Statistische Mustererkennung WS 2022

V 1.1, Stand: 05-12-2022

Thomas Melzer

thomas.melzer@geo.tuwien.ac.at

Sollten Sie im Skriptum einen Fehler entdecken, so bitte ich Sie, mir dies an obige e-mail Adresse mitzuteilen. An dieser Stelle vielen Dank an alle, die durch ihre Rückmeldungen dazu beigetragen haben, die Qualität dieses Skriptums zu verbessern (im speziellen an Andreas Roncat, Irene Teubner, Michael Melcher und Roland Lindorfer).

Change Log

V 1.05: 31.10.2022: Stetige Verteilungen: Student's t-Verteilung wurde hinzugefügt, Pareto-Verteilung wurde überarbeitet.

V 1.06: 13.11.2022: Abb. 19, erste Ableitung der ll (grüne Kurve) war falsch (mit positiver Ableitung) dargestellt; korrigiert.

V 1.1: 5.12.2022: Der Abschnitt über Entscheidungstheorie – insbesondere die Herleitung und Diskussion des *conditional risk* – wurde überarbeitet und ergänzt.

Literaturhinweise

- C. Bishop, **Pattern Recognition and Machine Learning**, Springer, 2006
  Gute und ausführliche Einführung in den modernen, "bayesianisch" geprägten Zugang zur Mustererkennung, einschließlich Parameterschätzung, Klassifizierung und Regression. Elektronisch kostenlos verfügbar.

- T. Hastie, R. Tibshirani, J. Friedman, **The Elements of Statistical Learning**, Springer, 2001
  Ein Klassiker. Sehr gute und ausführliche Behandlung linearer und kernelbasierter Verfahren, Bayes-Inferenz wird allerdings kaum behandelt. Elektronisch kostenlos verfügbar.

- E.T. Jaynes. **Probability Theory. The Logic of Science.**, Cambridge, 2003
  Herleitung und Rechtfertigung der Wahrscheinlichkeitstheorie als Erweiterung der Aussagen-Logik; eine mit Leidenschaft und ohne Selbstzweifel verfasste Kampfschrift für den Bayesianismus. Streicht äußerst interessante und wichtige Bezüge zwischen Wahrscheinlichkeitstheorie, Wissenschaftstheorie und Anwendungen (Jaynes ist Physiker) heraus. Äußerst empfehlenswert, setzt jedoch gute Mathematik-Grundkenntnisse und eine gewisse Ausdauer voraus.

- Gerd Gigerenzer. **Calculated Risks**, 2002, Übersetzung: **Das Einmaleins der Skepsis**, BTV, 2004.
  Gigerenzer ist ein deutscher Psychologe und profilierter Kämpfer gegen die "Rezeptstatistik". Bekannt wurde er durch seine fundierte Kritik an Brustkrebs-Screenings, ein Thema, welches auch in diesem Buch behandelt wird. Das Buch bietet zwar eine geraffte, populärwissenschaftliche Einführung in die Grundlagen statistisch basierter Entscheidungsfindung (insbesondere die Bayes-Regel); sein zentrales Thema ist jedoch die psychologische und politische Dimension statistischer Argumentation, insbesondere die weitverbreiteten Fehlinterpretationen bzw. falschen Anwendungen statistischer Tests, und die Frage, welche Ursachen diese haben und wie sie sich vermeiden lassen.

- **Epidemiology An Introduction** von Kenneth J. Rothman.
  Eine gute, kompakte Einführung in die Epidemiologie. Konzentriert sich auf die relevanten Konzepte (Kausalität, Arten von Studien, *Confounding*, Stratifizierung, epidemiologische Effektgrößen), mathematisch nicht sehr anspruchsvoll.

- **Der Hund, der Eier legt** von Dubben und Beck-Bornholdt und **Lügen mit Zahlen** von Bosbach und Korff sind zwei weitere äußerst empfehlenswerte populärwissenschaftliche Titel, die sich mit fehlerhaftem Gebrauch bzw. dem Missbrauch der Statistik in der Praxis auseinandersetzten, ersterer eher im wissenschaftlichen (insbesondere im Bereich *life sciences*), zweiterer eher im politischen Bereich.

- **The Black Swan** von Nassim Nicholas Taleb ist bedingt empfehlenswert. Das Thema ist die prinzipielle Nichtvorhersagbarkeit nichttrivialer Phänomene mittels Statistik. Der Autor hat das Talent, in diesem Kontext relevante, aber selten formulierte logische und statistische Grundlagen prägnant und unterhaltsam zu präsentieren. Auch hat er als ehemaliger *trader* praktische Einsichten in die Funktionsweise der Finanzwelt, die er sich dem Leser auf eloquent-zynische Art mitzuteilen nicht scheut. Allerdings ist er doch sehr von seiner eigenen Brillanz berauscht, was die Lektüre mitunter zu einer Herausforderung werden lassen kann.

- **Thinking Fast and Slow** von Daniel Kahneman befasst sich mit den evolutionär begründeten Eigenheiten des menschlichen Risikobeurteilungs- und Entscheidungsfindungssystems, und den resultierenden Verzerrungen, Irrtümern und Möglichkeiten zu systematischer Manipulation. Fazit: Menschen sind schlechte Entscheider und sind nur schwer in der Lage, Wahrscheinlichkeiten richtig einzuschätzen.

- Ian Hacking, **An Introduction to Probability and Inductive Logic**, Cambdridge University Press, 2001
  Hacking ist ein bekannter Philosoph und Wissenschaftstheoretiker, der sich seit vielen Jahrzehnten mit dem Thema *Wahrscheinlichkeit* auseinandersetzt. Dies ist eines seiner neueren und im Vergleich zu seinen philosophischen Essays leicht zugänglichen Werke, welches einen hervorragenden Überblick über die grundlegenden Ideen und Konzepte und die wichtigsten Schulen bzw. Protagonisten der Wahrscheinlichkeitstheorie bietet.

- **Entstehung und Entwicklung einer wissenschaftlichen Tatsache** von Ludvik Fleck ist ein stiller Klassiker der Wissenschaftstheorie aus den 1930ern. Fleck widerspricht der Auffassung von Wissenschaft als neutralem Erkenntisprozess bzw. Anhäufung von ewigen, unwandelbaren Gewissheiten, sondern beschreibt diese als sozialen Prozess, einen in einem *Denkkollektiv* vorherrschenden *Denkstil*, der Vorläufer des Kuhnschen Paradigmas. Ein wichtiges und hochaktuelles Buch, welches jeder, der auch nur entfernt mit wissenschaftstheoretischen Fragen zu hat, gelesen haben sollte.

Was ist Statistische Mustererkennung (SME)?

- Aufgabe: Computergestützte **Klassifizierung** von Objekten bzw. Prozessen anhand quantitativer **Merkmale** (*features*).

  Beispiele für Klassen:

  – Gesichter (Zugangsberechtigung)
  – Buchstaben (OCR)
  – Herztätigkeit eines Patienten (gesund vs. an Vorhofflimmern erkrankt)
  – Bewegungslinien (Trajektorien) von Sportlern (Effizienzbeurteilung).

In der Praxis wird nicht auf den interessierenden Objekten bzw. Prozessen selbst, sondern auf durch Messung erhaltenen Signalen gearbeitet (Bild eines Gesichts, eingescannter Buchstabe, EKG, Ausgabe eines Personentrackers). Je nach Autor werden diese Signale oder aber Klassen von Signalen[1] als **Muster** *(pattern)* bezeichnet.

Wir werden im ersteren Fall, d.h. wenn mit Muster die einem spezifischen Objekt zugeordneten Messungen gemeint sind, von **Musterinstanzen** sprechen:

Welt (distale Objekte) $\rightarrow$ Messung $\rightarrow$ Signal $\rightarrow$ Digitalisierung $\rightarrow$ Computersystem (proximale Musterinstanz).

[1] Z.B. der typische Verlauf der EKG-Kurve im Fall von Vorhofflimmern

- Muster werden durch **Merkmale** beschrieben. Dies können unmittelbare Messgrößen oder durch Transformation aus diesen erhaltene Größen sein (Merkmalsextraktion). Personen könnten z.B. durch Merkmale wie Alter und Körpergröße beschrieben werden, Audiosignale durch ihre Fourier-Koeffizienten usw. Der konkrete Wert, den ein Merkmal für eine gegebene Musterinstanz annimmt, wird als **Merkmalsausprägung** (auch: **Realisierung**, eng:*realisation*) bezeichnet (Claudia ist 17 Jahre alt und 1,60m groß).

- In der SME werden Merkmale als stetige oder diskrete **Zufallsvariablen** aufgefasst, welche in **Merkmalsvektoren** (*feature vectors*) zusammengefasst werden. Einer konkreten Merkmalsausprägung entspricht somit eine Ausprägung (Messung) des korrespondierenden Merkmalsvektors (z.B. $\mathbf{x} = (17, 1.60)^T$).

- Die in der SME verwendeten Merkmale haben i.a. kardinales **Skalenniveau** (quantitative Daten), d.h. es können Aussagen über die

  1. relative Häufigkeit (es gibt mehr Männer als Frauen - Nominalskala)
  2. relative Ordnung (Claudia ist jünger als Paul - Ordinalskala)
  3. Ähnlichkeit (Claudia ist 3 Monate jünger als Paul - Intervallskala)
     sowie möglicherweise
  4. das Verhältnis (Egon ist doppelt so alt wie Claudia - Verhältnisskala, absoluter Nullpunkt erforderlich)

  von Merkmalsausprägungen gemacht werden.

  Die Stärke der Skala nimmt in obiger Liste nach unten zu, d.h. jedes Skalenniveau i impliziert alle Skalenniveaus $< i$. Von einer Kardinal-Skala spricht man, wenn die Merkmale zumindest auf Intervall-Skalen-Niveau vorliegen, also Abstände (Metriken) sinnvoll interpretiert werden können. Dies ist im technisch-naturwissenschaftlichen Bereich typischerweise der Fall.

- **Merkmalsextraktion** *(feature extraction)*

  Ein Merkmal kann als Abbildung φ aus dem Muster-Raum (*pattern space*) P in den **Merkmalsraum** (*feature space*) $\mathcal{X}$ verstanden werden:

$$\varphi : P \rightarrow \mathcal{X} \tag{1}$$

  Die Merkmalsausprägungen sind dann gerade die Elemente von $\mathcal{X}$, welche durch **Merkmalsberechnung** (*feature computation*) als Bilder der Elemente von P erhalten werden.

  Der Begriff der **Merkmalsextraktion** (*feature extraction*) wird in der Literatur nicht einheitlich verwendet. Im engeren Sinn versteht man darunter die Auswahl oder Bestimmung der Abbildungsfunktion φ . Im weiteren Sinn wird unter Merkmalsextraktion auch die Merkmalsberechnung verstanden (insbesondere im Bereich Bildverarbeitung/Computer Vision).

- Bei der **Merkmalsselektion** (*feature selection*) geht es - im Unterschied zur Merkmalsextraktion - darum, aus einer gegebenen Menge von Merkmalen $\{\varphi_1, \ldots, \varphi_N\}$, eine kleine, bzg. der gegebenen Klassifizierungsaufgabe maximal "informative" Untermenge auszuwählen.

Verwandte Gebiete

- **Nichtmetrische Methoden der Mustererkennung:**
  - Entscheidungsbäume (decision trees): für nominale, qualitative Attribute (z.B. Farbe, Geschmack).
  - Strukturelle und Syntaktische Mustererkennung: Muster werden hierarchisch durch Regelanwendung aus sog. Primitiven erzeugt (*parsing*).

- **Statistik:** Die SME bedient sich statistischer Methoden, beschränkt sich jedoch nicht auf diese. Implementierbarkeit, Performance und numerische Stabilität der Algorithmen spielen in der SME eine wichtige Rolle.

- **Machine Learning:** "Estimating an unknown dependency or structure of a system using a limited number of observations." (Cherkassky)
  - Regression
  - Klassifizierung
  - Dichteschätzung (density estimation)
  - Clustering/Vektorquantisierung

Grundbegriffe

- **Regression und Klassifizierung** gehören zur Kategorie der überwachten (supervised) Verfahren. Hier wird versucht, anhand von gegebenen Paaren von Merkmalsausprägungen $\mathbf{x}_i$ und zugeordneten abhängigen Werten y_i den funktionalen Zusammenhang zwischen den Größen $y = f(\mathbf{x}, \theta)$ zu bestimmen.

  Während die Merkmale und die abhängige Größe y beobachtbar sind, kann die Funktion f (das **Modell**) auch von weiteren, nicht beobachtbaren Größen θ abhängen, den sogenannten Parametern. Beispiel lineares Modell: $y = ax + b$, mit $\theta = (a, b)$.

Je nach Disziplin und Kontext sind verschiedene Bezeichnungen für die Größen $(\mathbf{x}, y)$ gebräuchlich, z.B

- unabhängige Variable vs. abhängige Variable (Mathematik)
- Input-Variable vs. Output-Variable
- Merkmals-Variable vs. Target-Variable (Machine Learning, Neuronale Netze)
- *explanatory* / *predictor variable* vs. *response variable* (Statistik)

- Die **Entscheidungsfunktion** (*decision rule*) $\alpha(x) : x \mapsto a$ assoziiert mit jeder Merkmalsausprägung $\mathbf{x}$ eine bestimmte Aktion $a \in \mathcal{A}$. Bei der **Klassierung** besteht die Aktion in der Zuweisung eines Klassenlabels $l_j \in \mathcal{C}$, d.h. $\mathcal{A} = \mathcal{C}$. Die Entscheidungsfunktion bezeichnet man in diesem Fall als **Klassifikator** (*classifier*).
  - $\alpha()$ partitioniert den Merkmalsraum vollständig in $|\mathcal{C}|$ disjunkte **Entscheidungs-Regionen** (*decision regions*) $\mathcal{R}_i$, wobei

$$\mathcal{R}_i = \{x : \alpha(x) = l_i\}. \tag{2}$$

  - Die Grenze zwischen jeweils zwei Entscheidungsregionen wird als **Entscheidungsgrenze** (*decision boundary*) bezeichnet.
  - Die Entscheidungsregionen müssen nicht zusammenhängend sein.

- Lässt man als Argument von α auch Mengen von Beobachtungen zu, und besteht die Aufgabe der Entscheidungsfunktion darin, den den Beobachtungen zugrundeliegenden Parametervektor θ zu bestimmen ($\mathcal{A} = \mathbf{\Theta}$), so bezeichnet man die Entscheidungsfunktion $\alpha()$ als **Schätzer** (*estimator*).

- Die Kosten einer Entscheidung bezeichnet man als **loss** $L(a, \alpha(\mathbf{x}))$, wobei a für den wahren Wert steht.

- **Klassifikation**
  Systematische Einteilung von Objekten (oder Prozessen) in Klassen anhand vorgegebener Kriterien (z.B. morphologisch vs. phylogenetisch in der Biologie: Taxonomie), z.B.
  Klasse (Säugetier) - Ordnung (Raubtier) - Familie (Katze) - Spezies

- **Klassifizierung**
  Überbegriff. Im engeren Sinne:

  – Erstellen einer Klassifikation (unüberwachtes Lernen)
  – Rekonstruieren einer bestehenden Klassifikation (überwachtes Lernen)

- **Klassierung**
  Zuweisen eines Objekts an eine Klasse einer gegebenen Klassifikation durch Anwenden eines Klassikationsschemas (z.B. Nachschlagen im Pilzbestimmungsbuch). Techn. Implementierung := Klassifikator (*classifier*)

Merkmalsbasierte Klassifizierung: ein Beispiel

In einer Fischfabrik soll automatisch anhand eines Grauwertbilds zwischen Lachsen und Brassen unterschieden werden. Das System muss also im laufenden Betrieb pro Fisch (Muster) folgende Arbeitsschritte durchlaufen:

1. Sensor-Messung (Bildaufnahme)
2. Vorverarbeitung (z.B. Rauschfilterung)
3. Segmentierung, Labeling
4. Merkmalsberechnung (Helligkeit, Länge)
5. Klassierung (Zuweisung an eine gegebene Klasse)
6. Weiterverarbeitung

- Design/Implementierung des Systems

  Wir beschäftigen uns im folgenden nur mit den Punkten 4 und 5 (Merkmalsauswahl und Auswahl/Training des Klassifikators). Nehmen wir an, dass je 100 Brassen und Lachse vermessen wurden, und uns somit also 200 korrekt mit ihrer Klassenzugehörigkeit "gelabelte" Merkmalsvektoren zur Verfügung stehen (Trainings/Design-Set).

  Die Güte eines Merkmals hängt davon ab, a) wie einfach/schnell es berechnet werden kann und b), wie "diskriminativ" es ist, d.h., wie gut es zwischen den interessierenden Klassen unterscheidet. b) lässt sich z.B. mit Hilfe eines Histogramms visualisieren, in welchem auf der Abszisse die Merkmalsausprägungen und auf der Ordinate die beobachteten Häufigkeiten für jede Merkmalsausprägung (separat für jede Klasse!) aufgetragen werden. Im Idealfall sollten die Histogramme der unterschiedlichen Klassen nicht (oder nur wenig) überlappen.

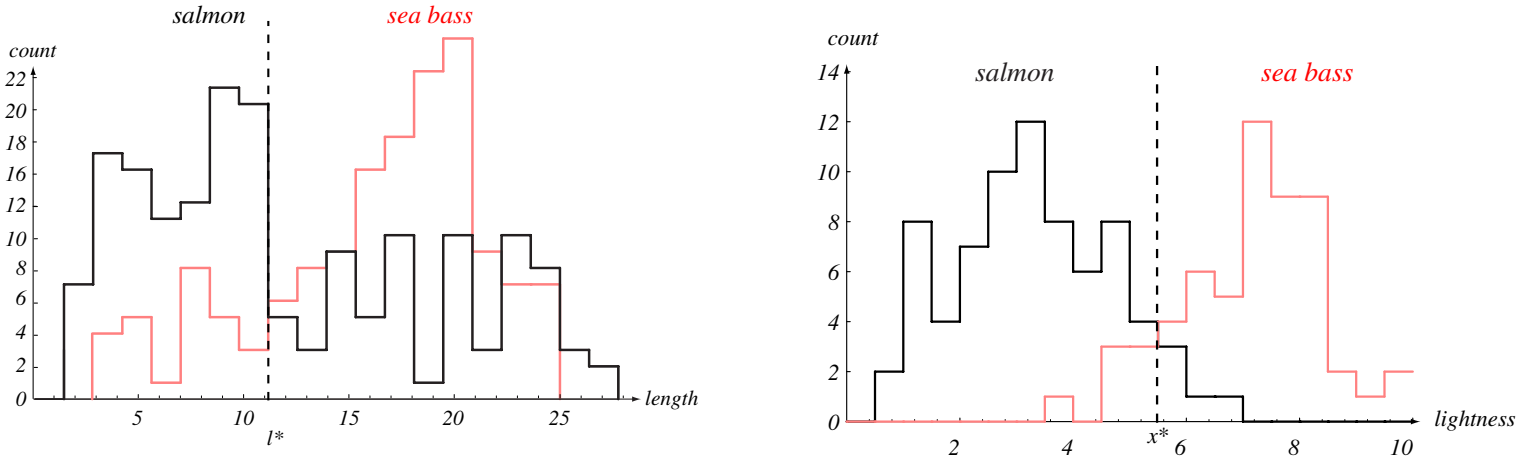


Abbildung 1: Histogramme der Häufigkeiten der gemessenen Längen (links) und Helligkeiten (rechts) für Lachse (schwarz) und Brassen (rot). Obwohl Lachse eher länger als Brassen sind, ist das Merkmal Länge für sich allein nur schlecht geeignet, um zwischen den beiden Fischarten zu unterscheiden. Die klassenspezifischen Ausprägungen des Merkmals Helligkeit überlappen sich zwar in geringerem Maße, jedoch lässt auch dieses Merkmal keine eindeutige, fehlerfreie Klassifizierung bzg. der gegebenen Klassenzugehörigkeiten (class labels) zu.

Nachdem man sich für einen bestimmten Klassifikator (Modell) entschieden hat, muss dieser noch auf den vorhandenen Daten trainiert werden (das Modell wird an die Daten *gefittet*); z.B. könnte die Gerade in Fig. 2 mittels *least squares* (Methode der kleinsten Quadrate) bestimmt werden.

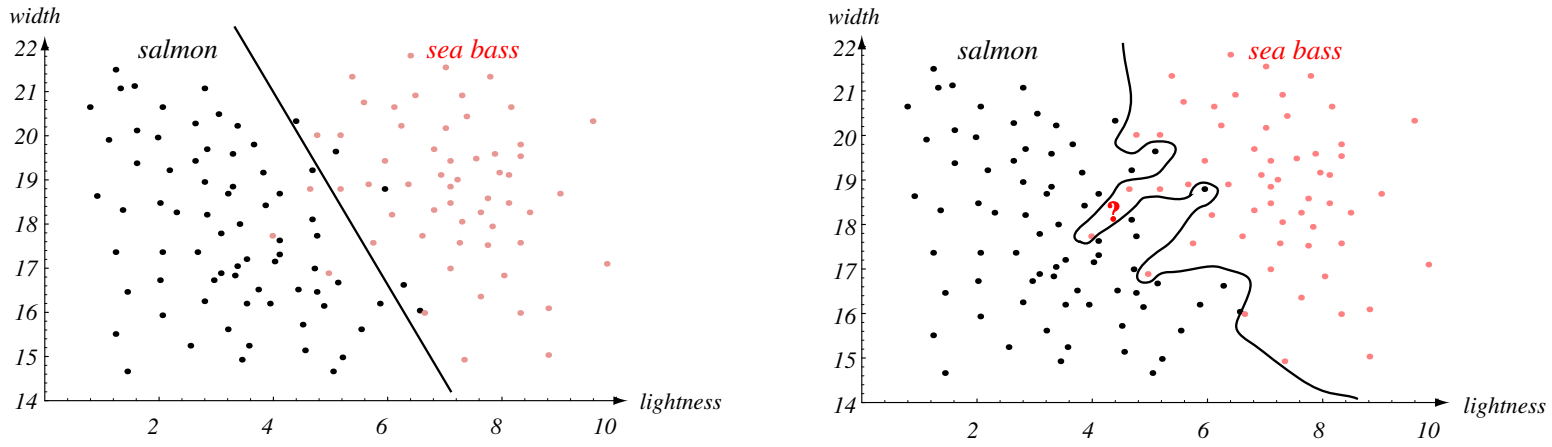


Abbildung 2: Die Kombination mehrerer Merkmale führt oft zu besseren Ergebnissen. Die beiden Klassen sind im zwei-dimensionalen Merkmalsraum (Länge/Helligkeit) bereits recht gut separiert. Das nächste Problem ist die Auswahl eines geeigneten Klassifikators (Modells). Links ist ein Beispiel für einen einfachen, linearen Klassifikator zu sehen: dieser ist offensichtlich nicht in der Lage, die beiden Klassen fehlerfrei zu unterscheiden. Der Klassifikator rechts leistet zwar eine fehlerfreie Klassifikation der Trainingsdaten, jedoch auf Kosten einer komplexen Entscheidungsgrenze.

