Contents

1 Statistical Methods IV: Median

Course Information

- **Instructor:** Shyamal K De
- **Department:** ASU
- **Email:** skd.isical@gmail.com

- **Marking Scheme:** Class Test 10%, Project 15%, Midterm 25%, End Semester 50%

Author Information

- **Name:** Mohammad Shaan
- **Academic Year:** B.Stat II Year
- **Email:** mdsworld2006@.com

1.1 Computation: what is *the* sample median, really?

Up to now, the sample median $\hat{\mu} = T(F_n)$ was defined abstractly via several equivalent population characterizations. However, when one plugs in the *empirical CDF* F_n:
- those definitions *need not give a unique solution*;
- therefore, a *convention* is imposed.

1.1.1 Procedure

1. Order the data:
$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}.$$
2. Define:
$$\hat{\mu} = \frac{x_{[(n+1)/2]} + x_{[(n+2)/2]}}{2}.$$

1.1.2 Interpretation

- If n is odd, both indices coincide and $\hat{\mu}$ is the usual middle order statistic.
- If n is even, $\hat{\mu}$ is the average of the two central observations.

1.1.3 Why this matters

- This convention ensures symmetry and equivariance.
- It matches the estimating–equation viewpoint $\hat{R}(\hat{\mu}) = 0$.

1.1.4 Important remark

In the multivariate case there is no natural ordering.

This is not a cosmetic issue. Everything above relies fundamentally on order. Once order disappears, the notion of a median becomes genuinely geometric.

1.2 Robustness: why statisticians love the median

Two core robustness concepts appear here.

1.2.1 Breakdown point $= \frac{1}{2}$

The *asymptotic breakdown point* is the smallest fraction of contamination that can drive an estimator arbitrarily far.

For the median,

$$\varepsilon^* = \frac{1}{2}.$$

Meaning

- Almost half the data can be replaced by arbitrarily bad outliers.
- The median still does not explode.

No location estimator can do better. This is *maximal robustness.*

1.2.2 Bounded influence function

The influence function of the median is

$$\mathrm{IF}(x; T, F) = \delta^{-1} S(x - T(F)), \qquad \delta = 2f(\mu).$$

Key features

- It takes only three values: $(\pm\delta^{-1}, 0)$.
- It does *not* grow as $|x| \to \infty$.

Interpretation

- A single extreme outlier has a *limited effect*.
- Contrast this with the mean, whose influence function is linear and unbounded.

This formally explains why the median resists outliers.

1.3 Asymptotic efficiency: the price of robustness

Now the comparison with the sample mean begins.

1.3.1 Assumption

- The distribution F has finite variance σ^2.

Then the sample mean satisfies

$$\sqrt{n}(\bar{X} - \mu) \;\to\; N(0, \sigma^2),$$

whereas the sample median satisfies

$$\sqrt{n}(\hat{\mu} - \mu) \;\to\; N\left(0, \frac{1}{4f(\mu)^2}\right).$$

1.3.2 Asymptotic Relative Efficiency (ARE)

The ARE of the median relative to the mean is defined as

$$\mathrm{ARE}(\mathrm{median}, \mathrm{mean}) = \frac{\mathrm{Var}(\mathrm{mean})}{\mathrm{Var}(\mathrm{median})} = 4f(\mu)^2\sigma^2.$$

Interpretation

- $\mathrm{ARE} < 1$: the median is less efficient.
- $\mathrm{ARE} > 1$: the median is more efficient.

Examples

- **Normal** $N(\mu, \sigma^2)$:

$$f(\mu) = \frac{1}{\sqrt{2\pi}\sigma} \quad \Rightarrow \quad \mathrm{ARE} \approx 0.64.$$

- **Heavy-tailed distributions**:
  - t_3: $\mathrm{ARE} \approx 1.62$,
  - Laplace: $\mathrm{ARE} = 2$.

Under Gaussian noise, the mean wins. Under heavy tails, the median dominates.

This is the classic *robustness–efficiency tradeoff*.

1.4 Estimating the variance: a practical headache

To construct confidence intervals for μ using the asymptotic normality of $\hat{\mu}$, one needs

$$\delta = 2f(\mu).$$

1.4.1 Problem

- The density value $f(\mu)$ is unknown.
- Density estimation at a single point is unstable.

Estimation of δ from the data is difficult.

1.5 Exact, distribution-free confidence intervals (the clever trick)

Instead of estimating δ, one can invert the *sign test*.

For a continuous distribution F,

$$P\big(x_{(i)} < \mu < x_{(n+1-i)}\big) = P\left(i \leq \frac{n\hat{R}(\mu)+1}{2} \leq n-i\right).$$

Since

$$\frac{n\hat{R}(\mu)+1}{2} \sim \mathrm{Bin}\Big(n, \tfrac{1}{2}\Big),$$

it follows that

$$P\big(x_{(i)} < \mu < x_{(n+1-i)}\big) = \sum_{j=i}^{n-i} \binom{n}{j} 2^{-n}.$$

1.5.1 Interpretation

- The confidence interval is an *order-statistic interval*.
- The coverage probability is an exact binomial tail.
- No density estimation is required.
- The procedure is fully distribution-free.

This is one of the deepest practical advantages of the median.

