

Beitrag zur robusten Parameterschätzung

Iteratively reweighted least squares revisited

Bastian Erdnütz

Fraunhofer-IOSB (Institut für Optronik, Systemtechnik und Bildverarbeitung)
Fraunhoferstr. 1, D-76131 Karlsruhe

Zusammenfassung Die Kleinste-Quadrate-Schätzung ist optimal für normalverteilte Messfehler, jedoch anfällig gegenüber groben Messfehlern. M-Schätzer können eine endlastigere Fehlerverteilung berücksichtigen, was sie robuster gegenüber groben Messfehlern macht. In diesem Beitrag wird eine in der Notation einfachere Beschreibung der klassischen Theorie der robusten M-Schätzer vorgestellt und für den Fall von gleichverteilten Ausreißer durchgesprochen. Darüber hinaus wird eine Familie bekannter robuster Verlustfunktionen in diese Notation übersetzt und Verbindungen zu einer Kernel-Lifting-Methode aufgezeigt, die als Alternative zum üblichen IRLS-Algorithmus zur Berechnung von M-Schätzern verwendet werden kann.

Schlüsselwörter Kleinste Quadrate, Robuste Schätzung, IRLS

Abstract The least squares estimator is optimal for normally distributed measurement errors, but it can break down under gross measurement errors. M-estimators can take fat-tailed error distribution into account, which makes them more robust to gross measurement errors. In this paper, a simpler description of the classical theory of robust M-estimators is presented and used to describe M-estimators for uniformly distributed outliers. In addition, a family of well known robust loss functions is presented in this notation and connections to a kernel lifting method are shown, which can be used as an alternative to the usual IRLS algorithm for calculating the M-estimators.

Keywords Least squares, robust estimation, IRLS

1 Der Kleinste-Quadrate-Sch  tzer

Der Kleinste-Quadrate-Sch  tzer ergibt sich als Maximum-Likelihood-Sch  tzer der Normalverteilung. Sind Beobachtungen $\mathbf{y} \in \mathbb{R}^N$ N -dimensional normalverteilt mit Erwartungswert $\boldsymbol{\mu} \in \mathbb{R}^N$ und Kovarianzmatrix $\Sigma \in \mathbb{R}^{N \times N}$ (geschrieben: $\mathbf{y} \sim \mathcal{N}_N(\boldsymbol{\mu}, \Sigma)$), so entspricht die *Likelihood* gegeben der Beobachtungen \mathbf{y} der Wahrscheinlichkeitsdichte von \mathbf{y} :

$$l(\boldsymbol{\mu}, \Sigma | \mathbf{y}) = p(\mathbf{y} | \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left(-\frac{1}{2}\|\mathbf{y} - \boldsymbol{\mu}\|_{\Sigma^{-1}}^2\right). \quad (1)$$

Hierbei steht $|\cdot|$ f  r die Determinante und $\|\mathbf{x}\|_A = \sqrt{\mathbf{x}^\top A \mathbf{x}}$ f  r die Norm bzgl. einer Metrik A . Die Likelihood ist maximal, wenn die negative Log-Likelihood

$$\lambda(\boldsymbol{\mu}, \Sigma | \mathbf{y}) = -\log(l(\boldsymbol{\mu}, \Sigma | \mathbf{y})) = \frac{1}{2} \log(|2\pi\Sigma|) + \frac{1}{2} \|\mathbf{y} - \boldsymbol{\mu}\|_{\Sigma^{-1}}^2 \quad (2)$$

minimal ist.

Ein paar Spezialf  lle werden nun genauer betrachtet. Liegen N unabh  ngig identisch eindimensional normalverteilte Beobachtungen $y_i \sim \mathcal{N}(\mu, \sigma^2)$ vor (also $\boldsymbol{\mu} = \mu \mathbf{1}$ der mit dem Faktor $\mu \in \mathbb{R}$ skalierte Konstant-1-Vektor und $\Sigma = \sigma^2 I$ die mit dem Faktor σ^2 skalierte Einheitsmatrix), so ist

$$\lambda(\mu, \sigma^2 | \mathbf{y}) = C + \frac{N}{2} \log(\sigma^2) + \frac{1}{2\sigma^2} \sum_i (y_i - \mu)^2 \quad (3)$$

mit der Konstante $C = \frac{N}{2} \log(2\pi)$. Unabh  ngig von der Wahl von σ^2 ist $\sum_i (y_i - \mu)^2$ minimal, wenn $\mu = \bar{y} = \frac{1}{N} \sum_i y_i$ der Mittelwert der Beobachtungen y_i ist. Dies ist der Maximum-Likelihood-Sch  tzer des Erwartungswerts μ der Beobachtungen y_i . Da $\mu = \bar{y}$ die Summe der Quadrate $\sum_i (y_i - \mu)^2$ minimiert, wird er auch als *Kleinste-Quadrate-Sch  tzer* bezeichnet.

