Dieses PDF ist keine Musterlösung zur Prüfung sondern eine Hilfestellung. Hier werden modifizierte Prüfungsbeispiele herangezogen, um zu zeigen, wie diese berechnet werden und zu verstehen sind. Eine derart ausführliche Beantwortung der Prüfungsbeispiele war weder bei der letzten noch wird sie bei der kommenden Prüfung erwartet.

Wichtiges zur Prüfung

TUGonline ist ein vorläufiger Prüfungstermin eingetragen. Wie auch beim ersten Antritt sollten sich die Prüfungskandidaten anmelden. Bei wem sich dieser Termin nicht ausgeht oder wer die Prüfung aus technischen Gründen nicht digital abhalten kann, soll dies schnellstmöglich per E-Mail bekanntgeben, damit wir hier eine Lösung finden können. Ein nicht entschuldigtes Fernbleiben wird als "Nicht Angetreten"gewertet und führt zum Verlust eines Antritts (womöglich des zweiten und somit letzten → führt zur Gesamtnote "Nicht Genügend").

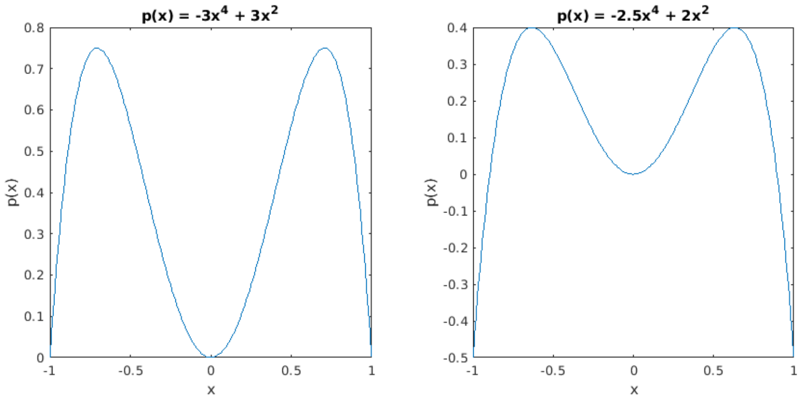
Die zweite Prüfung wird in der gleichen Weise wie die erste abgehalten werden. Es wird allerdings die Kritik der Studierenden miteinbezogen (Länge der Beispiele, Dauer der Prüfung, etc.). Es wird zeitnah zur Prüfung eine aktualisierte Version des "Infoblattes zur Prüfung" auf TeachCenter hochgeladen.

Die Prüfungskandidaten sollen sich schnellstmöglich (am besten sofort) zur Prüfung anmelden, damit sie alle prüfungsrelevanten E-Mails erhalten, da diese nur an die zur Prüfung angemeldeten Studenten ausgesendet werden.)

1 Wahrscheinlichkeitsdichten

1) Wahrscheinlichkeitsverteilungen

a) Gegeben seien die Funktionen in der nachstehenden Abbildung. Welche dieser beiden Funktionen taugt als Wahrscheinlichkeitsdichte für das Intervall [-1,1]. Können beide Funktionen als Wahrscheinlichkeitsdichte genutzt werden? Wieso kann man diese nutzen? Wieso kann man diese nicht nutzen? Falls eine Funktion nicht als Wahrscheinlichkeitsdichte taugt, gäbe es eine Möglichkeit diese doch noch zu verwenden?



b) Gegeben sei die Wahrscheinlichkeitsdichte

$$p(x) = \delta(x-a) = \begin{cases} \infty & \text{für } a = x \\ 0 & \text{sonst} \end{cases} .$$

Berechnen Sie die charakteristische Funktion dieser Wahrscheinlichkeitsdichte.

Hinweis: Die Delta-Distribution $\delta(x-a)$ hat folgenden Einfluss auf ein Integral

$$\int_{-\infty}^{+\infty} \delta(x-a) f(x) \mathrm{d}x = f(a)$$

c) Damit eine Funktion überhaupt als Wahrscheinlichkeitsdichte herangezogen werden kann, muss diese zwei Bedingungen erfüllen. Diese beiden Bedingungen sind die Normiertheit und die Nichtnegativität ($p(x) \geq 0$ für jedes x im betrachteten Intervall) der jeweiligen Funktion. Diese beiden Eigenschaften hätten wir zu Beginn der Rechnung überprüfen sollen, um festzustellen, ob weitere Berechnungen überhaupt sinnvoll sind. Erfüllt die gegebene Wahrscheinlichkeitsdichte die Bedingung der Nichtnegativität? Ist diese Wahrscheinlichkeitsdichte normiert? Ob Sie zur Überprüfung der Normiertheit die Wahrscheinlichkeitsdichte oder die charakteristische Funktion verwenden, liegt bei Ihnen.

d) Berechnen Sie als Nächstes den Mittelwert und die Varianz dieser Wahrscheinlichkeitsdichte. Es obliegt Ihnen, ob Sie für diese Berechnungen die charakteristische Funktion oder die Wahrscheinlichkeitsdichte nutzen wollen.

Wie sind das Ergebnis für den Erwartungswert und die Varianz zu verstehen? Hätte man diese Ergebnisse auch an der Wahrscheinlichkeitsdichte ablesen können? Wenn ja, wie?

e) Gegeben sei folgende Kovarianzmatrix C

$$C = \begin{pmatrix} \mathrm{Var(X)} & \mathrm{Cov(X,Y)} & \mathrm{Cov(X,Z)} \\ \mathrm{Cov(Y,X)} & \mathrm{Var(Y)} & \mathrm{Cov(Y,Z)} \\ \mathrm{Cov(Z,X)} & \mathrm{Cov(Z,Y)} & \mathrm{Var(Z)} \end{pmatrix} = \begin{pmatrix} 0.14 & 2.50 & 0.05 \\ 2.50 & 1.20 & 0.00 \\ 0.05 & 0.00 & 4.20 \end{pmatrix},$$

wobei Var() die Varianz und Cov(,) die Kovarianz bezeichnen.

