On Combinations of Sign and Peripherality Transformations for Robust Analysis of Hilbert-valued data

SM, SC

August 12, 2016

1 Introduction

???? I will write this section once the rest of the work is complete ????

The scope of Hilbert-valued data is considerable, and encompasses the univariate and multivariate finite-dimensional observations traditionally studied in Statistics, as well as certain kinds of functional data. For $\mathbb{R}^p$-valued data, the classical approach is to model the joint distribution of the observations using a probability distribution known up to finite dimensional parameters. If the correct probability distribution is used, there are often theoretical optimality guarantees. However in practice, the correct distribution is rarely, if ever, known; there may be aberrant observations that do not follow the pattern of the bulk of the data but whose influence on parameter estimates and eventual inference can be considerable; computational practicability and costs and complexity of patterns in the data may limit choices of statistical methods to adopt. The analysis of functional data is in a developmental stage, but the above comments on practical issues are applicable for such data also.

One way of circumventing the practical issues outlined above is to use *robust* techniques of data analysis, where theoretical optimality under restrictive idealized conditions is not of interest. Instead, efforts are made to make the resulting estimation and inference insensitive to technical assumptions, outliers in the data, and computationally feasible. Naturally, subject to these robustness guarantees, the goal is to use the most efficient techniques possible. The following very simple example recapitulates some very well known facts, but serves to illustrate the main drive of this paper.

Example 1.1(*Testing for a location parameter):* Suppose we have real valued observations $Y_1, \ldots, Y_n$, which from the domain of study that gave rise to these data, are known to be distributed around an unknown parameter $\mu \in \mathbb{R}$, and the scientific question of interest lies in evaluating whether the hypothesis $\mu = \mu_0$ for some given constant μ_0 is tenable or not.

Under the assumption that the observations $Y_1, \ldots, Y_n$ are independent, identically distributed (i.i.d. hereafter) as $N(\mu, \sigma^2)$, the Gaussian distribution with mean μ and variance σ^2, the optimal procedure is to use the t-test based on the statistic

$$T_n = n^{1/2}(\bar{Y}_n - \mu_0)/S_{n-1}$$

and the Student's t-distribution. Here $\bar{Y}_n = n^{-1}\sum_{i=1}^n Y_i$ and $S_{n-1}^2 = (n-1)^{-1}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$. However, note that the assumptions of the data being independent, or identically distributed, or following the Gaussian distribution, or having the same variance for all observations, are rarely justified from the domain of study, and are merely choices of convenience. Also, one or more outlying observations among $Y_1, \ldots, Y_n$ can have extreme effect on either or both of $\bar{Y}_n$ and S_{n-1}^2, and result in a numeric value that completely misrepresents the nature of the data and misleading inference relating to the scientific query.

Alternatives to the above is to use the *sign*, and possibly also the *rank* of each observation. Let $\mathcal{I}_A$ be the indicator of A, that is, $\mathcal{I}_A = 1$ if A is true and is zero otherwise. Define the following:

$$(a) \text{ Sign function } S(y; \nu) = |y - \nu|^{-1}(y - \nu)\mathcal{I}_{\{y \neq \nu\}},$$
$$(b) \text{ Rank function } R(y; Y_1, \ldots, Y_n) = \sum_{i=1}^n \mathcal{I}_{\{y \leq Y_i\}}.$$

Then $S_i = S(Y_i; \mu_0)$ indicates whether the i^{th} observation is greater than μ_0 or not, and $R_i = R(Y_i; Y_1, \ldots, Y_n)$ is the rank that the i^{th} observation takes among the entire data when sorted in an ascending order.

