Robuste Statistik

Musterlösung zu Blatt 2

Aufgabe 2.1:

a)

Um die Äquivarianz- bzw. Invarianz-Eigenschaften des MAD herleiten zu können, ist es essentiell sich zunächst die Äquivarianz- und Invarianz-Eigenschaften des Medians anzuschauen.

Lokationseigenschaften (Median):
Sei $z_n = y_n + l,\ n = 1, ..., N,\ l \in \mathbb{R}$
$\Rightarrow$ für ungerades N

$$med(\boldsymbol{z}) = z_{\left(\frac{n+1}{2}\right)} = y_{\left(\frac{n+1}{2}\right)} + l = med(\boldsymbol{y}) + l$$

bzw. für gerades N

$$med(\boldsymbol{z}) = \left[z_{\left(\frac{n}{2}\right)}, z_{\left(\frac{n}{2}+1\right)}\right] = \left[y_{\left(\frac{n}{2}\right)} + l, y_{\left(\frac{n}{2}+1\right)} + l\right] = med(\boldsymbol{y}) + l$$

Diese Gleichheiten gelten, weil sich die Ordnung der Elemente durch die Addition einer Konstanten auf jedes y_n nicht ändert.
$\Rightarrow$ Der Median ist lokations-äquivariant.

Skaleneigenschaften (Median):
Sei $z_n = s \cdot y_n, n = 1, ..., N,\ s > 0$
$\Rightarrow$ für ungerades N

$$med(\boldsymbol{z}) = z_{\left(\frac{n+1}{2}\right)} = s \cdot y_{\left(\frac{n+1}{2}\right)} = s \cdot med(\boldsymbol{y})$$

bzw. für gerades N

$$med(\boldsymbol{z}) = \left[z_{\left(\frac{n}{2}\right)}, z_{\left(\frac{n}{2}+1\right)}\right] = \left[s \cdot y_{\left(\frac{n}{2}\right)}, s \cdot y_{\left(\frac{n}{2}+1\right)}\right] = s \cdot med(\boldsymbol{y})$$

Diese Gleichheiten gelten, weil sich die Ordnung der Elemente durch die Multiplikation einer positiven Konstanten mit jedem y_n nicht ändert.
$\Rightarrow$ Der Median ist skalen-äquivariant.

Lokationseigenschaften (MAD):
Sei $z_n = y_n + l,\ n = 1, ..., N,\ l \in \mathbb{R}$

$$\begin{aligned} MAD(\boldsymbol{z}) &= med(|z_1 - med(\boldsymbol{z})|, ..., |z_N - med(\boldsymbol{z})|) \\ &= med(|y_1 + l - med(\boldsymbol{y}) - l|, ..., |y_N + l - med(\boldsymbol{y}) - l|) \\ &= med(|y_1 - med(\boldsymbol{y})|, ..., |y_N - med(\boldsymbol{y})|) \\ &= MAD(\boldsymbol{y}) \end{aligned}$$

$\Rightarrow$ Der MAD ist lokations-invariant.

Skaleneigenschaften (MAD):
Sei $z_n = s \cdot y_n, n = 1, ..., N,\ s > 0$

$$\begin{aligned} MAD(\boldsymbol{z}) &= med(|z_1 - med(\boldsymbol{z})|, ..., |z_N - med(\boldsymbol{z})|) \\ &= med(|s \cdot y_1 - med(s \cdot \boldsymbol{y})|, ..., |s \cdot y_N - med(s \cdot \boldsymbol{y})|) \\ &= med(|s \cdot y_1 - s \cdot med(\boldsymbol{y})|, ..., |s \cdot y_N - s \cdot med(\boldsymbol{y})|) \\ &= med(s \cdot |y_1 - med(\boldsymbol{y})|, ..., s \cdot |y_N - med(\boldsymbol{y})|) \\ &= s \cdot med(|y_1 - med(\boldsymbol{y})|, ..., |y_N - med(\boldsymbol{y})|) \\ &= s \cdot MAD(\boldsymbol{y}) \end{aligned}$$

$\Rightarrow$ Der MAD ist skalen-äquivariant.