Im zweiten betrachteten Fall ist $\mathbf{y} \sim \mathcal{N}_N(\boldsymbol{\mu}(\mathbf{x}), s^2 Q)$ mit linearem $\boldsymbol{\mu}(\mathbf{x}) = \mathbf{a} + A\mathbf{x}$, sowie $\mathbf{a} \in \mathbb{R}^N$, $\mathbf{x} \in \mathbb{R}^M$, $A \in \mathbb{R}^{M \times N}$, $Q \in \mathbb{R}^{N \times N}$ symmetrisch positiv definit und $s^2 > 0$. \mathbf{x} ist ein zu sch  tzender Parametervektor und \mathbf{a} , A , Q beschreiben das als bekannt vorausgesetzte

stochastische Modell der Beobachtungen. In dem Fall ist

$$\lambda(x, s^2 | \mathbf{y}) = C + \frac{N}{2} \log(s^2) + \frac{1}{2s^2} \|\mathbf{y} - \boldsymbol{\mu}(x)\|_{Q^{-1}}^2 \quad (4)$$

mit der Konstante $C = \frac{1}{2} \log(|2\pi Q|)$, wobei $\|\mathbf{y} - \boldsymbol{\mu}(x)\|_{Q^{-1}}^2$ unabhängig von s^2 minimal wird, wenn

$$\mathbf{x} = (A^\top Q^{-1} A)^{-1} A^\top Q^{-1} (\mathbf{y} - \mathbf{a}) \quad (5)$$

ist. (5) ist der Schätzer des linearen Gauß-Markoff-Modells, vgl. [1, Gl. (4.41)], und stellt gewissermaßen die Basis der Ausgleichsrechnung dar.

Schließlich wird noch ein dritter Fall betrachtet, in dem die Beobachtungen $\mathbf{y} = (\mathbf{y}_i)_i$ in n -dimensionale stochastisch unabhängige Beobachtungsgruppen $\mathbf{y}_i \sim \mathcal{N}_n(\boldsymbol{\mu}_i(x), s^2 Q_i)$ mit $\boldsymbol{\mu}_i(x) = \mathbf{a}_i + A_i x$ zerfallen. In dem Fall ist

$$\lambda(x, s^2 | \mathbf{y}) = C + \frac{N}{2} \log(s^2) + \frac{1}{2s^2} \sum_i \|\mathbf{y}_i - \boldsymbol{\mu}_i(x)\|_{Q_i^{-1}}^2 \quad (6)$$

mit der Konstante $C = \frac{1}{2} \sum_i \log(|2\pi Q_i|)$, wobei N die Gesamtdimension aller Beobachtungen ist. $\sum_i \|\mathbf{y}_i - \boldsymbol{\mu}_i(x)\|_{Q_i^{-1}}^2$ wird dann unabhängig von s^2 minimal, wenn

$$\mathbf{x} = \left(\sum_i A_i^\top Q_i A_i \right)^{-1} \sum_i A_i^\top Q_i (\mathbf{y}_i - \mathbf{a}_i) \quad (7)$$

ist. Häufig wird (5) intern mit (7) berechnet, um die üblicherweise vorhandene Block-Diagonal-Struktur von Q algorithmisch effizient zu nutzen. Im Folgenden wird sich diese Darstellung jedoch auch methodisch als sinnvoll erweisen.

2 Robuste Schätzung

Problematisch am Kleinste-Quadrate-Schätzer ist seine Anfälligkeit ggü. Ausreißern. Ein einziger grob falscher Messwert kann den Mittelwert beliebig weit verschieben. Huber hat daher in [2] vorgeschlagen, statt wie in (3) die Summe $\sum_i r_i^2$ der Quadrate der Residuen

$r_i = y_i - \mu$, die Summe $\sum_i \rho(r_i)$ anderer Verlustfunktionen ρ der Residuen zu minimieren, um dadurch robustere Sch  tzer zu erhalten, die er als M-Sch  tzer bezeichnet. Bspw. f  hrt die Minimierung der Summe der Absolutresiduen mit $\rho(r) = |r|$ auf den bekannterma  en robusten Mediansch  tzer $\mu = \text{med}_i(y_i)$. Huber basiert viele seiner Untersuchungen auf die Einflussfunktion $\psi(r) = \rho'(r)$. Passend zu den Ergebnissen von Huber gibt es den IRLS-Algorithmus (iteratively reweighted least squares, vermutlich auf unver  ffentlichte Arbeiten von Tukey zur  ckgehend, vgl. [3]), der mithilfe der Gewichtsfunktion $w(r) = \psi(r)/r$ Summen vom Typ $\sum_i \rho(r_i)$ mit oft nur wenigen Iterationen minimieren kann. Eine derartige Darstellung der Theorie findet sich z. B. in [1, Kap. 4.7.4].