- Modellkomplexität

  Das Ziel des Designs/Trainings besteht letztendlich nicht darin, die Trainingsdaten, sondern die Gesamtheit aller Muster (bzw. aller möglichen Merkmalsausprägungen) korrekt bzw. mit möglichst geringem "mittleren Fehler" zu klassifizieren; man spricht in diesem Zusammenhang auch von der **Generalisierungsfähigkeit** des Klassifikators.

  Während zu einfache Modelle zu schlechten Ergebnissen bereits auf dem Trainingsset führen, weil sie die den Daten zugrundeliegende Struktur nicht erklären können (*underfitting*), sind zu komplexe Modelle sehr sensitiv bzg. der Auswahl der Trainingsdaten sowie bzg. zufälliger Messfehler (Rauschen) in den Trainingsdaten, was ebenfalls zu schlechter Generalisierungsfähigkeit (hoher Prozentsatz falscher Klassierungen auf nicht im Trainingsset enthaltenen Daten) führen kann (*overfitting*).

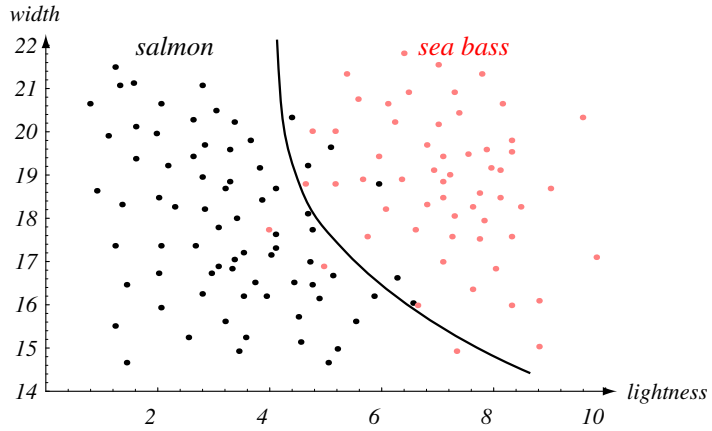


Abbildung 3: Beispiel für einen quadratischen Klassifikator mittlerer Komplexität.

- Statistische Charakterisierung der Modellgüte

  Die Minimierung des “mittleren Fehlers” eines Klassifikators ist möglich, falls die statistische Verteilung (Dichtefunktion) der Merkmale bekannt ist oder zumindest geschätzt werden kann. Dies motiviert den Einsatz statistischer Methoden zum Design optimaler Klassifikatoren (mit minimalem mittleren Fehler) sowie zur Dichteschätzung.

K-Nearest Neighbor Klassifikator (K-NN)

K-NN ist ein klassischer Vertreter sogenannter nicht-parametrischer Verfahren: diese treffen keine Annahme über die parametrische Form der zugrundeliegenden Verteilungen (z.B. Normalverteilung) bzw. gehen nicht von einem (spezifischen) Modell der interessierenden Funktion aus.

Sei im folgenden $\mathcal{S}_{Tr} = \{\mathbf{X}, \mathbf{y}\}$ ein Trainingsset, wobei $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N) \in \mathbf{R}^{d \times N}$ die Spaltenmatrix der Merkmalsvektoren und $\mathbf{y} = (y_1, \ldots, y_N)^T \in \mathbf{R}^N$ den Vektor korrespondierender Klassen-Labels bezeichne ($y_i \in \{l_1, \ldots, l_c\}$).

- Der NN Algorithmus Der NN-1 (kurz NN) Algorithmus weist einem neuen Merkmalsvektor $\mathbf{x}$ einfach das Klassen-Label des ähnlichsten Trainingsvektors zu:

$$\begin{aligned} \alpha(\mathbf{x}) &= y_s, \quad \text{wobei} && (3) \\ s &= \arg\min_i \|\mathbf{x} - \mathbf{x}_i\|, \quad 1 \leq i \leq N && (4) \end{aligned}$$

Hierdurch wird eine sogenannte *Voronoi-Tessellation* des Merkmalsraums induziert; das Einzugsgebiet des i-ten Trainingsvektors

$$P_i = \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{x}_i\| \leq \|\mathbf{x} - \mathbf{x}_j\|, \quad 1 \leq j \leq N\} \qquad (5)$$

wird auch als Voronoi-Polyeder (eng: polyhedron) von $\mathbf{x}_i$ bezeichnet.

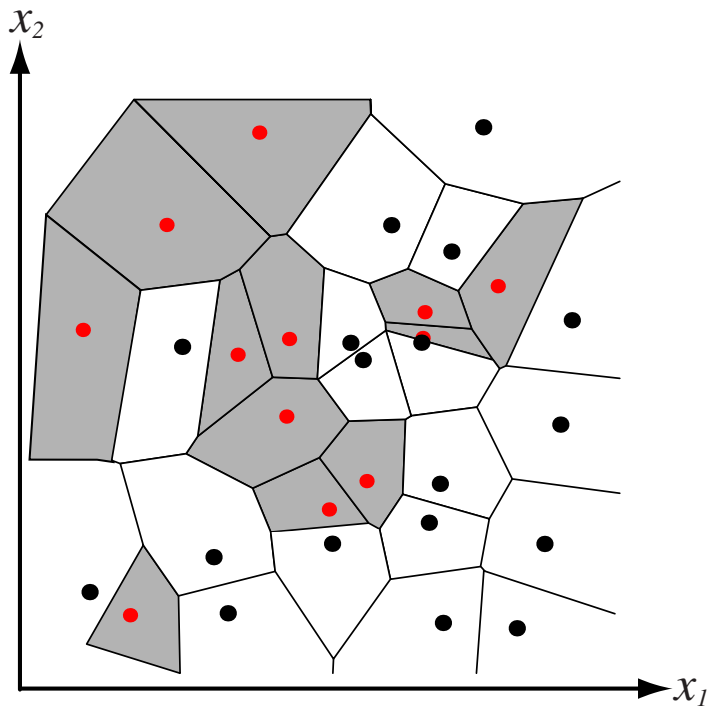


Abbildung 4: Voronoi-Tessellation des $\mathbb{R}^2$ für ein binäres Klassifizierungsproblem. Die Entscheidungsregion der Klasse ω_1 (grau unterlegt dargestellt) ist die Vereinigung aller Voronoi-Polyehedra der zur Klasse gehörigen Trainingsvektoren (rot dargestellt).

- Der K-NN Algorithmus

  Hier werden für einen zu klassifizierenden Merkmalsvektor $\mathbf{x}$ zunächst die K ähnlichsten Trainingsvektoren bestimmt. Gehören k_j dieser Vektoren zur Klasse ω_j (wobei $\sum_{j=1}^{c} k_j = K$ gelten muss), so wird für die Klasse mit dem größten Anteil an "Repräsentanten" entschieden:

$$\alpha(\mathbf{x}) = i, \quad \text{wobei} \tag{6}$$

$$i = \arg\max_j k_j, \quad 1 \leq j \leq c. \tag{7}$$

- Eigenschaften des K-NN Klassifikators

  K-NN erfordert kein Training im eigentlichen Sinn, sondern speichert einfach das gesamte Trainingsset als "Referenz-Menge" ab. Das Verfahren abstrahiert also nicht über das Trainingsset (im Sinne einer kompakten Repräsentation des zugrundeliegenden Datengenerators), sondern lernt es auswendig (*rote learning*). Sowohl Speicher- als auch Laufzeitaufwand wachsen linear mit der Größe des Trainingssets ($O(N)$).

Test- vs. Trainingsfehler, Modellkomplexität

- Ein naheliegendes Maß für die Güte eines Klassifikators $\alpha()$ ist die Anzahl der falsch klassierten Musterinstanzen. Wird diese auf die für das Training verwendeten Daten (*training set*) bezogen, so spricht man vom **Trainingsfehler**. Formal entspricht dies der Verwendung einer *0/1 loss*-Funktion (j bezeichne den Index der wahren Klasse):

$$L(j, \alpha(x)) = 1 - \delta_{\alpha(x),j} = \begin{cases} 1 \text{ falls } \alpha(x) \neq j \\ 0 \text{ falls } \alpha(x) = j. \end{cases} \tag{8}$$

  Sollen auf unterschiedlich großen Datenmengen trainierte Klassifikatoren miteinander verglichen werden, so ist sinnvollerweise der Anteil (und nicht die Anzahl) der falsch klassierten Musterinstanzen heranzuziehen.

- Es interessiert jedoch die Leistungsfähigkeit des trainierten Klassifikators auf der Gesamtheit aller möglichen Mustervektoren. Diese lässt sich abschätzen, indem man den Klassifikator auf eine repräsentative Datenmenge anwendet, die nicht fürs Training verwendet wurde (*test set*); den Anteil der Falschklassierungen bezogen auf diese Menge nennt man entsprechend **Testfehler**.

- Der Trainingsfehler unterschätzt i.a. den Testfehler, und zwar umso mehr, je komplexer und somit flexibler das Modell ist, d.h. je besser es sich an eine gegebene Datenmenge anpassen lässt. Führt eine Erhöhung der **Modellkomplexität** zu einer Verringerung des Trainingsfehlers, aber zu einer Erhöhung des Testfehlers, so liegt *overfitting* vor.

- Wird eine von der Trainingsmenge unabhängige Datenmenge verwendet, um die optimale Komplexität des Klassifikators zu ermitteln ($k = 7$ in Abb. 5), spricht man von einer **Validierungsmenge** *(validation set)*.

  Die Trainings-, Test- und Validierungsmenge sollten idealerweise disjunkt sein. Sollte dies nicht möglich sein, kann *cross validation* verwendet werden, um die optimale Modell-Komplexität auf dem Trainingsset zu bestimmen.

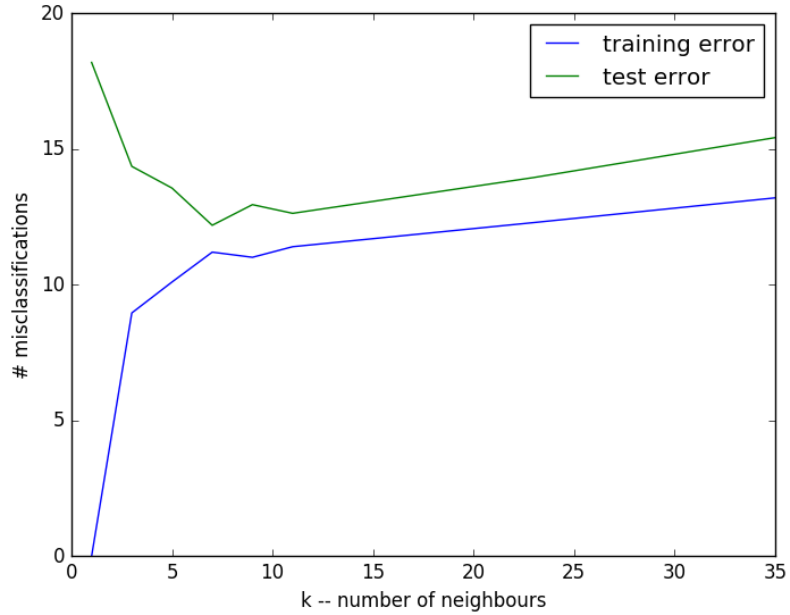


Abbildung 5: Test- vs. Trainingsfehler eines kNN-Klassifikators als Funktion des Parameters k. Für großes k ist das Modell sehr starr, und passt sich kaum den Daten an. Je kleiner k, desto flexibler das Modell: der Trainingsfehler fällt monoton bis auf 0 für k=1. Der Testfehler hingegen fällt zunächst mit fallendem k (bis ca. 7), steigt dann aber wieder.

- Die folgenden zwei Abbildungen illustrieren Test- und Trainingsfehler für Polynomregression. Der ursprüngliche Merkmalsraum umfasst das reelle Intervall $[0 \ldots 1]$, die Zielfunktion ist $\sin(x)$. Das Trainingsset umfasst 10 Punkte, welche in Abb. 6 äquidistant, in Abb. 7 zufällig aus $[0 \ldots 1]$ gewählt wurden. In der unteren Zeile ist jeweils der Grad des gefitteten Polynoms angegeben, das Polynom mit minimalem Testfehler ist mit $*$ gekennzeichnet. Die Fehler sind logarithmisch angegeben.

  In Abb. 7 kommt es außerhalb des von Trainingsdaten abgedeckten Bereichs für höhere Polynomgrade zu exzessivem Überschwingen.

  Wiederholte Versuche mit unterschiedlichen Trainingssets führen für niedere Polynomgrade zu ähnlichen Ergebnissen mit relativ hohem Testfehler, hohe Polynomgrade hingegen zu stark unterschiedlichen, die Details des jeweiligen Trainingsset reflektierenden Ergebnissen (*model bias* vs. *model variance*).

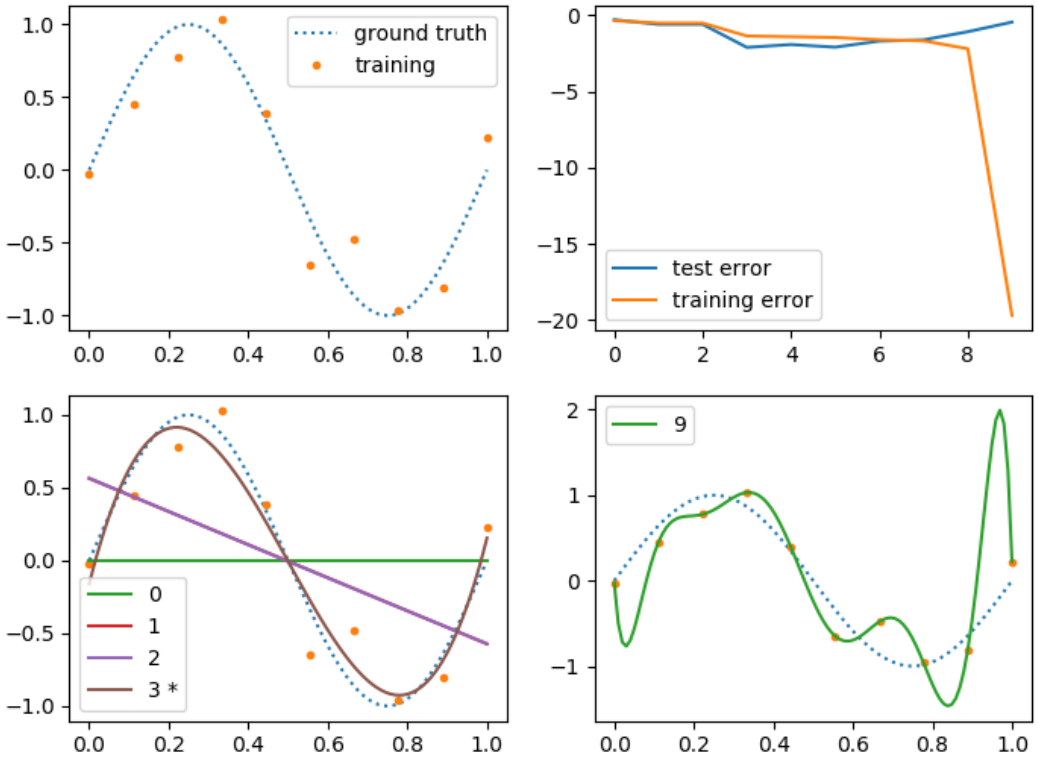


Abbildung 6: Polynomfit einer verrauschten (std 0.3) Sinusperiode mit unterschiedlichen Polynomgraden. Die x-Werte des Trainingssets decken den Merkmalsraum äquidistant ab.

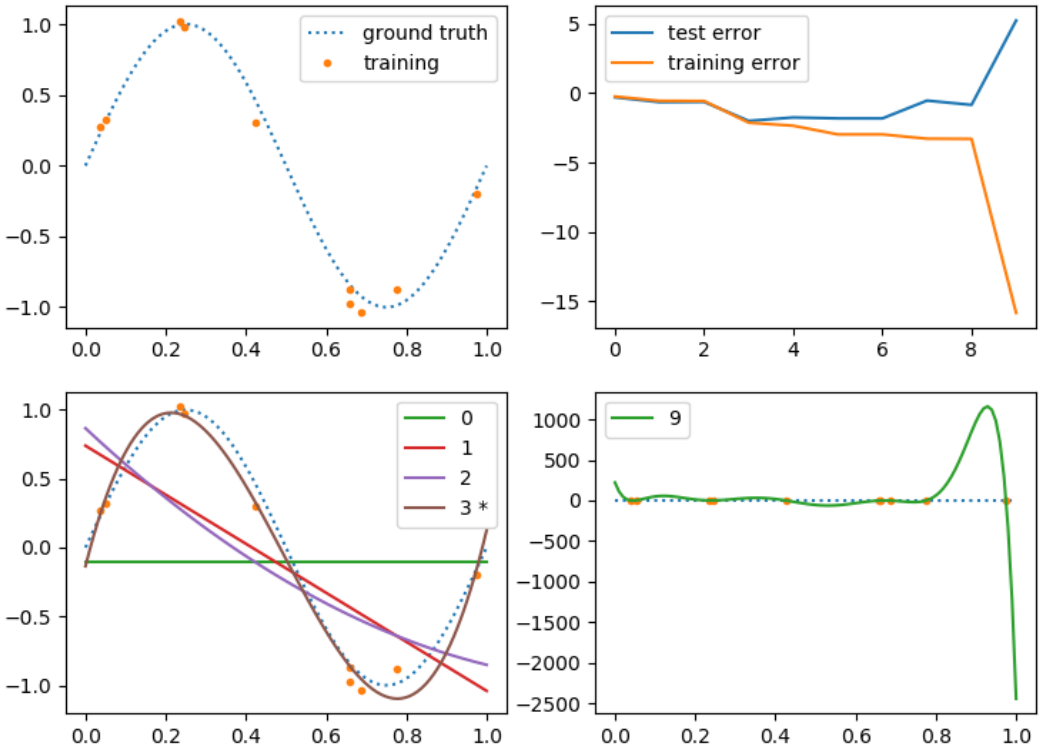


Abbildung 7: Polynomfit einer verrauschten (0.1 std) Sinusperiode mit unterschiedlichen Polynomgraden. Die x-Werte des Trainingssets wurden zufällig dem Bereich $[0 \ldots 1]$ entnommen.

- **Die Qualität eines Schätzers** lässt sich, wie wir später ausführlicher diskutieren werden, statistisch durch den *mean squared error (mse)* beschreiben, den Erwartungswert des L_2-*loss* bzg. der Realisationen des Trainingssets. Der mse zerfällt in die Summe des quadrierten *bias* (Abweichung des Mittels des Schätzers vom wahren Wert) und die Varianz *var* des Schätzers.

$$mse = bias^2 + var \tag{9}$$

  Diese Größen lassen sich sowohl bezüglich der geschätzten Parameter (Regressionskoeffizienten) als auch – wie im vorliegenden Fall – bezüglich der rekonstruierten Funktionen berechnen.

  Ein zu starres Modell, welches die zugrundeliegende Struktur der Daten nicht beschreiben kann, führt zu hohem *bias* und relativ ähnlichen Werten für Test- und Trainingsfehler: *underfitting*.

Ein zu flexibles Modell drückt sich in geringem *bias* und *hoher Varianz* aus und führt i.a. zu deutlich unterschiedlichen Werten für Test- und Trainingsfehler: *overfitting*.

Die optimale Modellkomplexität minimiert den *mse*, also die Summe von *bias* und Varianz.

Abb. 8 illustriert diese Zusammenhänge für unser Polynombeispiel. Man beachte, wie der *bias* mit steigendem Polynomgrad fällt, die Varianz hingegen steigt. Die Summe der beiden wird für den Polynomgrad 3 minimal.

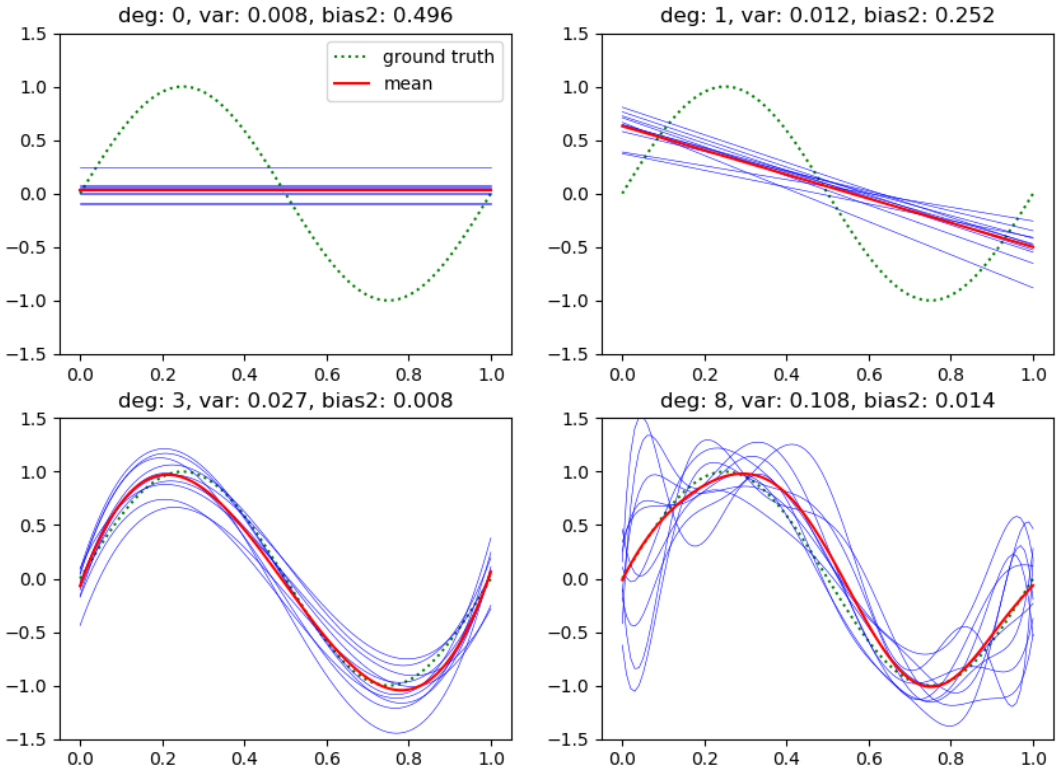


Abbildung 8: Illustration der Auswirkung der Modellkomplexität auf den *bias* und die Varianz des Schätzers. Die Werte wurden für jeweils 10 mit 0.3 std verrauschte Stichproben der Größe 10 berechnet.

