIP-ARMA model can be viewed as robust filter cleaner, where cleaned process is given by:

\[ \tilde{X}_n = Y_n - \tilde{Z}_n + \sigma_{\tilde{Z}} \psi \left( \frac{\tilde{Z}_n}{\sigma_{\tilde{Z}}} \right), \quad n = 1, \ldots, N \]

- \( \tilde{Z}_n \) is the residual from the BIP-ARMA model
- \( \sigma_{\tilde{Z}} \) is a robust scale of \( \tilde{Z}_n \)

Advantages of the BIP-ARMA Estimates over Filtered ARMA Estimates

- consistent
- tractable asymptotic expression and confidence intervals
- influence function, maximum bias curve, breakdown and efficiency computable
Intracranial Pressure (ICP)

![ICP measurement graph]

four hour excerpt of an ICP measurement

artifacts and nonstationary signal
Intracranial Pressure (ICP)

excerpt of ICP signal

intrinsic mode functions and residual

nonstationary $\rightarrow$ empirical mode decomposition $\rightarrow$ intrinsic mode functions (IMF)
Robust Estimation Application Example
Artifact Removal in Intracranial Pressure Signal

Intracranial Pressure (ICP)

- ARMA BIP-$\tau$ parameter estimation and robust model order selection [Muma,2014]
  $\rightarrow$ ARMA(2,1)
Robust Estimation Application Example
Artifact Removal in Intracranial Pressure Signal

Intracranial Pressure (ICP)

- first intrinsic mode function (IMF)

- ARMA BIP-τ artifact removal
Intracranial Pressure (ICP)

![Graph showing IMF1 and IMF1 cleaned over time](image)

- first intrinsic mode function (IMF)
- ARMA BIP-$\tau$ artifact removal
Intracranial Pressure (ICP)

![ICP graph]

- Four hour excerpt of an ICP measurement
- ARMA BIP-τ artifact removal for all IMFs → back transform
Robust Estimation Application Example
Artifact Removal in Intracranial Pressure Signal

Intracranial Pressure (ICP)

![Graph showing ICP measured and cleaned over time.]

four hour excerpt of an ICP measurement

- ARMA BIP-\(\tau\) artifact removal for all IMFs \(\rightarrow\) back transform

M. Muma, A.M. Zoubir, Robust ARMA Parameter Estimation by Bounded Innovation \(\tau\)–estimation., submitted.
Recent Trends in Robust Signal Processing

Advances in Robust Detection
Outline

- Robust Hypothesis Testing
  - Robust Hypothesis Testing with $\alpha$–divergence
- Robust Distributed Hypothesis Testing
  - Robust Decentralized Hypothesis Testing
- Robust Sequential Detection (Outlook)
Motivation
Robustness in Engineering World

Forest Fire Detection

Source: Matt Heavner, University of Alaska Southeast
Motivation
Distributed Hypothesis Testing

Definition:
Detection of events of interest with the help of geographically distributed sensors

Advantages:
- Detection accuracy increases with the number of sensors
- Failure of a single sensor does not jeopardize the sensor network
- Possible to infer the state of events in different geographical locations
Motivation
Mathematical Robustness Formulation

Optimum decision making:
Statistical model is correct

\[
\hat{\delta} = \arg \min_{\delta} f(\delta) \quad \delta: \text{decision rule}
\]

Robust decision making:
Statistical model includes uncertainties

\[
(\hat{\delta}, \hat{P}_0, \hat{P}_1) = \arg \min_{\delta} \max_{P_0, P_1} f(\delta, P_0, P_1) \quad P_0, P_1: \text{distributions}
\]

Robust distributed decision making:
Statistical model includes uncertainties for every sensor

\[
(\hat{\delta}, \hat{\gamma}, \hat{P}_0, \hat{P}_1) = \arg \min_{\delta, \gamma} \max_{P_0, P_1} f(\delta, \gamma, P_0, P_1) \quad \gamma: \text{fusion rule}
\]
Outline

- Robust Hypothesis Testing
  - Robust Hypothesis Testing with $\alpha$–divergence
- Robust Distributed Hypothesis Testing
  - Robust Decentralized Hypothesis Testing
- Robust Sequential Detection (Outlook)
A ROBUST VERSION OF THE PROBABILITY RATIO TEST

By Peter J. Huber

Swiss Federal Institute of Technology, Zurich

1. Introduction and summary. A statistical procedure is called robust, if its performance is insensitive to small deviations of the actual situation from the idealized theoretical model. In particular, a robust procedure should be insensitive to the presence of a few "bad" observations; that is, a small minority of the observations should never be able to override the evidence of the majority. (But at the same time the discordant minority might be a prime source of information for improving the theoretical model.)

The classical probability ratio test is not robust in this sense: a single factor \( p_i(x_i)/p_0(x_i) \) equal (or almost equal) to 0 or \( \infty \) may upset the test statistic \( T(x) = \prod_i p_i(x_i)/p_0(x_i) \). This leads to the conjecture that appropriate robust substitutes to both fixed sample size and sequential probability ratio tests might be obtained by censoring the single factors at some fixed numbers \( c' < c \). Thus, one would replace the test statistic by \( T'(x) = \prod_i \delta(x_i) \), where \( \delta(x_i) = \max \{ c', \min (c', p_i(x_i)/p_0(x_i)) \} \).

The problem of robustly testing a simple hypothesis \( P_0 \) against a simple alternative \( P_1 \) may be formalized by assuming that the true underlying distribution lies in some neighborhood of either of the idealized model distributions \( P_0 \) or \( P_1 \).

The present paper exhibits two different types of such neighborhoods for which the above mentioned test, to be called censored probability ratio test, is most robust in a well defined minimax sense.

The problem solved here originated through the earlier paper Huber (1964), over the question how to test hypotheses about the mean of contaminated normal distributions.

2. Setup of the problem. Let \((\mathfrak{X}, \mathfrak{A})\) be a measurable space, and let \( P_0 \), \( P_1 \) be two distinct probability measures on it, having densities \( p_0 \), \( p_1 \) with respect to some measure \( \mu \), e.g., \( \mu = \Theta P_0 + P_1 \). In order to formalize the possibility of unknown small deviations from the idealized models \( P_i \) we blow them up to composite hypotheses

\[
\Theta_i = [Q | Q = (1 - \epsilon_i)P_0 + \epsilon_iH_i, H_i \in \mathfrak{X}], \quad \epsilon_i \geq 0, \quad (i = 0, 1),
\]

where \( 0 \leq \epsilon_i < 1 \) are fixed numbers, and \( \mathfrak{X} \) denotes the class of all probability measures on \((\mathfrak{X}, \mathfrak{A})\). We shall always assume that \( \Theta_0 \) and \( \Theta_1 \) do not overlap (c.f. the remark after Lemma 2 below).

Let \( \varphi \) be any test between \( \Theta_0 \) and \( \Theta_1 \), rejecting \( \Theta_1 \) with conditional probability \( \delta(x) \), given that \( x = (x_1, \ldots, x_n) \) has been observed. Assume that a loss \( L_i > 0 \) is incurred if \( \varphi \) is falsely rejected, then the expected loss, or risk, is \( R(Q, \varphi) \)
Robust Hypothesis Testing
Uncertainty Sets

Outliers:

Model Mismatch:

\[
\begin{align*}
\{f, g\} & \quad \text{Correct measurements} \\
& \quad \text{Biased measurements}
\end{align*}
\]
Abstract—This paper considers the design of a minimax test for two hypotheses where the nominal probability density of the observa-
tion are located in neighborhoods of the probability density, thereby minimizing any deviation from the nominal density. The minimax test is formulated as a saddle-point problem which is characterized.

The robust best Available transformation which minimizes the nominal likelihood ratio in the vicinity of a saddle point is illustrated by considering the transmission of binary data in the pres-
cence of additive noise.

Index Terms—Robust hypothesis testing, saddle point, minimax problem, concave optimization, saddle point.

I. INTRODUCTION

Robust hypothesis testing and signal detection problems have been examined in detail over the last 40 years [1], [2]. The purpose of such studies is to design tests or detectors which are insensitive to modeling errors. Specifically, whereas standard Bayesian or Neyman-Pearson tests are designed for nominal observation probability density functions, their performance may degrade rapidly when the actual model deviates only moderately from the nominal model. To guard against modelling errors, a minimax framework is usually adopted for selecting tests or detectors. In this context, the goal is to develop a test that maximizes the worst-case performance for all observation models in a properly specified neighborhood of the nominal model. For robust hypothesis testing, the neighborhood of the nominal model under which hypothesis corresponds either to a contamination model or a perturbation model. A common measure for the nominal model is the Kullback-Leibler metric or a variant thereof. (Robust) showed that the minimax detector applies a clipping transformation to the nominal likelihood ratio function. The clipping effect is achieved by clamping small portions of the probability mass under each hypothesis to the tail portions where the density is low. Any relaxation of this assumption results in a significant degradation in test performance.

We adopt here a minimax formulation of the robust hypothesis test problem that yields the same type of [15]-[17]. The only difference is that the neighborhood of the actual observation probability density is located under each hypothesis is defined by placing an upper bound on the relative entropy of the actual density with respect to the nominal density. To justify this choice of the relative entropy as a measure of proximity between statistical models, observe that Fisher’s work addresses primarily situations where statistical model is obtained directly from imperfect data, possibly contaminated by outliers. However, there exist also situations where the density estimated from hypo-
thesis testing are model based, arising from physical consider-
tions, possibly with a few unknown parameters which are esti-
mated from the data. In this context, the relative entropy is a natural metric for model mismatch, since it provides the under-
dlying metric for establishing the convergence of the expecta-
tion-maximization method [18]. In statistical reasoning, a test from a differential geometric viewpoint, it is argued in [19] that the relative entropy is a measure of the distance between statistical model.

More recently, in the context of estimation and detection, it was shown in [7], [8] that minimum filters based on a rela-
tive entropy tolerance take the form of minimax filters, which have well-known robustness properties. By selecting the relative entropy as a measure of model mismatch, a relative entropy metric was also adopted recently in [9] as an alternative to robust estimators.