Somit erfüllt der MAD die Anforderungen an einen Skalenschätzer. Zusätzlich sehen wir, dass der Median die Anforderungen an einen Punktschätzer erfüllt.

b)

Der Modalwert ist derjenige Wert, der in der Stichprobe $y_1, ..., y_N$ am häufigsten angenommen wird. Werden mehrere Werte gleichhäufig und keiner häufiger angenommen, so ist der Modalwert nicht eindeutig.

Lokationseigenschaften (Modalwert):
Sei $z_n = y_n + l,\ n = 1, ..., N,\ l \in \mathbb{R}$
Die Häufigkeitsverteilung der Werte $z_1, ..., z_N$ ist identisch zu der Häufigkeitsverteilung der Werte $y_1, ..., y_N$, da sich Häufigkeitsverteilungen nicht ändern, wenn auf alle Werte die gleiche Konstante addiert wird. Somit ist $mod(\boldsymbol{z}) = mod(\boldsymbol{y}) + l$. Der Modalwert ist somit lokations-äquivariant.

Skaleneigenschaften (Modalwert):
Sei $z_n = s \cdot y_n, n = 1, ..., N,\ s > 0$
Die Häufigkeitsverteilung der Werte $z_1, ..., z_N$ ist identisch zu der Häufigkeitsverteilung der Werte $y_1, ..., y_N$, da sich Häufigkeitsverteilungen nicht ändern, wenn alle Werte mit der gleichen positiven Konstanten multipliziert werden. Somit ist $mod(\boldsymbol{z}) = s \cdot mod(\boldsymbol{y})$. Der Modalwert ist somit skalen-äquivariant.

Somit erfüllt der Modalwert die Anforderungen an einen Lokationsschätzer.

Aufgabe 2.2:
a)

Lokationseigenschaften:
Sei $z_n = y_n + l,\ n = 1, ..., N,\ l \in \mathbb{R}$

$$\begin{aligned} \hat{\theta}_1(\boldsymbol{z}) &= \left(\frac{1}{N} \sum_{n=1}^{N} |z_n - \bar{z}| \right)^2 \\ &= \left(\frac{1}{N} \sum_{n=1}^{N} |y_n + l - \bar{y} - l| \right)^2 \\ &= \left(\frac{1}{N} \sum_{n=1}^{N} |y_n - \bar{y}| \right)^2 \\ &= \hat{\theta}_1(\boldsymbol{y}) \end{aligned}$$

$\Rightarrow \hat{\theta}_1$ ist lokations-invariant.

Skaleneigenschaften:
Sei $z_n = s \cdot y_n, n = 1, ..., N,\ s > 0$

$$\begin{aligned}
\hat{\theta}_1(\boldsymbol{z}) &= \left(\frac{1}{N} \sum_{n=1}^{N} |z_n - \bar{z}| \right)^2 \\
&= \left(\frac{1}{N} \sum_{n=1}^{N} |s \cdot y_n - s \cdot \bar{y}| \right)^2 \\
&= \left(\frac{1}{N} \sum_{n=1}^{N} s \cdot |y_n - \bar{y}| \right)^2 \\
&= \left(\frac{s}{N} \sum_{n=1}^{N} |y_n - \bar{y}| \right)^2 \\
&= s^2 \cdot \left(\frac{1}{N} \sum_{n=1}^{N} |y_n - \bar{y}| \right)^2 \\
&= s^2 \cdot \hat{\theta}_1(\boldsymbol{y})
\end{aligned}$$

$\Rightarrow \hat{\theta}_1$ ist weder skalen-äquivariant noch skalen-invariant.

Somit erfüllt $\hat{\theta}_1$ weder die Anforderungen an einen Lokationsschätzer noch an einen Skalenschätzer.

b)

