Auswertung Projektaufgabe Block 2

Paul Reichelt (527691), Raphael Zöllner (527767)

# Theoretische Grundlagen

Zur Analyse von Verteilungen betrachtet man deren Schiefe und Moden.

Unter der Schiefe einer Verteilung versteht man deren Neigungsstärke. Sie zeigt an wie stark die Verteilung nach links oder rechts geneigt ist. Jede nichtsymmetrische Verteilung bezeichnet man als schief.

Die Häufungspunkte einer Verteilung bezeichnet man als Moden.

Um die Dichte einer Verteilung zu schätzen nutzt man sogenannte Kerndichteschätzer. Diese sind die Faltung des Kerns mit dem empirischen Maß. Kerne sind dabei Funktionen mit und Regularitätsbedingungen. Einige bekannte Kerne sind z.B. Rechteckskern, Gaußkern und Epanechnikovkern.

Unter dem allgemeinen Regressionsmodell

wurde ein Glättungsverfahren zur Schätzung der Daten ohne Messfehler mit Hilfe der Methode der nächsten Nachbarn implementiert. Der geschätzte Datenwert ist dabei der Mittelwert der k umliegenden Ausgangsdaten.

# Algorithmische Problemstellung & Darstellung der Ergebnisse

Im ersten Teil der Aufgabe haben wir eingelesene Daten analysiert. Hierfür haben wir deren empirische Verteilungsfunktion und das Histogramm betrachtet. Zudem haben wir mit Hilfe der density Funktion verschiedene Kerndichteschätzer mit unterschiedlichen Bandweiten auf die gegebenen Daten angewendet und das Ergebnis graphisch aufgearbeitet. Anschließend haben wir einen Test auf Normalverteilungshypothese durchgeführt.

Im zweiten Teil haben wir uns mit Glättungsverfahren auseinandergesetzt. Dafür haben wir die Methode der nächsten Nachbarn auf einen zuvor eingelesenen Vektor mit verrauschten Daten angewendet.

smooth <- function(k, vals){

newVals <- numeric(0)

for(i in 1:k){

newVals[i] <- mean(vals[0:(i+k)])

newVals[length(vals)-(i-1)] <-

mean(vals[(length(vals-i-1)-k):length(vals)])

}

for(i in (k+1):(length(vals)-k)){

newVals[i] <- mean(vals[(i-k):(i+k)])

}

newVals

}

Danach haben wir dieses Verfahren auf zweidimensionale Ausgangsdaten erweitert. Unter den nächsten Nachbarn werden hierbei alle Datenpunkte im Abstand von zum Ausgangspunkt verstanden. Diese Methode haben wir auf ein verrauschtes Bild angewandt, indem wir die Grauwerte der Pixel als zweidimensionale Matrix aufgefasst haben.

smooth2 <- function(vals,k){

uniI <- 0

uniJ <- 0

for(i in -k:k){

for(j in -k:k){

len <- norm(c(i,j), type="2")

if(len <= k && len > 0)

{

uniI <- c(uniI, i)

uniJ <- c(uniJ, j)

}

}

}

newVals <- matrix(nrow = dim(vals)[1], ncol = dim(vals)[2])

for(i in 1:dim(vals)[1]){

for(j in 1:dim(vals)[2]){

circleI <- uniI+i

circleJ <- uniJ+j

list <- numeric(0)

for(k in 1:length(circleI)){

if(circleI[k]>0 && circleJ[k]>0 && circleI[k]<=dim(vals)[1]

&& circleJ[k]<=dim(vals)[2]){

list <- c(list,vals[circleI[k],circleJ[k]])

}

}

newVals[i,j] <- mean(list)

}

}

newVals

}

# Diskussion und Auswertung

Die uns vorliegenden Daten aus der Datei snowfall.txt haben wir eingelesen und aufgearbeitet. Dafür haben wir die empirische Verteilungsfunktion und ein Histogramm erstellt (siehe Abb. 1 und 2).

Aufgrund des Histogramms vermuten wir, dass die gegebene Verteilung zu unseren Daten 3 Moden besitzt, da sich eine Häufung von Werten in 3 Bereichen erkennbar ist. Diese Vermutung ist auch an der empirischen Verteilungsfunktion erkennbar, da die Steigung dieser in den gleichen Bereichen abflacht.

Zudem lässt uns das Histogramm vermuten, dass die Verteilung leicht linksschief, also nahezu symmetrisch ist.

Anschließend haben wir dem Histogramm Kerndichteschätzer mit Epanechnikovkern und Gaußkern hinzugefügt. Als Bandweiten nutzten wir 3, 5 und den Standardwert. Die Kerndichteschätzer bestätigen unsere Vermutung. Das Ergebnis lässt darauf schließen, dass die Bandweite 3 die beste Schätzung ist.