Robust hypothesis testing problems have been examined in detail over the last 40 years [1], [2]. The purpose of such studies is to design tests or detectors which are insensitive to modeling errors. Specifically, whereas standard Bayesian or Neyman-Pearson tests are designed for nominal observation probability density functions, their performance may degrade rapidly when the actual model deviates only moderately from the nominal model. To guard against modelling errors, a minimax framework is usually adopted for selecting tests or detectors. In this context, the goal is to develop a test that maximizes the worst-case performance for all observation models in a properly specified neighborhood of the nominal model. For robust hypothesis testing, the neighborhood of the nominal model under which hypothesis corresponds either to a contamination model or a perturbation model. A common measure for the nominal model is the Kullback-Leibler metric or a variant thereof. (Robust) showed that the minimax detector applies a clipping transformation to the nominal likelihood ratio function. The clipping effect is achieved by clamping small portions of the probability mass under each hypothesis to the tail portions where the density is low. Any relaxation of this assumption results in a significant degradation in test performance.

We adopt here a minimax formulation of the robust hypothesis test problem that yields the same type of [15]-[17]. The only difference is that the neighborhood of the actual observation probability density is located under each hypothesis is defined by placing an upper bound on the relative entropy of the actual density with respect to the nominal density. To justify this choice of the relative entropy as a measure of proximity between statistical models, observe that Fisher’s work addresses primarily situations where statistical model is obtained directly from imperfect data, possibly contaminated by outliers. However, there exist also situations where the density estimated from hypo-
thesis testing are model based, arising from physical consider-
tions, possibly with a few unknown parameters which are esti-
mated from the data. In this context, the relative entropy is a natural metric for model mismatch, since it provides the under-
dlying metric for establishing the convergence of the expecta-
tion-maximization method [18]. In statistical reasoning, a test from a differential geometric viewpoint, it is argued in [19] that the relative entropy is a measure of the distance between statistical model.

More recently, in the context of estimation and detection, it was shown in [7], [8] that minimum filters based on a rela-
tive entropy tolerance take the form of minimax filters, which have well-known robustness properties. By selecting the relative entropy as a measure of model mismatch, a relative entropy metric was also adopted recently in [9] as an alternative to robust estimators.
State-of-the-Art:

- A minimax robust hypothesis testing scheme based on KL-divergence
- Valid **if and only if**
  - Nominal probability density functions are symmetric, i.e. $f_0(y) = f_1(-y) \forall y$
  - The nominal likelihood ratio function $l = f_1/f_0$ is increasing
  - Robustness parameters are equal $\epsilon_0 = \epsilon_1$
  - Nominal likelihood ratio threshold equals 1, i.e. $\rho = 1$

Recent Result:

- Minimax robust hypothesis testing scheme based on general $\alpha$-divergence
- Valid for asymmetric distributions and “degrees of robustness”
- Arbitrary likelihood ratio thresholds, i.e. $\rho \neq 1$
Robust Hypothesis Testing

Minimax Robust Test with the $\alpha$–divergence

The $\alpha$-divergence:

$$D(g, f ; \alpha) = \frac{1}{\alpha(1 - \alpha)} \left(1 - \int_{\mathbb{R}} \left(\frac{g}{f}\right)^\alpha f \, d\mu\right), \quad \alpha \in \mathbb{R} \setminus \{0, 1\}$$

- A smooth distance for every $\alpha$
- Includes many commonly used distances as special cases
  - $\alpha = 2$: $\chi^2$ distance
  - $\alpha = 1/2$: squared Hellinger distance
  - $\alpha = -1$: inverse Pearson distance
  - $\alpha \to 1$: KL-divergence
  - ...

Robust Hypothesis Testing
Minimax Robust Test with $\alpha$–divergence

Robust detection:

Minimax Robust Test Design

Problem Definition:

Maximization:
\[
\hat{g}_0 = \arg \sup_{g_0 \in \mathcal{G}_0} P_F(\delta, g_0) \quad \text{s.t.} \quad g_0 > 0, \quad \int g_0 \, d\mu = 1
\]
\[
\hat{g}_1 = \arg \sup_{g_1 \in \mathcal{G}_1} P_M(\delta, g_1) \quad \text{s.t.} \quad g_1 > 0, \quad \int g_1 \, d\mu = 1
\]

Minimization:
\[
\hat{\delta} = \arg \min_{\delta \in \Delta} P_E(\delta, \hat{g}_0, \hat{g}_1)
\]
Robust Hypothesis Testing
Minimax Robust Test with the $\alpha$-divergence

Uncertainty sets:
Robust Hypothesis Testing

Minimax Robust Test with the $\alpha$–divergence

Robust detection:

\[ f_0, f_1 \rightarrow \text{Minimax Robust Test Design} \rightarrow \hat{\delta}, \hat{g}_0, \hat{g}_1 \]

Problem Definition:

Maximization:
\[
\hat{g}_0 = \arg \sup_{g_0 \in G_0} P_F(\delta, g_0) \quad \text{s.t. } g_0 > 0, \int g_0 \, d\mu = 1 \\
\hat{g}_1 = \arg \sup_{g_1 \in G_1} P_M(\delta, g_1) \quad \text{s.t. } g_1 > 0, \int g_1 \, d\mu = 1
\]

Minimization:
\[
\hat{\delta} = \arg \min_{\delta \in \Delta} P_E(\delta, \hat{g}_0, \hat{g}_1)
\]
Robust Hypothesis Testing

Minimax Robust Test with the $\alpha$-divergence

The solutions of the minimax optimization problem:

$$\hat{\delta}(y) = \begin{cases} 0, & l(y) < \rho l_l \\ \frac{l^\alpha - 1(l(y)/\rho)^{1-\alpha} - 1}{(l^\alpha - 1(k u)^{\alpha-1})(l(y)/\rho)^{1-\alpha+k^{\alpha-1}1}}, & \rho l_l \leq l(y) \leq \rho l_u \\ 1, & l(y) > \rho l_u \end{cases}$$

$$\hat{g}_0(y) = \begin{cases} \frac{l}{z} f_0(y), & l(y) < \rho l_l \\ \Phi_0 f_0(y), & \rho l_l \leq l(y) \leq \rho l_u \\ k \frac{l u}{z} f_0(y), & l(y) > \rho l_u \end{cases}$$

$$\hat{g}_1(y) = \begin{cases} \frac{1}{z} f_1(y), & l(y) < \rho l_l \\ \rho \Phi_1 f_1(y), & \rho l_l \leq l(y) \leq \rho l_u \\ k \frac{1}{z} f_1(y), & l(y) > \rho l_u \end{cases}$$
Robust Hypothesis Testing
Minimax Robust Test with the $\alpha$–divergence

Robust likelihood ratio:

$$\hat{l}(y) = \frac{\hat{g}_1(y)}{\hat{g}_0(y)} = \begin{cases} \frac{1}{l_1} l(y), & l(y) < \rho_l \\ \rho, & \rho_l \leq l(y) \leq \rho_u \\ \frac{1}{l_u} l(y), & l(y) > \rho_u \end{cases}$$
Robust Hypothesis Testing

Minimax Robust Test with the $\alpha$–divergence

$l_l$, $l_u$, $\Phi_{\{0,1\}}$, $z$ and $k$ are determined by non-linear equations:

$$
\frac{1}{z^\alpha} \left( l_l^\alpha \int_{I_1} f_0 \, d\mu + \int_{I_2} z \Phi_0^\alpha f_0 \, d\mu + (kl_u)^\alpha \int_{I_3} f_0 \, d\mu \right) = \alpha(1 - \alpha)\epsilon_0
$$

$$
\frac{1}{z^\alpha} \left( \int_{I_1} f_1 \, d\mu + \int_{I_2} z \Phi_1^\alpha f_1 \, d\mu + k^\alpha \int_{I_3} f_1 \, d\mu \right) = \alpha(1 - \alpha)\epsilon_1
$$

with

$$
I_1 = \{l(y) < \rho l_l\}
$$

$$
I_2 = \{\rho l_l \leq l(y) \leq \rho l_u\}
$$

$$
I_3 = \{l(y) > \rho l_u\}
$$

Simulation setup:

- $f_0 \sim \mathcal{N}(-1, 1)$, $f_1 \sim \mathcal{N}(1, 2)$
- $\epsilon_0 = 0.02$ and $\epsilon_1 = 0.01$, $\alpha = 3$, $\rho = 1$
Robust Hypothesis Testing
Least Favorable Densities

Simulation setup:

- $f_0 \sim \mathcal{N}(-1, 1)$, $f_1 \sim \mathcal{N}(1, 2)$
- $\epsilon_0 = 0.02$ and $\epsilon_1 = 0.01$, $\alpha = 3$, $\rho = 1.2$
**Simulation setup:**

- \( f_0 \sim \mathcal{N}(-1, 1), \ f_1 \sim \mathcal{N}(1, 2) \)
- \( \epsilon_0 = 0.02, \ \epsilon_1 = 0.01, \ \rho = 1 \)
Robust Hypothesis Testing
  • Robust Hypothesis Testing with $\alpha$–divergence

Robust Distributed Hypothesis Testing
  • Robust Decentralized Hypothesis Testing
  • Minimax Decentralized Hypothesis Testing

Robust Sequential Detection (Outlook)
Robust Decentralized Hypothesis Testing
Parallel Network Topology
Robust Decentralized Hypothesis Testing
Parallel Network Topology
Robust Decentralized Hypothesis Testing
Parallel Network Topology
Robust Decentralized Hypothesis Testing

Related Work

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 42, NO. 1, JANUARY 1996

Minimax Robust Decentralized Detection

Vemulpal V. Veeravalli, Member, IEEE, Tamer Bagci, Fellow, IEEE,
and H. Vincent Poor, Fellow, IEEE

Abstract—Decentralized detection problems are studied where the sensor distributions are not specified completely. The sensor distributions are assumed to belong to known uncertainty classes. It is shown for a broad class of such problems that a set of best favorable distributions for the hypothesis. It is shown that the corresponding minimax robust test are solutions to single decentralized detection problems for which the sensor distributions are specified to be best favorable distributions.

Index Terms—Decentralized detection, robust hypothesis testing, best favorable distributions, minimax optimization.