Lokationseigenschaften:
Sei $z_n = y_n + l,\ n = 1, ..., N,\ l \in \mathbb{R}$

$$\begin{aligned}
\hat{\theta}_2(\boldsymbol{z}) &= \sqrt{\overline{z^2} - \bar{z}^2} \\
&= \sqrt{\frac{1}{N} \sum_{n=1}^{N} z_n^2 - \left(\frac{1}{N} \sum_{n=1}^{N} z_n \right)^2} \\
&= \sqrt{\frac{1}{N} \sum_{n=1}^{N} (y_n + l)^2 - \left(\frac{1}{N} \sum_{n=1}^{N} (y_n + l) \right)^2} \\
&= \sqrt{\frac{1}{N} \sum_{n=1}^{N} (y_n^2 + 2ly_n + l^2) - \left(\frac{1}{N} \left(\sum_{n=1}^{N} y_n + Nl \right) \right)^2} \\
&= \sqrt{\frac{1}{N} \sum_{n=1}^{N} y_n^2 + \frac{1}{N} \sum_{n=1}^{N} 2ly_n + \frac{1}{N} \sum_{n=1}^{N} l^2 - \left(\frac{1}{N} \sum_{n=1}^{N} y_n + l \right)^2} \\
&= \sqrt{\frac{1}{N} \sum_{n=1}^{N} y_n^2 + \frac{2l}{N} \sum_{n=1}^{N} y_n + l^2 - \left(\frac{1}{N} \sum_{n=1}^{N} y_n \right)^2 - \frac{2l}{N} \sum_{n=1}^{N} y_n - l^2} \\
&= \sqrt{\frac{1}{N} \sum_{n=1}^{N} y_n^2 - \left(\frac{1}{N} \sum_{n=1}^{N} y_n \right)^2} \\
&= \sqrt{\overline{y^2} - \bar{y}^2} \\
&= \hat{\theta}_2(\boldsymbol{y})
\end{aligned}$$

$\Rightarrow \hat{\theta}_2$ ist lokations-invariant.

Skaleneigenschaften:
Sei $z_n = s \cdot y_n, n = 1, ..., N,\ s > 0$

$$\begin{aligned}
\hat{\theta}_2(\boldsymbol{z}) &= \sqrt{\overline{z^2} - \bar{z}^2} \\
&= \sqrt{\frac{1}{N}\sum_{n=1}^{N} z_n^2 - \left(\frac{1}{N}\sum_{n=1}^{N} z_n\right)^2} \\
&= \sqrt{\frac{1}{N}\sum_{n=1}^{N} (s \cdot y_n)^2 - \left(\frac{1}{N}\sum_{n=1}^{N} s \cdot y_n\right)^2} \\
&= \sqrt{\frac{1}{N}\sum_{n=1}^{N} s^2 \cdot y_n^2 - \left(\frac{s}{N}\sum_{n=1}^{N} y_n\right)^2} \\
&= \sqrt{s^2 \cdot \frac{1}{N}\sum_{n=1}^{N} y_n^2 - s^2 \cdot \left(\frac{1}{N}\sum_{n=1}^{N} y_n\right)^2} \\
&= \sqrt{s^2 \cdot \left(\frac{1}{N}\sum_{n=1}^{N} y_n^2 - \left(\frac{1}{N}\sum_{n=1}^{N} y_n\right)^2\right)} \\
&= s \cdot \sqrt{\frac{1}{N}\sum_{n=1}^{N} y_n^2 - \left(\frac{1}{N}\sum_{n=1}^{N} y_n\right)^2} \\
&= s \cdot \sqrt{\overline{y^2} - \bar{y}^2} \\
&= s \cdot \hat{\theta}_2(\boldsymbol{y})
\end{aligned}$$

$\Rightarrow \hat{\theta}_2$ ist skalen-äquivariant.

Somit erfüllt $\hat{\theta}_2$ die Anforderungen an einen Skalenschätzer.

c)

Man kann sich leicht überlegen, dass die in Aufgabe 2.1 gezeigte Lokations- und Skalen-Äquivarianz des Medians auch für allgemeines p-Quantil gilt, siehe dazu auch die Darstellung in 3.1 im Skript zur Vorlesung.