The *sign test statistic* corresponding to the above location testing problem is given by $S = \sum_{i=1}^n S_i$ and under the very general assumption that the Y_i's are independent with distribution function F_i, and all the F_i's have a common median μ_0, the null distribution of S is Binomial $(n, 0.5)$. This testing procedure is also indeed optimal in a very broad framework, see **???? Need the appropriate reference from hajek, also need the conditions from there ????**

However, in many situations some additional information is known beyond the extremely relaxed conditions required by the sign test, and the Wilcoxon signed-rank test, is found to be considerably more efficient than the sign test. These are well known results from several decades back, see **???? Standard Hollander and Wolfe type refernce ????**for details.

□

Example 1.1 motivates the present paper: which is a study on combinations of the equivalents of the sign function and rank function in Hilbert spaces, that leads to robust statistical procedures of considerable efficiency in a wide variety of problems. Very interestingly, we find that for location parameters, using in conjunction with the sign function a weight function that assigns *more* mass to observations closer to the center, results in robust and considerably efficient estimators and test statistics. On the other hand, for scale or dispersion parameters, in conjunction with the sign function using a weight function that assigns *less* mass to observations closer to the center, results in similar properties.

Notice that the robust and potentially optimal procedures outlined in Example 1.1 above are based on transformations of the original data that preserved the structures of interest. In the rest of this paper we study equivalent transformations on general Hilbert spaces, and use the transformed data for (*i*) robust location estimation, (*ii*) testing for a location parameter in high dimensions and Hilbert spaces, (*iii*) robust dispersion estimation, (*iv*) robust principal component analysis (RPCA), (*v*) RPCA in dimension reduction and learning, (*vi*) robust analysis of functional data.

Given the considerable breadth of these topics, we present algorithmic details, some illustrative examples, and a selection of theoretical results. Considerably deeper studies in any of the above topics is feasible, however we postpone deep-dives into any specific application to future papers in order to make the main methodological drivers of this paper more easily understandable.

In Section 2 we present the *sign function* and the *peripherality function* for Hilbert spaces. The former is well known, and variants of the latter are also well known in the literature. One major contribution of this paper is in the proposal of combining the two in different ways to achieve simultaneous robustness and efficiency targets. Then on the following sections we present details on the six applications listed above.

2 The sign and peripherality functions

We consider data $X_1, \ldots, X_n$ from some subset $\mathcal{X}$ of a real separable Hilbert $\mathcal{H}$.

We consider two functions defined below. First, we consider the *sign function* $S : \mathcal{H} \times \mathcal{H} \to \mathcal{H}$, defined as

$$S(x; \mu_x) = ||x - \mu_x||^{-1}(x - \mu_x)\mathcal{I}_{\{x \neq \mu_x\}}.$$

This sign function is defined with respect to the *location parameter* $\mu_x \in \mathcal{H}$, and the norm $|| \cdot ||$ used above is the norm of the underlying Hilbert space. This is a direct generalization of the real-valued case of the indicator of whether the point x is to the right, left or at μ_x, described in Example 1.1. This function has been used many times in statistics, see **???? insert several references. ????**

We next describe the *peripherality function*, for which some mathematical preliminaries are necessary for easier exposition. Let $(\Omega, \mathcal{A}, \alpha)$ be a probability space, and let $\mathcal{B}$ be the Borel σ-algebra generated by the norm topology of $\mathcal{H}$. A $\mathcal{H}$-valued random variable is a mapping $X : \Omega \to \mathcal{H}$ such that for every $B \in \mathcal{B}$, $X^{-1}(B) \in \mathcal{A}$. It is easy to see that $\alpha_x = \alpha(X^{-1}(\cdot))$ is a probability measure on the measurable space $(\mathcal{H}, \mathcal{B})$. Mathematical details about such probability measures on Hilbert spaces are available from a number of places, including **???? BLSP notes (my primary reference), Gross, Segal, and what not. ????**