Welche Informationen können Sie dieser Kovarianzmatrix entnehmen? Welche Werte sind korreliert? Welche nicht? Kann man über die Korrelation der verschiedenen Größen sonst noch etwas aussagen? Welche Werte streuen besonders stark? Welche eher weniger stark?

Bei genauerer Betrachtung der Kovarianzen in dieser Matrix scheint ein Widerspruch vorzuliegen. Welcher könnte das sein?

Lösungen

a) Jede Wahrscheinlichkeitsdichte **muss** folgende drei Bedingungen erfüllen:

1. Die Wahrscheinlichkeitsdichte muss integrierbar sein. Der Grund hierfür ist, dass die Wahrscheinlichkeit das Integral über die Wahrscheinlichkeitsdichte ist.
2. Die Wahrscheinlichkeitsdichte ist nicht negativ. Diese Bedingung muss erfüllt sein, da Wahrscheinlichkeiten nur Werte im Intervall [0,1] annehmen können/dürfen. Bei negativen Werten gäbe es Bereiche in denen wir negative Wahrscheinlichkeiten erhalten würden.
3. Die Wahrscheinlichkeitsdichte muss normiert/normierbar sein. Streng genommen können nur normierte Funktionen als Wahrscheinlichkeitsdichten verwendet werden. Wir können dieses Kriterium auch etwas freier auslegen. → Sofern die Funktion normierbar ist, können wir diese einfach selber normieren.

Um zu bestimmen, welche der gegebenen Funktionen als Wahrscheinlichkeitsdichte verwendet werden kann, überprüfen wir nun die Erfüllung dieser drei Kriterien

Linke Abbildung: Diese Funktion lässt sich integrieren. Sie ist überall größer oder gleich null. Falls diese Funktion normiert ist, können wir sie als Wahrscheinlichkeitsdichte verwenden (strenge Auslegung der Normiertheit), andernfalls müssten wir sie noch normieren (freiere Auslegung der Normiertheit).

Rechte Abbildung: Auch diese Funktion lässt sich integrieren. Sie ist allerdings für Werte um $x \pm 1$ negativ und kann nicht als Wahrscheinlichkeitsdichte verwendet werden. Um diese Funktion doch verwenden zu können, könnten wir uns z.B. auf das Intervall $[-0.5, +0.5]$ beschränken oder eventuell eine passende Konstante addieren (z.B. p'(x) = p(x) + c). Den Parameter c müssten wir allerdings so wählen, dass die Funktion keine negativen Werte mehr aufweist. Für die Normierung gilt das gleiche wie für die vorherige Funktion

Kommentar: Einige haben die linke Funktion ausgeschlossen, da p(x=0) = 0. Dies ist allerdings kein Ausschlusskriterium. Dies bedeutet nur, dass der Wert x = 0 nicht angenommen werden kann. Dies soll durch folgendes Beispiel verbildlicht werden:

Beispiel: Wir gehen von drei Kisten mit der Kantenlänge 20 entlang der x-Achse aus. Wir stellen diese drei Kisten genau nebeneinander. Und versuchen nun Kugeln in diese Kisten zu werfen. Am Ende messen wir die x-Koordinate der **in** den Kisten gelandeten Kugeln. Falls nun auf der mittleren Kiste ein Deckel ist, können in dieser keine Kugeln landen. Somit wäre die Wahrscheinlichkeit, dass Kugeln sich im Intervall [20,40] befinden, gleich null.

Außerdem sollte man sich überlegen, ob Änderungen wie z.B. für die rechte Abbildung Sinn machen oder, ob man nicht doch lieber eine ganz andere Wahrscheinlichkeitsdichte verwenden sollte.

b) Die charakteristische Funktion wird über folgendes Integral berechnet

$$\Phi(\omega) = \int_{-\infty}^{+\infty} p(x) e^{i\omega x} dx. \tag{1}$$

Wir setzen nun die Wahrscheinlichkeitsdichte aus der Angabe ein und erhalten

$$\Phi(\omega) = \int_{-\infty}^{+\infty} \delta(x-a) e^{i\omega x} dx = e^{i\omega a}. \tag{2}$$

c) Nun sollen wir überprüfen, ob diese Wahrscheinlichkeitsdichte geeignet ist. Hierzu müssen die in **(a)** genannten Kriterien erfüllt sein. Das erste Kriterium wird in unserer LV immer erfüllt sein (in anderen LVs muss das Kriterium eventuell überprüft werden). Das zweite Kriterium ist auch erfüllt (kann in diesem Fall direkt in der Angabe abgelesen werden). Für das dritte Kriterium müssen wir allerdings etwas rechnen. Hierfür haben wir drei Möglichkeiten.

1. Wir **wissen**, dass
$$Z = \int_{-\infty}^{+\infty} \delta(x-a) dx = 1. \tag{3}$$
2. Wir verwenden $f(x) = 1$ und erhalten mithilfe des Hinweises
$$Z = \int_{-\infty}^{+\infty} 1 \cdot \delta(x-a) dx = 1. \tag{4}$$
3. Wir verwenden die charakteristische Funktion und erhalten
$$Z = i^{-0} \Phi(\omega = 0) = e^0 = 1. \tag{5}$$

Kommentar: Einige haben versucht, dass Integral der ersten Methode zu berechnen und dabei $Z = \infty$ erhalten, was zu einem ein unbestimmten Ausdruck ($Z \approx \infty \cdot 0$) führt. Dieser kann nicht direkt ausgewertet werden sondern nur mittels spezieller Methoden.

d) Wir berechnen den Erwartungswert (E) und die Varianz (Var) zunächst über die Wahrscheinlichkeitsdichte

$$\mathrm{E} = \langle x \rangle = \int_{-\infty}^{+\infty} x \delta(x-a) dx = a \tag{6}$$

$$\mathrm{Var} = \left\langle x^2 \right\rangle - \langle x \rangle^2 = \int_{-\infty}^{+\infty} x^2 \delta(x-a) dx - a^2 = a^2 - a^2 = 0. \tag{7}$$

Wir haben hier erneut den Hinweis sowie $f(x) = x$ und $f(x) = x^2$ verwendet.