In diesem Artikel wird eine andere Darstellung n  her an [4, Kap. 3.3] pr  sentiert, bei der ρ statt in den Residuen r_i in den halben quadrierten Residuen $\Omega_i = \frac{1}{2}r_i^2$ parametrisiert wird. Dies erlaubt eine einfachere Darstellung der Theorie bei mehrdimensionalen Beobachtungsgruppen.

2.1 Die Verteilung $\mathcal{W}_{w,n}$

In dieser Darstellung werden Verteilungen mit einem gewissen Grad an Symmetrie um ihr Zentrum μ betrachtet. Diese Verteilungen sollen durch die halbe quadratische Mahalanobisdistanz

$$\Omega_{\mu,\Sigma}(\mathbf{y}) = \frac{1}{2} \|\mathbf{y} - \mu\|_{\Sigma^{-1}}^2 \quad (8)$$

faktorisieren. Dazu wird eine n -dimensionale Verteilung $\mathcal{W}_{w,n}(\mu, \Sigma)$ mit Lageparameter $\mu \in \mathbb{R}^n$ und Skalenparameter $\Sigma \in \mathbb{R}^{n \times n}$ auf Basis einer integrierbaren Gewichtsfunktion $w : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ definiert. Diese soll die Wahrscheinlichkeitsdichte

$$p_{\mathcal{W}_{w,n}(\mu,\Sigma)}(\mathbf{y}) = \frac{\Gamma(\frac{n}{2})}{\sqrt{|2\pi\Sigma|}} \cdot p_{w,n}(\Omega_{\mu,\Sigma}(\mathbf{y})) \quad (9)$$

mit der Eulerschen Gammafunktion $\Gamma(k) = \int_0^\infty \exp(-t) t^{k-1} dt$ und

$$p_{w,n}(\Omega) = \frac{\exp(-\rho_w(\Omega))}{\Gamma_w(\frac{n}{2})} \quad \text{mit} \quad \rho_w(\Omega) = \int_0^\Omega w(\omega) d\omega \quad (10)$$

sowie der Normierungskonstante

$$\Gamma_w(k) = \int_0^\infty \exp(-\rho_w(t)) t^{k-1} dt \quad (11)$$

haben. Dieser Zugang unterscheidet sich von [4, Kap. 3.3] nur dahingehend, dass ρ in $\frac{1}{2}\|\mathbf{y} - \boldsymbol{\mu}\|_{\Sigma^{-1}}^2$ parametrisiert ist statt in $\|\mathbf{y} - \boldsymbol{\mu}\|_{\Sigma^{-1}}^2$ (ohne dem Vorfaktor $\frac{1}{2}$). Dieses Vorgehen bringt deutliche Vorteile in der Notation mit sich, u. a. den auch in [5, Gl. (2)] angedeuteten einfachen Zusammengang $\rho'_w(\Omega) = w(\Omega)$ zwischen Verlustfunktion ρ_w und zugehöriger Gewichtsfunktion w .

Nicht für alle Gewichtsfunktionen w ist $\mathcal{W}_{w,n}(\boldsymbol{\mu}, \Sigma)$ eine Wahrscheinlichkeitsverteilung. $w(\Omega) \cdot \Omega$ ist die *gewichtete* halbe quadratische Mahalanobisdistanz, an deren Grenzwert $\Omega_w = \lim_{\Omega \rightarrow \infty} w(\Omega) \cdot \Omega$ sich ablesen lässt, ob das Integral $\Gamma_w(n/2)$ konvergiert und $\mathcal{W}_{w,n}(\boldsymbol{\mu}, \Sigma)$ damit zu einer echten Wahrscheinlichkeitsverteilung macht. Existiert der Grenzwert Ω_w und gilt für diesen $\Omega_w > n/2$ (inkl. $\Omega_w = \infty$), so konvergiert $\Gamma_w(n/2)$. Gilt für den Grenzwert dagegen $\Omega_w < n/2$, so divergiert $\Gamma_w(n/2)$. Für $\Omega_w = n/2$ oder falls Ω_w nicht existiert, ist eine genauere Untersuchung notwendig. Divergiert $\Gamma_w(n/2)$, kann mit der uneigentlichen Wahrscheinlichkeitsdichte $h \exp(-\rho_w(\Omega_{\boldsymbol{\mu}, \Sigma}(x)))$ mit unbestimmtem Skalierungsfaktor h gearbeitet werden oder mit der auf $\Omega_{\boldsymbol{\mu}, \Sigma}(x) \leq \Omega_{\max}$ eingeschränkten Wahrscheinlichkeitsdichte, die entsteht, in dem man die Gewichtsfunktion formal mit $w(\Omega) = \infty$ für $\Omega > \Omega_{\max}$ anpasst, wodurch $\rho_w(\Omega) = \infty$ und damit $p_{w,n}(\Omega) = 0$ für $\Omega > \Omega_{\max}$ werden. Der Normierungsfaktor $\Gamma_w(n/2)$ in $p_{w,n}$ kann dann auch durch das unvollständige Integral $\gamma_w(n/2, \Omega_{\max})$ mit der ursprünglichen Gewichtsfunktion w ohne Anpassung ab Ω_{\max} ersetzt werden, für das gilt:

$$\gamma_w(k, T) = \int_0^T \exp(-\rho_w(t)) t^{k-1} dt. \quad (12)$$

Ist $\Omega_w > (n+1)/2$ existiert der Erwartungswert der $\mathcal{W}_{w,n}(\boldsymbol{\mu}, \Sigma)$ -Verteilung und ist $\boldsymbol{\mu}$. Ist $\Omega_w > (n+2)/2$ existiert auch die Kovarianzmatrix der Verteilung und ist $s_{w,n}^2 \Sigma$ mit dem Skalierungsfaktor

$$s_{w,n}^2 = \frac{\Gamma_w(\frac{n}{2} + 1)}{\frac{n}{2} \Gamma_w(\frac{n}{2})}, \quad \text{bzw.} \quad s_{w,n}^2 = \frac{\gamma_w(\frac{n}{2} + 1, \Omega_{\max})}{\frac{n}{2} \gamma_w(\frac{n}{2}, \Omega_{\max})}. \quad (13)$$

Ist $\Omega_w > (n + 3)/2$ existieren die 3. Zentralmomente der Verteilung und verschwinden, d. h. die Verteilung ist *symmetrisch*. Ist Ω_w exakt $(n + 1)/2$, $(n + 2)/2$ oder $(n + 3)/2$ ist jeweils eine genauere Untersuchung notwendig, ob die entsprechenden Momente existieren. Ist Ω_w dagegen kleiner, existieren sie nicht.

Die Verteilungen $\mathcal{W}_{w,n}(\boldsymbol{\mu}, \Sigma)$ haben ihren Modus in $\boldsymbol{\mu}$ und fallen um diesen herum symmetrisch ab. Dadurch ist zu erwarten, dass die Maximum-Likelihood-Methode auf diesem Verteilungstyp sinnvolle Ergebnisse liefert, da sich Fehler in alle Richtungen symmetrisch ausgleichen k  nnen.

F  r die konstante Gewichtsfunktion $w(\Omega) = s^{-2}$ ist $\Gamma_w(\frac{n}{2}) = s^n \Gamma(\frac{n}{2})$ und $\mathcal{W}_{w,n}(\boldsymbol{\mu}, \Sigma) = \mathcal{N}_n(\boldsymbol{\mu}, s^2 \Sigma)$ die n -dimensionale Normalverteilung mit Erwartungswert $\boldsymbol{\mu}$ und Kovarianzmatrix $s^2 \Sigma$. Insbesondere ist $\mathcal{W}_{w,n}(\boldsymbol{\mu}, \Sigma) = \mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$ f  r die konstante Gewichtsfunktion $w(\Omega) = 1$. F  r die konstante Gewichtsfunktion $w(\Omega) = 0$ ergibt sich die uneigentliche Gleichverteilung oder die Gleichverteilung auf $\Omega < \Omega_{\max}$, wenn sie mit $w(\Omega) = \infty$ f  r $\Omega > \Omega_{\max}$ bei Ω_{\max} abgeschnitten wird.

2.2 Iteratively Reweighted Least Squares (IRLS)

Mit $\boldsymbol{\mu} = \boldsymbol{\mu}(x)$ und $\Sigma = s^2 Q$ ist die negative Log-Likelihood von (9)

$$\lambda_{\mathcal{W}_{w,n}(\boldsymbol{\mu}(x), s^2 Q)}(x, s^2 | \mathbf{y}) = C + \frac{n}{2} \log(s^2) + \rho_w(s^{-2} \Omega_{\boldsymbol{\mu}(x), Q}(\mathbf{y})) \quad (14)$$

mit $C = \log\left(\frac{\Gamma_w(n/2)}{\Gamma(n/2)} \sqrt{|2\pi Q|}\right)$. F  r die n -dimensionalen Beobachtungsgruppen $\mathbf{y}_i \sim \mathcal{W}_{w,n}(\boldsymbol{\mu}_i(x), s^2 Q_i)$ ergibt sich analog zu (6) die zu minimierende negative Log-Likelihood