Let $\mathcal{M}$ be a set of probability measures on $\mathcal{H}$. A *peripherality function* $P : \mathcal{H} \times \mathcal{M} \to \mathbb{R}$, is a function that satisfies the following condition:
For every probability measure $F \in \mathcal{M}$, there exists a constant $\mu_F \in \mathcal{H}$ such that for every $t \in [0, 1]$ and every $x \in \mathcal{H}$

$$P\Big(\mu_F; F\Big) \leq P\Big(\mu_F + t(x - \mu_F); F\Big).$$

That is, for every fixed F, the peripherality function achieves a minimum at μ_F, and is non-decreasing in every direction away from μ_F. If we impose the practical restriction that $\inf_x P(x; F)$ is finite and bounded below, then we may as well impose without loss of generality $P(\mu_F; F) = 0$ and consequently $P(x; F) \geq 0$ for all $x \in \mathcal{H}$ and $F \in \mathcal{M}$. In many cases of interest, $P(\cdot; \cdot)$ is uniformly bounded above as well.

The peripherality function quantifies whether the point x is near or far from μ_F. We will impose additional conditions on this function as we proceed, but it can be seem immediately that any distance measure between x and μ_F satisfies the bare minimum requirement mentioned above.

In this paper, we demonstrate certain interesting applications arising from composing the sign function and the peripherality function together, to form the *signed-peripherality function* $\kappa(\cdot)$. We define this function with three parameters $\mu_x \in \mathcal{H}$, $F \in \mathcal{M}$ and $\mu_y \in \mathcal{H}$, argument $x \in \mathcal{H}$ and range $\mathcal{H}$. More precisely, we use two functions $\kappa_s : \mathcal{H} \to \mathcal{H}$, $\kappa_p : \mathcal{H} \to \mathcal{H}$ that are respectively composed with the sign transformation and the peripherality function, and then multiplied together to obtain the function $\kappa : \mathcal{H} \times \mathcal{H} \times \mathcal{M} \times \mathcal{H} \times \mathcal{H} \to \mathcal{H}$ defined as

$$\kappa(x; \mu_x, F, \mu_y) = k_s(S(x; \mu_x))k_p(P(x; F)) + \mu_y.$$

Within this very generic framework, we will explore two simple choices. We consider $\kappa_s(x) = x$, thus this is fixed to be the identity transformation. The two alternatives we consider for κ_p are $\kappa_p(x) = x$ and $\kappa_p(x) = \exp(-x)$, thus one is linearly increasing with x while the other is exponentially decreasing.

Notice that if we consider $\mu_y = \mu_F = \mu_x$, $\kappa_s(x) = \kappa_p(x) = x$, and take the very simple peripherality function $P(x; F) = ||x - \mu_F||$, we have $\kappa(x; \mu_x, F, \mu_y) \equiv x$ for all choices of parameters μ_x, F, μ_y. Consequently, under this choice of parameters for the κ-transformation, analyzing a dataset $\{X_1, \ldots, X_n\}$ and its κ-transformed version $\{Y_i = \kappa(X_i; \ldots),\ i = 1, \ldots, n\}$ are equivalent. However, in this paper we illustrate how other choices of the peripherality function lead to interesting robustness results. We have deliberately set the location parameters μ_x, μ_F, μ_y to be potentially non-identical, this additional flexibility has some advantage for robust data analysis. In many applications, the value of these three parameters may be identical, which leads to no conflict in our framework.

A whole class of peripherality functions can be defined from *data - depth*, which are center-outward ranking of multivariate data. Data-depths have been extensively used in statistics also, see **???? multiple references. ????**Peripherality functions can be defined as some inverse ranking based on data depth, and the concept of *outlyingness* associated with data depth is essentially same as what we use in this paper. We use the term *peripherality* to keep track of the difference in application contexts and technical assumptions. As an additional objective of this paper, we discuss some properties and uses of data-depth in real, separable Hilbert spaces.