Wir können alternativ auch die charakteristische Funktion verwenden um Erwartungswert und Varianz zu bestimmen. Hierzu müssen wir folgende Schritte durchführen (siehe Übungszettel 3):

1. Berechnen der charakteristischen Funktion (haben wir in **(a)** gemacht)
2. Die charakteristische Funktion für das n-te Moment n-mal nach ω ableiten.
3. Nach dem Ableiten ω null setzen und durch i^n dividieren

In diesem Beispiel sind diese Schritte sehr einfach durchzuführen und können auch im Kopf durchgeführt werden. Wir erhalten

$$E = \langle x \rangle = \frac{ia \cdot e^{ia0}}{i^1} = a \tag{8}$$

$$\mathbf{Var} = \frac{(ia)^2 \cdot e^{ia0}}{i^2} - \left(\frac{ia \cdot e^{ia0}}{i^1} \right)^2 = a^2 - a^2 = 0. \tag{9}$$

Ist dieses Ergebnis überraschend? Die Antwort sollte "Nein" lauten. Entsprechend unserer Wahrscheinlichkeitsdichte kann nur der Wert $x = a$ angenommen werden. Damit **muss** der Erwartungswert a und die $Varianz$ null sein.

SEHR WICHTIG:

- Die Normierung ist das **nullte** Moment einer Wahrscheinlichkeitsverteilung
- Der Erwartungswert ist das **erste** Moment einer Wahrscheinlichkeitsverteilung
- Die Varianz ist das **zweite zentrale** Moment einer Wahrscheinlichkeitsverteilung (daher das $-\langle x \rangle^2$ in der Berechnung der Varianz)

Die Varianz kann man sich folgendermaßen merken: Die Varianz ist die mittlere quadratische Abweichung der Zufallsvariablen vom Mittelwert. $\rightarrow$ Die Varianz hat also etwas mit dem Mittelwert (Erwartungswert) $\langle x \rangle$ zu tun.

Zusatzinfo: Wer sich nun fragt, wo eine solche Wahrscheinlichkeitsdichte überhaupt Sinn macht, kann sich überlegen, wie man einen Würfel oder eine Münze mit dieser Wahrscheinlichkeitsdichte beschreiben kann.

e) Wir beginnen zunächst mit der Streuung. Entsprechend der Kovarianzmatrix können wir ablesen, dass der Parameter X am wenigsten und der Parameter Z am stärksten streuen. Die Streuung des Parameters Y liegt irgendwo dazwischen.

Die Kovarianzen sind ein Maß dafür, wie stark zwei Parameter miteinander korrelieren. Anhand der Kovarianzmatrix können wir vermuten, dass die Parameter X und Y am stärksten korrelieren und die Parameter Y und Z gar nicht korrelieren, während die Parameter X und Z schwach korrelieren.

Bei Betrachtung der Kovarianzmatrix macht es zunächst keinen Sinn, dass das Parameterpaar (Y,Z) nicht korrelieren, während die übrigen Paare es tun. Dies muss aber kein Fehler sein. Wir diskutieren hierzu zwei Beispiele:

Ungenauigkeit des Ergebnisses: Die Varianzen und Kovarianzen in der Angabe sind auf die zweite Nachkommastelle gerundet. Wir könnten daher eventuell die Korrelation dieses einen Wertepaares auf null gerundet haben (z.B. Cov(Y,Z) = 0,002). Hier sollte man darauf achten, dass solche "kleinen"Rundungsfehler in einigen Fällen das Ergebnis stark ändern könnten.

Strom in Abhängigkeit des Widerstandes und der Spannung: Strom (I), Spannung (U) und Widerstand (R) eines ohmschen Widerstandes hängen über folgende Gleichung zusammen

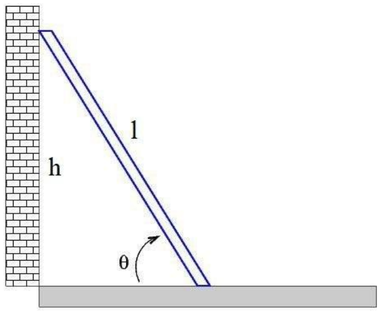
$$\frac{U}{R} = I. \tag{10}$$

Eine höhere Spannung bedeutet einen höheren Strom ein höherer Widerstand bedeutet einen geringeren Strom. Allerdings ändert die angelegte Spannung nicht den Widerstand. Ähnlich zu unserer Kovarianzmatrix hängen auch hier die Wertepaare (U,I) sowie (R,I) zusammen, während das Wertepaar (U,R) nicht zusammenhängt.

Kommentar: Wie im Vorlesungsskript besprochen hängt die Kovarianz von verschiedenen Faktoren ab, wodurch ein Vergleich schwierig werden kann. In diesem Fall kann es durchaus von Vorteil sein, die Korrelation nicht anhand der Kovarianz zu bewerten, sondern sich die Mühe zu machen und zusätzlich auch noch die Korrelationskoeffizienten zu berechnen.

2) Wahrscheinlichkeitsverteilungen

Jemand lehnt eine Leiter an die Wand. Der Winkel Θ den die Leiter mit dem Boden einschließt sei gleichverteilt und liege in dem Intervall $\left[0, \frac{\pi}{4}\right]$ rad. Bei einer Leiter kommt es nun aber nicht darauf an, welchen Winkel Θ diese mit dem Boden einschließt, sondern wie hoch diese reicht. Daher interessiert uns nicht die Gleichverteilung $p(\Theta)$ sondern die Wahrscheinlichkeitsdichte $p(h)$.



a) Berechnen Sie daher durch geeignete Variablentransformation die Wahrscheinlichkeitsdichte $p(h)$. Vergessen Sie dabei nicht anzugeben, für welche Werte von h diese Wahrscheinlichkeitsdichte gilt. Warum ist bei der Variablentransformation eine Ableitung zu berücksichtigen?