1.6 Equivariance: how the median behaves under transformations

For a location functional T, we desire

$$T(F_{aX+b}) = aT(F_X) + b.$$

1.6.1 Meaning

- Shifting the data shifts the estimator.
- Rescaling the data rescales the estimator.

The text notes that this holds for the median when f is smooth near μ, but in fact something stronger is true.

If g is *strictly monotone*, then

$$T(F_{g(X)}) = g(T(F_X)).$$

1.6.2 Interpretation

- Logarithmic, exponential, and power transformations commute with the median.
- The median transforms *exactly* as the data do.

The mean does *not* enjoy this property.

1.7 Influence Function: why the median resists outliers

1.7.1 What is the influence function, conceptually?

The **influence function (IF)** answers one precise question:

> If I contaminate the distribution F by an infinitesimal amount of mass at the point x, how much does my estimator move?

Formally,

$$\mathrm{IF}(x; T, F) = \frac{d}{d\varepsilon} T\big((1-\varepsilon)F + \varepsilon\Delta_x\big)\Big|_{\varepsilon=0}.$$

Here:

- T denotes the estimator viewed as a functional (here: the median),
- Δ_x denotes the point mass at x.

The influence function should be thought of as the **first-order sensitivity** of the estimator to a single outlier placed at x.

1.7.2 Why the sign function appears for the median

The population median $\mu = T(F)$ satisfies the estimating equation

$$\mathbb{E}_F[S(X-\mu)] = 0.$$

This equation *defines* the median.
Now contaminate the distribution slightly:

$$F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_x.$$

The contaminated median $T(F_\varepsilon)$ must still satisfy

$$\mathbb{E}_{F_\varepsilon}\left[S(X - T(F_\varepsilon))\right] = 0.$$

Linearizing this equation around $\varepsilon = 0$ is standard **M-estimator calculus**. The structure of the estimating equation forces the **sign function** to appear.

1.7.3 Where the constant $\delta = 2f(\mu)$ comes from

Differentiate the estimating equation with respect to t:

$$\frac{d}{dt}\mathbb{E}[S(X-t)]\Big|_{t=\mu} = -2f(\mu).$$

Explanation

- The function $S(X-t)$ jumps at $X = t$.
- The derivative picks up mass from the density at $t = \mu$.
- The resulting slope is exactly $-2f(\mu)$.

Conclusion: the median reacts **more strongly** when the density at the median is small, i.e. when the distribution is flat near μ.

1.7.4 The final influence function for the median

Putting everything together,

$$\boxed{\mathrm{IF}(x;T,F) = \frac{1}{2f(\mu)}\,S(x-\mu)}$$

Explicitly,

$$\mathrm{IF}(x;T,F) = \begin{cases} +\dfrac{1}{2f(\mu)}, & x > \mu, \\ 0, & x = \mu, \\ -\dfrac{1}{2f(\mu)}, & x < \mu. \end{cases}$$

1.7.5 Why this is a *big deal*

The influence function is bounded

No matter how large x is,

$$|\mathrm{IF}(x;T,F)| \leq \frac{1}{2f(\mu)}.$$

This is the **formal reason** the median is robust.

Compare this with the mean:

$$\mathrm{IF}_{\text{mean}}(x) = x - \mu,$$

which diverges as $|x| \to \infty$.

Interpretation in plain language

- An observation far above the median pushes it upward by a *fixed amount*.
- An observation far below the median pushes it downward by the *same fixed amount*.
- Extreme values are **automatically clipped**.

The median does not care *how far* the outlier is—only **which side** it lies on.

1.7.6 Connection to the breakdown point

Because the influence function is bounded,

- a single outlier cannot destroy the estimator;
- one needs **half the data** on one side to move the median arbitrarily far.

This is why the breakdown point equals **1/2**.

1.7.7 Mental picture (very important)

Visualize the median as a **balance point**:

- points to the right push right,
- points to the left push left,
- push strength is constant,
- only the *number* of points matters, not their magnitude.

This geometric intuition is encoded exactly by the **sign function** in the influence function.

1.7.8 Why $f(\mu)$ matters

- If $f(\mu)$ is small:
  - the distribution is flat near the median;

  - a small perturbation shifts the median substantially.
- If $f(\mu)$ is large:
  - the median is well anchored.

This also explains the variance formula:

$$\mathrm{Var}(\hat{\mu}) \approx \frac{1}{4nf(\mu)^2}.$$

1.8 Deriving the influence function by linearization

1.8.1 The equation we want to linearize

The median is defined by the **estimating equation**

$$\Psi(t, F) := \mathbb{E}_F[S(X - t)] = 0.$$

The true median μ satisfies

$$\Psi(\mu, F) = 0.$$

Now contaminate the distribution:

$$F_\varepsilon = (1 - \varepsilon)F + \varepsilon\Delta_x,$$

and let the corresponding median be

$$t_\varepsilon := T(F_\varepsilon).$$

By definition, t_ε satisfies

$$\Psi(t_\varepsilon, F_\varepsilon) = 0.$$

This is the equation we will **linearize around** $(\mu, 0)$.

1.8.2 What "linearize" means here

Linearize means:

Take a first-order Taylor expansion of $\Psi(t_\varepsilon, F_\varepsilon)$ around $t = \mu$ and $\varepsilon = 0$.