old material:*In this paper, we consider a few illustrative cases of the use of the κ-transformation. Suppose the data at hand is $X_1, \ldots, X_n$, and we define $Y_i = \kappa(X_i; \mu_X, F, \mu_Y)$ for some choice of parameters μ_X, F, μ_Y. For interpretability and convenience, we assume that $\mathbb{E}S(X_i; \mu_X)P(X_i; F) = 0$, thus $\mathbb{E}Y_i = \mu_Y$. We thus have*

$$\begin{aligned}\mathbb{V}Y_i &= \mathbb{E}P(X_i; F)^2 S(X_i; \mu_X) S(X_i; \mu_X)^T \\ &= \mathbb{E}P(X_i; F)^2 ||X_i - \mu_X||^{-2} (X_i - \mu_X)(X_i - \mu_X)^T.\end{aligned}$$

???? Need to include (a) Biman-PC idea for affine equivariance for $p \ll n$, (b) kernel versions as an example of generalization. (c) anything else? ????

3 The robust location problem

3.1 General approach

Here are things we need to do:

1. Propose robust location estimation in a general Hilbert space $\mathcal{H}$, based on $\kappa_s(x) = x$ and $\kappa_p(x) = \exp(-x)$. I am using $\exp(-x)$ and not $1/x$ (on which you have done some simulations successfully I think) so that we do not have unboundedness issues.

2. Suggest that this may be a better location estimate than the spatial median that is obtained by just minimizing the sign function.

3. Show whatever theoretical results we have. Make these completely error-free, so no handwaving or loose, non-crisp arguments.

4. Do the infinite dimensional case also. Our strongest point in this paper is being able to handle robustness in infinite dimensions. There may not be any useful or usable definition of invariance or equivariance in infinite dimensions. So separate the results for the $p \ll n$ case and the high-dimensional case, if necessary.

5. Show simulation results.

6. Do one running data analysis in infinite dimensions. Working with extreme precipitation or temperatures is a good idea, there is a potentially a need for robust location estimation there.

7. Let us finish one section first, then move to another section.

3.2 The following are concrete suggestions:

1. Describe what is being observed, ie, the main aspects of the data. For example, are you assuming iid data throughout?

2. After you have described the data, define the parameter you are interested in. The population mean, population median, the population equivalent of the minimizer of the weighted sign function are not all same. However, each of them can be a candidate for a *location parameter*. Don't skip the mathematical details, include everything.

3. Then propose your estimator for the parameter of your choice.

4. Now explore properties of your proposed estimator. Simulate data and show that your estimator is better than the (sample) mean in terms of robustness, and better than the (sample) median in terms of efficiency. The following steps describe what you might simulate.

5. Use the el Nino data as a template. Do your robust PCA, and extract the first 5 or so PCs with their loadings $\beta_1, \beta_2, \ldots, \beta_5$. This will form the foundation for the simulation. For the time index of the data, make it $t \in [0, 1)$ to avoid technical difficulties. That is, consider June as $t = 0$, July as $t = 1/12$ and so on. This simulation will also form the spine of the paper for each application (we will use this simulated data for all the sections), so do this carefully. Use `set.seed`, so we can reproduce results.

6. We will generate n curves indexed by $i = 1, \ldots, n$, all of which will be sampled at time points $t_1, \ldots, t_T$. That is, the observation times for all the curves are same. You can take $T = 12$ and $t_j = (j - 1)/12$.

7. For simulation, first generate $b_{ij} \sim N(\beta_j, V_j)$ for $j = 1, \ldots, 5$, and define the simulated signal for the i^{th} curve as

$$\xi_i(t) = \sum_{j=1}^{5} b_{ij} PC_j(t), \; i = 1, \ldots, n.$$

   Use V_2 very large, so that in different signals there is considerable variability in the signal.

8. Add a iid high-variance noise, say t_2 or something to each observed data point t_{ik}, $k = 1, \ldots, T_i$, $i = 1, \ldots, n$. Thus, the final simulated data is

$$Y_i(t_{ik}) = \xi(t_{ik}) + \varepsilon(t_{ik}).$$

9. While this is supposed to be functional data, all we have really is a $n \times T$ table. Get the T-dimensional sample mean (minimizer of squared Euclidean norm), and median (ie, the minimizer of the Euclidean norm, ie, Haldane's median, ie, the quantity that solves $\sum S(Y_i; \theta) = 0$.) Also get your proposed weighted median.

