Randolf Altmeyer

BZQ II: Stochastikpraktikum
Wintersemester 2016
Humboldt-Universität zu Berlin

Projektaufgaben Block 2

1 Nichtparametrisches Testen (5P)

1. In einer Zwillingsstudie wurde von 8 Zwillingspaaren jeweils ein Zwilling zufällig ausgewählt für den Besuch eines Kindergartens. Der andere Zwilling blieb zu Hause. Am Ende der Studie wurden die sozialen Fähigkeiten beider Kinder überprüft. Die Testergebnisse in der Kindergartengruppe und in der zweiten Gruppe sind 82,69,73,43,58,56,76,65 bzw. 63,42,74,37,51,43,80,62 (geordnet entsprechend den Zwillingspaaren). Verwende den zweiseitigen T-Test (in R: `t.test`) und den zweiseitigen Wilcoxon-Vorzeichen-Test (in R: `wilcox.test`), um zu überprüfen, ob die Testergebnisse signifikant sind für das Fehlerniveau $\alpha = 0.05$, um die Nullhypothese zu verwerfen, dass der Kindergartenbesuch keinen nennenswerten Einfluss auf die sozialen Fähigkeiten hat. Ändert sich etwas, wenn man einseitige Tests benutzt?

2. Seien $X_1, \dots, X_n \overset{iid}{\sim} \theta + \varepsilon$, $n = 30$, wobei ε eine symmetrische Verteilung hat. Wähle eine Verteilung für ε und schätze für den T-Test und den Wilcoxon-Vorzeichen-Test die Wahrscheinlichkeit den Test abzulehnen mit einer Monte-Carlo-Simulation für $\theta \in \{0, 0.5, 1\}$ und $\alpha = 0.05$. Hängen die Ergebnisse von ε ab?

   Hinweis: Verwende ohne Beweis, dass beide Tests jeweils die Form $\mathbf{1}(p(X_1, \dots, X_n) < \alpha)$ haben für entsprechende p-Werte.

2 Dichteschätzung (5P)

1. Kerndichteschätzer:

   (a) Implementiere einen allgemeinen Kerndichteschätzer $\hat{f}_n(x)$ für gegebene Daten $X_1, \dots, X_n$ an einer Stelle x für einen Kern K und eine Bandweite h.

   (b) Wähle eine beliebige Dichte f und erzeuge Zufallszahlen $X_1, \dots, X_n$ bezüglich f. Teste die Methode aus (a) für verschiedene K und h. Plotte dazu die wahre Dichte zusammen mit der Geschätzten.

   (c) Seien die $X_1, \dots, X_n$ gegeben durch die Eruptionsdauern im Datensatz `faithful`. Teste die Methode aus (a) für den Gaußkern und verschiedene h.

2. Kreuzvalidierung zur Bandweitenwahl:

   (a) Zeige für den Schätzer $\hat{J}(h)$, $h > 0$, im Kreuzvalidierungskriterium unter geeigneten Annahmen an den Kern K und die Dichte f (welche Annahmen?), dass $\mathbb{E}[\hat{J}(h)] = J(h)$.

(b) Implementiere den Schätzer $\hat{J}(h)$. Finde damit eine „optimale“ Bandweite h^* für die Daten in 1(c). Vergleiche h^* mit dem Ergebnis der R-Methode bw.ucv.

Freiwillig (2 Extrapunkte): Implementiere zusätzlich zum Kerndichteschätzer in 1. die $knn(m)$-Methode (m-nächste Nachbarn). Dabei ist $\hat{f}_{knn}(x) = \frac{1}{nh(x,m)} \sum_{k=1}^{n} K(\frac{X_k - x}{h(x,m)})$ für einen Kern K und $h(x,m)$ ist der Abstand zwischen x und dem m-nächsten Nachbarn. Vergleiche mit den Ergebnissen in 1. Gibt es Unterschiede?

3 Bildentrauschen (10P)

1. Lies die ersten 4 Abschnitte in der offiziellen Dokumentation des R-Packets *EBImage* (siehe web), um die Hauptfunktionen zu lernen. Installiere und lade das Packet wie in der Dokumentation beschrieben.

2. Lade das Bild lena.png (siehe web) und wandle es in Graustufen um.

3. Erzeuge ein verrauschtes Bild mit Pixeln $Y = (Y_{ij})_{i=1,\dots,m,j=1,\dots,p}$ durch $Y_{ij} = f(i,j) + \varepsilon_{ij}$, wobei $(f(i,j))_{i=1,\dots,m,\ j=1,\dots,p}$ die Pixel des ursprünglichen Bildes sind und $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$ für ein beliebiges Rauschniveau $\sigma > 0$. Passe Y gegebenenfalls an, damit $Y_{ij} \in [0,1]$ gilt.