Betrachten wir jedoch den Shapiro-Wilk-Tests auf Normalverteilung, so liefert dieser keine signifikanten Ergebnisse die Normalverteilungshypothese zu verwerfen, da .

Im zweiten Teil glätten wir verrauschte Daten in Form eines Vektors mit Hilfe eines Schätzers ähnlich dem Nadaraya-Watson-Schätzer. Auf die uns vorliegenden Daten aus der Datei noisysignal.txt haben wir unsere Methode zur Glättung angewandt mit einer Bandweite von 5. Das Ergebnis haben wir in den Abbildungen 5 veranschaulicht. Als vergleich sieht man in Abbildung 4 die verrauchten Daten. Wie man sieht, konnten wir die unsortiert wirkende Punktewolke in eine geglättete Kurve umgewandeln.

Des Weiteren lag uns ein verrauschtes Schwarzweißbild vor (siehe Abbildung 7). Dieses haben wir zunächst in eine Matrix umgewandelt, wobei diese die Graustufen der jeweiligen Pixel als Einträge enthält. Auf diese verrauschten Daten haben wir unsere erweiterte Implementierung des Glättungsverfahrens angewandt mit Bandbreiten 2, 3, 5 und 8. Man sieht, dass das Bild mit steigender Bandbreite zwar deutlicher wird, aber auch verschwommener. Die geglätteten Bilder haben wir in Abbildung 8 bis 14 dargestellt. Dabei kann man deutlich erkennen, dass bei einem Radius von 2 und 3 das Bild noch undersmoothed vorliegt und bei 8 bereits oversmoothed. Das beste Ergebnis liefert unserer Meinung nach die Bandbreite 5 mit der euklidischen Norm, da Konturen noch deutlich erkennbar sind und das Bild nicht zu stark verschwommen ist.

# Anlagen

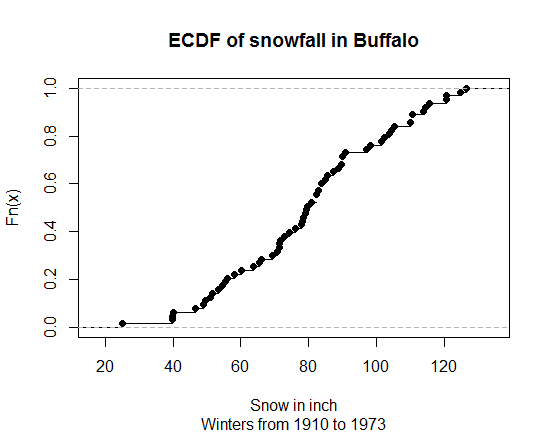
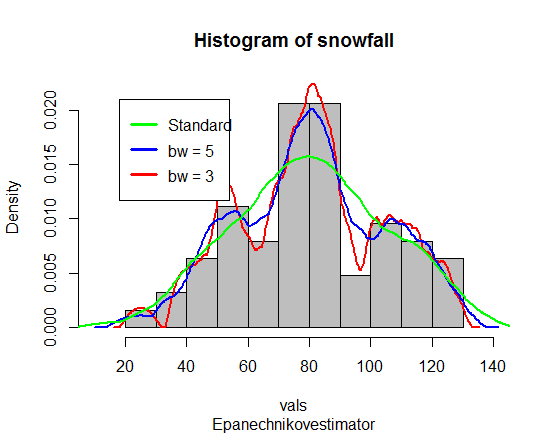
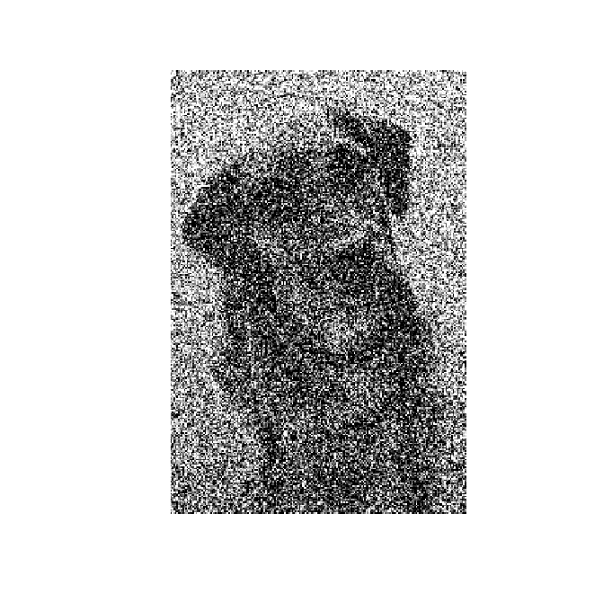
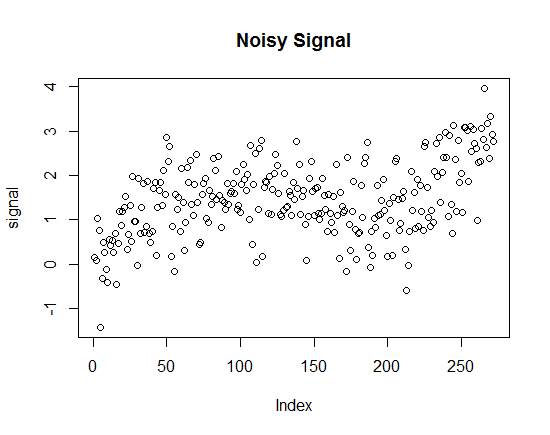
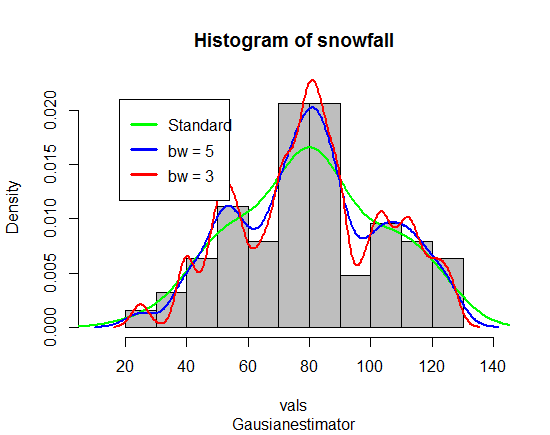
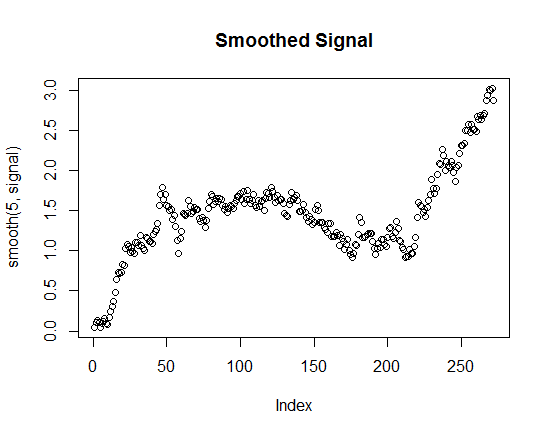