Nothing more than first-order calculus.

1.8.3 Expansion with respect to both arguments

Write
$$0 = \Psi(t_\varepsilon, F_\varepsilon) \approx \Psi(\mu, F) + \frac{\partial \Psi}{\partial t}\Big|_{\mu,F}(t_\varepsilon - \mu) + \frac{\partial \Psi}{\partial \varepsilon}\Big|_{\mu,0}\varepsilon.$$

Since $\Psi(\mu, F) = 0$, this reduces to

$$0 \approx \frac{\partial \Psi}{\partial t}(\mu, F)(t_\varepsilon - \mu) + \frac{\partial \Psi}{\partial \varepsilon}(\mu, 0)\,\varepsilon.$$

We now compute both derivatives explicitly.

1.8.4 Derivative with respect to t

Recall that
$$\Psi(t, F) = \mathbb{E}_F[S(X - t)].$$

As t increases, the sign function flips at $X = t$. The derivative comes entirely from this jump:

$$\frac{\partial}{\partial t}\Psi(t, F) = -2f(t).$$

Evaluated at $t = \mu$,

$$\frac{\partial \Psi}{\partial t}(\mu, F) = -2f(\mu) = -\delta.$$

This is exactly where the constant $\delta = 2f(\mu)$ comes from.

1.8.5 Derivative with respect to ε

Using the contaminated distribution,

$$\Psi(t, F_\varepsilon) = (1 - \varepsilon)\mathbb{E}_F[S(X - t)] + \varepsilon S(x - t).$$

Differentiate with respect to ε:

$$\frac{\partial \Psi}{\partial \varepsilon} = -\mathbb{E}_F[S(X - t)] + S(x - t).$$

At $t = \mu$, since $\mathbb{E}_F[S(X - \mu)] = 0$,

$$\frac{\partial \Psi}{\partial \varepsilon}(\mu, 0) = S(x - \mu).$$

1.8.6 Putting the pieces together

Insert both derivatives into the linearized equation:

$$0 \approx (-2f(\mu))(t_\varepsilon - \mu) + \varepsilon S(x - \mu).$$

Solving for $t_\varepsilon - \mu$ gives

$$t_\varepsilon - \mu \approx \frac{\varepsilon}{2f(\mu)} S(x - \mu).$$

1.8.7 Definition of the influence function

By definition,

$$\mathrm{IF}(x; T, F) = \left.\frac{d}{d\varepsilon} t_\varepsilon \right|_{\varepsilon=0}.$$

From the expansion above,

$$\boxed{\mathrm{IF}(x; T, F) = \frac{1}{2f(\mu)} S(x - \mu)}$$

1.8.8 What "standard M-estimator calculus" really means

Whenever an estimator $T(F)$ is defined by

$$\mathbb{E}_F[\psi(X, T(F))] = 0,$$

the influence function is

$$\mathrm{IF}(x) = \left(\mathbb{E}\left[\frac{\partial}{\partial t} \psi(X, t) \Big|_{t=\mu} \right] \right)^{-1} \psi(x, \mu).$$

For the median,

- $\psi(x, t) = S(x - t)$,
- the derivative equals $-2f(\mu)$.

Everything derived above is a concrete instance of this general rule.

1.8.9 Intuition check

A tiny contamination moves the median proportionally to the sign of the contaminating point, scaled by how steep the CDF is at the median.

No higher-order effects matter at first order. That is the entire point of linearization.

1.9 What is ARE (Asymptotic Relative Efficiency)?

ARE — Asymptotic Relative Efficiency — is a precise asymptotic notion for comparing two estimators *when the sample size is large*. It answers one sharp question:

> How many samples does estimator A need to match the precision of estimator B?

1.9.1 The formal definition

Suppose two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ estimate the same parameter θ, and both are $\sqrt{n}$-consistent and asymptotically normal:

$$\sqrt{n}(\hat{\theta}_i - \theta) \xrightarrow{d} N(0, V_i), \qquad i = 1, 2.$$

Then the **ARE of $\hat{\theta}_1$ relative to $\hat{\theta}_2$** is defined as

$$\boxed{\mathrm{ARE}(\hat{\theta}_1, \hat{\theta}_2) = \frac{V_2}{V_1}}$$

Interpretation

- $\mathrm{ARE} = 1$: the estimators are equally efficient.
- $\mathrm{ARE} < 1$: $\hat{\theta}_1$ is less efficient.
- $\mathrm{ARE} > 1$: $\hat{\theta}_1$ is more efficient.

1.9.2 Why it is called *relative*

If $\mathrm{ARE} = 0.64$, then estimator 1 requires approximately

$$\frac{1}{0.64} \approx 1.56$$

times **more data** to achieve the same asymptotic accuracy as estimator 2.

ARE is therefore literally a **sample-size exchange rate**.