10. Show that the weighted median and original median are robust, and the weighted median is closer than the median to the "population mean"

$$\sum_{j=1}^{5} \beta_j PC_j(t), t = 1, \ldots, T.$$

11. Remember that although we are calling this FDA, this is really 12-dim data analysis.

3.3 Subho's write-up

Consider the general situation of estimating or testing for the location parameter of an elliptical distribution using weighted sign vectors. **???? What is an elliptical distribution? In any scientific writing, for any technical term you must (a) either define it if it is brand new, (b) give the formula with proper citations if somebody has already defined it (in this case, there should be a definition in Fang's book, maybe in several other places), (c) prove every claim that you are making, or provide a reference from a trustworthy source. ????**For now the only condition we impose on these weights, say $w(\cdot)$, is that they need to be scalar-valued affine equivariant and square-integrable functions of the data, or in other words functions of the norm of the standardized random variable $\mathbf{Z}$. **???? I am not sure if in FDA context, "affine equivariance" is defined. Find out. If not, then what you have above is a good definition. ????**In that sense $w(\mathbf{X})$ can be equivalently written as $f(r)$, with $r = \|\mathbf{Z}\|$. **???? Are we always working with square integrable functions of the data? (to begin with, what do you mean by that? Do you mean that the data is iid random variables all of which have finite second moments? What is a standardized random variable? ????**

The simplest use of weighted signs here would be to construct a robust alternative to the Hotelling's T^2 test **???? reference ????**using their sample mean vector and covariance matrix. **???? what sort of robust alternative? provide details? always provide 100% exact mathematicalstatements ????**Formally, this means testing for $H_0 : \boldsymbol{\mu} = \mathbf{0}_p$ vs. $H_1 : \boldsymbol{\mu} \neq \mathbf{0}_p$ based on the test statistic: **???? are you testing for the mean? Or a different parameter? If it is the latter, then it's not the same test. Requires discussion and clarification ????**

$$T_{n,w} = n\bar{\mathbf{X}}_w^T (Cov(X_w))^{-1} \bar{\mathbf{X}}_w$$

with $\bar{\mathbf{X}}_w = \sum_{i=1}^n \mathbf{X}_{w,i}/n$ and $\mathbf{X}_{w,i} = w(\mathbf{X}_i)\mathbf{S}(\mathbf{X}_i)$ for $i = 1, 2, ..., n$. However, the following holds true for this weighted sign test:

Proposition 3.1. *Consider n random variables $Z = (\mathbf{Z}_1, ..., \mathbf{Z}_n)^T$ distributed independently and identically as $\mathcal{E}(\boldsymbol{\mu}, kI_p, G); k \in \mathbb{R}$, and the class of hypothesis tests defined above. Then, given any $\alpha \in (0, 1)$, local power at $\boldsymbol{\mu} \neq \mathbf{0}_p$ for the level-α test based on $T_{n,w}$ is maximum when $w(\mathbf{Z}_1) = c$, a constant independent of $\mathbf{Z}_1$.*

This essentially means that power-wise the (unweighted) spatial sign test [**?**] is optimal in the given class of hypothesis tests when the data comes from a spherically symmetric distribution. Our simulations show that this empirically holds for non-spherical but elliptic distributions as well.

3.4 The weighted spatial median

Our weight functions are affine equivariant functions of the data, i.e. they are not affected by the population location parameter $\boldsymbol{\mu}$. This ensures that there exists a unique solution of the following of optimization problem:

$$\boldsymbol{\mu}_w = \arg \min_{\boldsymbol{\mu}_0 \in \mathbb{R}^p} E(w(\mathbf{X})|\mathbf{X} - \boldsymbol{\mu}_0|)$$

This can be seen as a generalization of the Fermat-Weber location problem (which has the spatial median [**? ?**] as the solution) using data-dependent weights. We call its solution, $\boldsymbol{\mu}_w$, the *weighted spatial median* of F. In a sample setup it is estimated by iteratively solving the equation $\sum_{i=1}^n w(\mathbf{X}_i)\mathbf{S}(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_w)/n = \mathbf{0}_p$.