Bonus: Ganz allgemein kann man auch zeigen, dass das p-Quantil äquivariant bezüglich streng monotoner, bijektiver Transformationen ist, d.h. für streng monotones, bijektives $f : \mathbb{R} \to \mathbb{R}$ gilt

$$f(\tilde{y})_p = f(\tilde{y}_p). \tag{1}$$

Dabei ist $f(y)$ die koordinatenweise Anwendung der Abbildung auf f auf den Datensatz y. Formel (1) lässt sich wie folgt zeigen:

$$\begin{aligned}
f(\tilde{y})_p =& \Big\{\mu \in \mathbb{R}; \#\{n; f(y_n) \leq \mu\} \geq pN \text{ und } \#\{n; f(y_n) \geq \mu\} \geq (1-p)N\Big\} \\
=& \Big\{\mu \in \mathbb{R}; \#\{n; y_n \leq f^{-1}(\mu)\} \geq pN \text{ und } \#\{n; y_n \geq f^{-1}(\mu)\} \geq (1-p)N\Big\} \\
=& \Big\{f(\mu) \in \mathbb{R}; \#\{n; y_n \leq \mu\} \geq pN \text{ und } \#\{n; y_n \geq \mu\} \geq (1-p)N\Big\} \\
=& f\left(\Big\{\mu \in \mathbb{R}; \#\{n; y_n \leq \mu\} \geq pN \text{ und } \#\{n; y_n \geq \mu\} \geq (1-p)N\Big\}\right) = f(\tilde{y}_p).
\end{aligned}$$

Wenn wir $f(x) = sx + l$ für $s > 0$ wählen, folgt, dass p-Quantile Lageschätzer sind.

Man beachte, dass der Schätzer $\hat{\theta}_3$ nur sinnvoll ist, falls $y_i > 0$ für beliebige $i = 1, \ldots, N$.
Somit gilt insgesamt für $\hat{\theta}_3$:

Lokationseigenschaften:
Sei $z_n = y_n + l,\ n = 1, ..., N,\ l \in \mathbb{R}$

$$\begin{aligned}\hat{\theta}_3(\boldsymbol{z}) &= \frac{\tilde{z}_{0.75} - \tilde{z}_{0.25}}{\tilde{z}_{0.5}} \\ &= \frac{(\tilde{y}_{0.75} + l) - (\tilde{y}_{0.25} + l)}{\tilde{y}_{0.5} + l} \\ &= \frac{\tilde{y}_{0.75} - \tilde{y}_{0.25}}{\tilde{y}_{0.5} + l}\end{aligned}$$

$\Rightarrow \hat{\theta}_3$ ist weder lokations-invariant noch lokations-äquivariant.

Skaleneigenschaften:
Sei $z_n = s \cdot y_n, n = 1, ..., N,\ s > 0$

$$\begin{aligned}\hat{\theta}_3(\boldsymbol{z}) &= \frac{\tilde{z}_{0.75} - \tilde{z}_{0.25}}{\tilde{z}_{0.5}} \\ &= \frac{s \cdot \tilde{y}_{0.75} - s \cdot \tilde{y}_{0.25}}{s \cdot \tilde{y}_{0.5}} \\ &= \frac{s \cdot (\tilde{y}_{0.75} - \tilde{y}_{0.25})}{s \cdot \tilde{y}_{0.5}} \\ &= \frac{\tilde{y}_{0.75} - \tilde{y}_{0.25}}{\tilde{y}_{0.5}} \\ &= \hat{\theta}_3(\boldsymbol{y})\end{aligned}$$

$\Rightarrow \hat{\theta}_3$ ist skalen-invariant.

Somit erfüllt $\hat{\theta}_3$ weder die Anforderungen an einen Lokationsschätzer noch an einen Skalenschätzer.

Hinweis: Die Lokations- und Äquivarianzeigenschaften können auch gleichzeitig gezeigt werden, z.B.

$$\begin{aligned}&\hat{\theta} \text{ ist ein Lageschätzer, also skalen- und lokationsäquivariant} \\ \Leftrightarrow\ &\forall y \in \mathbb{R}^N \forall l \in \mathbb{R} \forall s > 0 :\ \hat{\theta}(sy + l) = s\hat{\theta}(y) + l,\end{aligned}$$

oder z.B.

$$\begin{aligned}&\hat{\theta} \text{ ist ein Skalenschätzer, also skalenäquivariant und lokationsinvariant} \\ \Leftrightarrow\ &\forall y \in \mathbb{R}^N \forall l \in \mathbb{R} \forall s > 0 :\ \hat{\theta}(sy + l) = s\hat{\theta}(y).\end{aligned}$$