Hinweis: Es gilt $\frac{\delta}{\delta x}\arcsin(x) = \frac{1}{\sqrt{1-x^2}}$

b) Berechnen Sie nun die Wahrscheinlichkeitsdichte an der Stelle $h = \frac{\sqrt{3}}{2}l$. Die **Funktion** $p(h)$ liefert hier zwei Werte als Ergebnis. Liefert auch die Wahrscheinlichkeitsdichte $p(h)$ diese beiden Werte? Begründen Sie Ihre Antwort. Worauf müssen Sie bei der Berechnung dieser Wahrscheinlichkeitsdichte im Hinblick auf die möglichen Ergebnisse besonders achten?

c) Als nächstes sind Varianz und Kovarianz der Parameter h und Θ zu bestimmen. Dies ist allerdings zu aufwendig, weshalb wir Ihnen stattdessen die folgende Kovarianzmatrix (samt fiktiver Einträge) zur Verfügung stellen

$$C = \begin{pmatrix} \text{Var}(\Theta) & \text{Cov}(h,\Theta) \\ \text{Cov}(\Theta,h) & \text{Var}(h) \end{pmatrix} = \begin{pmatrix} 0.04 & 0.00 \\ 0.00 & 10.20 \end{pmatrix} .$$

Hierbei ist Var() die Varianz der entsprechenden Größe. Cov(,) bezeichnet die Kovarianz der beiden betroffenen Größen. Ihre Aufgabe ist es nun sich zu überlegen, ob die hier gegebenen Werte Sinn machen oder nicht. Überlegen Sie sich hierzu wie h und Θ zusammenhängen und welche Ergebnisse Sie sich für die Varianzen und die Kovarianz erwarten würden. Begründen Sie Ihre Entscheidung.

Lösung

a) Wir müssen folgende Transformation durchführen

$$p_h(h) = p_\Theta(\Theta)\left|\frac{\delta\Theta}{\delta x}\right| . \tag{11}$$

Der Zusammenhang zwischen Θ und h ist gegeben durch

$$\frac{h}{l} = \sin(\Theta) \longrightarrow \Theta = \arcsin\left(\frac{h}{l}\right) \tag{12}$$

Entsprechend der Angabe ist unser Winkel gleichverteilt. Unter Verwendung des Hinweises erhalten wir nun

$$p_h(h) = \frac{1}{\frac{\pi}{4} - 0} \cdot \frac{1}{\sqrt{1 - \left(\frac{h}{l}\right)^2}} = \frac{4}{\pi\sqrt{1 - \left(\frac{h}{l}\right)^2}}. \tag{13}$$

h liegt im Intervall $\left[0, l\sin\left(\frac{\pi}{4}\right)\right] = \left[0, \frac{l}{\sqrt{2}}\right]$. Die Ableitung folgt aus dem Transformieren eines Integrals (siehe Feedbackr-Antworten auf TeachCenter).

b) Wir erhalten nun das Ergebnis

$$p\left(h = \frac{\sqrt{3}}{2}l\right) = \frac{4}{\pi\sqrt{1 - \frac{3}{4}}} = \pm\frac{8}{\pi} \tag{14}$$

Die Wahrscheinlichkeitsdichte und die Funktion liefern nicht die gleichen Ergebnisse. Wie wir im vorherigen Beispiel besprochen haben, darf eine Wahrscheinlichkeitsdichte keine negativen Werte liefern. Für die Wahrscheinlichkeitsdichte dürfen wir daher nur das positive Ergebnis der "Funktion" p(h) verwenden.

c) Die Einträge in dieser Kovarianzmatrix können auf den ersten Blick so nicht stimmen. Da die Höhe der Leiter von dem Winkel Θ abhängt sollte die Kovarianz einen Wert größer null annehmen (je größer der Winkel Θ desto größer die Höhe h). Außerdem scheinen auch die Varianzen nicht zu stimmen. Bei stark streuenden Winkeln wäre zu erwarten, dass auch die Höhe stärker streuen sollte.

Kommentar: Hier wäre zu überlegen woher die Diskrepanz zwischen unseren Erwartungen und den gegebenen Werten stammen könnte. Sind diese Werte falsch berechnet oder stimmen sie und unsere Erwartungen sind einfach falsch (vgl. Kovarianzmatrix des vorangegangen Beispiel).

2 Parameterschätzen

1) Parameterschätzen

a) Es soll mithilfe des MaxEnt-Prinzips ein Prior bestimmt werden. Die Lagrangefunktion, die wir hierzu lösen sollen, laute

$$\mathcal{L} = -\int p(y)\ln(p(y))\mathrm{dy} - \lambda_0 \left[\int p(y)\mathrm{dy} - 1\right] - \lambda_1 \left[\int y\, p(y)\mathrm{dy} - 10\right] - \lambda_2 \left[\int (y-10)^2\, p(y)\mathrm{dy} - 2.5\right]$$

Das Berechnen des Priors ist hier nicht von Interesse. Wir interessieren uns lediglich für den Informationsgehalt des Priors. Welche Nebenbedingungen werden über die Parameter λ_i eingeführt? Welche weiteren Informationen über diese Nebenbedingungen können Sie der Lagrangefunktion entnehmen?

b) Wir messen in einem Experiment eine Größe y. Wie es bei vielen Experimenten üblich ist, erhalten wir nicht den wahren Wert y sondern einen verrauschten Wert z. Unser Messergebnis sei durch ein additives Gaußsches Rauschen verfälscht

$$z = y + \eta.$$

Geben Sie die Likelihood für dieses Problem an. Der Wert von y folgt der nachstehenden Funktion. Drücken Sie daher y in Ihrer Likelihood durch diese Funktion aus

$$y = Ax + B.$$

Wir sind nun interessiert am Maximum-Likelihood-Schätzwert für x. Nutzen Sie daher die Maximum-Likelihood-Methode, um diesen Schätzwert zu bestimmen. A und B sind dabei folgendermaßen zu wählen: Der Wert für A entspricht ihrem Geburts**monat**, während jener für B Ihrem Geburts**tag** entspricht.