The following theorem shows that the sample weighted spatial median $\hat{\boldsymbol{\mu}}_w$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\mu}_w$, and gives its asymptotic distribution:

Theorem 3.1. *Let A_w, B_w be two matrices, dependent on the weight function w such that*

$$A_w = E\left[\frac{w(\boldsymbol{\epsilon})}{\|\boldsymbol{\epsilon}\|}\left(1 - \frac{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T}{\|\boldsymbol{\epsilon}\|^2}\right)\right]; \quad B_w = E\left[\frac{(w(\boldsymbol{\epsilon}))^2\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T}{\|\boldsymbol{\epsilon}\|^2}\right]$$

where $\boldsymbol{\epsilon} \sim \mathcal{E}(\mathbf{0}_p, \Sigma, G)$. Then

$$\sqrt{n}(\hat{\boldsymbol{\mu}}_w - \boldsymbol{\mu}_w) \rightsquigarrow N_p(\mathbf{0}_p, A_w^{-1}B_wA_w^{-1}) \tag{1}$$

We provide a sketch of its proof in the supplementary material, which generalizes equivalent results for the spatial median [**?**]. Setting $w(\boldsymbol{\epsilon}) = 1$ above yields the asymptotic covariance matrix for the spatial median. Following this, the asymptotic relative efficiency (ARE) of $\boldsymbol{\mu}_w$ corresponding to some non-uniform weight function with respect to the spatial median, say $\boldsymbol{\mu}_s$ will be:

$$ARE(\boldsymbol{\mu}_w, \boldsymbol{\mu}_s) = \left[\frac{\det(A^{-1}BA^{-1})}{\det(A_w^{-1}B_wA_w^{-1})}\right]^{1/p}$$

with $A = E[1/\|\boldsymbol{\epsilon}\|(I_p - \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T/\|\boldsymbol{\epsilon}\|^2)]$ and $B = E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T/\|\boldsymbol{\epsilon}\|^2]$. This is further simplified under spherical symmetry:

Corollary 3.1. *For the spherical distribution $\mathcal{E}(\boldsymbol{\mu}, kI_p, G); k \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^p$, we have*

$$ARE(\boldsymbol{\mu}_w, \boldsymbol{\mu}_s) = \frac{\left[E\left(\frac{f(r)}{r}\right)\right]^2}{Ef^2(r)\left[E\left(\frac{1}{r}\right)\right]^2}$$

4 A high-dimensional test of location

It is possible to take an alternative approach to the location testing problem by using the covariance-type U-statistic $C_{n,w} = \sum_{i=1}^n \sum_{j=1}^{i-1} \mathbf{X}_{w,i}^T \mathbf{X}_{w,j}$. This class of test statistics are especially attractive since they are readily generalized to cover high-dimensional situations, i.e. when $p > n$. The Chen and Qin (CQ) high-dimensional test of location for multivariate normal $\mathbf{X}_i$ [**?**] is a special case of this test that uses the statistic $C_n = \sum_{i=1}^n \sum_{j=1}^{i-1} \mathbf{X}_i^T \mathbf{X}_j$, and a recent paper ([**?**], from here on referred to as WPL test) shows that one can improve upon the power of the CQ test for non-gaussian elliptical distributions by using spatial signs $\mathbf{S}(\mathbf{X}_i)$ in place of the actual variables.