$$\lambda(x, s^2 | \mathbf{y}) = C + \frac{N}{2} \log(s^2) + \sum_i \rho_w(s^{-2} \Omega_i) \quad (15)$$

mit $C = \sum_i \log\left(\frac{\Gamma_w(n/2)}{\Gamma(n/2)} \sqrt{|2\pi Q_i|}\right)$ und $\Omega_i = \Omega_{\boldsymbol{\mu}_i(x), Q_i}(\mathbf{y}_i)$. Die Beobachtungsgruppen \mathbf{y}_i k  nnten auch unterschiedliche Dimensionen n_i und unterschiedliche Gewichtsfunktionen w_i haben, z. B. wenn unterschiedliche Beobachtungstypen wie 2D-Featurepunkte und 3D-GNSS-Messungen miteinander kombiniert werden, oder wenn die Beobachtungen gegen qualitativ unterschiedliche Ausre   er anf  llig sind. In

dem Fall ist $N = \sum_i n_i$ in Formel (15) die Gesamtdimension aller Beobachtungen.

Damit (15) minimal in \mathbf{x} ist, muss

$$0 = \frac{\partial}{\partial \mathbf{x}} \lambda(\mathbf{x}, s^2 | \mathbf{y}) = s^{-2} \sum_i \rho'_w(s^{-2} \Omega_i) \frac{\partial}{\partial \mathbf{x}} \Omega_i \quad (16)$$

$$= s^{-2} \sum_i w(s^{-2} \Omega_i) (\mathbf{y}_i - \boldsymbol{\mu}_i(\mathbf{x}))^\top Q_i^{-1} \frac{\partial}{\partial \mathbf{x}} \boldsymbol{\mu}_i(\mathbf{x}) \quad (17)$$

sein, und mit $\Omega_i = \Omega_{\boldsymbol{\mu}_i(\mathbf{x}), Q_i}(\mathbf{y}_i)$ und $w_i = w(\Omega_i / s^2)$ ist das im linearen Gauß-Markoff-Modell $\boldsymbol{\mu}_i(\mathbf{x}) = \mathbf{a}_i + A_i \mathbf{x}$ erfüllt, wenn

$$\mathbf{x} = \left(\sum_i w_i A_i^\top Q_i^{-1} A_i \right)^{-1} \sum_i w_i A_i^\top Q_i^{-1} (\mathbf{y}_i - \mathbf{a}_i) \quad (18)$$

ist. Abgesehen von den Gewichten w_i entspricht diese Formel gerade (7). Jedoch ist zu beachten, dass hier die w_i selbst sowohl von \mathbf{x} als auch von s^2 abhängen. Dennoch können startend von Näherungswerten \mathbf{x} und s^2 iterativ die Gewichte w_i berechnet werden und damit eine verbesserte Lösung für \mathbf{x} berechnet werden. Dies ist der IRLS-Algorithmus.

Um s^2 robust zu schätzen, gibt es mehrere Möglichkeiten, z. B. [1, Kap. 4.7.3] oder den mit leichtem Bias versehenen Maximum-Likelihood-Schätzer, der entsteht, wenn (15) nach s^2 abgeleitet und dessen Nullstelle berechnet wird. Das führt auf

$$s^2 = \frac{2}{N} \sum_i w_i \Omega_i \quad (19)$$

wobei zu beachten ist, dass $w_i = w(\Omega_i / s^2)$ selbst von s^2 abhängt und (19) aufgefasst als Fixpunktgleichung $v = f(v) = \frac{2}{N} \sum_i w(\Omega_i / v) \Omega_i$ mit $v = s^2$ nicht zwingend konvergieren muss.

2.3 Bekannte Gewichtsfunktionen

Barron [6] hat eine Funktionsfamilie aufgezeigt, die viele der in der Literatur bekannten Schätzer umfasst. In der hier gewählten Darstellung hat sie die Form

$$w_{\beta, s^2, k}(\Omega) = \frac{k}{s^2} w_{\beta, 1, 1} \left(\frac{\Omega}{s^2} \right) \quad \text{mit} \quad w_{\beta, 1, 1}(\Omega) = \left(1 + \frac{\Omega}{\beta} \right)^{-\beta} \quad (20)$$

für $0 < \beta < \infty$ mit den Grenzwerten $w_{0,1,1}(\Omega) = 1$ und $w_{\infty,1,1}(\Omega) = \exp(-\Omega)$. Die Familie umfasst die Normalverteilungen $w_{0,1,s^{-2}}$, den geglätteten Huber-Schätzer $w_{1/2,s^2,s^2}$, die n -dimensionale Cauchy-Verteilung $w_{1,s^2/2,(n+1)/2}$, den Geman-McClure-Schätzer $w_{2,s^2,1}$ und den Welsch-Schätzer w_{∞,s^2,s^2} . Es gilt