c) Ist dieser Schätzwert verzerrt? Führen Sie hierzu die Rechnung explizit durch. Worauf könnte ein verzerrter Schätzwert hinweisen? Ist es ratsam einen verzerrten Schätzwert zu verwenden oder sollte man hiervon abraten? Überlegen Sie sich hierzu welchen Wert ihr bester Schätzwert τ_{ML} für τ im Idealfall liefern sollte. Denken Sie dabei nicht zu kompliziert.

d) Zusätzlich zum Maximum-Likelihood-Schätzwert wollen wir uns auch noch den Maximum-a-posteriori-Schätzwert berechnen. Hierzu gehen wir davon aus, dass der Prior folgende Form besitzt

$$p(\alpha|\mathcal{B}) = \frac{1}{\alpha}.$$

Berechnen Sie nun den Maximum-a-Posteriori-Schätzwert. Welchen Einfluss hat der Prior auf den Schätzwert? Ist der Maximum-a-Posteriori-Schätzwert kleiner, gleich groß oder größer als der Maximum-Likelihood-Schätzwert? Können Sie diese Änderung anhand des Priors erklären?

Erklären Sie außerdem, warum jemand den Maximum-Likelihood-Schätzwert verwenden sollte oder warum doch der Maximum-a-posteriori-Schätzwert zu bevorzugen ist.

e) Wir sind mit Ihrer Abschätzung für den Maximum-Likelihood-Schätzwert nicht so ganz zufrieden, da dieser nur für den Fall eines flachen Priors eine vernünftige Lösung ist. Daher möchten wir von Ihnen, dass Sie Ihr α_{ML} benutzen, um einen besseren Prior zu bestimmen und den Maximum-a-posteriori-Schätzwert zu berechnen. Sind Sie bereit dies zu tun? Falls ja, wieso ist es möglich Ihr Ergebnis auf diese Weise zu verbessern? Falls nein, wieso weigern Sie sich? Haben wir in unseren Überlegungen einen Fehler gemacht? Ist dieses Vorgehen ratsam oder nicht?

Lösung

a) Wir können aus der Lagrangefunktion folgende Informationen ablesen.

- λ_0 führt die Nebenbedingung *Normierung = 1* ein.
- λ_1 führt die Nebenbedingung *Erwartungswert = 10* ein.
- λ_2 führt die Nebenbedingung *Varianz = 2.5* ein.

Kommentar: Wie bereits besprochen ist die Varianz das zweite zentrale Moment und hängt mit dem Mittelwert zusammen. Dies können wir auch an dieser Lagrangefunktion wunderbar sehen.

Es kann durchaus Sinn machen sich im Vorfeld einige Gedanken zu den Berechnungen zu machen. Für den Fall, dass jemand das MaxEnt-Prinzip anwendet und mit der gegebenen Lagrangefunktion eine Wahrscheinlichkeitsdichte berechnet, kann das Ergebnis am Ende überprüft werden. Die resultierende Wahrscheinlichkeitsdichte sollte normiert sein und einen Erwartungswert von 10 sowie eine Varianz von 2,5 aufweisen. Falls sie dies nicht tut, hat man sich wohl verrechnet.

b) Das Rauschen ist ein Zufallsprozess und wird durch eine Normalverteilung beschrieben (Gaußsches Rauschen). Damit ist unsere Wahrscheinlichkietsdichte gegeben durch

$$p(\eta|\mathcal{B}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\eta^2}{2\sigma^2}} \tag{15}$$

Aus der Angabe wissen wir nun, dass

$$\eta = z - y \tag{16}$$

Einsetzen in die Wahrscheinlichkeitsdichte liefert nun

$$p(\eta|\mathcal{B}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-y)^2}{2\sigma^2}} \tag{17}$$

In dieser Wahrscheinlichkeitsdichte kommt η nicht mehr vor. Wir müssen diese daher umschreiben. Wir überlegen uns daher, was gegeben ist und was wir bestimmen möchten. In unserem Fall sei z ein Messwert und y der unbekannte wahre Wert. Wir sind an einer Likelihood interessiert. Wir suchen daher eine Wahrscheinlichkeitsdichte, die uns für einen wahren Wert y die Wahrscheinlichkeit liefert, z zu messen. Die Wahrscheinlichkeitsdichte ist somit gegeben durch

$$p(z|y,\mathcal{B}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-y)^2}{2\sigma^2}}. \tag{18}$$

Wichtig ist hier zu beachten, dass die hier auftretende Varianz jene des Rauschens ist ($\sigma = \sigma_\eta$). Entsprechend der Angabe können wir y als Funktion von x ausdrücken und erhalten

$$p(z|x,\mathcal{B}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-Ax-B)^2}{2\sigma^2}}. \tag{19}$$

Nun berechnen wir den Maximum-Likelihood-Schätzwert. Hierzu müssen wir folgende Schritte durchführen:

1. Aufstellen der Likelihood (Haben wir bereits erledigt)
2. Umschreiben der Likelihhood (falls möglich) und Logarithmieren der Likelihood (das Logarithmieren muss nicht getan werden, kann aber die Rechnung vereinfachen)
3. Ableiten **nach dem gesuchten Parameter** und Nullsetzen der Ableitung
4. Umformen nach dem gesuchten Parameter

Angenommen dass es N Messpunkte gibt, die zusätzlich unabhängig voneinander sind, können wir die Likelihood umschreiben und erhalten

$$p(\vec{z}|x,\mathcal{B}) = p(z_1,z_2,...z_N|x,\mathcal{B}) = \prod_{j=1}^{N} p(\vec{z}_i|x,\mathcal{B}) \tag{20}$$