I. INTRODUCTION

The design of optimal decision rules in detection (hypoth

thesis testing) problems requires the knowledge of the conditional probability distributions of the observations, given each hypothesis. In many applications, however, the probability distributions are not specified completely. In those cases, the probability distributions are usually specified to belong to classes (sets) of distributions, often termed uncertainty classes. One way to design decision rules when the probability distributions are given to belong to uncertainty classes is the minimax approach, where the goal is to minimize the worst case performance over the uncertainty classes. The decision rules thus obtained are said to be robust to the uncertainties in the probability distributions.

Minimax robust detection problems with two hypotheses and with centralized information have been the subject of numerous papers for a survey of results in this area, see [2]. The solutions to these problems invariably involve identifying a pair of best favorable distributions (LFD's), and subsequently designing a simple hypothesis test between the LFD's.

In this paper, we attempt to find a more comprehensive solution to robust decentralized detection problems. We study both cases with and without a fusion center. For the case without a fusion center, we give a solution to the minimax robust detection problem for the general case of finite number of sensors, finite observation block length, and nonidentical sensor characteristics. This solution covers all the block detection cases considered in [5] and [6]. Furthermore, our analysis is not restricted to Bayesian detection. For the case when no fusion center is present, we extend the work in [4] to more than two sensors and more general cost functions. We also give sufficient conditions for the decoupling of the minimax robust detection problem.

The remainder of this paper is organized as follows. In Section II, we give a detailed introduction to robust centralized detection. The purpose of this introduction is twofold: first, we believe that we have provided a framework whereby most of the previous results in robust centralized detection are unified; second, the results here are used explicitly in the solution to the decentralized problems in the subsequent sections. In Section III, we consider decentralized detection problems where a fusion center is present, and in Section IV, we consider the case where the fusion center is absent. Finally, in Section V, we give some concluding remarks. An Appendix contains the proof of one of the results is included at the end of the paper.
Robust Decentralized Hypothesis Testing

Related Work

State-of-the-Art:

- Random variables $Y_i$, corresponding to $y_i$ are
  - Mutually independent
  - Follow a distribution which belongs to the uncertainty set $G^i_j$ for hypothesis $H_j$
  - Possibly multi-variate
- Sensor decisions are possibly multi-level
Robust Decentralized Hypothesis Testing

Related Work

State-of-the-Art:

▶ Random variables $Y_i$, corresponding to $y_i$ are
  ● Mutually independent
  ● Follow a distribution which belongs to the
    uncertainty set $G_j^i$ for hypothesis $H_j$
  ● Possibly multi-variate

▶ Sensor decisions are possibly multi-level

Assumptions:

▶ Every pair of uncertainty sets $(G_0^i, G_1^i)$ satisfies the joint stochastic boundedness property

▶ Sensors must employ monotone likelihood ratio tests (MLRTs)
Robust Decentralized Hypothesis Testing

Related Work

Joint stochastic boundedness property:

\[ P_F(\hat{\delta}(t), g_0) \leq P_F(\hat{\delta}(t), \hat{g}_0) \quad \forall t, \forall g_0 \in \mathcal{G}_0 \]
\[ P_M(\hat{\delta}(t), g_1) \leq P_M(\hat{\delta}(t), \hat{g}_1) \quad \forall t, \forall g_1 \in \mathcal{G}_1 \]

Monotone likelihood ratio test (MLRT):

\[ u = \begin{cases} 
-4 & 1 \\
-2 & 2 \\
2 & 3 \\
4 & 4 \\
6 & 5 \\
10 & 6 \\
15 & 7 \\
\end{cases} \]
Recent Result:
- The same scheme which does not require the given two assumptions

Implications:
- No joint stochastic boundedness property:
  - Allows different robust tests to be jointly employed
- No monotone likelihood ratio test:
  - The sensors have the flexibility to choose the quantization method
  - The fusion center must employ a certain permutation function

Outline

☑ Robust Hypothesis Testing
  ● Robust Hypothesis Testing with $\alpha$–divergence
☑ Robust Distributed Hypothesis Testing
  ● Robust Decentralized Hypothesis Testing
☐ Robust Sequential Detection (Outlook)
Sequential Tests

What Constitutes a Sequential Test?

Test statistic:

\[ T_n = \prod_{i=1}^{n} T(x_i) = T_{n-1} T(x_n) \]
Sequential Tests

What Constitutes a Sequential Test?

Test statistic:

\[ T_n = \prod_{i=1}^{n} T(x_i) = T_{n-1} T(x_n) \]

Stopping and decision rules:

\[ T_n(x_1, \ldots, x_n) \in \begin{cases} 
S_0 & \text{decision for } \mathcal{H}_0, \ \delta = 0 \\
S_1 & \text{decision for } \mathcal{H}_1, \ \delta = 1 \\
C & \text{continue testing} 
\end{cases} \]
Sequential Tests

What Constitutes a Sequential Test?

Test statistic:

\[
T_n = \prod_{i=1}^{n} T(x_i) = T_{n-1} T(x_n)
\]

Stopping and decision rules:

\[
T_n(x_1, \ldots, x_n) \in \begin{cases} 
S_0 & \text{decision for } H_0, \ \delta = 0 \\
S_1 & \text{decision for } H_1, \ \delta = 1 \\
C & \text{continue testing}
\end{cases}
\]

Stopping time:

\[
\tau = \min_{n \geq 1} \{n : T_n \in S_0 \cup S_1\}
\]
Sequential Tests

Example

Test statistic

\[
T_n = \prod_{i=1}^{n} \frac{p_1(x_i)}{p_0(x_i)}
\]

Stopping/decision regions

\begin{align*}
S_0 &= (-\infty, B] \\
S_1 &= [A, \infty) \\
C &= (B, A)
\end{align*}

Stopping time

\[T_n \notin (B, A)\]
Sequential Tests

Example

Three performance measures:

\[ P_0[\delta = 1] \quad P_1[\delta = 0] \quad \tau = \min_n \{ n \geq 1 : T_n \notin (B, A) \} \]

error probabilities \quad run-length

(stopping time)
Three uncertainty sets $\mathcal{P}$, $\mathcal{P}_0$, $\mathcal{P}_1$

$$\min_{\delta, \tau} \max_{P \in \mathcal{P}} E_P[\tau] \quad \text{s.t.} \quad \max_{P_0 \in \mathcal{P}_0} P_0[\delta_\tau = 1] \leq \alpha_0, \quad \max_{P_1 \in \mathcal{P}_1} P_1[\delta_\tau = 0] \leq \alpha_1$$
Minimax Sequential Tests

Problem Formulation

Three uncertainty sets $\mathcal{P}, \mathcal{P}_0, \mathcal{P}_1$

$$\min_{\delta, \tau} \max_{P \in \mathcal{P}} E_P[\tau] \quad \text{s.t.} \quad \max_{P_0 \in \mathcal{P}_0} P_0[\delta \tau = 1] \leq \alpha_0, \quad \max_{P_1 \in \mathcal{P}_1} P_1[\delta \tau = 0] \leq \alpha_1$$

$\mathcal{P}_{0,1}$: Uncertainty about distributions under each hypothesis.

$\mathcal{P}$: Uncertainty about true distribution.

Need not be the same! Example:

Bound error probabilities under each hypothesis, but expected run-length over larger uncertainty set. ("At least stop in time.")
Minimax Sequential Tests

Problem Formulation

Three uncertainty sets $\mathcal{P}, \mathcal{P}_0, \mathcal{P}_1$

$$\min_{\delta, \tau} \max_{P \in \mathcal{P}} E_P[\tau] \quad \text{s.t.} \quad \max_{P_0 \in \mathcal{P}_0} P_0[\delta_\tau = 1] \leq \alpha_0, \quad \max_{P_1 \in \mathcal{P}_1} P_1[\delta_\tau = 0] \leq \alpha_1$$

Under mild assumptions the solution is characterized by

$$\rho(z_0, z_1) = \min \left\{ \lambda_0 z_0, \lambda_1 z_1, 1 + \max_{P, P_0, P_1} \int \rho \left( z_0 \frac{p_0(x)}{p(x)}, z_1 \frac{p_1(x)}{p(x)} \right) p(x) \, dx \right\}$$
Minimax Sequential Tests

Problem Formulation

Three uncertainty sets $\mathcal{P}, \mathcal{P}_0, \mathcal{P}_1$

$$\min_{\delta, \tau} \max_{P \in \mathcal{P}} E_P[\tau] \quad \text{s.t.} \quad \max_{P_0 \in \mathcal{P}_0} P_0[\delta, \tau = 1] \leq \alpha_0, \quad \max_{P_1 \in \mathcal{P}_1} P_1[\delta, \tau = 0] \leq \alpha_1$$

Under mild assumptions the solution is characterized by

$$\rho(z_0, z_1) = \min \left\{ \lambda_0 z_0, \lambda_1 z_1, 1 + \max_{P, P_0, P_1} \int \rho \left( z_0 \frac{p_0(x)}{p(x)}, z_1 \frac{p_1(x)}{p(x)} \right) p(x) \, dx \right\}$$

- stopping rule
- test statistic
- true distribution

- Joint optimization over stopping rule, test statistic and distributions
- $z_0, z_1$ denote likelihood ratios $p_0(x_1, \ldots)/p(x_1, \ldots)$ and $p_1(x_1, \ldots)/p(x_1, \ldots)$. 
Optimal test statistic is 2-dimensional.

\[
\begin{align*}
\log z_0 & \\
\log z_1 & \\
s_0 & \\
s_1 & 
\end{align*}
\]
Least favorable densities depend on the current value of the test statistic.

Detect shift in standard normal distribution: \( H_0 : \mu = 0, \quad H_1 : \mu = 1 \)

Minimize worst expected run-length under \( H_0 \), allowing for 20% outliers.
Least favorable densities depend on the current value of the test statistic.