1.9.3 ARE for the median versus the mean

For a symmetric distribution F:

- **Mean:**
$$\sqrt{n}(\bar{X} - \mu) \to N(0, \sigma^2).$$
- **Median:**
$$\sqrt{n}(\hat{\mu} - \mu) \to N\left(0, \frac{1}{4f(\mu)^2}\right).$$

Therefore,

$$\boxed{\text{ARE(median, mean)} = \frac{\sigma^2}{\frac{1}{4f(\mu)^2}} = 4f(\mu)^2\sigma^2}$$

1.9.4 Concrete examples

- **Normal distribution** $N(\mu, \sigma^2)$
Since
$$f(\mu) = \frac{1}{\sqrt{2\pi}\sigma},$$
we obtain
$$\text{ARE} = \frac{2}{\pi} \approx 0.64.$$
The median loses efficiency under Gaussian noise.
- **Heavy-tailed distributions**
  - t_3: ARE ≈ 1.62,
  - Laplace: ARE $= 2$.

  In these cases, the median wins decisively.

1.9.5 What ARE is *not*

- Not a finite-sample guarantee.
- Not a robustness measure.
- Not about bias.

ARE is **purely about asymptotic variance**.

1.9.6 Takeaway

The mean is optimal under Gaussian noise.
The median sacrifices efficiency to gain robustness — and under heavy tails, that sacrifice becomes a win.

1.10 Affine Equivariance of the Estimate; Transformation–Retransformation (TR) Median

1.10.1 Formal setup and definitions

Let $X \in \mathbb{R}^p$ be a random vector with distribution F_X.

Multivariate location functional

A **multivariate location functional** is a mapping

$$T : \mathcal{F}_p \to \mathbb{R}^p,$$

where $\mathcal{F}_p$ is a suitable class of distributions on $\mathbb{R}^p$.

Examples.

- Mean vector: $T(F_X) = \mathbb{E}[X]$.
- Vector of marginal medians:

$$T(F_X) = \big(\mathrm{med}(X_1), \dots, \mathrm{med}(X_p)\big)^\top.$$

Affine transformations of distributions

For a full-rank matrix $A \in \mathbb{R}^{p\times p}$ and a vector $b \in \mathbb{R}^p$, define

$$Y = AX + b.$$

The distribution of Y is denoted by F_{AX+b}.

Affine equivariance (formal definition)

A location functional T is **affine equivariant** if

$$\boxed{T(F_{AX+b}) = A\,T(F_X) + b}$$

for all full-rank A and all b.

This is a **structural requirement**, not a statistical convenience.

1.10.2 Why affine equivariance is expected

Formal meaning

Affine transformations include:

- translations,
- rescalings,
- rotations,
- shears,

- and any composition of the above.

Affine equivariance ensures that the estimator **respects the geometry of the data**.

Intuition

If you rotate your coordinate system, your notion of "center" should rotate with it. If you stretch the axes, the center should stretch accordingly.

If this fails, the estimator is coordinate-dependent and geometrically inconsistent.

1.10.3 The vector of marginal medians is *not* affine equivariant

Definition of the marginal median vector

$$T(F_X) = \begin{pmatrix} \mathrm{med}(X_1) \\ \vdots \\ \mathrm{med}(X_p) \end{pmatrix}.$$

Each component is computed **independently**, ignoring dependence among coordinates.

Failure of affine equivariance (formal argument)

Consider an affine transformation

$$Y = AX + b, \qquad A = (a_{ij})_{i,j=1}^{p}.$$

Then

$$Y_i = \sum_{j=1}^{p} a_{ij} X_j + b_i.$$

The median of Y_i satisfies

$$\mathrm{med}(Y_i) = \mathrm{med}\left(\sum_{j=1}^{p} a_{ij} X_j\right) + b_i.$$

In general,

$$\mathrm{med}\left(\sum_{j=1}^{p} a_{ij} X_j\right) \neq \sum_{j=1}^{p} a_{ij}\, \mathrm{med}(X_j),$$

unless **only one term is present**.

Hence,

$$T(F_{AX+b}) \neq AT(F_X) + b$$

for general A.

Special case where it works

If A is diagonal with nonzero entries,

$$A = \operatorname{diag}(a_1, \dots, a_p),$$

then

$$Y_i = a_i X_i + b_i,$$

and since univariate medians are affine equivariant,

$$\operatorname{med}(Y_i) = a_i \operatorname{med}(X_i) + b_i.$$

Thus affine equivariance holds **only for diagonal** A.

Intuition. Marginal medians treat each axis as sacred. The moment you mix coordinates (rotation, shear), the estimator refuses to cooperate.

This is not a bug—it is a **structural limitation** of marginal thinking in multivariate space.

1.10.4 Invariant Coordinate System (ICS) functional

To repair this, we introduce a **coordinate-aware transformation**.

Definition (formal)

A matrix-valued functional

$$G : \mathcal{F}_p \to \mathbb{R}^{p \times p}$$

is called an **Invariant Coordinate System (ICS)** functional if

$$\boxed{G(F_{AX+b}) = G(F_X)A^{-1}}$$

for all full-rank A and b.

Interpretation

- $G(F_X)$ chooses a **data-dependent coordinate system**.
- Under affine transformation, the coordinate system adapts **contragrediently**.

Intuition. Think of $G(F_X)$ as saying:

> "Before doing anything, I will rotate and scale the data into a canonical position."

If the data are transformed, the canonicalizer compensates exactly.