- Effekt der Rauschvarianz / SNR

  Stärkeres Rauschen führt dazu, dass die Trainingsdaten weniger brauchbare Information über die gesuchte Funktion enthalten und somit deren Potential, eine komplexe Modellinstanz festzulegen, reduziert. Für das oben diskutierte Polynombeispiel wurde für 3 verschiedene Rausch-Niveaus σ 100-mal die optimale Modellkomplexität für mit σ zufällig verrauschte Trainingsdaten bestimmt; das Optimalitätskriterium war der mse bzg. der bekannten wahren Funktion. Die Ergebnisse sind in der nachstehenden Tabelle wiedergegeben.

Die erste Spalte enthält die Standardabweichung des Rauschens σ, die zweite Spalte die mittlere Anzahl der optimalen Freiheitsgrade FG, und anschließend das Histogramm der optimalen FG, z.b. wurde für das Rausch-Niveau 0.3 in den 100 Versuchen ein optimales Modell mit im Mittel 4.02 FG gefunden, wobei nie ein Modell mit 1, 2 oder 3 FGen gewählt wurde, 52 ein Modell mit 3 FG usw.

σ	$\overline{dof}$	1	2	3	4	5	6	7	8	9	10
0.3	4.02	0	0	0	52	12	25	8	0	2	1
0.6	3.39	0	6	0	65	14	9	5	1		
0.9	2.65	3	22	0	62	9	3	1			

Man sieht, dass mit steigendem Rausch-Niveau tendenziell weniger komplexe Modelle das optimale Ergebnis (minimalen mse) liefern.

k-fache Kreuzvalidierung *(k-fold cross validation)*

Die Trainingsmenge S_{Tr} wird in k möglichst gleich große Teilmengen $S_1, \ldots, S_k$ zerlegt.[2] Bezeichne $U_i = S_{Tr} \setminus S_i, 1 \leq i \leq k$ die Trainingsmenge ohne den i-ten Teil S_i. Sei weiters C die Menge der möglichen Werte des Komplexitätsparameters, sowie $\alpha_c(.)$ eine Modellinstanz des Klassifikators mit Komplexität c.[3] Es wird nun für jede Wahl von c auf den U_i trainiert, während auf den korrespondierenden S_i der Loss bestimmt wird. Der Spezialfall $k = N$ wird aus naheliegenden Gründen als *leave-one-out* bezeichnet.

Eine mögliche Zerlegung in Trainings- und Validierungsuntermengen für den Fall $k = 5$ ist in Abb. 9 dargestellt.

[2] Dieses k hat nichts mit jenem in kNN zu tun!

[3] Ebenso hat dieses c nichts mit der Anzahl der Klassen zu tun.

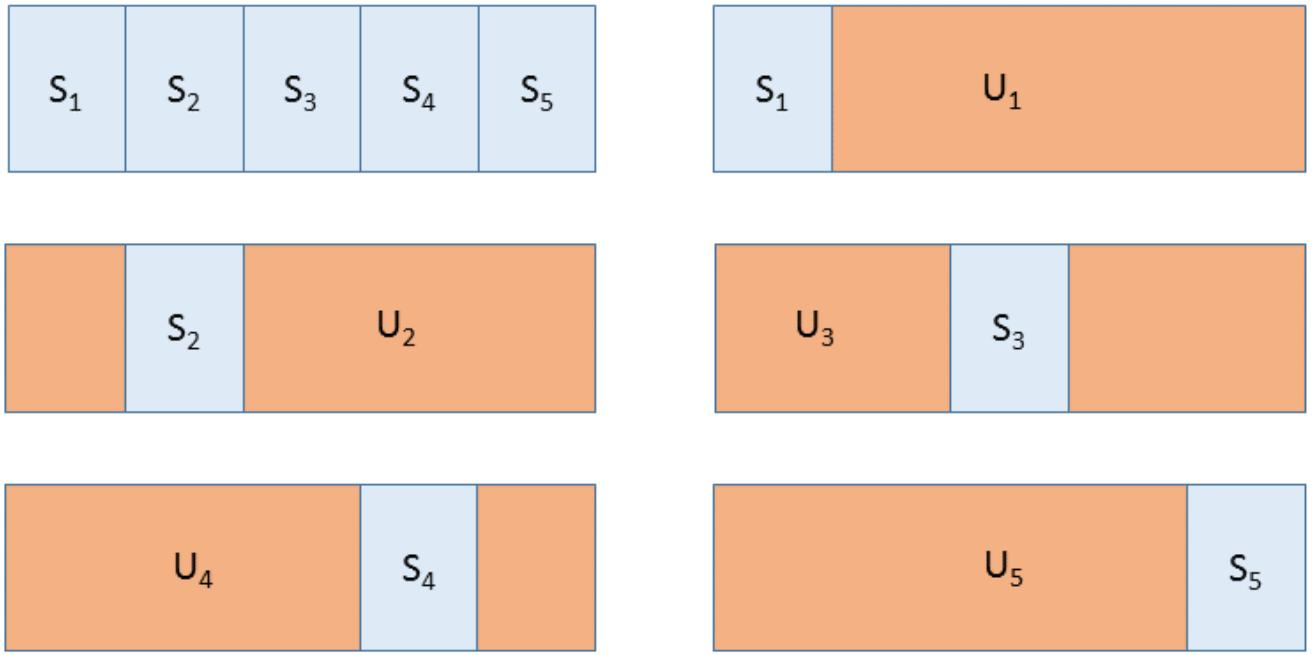


Abbildung 9: Schematische Darstellung der 5-fachen *cross validation*. Oben links: Partitionierung der Trainingsmenge S_{Tr} in 5 gleich große Teile S_i (*folds*). Rest: die 5 möglichen Zerlegungen in Trainingsset U_i mit $|U_i| = 4/5|S_{Tr}|$ und Validierungsset S_i mit $|S_i| = 1/5|S_{Tr}|$.

- for c in C
  - for i = 1 to k do
    - Trainiere $\alpha_{c,i}(.)$ auf U_i
    - Berechne den Validierungsfehler $err_{c,i} = \sum_{j \in S_i} L(y_j, \alpha_{c,i}(\mathbf{x_j}))$ von $\alpha_{c,i}(.)$ auf S_i (z.B. Anzahl der falsch klassierten Musterinstanzen bei 0/1-Loss)
  - Berechne err_c als mittleren Loss, z.B. als Anteil der falsch klassierten Musterinstanzen $\frac{1}{|S_{Tr}|} \sum_{i=1}^{k} err_{c,i}$

Wähle den Komplexitätsparameter mit minimalem *cross validation error* $c = \arg\min err_c$.

Obiges Schema lässt sich genauso auf Regressionsprobleme anwenden; statt dem 0/1-Loss wählt man dazu z.B. den quadratischen Loss $L(y, \alpha(x)) = (y - \alpha(x))^2$.

- **Analytische Kreuzvalidierungsverfahren** versuchen, den Validierungsfehler als Funktion des Trainingsfehlers, der Größe des Trainingssets und des Rausch-Niveaus zu bestimmen. Beispielhaft sehr hier die CP-Statistik für lineare Modelle erwähnt

$$CP = err_c + 2\ c\ \sigma^2/N, \tag{10}$$

wobei err_c den mittleren Trainingsfehler, c die Anzahl der Freiheitsgrade des Modells (Koeffizienten), N die Größe des Trainingssets und σ^2 die Rauschvarianz bezeichnen.

Akaike Information Criterion AIC sowie *Bayesian Information Criterion* sind weitere bekannte, allgemeinere Kriterien (wobei das CP-Kriterium einen Spezialfall des AIC darstellt).

Grundbegriffe der Wahrscheinlichkeitstheorie

- Ein **Elementarereignis** ist das Eintreten bzw. das Beobachten einer (evt. mehrdimensionalen) Merkmalsausprägung an einem Element (Merkmalsträger) einer Grundgesamtheit (Population), z.B. die geworfene Augenzahl beim Würfeln, das Geschlecht einer Person etc.

- Die Menge aller Elementarereignisse wird als **Stichprobenraum** $\Omega = \{e_1, ..., e_n\}$ bezeichnet, für die beiden obigen Beispiele wäre dies $\Omega = \{\boxdot, ..., \boxplus\}$, bzw. $\Omega = \{"maennlich", "weiblich"\}$.

  Der Stichprobenraum ist das wahrscheinlichkeitstheoretische Pendant zum Merkmalsraum; ein Elementarereignis entspricht einer Realisierung eines (distalen) Merkmals.

- **Ereignisse** sind Mengen von Elementarereignissen, z.B ist das Ereignis "Augenzahl gerade"durch $\{\boxdot, \boxdot, \boxdot\}$ gegeben. Die Menge aller interessierenden Ereignisse wird als **Ereignisraum** Σ bezeichnet.

- Das Herbeiführen und Beobachten eines Elementarereignisses unter kontrollierten Bedingungen, aber mit nicht feststehendem (zufälligem) Ergebnis, bezeichnet man als **Zufallsexperiment**, z.B. Münzwurf. Es wird i.a. gefordert, dass ein Zufallsexperiment unendlich oft wiederholbar sein soll. Das im Rahmen des Zufallsexperiments beobachtete Elementarereignis bezeichnet man auch als dessen **Ausfall**.

- Die **Wahrscheinlichkeiten** $P(e_i)$ für Elementarereignisse werden auf unterschiedliche Weise konstruiert:
  - Sie werden empirisch als relative Häufigkeiten von Ereignissen (Anzahl der interessierenden durch Anzahl der möglichen Vorkommnisse) bezüglich einer gegebenen, endlichen Grundgesamtheit ermittelt, z.B. Anteil der 59-jährigen an der Wiener Bevölkerung.

– Sie werden aus Symmetrieüberlegungen als gleichverteilt angenommen, z.B. $P(e_i) = 1/|\Omega| = 1/6$ für den Fall eines 6-seitigen Würfels (**klassischer Wahrscheinlichkeitsbegriff**, **Laplacesches Indifferenzprinzip)**.

– In weiterer Folge fallen auch auf kombinatorischen Argumenten beruhende Wahrscheinlichkeitszuweisungen, insbesondere durch Anwendung der hypergeometrischen Verteilung (z.B. Lotto 6 aus 45), in diese Kategorie.

– Wahrscheinlichkeiten werden als Grenzwert der relativen Häufigkeit des interessierenden Ereignisses bei einer theoretisch unendlichen Anzahl von unabhängigen Wiederholungen eines zugrundeliegenden Zufallsexperiments aufgefasst, z.B $P(⚄) = 0.2$, wenn unter 1000 Würfen 200 ⚄er vorkommen (**frequentistischer Wahrscheinlichkeitsbegriff**).

– Subjektive Wahrscheinlichkeiten lassen sich z.B. durch Auffinden einer als äquivalent empfundenen Wette kalibrieren.

- **Axiomatische Definition der Wahrscheinlichkeit**
  Die Wahrscheinlichkeit $P(A)$ eines Ereignisses A ist durch eine Funktion $P : \Sigma \to \mathbf{R}$ gegeben. Die klassischen *Kolmogorov*-Axiome fordern, dass

  – $P \in [0..1]$
  – $P(\Omega) = 1$
  – $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ für paarweise disjunkte Ereignisse A_i (σ-*Additivität*), was natürlich einfache Additivität impliziert:
    $P(A \cup B) = P(A) + P(B)$ für disjunkte Ereignisse $A, B \subset \Sigma$ mit $A \cap B = \emptyset$

  Es sei jedoch darauf hingewiesen, dass auch andere Axiomatisierungen des Wahrscheinlichkeitsbegriffs möglich sind (siehe im speziellen *Jaynes, Logic of Science*).

- Unter einer **Zufallsvariable** (*random variable)* X versteht man eine Abbildung $X : \Omega \rightarrow \Omega' \subseteq$ **R**. Zufallsvariablen kodieren Elementarereignisse; sie stellen formal den Zusammenhang zwischen Ereignissen bezüglich distaler Objekte (Würfel, Gruppe von Personen) und numerisch kodierten Merkmalsausprägungen dieser Objekte her, z.B. $X(\boxdot) = 3, X(\text{"}weiblich\text{"}) = 0$.

  Des weiteren legt eine Zufallsvariable via

$$P_X(X \in r) = P\left(X^{-1}(r)\right) = P(\{e : X(e) = r\}) \tag{11}$$

  fest, wie sich die Wahrscheinlichkeitsmasse 1 auf Teilmengen (also: Ereignisse) $r \subset$ **R** verteilt: X legt die **Verteilung** des kodierten Merkmals fest.

Man beachte, dass $P_X()$ auf dem Bildbereich $\mathbf{R}$, $P()$ jedoch auf dem ursprünglichen Stichprobenraum definiert ist. Wir werden im folgenden die kürzere Schreibweise $P()$ statt $P_X()$ verwenden, wenn $P()$ aus dem Kontext eindeutig bestimmt ist. Dies ist in Ausdrücken wie $P(X > 3)$ (durch die explizite Angabe der Zufallsvariable X) stets der Fall.

Diskrete Verteilungen

- Eine Verteilung heißt **diskret**, wenn die Anzahl der Elementarereignisse (der möglichen Versuchsausfälle) $|\Omega|$ endlich oder abzählbar ist. Elementarereignisse (Merkmalsausprägungen) werden typischerweise durch ganze Zahlen i kodiert, wobei dieser Zusammenhang formal durch eine **diskrete Zufallsvariable** $X(e_i) = i$ hergestellt wird.

Sei im folgenden $X(\Omega) = \Omega' \subset \mathbb{N}$.

- Beispiele für diskrete Verteilungen
  - Geschlecht eines Probanden
    $X(\text{"}weiblich\text{"}) = 0$, $X(\text{"}maennlich\text{"}) = 1$
    $\Omega' = \{0, 1\}, |\Omega| = 2$
  - Augenzahl beim Würfeln
    $\Omega' = \{1, 2, 3, 4, 5, 6\}, |\Omega| = |\Omega'| = 6$
  - Anzahl der pro Sekunde gemessenen Teilchen eines radioaktiven Zerfallsprozesses
    $\Omega' = \mathbb{N}, |\Omega| = |\Omega'| = \aleph_0$

- Die Wahrscheinlichkeit, dass das Elementarereignis $i \in \Omega'$ eintritt, ist durch die **Wahrscheinlichkeitsfunktion, WF** (*probability mass function, pmf)*

$$p_i = p(i) = P(X = i) \tag{12}$$

gegeben. Obige Schreibweise werden wir bevorzugt verwenden, wenn der Stichprobenraum eine Teilmenge der natürlichen Zahlen darstellt. Eine allgemeinere Formulierung für beliebige (auch überabzählbare) Stichproben- bzw. Merkmalsräume Ω' ist

$$p(x) = p_X(x). \tag{13}$$

(Im Falle stetiger Verteilungen bezeichnen diese allerdings keine Wahrscheinlichkeiten, sondern Wahrscheinlichkeitsdichten - siehe dort).

Die Verteilung ist durch die Gesamtheit aller p_i festgelegt, wobei $p_i \geq 0$ und $\sum_{i \in \Omega'} p_i = 1$ gelten muss.

Beispiel: Bernoulli-Verteilung

Die Bernoulli-Verteilung $B(1, \theta)$ mit Parameter $0 \leq \theta \leq 1$ beschreibt einen Zufallsversuch, der nur zwei mögliche Ausfälle haben kann (z.B. Münzwurf). Für eine Bernoulli-verteilte Zufallsvariable $X \sim B(1, \theta)$ gilt:

$$P(X = 1) = \theta, P(X = 0) = 1 - \theta \tag{14}$$

bzw. in kompakter Darstellung

$$P(X = x) = p(x) = \theta^x \ (1 - \theta)^{(1-x)} \tag{15}$$

- Wir werden häufig p Zufallsvariablen simultan betrachten. Bei nicht mehr als 3 Variablen bezeichnen wir diese meist mit X, Y, Z, bei $p > 3$ mit X_i, und fassen diese in einem p-dimensionaler **Zufallsvektor** (*random vector*)

  $\vec{X} = (X_1, \ldots, X_p)^T = \begin{pmatrix} X_1 \\ \ldots \\ X_p \end{pmatrix}$ zusammen. Die Realisierungen dieser Zufallsvektoren sind die **Merkmalsvektoren** $\mathbf{x} = (x_1, \ldots, x_p)^T = \begin{pmatrix} x_1 \\ \ldots \\ x_p \end{pmatrix}$

  Vektoren (ausgenommen Zufallsvektoren) werden im folgenden mit fetten Kleinbuchstaben bezeichnet und stets als Spaltenvektoren aufgefasst.

Beispiel: Kategorische (Multinoulli) Verteilung

Diese ist eine Verallgemeinerung der Bernoulli-Verteilung auf $k \geq 3$ mögliche Ausfälle, d.h. $X \in \{1, \ldots, k\}$ mit $P(X = i) = \theta_i$, $\theta_1 + \ldots + \theta_k = 1$. Eine naheliegende Wahl für den Stichprobenraum wäre $\Omega' = \{1, \ldots, k\}$. Eine elegante Alternative besteht darin, die i-te mögliche Ausprägung durch den i-ten kanonischen Einheitsvektor darzustellen

$$\mathbf{x}_i = (x_{1,i}, \ldots, x_{k,i})^T = (\delta_{1,i}, \ldots, \delta_{k,i})^T , \tag{16}$$

z.B

$$\mathbf{x}_2 = \mathbf{e}_2 = (0, 1, 0, \ldots, 0)^T \tag{17}$$

(1-aus-k Kodierung). Der Stichprobenraum ist in diesem Fall durch die

k Einheitsvektoren $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ gegeben. Wir haben somit

$$P(X = i) = P(\vec{X} = \mathbf{x}_i) = p(\mathbf{x}_i) = \prod_{j=1}^{k} {\theta_j}^{x_{ji}} = \theta_i. \tag{18}$$

- Die **(kumulative) Verteilungsfunktion** ist durch

$$F_X(k) = P(X \leq k) = \sum_{i=-\infty}^{k} p_i \tag{19}$$

gegeben. Gibt z.B. $P(X = i) = p_i$ die Wahrscheinlichkeit an, beim Würfeln die Augenzahl i zu erhalten, so ist $F_X(4)$ die Wahrscheinlichkeit, eine Augenzahl kleiner oder gleich 4 zu erhalten. $F_X(.)$ ist monoton wachsend, mit $\lim_{k\to\infty} F_X(k) = 1$ und $\lim_{k\to-\infty} F_X(k) = 0$.

- Seien X, Y zwei diskrete Zufallsvariablen, und bezeichne weiters A ein Elementarereignis bzg. X (z.B. $X = i$) und B ein Elementarereignis bzg Y (z.B. $Y = j$). Die Wahrscheinlichkeit, dass die Ereignisse A und B gemeinsam auftreten, ist durch die **Verbundwahrscheinlichkeit** (*joint probability*)

$$p_{ij} = P(A, B) = P(A \cap B) \tag{20}$$

  gegeben.

Randverteilung und Unabhängigkeit

- **Beispiel: Länge und Helligkeit von Lachsen**
  Seien X und Y zwei diskrete Zufallsvariablen, welche die Verteilung der Länge (X) und Helligkeit (Y) von Lachsen beschreiben, wobei wir von $n_X = 4$ Längen- und $n_Y = 2$ Helligkeitsstufen ausgehen.

  Seien weiters $p_i = P(X = i)$ und $p_j = P(Y = j)$ die entsprechenden Wahrscheinlichkeitsfunktionen, wobei wir annehmen, dass beide Helligkeitsstufen gleich wahrscheinlich sind und sich die Längen wie im folgenden Histogramm dargestellt verteilen:

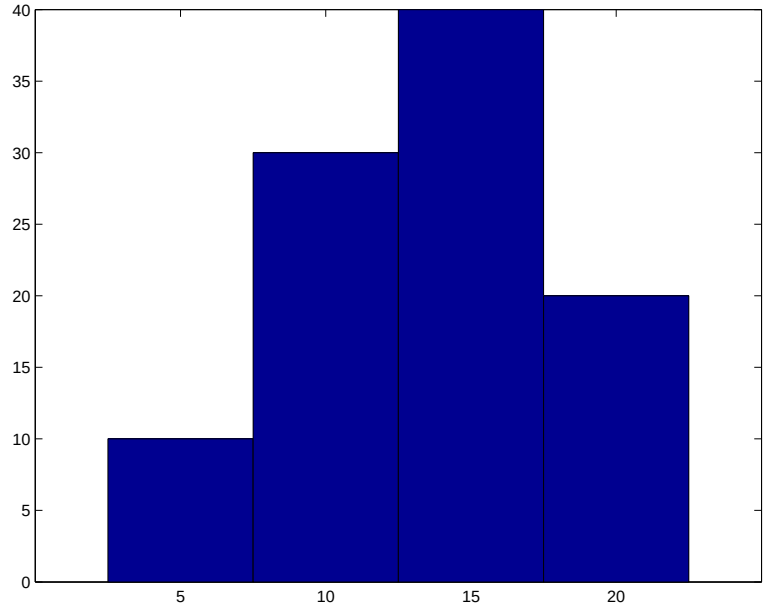


Abbildung 10: Histogramm der Längen (Ordinate = p_i*100).