Detect shift in standard normal distribution:

\[ H_0 : \mu = 0, \quad H_1 : \mu = 1 \]

\[ P = 0.8P_0 + 0.2H, \quad H : \text{arbitrary distribution} \]
Least favorable density $q_z(x)$ for

$$(z_0, z_1) = (1, 1) \quad \text{no preference for either hypothesis}$$
Least favorable density \( q_z(x) \) for 

\[(z_0, z_1) = (10, 1) \] preference for null hypothesis
Least favorable density $q_z(x)$ for $(z_0, z_1) = (1, 10)$ preference for alternative hypothesis
Detect shift in standard normal distribution:
\( \mathcal{H}_0 : \mu = 0, \quad \mathcal{H}_1 : \mu = 1 \)

Error probabilities:
\[ \alpha_0 = \alpha_1 = 0.05 \]

Minimax fixed sample size detector

vs.

Asymptotically minimax sequential detector
Conclusions

- Newly developed robust schemes are general in terms of the choice of nominal distributions, robustness parameters, Bayesian priors, and the distances.
- Robust decentralized hypothesis schemes generalizes previous work under more realistic assumptions.
- Theoretical bounds exist for minimax decision making in decentralized detection networks.
- Robust sequential tests are more intricate than their fixed sample size counterparts.
- ...but significantly reduce the required number of samples for moderate contamination.
Recent Trends in Robust Signal Processing

Robust, Scalable and Fast Bootstrap Method for Analyzing Large Scale Data
Large Scale Data Analysis

- We live in an era of data deluge
- The lack of scalability of the conventional signal processing (SP) and machine learning techniques and the complexity of data form a bottleneck in the search of relevant information.
- Data may be so BIG that it is not possible to store and process all the data in the same unit.
- We leverage the rich field of robust statistical signal processing to extract relevant information from high-volume and high-dimensional data
Data Analysis problem at hand

Let \( X = (x_1 \cdots x_n) \in \mathbb{R}^{d \times n} \) be a large volume and high dimensional data set that can only be processed and stored via parallel and distributed architectures.

Let \( \hat{\theta}_n \) be an estimator of a parameter of interest \( \theta \in \mathbb{R}^d \) based on the observed big data \( X \).

Statistical inference and analysis of such large scale data sets is crucial in order to quantify statistical correctness of parameter estimates \( \hat{\theta}_n \) (e.g., via Confidence Intervals) and testing hypotheses.

The Problem: Performing statistical inference on massive data sets is not computationally feasible using the conventional statistical inference methodology such as the bootstrap.
Data Analysis problem at hand

- Confidence intervals may be more useful information than point estimate for Big Data.
- We introduce scalable, robust and computationally efficient bootstrap techniques for computing confidence intervals for big data.
A robust bootstrap method is presented. It facilitates bootstrap analysis of very large scale data. It is suited for estimators that can be expressed as a solution to fixed-point equations (e.g. M-estimator, MM-estimator, S-estimator, FastICA estimator).

The method is statistically convergent and quantitatively robust

Key properties:
1. Scalable to very large volume and high-dimensional data sets (Big Data).
2. Compatible with distributed data storage systems and distributed and parallel processing architectures.
3. Fast to compute as the fixed-point estimating equations do not need to be (re)solved for each bootstrap sample.
4. Statistically highly robust against outliers.
The Conventional Bootstrap

- The Bootstrap method [Efron, 1979] combines statistics and high-speed computational techniques, storage and resampling in order to find properties of estimators.

- Computational capabilities and resampling are used to compensate for the lack of knowledge on statistical properties or underlying distribution models.
The Conventional Bootstrap

The bootstrap method [Efron, 1979] is a consistent and reliable method of constructing confidence intervals for statistical estimates (e.g., by bootstrap percentile method, BCA method, etc.).

1. Generate \( r \) bootstrap samples \( X^* \) of size \( n \) by resampling with replacement from the original data set \( X \).
2. Compute \( \hat{\theta}_n^* \) on each bootstrap sample \( X^* \).
3. Use the population of bootstrap replications \( \hat{\theta}_n^* \) to estimate the desired confidence intervals \( \xi \).

\[
X = (x_1 \cdots x_n) \\
X^{*(1)} = (x_1^{*(1)} \cdots x_n^{*(1)}) \\
\hat{\theta}_n^{*(1)} \\
X^{*(2)} = (x_1^{*(2)} \cdots x_n^{*(2)}) \\
\hat{\theta}_n^{*(2)} \\
\vdots \\
X^{*(B)} = (x_1^{*(B)} \cdots x_n^{*(B)}) \\
\hat{\theta}_n^{*(B)} \\
\xi(\hat{\theta}_n^{*(1)}, \hat{\theta}_n^{*(2)}, \ldots, \hat{\theta}_n^{*(B)})
\]
Remarks on computation

- The simplest method of finding confidence intervals for an unknown parameter is to take $\alpha/2$ and $1 - \alpha/2$ quantiles of the bootstrap distribution of the estimator $\hat{\theta}_n$ as endpoints of the $100(1 - \alpha)\%$ confidence interval.

- Simple example: If we have 100 bootstrap estimates of $\theta$, we will rank order them and the 90% confidence interval is found by choosing the 5th smallest and 95th value as the end points of the interval.

- In other words, trim $\alpha/2\%$ of the smallest and largest values off to find the confidence interval.
Remarks on computation

- Expected number of distinct data points in each resample is about $0.632n$
- Computational complexity typically scales with the number of distinct data points
- For a dataset of 1 TB, resample size would be 632 GB
- Parameter estimation and quality assessment is done for each resample
- Methods for reducing the number of resamples [Efron, 1979] and subsampling methods ($m$ out of $n$ bootstrap) have been proposed [Bickel, 1997]
The Conventional Bootstrap

- **Advantages:** Accurate for a wide range of estimators $\hat{\theta}_n$
- Automatic, because does not require any manipulation of the estimation equations
- **Disadvantages:** Computationally very costly since the estimator $\hat{\theta}_n$ is recomputed for each bootstrap sample
- Not robust in the face of outliers (highly deviating data)
- The size of each bootstrap sample is the same as the original big data set
- Not scalable, not suitable for distributed storage and (parallel) computing architectures.
The bag of little bootstraps (BLB) is a newly proposed bootstrap scheme [Kleiner, Jordan, et al., 2014] aiming to make the naive bootstrap method suitable for analysis of Big Data. In this method:

- Disjoint subsamples of significantly smaller size $b = \{ \lfloor n^\gamma \rfloor | \gamma \in [0.6, 0.9] \}$ are drawn from the original Big Data set. Subsamples may be kept in/sent to in distributed storage and processing units for parallel computations.

- In each unit, bootstrap samples are constructed by assigning a random weight vector $\mathbf{n}^* = (n_1^*, \ldots, n_b^*)$ from $\text{Multinomial}(n, (1/b)\mathbf{1}_b)$ to the distinct data points of the subsample, where $\sum_{i=1}^b n_i^* = n$. 
The BLB procedure

1. Draw \( s \) subsamples \( \mathbf{\tilde{X}} = (\mathbf{\tilde{x}}_1, \ldots, \mathbf{\tilde{x}}_b) \) of smaller size \( b = \{\lfloor n^{\gamma} \rfloor | \gamma \in [0.6, 0.9] \} \) by randomly sampling without replacement from \( \mathbf{X} \).

for each subsample \( \mathbf{\tilde{X}} \)

2. Generate \( r \) bootstrap samples \( \mathbf{X}^* = (\mathbf{\tilde{X}}; \mathbf{n}^*) \) by assigning a random weight vector \( \mathbf{n}^* = (n_1^*, \ldots, n_b^*) \) from \( \text{Multinomial}(n, (1/b)\mathbf{1}_b) \) to data points of \( \mathbf{\tilde{X}} \).

3. Compute the estimator \( \hat{\theta}_n^* \) based on each \( \mathbf{X}^* \).

4. Use the population of \( r \) bootstrap replications \( \hat{\theta}_n^* \) to estimate the bootstrap confidence interval \( \hat{\xi}^* \) (e.g., by bootstrap percentile method).

end

5. Average the computed values of \( \hat{\xi}^* \) over the subsamples, i.e.,

\[
\hat{\xi}^* = \frac{1}{s} \sum_{k=1}^{s} \xi^{*(k)}. 
\]
The Bag of Little Bootstraps (BLB)

\[ X = (x_1 \cdots x_n) \]

- \( \tilde{X}^{(k)} \), \( k = 1, \ldots, s \) denotes the disjoint subsamples
- \( X^{*(kj)} \), \( j = 1, \ldots, r \) corresponds to the \( j \)th BLB sample generated based on the subsample \( k \).

\[ \tilde{\xi}^{(k)} \]

\[ \xi^{*(1)} \]

\[ \xi = \frac{1}{s} \sum_{i=1}^{s} \tilde{\xi}^{(i)} \]
The Bag of Little Bootstraps (Remarks)

The distinct data of the subsamples in Step 2 allows the original Big data set to be stored in distributed storage systems.

In Step 2, subsamples $\check{X} = (\check{x}_1 \cdots \check{x}_b)$ can be processed in parallel using different computing nodes.

$X^* = (\check{X}; n^*)$ resembles a conventional bootstrap sample of size $n$ with at most $b = \{\lfloor n^\gamma \rfloor | \gamma \in [0.6, 0.9]\}$ distinct data points. Element $n_i^*$ of $n^* = (n_1^*, \ldots, n_b^*)$ denotes the multiplicity of original subsample data point $\check{x}_i$ at the bootstrap sample $X^*$.

BLB is computationally less complex than the conventional bootstrap. E.g., in the BLB scheme, $E_{F_n^*}[X^*]$ is computed by $b$ summations (+) and $b$ multiplications ($\times$) as $E_{F_n^*}[X^*] = \frac{1}{n} \sum_{i=1}^{b} n_i^* \check{x}_i$, whereas in conventional method $n$ summations (+) are needed as $E_{F_n^*}[X^*] = \frac{1}{n} \sum_{i=1}^{n} x_i^*$. 
The Bag of Little Bootstraps (BLB)

Advantages

▶ In comparison with the conventional bootstrap, less computational resources are needed.
▶ BLB is scalable and well suited for distributed computing architectures and storage systems.