$$\rho_{w_{\beta,s^2,k}}(\Omega) = k \rho_{w_{\beta,1,1}}\left(\frac{\Omega}{s^2}\right) \quad (21)$$

$$\rho_{w_{\beta,1,1}}(\Omega) = \frac{\beta}{1-\beta} \left(\left(1 + \frac{\Omega}{\beta}\right)^{1-\beta} - 1 \right) \quad (22)$$

mit den Grenzwerten $\rho_{w_{0,1,1}}(\Omega) = \Omega$ und $\rho_{w_{\infty,1,1}}(\Omega) = 1 - \exp(-\Omega)$ und dem Sonderfall $\rho_{w_{1,1,1}}(\Omega) = \log(1 + \Omega)$.

$w_{\beta,s^2,k}$ führt für $\beta < 1$ auf eine Wahrscheinlichkeitsverteilung, zu der alle Momente existieren. Für $\beta > 1$ lässt sich die entstehende Verteilung dagegen nicht normieren und nur als uneigentliche oder abgeschnittene Wahrscheinlichkeitsverteilung verwenden. Für $\beta = 1$ hängt die Situation von dem Wert von k ab. Für $k > \frac{n}{2}$ entsteht eine n -dimensionale Wahrscheinlichkeitsverteilung zu der nur genau die Momente kleiner als $2k - n$ existieren. Für $k \leq \frac{n}{2}$ lässt sich die entstehende n -dimensionale Verteilung dagegen wieder nicht normieren.

[6] schlägt vor, startend von $\beta = 0$ schrittweise $\beta \rightarrow \infty$ laufen zu lassen, wodurch Ausreißer zunehmend abgewertet werden. In der hier gewählten Darstellung zeigt sich ein auffälliger Zusammenhang der Funktionsfamilie (20) zur bekannten Approximation $\exp(x) = \lim_{n \rightarrow \infty} (1 + \frac{x}{n})^n$ der Exponentialfunktion in $w_{\beta,1,1}$.

2.4 Mischungen von $\mathcal{W}_{w_i,n}$

Die Mischung zweier Zufallsvariablen y_0 und y_1 ist die Zufallsvariable $y = y_I$ mit dem zufälligen Index $I \sim \mathcal{B}(\varepsilon)$, der mit Wahrscheinlichkeit $\varepsilon \in [0,1]$ den Wert 1 annimmt und mit Wahrscheinlichkeit $1 - \varepsilon$ den Wert 0. Die Wahrscheinlichkeitsdichte p_y von y ist

$$p_y(y) = (1 - \varepsilon) p_{y_0}(y) + \varepsilon p_{y_1}(y) , \quad (23)$$

wobei p_{y_i} für $i = 0,1$ jeweils die Wahrscheinlichkeitsdichte von y_i ist. Sind $y_i \sim \mathcal{W}_{w_i,n}(\mu, \Sigma)$, so ist auch $y \sim \mathcal{W}_{w,n}(\mu, \Sigma)$ mit

$$p_{w,n}(\Omega) = (1 - \varepsilon) p_{w_0,n}(\Omega) + \varepsilon p_{w_1,n}(\Omega) . \quad (24)$$

Aus (10) folgt $p'_{w,n}(\Omega) = -p_{w,n}(\Omega) \cdot w(\Omega)$ und leitet man (24) nach Ω ab, erhält man nach Multiplikation mit -1

$$p_{w,n}(\Omega) w(\Omega) = (1 - \varepsilon) p_{w_0,n}(\Omega) w_0(\Omega) + \varepsilon p_{w_1,n}(\Omega) w_1(\Omega) . \quad (25)$$

Löst man das nach $w(\Omega)$ auf und ersetzt darin $p_{w,n}(\Omega)$ mit (24), so folgt

$$w(\Omega) = \frac{(1 - \varepsilon) p_{w_0,n}(\Omega) w_0(\Omega) + \varepsilon p_{w_1,n}(\Omega) w_1(\Omega)}{(1 - \varepsilon) p_{w_0,n}(\Omega) + \varepsilon p_{w_1,n}(\Omega)} . \quad (26)$$

Folgt y_0 der Wahrscheinlichkeitsverteilung der Inlier und y_1 der Wahrscheinlichkeitsverteilung der Ausreißer, so folgt y der Wahrscheinlichkeitsverteilung, die entsteht, wenn man Inlier mit einem Anteil von ε an Ausreißern kontaminiert.