1.10.5 Transformation–Retransformation (TR) median

Definition (formal)

Let T be the **vector of marginal medians**. Define the **TR median functional**:

$$\boxed{T_{\mathrm{TR}}(F_X) = G(F_X)^{-1} T\Big(F_{G(F_X)X}\Big)}$$

This is a three-step procedure:

1. Transform the data using $G(F_X)$.
2. Compute marginal medians.
3. Retransform back.

Proof of affine equivariance

Let $Y = AX + b$. Then

$$\begin{aligned} T_{\mathrm{TR}}(F_Y) &= G(F_Y)^{-1} T(F_{G(F_Y)Y}) \\ &= (G(F_X)A^{-1})^{-1} T(F_{G(F_X)A^{-1}(AX+b)}) \\ &= AG(F_X)^{-1} T(F_{G(F_X)X+G(F_X)A^{-1}b}). \end{aligned}$$

Using translation equivariance of marginal medians,

$$T(F_{Z+c}) = T(F_Z) + c,$$

we obtain

$$T_{\mathrm{TR}}(F_Y) = AT_{\mathrm{TR}}(F_X) + b.$$

Thus T_{TR} is **affine equivariant**.

Intuition. This is "median, but done in the right coordinates."

Instead of forcing medians to understand geometry, we **teach geometry first**, then compute medians, then come back.

1.10.6 Common pitfalls

- Believing marginal robustness implies multivariate robustness.
- Ignoring coordinate dependence.
- Assuming medians behave like means under linear mixing.

The TR construction fixes all three.

1.11 Why the Transformation–Retransformation (TR) Concept Matters

The **Transformation–Retransformation (TR)** idea is not cosmetic. It fixes a *structural flaw* in naive multivariate robust estimators and does so in a way that is mathematically principled, geometrically honest, and practically useful.

1.11.1 What problem does TR actually solve?

The blunt truth

Most "simple" multivariate robust estimators (such as the vector of marginal medians) are **coordinate artifacts**. Their output depends on how you choose your axes, not on the intrinsic geometry of the data cloud.

This is unacceptable when data live in $\mathbb{R}^p$ as geometry, not as p unrelated columns.

What affine equivariance really buys you

> Two analysts using different linear coordinate systems will report the same center, up to the same transformation.

This is not philosophical. It is **statistical reproducibility under reparameterization**.

Without affine equivariance:

- rotating the data changes the estimator,
- mixing variables changes the estimator,
- scientific conclusions depend on arbitrary preprocessing.

TR restores this invariance **without sacrificing robustness**.

1.11.2 What TR is doing conceptually

TR is a **change-of-coordinates strategy**:

1. **Find the geometry of the data.**
   Use an ICS functional $G(F)$ to identify directions, scales, and shape.
2. **Move to canonical coordinates.**
   Transform the data so the cloud is standardized (often spherical or axis-aligned).
3. **Apply a simple robust estimator.**
   Compute marginal medians where they actually make sense.

4. **Transform back.**
   Return to the original space.

The estimator stays simple. The *space* is made intelligent.

Mental picture

Imagine a **tilted elliptical cloud** of points in two dimensions:

- the true center is the center of the ellipse,
- the cloud is rotated by, say, 30°.

Now compute:

- the median of X_1,
- the median of X_2.

These medians are taken **along the coordinate axes**, not along the geometry of the cloud.

What goes wrong

- correlation is ignored,
- rotating the cloud changes the reported "center",
- the estimator is not intrinsic to the data.

This is the concrete failure of affine equivariance.

Step-by-step visual logic

1. **Before TR.**
   The data cloud is elongated and rotated.
2. **Apply $G(F)$.**
   Rotation and scaling are undone; the ellipse becomes roughly spherical.
3. **Compute marginal medians.**
   Axes now align with structure; medians behave sensibly.
4. **Retransform.**
   The center is mapped back to the original geometry.

The final point:

- sits at the geometric center,
- moves correctly under affine transformations,
- remains robust to outliers.

1.11.3 Why this matters statistically

Robustness plus equivariance is rare

- **Mean**: affine equivariant, not robust.
- **Marginal medians**: robust, not affine equivariant.

TR delivers **both**, provided the ICS functional is well chosen.

In practice, TR enables

- meaningful multivariate medians,
- robust PCA-like constructions,
- coordinate-free comparison across studies,
- geometrically honest inference.

This is why TR ideas appear repeatedly in robust multivariate analysis, outlier detection, and high-dimensional statistics.

1.11.4 A clean intuition to remember

TR does not fix the estimator.
It fixes the coordinate system in which the estimator is allowed to act.

Once you see it this way, the idea stops looking clever and starts looking inevitable.

Robust statistics fails when it ignores geometry. TR is geometry-aware robustness.

1.12 The Spatial Median: Geometry-Aware Robustness in $\mathbb{R}^p$

1.12.1 What the spatial median is—and why it exists

Formal definition (sample version)

Given data points $x_1, \dots, x_n \in \mathbb{R}^p$, define

$$D_n(t) = \frac{1}{n}\sum_{i=1}^{n}\left(|x_i - t| - |x_i|\right), \qquad t \in \mathbb{R}^p,$$

where $|\cdot|$ denotes the Euclidean norm.

The **spatial median** $\hat{\mu}$ is any minimizer of $D_n(t)$.