Disadvantages

▶ The estimating equations need to be (re)solved for all bootstrap samples of all bags (overall $s \times r$ times). This is prohibitively expensive especially when a full optimization problem need to be numerically solved (e.g. matrix inversion or fixed-point iterative algorithm).
▶ The method is not statistically robust, in the sense that outlier contamination of only one subsample deteriorates the end result of the whole scheme.
The Fast and Robust Bootstrap (FRB)

FRB method [Salibian-Barrera, et al., 2008] is applicable for estimators \( \hat{\theta}_n \in \mathbb{R}^d \) that can be expressed as a solution to a system of smooth Fixed Point estimation equations

\[
\hat{\theta}_n = Q(\hat{\theta}_n; X),
\]

(5)

where \( Q : \mathbb{R}^d \to \mathbb{R}^d \) and \( Q() \) is continuous and differentiable The bootstrap replicated estimator \( \hat{\theta}_n^* \) then solves

\[
\hat{\theta}_n^* = Q(\hat{\theta}_n^*; X^*),
\]

(6)

Fixed point estimating equations usually converge fast, in few iterations.
Instead of computing $\hat{\theta}_n^*$, we compute an approximation:

$$\hat{\theta}_n^{1*} = Q(\hat{\theta}_n; X^*),$$

(7)

Since the distribution of $\hat{\theta}_n^{1*}$ typically underestimates the sampling variability of $\hat{\theta}_n$, a linear correction based on Taylor approximation of function $Q$ is applied as follows:

$$\hat{\theta}_n^{R*} = \hat{\theta}_n + \left[ I - \nabla Q(\hat{\theta}_n; X) \right]^{-1} (\hat{\theta}_n^{1*} - \hat{\theta}_n),$$

(8)

where $\nabla Q(\cdot) \in \mathbb{R}^{d \times d}$ is the matrix of partial derivatives w.r.t. $\hat{\theta}_n$. 

\[ X = (x_1 \cdots x_n) \rightarrow \hat{\theta}_n \]

\[
\begin{align*}
Q(\hat{\theta}_n; X^{(1)}) & \rightarrow \hat{\theta}_n^{R*(1)} \\
Q(\hat{\theta}_n; X^{(2)}) & \rightarrow \hat{\theta}_n^{R*(2)} \\
& \vdots \\
Q(\hat{\theta}_n; X^{(r)}) & \rightarrow \hat{\theta}_n^{R*(r)} \\
\end{align*}
\]

$$\xi(\hat{\theta}_n^{R*(1)}, \hat{\theta}_n^{R*(2)}, \ldots, \hat{\theta}_n^{R*(r)})$$
The Fast and Robust Bootstrap (FRB)

Advantages:

- Fast to compute, as the initial estimator $\hat{\theta}_n$ is computed only once (e.g., for the full data set $X$). One step improvement $\hat{\theta}_n^{1*}$ requires only one iteration of FP equation.
- Robust against outliers. For instance in case of the MM-regression estimator, it has been shown that equation (8) remains bounded if $\hat{\theta}_n$ is a reliable estimate of $\theta$ and there are only $p$ (the dimension of the regression model) non-outlier data points in the bootstrap sample $X^*$. 

Disadvantages:

- Not scalable and difficult to parallelize across distributed computing systems. $\hat{\theta}_n$ computed from large scale data $X$
A new bootstrap scheme is presented that is suitable for analyzing large multivariate data sets. The BLFRB method [S. Basiri, E. Ollila, V. Koivunen, 2015] combines the desirable properties of the BLB and FRB methods as it is:

1. Scalable to large volume data sets and compatible with distributed data storage and processing architectures.

2. Less complex and fast to compute as the estimating equations are computed only once for each bag.

3. Statistically robust and works reliably in the face of outliers. Bootstrap analysis of fixed-point robust estimators (e.g. S-estimator, MM-estimator, etc.) is facilitated, thanks to the low complexity of the scheme.
Recall that the main computational burden of the BLB scheme is in step 3 of the algorithm where, the estimating equations need to be (re)solved for each bootstrap sample $X^*$. 
Scalable, Fast and Robust Bootstrap for Big Data

Such computational complexity can be drastically reduced by computing the FRB replications instead. This can be done locally for each distinct dataset:

- Let $\hat{\theta}_{n,b}$ be a solution to $\hat{\theta}_n = Q(\hat{\theta}_n; X)$, for subsample $\tilde{X} \in \mathbb{R}^{d \times b}$:

$$\hat{\theta}_{n,b} = Q(\hat{\theta}_{n,b}; \tilde{X}). \quad (9)$$

- Let $X^* \in \mathbb{R}^{d \times n}$ be a bootstrap sample of size $n$ randomly resampled with replacement from distinct data subset $\tilde{X}$ of size $b$;

- The FRB replication of $\hat{\theta}_{n,b}$ can be obtained by

$$\hat{\theta}_{n,b}^{R*} = \hat{\theta}_{n,b} + [I - \nabla Q(\hat{\theta}_{n,b}; \tilde{X})]^{-1} (\hat{\theta}_{1*}^{n,b} - \hat{\theta}_{n,b}), \quad (10)$$

where $\hat{\theta}_{1*}^{n,b} = Q(\hat{\theta}_{n,b}; X^*)$ is the one-step estimator and $\nabla Q(\cdot) \in \mathbb{R}^{d \times d}$ is the matrix of partial derivatives w.r.t. $\hat{\theta}_{n,b}$.

- Note: The initial estimate $\hat{\theta}_{n,b}$ and correction $[I - \nabla Q(\hat{\theta}_{n,b}; \tilde{X})]^{-1}$ are computed only once for each distinct data subset.
BLFRB Procedure

1: Draw \( s \) disjoint subsamples \( \tilde{X} = (\tilde{x}_1, \ldots, \tilde{x}_b) \) of smaller size \( b = \lfloor n^\gamma \rfloor | \gamma \in [0.6, 0.9] \).

for each subsample \( \tilde{X} \)

2: Generate \( r \) bootstrap samples \( X^* = (\tilde{X}; n^*) \) according to the BLB procedure.

3: \( a \): Find the estimate \( \hat{\theta}_{n,b} \) based on \( \tilde{X} \).

\( b \): For each bootstrap sample \( X^* \) compute the FRB replication \( \hat{\theta}_{n,b}^{R*} \) using \( \hat{\theta}_{n,b} \).

4: Compute the bootstrap confidence intervals \( \hat{\xi}^* \) based on the population of \( r \) FRB replicated values \( \hat{\theta}_{n,b}^{R*} \).

end

5: Average the computed values of \( \hat{\xi}^* \) over the subsamples, i.e.,

\[
\hat{\xi}^* = \frac{1}{s} \sum_{k=1}^{s} \hat{\xi}^*(k).
\]
Example formulation of BLFRB for MM-estimator of linear regression

Let $\mathbf{X} = \{(y_1, \mathbf{z}_1^\top)^\top, \ldots, (y_n, \mathbf{z}_n^\top)^\top\}$, $\mathbf{z}_i \in \mathbb{R}^p$, be a sample of independent random vectors that follow the linear model:

$$y_i = \mathbf{z}_i^\top \mathbf{\theta} + \sigma_0 e_i \quad \text{for} \quad i = 1, \ldots, n,$$

where:

- $\mathbf{\theta} \in \mathbb{R}^p$: The unknown parameter vector.
- $e_i$: Noise terms are i.i.d. random variables from a symmetric distribution with unit scale.
Highly robust MM-estimators [V. J. Yohai, 1987] proceed in 3 stages: (1) initial highly robust estimate that is not necessarily efficient is found; (2) M-estimate of error scale is computed based on residuals; (3) M-estimate of the parameter vector is computed.

Two loss functions $\rho_0 : \mathbb{R} \to \mathbb{R}^+$ and $\rho_1 : \mathbb{R} \to \mathbb{R}^+$ are used which determine the breakdown point and efficiency of the estimator. The $\rho_0(\cdot)$ and $\rho_1(\cdot)$ functions are

- Symmetric,
- Twice continuously differentiable with $\rho(0) = 0$,
- Strictly increasing on $[0, c]$ and constant on $[c, \infty)$ for some constant $c$.
- $\psi = \rho'$
Example formulation of BLFRB for MM-estimator of linear regression

The MM-estimate of $\hat{\theta}_n$ satisfies

$$\frac{1}{n} \sum_{i=1}^{n} \psi_1 \left( \frac{y_i - z_i^\top \hat{\theta}_n}{\hat{\sigma}_n} \right) z_i = 0$$  \hspace{1cm} (12)$$

where $\hat{\sigma}_n$ is a S-estimate of scale defined as follows.

Consider M-estimate of scale $\hat{s}_n(\theta)$ obtained as a solution to:

$$\frac{1}{n} \sum_{i=1}^{n} \rho_0 \left( \frac{y_i - z_i^\top \theta}{\hat{s}_n(\theta)} \right) = m,$$  \hspace{1cm} (13)$$

where $m = \rho_0(\infty)/2$ is a constant. Let $\tilde{\theta}_n$ be the argument that minimizes $\hat{s}_n(\theta)$,

$$\tilde{\theta}_n = \arg \min_{\theta \in \mathbb{R}^p} \hat{s}_n(\theta),$$

then $\hat{\sigma}_n = \hat{s}_n(\tilde{\theta}_n)$. 
Example formulation of BLFRB for MM-estimator of linear regression

Simple computations yield the following FP representation of (12) and (13):

$$\hat{\theta}_n = \left( \sum_{i=1}^{n} \omega_i z_i z_i^\top \right)^{-1} \sum_{i=1}^{n} \omega_i z_i y_i,$$  \hspace{1cm} (14)

$$\hat{\sigma}_n = \sum_{i=1}^{n} \nu_i (y_i - z_i^\top \tilde{\theta}_n),$$  \hspace{1cm} (15)

where

$$r_i = y_i - z_i^\top \hat{\theta}_n, \quad \hat{r}_i = y_i - z_i^\top \tilde{\theta}_n,$$

$$\omega_i = \rho'_1(r_i/\hat{\sigma}_n)/r_i \quad \text{and} \quad \nu_i = \frac{\hat{\sigma}_n}{nm} \rho_0(\hat{r}_i/\hat{\sigma}_n)/\hat{r}_i.$$
Example formulation of BLFRB for MM-estimator of linear regression