4. Betrachte den Nadaraya-Watson-Schätzer aus der Vorlesung.

   (a) Zeige, dass $K_2(x,y) = K(x)K(y)$, $x, y \in \mathbb{R}$, einen zweidimensionalen Kern definiert, wenn K ein eindimensionaler Kern ist.

   (b) Gib einen geeigneten Nadaraya-Watson-Schätzer $(\hat{f}_{mp}(i,j))_{i=1,\dots,m,j=1,\dots,p}$ an für $(f(i,j))_{i=1,\dots,m,j=1,\dots,p}$ mit dem Kern K_2 aus (a). Begründe deine Wahl.
   Hinweis: Anders als in der Vorlesung werten wir $\hat{f}_{mp}$ nur auf den Designpunkten (i,j), $i = 1,\dots,m, j = 1,\dots,p$, aus.

   (c) Zeige, dass $\hat{f}_{mp}(i,j) = (MYN^\top)_{ij}$ gilt für zwei Matrizen $M \in \mathbb{R}^{m\times m}$ und $N \in \mathbb{R}^{p\times p}$, die nicht von den Daten Y, sondern nur vom Kern und der Bandweite abhängen.

   (d) Implementiere den Nadaraya-Watson-Schätzer mit der Methode aus (c) für einen allgemeinen Kern K. Erzeuge die Matrizen M, N mit der R-Funktion **outer**.

5. Berechne $\hat{f}_{mp}$ für die Kerne $K(x) = K_{Gauss}(x) = 1/\sqrt{2\pi} \cdot \exp(-x^2/2)$ und $K(x) = K_{rect}(x) = 1/2 \cdot \mathbf{1}[-1,1](x)$ und verschiedene Bandweiten. Kommentiere die Unterschiede für verschiedene Kerne, Bandweiten und Rauschniveaus σ, insbesondere auch für Pixel am Rand, Farbflächen und Kanten im Bild.

6. Wiederhole dieselbe Analyse für das Bild porsche.jpg von der Webseite. Gibt es Unterschiede zur Analyse des ersten Bildes?

7. Fragen zum Rauschen $(\varepsilon_{ij})_{i=1,\dots,m,j=1,\dots,p}$:

   (a) Untersuche und beschreibe (kurz!) wie robust der Nadaraya-Watson-Schätzer bezüglich dem Entrauschen eines Bildes ist, wenn die ε_{ij} *nicht* normalverteilt sind.

   (b) Gib einen Schätzer an, der robuster ist (d.h. weniger abhängig von der Verteilung der ε_{ij}) als der Nadaraya-Watson-Schätzer.
   Tipp: Für den Nadaraya-Watson-Schätzer gilt

$$\hat{f}_n(i,j) = \operatorname{argmin}_{y\in\mathbb{R}} \left(\sum_{l_1=1}^{m} \sum_{l_2=1}^{p} \left(Y_{l_1 l_2} - y\right)^2 w_{l_1 l_2}(i,j) \right).$$

   Wie kann man diese Definition anpassen?

(c) *Freiwillig (bis zu 5 Extrapunkte):* Implementiere den neuen Schätzer und teste ihn an den obigen Bildern. Verwende nicht-normalverteiltes Rauschen.

Hinweise:

- Alternativ kann der Nadaraya-Watson-Schätzer auch mit einer Schleife implementiert werden, d.h. wir berechnen $\hat{f}_{mp}(i,j)$ für jedes Pixel getrennt. Allerdings ist der Rechenaufwand pro Paar (i,j) relativ hoch und damit schon bei wenigen Pixeln nicht mehr akzeptabel.
- Verwende beim Testen sehr kleine Bilder. Betrachte dafür Teilausschnitte, z.B. mit lena[299:376, 224:301].
- Für die Programmabgabe bitte alle Bilder mit relativen Pfadangaben aus dem darüberliegenden Ordner laden, z.B. ../lena.png.

Freiwillig (bis zu 5 Extrapunkte): Recherchiere andere Methoden zum Bildentrauschen und erkläre kurz (!) die Hauptideen, eventuell auch mit Beispielcode in R.

Hinweise zur Abgabe:

Alle Dateien (pdf der Auswertung, R-Code für jede einzelne Aufgabe in eigener Datei, aber OHNE Bilder!!) „gepackt“ (z.B. mit zip) mit dem Namen *UE2_Student1_Student2* bis zum 06.12. per Email an *altmeyrx@math.hu-berlin.de* schicken.