Wir haben in unserem Beispiel einen einzigen Messpunkt. Daher gilt N=1. In unserem Fall erleichtert der Logarithmus die Berechnung und wir erhalten

$$\frac{\mathrm{d}}{dx}\ln\left(p(\vec{z}|x,\mathcal{B})\right) = ... = \frac{\mathrm{d}}{dx}\frac{(z-Ax-B)^2}{2\sigma^2} \stackrel{!}{=} 0 \tag{21}$$

Wir führen nun diese Ableitung durch und formen um. Wir erhalten nun als Maximum-Likelihood-Schätzwert

$$x_{\mathrm{ML}} = \frac{z-B}{A} \tag{22}$$

Kommentar: Mit den Berechnungen hier haben wir nur festgestellt, dass x_{ML} ein Extremum ist. In den bisherigen Beispielen hat dies ausgereicht, da unsere Wahrscheinlichkeitsdichten weder Minima noch Sattelpunkte hatten. Allerdings könnten Wahrscheinlichkeitsdichten auch diese Extrema aufweisen (vgl. Abbildungen im ersten Beispiel). In diesem Fall kann man die zweite Ableitung heranziehen (Stichwort Kurvendiskussion) oder man setzt die verschiedenen Ergebnisse in die Wahrscheinlichkeitsdichte ein und bestimmt auf diese Art das Maximum. Wie mit zwei gleichwertigen Maxima (siehe erneut Beispiel 1) umgegangen wird hängt von der Situation ab.

c) Um zu bestimmen, ob ein unverzerrter Schätzwert vorliegt, müssen wir überprüfen, ob $\langle x_{\mathrm{ML}}\rangle = x$ gilt. Hierzu berechnen wir folgendes Integral

$$\langle x_{\mathrm{ML}}\rangle = \int_{-\infty}^{+\infty} x_{\mathrm{ML}}\, p(z|x,\mathcal{B})\mathrm{dz}. \tag{23}$$

Wir setzen unsere Wahrscheinlichkeitsdichte und unseren ML-Schätzwert ein und führen die Substitution $s = z - B$ durch. Damit erhalten wir nun

$$\langle x_{\mathrm{ML}}\rangle = \int_{-\infty}^{+\infty} \frac{s}{A}\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(s-Ax)^2}{2\sigma^2}}\mathrm{ds} \tag{24}$$

Nun haben wir zwei Möglichkeiten diesen Ausdruck auszuwerten. Die erste Möglichkeit wäre das direkte Berechnen des Integrals, die zweite Möglichkeit wäre festzustellen, dass wir hier nichts anderes tun, als den Mittelwert einer Normalverteilung mit dem Mittelwert $\mu = Ax$ zu berechnen. Somit können wir das Ergebnis sofort ablesen und erhalten sofort

$$\langle x_{\mathrm{ML}} \rangle = \int_{-\infty}^{+\infty} \frac{s}{A} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(s-Ax)^2}{2\sigma^2}} \mathrm{d}s = \frac{Ax}{A} = x \tag{25}$$

Unser Mittelwert ist somit tatsächlich unverzerrt. Welchen Wert sollte unser Schätzwert nun annehmen? In diesem Beispiel wollen wir x abschätzen, damit sollte unser Schätzwert idealerweise x sein. Worauf könnte ein verzerrter Schätzwert hinweisen? Ein verzerrter Schätzwert könnte darauf hinweisen, dass im Experiment oder in der Rechnung etwas furchtbar schief gelaufen ist. Z.B. könnte man sich verrechnet haben. Ein Beispiel für einen Bias in einem Experiment (der zu einem falschen Schätzwert führen würde) wäre das Messen der Körpergröße mit angezogenen Schuhen. Dies würde z.B. bei High Heels zu falschen Ergebnissen führen. Einen falschen Schätzwert sollte man daher keinesfalls verwenden.

Kommentar: Wie könnte man überprüfen, ob ein Schätzwert richtig bestimmt wurde? Eine Möglichkeit wäre z.B. verschiedene Experimente durchzuführen und zu untersuchen, ob diese Experimente den gleichen Schätzwert liefern. Ein einfaches Beispiel wäre die Widerstandsbestimmung. Einerseits könnte man den Widerstand direkt messen, andererseits könnte man Strom und Spannung bestimmen und mittels ohmschen Gesetz den Widerstand berechnen. Eine andere Möglichkeit wäre das Vorhersagen von Ergebnissen. Z.B.: Wie zuvor messen wir den Widerstand direkt. Nun legen wir eine Spannung an. Mit dem ohmschen Gesetz können wir voraussagen wie groß der Stromfluss sein sollte. Falls bei diesen beiden Vorgehensweisen eine Diskrepanz auftritt, stimmt vermutlich etwas mit dem Schätzwert nicht.

d) Nun berechnen wir den Maximum-a-posteriori-Schätzwert. Der Maximum-Posterior-Schätzwert wird auf die gleiche Weise wie der Maximum-Likelihood-Schätzwert berechnet. Der einzige Unterschied besteht darin, dass wir nun den Posterior anstelle der Likelihood verwenden. Stark vereinfacht formuliert könnten wir hier einfach sagen: Ersetze in den Schritten 1-4 aus **()b)** das Wort Likelihood durch Posterior. Wir bestimmen zunächst den Posterior. Dieser ergibt sich in unserem Beispiel zu

$$p(x|z,\mathcal{B}) = \frac{1}{Z} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-Ax-B)^2}{2\sigma^2}} \frac{1}{x} \tag{26}$$

Der Nenner Z muss für das Parameterschätzen nicht berechnet werden, da dieser nicht vom gesuchten Parameter abhängt und im Laufe der Rechnung rausfällt. Wir Logarithmieren nun den Posterior und leiten diesen ab

$$\frac{\mathrm{d}}{dx} \ln\left(p(x|z,\mathcal{B})\right) = ... = \frac{\mathrm{d}}{dx}\left[\frac{(z-Ax-B)^2}{2\sigma^2} - ln(x)\right] \stackrel{!}{=} 0 \tag{27}$$