	1	2	3	4
p_i	0.1	0.3	0.4	0.2
p_j	0.5	0.5		

Tabelle 1: Wahrscheinlichkeitsfunktionen für Länge X und Helligkeit Y.

Y / X	1	2	3	4	$p_{.,j}$
1	0.08	0.12	0.15	0.15	0.5
2	0.02	0.18	0.25	0.05	0.5
$p_{i,.}$	0.1	0.3	0.4	0.2	1

Tabelle 2: Verbundwahrscheinlichkeiten p_{ij}

- Die **Randverteilung** (*marginal distribution*) von X, $p_{i,.}$, erhält man aus p_{ij}, indem man für jede Merkmalsausprägung (jedes Elementarereignis) bzg. X über alle möglichen Merkmalsausprägungen bzg. Y summiert:

$$p_i = p_{i,.} = \sum_{j=1}^{n_Y} p_{ij} \tag{21}$$

  Analog erhält man die Randverteilung von Y, $p_{.,j}$.

Y / X	1	2	3	4	$p_{.,j}$
1	0.05	0.15	0.2	0.1	0.5
2	0.05	0.15	0.2	0.1	0.5
$p_{i,.}$	0.1	0.3	0.4	0.2	1

Tabelle 3: Verbundwahrscheinlichkeiten im Falle der Unabhängigkeit von X, Y.

- Im Falle der **Unabhängigkeit** (*independence*) von X, Y gilt

$$p_{ij} = p_{i,.}\, p_{.,j}, \tag{22}$$

für $1 \leq i \leq n_X, 1 \leq j \leq n_Y$, d.h., die Verbundwahrscheinlichkeiten ergeben sich als das Produkt der korrespondierenden Randverteilungen.

- **Bedingte Wahrscheinlichkeit** (*conditional probability*)
  Bezeichne A das Ereignis $X = i$ und B das Ereignis $Y = j$.

  Die bedingte Wahrscheinlichkeit von A unter B, $P(A|B)$, (d.h. die Wahrscheinlichkeit, dass A eintritt, nachdem B bereits eingetreten ist), ist gegeben durch

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(X=i, Y=j)}{P(Y=j)} = \frac{p_{ij}}{p_{.,j}}. \tag{23}$$

  Sind die bedingten Wahrscheinlichkeiten und die Randverteilungen bekannt, so kann die Verbundwahrscheinlichkeit wie folgt berechnet werden

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A). \tag{24}$$

- Sind X, Y unabhängig, so gilt (für zugeordnete Ereignisse A, B)

$$P(A, B) = P(A|B)P(B) = P(A)P(B) \tag{25}$$

und somit

$$P(A|B) = P(A) \tag{26}$$

- Für festes j erhält man die **bedingte Verteilung** von X unter $Y = j$.

	1	2	3	4	
$P(X = i\|Y = 1)$	0.16	0.24	0.30	0.30	1
$P(X = i\|Y = 2)$	0.04	0.36	0.50	0.10	1

Tabelle 4: Bedingte Verteilungen von X (für Tab. 2).

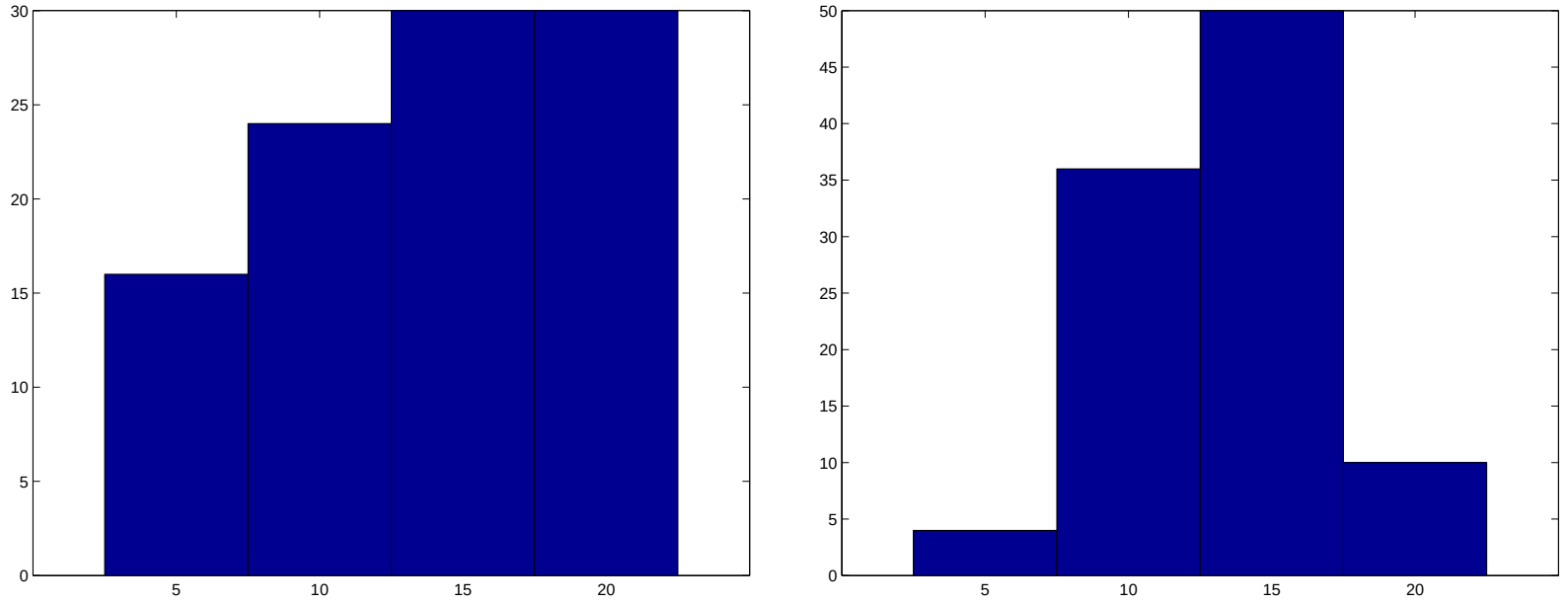


Abbildung 11: Bedingte Verteilungen $P(X = i|Y = 1)$ (links) und $P(X = i|Y = 2)$ (rechts) für die Verbundwahrscheinlichkeiten in Tab. 2.

- **Bedingte Unabhängigkeit**

  Weiß man, dass C eingetreten ist, so erhält man die bedingte Wahrscheinlichkeit von A, gegeben B und C, mit

$$P(A|B,C) = \frac{P(A,B|C)}{P(B|C)}. \tag{27}$$

  Man beachte, dass ',' stärker bindet als '|', d.h. $P(A|B,C) = P(A|(B,C))$.

  Analog zu Gl. 25 – 26 bezeichnet man und A und B als **bedingt unabhängig** gegeben C, wenn

$$P(A,B|C) = P(A|C)\ P(B|C) \tag{28}$$

  bzw.

$$P(A|B,C) = P(A|C). \tag{29}$$

Unabhängige Wiederholungen eines Zufallsversuchs

Einen wichtigen Spezialfall einer Verbundverteilung stellt die N-malige Wiederholung eines Zufallsversuchs dar, wobei die die N Wiederholungen oder *trials* als unabhängig angenommen werden. Im Englischen spricht man auch von *independent, identically distributed (iid)* Variablen. Der zugehörige Stichprobenraum ergibt sich formal als N-faches kartesisches Produkt des ursprünglichen Stichprobenraums mit sich selbst.

Meist nimmt man an, dass der Ausfall der Versuchswiederholungen von einem oder mehreren Parametern abhängt. In diesem Fall werden die Wiederholungen als bedingt unabhängig bzg. des Parameters angenommen.

Beispiel: Bernoulli-Verteilung

Dies sei am Beispiel einer Bernoulli-Verteilung erörtert, sagen wir, dem N-maligen Werfen einer Münze. Die WF faktorisiert aufgrund der iid-Annahme gemäß Gl. 28 in das Produkt von N identischen Bernoulli-WFs

$$p(x_1, \ldots, x_N|\theta) = \prod_{i=1}^{N} p(x_i|\theta) = \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{(1-x_i)} = \theta^m(1-\theta)^{N-m}, \tag{30}$$

wobei $m = x_1 + \ldots + x_n$ die Anzahl der günstigen Versuchsausfälle bezeichne.

Produkt- und Summenregel der Wahrscheinlichkeitsrechnung

Seien A und B beliebige Ereignisse (also nicht notwendigerweise Elementarereignisse) bezüglich der Zufallsvariablen X resp. Y.

- **Produktregel** (*product rule*)

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A) \tag{31}$$

(siehe Gleichung 24).

- **Summenregel** (*sum rule*)

  Die Wahrscheinlichkeit, dass Ereignis A oder Ereignis B eintritt, ist die Summe der Einzelwahrscheinlichkeiten minus der Wahrscheinlichkeit, dass sowohl A als auch B eintritt:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{32}$$

  Z.B. gilt für die Verbundverteilung in Tabelle 2:
  $P(X = 1 \cup Y = 1) = p_{1,.} + p_{.,1} - p_{1,1} = 0.1 + 0.5 - 0.08 = 0.52.$

- **Erweiterte Summenregel** (auch: *law of total probability*)

  Wenn die Ereignisse $B_1, .., B_n$ eine Partitionierung des Stichprobenraums darstellen, d.h. $\cup_{i=1}^{n} B_i = \Omega$ und $B_i \cap B_j = \emptyset$ für $i \neq j$, dann gilt:

$$\sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i) = P(A) \tag{33}$$

Das Urnenmodell

Gegeben sei eine Urne, welche M rote und $N - M$ weiße Kugeln enthalte. Bezeichne R_i das Ereignis: *die i-te gezogene Kugel ist rot*[4]. Die Wahrscheinlichkeit, zuerst eine rote bzw. weiße Kugel zu ziehen, ist gemäß dem Indifferenzprinzip durch $P(R_1) = M/N$ bzw. $P(W_1) = (N - M)/N$ gegeben.

[4]Korrekter, wenn auch schlechter lesbar wäre die Notation $X_i = R$. R_i ist im folgenden also eine symbolische Konstante, keine Zufallsvariable.

- **Ohne Zurücklegen** (*without replacement*)

  Wie groß ist nun die Wahrscheinlichkeit, nachdem eine rote Kugel gezogen und diese nicht in die Urne zurückgelegt wurde, eine weitere rote $P(R_1, R_2)$ bzw. eine weiße $P(R_1, W_2)$ zu erhalten? Wir bemerken, dass die Grundgesamtheit nach der ersten (aber vor der zweiten) Entnahme nur mehr $M - 1$ rote Kugeln enthält. Eine weitere Anwendung des Indifferenzprinzips in Verbindung mit der Produktregel ergibt:

$$
\begin{aligned}
P(R_1, R_2) &= P(R_2|R_1)P(R_1) = \frac{M-1}{N-1}\frac{M}{N} \\
P(R_1, W_2) &= P(W_2|R_1)P(R_1) = \frac{N-M}{N-1}\frac{M}{N}
\end{aligned}
$$

Es lässt sich leicht zeigen, dass die Wahrscheinlichkeit, eine gegebene Sequenz aus r roten und $w = n-r$ weißen Kugeln zu ziehen, für alle $\begin{pmatrix} n \\ r \end{pmatrix}$ möglichen Sequenzen gleich ist, z.b. $P(R_1, R_2, W_3) = P(R_1, W_2, R_3)$.

Die Wahrscheinlichkeit, bei n Entnahmen ohne Zurücklegen aus einer Urne mit ursprünglich M roten und $N - M$ weißen Kugeln irgendeine Sequenz aus r roten und somit $w = n - r$ weißen Kugeln zu erhalten, ist durch die **hypergeometrische Verteilung** $X \sim H(N, M, n)$ mit

$$P(X = r) = h(r|N, M, n) = \frac{\begin{pmatrix} M \\ r \end{pmatrix} \begin{pmatrix} N - M \\ n - r \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} \tag{34}$$

gegeben.

Ist das Ergebnis der ersten Entnahme nicht bekannt, so berechnet sich die Wahrscheinlichkeit für rot im zweiten Zug aus der erweiterten Summenregel mit

$$P(R_2) = P(R_1, R_2) + P(W_1, R_2) = \frac{M}{N}, \tag{35}$$

ebenso für rot bzw. weiß im n-ten Zug, wenn die Ergebnisse der vorhergehenden Entnahmen nicht bekannt sind.

Die einzelnen Entnahmen sind im Urnenmodell ohne Zurücklegen also nicht unabhängig voneinander $P(R_1, W_2) \neq P(R_1)P(W_2)$, d.h. die Wahrscheinlichkeitsfunktion faktorisiert nicht in das Produkt der Randwahrscheinlichkeiten. Allerdings ist sie invariant gegenüber einer Permutation der Argumente $P(R_1, W_2) = P(W_1, R_2)$; dies wird als *exchangeability* bezeichnet. Unabhängigkeit impliziert *exchangeability*, aber nicht umgekehrt.

- **Mit Zurücklegen** (*with replacement*)

  Hierbei wird angenommen, dass bereits gezogene Kugeln vor der nächsten Entnahme wieder in die Urne zurückgelegt werden, wodurch die Entnahmen statistisch unabhängig werden sollen. Diese Vorstellung ist allerdings nicht unproblematisch. Sauberer ist es, N und M unendlich groß werden zu lassen, während $\theta = \frac{M}{N}$ konstant gehalten wird. θ gibt unter dieser Annahme die Wahrscheinlichkeit an, unabhängig von vorangegangenen Zügen eine rote Kugel zu erhalten (Bernoulli-Verteilung).

Die Wahrscheinlichkeit, bei n **unabhängigen** Entnahmen irgendeine Sequenz aus r roten und somit $w = n - r$ weißen Kugeln zu beobachten, erhält man somit als Limit der hypergeometrischen Verteilung für $N \to \infty$, für $\frac{M}{N} = \theta$ konstant. Dieses Limit ist durch die **Binomialverteilung** $X \sim B(\theta, n)$

$$P(X = r) = B(r|n, \theta) = \left(\begin{array}{c} n \\ r \end{array} \right) \theta^r (1 - \theta)^{n-r} \tag{36}$$

gegeben[5] (ohne Beweis).

[5]Man beachte, dass $B(n, \theta)$ die Verteilung, $B(r|n, \theta)$ hingegen die Wahrscheinlichkeitsfunktion bezeichnet.

Summe zweier unabhängiger diskreter Zufallsvariablen

- Für die Summe $Z = X+Y$ zweier unabhängiger diskreter Zufallsvariablen (z.B. Summe der Augenzahlen beim Würfeln) gilt:

$$\begin{aligned} P(Z=z) &= P(X+Y=z) = \sum_i P(X=i, Y=z-i) \\ &= \sum_i p_{i,z-i} = \sum_i p_{i,.} p_{.,z-i} \end{aligned} \tag{37}$$

  d.h. die Wahrscheinlichkeitsfunktion der Summe erhält man als **Faltung** der Wahrscheinlichkeitsfunktionen der Summanden.

Beispiel: Binomialverteilung

Die Binomialverteilung $Z \sim B(n, \theta)$ gibt die Wahrscheinlichkeit an, dass k von n Wiederholungen positiv ausfallen, wenn ein einzelnes Experiment unabhängig mit Wahrscheinlichkeit θ positiv ausfällt. Die binomialverteilte Zufallsvariable Z erhält man als Summe von n unabhängigen $X_i \sim B(1, \theta)$ Bernoulli-Variablen:

$$Z = \sum_{i=1}^{n} X_i \tag{38}$$

Stetige Verteilungen

- **Stetige Zufallsvariable**
  Elementarereignisse werden durch reelle Zahlen kodiert, z.B. Körpergröße von 1.6m: $X == 1.6$, Ereignisse durch Teilmengen des **R**, z.B. Größe zwischen 1.5m und 1.7m: $X \in [1.5, 1.7]$

- **Verteilungsfunktion, VF** (*cumulative distribution function, cdf*)
  $F_X(x) = P(X \leq x)$ gibt die Wahrscheinlichkeit an, dass eine Beobachtung in das Intervall $(-\infty, x]$ fällt.

- **Dichtefunktion, DF** (*probability density function, pdf*)
  Im Falle einer stetigen Verteilung lässt sich $F_X(x)$ als Integral einer nichtnegativen Dichtefunktion $p_X(x)$ darstellen: $F_X(x) = \int_{-\infty}^{x} p_X(x')dx'$.

- Beziehung zwischen Zufallsvariable und VF bzw. DF

  Eine Zufallsvariable kann als eine spezielle Repräsentation einer Verteilung betrachtet werden. Der Zusammenhang ist durch

$$P(a \leq X \leq b) = \int_a^b p_X(x)dx = F_X(b) - F_X(a) \tag{39}$$

  gegeben.

  Man beachte, dass die diskrete Wahrscheinlichkeitsfunktion p_i tatsächlich die Wahrscheinlichkeit des Eintretens eines Ereignisses angibt, während ihr stetiges Gegenstück, die Dichtefunktion $p_X(x), x \in \mathbf{R}$, nicht als Wahrscheinlichkeit interpretiert werden kann; insbesondere gilt im Falle einer stetigen Zufallsvariablen X

$$P(X = \alpha) = \int_\alpha^\alpha p_X(x)dx = 0 \ (\forall \alpha \in \mathbf{R}). \tag{40}$$

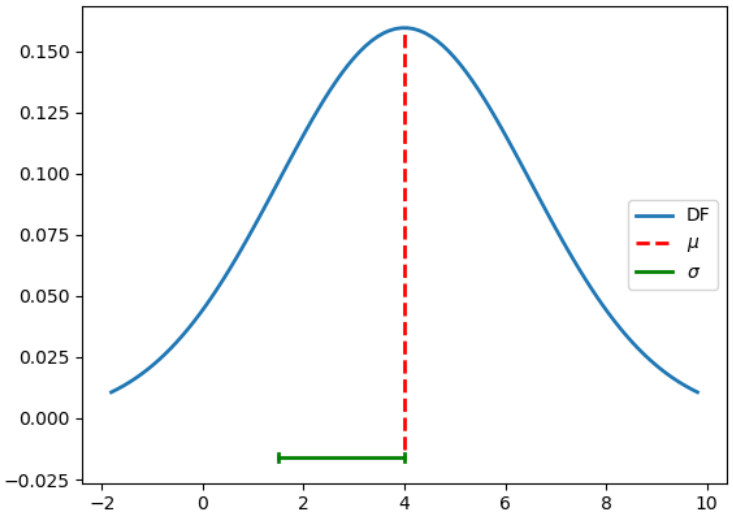
- **Quantile**

  Für das α-Quantil x_α gilt, dass ein α-Anteil der Daten kleiner und ein $(1-\alpha)$-Anteil der Daten größer als x_α ist: $F(x_\alpha) = P(X \leq x_\alpha) = \alpha$. Die Quantilfunktion ist also die Umkehrfunktion der VF.

- Eigenschaften der VF und DF:
  - Die VF $F(x)$ ist monoton wachsend
  - $\lim_{x \to -\infty} F(x) = 0$ und $\lim_{x \to \infty} F(x) = 1$
  - $p(x) \geq 0$ $(\forall x \in \mathbf{R})$
  - Die DF ist die erste Ableitung der VF: $p(x) = dF(x)/dx$.

Beispiel: Normalverteilung $N(\mu, \sigma^2)$

Die Normalverteilung ist eindeutig festgelegt durch Mittelwert μ und Varianz σ^2 bzw. Standardabweichung σ, hier: $\mu = 4$ und $\sigma = 2.5$.



$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Quantile der Standard-Normalverteilung $N(0,1)$

Einige in der Praxis wichtige Quantile:

α	x_α
0.5	0
0.95	1.64
0.975	1.96

D.h, im Intervall $[-2,2]$ liegen ca. 95% der Wahrscheinlichkeitsmasse.

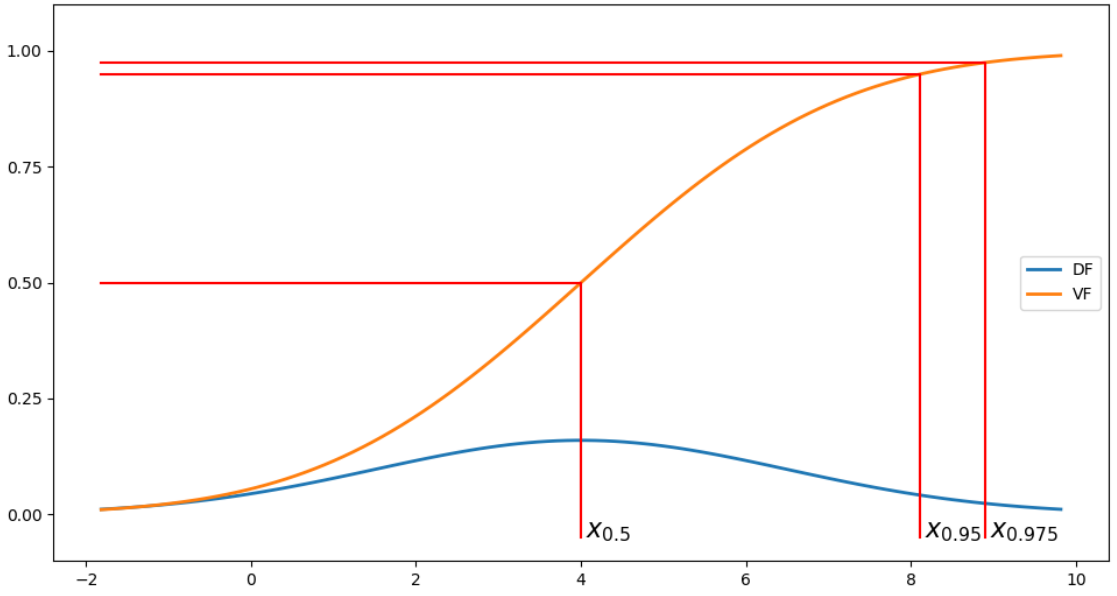


Abbildung 12: Ausgewählte Quantile der Normalverteilung $N(4, 2.5^2)$.

Z-Standardisierung

Eine normalverteilte Zufallsvariable $X \sim N(\mu, \sigma^2)$ lässt sich mittels

$$Z = \frac{X - \mu}{\sigma} \tag{41}$$

in eine standard-normalverteilte Zufallsvariable $Z \sim N(0, 1)$ transformieren.