Given these, and some mild regularity conditions, the following holds for our generalized test statistic $C_{n,w}$ under H_0 as $n, p \to \infty$:

$$\frac{C_{n,w}}{\sqrt{\frac{n(n-1)}{2}\mathrm{Tr}(B_w^2)}} \rightsquigarrow N(0,1) \tag{2}$$

and under contiguous alternatives $H_1 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$,

$$\frac{C_{n,w} - \frac{n(n-1)}{2}\boldsymbol{\mu}_0^T A_w^2 \boldsymbol{\mu}_0 (1 + o(1))}{\sqrt{\frac{n(n-1)}{2}\mathrm{Tr}(B_w^2)}} \rightsquigarrow N(0,1) \tag{3}$$

we provide the details behind deriving these two results in the supplementary material, which involve modified regularity conditions and sketches of proofs along the lines of [**?**].

Following this, the ARE of this test statistic with respect to its unweighted version, i.e. the WPL statistic, is expressed as:

$$ARE(C_{n,w}, \mathrm{WPL}; \boldsymbol{\mu}_0) = \frac{\boldsymbol{\mu}_0^T A_w^2 \boldsymbol{\mu}_0}{\boldsymbol{\mu}_0^T A^2 \boldsymbol{\mu}_0} \sqrt{\frac{\mathrm{Tr}(B^2)}{\mathrm{Tr}(B_w^2)}} (1 + o(1))$$

when $\Sigma = kI_p$, this again simplifies to $E^2(f(r)/r)/[Ef^2(r).E^2(1/r)]$.

(need to write the sketch in supplementary material)

5 SC: (older material) Testing for a location parameter

Suppose $W \in \{0, 1\}$, $Z \in \mathbb{R}^p$ and $U \in \mathbb{R}^p$ are independent random variables, with W having a Bernoulli distribution with parameter θ, and each element of U is an independent standard Cauchy random variable. The random vector Z follows a mean-zero, variance matrix Σ Normal distribution $Z \sim N_p(0, \Sigma)$. Our choices of Σ in the details below include cases where $Cov(Z_i, Z_j) = \rho^{|i-j|}$ for some $\rho \in [0, 1)$ which includes as special case the identity matrix $\mathbb{I}_p$ corresponding to $\rho = 0$, and where $Cov(Z_i, Z_j) = \rho$.

We define $X = \mu + WZ + (1-W)U$, and let the observed data be independent, identically distributed copies $X_1, \ldots, X_n$ copies of X. Note that the conditional distribution of $(X|W = 0)$ does not have finite moments. Suppose our goal is to conduct inference on μ, in particular, to test the hypothesis that $\mu = 0 \in \mathbb{R}^p$. The sample mean and variance from the observed data $X_1, \ldots, X_n$, denoted respectively by $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ and $S_X = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T$. These are well-defined quantities, even though their expectations may not exist. Assume $p < n$, and thus S_x is invertible.

One option for inference on μ is to ignore the heavy-tailed component of the data, and to use Hotteling's T^2 statistic given by $T^2 = n(\bar{X} - \mu_0)^T S_X^{-1}(\bar{X} - \mu_0)$.

Other, more robust options, is to use the sign function exclusively, or any variety of data-depth functions exclusively. **???? suggest 2 or 3 depth functions used earlier for one-sample mean testing, provide references. List details on how these methods would be implemented, just do what these references do ????**

Another alternative is to use the κ-transformation proposed in this paper. We begin with a robust estimator of μ, for example the co-ordinatewise median, to use as μ_X and μ_F. As F, we use the p-dimensional Normal distribution with mean μ_F and identity covariance matrix. One simple peripherality function is $||x - mu_F||/(1 + ||x - mu_F||)$, which is bounded above and below and achieves a minima at μ_F. We take μ_Y to be μ also, and we may again use the co-ordinatewise median or a different statistic as its estimator like simple sample mean of $Y_i = \kappa(X_i)$. (Use the sample mean: we want to sell the idea that after the kappa transformation all standard classical methods like sample moments can be used).

We should be able to show that

1. Classical methods like Hotteling is a disaster because of outliers.

2. Sign and depth-based methods lack power, which becomes quite bad when ρ is considerably away from zero.