The subtraction of $|x_i|$ is constant in t and does **not** affect the minimizer. It is included solely to ensure finiteness of expectations in the population version.

Population version

Let $X \sim F_X$. Define

$$D(t) = \mathbb{E}\big[|X - t| - |X|\big].$$

The **spatial median functional**

$$\mu = T(F_X)$$

is the unique minimizer of $D(t)$, under the stated assumptions.

Intuition

The spatial median minimizes the **average distance** to the data cloud:

- the mean minimizes average *squared* distance,
- the spatial median minimizes average *distance*.

This single change replaces sensitivity to magnitude by sensitivity to geometry.

1.12.2 Assumptions and why they matter

Assumption 1: uniqueness

The minimizer μ of $D(t)$ is unique.

Why needed: Without uniqueness, asymptotic expansions and limiting distributions are not well-defined.

Geometric intuition: In dimension ≥ 2, surrounding geometry pins the center down; on a line, it can slide.

Assumption 2: smoothness of the density

F_X has a bounded and continuous density at μ.

Why needed: Ensures Taylor expansion of $D(t)$ and finiteness of expectations involving $|X - \mu|^{-1}$.

1.12.3 Local quadratic expansion of the objective

Formal expansion

Under the assumptions,

$$D(t) = D(\mu) + \frac{1}{2}(t - \mu)^\top \Gamma (t - \mu) + o(|t - \mu|^2),$$

where

$$\Gamma = \mathbb{E}\left[\frac{1}{|X - \mu|}\left(I_p - \frac{(X - \mu)(X - \mu)^\top}{|X - \mu|^2}\right)\right].$$

Why this matrix appears

The gradient of $|x - t|$ is

$$\nabla_t |x - t| = -\frac{x - t}{|x - t|},$$

and the Hessian introduces the projection matrix

$$I_p - \frac{(x - \mu)(x - \mu)^\top}{|x - \mu|^2},$$

which projects orthogonally to the direction $x - \mu$.

Intuition

Near the spatial median, the objective is quadratic—but curvature depends on direction. The bowl is anisotropic and reflects how points surround the center.

1.12.4 Spatial sign and centered rank

Spatial sign

$$S(t) = \begin{cases} t/|t|, & t \neq 0, \\ 0, & t = 0. \end{cases}$$

This is a **direction-only** object.

Centered rank

$$\hat{R}(t) = \frac{1}{n} \sum_{i=1}^{n} S(t - x_i).$$

Key properties

- lies inside the unit p-ball,
- ignores magnitude completely,
- retains geometric direction.

Each observation pulls with **unit force**, no matter how far away it is.

1.12.5 Spatial median as a zero of the rank function

The spatial median satisfies

$$\hat{R}(\hat{\mu}) = 0.$$

This is the multivariate analogue of balancing signs to the left and right in one dimension.

1.12.6 Asymptotic distribution

Central limit theorem for ranks

$$\sqrt{n}\,\hat{R}(\mu) \xrightarrow{d} N_p(0,\Omega), \qquad \Omega = \mathbb{E}\left[\frac{(X-\mu)(X-\mu)^\top}{|X-\mu|^2}\right].$$

Asymptotic normality of the estimator

$$\hat{\mu} = \mu + \Gamma^{-1}\hat{R}(\mu) + o_P(n^{-1/2}),$$

hence

$$\sqrt{n}(\hat{\mu}-\mu) \xrightarrow{d} N_p(0,\Gamma^{-1}\Omega\Gamma^{-1}).$$

Intuition

- Ω measures directional variability,
- Γ^{-1} converts force imbalance into displacement.

Think: **force** $\rightarrow$ **motion**, governed by curvature.

What to see

- the spatial median sits where directional pulls cancel,
- distant outliers barely move it,
- the center reflects shape, not extremes.

1.12.7 Computation: Weiszfeld algorithm

Iteration step

$$\mu \leftarrow \mu + \left(\sum_{i=1}^{n} \frac{1}{|x_i-\mu|}\right)^{-1} \hat{R}(\mu).$$

This is a fixed-point iteration derived from the estimating equation.

Practical note

The classical algorithm may fail if μ coincides with a data point. Modified variants guarantee monotone convergence.

1.12.8 Robustness properties

Breakdown point

The spatial median has asymptotic breakdown point $1/2$. No location estimator can do better.

Influence function

$$\mathrm{IF}(x;T,F) = -\Gamma^{-1}S(x-\mu).$$

This influence function is **bounded**.

Intuition

Magnitude is ignored; only direction matters. One bad point cannot dominate.

1.12.9 Efficiency tradeoff

If the covariance matrix Σ exists, the ARE relative to the mean is

$$\mathrm{ARE} = \frac{1}{p}\frac{|\Sigma|}{|\Gamma^{-1}\Omega\Gamma^{-1}|}.$$

Meaning

- some efficiency is lost under perfect Gaussianity,
- massive stability is gained under contamination.

This is not a defect—it is a deliberate design choice.