Die Umkehrung der obigen Beziehung kann verwendet werden, um die Quantile von $N(\mu, \sigma^2)$ aus jenen von $N(0, 1)$ zu berechnen. Z.B. liegen für jede Normalverteilung $N(\mu, \sigma^2)$ (ca.) 95% der Wahrscheinlichkeitsmasse im Intervall $[\mu - 2\sigma, \mu + 2\sigma]$.

Varianz als Skalenparameter: die Tschebyscheff-Ungleichung

Für eine Zufallsvariable X mit Mittelwert μ und Varianz σ^2 gilt für alle $k \in \mathbf{R}$

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2} \tag{42}$$

Für $k = 2\,\sigma$ erhält man

$$P(|X - \mu| \geq 2\,\sigma) \leq \frac{1}{4} \tag{43}$$

Drei Viertel der Verteilungsmasse müssen also innerhalb des Intervalls $[\mu - 2\sigma, \mu + 2\sigma]$ liegen, gleich um welche Verteilung es sich handelt.

Beispiel: Student's t-Verteilung

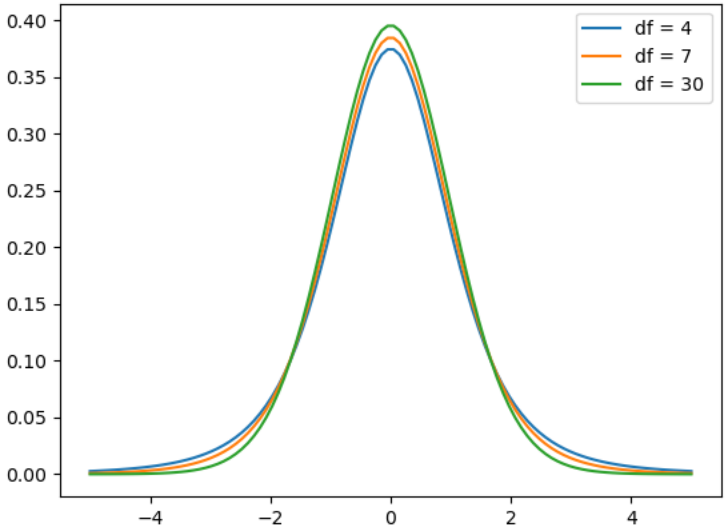


Abbildung 13: Student's t-Verteilung für verschiedene Freiheitsgrade FG; die DF hängt nur vom Parameter FG ab. Für kleine FG ist sie schmalgipfliger und langschwänziger als die Standardnormalverteilung, für große FG nähert sie sich dieser an.

DF	μ	σ^2	A
-	0	1	0.997
4	0	2	0.987
7	0	1.4	0.991
30	0	1.07	0.996

Tabelle 5: Anteil der Beobachtungen im Intervall $[\mu - 3\sigma, \mu + 3\sigma]$ für verschiedene Student's t-Verteilungen. 1ste Zeile: Standardnormalverteilung.

Beispiel: Pareto-Verteilung

$$p(x) = \frac{\alpha \; x_{min}^{\alpha}}{x^{\alpha+1}}, \; x \geq x_{min} \tag{44}$$

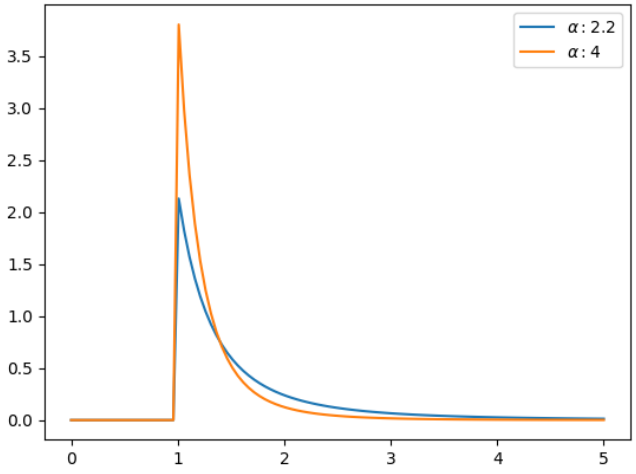


Abbildung 14: Pareto-DF für $x_{min} = 1$ und verschiedene Werte von α.

Die Pareto-Verteilung folgt einem Potenzgesetz (*power law*); sie ist asymmetrisch und deutlich langschwänziger als die Normalverteilung (bzw. eine Exponentialverteilung), extreme Beobachtungen – sog. *black swans* – sind somit deutlich wahrscheinlicher. Dies gilt im speziellen für Parameterwerte $\alpha <= 2$, für welche die Varianz unendlich wird. Für $\alpha \to \infty$ nähert sich die Pareto-DF dem Dirac-Stoß an.

Für $\alpha = 1.16$ erhält man das **Pareto-Prinzip**, wonach nur 20% aller Beobachtungen für 80% der Merkmalssumme verantwortlich sind, z.B. wenn 80% des Einkommens auf nur 20% der Bevölkerung entfallen.

α	μ	σ^2	A
-	0	1	0.997
1.16	7.25	∞	1
2.2	1.83	7.64	0.994
4	1.33	0.22	0.982
10	1.11	0.02	0.981

Tabelle 6: Anteil der Beobachtungen im Intervall $[\mu - 3\sigma, \mu + 3\sigma]$ für verschiedene Pareto-Verteilungen. 1ste Zeile: Standard-Normalverteilung.

- **Zufallsvariable vs. Variable – Wiederholung**

  Zufallsvariablen (*random variable*) beschreiben formal die zugrunde liegende Wahrscheinlichkeitsstruktur (Verteilung) eines Merkmals. Kodieren wir z.B. ein Merkmal durch die Zufallsvariable X, so bedeutet $X \sim N(\mu, \sigma^2)$, dass die Merkmalsausprägungen einer Normalverteilung folgen.

  Zufallsvariablen sind von "kontrollierten" Variablen zu unterscheiden, welche z.B. als Integrationsgrenzen oder als Laufvariablen verwendet werden; insbesondere sind die Argumente x von $F(x)$ und $p(x)$ keine Zufallsvariablen.

  In der Praxis wird diese Unterscheidung jedoch nicht immer getroffen.

- **Transformation von Zufallsvariablen**

  Sei X eine Zufallsvariable mit zugehöriger DF $p_X(x)$. Wir werden im folgenden häufig mit dem Problem konfrontiert sein, die DF $p_Y(y)$ einer gemäß $Y = f(X)$ transformierten Zufallsvariablen Y zu bestimmen. Betrachten wir zunächst den univariaten Fall. Damit die lokale Wahrscheinlichkeitsmasse im Punkt x unter der Transformation $f(.)$ erhalten bleibt, muss eine Änderung des differentiellen Längenelements dx in dy durch eine korrespondierende Änderung der DF kompensiert werden, sodass

$$|dy|\, p_Y(y) = |dx|\, p_X(x) \quad \text{bzw.} \tag{45}$$

$$\begin{aligned} p_Y(y) &= p_X(x)\left|\frac{dx}{dy}\right| \\ &= p_X\left(g(y)\right)|g'(y)| = p_X(g(y))\frac{1}{|f'(g(y))|}, \end{aligned} \tag{46}$$

wobei $g(y) = f^{-1}(y)$ die Inverse von $f(x)$ ist. Die Differentiale bzw. Ableitungen müssen absolut genommen werden, damit obiges Argument auch für fallendes $f(x)$ richtig bleibt.

Für höherdimensionale Merkmalsräume $\mathbf{x}, \mathbf{y} \in \mathbf{R}^p$ ist $|g'(y)|$ als Absolutbetrag der Funktionaldeterminante von $g : \mathbf{R}^p \rightarrow \mathbf{R}^p$ aufzufassen.

Ist $f(.)$ nicht monoton, so muss in Gl. 46 über alle Äste der Umkehrfunktion summiert werden.

Beispiel: $Y = -X$

Die DF von $Y = -X$ ist durch die Spiegelung von $p_X(.)$ gegeben: $p_{-X}(y) = p_X(-y)$.

Beispiel: Skalierte Normalverteilung

Sei $X \sim N(\mu, \sigma^2), f(x) = a\,x$, und somit $g(y) = \frac{y}{a}$ und $g'(y) = \frac{1}{a}$. Wir haben

$$p_y(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\frac{y}{a}-\mu)^2}{2\sigma^2}} . \frac{1}{a} \tag{47}$$

$$= \frac{1}{\sqrt{2\pi}a\sigma} e^{-\frac{(y-a\mu)^2}{2(a\sigma)^2}} \tag{48}$$

In Worten: skalieren wir eine normalverteilte Zufallsvariable $X \sim N(\mu, \sigma^2)$ mit einem Faktor $a \neq 0$, so ist $Y = aX$ ebenfalls normalverteilt mit $Y \sim N\left(a\mu, (a\sigma)^2\right)$.

- **Summe zweier stetiger Zufallsvariablen**

  Die Dichtefunktion der Summe $Z = X + Y$ zweier unabhängiger stetiger Zufallsvariablen mit $p(x, y) = p_X(x)p_Y(y)$ erhält man - analog zum diskreten Fall - als Faltung der Randdichtefunktion von X mit jener von Y:

$$p_Z(z) = \int_{-\infty}^{+\infty} p_X(x)p_Y(z - x)dx \tag{49}$$

  Achtung: Die Summe der Zufallsvariablen darf nicht mit der Summe der Dichtefunktionen $p_X(.) + p_Y(.)$ verwechselt werden!

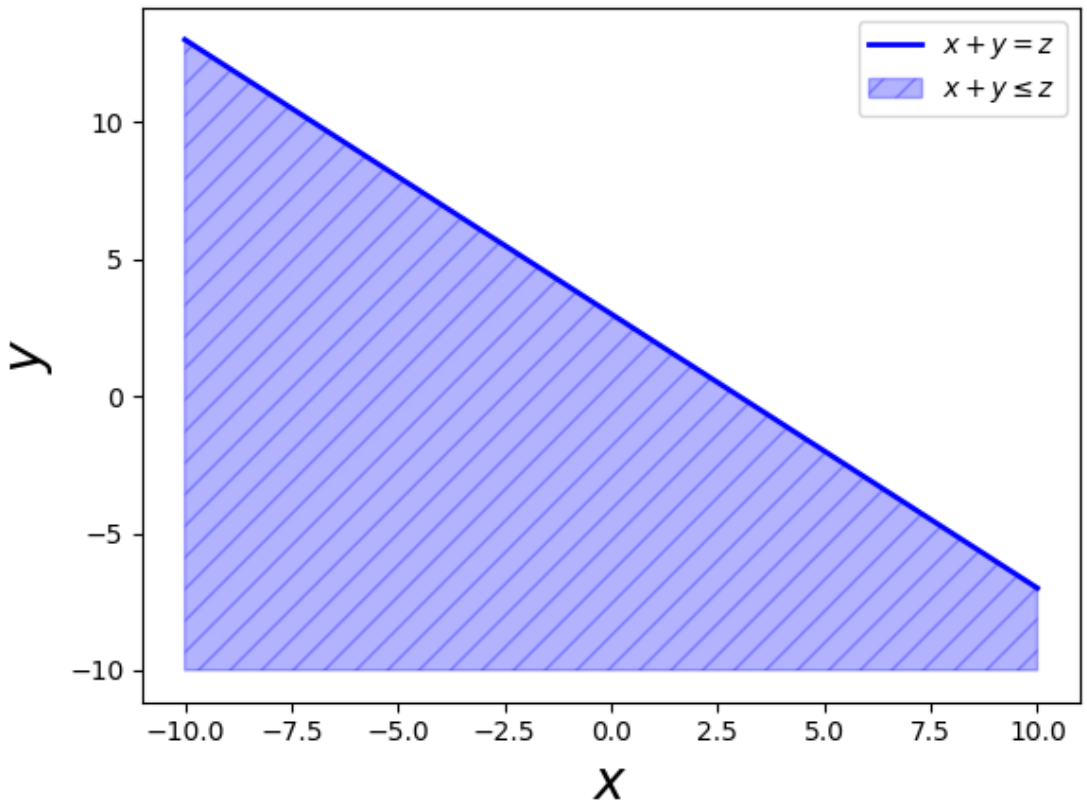


Abbildung 15: Fläche bzw. Gerade über welche die Verbund-DF $p(x, y)$ integriert werden muss, um $F_Z(2)$ bzw. $p_Z(2)$ zu erhalten.

Der Beweis erfolgt am einfachsten über die Siebeigensschaft des Diracstoßes

$$\int_{-\infty}^{+\infty} \delta(x - x')f(x)dx = f(x'). \tag{50}$$

Wir haben:

$$\begin{aligned} p_Z(z) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x,y)\delta(z - x - y)dx\ dy & (51)\\ &= \int_{-\infty}^{+\infty} p_X(x) \int_{-\infty}^{+\infty} p_Y(y)\delta(z - x - y)dy\ dx & (52)\\ &= \int_{-\infty}^{+\infty} p_X(x)p_Y(z - x)dx. & (53) \end{aligned}$$

Beispiel: Summe von normalverteilten Zufallsvariablen

Die Summe zweier unabhängiger, normalverteilter Zufallsvariablen $Z = X + Y$, mit $X \sim N(\mu_x, \sigma_x^2)$, $Y \sim N(\mu_y, \sigma_y^2)$ ist wiederum normalverteilt mit

$$Z \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2). \tag{54}$$

Zum Beweis setzt man im Faltungsintegral Gl 49 zwei Normalverteilungs-DFen ein; eine direkte Berechnung des resultierenden Integrals ist nicht schwierig, aber aufwendig. Am einfachsten erfolgt der Beweis unter Verwendung der charakteristischen Funktionen (Fourier-Transformation) von X und Y.

Beispiel: Chi-Quadrat Verteilung

Die Summe der Quadrate von k unabhängig standard-normalverteilten Größen $X_i \sim N(0,1)$ ist χ^2 (sprich: ki Quadrat) verteilt mit k Freiheitsgraden:

$$Q = \sum_{i=1}^{k} X_i^2 \sim \chi^2(k). \tag{55}$$

Eine $\chi^2(k)$-Verteilung hat das Mittel $\mathcal{E}[\chi^2(k)] = k$ und die Varianz $Var[\chi^2(k)] = 2k$.

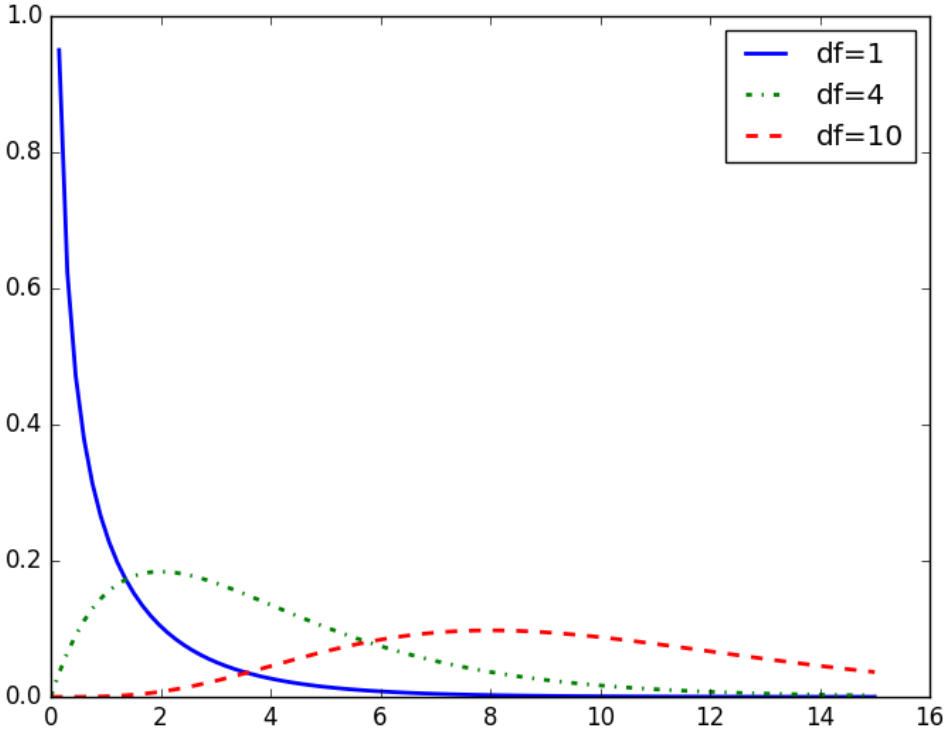


Abbildung 16: Dichtefunktion der χ^2-Verteilung für verschiedene Freiheitsgrade df=k. Für großes k geht die χ^2-Verteilung in eine Normalverteilung über.

Erwartungswerte

- Der Erwartungswert (*expectation*) $\mathcal{E}[]$ einer Funktion $Y = f(X)$ einer **stetigen** Zufallsvariablen X ist definiert als Funktional (Abbildung der Dichtefunktion auf einen Skalar)

$$\mathcal{E}[f(X)] = \int_{-\infty}^{\infty} f(x)p(x)dx, \tag{56}$$

  bzw. im bivariaten[6] Fall als

$$\mathcal{E}[f(X,Y)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)p(x,y)dxdy. \tag{57}$$

[6] Die Verallgemeinerung auf den allgemeinen multivariaten Fall werden wir später diskutieren.

Im **diskreten Fall** wird das Integral zur Summe über alle möglichen Elementarereignisse, und das Differential $p(x)dx$ zur Wahrscheinlichkeitsfunktion p_i

$$\mathcal{E}[f(X)] = \sum_{i \in \Omega'} f(i) p_i. \tag{58}$$

Im folgenden werden wir uns auf die Diskussion des stetigen Falles beschränken.

Ist $f(.)$ die Identitätsfunktion, so erhält man den Mittelwert von X (s.u.)

$$\mathcal{E}[X] = \int_{-\infty}^{\infty} x p(x) dx. \tag{59}$$

Interpretiert man $p(x)$ als Masseverteilung, so entspricht $\mathcal{E}[X]$ gerade dem Massenmittelpunkt.

Die allgemeinere Gleichung 56 liefert den Mittelwert $\mathcal{E}[Y]$ der transformierten Größe $Y = f(X)$, ohne die DF von Y explizit zu benötigen. Nehmen wir an, dass f monoton ist, so liefert Substitution $f(x) \rightarrow y$ unter Berücksichtigung des Transformationssatzes Gl. 46

$$\mathcal{E}[f(X)] = \int_{-\infty}^{\infty} f(x) p_X(x) dx = \int_{f(-\infty)}^{f(\infty)} y \; p_X(g(y)) \frac{\partial g}{\partial y} dy \quad (60)$$

$$= \int_{-\infty}^{\infty} y \; p_Y(y) dy = \mathcal{E}[Y], \quad (61)$$

wobei wir $g(y) = f^{-1}(y)$ gesetzt haben.

- **Indikatorfunktion**

  Die Wahrscheinlichkeit für ein Ereignis A bezüglich X lässt sich als Erwartungswert der **Indikatorfunktion**

$$f(x) = I_A(x) = \begin{cases} 1 \text{ für } x \in A \\ 0 \text{ sonst} \end{cases} \tag{62}$$

  ausdrücken, z.B. im stetigen Fall die Verteilungsfunktion

$$F(x) = P(X \in]-\infty, x]) = \mathcal{E}[I_{]-\infty,x]}(X)] = \int_{-\infty}^{x} p(x')dx' \tag{63}$$

  und im diskreten Fall die Wahrscheinlichkeitsfunktion

$$p_i = P(X = i) = P(X \in \{i\}) = \mathcal{E}[I_{\{i\}}(X)] \tag{64}$$

- **Momente**

  Für $f(X) = X^i$ erhält man das **Moment i-ter Ordnung** der Verteilung. Speziell erhält man für $i = 1$ den **Mittelwert** (*mean*) μ

$$\mu = \mathcal{E}[X] = \int_{-\infty}^{\infty} x \; p(x) dx \tag{65}$$

  Die **Varianz** σ^2 ergibt sich als zentrales Moment 2-ter Ordnung

$$\sigma^2 = Var[X] = \mathcal{E}[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx \tag{66}$$

  Es gilt außerdem

$$\sigma^2 = \mathcal{E}[X^2] - \mathcal{E}[X]^2. \tag{67}$$

- **Summe zweier Zufallsvariablen**

  Der Erwartungswert der Summe zweier Zufallsvariablen X, Y ist gleich der Summe der Erwartungswerte, im speziellen

$$\mathcal{E}[aX + bY] = a\mathcal{E}[X] + b\mathcal{E}[Y], \tag{68}$$

  für a, b konstant.

- **Produkt zweier Zufallsvariablen**

  Für **unabhängige** Zufallsvariablen X, Y gilt

$$\mathcal{E}[XY] = \mathcal{E}[X]\mathcal{E}[Y]. \tag{69}$$

- **Varianz der Summe zweier Zufallsvariablen**

$$\sigma^2_{X+Y} = \mathcal{E}[(X + Y - \mathcal{E}[X + Y])^2] = \sigma^2_x + \sigma^2_Y + 2\sigma_{XY}, \tag{70}$$

wobei σ_{XY} als **Kovarianz** (*covariance*) bezeichnet wird. Es gilt

$$\sigma_{XY} = \mathcal{E}[(X - \mathcal{E}[X])(Y - \mathcal{E}[Y])] = \mathcal{E}[XY] - \mathcal{E}[X]\mathcal{E}[Y]. \tag{71}$$

Im Falle der Unabhängigkeit von X, Y gilt $\sigma_{XY} = 0$, sodass

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y \tag{72}$$

- **Varianz einer skalierten Zufallsvariablen** aX

(a konstant):

$$Var[aX] = \sigma^2_{aX} = \mathcal{E}[(aX - \mathcal{E}[aX])^2] = a^2\sigma^2_X \tag{73}$$

- **Zentraler Grenzwertsatz** *(Central Limit Theorem)*

  Es wurde gezeigt, dass Mittelwert und Varianz einer Summe von unabhängigen Zufallsvariablen durch die Summe der Mittelwerte bzw. Varianzen gegeben sind. Wir wissen außerdem, dass man die Dichte- bzw. Wahrscheinlichkeitsfunktion einer Summe von unabhängigen Zufallsvariablen als deren Faltungsprodukt erhält.