2.5 Gleichverteilte Ausreißer

Es wird angenommen, dass die Inlier einer Normalverteilung mit Erwartungswert μ und Kovarianzmatrix Σ folgen, dass also konstant $w_0(\Omega) = 1$ ist. [4, Kap. 3.3] und [1, Fig. 4.13] schlagen beide die Gleichverteilung für die Ausreißer vor, geben jedoch nicht die dafür notwendige Gewichtsfunktion

$$w(\Omega) = (1 + k \exp(\Omega))^{-1} \quad (27)$$

an. Diese lässt sich aus (26) mit $w_0(\Omega) = 1$ und $w_1(\Omega) = 0$ durch kürzen des Zählers ermitteln, wobei sich $p_{w_1,n}(\Omega) = 1/\gamma_{w_1}(\frac{n}{2}, \Omega_{\max}) = \frac{n}{2} \Omega_{\max}^{-n/2}$ ergibt und $k = \varepsilon/(1 - \varepsilon) \cdot \Gamma(\frac{n}{2} + 1) \cdot \Omega_{\max}^{-n/2}$ substituiert wurde.

Für große k kann die Gewichtsfunktion (27) durch $w_{\infty,1,1/k}(\Omega) = \exp(-\Omega)/k$ angenähert werden, wobei k insbesondere dann groß wird, wenn der Ausreißeranteil ε groß ist. Diese Annäherung ist proportional zur Gewichtsfunktion des Welsch-Schätzers und liefert zu vorgegebenem s^2 daher im Grenzwert dieselben Ergebnisse.

Für (27) ergibt sich durch dividieren von (25) durch $w(\Omega)$ wegen $w_0(\Omega) = 1$ und $w_1(\Omega) = 0$

$$p_{w,n}(\Omega) = \frac{(1 - \varepsilon)p_{w_0,n}(\Omega)}{w(\Omega)} = \frac{1 - \varepsilon}{w(\Omega)} \cdot \frac{\exp(-\Omega)}{\Gamma(\frac{n}{2})} . \quad (28)$$

Ist I_i das Ereignis, dass es sich bei der i . Beobachtung um einen Inlier handelt, mit a-priori Wahrscheinlichkeit $P(I_i) = 1 - \varepsilon$, so ist mit (28) die a-posteriori Wahrscheinlichkeit zu gegebenem \mathbf{y}_i nach Bayes,

$$P(I_i | \mathbf{y}_i) = \frac{P(\mathbf{y}_i | I_i)P(I_i)}{P(\mathbf{y}_i)} = \frac{p_{w_0,n}(\Omega_i)(1 - \varepsilon)}{p_{w,n}(\Omega_i)} = w(\Omega_i) \quad (29)$$

mit $\Omega_i = \Omega_{\mu,\Sigma}(\mathbf{y}_i)$. Dadurch lassen sich die Summen der Form $\sum_i w_i T_i$ in (18) und (19) als empirische Erwartungswerte   ber die mit den Inlierwahrscheinlichkeiten $w_i = P(I_i | \mathbf{y}_i)$ gewichteten Beobachtungen \mathbf{y}_i auffassen. Diese Interpretation ist nur f  r die Gewichtsfunktion (27) m  glich, denn f  r andere Gewichtsfunktionen gilt im Allgemeinen $P(I_i | \mathbf{y}_i) \neq w(\Omega_i)$.

$W = \sum_i P(I_i | \mathbf{y}_i) = \sum_i w_i$ ist die a-posteriori zu erwartende Anzahl an Inliern, gegeben der Beobachtungen \mathbf{y}_i . Da diese mit der Anzahl M der n -dimensionalen Beobachtungen \mathbf{y}_i a-priori erwartungsgem    $E[W] = (1 - \varepsilon)M$ ist, ist bei gleichverteilten Ausre   ern zu vorgegebenem k

$$1 - \varepsilon = \frac{1}{M} \sum_i w_i = \bar{w} \quad (30)$$

ein Sch  tzer f  r den Inlieranteil.

Analog zu dem Vorgehen in [6] bietet es sich an, k startend von 0 schrittweise zu erh  hen, wodurch Ausre   er zunehmend abgewichtet werden. Mit (30) in (27) ist deren negative Log-Likelihood

$$\lambda(k | \mathbf{y}) = C - M \log(\bar{w}) + \sum_i \log(w_i) \quad (31)$$

mit von k unabh  ngigem C minimal in k , wenn

$$\frac{1}{M} \sum_i \log(w_i) - \log(\bar{w}) = \overline{\log(w)} - \log(\bar{w}) \quad (32)$$

minimal in k ist. Grunds  tzlich l  sst sich k durch Ableiten von (32) nach k und berechnen der Nullstellen bestimmen, jedoch f  hrt das auf komplizierte Formeln. Stattdessen ist es einfacher, k wachsen zu lassen, solange (32) f  llt und aufzuh  ren, sobald es zu steigen beginnt.