Ableiten, Umformen und Anwenden der kleinen Lösungsformel ergibt

$$\langle x_{\mathrm{MAP}} \rangle = \frac{B'}{2} \pm \sqrt{\frac{B'^2}{4} + C'}, \tag{28}$$

mit $B' = \frac{z-B}{A}$ und $C' = \frac{2\sigma^2}{A}$. Wir haben nun das Problem, dass unser MAP-Schätzwert zwei Ergebnisse liefert (siehe Kommentar zu Aufgabenteil **(b)**). Wir könnten mit diesem Ausdruck weiterrechnen oder uns mit diesem Ergebnis begnügen. Wichtig ist hier nicht das explizite Berechnen des Schätzwertes sondern zu wissen, welche Rechenschritte zum Erreichen des gesuchten Ergebnisses nötig sind (siehe die beschriebenen Rechenschritte in Aufgabenteil **(b)**). Außerdem sollte man folgende Argumentation nachvollziehen bzw. selber liefern können: Entsprechend des gegebenen Priors sind kleine Werte für x wahrscheinlicher als große Werte. Dies bedeutet für unseren MAP-Schätzwert, dass dieser kleiner als unser Maximum-Likelihood-Schätzwert ist.

Der MAP-Schätzwert berücksichtigt den Prior, welcher wiederum ein gewisses Vorwissen enthält, und sollte daher falls möglich immer bevorzugt werden.

Kommentar: Man könnte sich in zwei Fällen auch für den ML-Schätzwert aussprechen. Falls keinerlei Vorwissen vorherrscht (flacher Prior) genügt es, den ML-Schätzwert zu berechnen. Im Falle eines flachen Priors sind allerdings ML- und MAP-Schätzwert ident. Falls man keinen geeigneten Prior berechnen kann (nicht weiß wie das geht), wäre es eventuell besser den ML-Schätzwert zu berechnen, da man mit einem falschen Prior das Ergebniss stark verschlechtern kann.

e) Je nach zugeteilter Prüfung gab es hier verschiedenste Angaben. Einige hatten zu entscheiden, ob der berechnete ML-Schätzwert verwendet werden sollte, um einen besseren Prior zu bestimmen (z.B. mithilfe des MaxEnt-Prinzips). Andere mussten sich entscheiden, ob sie die Hälfte der Daten für einen ersten ML-Schätzwert verwenden, mit welchem sie einen besseren Prior abschätzen, um im nächsten Schritt mit diesem Prior und der verbleibenden Hälfte der Datenpunkte stattdessen einen MAP-Schätzwert zu bestimmen und die andere Hälfte der Daten für ihre Likelihood verwenden.

- Verbesserung des Prior mithilfe des ML-Schätzwertes: Von diesem Vorgehen ist **unbedingt** abzuraten. Da ein solches Vorgehen bedeuten würde, dass wir die Daten zweimal gemessen haben. Wir würden damit Daten aus dem Nichts erschaffen. Jeder Datenpunkt darf in der Rechnung ein einziges Mal verwendet werden !!!
- Aufteilen der Datenpunkte: Dieses Vorgehen wäre prinzipiell erlaubt, da jeder Datenpunkt nur einmal in die Rechnung eingeht. Allerdings entspräche dieses Vorgehen einfach einem Umschreiben des ursprünglichen Problems anstatt einer tatsächlichen Verbesserung des Ergebnisses.

3 Hypothesentests

1) Hypothesentest

Für die folgende Rechnung benötigen wir zwei Zahlen n und m, die wir folgendermaßen bestimmen: n sei Ihr Geburts**tag**, während m Ihr Geburts**monat** sei.

Sie haben n+m Würfel zur Verfügung. n Würfel sind konventionell fair. m Würfel haben die folgenden Wahrscheinlichkeiten (WSL) die entsprechenden Zahlen zu werfen.

Zahl	1	2	3	4	5	6
WSL	1/21	2/21	3/21	4/21	5/21	6/21

a) Sie wählen zufällig einen dieser Würfel und werfen ihn 50 Mal. Sie zählen dabei 4 Mal die Augenzahl 1. Stellen Sie eine Hypothese über die Fairness des Würfels auf und überprüfen Sie diese. Geben Sie hierzu eine Formel zur Berechnung der Posterior-Wahrscheinlichkeit an und bestimmen Sie das ODDS-Ratio.

Wie ist die Änderung der Posterior-Wahrscheinlichkeit gegenüber der Prior-Wahrscheinlichkeit zu verstehen? Ist der Würfel nun fairer/manipulierter als vor dem Würfeln oder hat die Wahrscheinlichkeit hier eine andere Bedeutung?

b) Bestimmen Sie außerdem die Posterior-Wahrscheinlichkeit dafür, dass der gewählte Würfel fair ist. Würden Sie anhand der Posterior-Wahrscheinlichkeit dieselbe Schlussfolgerung ziehen?

c) Welche Augenzahl/en würden Sie mitzählen, um mit möglichst wenig Versuchen ein aussagekräftiges Ergebnis zu erhalten? Begründen Sie Ihre Antwort.

d) Ein Kollege führt einen Versuch mit unbekanntem Würfel durch. Er teilt Ihnen mit, welchen Mittelwert und welche Standardabweichung er erwürfelt hat. Wie können Sie abschätzen, ob dieser Würfel fair oder manipuliert ist?