  Sind die Stichprobenelemente X_i iid normalverteilt, so ist deren Summe ebenfalls normalverteilt, und zwar mit

$$\sum_{i=1}^{N} X_i \sim N\left(\sum_{i=1}^{N} \mathcal{E}[X_i], \sum_{i=1}^{N} Var[X_i]\right). \tag{74}$$

Dieses Resultat gilt asymptotisch auch für nicht-normalverteilte, unabhängige Summanden. Der Grenzwert der Verteilung einer Folge von Summen von Zufallsvariablen bzw. Faltungen der korrespondierenden Dichtefunktionen ist durch den **zentralen Grenzwertsatz** gegeben:

Die Summe von N unabhängigen Zufallsvariablen X_i konvergiert (für $N \to \infty$) gegen eine Normalverteilung.[7]

[7] Mittelwert und Varianz wie in Gl. 74.

Parameterschätzung I: Frequentistischer Ansatz

- Aufgabe der **Parameterschätzung** (*parameter estimation*) ist die Bestimmung der Verteilungsparameter (z.B. μ, σ) anhand einer Stichprobe des Umfangs N, $\mathcal{D} = [x_1, .., x_N]$, wobei die Stichprobenelemente x_i als Realisierungen von N **unabhängig und identisch verteilten** (*iid, independent and identically distributed*) Zufallsvariablen X_i angenommen werden. Genauer gesagt, wird vorausgesetzt, dass die X_i bedingt unabhängig gegeben den wahren – aber unbekannten – Wert des gesuchten Parameters θ sind

$$p(x_1, .., x_N|\theta) = \prod_{i=1}^{N} p(x_i|\theta). \tag{75}$$

Eine Funktion einer Zufallsstichprobe $\hat{\theta} = f(X_1, ..., X_N)$ wird als **Statistik** bezeichnet; diese ist wiederum eine Zufallsvariable. Im Kontext der Parameterschätzung ist zu unterscheiden zwischen dem

- wahren Parameter (auch *estimand)* $\theta = g[X]$ als Funktional der wahren Verteilung (z.B. Erwartungswert), dem

- **Schätzer** bzw. der **Schätzfunktion** (*estimator*) $\hat{\theta} = f(X_1, ..., X_N)$, sowie dem

- **Schätzwert** (*estimate*) $\hat{t} = f(x_1, ..., x_N) = f(\mathcal{D})$ als Realisierung des Schätzers.

In der Literatur wird allerdings oft nicht deutlich zwischen Schätzer und Schätzwert unterschieden. Die Verteilung des Schätzers wird im Englischen als *sampling distribution* bezeichnet.

Die Stichprobenelemente X_i repräsentieren N Wiederholungen desselben Zufallsversuches X (oder, anders formuliert, N Messungen desselben Merkmals X an zufällig ausgewählten Populationsmitgliedern), z.B N-maliges Werfen einer Münze, oder Messung der Körpergröße von N zufällig ausgewählten Personen.

Die X_i folgen alle derselben Verteilung und besitzen daher dieselben Verteilungsparameter. Insbesondere gilt für beliebige Erwartungen $\mathcal{E}[h(X_i)] = \mathcal{E}[h(X_j)] = \mathcal{E}[h(X)]$.

Ist, wie im obigen Fall, die Unterscheidung zwischen den Wiederholungen nicht relevant, schreiben wir auch kurz X statt X_i.

Achtung: Das Produkt $X_i X_j$ ist nur im Falle $i \neq j$ unabhängig, jedoch für $i = j$ abhängig (da im letzteren Fall beide Zufallsvariablen für jede mögliche Realisierung denselben Wert annehmen müssen).

- **Die Maximum likelihood-Methode (ML)**

  Dies ist das wichtigste Verfahren, um zu einer Schätzfunktion zu gelangen. Ausgangspunkt ist die bedingte Dichtefunktion (bzw. Wahrscheinlichkeitsfunktion im diskreten Fall) der Stichprobe, gegeben den wahren Wert des Parameters θ

$$p(\mathcal{D}|\theta) = p(x_1, .., x_n|\theta) \tag{76}$$

  Dies ist, für gegebenen Parameter θ, eine Funktion der Stichprobe $\mathcal{D}$. ML fasst nun die Stichprobe (genauer: deren Realisation) als Funktion des gesuchten Parameters θ (likelihood-Funktion) auf

$$l(\theta) = p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(x_i|\theta) \tag{77}$$

wobei der letzte Schritt aus der bedingten Unabhängigkeit der X_i folgt.

ML wählt jenen Wert des Parameters θ^*, welcher die joint-likelihood Eq. 77 maximiert. Oft ist es einfacher, den Logarithmus von Eq. 77 zu maximieren; dies führt zur log-likelihood-Funktion

$$ll(\theta) = \log l(\theta) = \sum_{i=1}^{N} \log p(x_i|\theta). \tag{78}$$

Den ML-Schätzer θ^* erhält man dann durch Nullsetzen der ersten Ableitung der (log-)likelihood-Funktion und Auflösen nach θ

$$\frac{\partial l(\theta)}{\partial \theta} = 0 \tag{79}$$

Beispiel: Bernoulli Verteilung

Logarithmieren der DF bzw. likelihood

$$l(\theta) = p(x_1, ..x_N|\theta) = \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{(1-x_i)} \tag{80}$$

liefert

$$ll(\theta) = \sum_{i=1}^{N} x_i \log(\theta) + (1-x_i)\log((1-\theta)) \tag{81}$$

Ableiten nach θ und Nullsetzen der Ableitung liefert

$$\frac{\partial}{\partial\theta} ll(\theta) = \sum_{i=1}^{N} \frac{x_i}{\theta} - \frac{1-x_i}{1-\theta} = \left(\sum_{i=1}^{N} x_i\right) - N\theta = 0, \tag{82}$$

und somit

$$\theta = \frac{k}{N}, \tag{83}$$

wobei k die Anzahl der günstigen Ausfälle bezeichnet.

ML liefert somit als Schätzung des wahren Anteils den Anteil der günstigen Ausfälle in der Stichprobe.

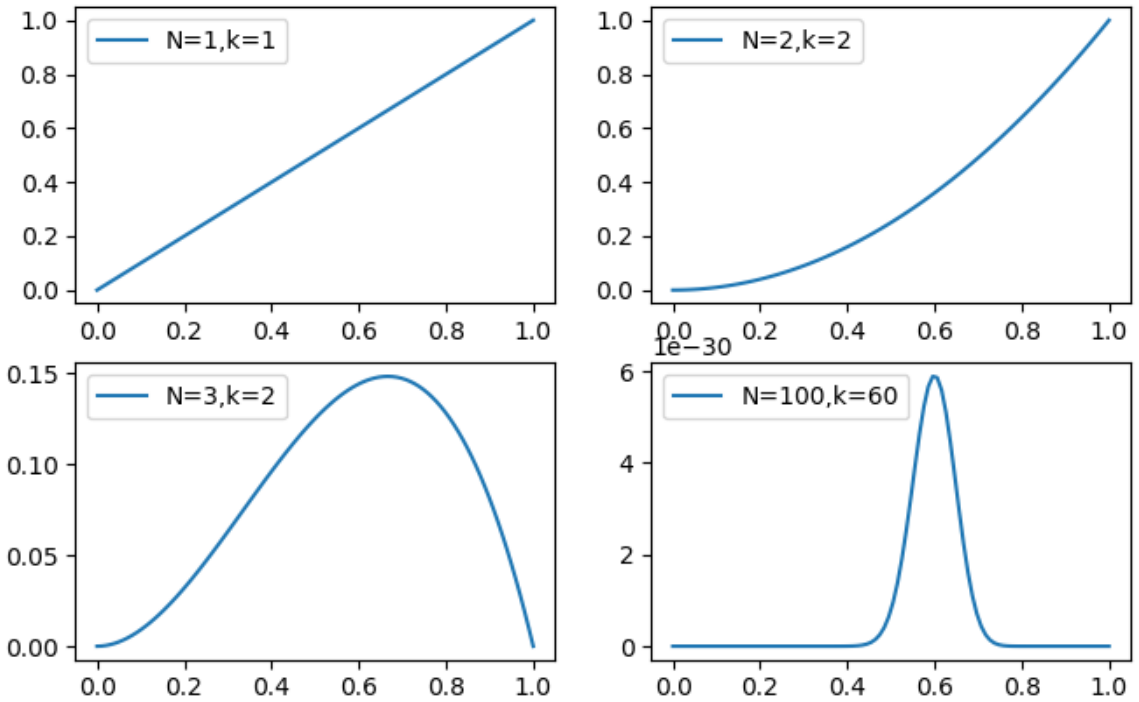


Abbildung 17: *likelihood* $l(\theta)$ der Bernoulli-Verteilung für verschiedene Werte von N und k.

Beispiel: Mittel der Normalverteilung

Die Dichtefunktion der Stichprobe gegeben μ (σ wird als bekannt vorausgesetzt) ist:

$$l(\mu) = p(x_1, ..x_n|\mu) = \frac{1}{(\sqrt{2\pi}\sigma)^N} \prod_{i=1}^{N} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{84}$$

Durch Logarithmieren erhalten wir

$$ll(\mu) = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2 + const, \tag{85}$$

wobei $const$ ausschließlich Terme enthält, die nicht vom gesuchten Parameter μ abhängen. Anstatt Gl. 85 zu maximieren, können wir

genauso gut

$$\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2 \tag{86}$$

minimieren. Anders formuliert: unter Annahme einer Normalverteilung erhalten wir den Schätzer des Populationsmittels, indem wir die Fehlerquadratsumme Eq. 86 minimieren. Bilden der ersten Ableitung bezüglich des Parameters μ und Nullsetzen derselben liefert

$$\frac{1}{2\sigma^2}\frac{\partial}{\partial\mu}\sum_{i=1}^{N}(x_i-\mu)^2 = 0 \tag{87}$$

$$\left(\sum_{i=1}^{N}x_i\right)-N\mu = 0 \tag{88}$$

$$\mu^* = \bar{x} = \frac{\sum_{i=1}^{N}x_i}{N} \tag{89}$$

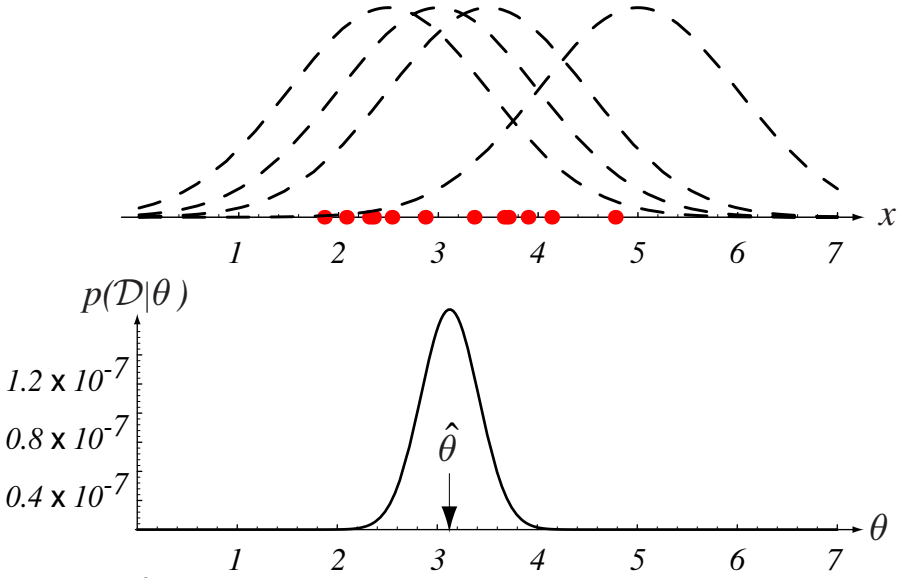


Abbildung 18: Beispiel zur ML-Parameterschätzung. Gesucht ist der Mittelwert $\theta = \mu$ einer $N(\mu, \sigma^2)$-Verteilung (σ^2 bekannt).
Oben: Trainingspunkte und Kandidaten für die generierende pdf.
Unten: Verlauf der joint-likelihood $p(\mathcal{D}|\theta)$. Diese wird mit zunehmendem N enger.
(Aus *Duda, Hart, Stork: Pattern Classification, 2nd ed.*)

Der ML-Schätzwert für den Mittelwert μ ist also das arithmetische Mittel $\bar{x}$. Dieser wurde für eine gegebene, aber beliebige Stichprobe $\mathcal{D}$ hergeleitet, und ist somit eine Realisierung des Schätzers. Um von diesem ausgehend eine Schätzfunktion zu erhalten, substituieren wir formal die x_i durch ihre korrespondierenden X_i, und erhalten auf diese Weise das sogenannte **Stichprobenmittel** (siehe unten).

- **Exkurs: Fisher Information**

  Die log-likelihood-Funktion ll (Logarithmus der bedingten DF der Stichprobenverteilung) stellt nicht nur die Grundlage für ML-Schätzer dar, sondern charakterisiert ganz allgemein die Eigenschaften von Schätzern der Parameter der Stichprobenverteilung $p(\mathcal{D}|\theta)$. Die *Fisher Information*

$$\mathcal{I}(\theta) = \mathcal{E}\left[(\frac{\partial}{\partial\theta}\log p(\mathcal{D}|\theta))^2\right] \quad (90)$$

$$= -\mathcal{E}\left[\frac{\partial^2}{\partial\theta^2}\log p(\mathcal{D}|\theta))\right] \quad (91)$$

  (mit Erwartungswert bzg. $p(\mathcal{D}|\theta)$) lässt sich unter milden Voraussetzungen sowohl als Varianz (90) als auch als mittlere Kurvatur (91) von ll interpretieren. Große Werte der Krümmung bedeuten einen schmalen Gipfel der likelihood bzg. von ll und somit eine geringe Unsicherheit bzg.

des zu schätzenden Parameters θ.

Die Fisher Information FI ist additiv, d.h. ist $\mathcal{I}(\theta)$ die FI einer Stichprobe vom Umfang 1, so ist die FI einer iid Stichprobe vom Umfang N durch $N\mathcal{I}(\theta)$ gegeben.

Die FI erlaubt es, die Varianz von Schätzern (genauer: ihrer *sampling distribution*) abzuschätzen. Die Varianz erwartungstreuer Schätzer ist gemäß dem Satz von **Cramer Rao** nach unten durch den Kehrwert der FI beschränkt

$$Var[\hat{\theta}] \geq \frac{1}{N\mathcal{I}(\theta)}. \tag{92}$$

Für ML-Schätzer wird die untere Grenze angenommen.

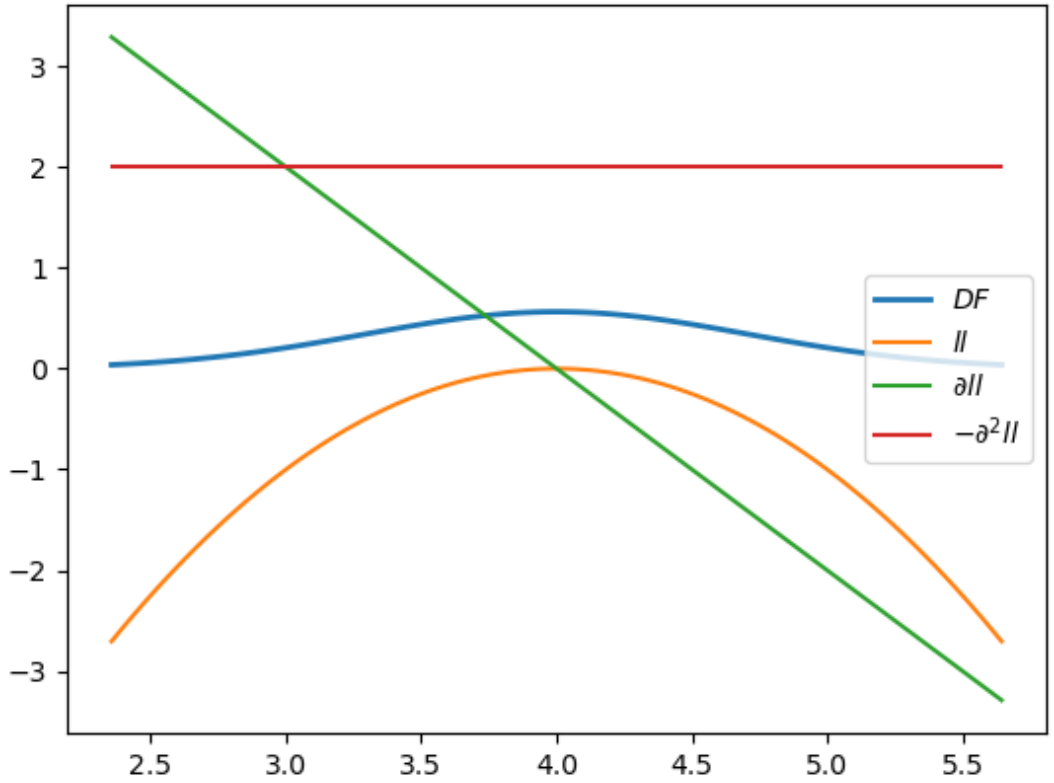


Abbildung 19: Normalverteilungs-DF, log-likelihood $ll(\mu)$, erste $\partial ll/\partial\mu$ und negative zweite Ableitung $-\partial^2 ll/\partial\mu^2$ bzg. μ. Da letztere konstant ist, ist sie auch gleich ihrem Erwartungswert bzg. der Stichprobe x und somit $\mathcal{I}(\mu)$.

- **Schätzung des Populationsmittels: Das Stichprobenmittel**

  Der wahre Mittelwert gemäß Eq. 65, welcher auch als **Populationsmittel** (*population mean*) bezeichnet wird, kann mittels des **Stichprobenmittels** (*sample mean*)

$$\hat{\mu} = \bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i \tag{93}$$

  geschätzt werden.

  $\hat{\mu}$ ist als Funktion einer Zufallsstichprobe (Statistik, Schätzer) selbst eine Zufallsgröße.

Eine Realisierung des Stichprobenmittels $\bar{X}$ – d.h. seinen Wert für ein konkretes sample $[x_1, \ldots, x_N]$ – werden wir im folgenden mit $\hat{m}$ bezeichnen:

$$\hat{m} = \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{94}$$

- **Erwartungstreue des Stichprobenmittels**

  $\hat{\mu} = \bar{X}$ ist **erwartungstreu** (*unbiased*), da

$$\mathcal{E}[\hat{\mu}] = \mathcal{E}[\bar{X}] = \frac{1}{N}\sum_{i=1}^{N}\mathcal{E}[X_i] = \frac{1}{N}\sum_{i=1}^{N}\mathcal{E}[X] = \mu, \tag{95}$$

  d.h. der Erwartungswert des Schätzers ist der gesuchte Parameter. Man beachte, dass der Erwartungswert hier bezüglich der Verteilung aller Stichproben des Umfangs N, d.h. einer N-dimensionalen Zufallsvariablen berechnet wird.

- **Varianz des Stichprobenmittels**

  Gemäß Eq. 72 (Unabhängigkeit der X_i!) und Eq. 73 berechnet sich die Varianz $\sigma^2_{\hat{\mu}}$ des Schätzers $\hat{\mu}$ als

$$\sigma^2_{\hat{\mu}} = \frac{1}{N^2} \sum_{i=1}^{N} \sigma^2_X = \frac{\sigma^2_X}{N}, \tag{96}$$

  σ^2_X bezeichnet hier die wahre (und für alle X_i identische) Populationsvarianz.

- **Asymptotische Verteilung des Stichprobenmittels**

  Die asymptotische Verteilung (die *sampling distribution*) des Stichprobenmittels $\bar{X}$ einer Stichprobe von N iid Beobachtungen mit $\mathcal{E}[X_i] = \mu$ und $Var[X_i] = \sigma^2$ folgt aus dem zentralen Grenzwertsatz:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{N}). \tag{97}$$

- **Eigenschaften von Schätzern**

  Sei $\hat{\theta}$ ein Schätzer des Parameters θ. Es sei noch einmal angemerkt, dass Erwartungswerte im Kontext von Schätzern sich auf die Verteilung aller Stichproben vom Umfang N beziehen, siehe Eq. 75.

  - **Erwartungstreue**
    Der *bias* ist definiert als

$$bias(\hat{\theta}) = \mathcal{E}[\hat{\theta}] - \theta. \tag{98}$$

    Im Falle der Erwartungstreue gilt *bias* $= 0$.

  - **Varianz** (*variance*)

$$var(\hat{\theta}) = \mathcal{E}\left[\left(\hat{\theta} - \mathcal{E}[\hat{\theta}]\right)^2\right] \tag{99}$$

- Die Standardabweichung eines Schätzers $se(\hat{\theta}) = \sqrt{var(\hat{\theta})}$ wird auch als dessen **Standardfehler** (standard error) bezeichnet.

- **Mean Squared Error MSE**

$$mse(\hat{\theta}) = \mathcal{E}[(\theta - \hat{\theta})^2] = bias^2(\hat{\theta}) + var(\hat{\theta}) \tag{100}$$

- **Effizienz**
  Je geringer die Varianz $var(\hat{\theta})$, desto effizienter ist $\hat{\theta}$.

- **(Asymptotische) Konsistenz**
  Der wahre Populationsparameter lässt sich für $N \to \infty$ beliebig genau schätzen. Hierfür ist notwendig, dass sowohl *bias* als auch *variance* (und somit der MSE) für $N \to \infty$ gegen 0 gehen.

– **Robustheit** (robustness)
  Unempfindlichkeit gegenüber Ausreißern (extremen Werten) in der Stichprobe. Das Stichprobenmittel ist z.B. nicht robust, da ein einzelner Ausreißer die Schätzung beliebig weit vom wahren Mittel wegziehen kann. Robuste Schätzer wie der Median sind jedoch i.a. weniger effizient als ihre nicht-robusten Gegenstücke.

Verwandte Größen sind **Richtigkeit** *(trueness)*, welche als Abwesenheit von *bias* – sprich: Erwartungstreue – definiert ist, und **Präzision** *(precision)*, welche üblicherweise – vor allem in der Statistik – als Kehrwert der Varianz – aufgefasst wird. **Genauigkeit** *(accuracy)* berücksichtigt sowohl Richtigkeit als auch Präzision (im MSE-Sinn). *Accuracy* wird manchmal auch im Sinne von Richtigkeit verwendet, daher ist Vorsicht angebracht!