Abbildung 3

Abbildung 4

Abbildung 5

Abbildung 6 Ausgangsbild noisyimage.bmp

Abbildung 2

Abbildung 1

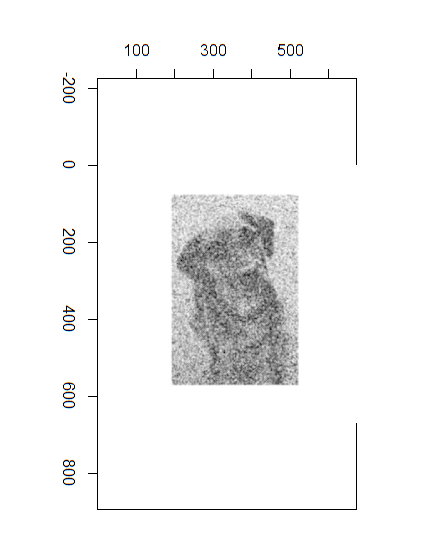
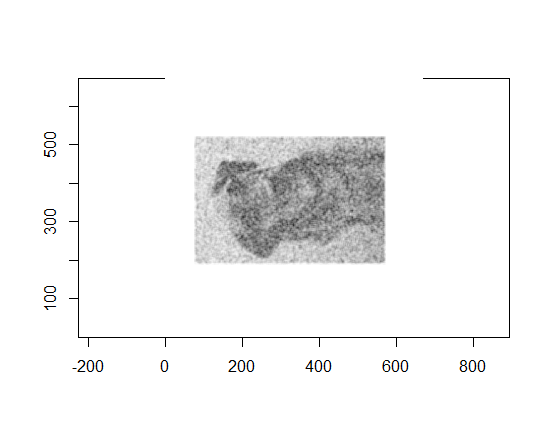
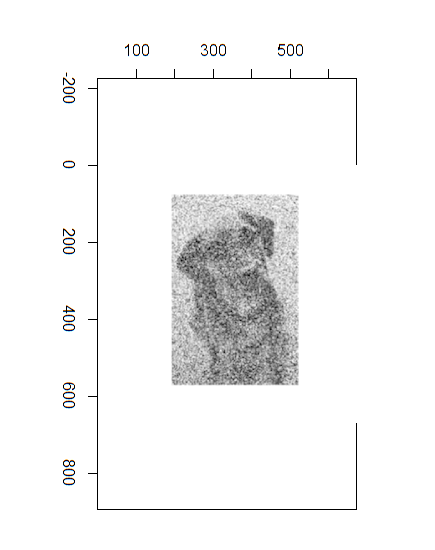
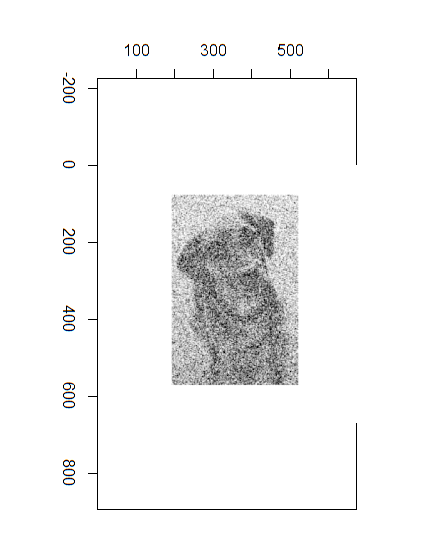