Lösung

a) Wir führen folgende Propositionen ein

- f/m: Der Würfel ist fair/manipuliert
- n: Die Gesamtanzahl an Würfelwürfen
- n_i: Die Anzahl wie oft bei n Würfen die Augenzahl i gewürfelt wurde

Die Posterior-Wahrscheinlichkeit für einen fairen Wüfel berechnet sich über

$$p(f|n,n_i,\mathcal{B}) = \frac{p(n_i|f,n,\mathcal{B})p(f|\mathcal{B})}{p(n_i|n,\mathcal{B})} = \frac{\binom{n}{n_i}\frac{1}{6}^{n_i}\left(1-\frac{1}{6}\right)^{n-n_i}\frac{n}{n+m}}{\binom{n}{n_i}\frac{1}{6}^{n_i}\left(1-\frac{1}{6}\right)^{n-n_i}\frac{n}{n+m}+\binom{n}{n_i}\frac{1}{21}^{n_i}\left(1-\frac{1}{21}\right)^{n-n_i}\frac{m}{n+m}} \tag{29}$$

In gleicher Weise erhalten wir für den manipulierten Würfel

$$p(m|n,n_i,\mathcal{B}) = \frac{p(n_i|m,n,\mathcal{B})p(m|\mathcal{B})}{p(n_i|n,\mathcal{B})} = \frac{\binom{n}{n_i}\frac{1}{21}^{n_i}\left(1-\frac{1}{21}\right)^{n-n_i}\frac{m}{n+m}}{\binom{n}{n_i}\frac{1}{6}^{n_i}\left(1-\frac{1}{6}\right)^{n-n_i}\frac{n}{n+m}+\binom{n}{n_i}\frac{1}{21}^{n_i}\left(1-\frac{1}{21}\right)^{n-n_i}\frac{m}{n+m}} \tag{30}$$

Das ODDS-Ratio sei nun

$$o = \frac{p(f|n,n_i,\mathcal{B})}{p(m|n,n_i,\mathcal{B})} = \frac{\frac{1}{6}^{n_i}\left(1-\frac{1}{6}\right)^{n-n_i}}{\frac{1}{21}^{n_i}\left(1-\frac{1}{21}\right)^{n-n_i}}\frac{n}{m}, \tag{31}$$

wobei wir die verschiedenen Terme im Zähler und Nenner bereits gekürzt und vereinfacht haben. Wir müssen nun nur noch die Werte für n,m und n_i einsetzen und das ODDS-Ratio ausrechnen. Ein ODDS-Ratio größer 1 spricht für einen fairen Würfel während ein ODDS-Ratio kleiner 1 für einen manipulierten spricht.

Kommentar: Hier sollte man sich noch einmal in Erinnerung rufen, was die einzelnen Beiträge bedeuten (i.e. ODDS-Ratio, Bayes-Faktor, Prior-ODDS).

b) Wir können die Posterior-Wahrscheinlichkeit für einen fairen Würfel aus **(a)** nehmen. Wir müssen dann nur noch die Werte für n,m und n_i einsetzen. Bei einer Posterior-Wahrscheinlichkeit größer 0,5 würde diese für einen fairen Würfel sprechen andernfalls würde diese eher für einen manipulierten Würfel sprechen.

c) Hier gibt es im Wesentlichen zwei verschiedene Zugänge.

Vergleiche fairen und manipulierten Würfel miteinander: Für diesen Zugang ist eine Augenzahl vollkommen ausreichend. Am besten wäre jene Augenzahl, deren Wahrscheinlichkeit am stärksten von $\frac{1}{6}$ abweicht. In diesem Fall wäre dies entweder die Augenzahl *1* **oder** die Augenzahl *6*.

Vergleiche zwei verschiedene Augenzahlen desselben Würfels miteinander: In diesem Fall müssen wir zwei Augenzahlen auswählen. Die geeignetsten wären jene die sich in der Wahrscheinlichkeit am stärksten von einander unterscheiden. In diesem Fall wären dies die Augenzahlen *1* **und** *6*.

Kommentar: Allgemein sollte man sich zunächst überlegen, was man zeigen möchte. In unserem Fall wollen wir einen fairen von einen manipulierten Würfel unterscheiden. Dazu sollten wir logischerweise jene Eigenschaften heranziehen, dies ich am stärksten unterscheiden. Das unsere Wahl auf die obengenannten Augenzahlen fällt, lässt sich leicht in Zahlen fassen. Angenommen wir würfeln 60 mal. Bei einem fairen Würfel würden wir erwarten, dass alle Augenzahlen etwa zehn mal vorkommen. Bei einem manipulierten Würfel würden wir erwarten, dass die Augenzahl *1* 3 mal $(60 \cdot \frac{1}{21})$ und die Augenzahl *6* 18 mal $(60 \cdot \frac{6}{21})$ vorkommen. Die Augenzahlen *3* und *4* sollten etwa 9 bzw. 12 mal vorkommen. WICHTIG: 60 Würfe ist eine vergleichsweise kleine Zahl. Man müsste hier in Realität eventuell eine höhere Zahl an Würfen betrachten.

d) Eine einfache Möglichkeit wäre das Berechnen des Mittelwertes und der Standardabweichung eines fairen und eines manipulierten Würfels. Eine noch einfachere Möglichkeit zur Abschätzung wäre folgende:

Bei einem fairen Würfel kommen alle Augenzahlen mit gleicher Wahrscheinlichkeit vor. Beim manipulierten Würfel kommen die hohen Augenzahlen öfter vor als die niedrigen. Der Mittelwert des manipulierten Würfels müsste somit höher sein als jener des fairen Würfels. Bei dem manipulierten Würfel ist der Mittelwert somit zu höheren Augenzahlen verschoben. Die Augenzahl 6 kommt sehr viel häufiger vor als die Augenzahl 1, ist aber näher am Mittelwert dran. Gleiches gilt für die Paarungen 2 und 5 sowie 3 und 4. Somit sind die Augenzahlen die häufiger Vorkommen näher am Mittelwert dran. Konsequenterweise müsste die Standardabweichung unseres manipulierten geringer sein als jene des fairen Würfels.

Wir können nun betrachten, ob Mittelwert und Standardabweichung jene eines fairen Würfels sind oder sich wie oben beschrieben verhalten, womit wir dann ein Indiz für die Fairness des Würfels hätten. Um hier tatsächlich eine vernünftige Abschätzung liefern zu können, wäre es allerdings ratsam trotzdem Mittelwert und Standardabweichung auch zu berechnen.

Kommentar: Wer möchte kann Mittelwert und Standardabweichung der beiden Würfel berechnen und die obigen Überlegungen überprüfen.