Ähnliches gilt für **Validität** *(validity)* und **Reliabilität** *(reliability)*; diese sind jedoch nicht einfach als Synonyme für Richtigkeit und Präzision anzusehen, sondern haben, ja nach Disziplin, eine etwas andere bzw. erweiterte Bedeutung.

- **Schätzung von Erwartungswerten: Gesetz der großen Zahl**

  Das Populationsmittel Eq. 65 ist als spezieller Erwartungswert definiert. Asymptotisch konsistente Schätzer für andere Erwartungswerte gemäß Eq.56 können analog als Stichprobenmittel konstruiert werden, sprich

$$Z_N = \overline{h(X)} = \frac{1}{N}\sum_{i=1}^{N} h(X_i) \tag{101}$$

  ist unter den üblichen Voraussetzungen (X_i iid) ein asymptotisch konsistenter Schätzer von $E[h(X)]$. Formal wird dies durch das **(schwache) Gesetz der großen Zahl** ausgedrückt:

$$\lim_{N\to\infty} P(|Z_N - \mathcal{E}[h(X)]| > \epsilon) = 0 \tag{102}$$

Für jedes (beliebig kleine, jedoch positive) ϵ und unabhängige X_i geht die Wahrscheinlichkeit, dass sich das Stichprobenmittel um mehr als ϵ vom Erwartungswert unterscheidet, mit wachsender Stichprobengröße gegen 0.

Wir betrachten im folgenden zwei Spezialfälle, die Schätzung von Anteilen $h(X) = I_A(X)$ und die Schätzung der Populationsvarianz $h(X) = (X - \mu)^2$.

- **Anteilsschätzer**

  Sei X die Augenzahl beim Würfeln und $h(x) = I_{\{4\}}(x)$ die Indikatorfunktion für das Elementarereignis *"Augenzahl 4"*. Z_N entspricht somit dem *"Anteil der 4er in einer Stichprobe vom Umfang N"*, und $\mathcal{E}[h(X)]$ dem wahren Anteil (sprich: der Wahrscheinlichkeit), einen 4er zu Würfeln.

  Schätzer des Anteils werden oft mit $Z_N = \hat{p}$ bezeichnet. Anteilsmerkmale sind binomialverteilt; die Verteilung des Schätzers wird in der Praxis aber oft durch eine Normalverteilung

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{N}\right) \tag{103}$$

  mit Mittel p (dem wahren Anteil) und Standardfehler $se = \sqrt{p(1-p)/N}$ angenähert.

- **Schätzung der Populationsvarianz**

$$\hat{\sigma}_X^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2 \tag{104}$$

$$\hat{\sigma}_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \hat{\mu})^2. \tag{105}$$

Beide Schätzer bezeichnet man als **Stichprobenvarianz** (*sample variance*) von X. Eq. 104 ist anwendbar, wenn das Populationsmittel μ bekannt ist. Muss es jedoch aus der Stichprobe geschätzt werden, unterschätzt Eq. 104 die Varianz; Eq. 105 korrigiert diesen *bias* und ist auch bei Verwendung des geschätzten Mittelwerts erwartungstreu.

Eine Realisierung von $\hat{\sigma}_X^2$ werden wir im folgenden mit $\hat{s}_X^2$ bezeichnen.

Form der Verteilung des Varianzschätzers

Sind die Stichprobenelemente iid **normalverteilt** mit $X_i \sim N(\mu, \sigma^2)$, so gilt für bekanntes Mittel μ:

$$\frac{1}{\sigma^2} \sum_{i=1}^{N} (X_i - \mu)^2 \sim \chi^2(N) \tag{106}$$

und somit

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (X_i - \mu)^2 \sim \frac{\sigma^2}{N} \chi^2(N). \tag{107}$$

Aus den Eigenschaften der Chi-Quadrat-Verteilung sowie der Varianz einer skalierten Zufallsvariable ergibt sich

$$\mathcal{E}[\hat{\sigma}^2] = \sigma^2 \tag{108}$$

$$Var[\hat{\sigma}^2] = \frac{2}{N}\sigma^4. \tag{109}$$

Bei Verwendung des Schätzers Gl. 105 (unbekanntes Mittel) ist N durch $N-1$ zu ersetzen.

Angabe der Genauigkeit von Schätzungen

- Ein interessierender Parameter θ ist durch Angabe des Schätzwerts allein i.a. nicht ausreichend bestimmt; es muss auch die mit dem Schätzer verbundene Unsicherheit angegeben werden; im einfachsten Fall kann dies durch Angabe des (geschätzten) Standardfehlers $se(\hat{\theta})$ geschehen. Wird z.B. der Mittelwert eines Merkmals mit bekannter Varianz σ^2 aus einer Stichprobe vom Umfang N geschätzt, so ist der Standardfehler durch $se(\bar{X}) = \sigma/\sqrt{N}$ gegeben, für Anteilsschätzer durch $\sqrt{p(1-p)/N}$.

  Ist die Verteilung des Schätzers bekannt, so lässt sich ein Bereich angeben, welcher den wahren Parameter mit einer gegebenen Wahrscheinlichkeit überdeckt, ein sogenanntes Konfidenz- bzw. Schätzintervall.

- **Einführende Betrachtungen zum Thema Konfidenzintervalle**

  Bei bekannter Verteilung lässt sich für einen interessierenden Parameter θ ein Intervall $[\theta - a, \theta + b]$ angegeben, welches θ enthält und einen $1-\alpha$-Anteil der Verteilung abdeckt. Dieses Intervall ist i.a. nicht eindeutig; wir wollen im folgenden davon ausgehen, dass das Intervall vom linken und rechten Schwanz der Verteilung eine Fläche von je $\alpha/2$ abschneidet. Wir haben somit

$$
\begin{aligned}
P(x_{\alpha/2} \leq X \leq x_{1-\alpha/2}) &= && (110)\\
P(x_{\alpha/2} - \theta \leq X - \theta \leq x_{1-\alpha/2} - \theta) &= && (111)\\
P(-a \leq X - \theta \leq b) &= && (112)\\
P(X - b \leq \theta \leq X + a) &= 1 - \alpha. && (113)
\end{aligned}
$$

Man beachte den Positions- und Vorzeichenenwechsel von a und b zwischen Gl. 112 und Gl. 113, welche ein sogenanntes 1-α - **Konfidenzintervall** für θ festlegt . Man beachte weiters, dass in Gl. 113 nun die Intervallgrenzen (als Funktionen von X) Zufallsgrößen sind, wohingegen die Größe θ, welche im Inneren des Intervalls liegt, eine (wenn auch unbekannte) Konstante ist.

$(1 - \alpha)$ bezeichnet man als **Überdeckungswahrscheinlichkeit** oder **Konfidenzzahl**.

Abb. 20 illustriert dies am Beispiel des Medians einer $\chi^2(3)$-Verteilung. Die Längen der beiden roten Teilintervalle a und b entsprechen genau dem Abstand vom Median zum 5 bzw. 95-Perzentil der Verteilung; aufgrund der Asymmetrie der χ^2-Verteilung sind diese Strecken jedoch ungleich lang; das rote Gesamtintervall $[x_{0.05}, x_{0.95}]$ gibt einen zentralen Bereich mit 0.9 Überdeckungswahrscheinlichkeit an. Die beiden grünen Intervalle entsprechen Realisierungen des Konfidenzintervalls (sogenannte **Schätzintervalle**) für die zwei Werte $X = x_{0.05}$ und $X = x_{0.95}$, welche gerade noch im zentralen 0.9 Überdeckungsbereich der Verteilung (rotes Intervall) liegen. Man sieht, dass diese beiden Schätzintervalle gerade noch den Parameter θ überdecken (man beachte auch die Vertauschung der Teilintervalle in den Schätzintervallen im Vergleich zum wahren Überdeckungsbereich); in jenen 10% der Fälle, wo X außerhalb des roten 0.9-Überdeckungsbereichs liegt, enthält das Konfidenzintervall den Parameter θ nicht.

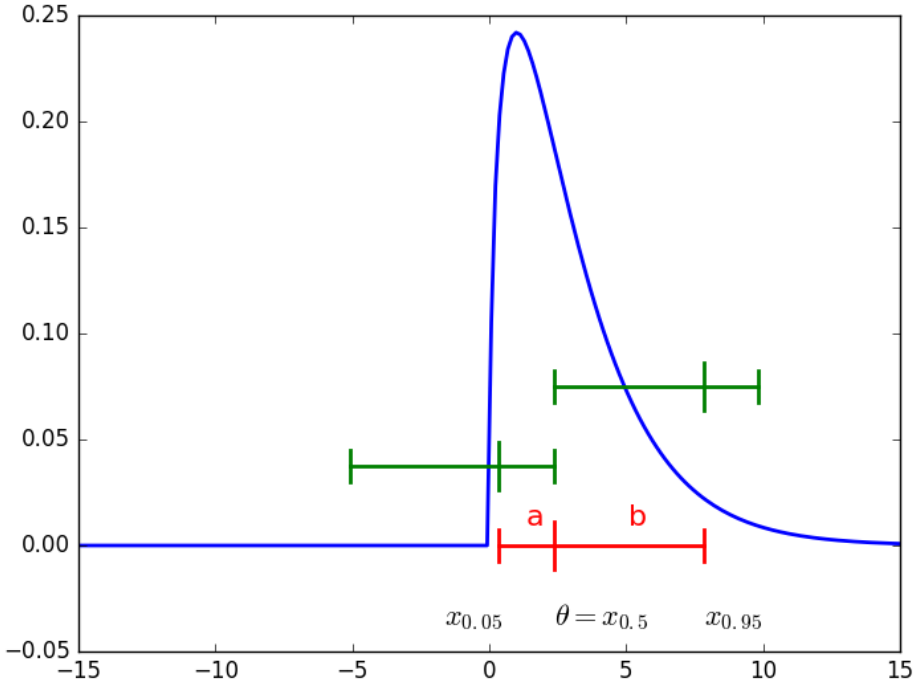


Abbildung 20: DF der $\chi^2(3)$ Verteilung. Die Markierungen auf der roten Strecke entsprechen dem 5, 50 und 95-Perzentil. Die grünen Strecken sind Schätzintervalle für $\theta = x_{0.5}$ für zwei extreme Beobachtungen an der unteren und oberen Grenze des zentralen 90%-Überdeckungsbereichs der Verteilung.

- **Konfidenzintervall am Beispiel des Stichprobenmittels**

  Wenn die Strichprobenelemente X_i iid normalverteilt sind, d.h, $X_i \sim N(\mu, \sigma^2), 1 \leq i \leq N$, so ist das Stichprobenmittel $\hat{\mu}$ ebenfalls normalverteilt mit $\hat{\mu} \sim N(\mu, \frac{\sigma^2}{N})$. Bezeichne z_α das α-Quantil der Standardnormalverteilung. Es gilt

$$P(z_{\alpha/2} \leq \frac{\hat{\mu} - \mu}{\frac{\sigma}{\sqrt{N}}} \leq z_{1-\alpha/2}) = 1 - \alpha, \tag{114}$$

  bzw. konkret für $\alpha = 0.05$:

$$P(z_{0.025} \leq \frac{\hat{\mu} - \mu}{\frac{\sigma}{\sqrt{N}}} \leq z_{0.975}) = 0.95, \tag{115}$$

  sprich: der Schätzer $\hat{\mu}$ liegt mit 95%iger Wahrscheinlichkeit (für 95 von

100 Stichproben) im Intervall:

$$[\mu + z_{0.025}\frac{\sigma}{\sqrt{N}}, \mu + z_{0.975}\frac{\sigma}{\sqrt{N}}]$$

bzw. unter Verwendung der Identität $z_\alpha = -z_{1-\alpha}$

$$[\mu - z_{0.975}\frac{\sigma}{\sqrt{N}}, \mu + z_{0.975}\frac{\sigma}{\sqrt{N}}] \tag{116}$$

Durch Umformung erhält man

$$P(\hat{\mu} - z_{0.975}\frac{\sigma}{\sqrt{N}} \leq \mu \leq \hat{\mu} + z_{0.975}\frac{\sigma}{\sqrt{N}}) = 0.95 \tag{117}$$

sprich: für 95 von 100 Stichproben überdeckt das obige 0.95-**Konfidenz-Intervall** (confidence interval) den wahren Populations-Parameter μ.

Für eine Realisierung $\hat{m}$ von $\hat{\mu}$ bezeichnet man

$$[\hat{m} - z_{0.975}\frac{\sigma}{\sqrt{N}}, \hat{m} + z_{0.975}\frac{\sigma}{\sqrt{N}}] \tag{118}$$

auch als **Schätz-Intervall**.

- Man beachte, dass für eine gegebene Stichprobe (Realisierung des Schätzers) das obige Intervall nichts über die Verteilung des Schätzers aussagt: der wahre Parameter wird entweder vom Schätz-Intervall überdeckt oder nicht.
- Ist die Populationsvarianz nicht bekannt, sondern muss diese aus der Stichprobe geschätzt werden, so ist die standardisierte Abweichung des Stichprobenmittels vom wahren Mittel $\frac{\hat{\mu}-\mu}{se(\hat{\mu})} = \frac{(\hat{\mu}-\mu)\sqrt{N}}{\hat{\sigma}}$ Student-t verteilt mit $N-1$ Freiheitsgraden. Da sich die Student-t Verteilung jedoch (als Funktion von N) der Normalverteilung recht schnell annähert, wird in der Praxis meist die Normalverteilung verwendet.
- Für Stichproben-Mittelwerte kann, auch wenn das Merkmal in der Grundgesamtheit nicht normalverteilt ist, bei nicht zu kleinen Stichproben die Normalverteilung angenommen werden (zentraler Grenzwertsatz).

- **Bootstrap**

  Der Bootstrap gehört zu den Monte-Carlo-Verfahren, ersetzt also analytische Verfahren durch eine Computersimulation.

  Die Stichprobe $\mathcal{D} = [x_1, .., x_N]$ wird als endliche Grundgesamtheit GG aufgefasst. Durch Ziehen mit Zurücklegen werden B *bootstrap samples* $\mathcal{D}^{*i}$ der Größe N aus GG gezogen (die Elemente sind identisch und unabhängig – iid – verteilt!). Auf diese wird die Schätzfunktion $f(.)$ angewandt, was in B sogenannten *bootstrap replicas* $f(\mathcal{D}^{*i}), 1 \leq i \leq B$, resultiert; diese sind unabhängige Realisationen der Schätzfunktion bzg. GG. Die Verteilung der $f(\mathcal{D}^{*i})$ kann nun als Grundlage zur Charakterisierung der Genauigkeit des Schätzers herangezogen werden: wir können z.B. die empirische Standardabweichung (als Schätzer des Standardfehlers), aber auch verschiedene Quantile etc. berechnen.

Binäre Klassifizierung

- Motivation: für die Diskussion von Signifikanztests weiter unten benötigen wir einige Begriffe aus der statistischen Testtheorie, die wir hier zunächst als Spezialfall der Klassifizierung einführen.

- Klassifizierung ist mit dem Problem befasst, festzustellen, zu welcher von mehreren gegebenen Populationen (Klassen, Kategorien) eine Beobachtung gehört, z.B. Landnutzungsklassen (Acker, Urban, Wald) in der Fernerkundung. In der statistischen Mustererkennung werden entweder die Verteilung der Klassen (generative Verfahren, z.B. LDA) oder die Grenzen zwischen den Klassen (diskriminative Verfahren, wie SVM, logistische Regression) im Merkmalsraum modelliert, wobei auch eine *reject option* (keine zuverlässige Klassifizierung möglich) vorgesehen sein kann.

Wir befassen uns zunächst mit binären Klassifizierungsproblemen (in der Signalverarbeitung spricht man in diesem Zusammenhang auch von **Detektion**), und bezeichnen die zwei interessierenden Klassen (genauer gesagt, die Annahme, dass eine Beobachtung einer dieser Klassen angehört) in Anlehnung an die statistische Testtheorie (s.u.) als **Null-Hypothese** H_0 und **Alternativ-Hypothese** H_1 . Null- und Alternativ-Hypothese sind i.a. nicht gleichwertig: H_0 bedeutet „Normalfall“ bzw. "keine Veränderung", H_1 hingegen "deutliche Veränderung".

Beispiele:

- Ist ein Patient gesund (H_0) oder krank (H_1)?
- Genügt ein Werkstück den Qualitätsanforderungen (H_0) oder ist es defekt (H_1)?
- Führt ein Tempolimit zu weniger Unfällen (H_1) oder nicht (H_0)?
- Verringert eine neue Krebstherapie die Mortalitätsrate (H_1) oder nicht?
- Wurde ein Radarecho von einem Objekt verursacht (H_1) oder handelt es sich um Rauschen (H_0)?

- **Kenngrößen von binären Klassifikatoren**

  Wir gehen im folgenden von zwei Populationen (Klassen) H_1 und H_0 aus, die anhand eines gemeinsamen Merkmals X unterschieden werden sollen. Sei T ein Test (binärer Klassifikator), der entscheiden soll, ob ein gegebenes Objekt zu H_1 gehört $(T = +)$ oder zu H_0 $(T = -)$. Es gibt 4 mögliche Kombinationen von Testergebnissen und wahren Klassenzugehörigkeiten:

	H_1	H_0
$+$	true positive (tp)	false positive (fp)
-	false negative (fn)	true negative (tn)

Tabelle 7: Tatsächliche Klassenzugehörigkeit (Spalten) vs. vorhergesagte Klassenzugehörigkeit (Zeilen)

Wenn in obiger Tabelle für jedes Ereignis (z.B tp) dessen Häufigkeit (z.B. $\#tp$) eingetragen wird, erhalten wir eine *Kontingenztafel*. Die nachfolgenden Wahrscheinlichkeiten lassen sich aus einer solchen Kontingenztafel berechnen (endliche Grundgesamtheit) bzw. schätzen (Stichprobe):

- **Sensitivität** (sensitivity, true positive rate tpr)
  $P(+|H_1)$, $\frac{\#tp}{\#tp+\#fn}$
- **Falsch-Negativ-Rate** (false negative rate, fnr)
  $P(-|H_1) = 1 - P(+|H_0)$, $\frac{\#fn}{\#tp+\#fn}$
- **Spezifität** (specificity, true negative rate, tnr)
  $P(-|H_0)$, $\frac{\#tn}{\#tn+\#fp}$
- **Falsch-Positiv-Rate** (false positive rate, fpr)
  $P(+|H_0) = 1 - P(-|H_0)$, $\frac{\#fp}{\#tn+\#fp}$

Man beachte, dass z.B

$$P(tp) = P(+, H_1) = P(+|H_1)P(H_1) = tprP(H_1) \tag{119}$$

Werden die Zeilen statt der Spalten als Referenz (bedingende Größen) verwendet, erhält man z.B. folgende Wahrscheinlichkeiten:

- **Positiver Vorhersagewert** (positive predictive value, ppv)
  $P(H_1|+)$, $\frac{\#tp}{\#tp+\#fp}$

- **Negativer Vorhersagewert** (negative predictive value, npv)
  $P(H_0|-)$, $\frac{\#tn}{\#tn+\#fn}$

- **Beispiel: Diagnose von Krankheiten**

  Wir untersuchen einen Test für eine Krankheit mit Prävalenz $P(H_1) = 0.001$ von 1 Promille (1 von 1000 in der Bevölkerung ist betroffen). Der Test habe eine fp-Rate $P(+|H_0) = 0.01$ von 1 Prozent und eine Sensitivität $P(+|H_1) = 1$ von 100 Prozent (d.h., jede Erkrankung wird detektiert). Am anschaulichsten – und am wenigsten anfällig für falsche Interpretationen – ist die Darstellung in Form einer Kontingenztafel für eine hypothetische endliche Grundgesamtheit, wie in Tabelle 8 gezeigt.

  Man bemerkt, dass – obwohl die fpr für sich genommen als recht gut erscheinen mag – die Wahrscheinlichkeit, dass ein positiver Testbefund mit einer tatsächlichen Erkrankung einhergeht, also der positive Vorhersagewert $P(H_1|+)$, nur $1/11$ beträgt. Das ist auf die geringe Prävalenz der Krankheit in der Bevölkerung zurückzuführen; ganz allgemein gilt, dass je seltener eine Krankheit, desto größer der Effekt von falsch positiven Klassierungen hinsichtlich der ppv. Positive Testergebnisse sind also

cum grano salis zu genießen; andererseits kann man sich im gegebenen Beispiel bei einem negativen Testergebnis zu 100% sicher sein, nicht erkrankt zu sein.

	H_1	H_0	$\sum$	
$+$	1	10	11	ppv: $1/11 = 0.09$
-	0	989	989	npv: $989/989 = 1$
$\sum$	1	999	1000	
	fnr: $0/1$=0	fpr: $10/999 = 0.01$		

Tabelle 8: Kontingenztafel mit ausgewählten Qualitätskennzahlen für eine hypothetische Population von $N = 1000$ (siehe Text).

Statistische Testtheorie

- Ziel eines statistischen Tests ist es, die Null-Hypothese zu verwerfen bzw. die Alternativ-Hypothese anzunehmen.

- Ein statistischer Test ist ein Verfahren zur induktiven Entscheidungsfindung bzw. -begründung; eine aufgrund der Stichprobe erfolgte Ablehnung der Null-Hypothese kann nicht als logischer (deduktiver) Beweis ihrer Falschheit betrachtet werden (ditto für deren Annahme).

- Eine falsch positive Entscheidung wird in der **statistischen Testtheorie** auch als α-Fehler oder **Fehler der ersten Art**, eine falsch negative Entscheidung als β-Fehler oder **Fehler der zweiten Art**, und die Sensitivität

als **Macht** (power) des Tests bezeichnet. α bzw. β bezeichnen die korrespondierenden bedingten Wahrscheinlichkeiten. Hier noch einmal die Entsprechungen zu den oben eingeführten Bezeichnungen.

- $P(+|H_0)$: fpr, α, Signifikanzniveau (siehe unten)
- $P(-|H_0)$: tnr, $1 - \alpha$
- $P(+|H_1)$: tpr, $1 - \beta$, Macht
- $P(-|H_1)$: fnr, β

- **Beispiel: Münzwurf**

  Im nachfolgenden Beispiel wird die Statistik *"Anzahl von Kopf in 100 Münzwürfen"* herangezogen, um zu entscheiden, ob eine Münze fair ist. Tatsächlich ist das betrachtete Merkmal binomialverteilt $B(100, 0.5)$, wir nutzen hier jedoch die Normalverteilungs-Approximation des Anteils mit Standardfehler $se = \sqrt{p(1-p)/100} = 0.5/10 = 0.05$.