1.13 Estimation of the Covariance Matrix; Affine Equivariance; TR Spatial Median

1.13.1 Asymptotic covariance of the spatial median

Formal result recalled

From earlier theory, the spatial median $\hat{\mu}$ satisfies

$$\sqrt{n}(\hat{\mu}-\mu) \xrightarrow{d} N_p\Big(0,\ \Gamma^{-1}\Omega\Gamma^{-1}\Big),$$

where

$$\Gamma = \mathbb{E}\left[\frac{1}{|X-\mu|}\left(I_p - \frac{(X-\mu)(X-\mu)^\top}{|X-\mu|^2}\right)\right], \qquad \Omega = \mathbb{E}\left[\frac{(X-\mu)(X-\mu)^\top}{|X-\mu|^2}\right].$$

Therefore, an **approximate covariance matrix** for $\hat{\mu}$ is

$$\boxed{\mathrm{Cov}(\hat{\mu}) \approx \frac{1}{n}\Gamma^{-1}\Omega\Gamma^{-1}.}$$

1.13.2 Sample estimators $\hat{\Gamma}$ and $\hat{\Omega}$

Plug-in principle

Since Γ and Ω are expectations under F_X, we estimate them by empirical averages, replacing

- μ with $\hat{\mu}$,
- expectations with sample means.

Thus,

$$\boxed{\hat{\Gamma} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{|x_i - \hat{\mu}|}\left[I_p - \frac{(x_i - \hat{\mu})(x_i - \hat{\mu})^\top}{|x_i - \hat{\mu}|^2}\right]}$$

and

$$\boxed{\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} \frac{(x_i - \hat{\mu})(x_i - \hat{\mu})^\top}{|x_i - \hat{\mu}|^2}}.$$

The resulting covariance estimator is

$$\boxed{\widehat{\mathrm{Cov}}(\hat{\mu}) = \frac{1}{n}\hat{\Gamma}^{-1}\hat{\Omega}\hat{\Gamma}^{-1}.}$$

Why these formulas are correct

- Each summand is bounded whenever $x_i \neq \hat{\mu}$.
- Expectations exist under the stated density assumptions.
- The law of large numbers ensures consistency.
- The estimator mirrors the asymptotic variance exactly.

This is direct M-estimation theory, not heuristic adjustment.

Intuition: what Γ and Ω mean

- Ω: **directional scatter** — only directions matter; distances are normalized.
- Γ: **local curvature** — resistance of the objective to movement.

Think of Ω as random force and Γ^{-1} as mechanical compliance.

1.13.3 Why the spatial median is *not* affine equivariant

Formal statement

The spatial median satisfies

$$T(F_{AX+b}) = AT(F_X) + b$$

only if A is orthogonal, i.e.

$$A^\top A = I_p.$$

Why this fails in general

The spatial median minimizes

$$\mathbb{E}|X - t|,$$

which depends explicitly on the **Euclidean norm**.

If A is not orthogonal,

$$|AX| \neq |X|,$$

so the objective itself changes shape under general affine maps.

Intuition

The spatial median is

- rotation invariant,
- reflection invariant,
- translation invariant,

but **not** scale- or shear-invariant. It respects angles, not full linear geometry.

1.13.4 Transformation–Retransformation (TR) spatial median

Motivation

Fix the geometry *before* computing the median.

1.13.5 Construction of the TR spatial median

Step 1: choose a scatter functional

Let $S(F)$ be a **scatter functional** (e.g. Tyler's shape matrix). Define

$$G(F) = S(F)^{-1/2}.$$

This standardizes the data.

Step 2: normalization property

By construction,

$$\boxed{G(F)\, S(F)\, G(F)^\top = I_p.}$$

The transformed data are therefore spherical.

Step 3: define the TR estimator

$$\boxed{T_{\mathrm{TR}}(F_X) = G(F_X)^{-1}\, T\Big(F_{G(F_X)X}\Big).}$$

Compute the spatial median after standardization, then map it back.

1.13.6 Why TR restores affine equivariance

Formal logic

- $S(F_{AX+b}) = AS(F_X)A^\top$,
- $G(F_{AX+b}) = G(F_X)A^{-1}$,
- the standardized geometry is invariant,
- retransformation restores the affine structure.

Hence,

$$\boxed{T_{\mathrm{TR}}(F_{AX+b}) = AT_{\mathrm{TR}}(F_X) + b.}$$

Intuition

The spatial median fails because it trusts raw coordinates.

Coordinates are negotiable. Geometry is not.

1.13.7 Relationship to known estimators

The **Hettmansperger–Randles median** combines

- the spatial median (robust location),
- Tyler's scatter (robust shape),
- TR geometry correction.

This is not ad hoc. It is the canonical way to unite robustness with affine invariance.

1.14 The Oja Median: A Geometric Notion of Multivariate Location

1.14.1 Restatement of the problem context

We are in **multivariate statistics**, studying a notion of multivariate location called the **Oja median**. The construction is geometric: it measures how a candidate point $t \in \mathbb{R}^p$ relates to the data cloud via volumes of simplices.

The goals are:

- define the **sample Oja median** via a minimization problem,
- define the corresponding **population (functional) version**,
- introduce the **multivariate sign and rank functions** needed for asymptotic analysis.

The development proceeds from geometry to probability, step by step.

1.14.2 Data, distribution, and notation

Formal setup

Let

$$X = (x_1, x_2, \ldots, x_n)' \quad \text{with } x_i \in \mathbb{R}^p$$

be a random sample from a p-variate distribution with cumulative distribution function

$$F : \mathbb{R}^p \to [0, 1].$$

Here p is the dimension and n the sample size.