Let $X^* = (\hat{X}; n^*)$ denote a BLFRB bootstrap sample based on subsample $\hat{X} = \{(\hat{y}_1, \hat{z}_1^T)^T, \ldots, (\hat{y}_b, \hat{z}_b^T)^T\}$, $\hat{z}_i \in \mathbb{R}^p$ and a weight vector $n^* = (n_1^* \cdots n_b^*) \in \mathbb{R}^b$,

$$\hat{\theta}_{n,b}^{1*} = \left( \sum_{i=1}^{b} n_i^* \hat{\omega}_i \hat{z}_i \hat{z}_i^T \right)^{-1} \sum_{i=1}^{b} n_i^* \hat{\omega}_i \hat{z}_i \hat{y}_i,$$

$$\hat{\sigma}_{n,b}^{1*} = \sum_{i=1}^{b} n_i^* \hat{\upsilon}_i (\hat{y}_i - \hat{z}_i^T \tilde{\theta}_{n,b}),$$

where

$$\hat{r}_i = \hat{y}_i - \hat{z}_i^T \hat{\theta}_{n,b}, \quad \tilde{r}_i = \hat{y}_i - \tilde{z}_i^T \tilde{\theta}_{n,b},$$

$$\check{\omega}_i = \rho'_1(\hat{r}_i/\hat{\sigma}_{n,b})/\hat{r}_i \quad \text{and} \quad \check{\upsilon}_i = \frac{\hat{\sigma}_{n,b}}{nm} \rho_0(\tilde{r}_i/\hat{\sigma}_{n,b})/\tilde{r}_i.$$
The BLFRB replications of $\hat{\theta}_{n,b}$ are obtained from the linearly corrected version of the one-step approximations in (16) and (17):

$$\hat{\theta}_{n,b}^{R*} = \hat{\theta}_{n,b} + M_{n,b}(\hat{\theta}_{n,b}^{1*} - \hat{\theta}_{n,b}) + d_{n,b}(\hat{\sigma}_{n,b}^{1*} - \hat{\sigma}_{n,b}),$$

(18)

where

$$M_{n,b} = \hat{\sigma}_{n,b} \left( \sum_{i=1}^{b} \rho''_{1}(\bar{r}_{i}/\hat{\sigma}_{n,b})\bar{z}_{i}\bar{z}_{i}^{\top} \right)^{-1} \sum_{i=1}^{b} \bar{\omega}_{i}\bar{z}_{i}\bar{z}_{i}^{\top},$$

$$d_{n,b} = k_{n,b}^{-1} \left( \sum_{i=1}^{b} \rho''_{1}(\bar{r}_{i}/\hat{\sigma}_{n,b})\bar{z}_{i}\bar{z}_{i}^{\top} \right)^{-1} \sum_{i=1}^{b} \rho'_{1}(\bar{r}_{i}/\hat{\sigma}_{n,b})\bar{r}_{i}\bar{z}_{i}$$

and

$$k_{n,b} = \frac{1}{nm} \sum_{i=1}^{b} \left( \rho'_{0}(\bar{r}_{i}/\hat{\sigma}_{n,b})\bar{r}_{i}/\hat{\sigma}_{n,b} \right).$$
STATISTICAL PROPERTIES

- Estimator is shown to be convergent (convergence in distribution).
- Statistical robustness of the estimator is proved. It inherits the breakdown point of the fixed point estimator used.
- Estimator can be designed to possess high efficiency (e.g. 95%).
- Proofs may be found in
  S. Basiri, E. Ollila, and V. Koivunen,
  *Robust, scalable and fast bootstrap method for analyzing large scale data*,
### Statistical Properties

Statistical robustness

<table>
<thead>
<tr>
<th>p</th>
<th>n</th>
<th>$\gamma = 0.6$</th>
<th>$\gamma = 0.7$</th>
<th>$\gamma = 0.8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>50000</td>
<td>0.425</td>
<td>0.475</td>
<td>0.491</td>
</tr>
<tr>
<td></td>
<td>200000</td>
<td>0.467</td>
<td>0.490</td>
<td>0.497</td>
</tr>
<tr>
<td></td>
<td>1000000</td>
<td>0.488</td>
<td>0.497</td>
<td>0.499</td>
</tr>
<tr>
<td>100</td>
<td>50000</td>
<td>0.349</td>
<td>0.449</td>
<td>0.483</td>
</tr>
<tr>
<td></td>
<td>200000</td>
<td>0.434</td>
<td>0.481</td>
<td>0.494</td>
</tr>
<tr>
<td></td>
<td>1000000</td>
<td>0.475</td>
<td>0.494</td>
<td>0.498</td>
</tr>
<tr>
<td>200</td>
<td>50000</td>
<td>0.197</td>
<td>0.398</td>
<td>0.465</td>
</tr>
<tr>
<td></td>
<td>200000</td>
<td>0.368</td>
<td>0.461</td>
<td>0.488</td>
</tr>
<tr>
<td></td>
<td>1000000</td>
<td>0.450</td>
<td>0.487</td>
<td>0.497</td>
</tr>
</tbody>
</table>

**Table:** Upper breakdown point of the BLFRB estimates of quantiles for MM-regression estimator with 50% breakdown point and 95% efficiency at the Gaussian model.
The bootstrap setup is as follows:

- The number of distinct data subsamples is $s = 25$,
- size of each subsample is $b = \lfloor n^\gamma \rfloor = 1946$ with $\gamma = 0.7$ ($n = 50000$)
- maximum number of bootstrap samples in each subsample module is $r_{\text{max}} = 300$.

We start from $r = 2$ and continually add a new set of bootstrap samples (while $r < r_{\text{max}}$) to subsample modules.
Figure: Relative errors of the BLB (solid line) and BLFRB (dashed line) methods w.r.t. the number of bootstrap samples $r$ are illustrated. Both methods perform equally well when there are no outliers in the data.
NUMERICAL EXAMPLES
Lack of robustness of BLB

To show lack of robustness of the BLB, we introduce outlier by randomly choosing one of the data points and multiplying it by a large $\alpha$. 

![Graph showing relative errors vs. number of bootstrap samples for BLB with different $\alpha$ values and clean data.]

- BLB, $\alpha = 1000$
- BLB, $\alpha = 500$
- BLB, Clean data
We severely contaminate the original data points of the first bag by multiplying 40% ([0.4 \times b] = 778) of the data points by $\alpha = 1000$. 

![Graph showing relative errors (ε) against number of bootstrap samples (r) for BLFRB, $\alpha = 1000$ and BLFRB, Clean data. The graph shows nearly negligible relative errors across different bootstrap sample sizes, indicating the robustness of the BLFRB method.](image)
The computational complexity of the BLB and BLFRB methods are compared.

- The same MM-estimator is used in both schemes.
- An identical computing system is used to compute the relative errors after each iteration of the algorithms.
- The required processing time is cumulatively recorded after each iteration of the algorithms.
NUMERICAL EXAMPLES
Computational complexity

Figure: The computed relative errors $\varepsilon$ after adding each set of new bootstrap samples are depicted w.r.t. the required processing time of computation. The BLFRB is significantly faster to compute as the (re)computation of the estimating equations is not needed in this method.
NUMERICAL EXAMPLES
Real world data

▶ We consider the simplified version of the Million Song Dataset (MSD), available on the UCI Machine Learning Repository.
▶ The data set $X = \{(y_1, z_1^\top)^\top, \ldots, (y_n, z_n^\top)^\top\}$ contains $n = 515345$ music tracks, where:
  ▶ $y_i$ (i.e., $i = 1, \ldots, n$) represents the released year of the $i$th song (i.e., ranging from 1922 to 2011).
  ▶ $z_i \in \mathbb{R}^p$ is a vector of $p = 90$ different audio features of each song.
  ▶ The features are the average and non-redundant covariance values of the timbre vectors of the song.
▶ Linear regression can be used to predict the released year of a song based on its audio features.
Considering the linear model $y_i = z_i^T \theta + \sigma_0 e_i$, we use BLFRB to conduct a fast, robust and scalable hypothesis test:

$$\mathcal{H}_0 : [\theta]_l = 0 \quad \text{vs.} \quad \mathcal{H}_1 : [\theta]_l \neq 0,$$

where $[\theta]_l$ (i.e., $l = 1, \ldots, p$) denotes the $l$th element of $\theta$.

The BLFRB test of level $\alpha$ rejects the null hypothesis if the computed $100(1 - \alpha)$% confidence interval does not contain 0.

We make a test on each feature coefficient $\theta_i$ and discard it if its confidence interval contains 0.
NUMERICAL EXAMPLES
Real world data

Here we run the BLFRB hypothesis test of level $\alpha = 0.05$ with the following bootstrap setup;

- Number of distinct data subsamples (bags) is $s = 51$,
- size of each subsample is $b = \lceil n^{\gamma} \rceil = 9964$ with $\gamma = 0.7$, $n = 515345$.
- number of bootstrap samples in each subsample module is $r = 500$.