3 Zusammenhang zur Lifting Methode

In [7, Kap.3.4] weist Zach auf eine Lifting-Methode hin, die mit einer geeigneten Kernelfunktion dieselben Ergebnisse wie IRLS liefern kann, jedoch teilweise einen größeren Konvergenzbereich aufweisen soll. In etwas angepasster Notation wird dazu

$$\min_x \sum_i \rho_w(\Omega_i) = \min_{x,w} \sum_i (w_i \Omega_i + \varphi_\Omega(w_i)) \quad (33)$$

mit $\Omega_i = \Omega_{\mu_i(x), \Sigma}(y_i)$ minimiert, wobei die linke Seite hier für die Lösung des IRLS-Algorithmus steht (ohne Berücksichtigung eines Skalenparameters s^2 oder Parameter der Gewichtsfunktion w) und die rechte Seite die ersatzweise zu minimierende Funktion der Kernelmethode darstellt. Damit die beiden Lösungen übereinstimmen muss die Kernelfunktion φ_Ω zur Gewichtsfunktion w auf der linken Seite passen. Für monoton fallende w ist dies in der hier gewählten Darstellung besonders einfach, denn mit der Umkehrfunktion $\Omega(w)$ von $w(\Omega)$ und $w_0 = w(0) = \max_\Omega w(\Omega)$ ist

$$\varphi_\Omega(w) = \int_w^{w_0} \Omega(\omega) d\omega . \quad (34)$$

Für gleichverteilte Ausreißer ist bspw. $\Omega(w) = \log\left(\frac{1-w}{kw}\right)$ durch auflösen von (27) nach Ω und durch integrieren ergibt sich hierfür

$$\varphi_\Omega(w) = (1-w) \log\left(\frac{1-w}{kw}\right) + \log(w) . \quad (35)$$

Analog lassen sich die Gewichtsfunktionen (20) nach Ω auflösen und integrieren, was mit elementaren Mitteln machbar ist, allerdings zu etwas sperrigen Ausdrücken führt.

4 Zusammenfassung

In diesem Artikel wurden die auf Huber [2] zurückgehenden M-Schätzer und der IRLS-Algorithmus (iteratively reweighted least squares) zu deren Berechnung betrachtet. Es wurde ein alternativer Zugang dazu gegeben, der die Verlustfunktionen und Gewichtsfunktionen im halben quadratischen Fehler parametrisiert. Dadurch entfallen viele der sonst notwendigen Zwischenschritte und die Darstellung

wird schlanker. Auch werden Querverbindungen sichtbar, die in der üblichen Darstellung verborgen bleiben. Diese ist zum einen ein Zusammenhang zur eulerschen Gammafunktion über (11), der auch z. B. in (13) bemerkbar wird; zum anderen ein Zusammenhang einer Familie bekannter robuster Gewichtsfunktionen [6] zur Approximation der Exponentialfunktion über (20); und schließlich ein Zusammenhang einer Kernel-Lifting-Methode [7], in dem die Verlustfunktion ρ_w des IRLS-Algorithmus auf symmetrische Weise mit der Kernelfunktion φ_Ω der Lifting-Methode über die Umkehrfunktion $\Omega(w)$ der Gewichtsfunktion $w(\Omega)$ zusammenhängt.

Des Weiteren wurde mit gleichverteilten Ausreißern eine sehr gut interpretierbare Gewichtsfunktion (27) durchgesprochen, die zwar an mehreren Stellen insbesondere zur anschaulichen Argumentation angeschnitten wird, aber deren Eigenschaften scheinbar nirgends ausführlich behandelt werden.

Literatur

1. W. Förstner and B. Wrobel, *Photogrammetric Computer Vision: Geometry, Orientation and Reconstruction*, ser. Geometry and Computing. Springer International Publishing, 2016.
2. P. J. Huber, “Robust Estimation of a Location Parameter,” *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.
3. P. W. Holland, “Weighted ridge regression: Combining ridge and robust regression methods,” National Bureau of Economic Research, Tech. Rep., 1973.
4. B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, “Bundle adjustment — a modern synthesis,” in *International Workshop on Vision Algorithms*, 2000, pp. 298–372.
5. C. Zach and G. Bourmaud, “Descending, lifting or smoothing: Secrets of robust cost optimization,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 547–562.
6. J. T. Barron, “A general and adaptive robust loss function,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4331–4339.
7. C. Zach, “Robust bundle adjustment revisited,” in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 772–787.