Abbildung 7 (euklidische Norm, Bandbreite 2)

Abbildung 8 (maximums Norm, Bandbreite 2)

Abbildung 10 (maximums Norm, Brandbreite 3)

Abbildung 9 (euklidische Norm, Bandbreite 3)

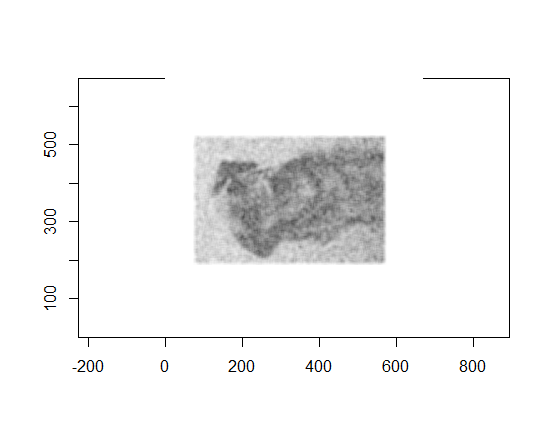
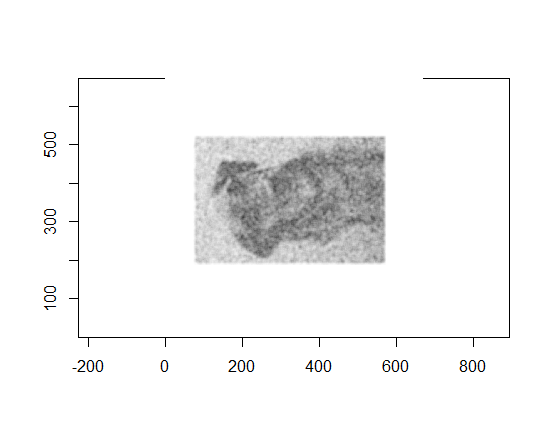
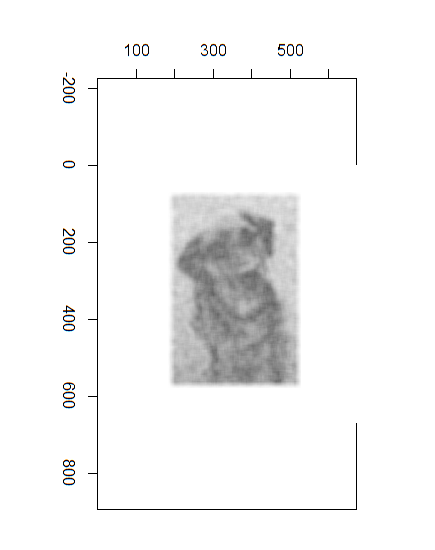
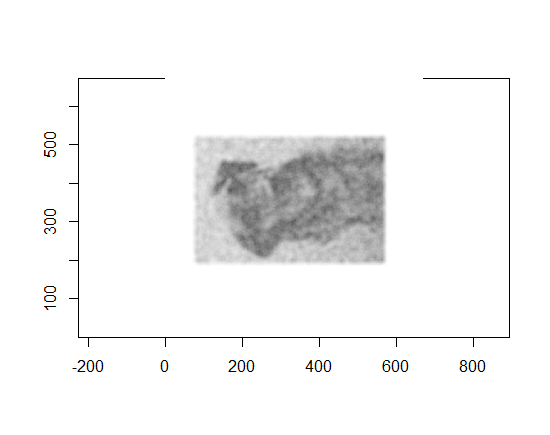


Abbildung 13 (euklidische Norm, Bandbreite 8)

Abbildung 14 (euklidische Norm, Bandbreite 8)

Abbildung 11 (euklidische Norm, Bandbreite 5)

Abbildung 12 (euklidische Norm, Bandbreite 5)