The null hypothesis is accepted for 6 features numbered: 32, 40, 44, 47, 54, 75. These results can be exploited to reduce the dimension of the data by excluding the ineffective variables from the regression analysis.
**Figure**: The 95% confidence intervals computed by BLFRB method is shown for some of the audio features of the MSD data set. The null hypothesis is accepted for those features having 0 inside the interval.
A new bootstrap method is presented with the aim of facilitating robust, scalable bootstrap analysis of large multivariate data sets. The proposed method is:

1. Scalable to large volume and high dimensional data sets.
2. Compatible with distributed data storage systems and distributed parallel processing architectures.
3. Robust in the face of outliers and heavy-tailed noise.
4. Significantly faster to compute than its only counterpart (i.e., the BLB scheme).
Further Emerging Topics in Robust SP

- Robust Statistical Inference for Big Data analytics (Volume, Velocity, Variety)
  - Fast, scalable and robust statistical methods for distributed computation and storage: bootstrapping
  - Sequential Inference for dynamic large scale data
  - Robust methods for tensors: decompositions, sparseness, tensor completion
- Random Matrix Theory and statistical robustness [R. Couillet, F. Pascal, 2011]
- Geodesic convexity [A. Wiesel, 2012]
- Regularized M-estimation [E. Ollila, D. Tyler, 2014]
- Robust inference in graphs
Overall Conclusions of the Tutorial

Non-parametric

Robust

Parametric
Overall Conclusions of the Tutorial

Optimality is important, but ... robustness is the engineer’s choice.
Additional Material: High-Dimensional Data
High-Dimensional Data

- High-dimensional data e.g. in Neuroimaging, Remote Sensing, Chemometrics, Environmetrics, Network Data, Internet Data, Data collected by mobile terminals
- Tensors represent such data as multi-way arrays. Tensor decompositions provide a unifying framework for multidimensional data analysis with simplified notations and algebras.
- Tensors in SP, major advances by N. Sidiropoulos, P. Comon, L de Lathauwer
- Sparsity constraints can be used for accurate signal recovery (e.g. compressed sensing) or to eliminate unnecessary redundant features of modern data sets (e.g. financial data, DNA micro arrays, network traffic flows, fMRIs).
- Robustness ensures the resistance to heavy-tailed errors or outliers that appear commonly in high-dimensional data.
Specially Structured Tensors

▶ Kruskal Tensor: \( \mathbf{X} \in \mathbb{R}^{I \times J \times K}, \gamma \in \mathbb{R}^{R}, \)

\[
\mathbf{X} \equiv [\gamma; \mathbf{A}, \mathbf{B}, \mathbf{C}] = \sum_{r=1}^{R} \gamma_r \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r.
\]

where \( \mathbf{a} \in \mathbb{R}^{I}, \mathbf{b} \in \mathbb{R}^{J} \) and \( \mathbf{c} \in \mathbb{R}^{K} \) form the unit-norm column vectors of \( \mathbf{A} \in \mathbb{R}^{I \times R}, \mathbf{B} \in \mathbb{R}^{J \times R}, \) and \( \mathbf{C} \in \mathbb{R}^{K \times R}. \)
Tensor decompositions

Approximate a tensor by a low-rank set of factors along each tensor mode

- CANDECOMP = Canonical Decomposition [Carrol and Chang, 1970],
  PARAFAC = Parallel Factors [Harshman, 1970]

- CANDECOMP / PARAFAC (CP) decomposition

\[ \mathcal{X} \approx [\gamma; A, B, C] = \sum_{r=1}^{R} \gamma_r a_r \circ b_r \circ c_r. \]
Tensor decompositions

- Tucker decomposition [Tucker, 1966]: Three-mode factor analysis, Three-mode PCA, or Orthogonal array decomposition

\[ \mathbf{X} \approx [\mathbf{G}; \mathbf{A}, \mathbf{B}, \mathbf{C}] = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr} a_p \circ b_q \circ c_r \]

Not Unique!

- HOSVD, HOII, HOPM, ...
CP decomposition

approximates a tensor $\mathbf{X} \in \mathbb{R}^{I \times J \times K}$ by a predicted tensor $\hat{\mathbf{X}}$ consisting of a sum of rank-1 tensors:

$$\hat{\mathbf{X}} \equiv [\gamma; \mathbf{A}, \mathbf{B}, \mathbf{C}] \triangleq \sum_{r=1}^{R} \gamma_r \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r.$$ 

Thus, we model $\mathbf{X}$ as

$$\mathbf{X} = \sum_{r=1}^{R} \gamma_r \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r + \mathbf{E} \quad (19)$$

where $\mathbf{a}_r \in \mathbb{R}^I$, $\mathbf{b}_r \in \mathbb{R}^J$ and $\mathbf{c}_r \in \mathbb{R}^K$ for $r = 1, \ldots, R$ form the unit-norm column vectors of $\mathbf{A} \in \mathbb{R}^{I \times R}$, $\mathbf{B} \in \mathbb{R}^{J \times R}$, and $\mathbf{C} \in \mathbb{R}^{K \times R}$ and the tensor $\mathbf{E} \in \mathbb{R}^{I \times J \times K}$ contains the error terms.

**GOAL:** to minimize

$$\| \mathbf{X} - \hat{\mathbf{X}} \|_F = \| \mathbf{X}_k - \hat{\mathbf{X}}_k \|_F$$
The model (19) can be expressed in a matrix form by unfolding the tensor into a matrix along any of the three modes. Unfolding the tensor $\mathbf{X}$ along the first mode yields a $I \times JK$-matrix denoted as $\mathbf{X}_{(1)}$ so that the equivalent representation of (19) is

$$\mathbf{X}_{(1)} = \mathbf{A}\Gamma(\mathbf{C} \odot \mathbf{B})^T + \mathbf{E}_{(1)},$$

(20)

where $\Gamma = \text{diag}(\gamma)$ and $\mathbf{E}_{(1)}$ denotes the unfolded $I \times JK$ matrix of $\mathbf{E}$. Similarly,

$$\mathbf{X}_{(2)} = \mathbf{B}\Gamma(\mathbf{C} \odot \mathbf{A})^T + \mathbf{E}_{(2)},$$

$$\mathbf{X}_{(3)} = \mathbf{C}\Gamma(\mathbf{B} \odot \mathbf{A})^T + \mathbf{E}_{(3)}$$
Consider the case that $B$ and $C$ are fixed and that $\gamma_r$’s are the scales of the columns of $A$, i.e., $a_r$’s are no-longer unit vectors, but $\gamma_r = \|a_r\|$. 

$$
\min_A \|X - [\gamma; A, B, C]\|^2 = \min_A \|X^{(1)} - A(C \odot B)^T\|^2 
$$

(21)

of which the LS solution is

$$
\hat{A} = X^{(1)}(C \odot B)((C^T C) \ast (B^T B))^\dagger 
$$

with $\dagger$ denoting the Moore-Penrose inverse.

Note that $(C \odot B)^T(C \odot B) = (B^T B) \ast (C^T C)$ where $\ast$ denotes pointwise multiplication.

ALS idea: Solve for each factor in an alternating manner, leaving all the others fixed.
ALS for CP decompositions

1. Initialize $\mathbf{B}$ and $\mathbf{C}$ by $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$
2. $\hat{\mathbf{A}} = \mathbf{X}(1)\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}$, where $\mathbf{Z} = \mathbf{C} \odot \mathbf{B}$
3. $\hat{\mathbf{B}} = \mathbf{X}(2)\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}$, where $\mathbf{Z} = \mathbf{C} \odot \hat{\mathbf{A}}$
4. $\hat{\mathbf{C}} = \mathbf{X}(3)\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}$, where $\mathbf{Z} = \hat{\mathbf{B}} \odot \hat{\mathbf{A}}$

▷ Repeat steps 2–4 until the relative change in fit is small.
generally means that only a few entries are non-zero.

**Two notions of Sparsity:**
1. The considerable number of data elements are zero or close to zero in their relative magnitude.

2. In regularization methods (e.g. ridge regression, LASSO), *sparsity* is used for the estimated regression parameters that are either shrunk towards zero or put to zero by increasing the penalty of model complexity.

Relation of two notions: The underlying sparsity of tensor data naturally implies that factor matrices of a decomposed tensor are sparse as well. Regularization methods successfully estimate tensor factors compared to the usual tensor estimation based on the least squares.
LASSO (Least Absolute Shrinkage and Selection Operator): [R. Tibshirani, 1996]

To obtain sparse solutions, we solve for $\mathbf{A}$ using $\ell_2 - \ell_1$ criterion function instead of $\ell_2$ criterion:

$$\hat{\mathbf{A}} \equiv \text{LASSO}(\mathbf{X}_{(1)}, \mathbf{Z}, \lambda) = \arg \min_{\mathbf{A}} \| \mathbf{X}_{(1)} - \mathbf{A} \mathbf{Z}^T \|_2^2 + \lambda \| \mathbf{A} \|_1$$
Robust tensor decompositions

- Outliers often occurring in high-dimensional data indicate some deviation from the model assumptions and add difficulty in data analysis.
- Tensor decompositions based on least squares are highly sensitive to outliers or heavy-tailed errors resulting in biased estimates.
- Very little work on robust estimators in the tensor community. Other than some work in the medical imaging, robust tensor factorization studies are found in [Vorobyov et al., 2005], [Pang and Yuan, 2010], [Chi and Kolda, 2011].
Robust Error Criteria

- **Least Absolute Deviations (LAD):** For the regression parameter \( \beta = (\beta_1, ..., \beta_p) \), the absolute value loss function is

\[
l(r(\beta)) = \sum_{j=1}^{h} |r(\beta)|
\]

- **Least Trimmed Squares Regression (LTS):** [Rousseeuw, 1984]

- **Tukey Loss Function**

\[
\rho(r(\beta)) = \min\{1, (1 - (r(\beta)/c)^2)^3\}
\]

where \( c = 3.4437 \) attaining 85% of efficiency at normal distribution.
Robust CP decompositions

- [Vorobyov et al., 2005], [Chi & Kolda, 2011]:

(LAD) Let $X = X_{(1)}$ in (21).

\[
Q_{L_1}(A, B, C) = \|X - AZ^\top\|_1 = \sum_{i=1}^{m} \sum_{j=1}^{n} |x_{ij} - z_j^\top a_i|.
\]

Note: $L_1$-loss is not bounded!
Robust and Sparse Tensor Decompositions

We present a tensor decomposition methods that enjoy both the properties of sparsity and robustness to outliers.

- CP Alternating LAD + LASSO (CPA-LAD LASSO)
Objective function for Robustness and Regularization:

\[
\sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} |x_{ij} - z_j^\top a_i| + \lambda_1 \|a_i\|_1 \right\} + \lambda_2 \|B\|_1 + \lambda_3 \|C\|_1.
\]

The minimum \( \hat{A} = (\hat{a}_1 \cdots \hat{a}_m)^\top \) can be found by

\[
\hat{a}_i = \min_{a} \left\{ \sum_{j=1}^{n} |x_{ij} - z_j^\top a| + \lambda_1 \|a\|_1 \right\}
\]

(22)

for \( i = 1, \ldots, m \), when \( B \) and \( C \) are fixed.
Selection of the shrinkage parameter

- Bayesian information criteria (BIC):

\[
\text{BIC}(\lambda) = 2N \ln \hat{\sigma} + w \cdot \text{df}(\lambda) \cdot \ln N
\]  

(23)

where \( N = I \cdot J \cdot K \), \( w(= \sqrt{2}) \) a weight assigned by the user, and \( \hat{\sigma} \) is a scale estimate of the residuals.

\[
\hat{\sigma}^2 = \text{ave}_{i,j}\{r_{i,i}^2\} \text{ for CPA-LASSO}, \\
\hat{\sigma} = \text{ave}_{i,j}\{|r_{i,j}|\} \text{ for CPA-LADLASSO},
\]

- Degrees of freedom of the model \( \text{df}(\lambda) \) :

sum of the number of non-zero elements in factor matrices (\( \hat{A}, \hat{B}, \hat{C} \)).
Simulations

Model
The observed three-way tensor is generated as $\mathbf{X} = \mathbf{X}_0 + \mathbf{E}$, where $\mathbf{X}_0 = \sum_{r=1}^{R} \gamma_r \mathbf{a}_r \odot \mathbf{b}_r \odot \mathbf{c}_r$ is the Kruskal tensor, $\mathbf{E}$ is the noise tensor and the rank $R$ is assumed to be known.