Intuition

Think of $x_1, \ldots, x_n$ as points in $\mathbb{R}^p$. We seek a notion of "center" that is robust and intrinsic to their geometry.

1.14.3 Volume of a p-simplex

Definition

Given $p + 1$ points

$$t_1, t_2, \ldots, t_{p+1} \in \mathbb{R}^p,$$

the volume of the simplex they determine is

$$V(t_1, \ldots, t_{p+1}) = \frac{1}{p!} \left| \det \begin{pmatrix} 1 & \cdots & 1 \\ t_1 & \cdots & t_{p+1} \end{pmatrix} \right|.$$

Why this formula is valid

- the determinant computes signed volume of a parallelepiped,
- the row of ones converts points to an affine representation,
- division by $p!$ converts parallelepiped volume to simplex volume,
- the absolute value removes orientation.

Low-dimensional intuition

- $p = 1$**:** two points define an interval, and

$$V(t_1, t_2) = |t_2 - t_1|.$$

- $p = 2$**:** three points form a triangle, and V is its area.

This confirms the formula generalizes length and area.

1.14.4 Sample Oja objective function

Definition

For a candidate location $t \in \mathbb{R}^p$, define

$$D_n(t) = \binom{n}{p}^{-1} \sum_{1 \le i_1 < \cdots < i_p \le n} V(x_{i_1}, \ldots, x_{i_p}, t).$$

Explanation of components

- the sum runs over all p-subsets of the sample,
- each term is the volume of the simplex formed by those points and t,
- the binomial factor normalizes the sum to an average.

Intuition

If t is central, simplices formed with the data tend to have small volume on average.

1.14.5 Sample Oja median

Definition

The **Oja median** $T(X)$ is any minimizer of

$$D_n(t), \qquad \text{i.e.} \qquad T(X) \in \arg\min_{t \in \mathbb{R}^p} D_n(t).$$

Intuition

- in one dimension, this reduces to the usual median,
- in higher dimensions, it minimizes average simplex volume.

1.14.6 Population (functional) version

Definition

Define the population objective

$$D(t) = E_F[V(X_1, \ldots, X_p, t)],$$

where $X_1, \ldots, X_p$ are i.i.d. with distribution F.

The **Oja functional** $T(F)$ is any minimizer of $D(t)$.

Existence of expectations

- $V(\cdot)$ grows at most linearly,
- finite first moments of F suffice for finiteness.

1.14.7 Assumptions for asymptotic theory

- **Uniqueness:** the minimizer $\mu = T(F)$ is unique,
- **Second moments:** $E|X|^2 < \infty$.

  These guarantee differentiability and quadratic approximation of $D(t)$.

1.14.8 Quadratic expansion of $D(t)$

Formal statement

Near μ,

$$D(t) = D(\mu) + \frac{1}{2}(t-\mu)'\Delta(t-\mu) + o(|t-\mu|^2),$$

where

$$\Delta = \frac{\partial^2}{\partial t\, \partial t'} D(t)\bigg|_{t=\mu}.$$

Interpretation

Δ is the Hessian of the objective; locally, $D(t)$ behaves like a quadratic bowl.

1.14.9 Indexing subsets

Define

$$Q = \{(i_1, \ldots, i_{p-1}) : 1 \le i_1 < \cdots < i_{p-1} \le n\}, \quad P = \{(i_1, \ldots, i_p) : 1 \le i_1 < \cdots < i_p \le n\}.$$

These index simplices of different orders.

1.14.10 Determinant decompositions

Definition

For $q \in Q$, define $e_q \in \mathbb{R}^p$ by

$$\det(x_{i_1}, \ldots, x_{i_{p-1}}, x) = e_q' x.$$

Similarly, for $p \in P$,

$$\det \begin{pmatrix} 1 & \cdots & 1 & 1 \\ x_{i_1} & \cdots & x_{i_p} & x \end{pmatrix} = d_{0p} + d_p' x.$$

Interpretation

Linearity of the determinant reduces volume calculations to linear forms.

1.14.11 Sample sign and rank functions

Definitions

$$\hat{S}(t) = \binom{n}{p}^{-1} \sum_{q \in Q} \operatorname{sign}(e_q' t)\, e_q,$$

$$\hat{R}(t) = \binom{n}{p}^{-1} \sum_{p \in P} \operatorname{sign}(d_{0p} + d_p' t)\, d_p.$$

Sign function

$$\operatorname{sign}(u) = \begin{cases} +1, & u > 0, \\ 0, & u = 0, \\ -1, & u < 0. \end{cases}$$

Role

These act as generalized gradients; the Oja median satisfies $\hat{R}(t) \approx 0$.

1.14.12 Population versions

$$S(t) = E[\operatorname{sign}(e_q' t) e_q], \qquad R(t) = E[\operatorname{sign}(d_{0p} + d_p' t) d_p].$$

These are theoretical objects used in asymptotic analysis.

1.14.13 Big picture intuition

How small are the simplices formed when this point is included?

The sign and rank functions translate geometry into algebra suitable for probability theory.

This is geometry wearing a statistics hat—and doing it properly.