The factor matrices and the true noise-free three-way tensor $\mathbf{X}_0$ is sparse.

The accuracy of the obtained estimate $\hat{\mathbf{X}}$ can be calculated by the normalized mean squared error (NMSE)

$$\text{NMSE}(\hat{\mathbf{X}}) = \frac{\|\mathbf{X}_0 - \hat{\mathbf{X}}\|_2^2}{\|\mathbf{X}_0\|_2^2}.$$
Simulations - Measure of performance

- 2 × 2 contingency table

<table>
<thead>
<tr>
<th></th>
<th>True A = 0</th>
<th>True A ≠ 0</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>n_{1C}</td>
<td>n_{1M}</td>
<td>n_{1}</td>
</tr>
<tr>
<td>≠ 0</td>
<td>n_{2M}</td>
<td>n_{2C}</td>
<td>n_{2}</td>
</tr>
<tr>
<td>sum</td>
<td>n_{1}'</td>
<td>n_{2}'</td>
<td>I · R</td>
</tr>
</tbody>
</table>

where $n_{1C}$ (resp. $n_{2C}$) is the number of entries in the estimate $\hat{A}$ “correctly classified” as being zero (resp. non-zero) and $n_{1M}$ (resp. $n_{2M}$) is the number of entries in $\hat{A}$ “misclassified” as being non-zero (resp. zero).

- The classification error rate: $\text{CER}(\hat{A}) = (n_{1M} + n_{2M})/(I \cdot R)$
- or recovery rate: $\text{RER}(\hat{A}) = 1 - \text{CER}(\hat{A})$. 

12th December 2015 | Aalto University and Technische Universität Darmstadt | V. Koivunen and A.M. Zoubir | 174
Simulations (for Sparsity and Robustness)

- Simulation Setting:
  The heavy-tailed noise tensor $\mathcal{E} \in \mathbb{R}^{1000 \times 20 \times 20}$ from the Cauchy distribution with symmetry center 0 and scale parameter 1/2 is added in place of the normal noise tensor to $\mathcal{X}_0$ generated in Simulation setting I.

- The penalty parameter is selected by minimizing the BIC with the weight $w = \sqrt{2}$ over a grid of $\lambda = \lambda_1 = \lambda_2 = \lambda_3$ values for the computational feasibility.
The robust sparse methods, LAD-LASSO shows excellent performance, whereas CP-ALS and the CP alternating LASSO methods yield poor estimates.
Conclusions

- Multi-linear techniques using tensor decompositions provide a unifying framework for the high-dimensional data analysis.
- Sparsity enables us to extract some essential features from a big data that are easily interpretable.
- Robust (regularized) tensor decompositions clearly improves the analysis and inference of multi-dimensional data.
Additional Material: Geodesic Convexity and Regularized M-estimator
Geodesic \((g-\text{convexity})\)

Geodesic path from \(\Sigma_0 \in \mathcal{H}(p)\) to \(\Sigma_1 \in \mathcal{H}(p)\)

\[\Sigma_t = \left(\Sigma_0^{-1/2}\Sigma_1\Sigma_0^{-1/2}\right)^t \Sigma_0^{1/2}\]

for \(t \in [0, 1]\).

where \(\Sigma_t \in \mathcal{H}(p)\) for \(0 \leq t \leq 1 \Rightarrow \mathcal{H}(p)\) forms a \(g\)-convex set (= all geodesic paths \(\Sigma_t\) lie in \(\mathcal{H}(p)\)).

- Main idea: change the parametric path going from \(\Sigma_0\) to \(\Sigma_1\).
- Mid-point of the path \(\Sigma_{1/2}\) is the geometric mean between \(\Sigma_0\) and \(\Sigma_1\).

**Example:** Paths between diagonal matrices \(\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2)\):

![Geodesic path](image1)

![Euclidean path](image2)
Geodesically (g-)convex function

A function \( h : \mathcal{H}(p) \to \mathbb{R} \) is \( g \)-convex function if

\[
h(\Sigma_t) \leq (1 - t) h(\Sigma_0) + t h(\Sigma_1) \quad \text{for} \quad t \in (0, 1).
\]

If the inequality is strict, then \( h \) is strictly \( g \)-convex. See [Wiesel, 2012].

**Note:** Def. of convexity of \( h(\Sigma) \) remains the same, i.e., w.r.t. to given path \( \Sigma_t \). Now geodesic instead of Euclidean path.

**Properties of \( g \)-convex functions**

1. any local minimum of \( h(\Sigma) \) over \( \mathcal{H}(p) \) is a global minimum.
2. If \( h \) is strictly \( g \)-convex and a minimum is in \( \mathcal{H}(p) \), then it is a unique minimum.
3. \( g \)-convex + \( g \)-convex = \( g \)-convex
Penalized $M$-estimating equation

The penalized cost fnc

$$\mathcal{L}_\alpha(\Sigma) = \frac{1}{n} \sum_{i=1}^{n} \rho(z_i^H \Sigma^{-1} z_i) - \ln |\Sigma^{-1}| + \alpha \text{Tr}(\Sigma^{-1})$$

where $\alpha \geq 0$ is a fixed regularization parameter.

**Condition 1.** [Zhang et al., 2013, Ollila and Tyler, 2014]

- $\rho(t)$ is nondecreasing and continuous for $0 < t < \infty$.
- $\rho(t)$ is $g$-convex (i.e., $\rho(e^x)$ is convex in $-\infty < x < \infty$)

**Note:** Tyler’s, Huber’s, Gaussian loss fnc $\rho(t)$ satisfies Cond. 1.

**Result 1:** [Ollila and Tyler, 2014]

If $\rho(t)$ satisfies Cond. 1, then $\mathcal{L}_\alpha(\Sigma)$ is strictly $g$-convex in $\Sigma \in \mathcal{H}(p)$. 

---

**Additional Material:**

Geodesic Convexity and Regularized M-estimator
Result 2: [Ollila and Tyler, 2014]

Assume $\rho(t)$ satisfies Condition 1. If $\rho(t)$ is bounded below and differentiable, then $L_\alpha(\Sigma)$ has a unique minimum in $\mathcal{H}(p)$, and the minimum corresponds to the unique solution of the regularized $M$-estimating eq.

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} u(z_i^H \hat{\Sigma}^{-1} z_i) z_i z_i^H + \alpha I.$$ 

If $u(t) = \rho'(t)$ is nonincreasing, then fixed-point algorithm converges.

Why regularization?

Low sample support (sample size $n$ small relative to dimension $p$) is a common problem in many applications. In this case, $M$-estimators do not exists or are estimated with a lot of error, whereas regularized $M$-estimators do not require any conditions on the data. Even $n = 1$ suffices and the algorithm always converges to unique minimum.
Example 1: regularized SCM

- Tuned Gaussian cost function $\rho(t) = \beta t$, where $\beta > 0$ is a fixed tuning parameter.
- The penalized cost function is

$$\mathcal{L}_\alpha(\Sigma) = \text{Tr}\{ (\beta \hat{\Sigma} + \alpha I)\Sigma^{-1} \} - \ln |\Sigma^{-1}|$$

where $\hat{\Sigma}$ denotes the sample covariance matrix (SCM).
- Due to Result 2, its unique minimizer is

$$\hat{\Sigma}_{\alpha,\beta} = \beta \hat{\Sigma} + \alpha I$$

which corresponds to shrinkage estimator due to [Ledoit and Wolf, 2004].
- Note: In [Ledoit and Wolf, 2004], $\hat{\Sigma}_{\alpha,\beta}$ was not shown to solve an optimization problem.
Example 2: regularized Tyler’s $M$-estimator

- Tuned Tyler’s cost fnc $\rho(t) = p\beta \log t$ for fixed $0 < \beta < 1$.
- The penalized Tyler’s cost fnc is
  \[
  L_\alpha(\Sigma) = \frac{\beta}{n} \sum_{i=1}^{n} \log(z_i^H \Sigma^{-1} z_i) - \ln |\Sigma^{-1}| + \alpha \text{Tr}(\Sigma^{-1}),
  \]
- The weight fnc is $u(t) = p\beta/t$, so the regularized $M$-estimating eq. is
  \[
  \hat{\Sigma} = \frac{p\beta}{n} \sum_{i=1}^{n} \frac{z_i z_i^H}{z_i^H \hat{\Sigma}^{-1} z_i} + \alpha I,
  \]
- Note: $\rho$ is not bounded below, so Result 2 does not hold.
  Conditions for existence and uniqueness for the regularized Tyler’s $M$-estimators can be found in [Ollila and Tyler, 2014, Theorem 4].
- Related works: [Abramovich and Spencer, 2007, Chen et al., 2011, Pascal et al., 2014]
Excerpt of Useful References
Literature

Robust Estimation Overview


Literature

Complex Valued and Multichannel Data

Literature

Complex Valued and Multichannel Data


Literature

Complex Valued and Multichannel Data


DOA for Stationary Data

Literature
DOA for Non-stationary Data


Literature

Multiuser Detection

Literature
Spectrum Sensing


Literature
Model Selection

Literature

Bootstrap


Literature

Distributed Data

Literature

Dependent Data

Literature
Dependent Data

Literature

Robust Detection

Robust Tensor Decomposition