  Wir nehmen an, dass der Anteil für die H_0 und H_1 jeweils normalverteilt mit Mittel $\mu_0 = 0.5$ bzw. $\mu_1 = 0.6$ und identischer Varianz $\sigma^2 = 0.05^2$ ist. Die H_0 wird akzeptiert, wenn die Ausprägung von X in ein symmetrisches Intervall $0.5 \pm 1.96 * 0.05$ fällt, welches die Verteilung von H_0 zu 95 % abdeckt. Dies ist in Abb. 21 dargestellt.

  Fällt eine Ausprägung hingegen in den blau schraffierten Ablehnungsbereich, so wird die H_1 angenommen, und die H_0 abgelehnt; in der Statistik spricht man von einem (auf α-Niveau) **signifikanten** Ergebnis, α wird in diesem Kontext **Signifikanzniveau** genannt.

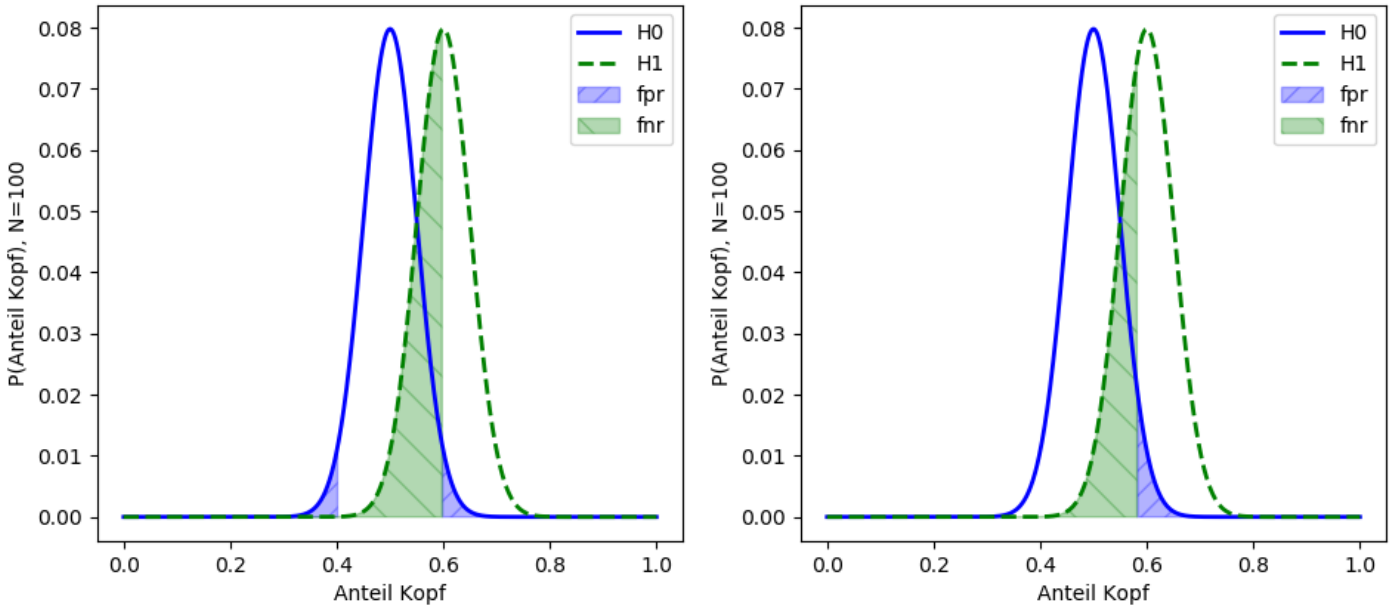


Abbildung 21: α (fpr) und β (fnr) für H_0 $X \sim N(0.5, 0.05^2)$ und H_1 $X \sim N(0.6, 0.05^2)$, für den zweiseitigen (links) und einseitigen (rechts) Fall. In beiden Fällen haben wir $\alpha = .05$, β ist jedoch für den einseitigen Test mit $\beta = 0.36$ geringer als für den zweiseitigen mit $\beta = 0.5$.

Man unterscheidet zwischen einseitigen (gerichteten) und zweiseitigen (ungerichteten) Hypothesentests. Ist man z.B. an einer Abweichung nach oben interessiert, so wird man die H_0 nur dann ablehnen, wenn die Teststatistik einen zu großen Wert annimmt. Einseitige Tests akzeptieren bei gleichem α eher die H_1 bei Beobachtungen, die in Richtung der H_1 liegen. Im obigen Beispiel würde ein einseitiger Test die Beobachtung 0.59 auf dem 0.05-Niveau als signifikant akzeptieren, ein zweiseitiger Test hingegen nicht. (Tatsächlich würden für $\mu = 0.6$ ein zweiseitiger Test auf dem 0.05-Niveau und ein einseitiger Test auf dem 0.025-Niveau fast identische Ergebnisse liefern, da Schätzwerte im linken blauen Schwanz der H_0 extrem unwahrscheinlich sind.)

Wir nehmen im folgenden den zweiseitigen Fall an.

Die fpr α sagt, wie wahrscheinlich es ist, dass eine faire Münze als unfair erkannt wird (hier: 0.05). Die fnr β hingegen sagt, wie wahrscheinlich es ist, dass ein tatsächlicher Unterschied nicht erkannt wird. Dies hängt u.a. von der Größe des in der H_1 postulierten Unterschieds zur H_0 ab: je größer der zu erkennende Unterschied, desto geringer β. Dieser Zusammenhang ist in Abb. 22 dargestellt . Die blaue Kurve entspricht dem in Abb. 21 dargestellten Fall; für $N = 100$ wird für $\mu_1 = 0.6$ eine unfaire Münze nur in 50% der Fälle detektiert.

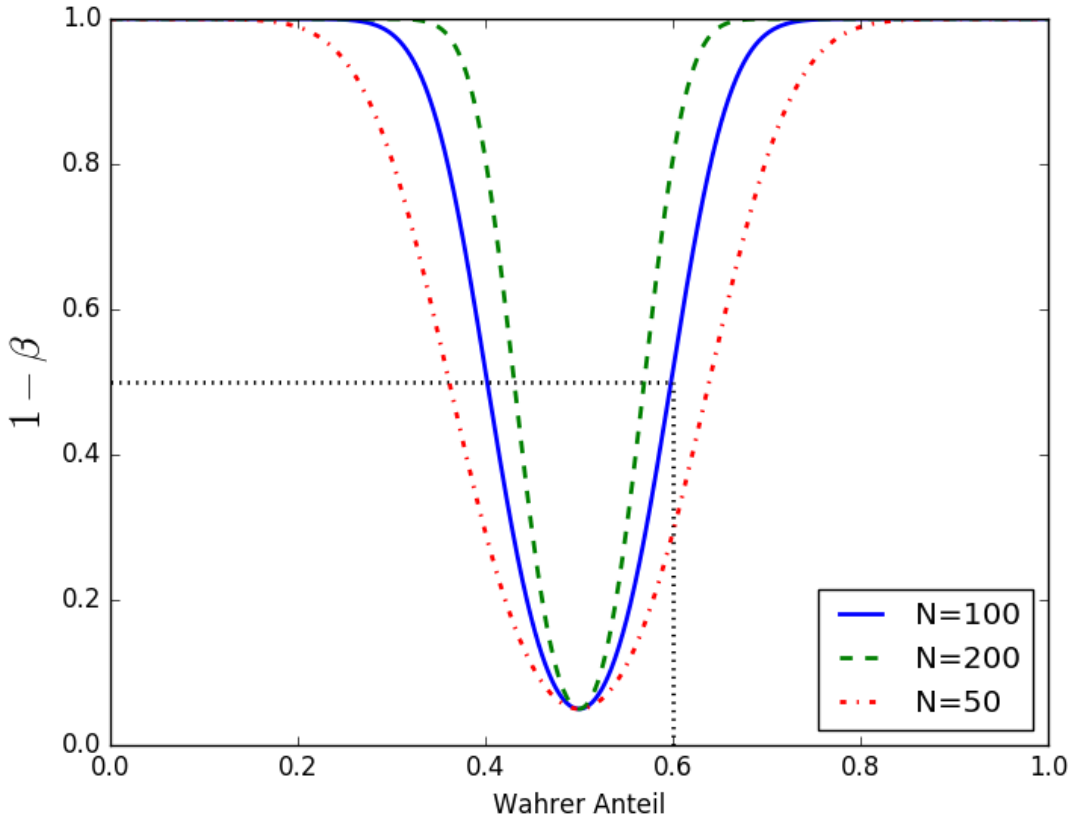


Abbildung 22: Sensitivität $1 - \beta$ (Macht) als Funktion des wahren Mittels μ_1 für unterschiedliche Stichprobengrößen (Standardfehler).

Weitere Möglichkeiten, β zu verringern (also die Sensitivität zu vergrößern), bestehen darin ein größeres α zuzulassen, oder den Standardfehler zu verringern (z.B. durch Vergrößerung des Stichprobenumfangs). Die Sensitivität $1 - \beta$ dargestellt als Funktion von α wird als ROC-Kurve bezeichnet (siehe auch Anhang A).

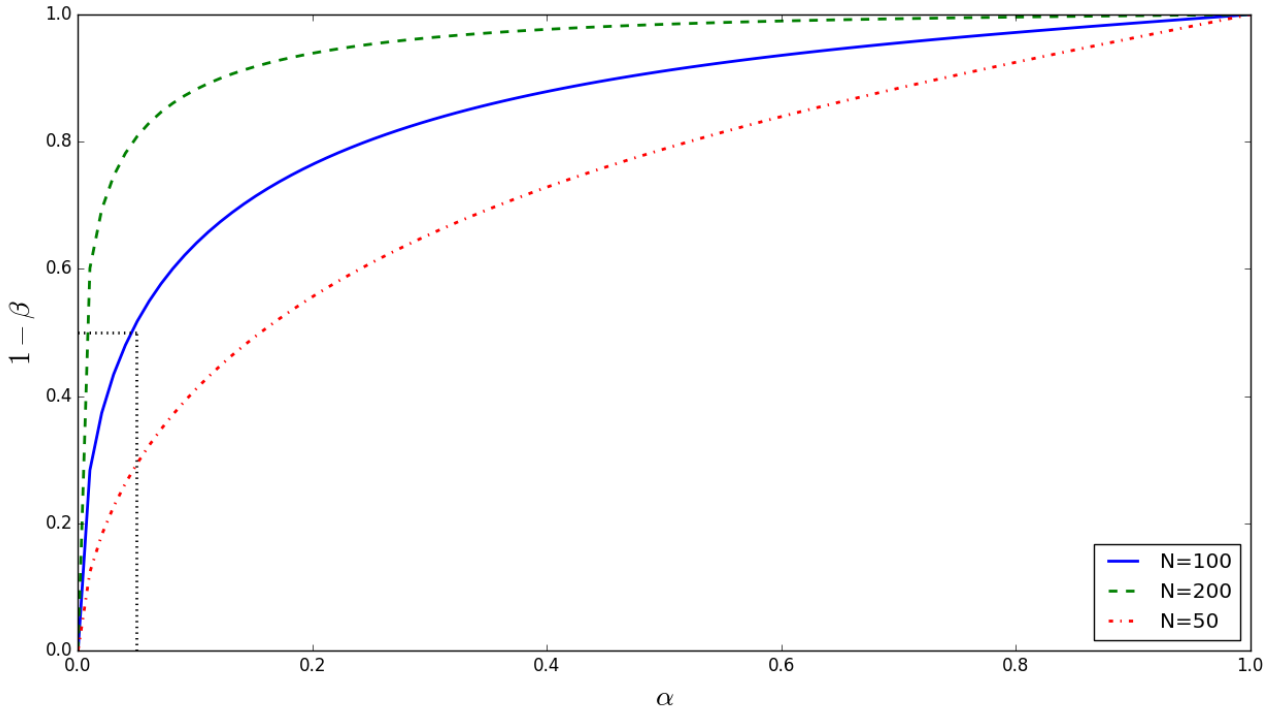


Abbildung 23: ROC-Kurven für $\mu_1 = 0.6$ und unterschiedliche Stichprobengrößen. Man sieht wiederum, dass für $N = 100$ mit einer fpr von $\alpha = 0.05$ eine Sensitivität von $1 - \beta = 0.5$ erzielt werden kann.

- **Hypothesentest am Beispiel des Stichprobenmittels**

  Angenommen, wir sind daran interessiert, ob eine neue Behandlungsmethode H_1 deutlich andere Ergebnisse als eine etablierte Methode H_0 mit Populationsmittel μ_0 liefert. Sei $\hat{m}$ sei das Mittel einer Stichprobe vom Umfang N, und wir möchten nun überprüfen, ob dieses mit der gegebenen Null-Hypothese $H_0 : \mu = \mu_0$ kompatibel ist. Wir betrachten dazu die hypothetische Verteilung des Schätzers unter H_0

$$\hat{\mu}|H_0 \sim N(\mu_0, \frac{\sigma^2}{N}) \tag{120}$$

  Ist für eine gegebene Stichprobe die Abweichung zwischen $\hat{m}$ und μ_0 zu groß, so wird man die H_0 nicht mehr akzeptieren. Setzen wir z.B. in Eq. 116 $\mu = \mu_0$, so liegt das Stichprobenmittel mit 95%iger Wahrschein-

lichkeit innerhalb des Intervalls

$$[\mu_0 - z_{0.975}\frac{\sigma}{\sqrt{N}}, \mu_0 + z_{0.975}\frac{\sigma}{\sqrt{N}}] \tag{121}$$

Angenommen, die H_0 ist wahr. Kommt $\hat{m}$ außerhalb des Intervalls Eq. 116 zu liegen, verwerfen wir die H_0, obwohl sie wahr ist: wir begehen einen **Fehler erster Art**. Die Wahrscheinlichkeit, dass dies geschieht, ist im obigen Beispiel $\alpha = 0.05$. Ein Testergebnis, welches in den Ablehnungsbereich der H_0 fällt, wird als **signifikant** bezeichnet, α als **Signifikanz-Niveau** (*significance level*) des Tests. Ein Quantil (Intervallgrenze), dessen Unter- bzw. Überschreiten zu einem Verwerfen der H_0 führt, bezeichnet man auch als **kritischen Wert**, das/die korrespondierenden Intervalle als **kritischen Bereich** (die Bereiche unterhalb der blau schraffierten Flächen in Abb. 21).

Für den oben formulierten Test gilt, dass für gegebenes Stichprobenmittel $\hat{m}$ alle Null-Hypothesen μ_0 auf dem $\alpha = 0.05$-Niveau akzeptiert werden, für welche μ_0 innerhalb des korrespondierenden Schätzintervalls liegt, im konkreten Fall:

$$\begin{aligned} \mu_0 \quad &\in \quad [\hat{m} - z_{0.975}\frac{\sigma}{\sqrt{N}}, \hat{m} + z_{0.975}\frac{\sigma}{\sqrt{N}}] \\ &\Longleftrightarrow \\ \hat{m} \quad &\in \quad [\mu_0 - z_{0.975}\frac{\sigma}{\sqrt{N}}, \mu_0 + z_{0.975}\frac{\sigma}{\sqrt{N}}]. \end{aligned}$$

Für den einseitigen Test verhält es sich ganz analog.

Konfidenzintervalle und Hypothesentests sind also eng miteinander verwandt, sie drücken beide die normierte Differenz zwischen hypothetischem Wert und Schätzwert jeweils relativ zu einer der beiden Größen aus: allgemein lässt sich ein $(1 - \alpha)$-Schätzintervall für den Parameter

θ bei gegebener Realisierung des Schätzers $\hat{t}$ als Menge jener hypothetischen Parameterwerte θ_0 auffassen, welche bei einem Hypothesentest nicht zu einem auf dem α-Niveau signifikanten Testergebnis führen.

Man beachte, dass die Aussage *"Das Ergebnis ist auf 0.05-Nivau signifikant."* deutlich weniger Information liefert als die Angabe des Schätzwertes plus Standardfehler und diese wiederum weniger als das korrespondierenden Schätzintervall (da dieses zusätzlich die Form der Verteilung berücksichtigt); letztere sind daher einem Signifikanztest stets vorzuziehen!

- **Likelihood Ratio**

  Sei $p_0(x) = p(x|H_0)$ die DF des Merkmals unter der H_0, sowie $p_1(x) = p(x|H_1)$ die DF des Merkmals unter der spezifischen H_1, z.B. die im obigen Münzwurfbeispiel gegebenen Glockenkurven. Für festes x und als Funktion der Klassenzugehörigkeit H_i betrachtet, bezeichnet man die DF als likelihood.

  Es ist nun naheliegend, sich bei gegebener Merkmalsausprägung x für jene Klasse mit größerer likelihood zu entscheiden, wobei aber ggf. Nebenbedingungen zu beachten sind.

Fordern wir z.B. eine maximale fpr von α, so entscheiden wir uns – gegeben die Beobachtung x – für die H_1 falls für den Quotienten der likelihoods – die sogenannte **likelihood ratio** (LR) – gilt:

$$LR(x) = \frac{p_1(x)}{p_0(x)} \geq \gamma, \quad \text{wobei} \tag{122}$$

$$P\left(LR(x) \geq \gamma | H_0\right) = \alpha \tag{123}$$

Es lässt sich zeigen, dass der obige Test von allen Tests mit $fpr = \alpha$ die größte Macht (Sensitivität) hat. Dieser Sachverhalt ist als **Neyman-Pearson-Lemma** bekannt.

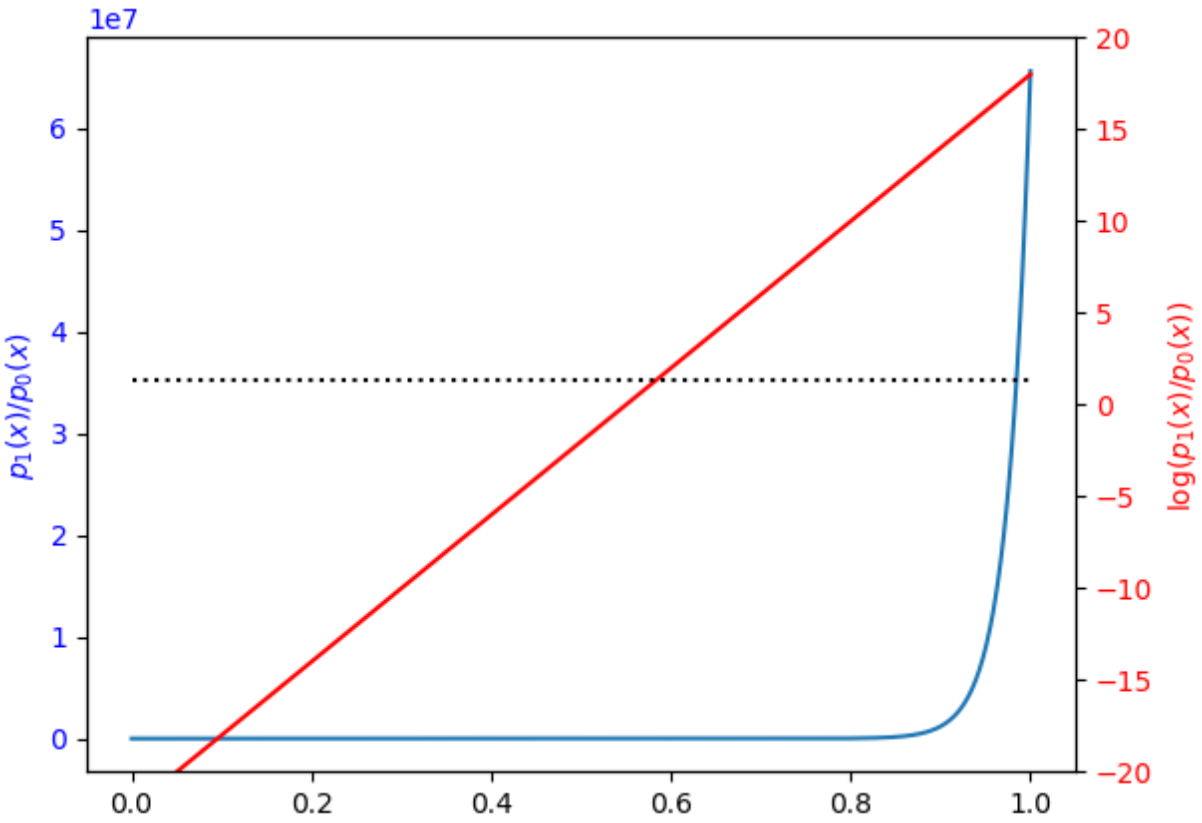


Abbildung 24: Likelihood ratio LR für das Münzwurf-Beispiel. Der Logarithmus der LR ist rot dargestellt, der Logarithmus des Schwellwerts $\gamma(\alpha = 0.05)$ schwarz gestrichelt.

In Abb. 24 ist die LR für unser Münzwurfbeispiel mit $\alpha = 0.05$ dargestellt. Die optimale Entscheidungsregion entspricht allen Punkten, für welche die logarithmierte LR (rote Kurve) über dem Schwellwert (der schwarz gestrichelten Linie) liegt, d.h. alle Punkte die rechts vom Schnittpunkt der schwarz gestrichelten und der roten Kurve liegen. Dies entspricht dem einseitigen Test auf 0.05-Niveau.

- **p-Wert** (p-value)

  Unter dem p-Wert eines Ereignisses (Merkmalsausprägung) x versteht man die Wahrscheinlichkeit, dass **unter Voraussetzung der** H_0 **!** x oder ein extremerer Wert (extrem im Sinne von: nicht mit der H_0 vereinbar) beobachtet wird. Der p-Wert darf nicht mit dem Signifikanz-Niveau α verwechselt werden: α (fpr) ist eine Eigenschaft des Tests, der p-Wert hingegen eine Eigenschaft einer konkreten Messung bzw. Stichprobe. Beobachtungen, deren p-Wert kleiner als das vorab gewählte Signifikanzniveau α ist, gelten als signifikant, d.h, die H_1 wird akzeptiert.

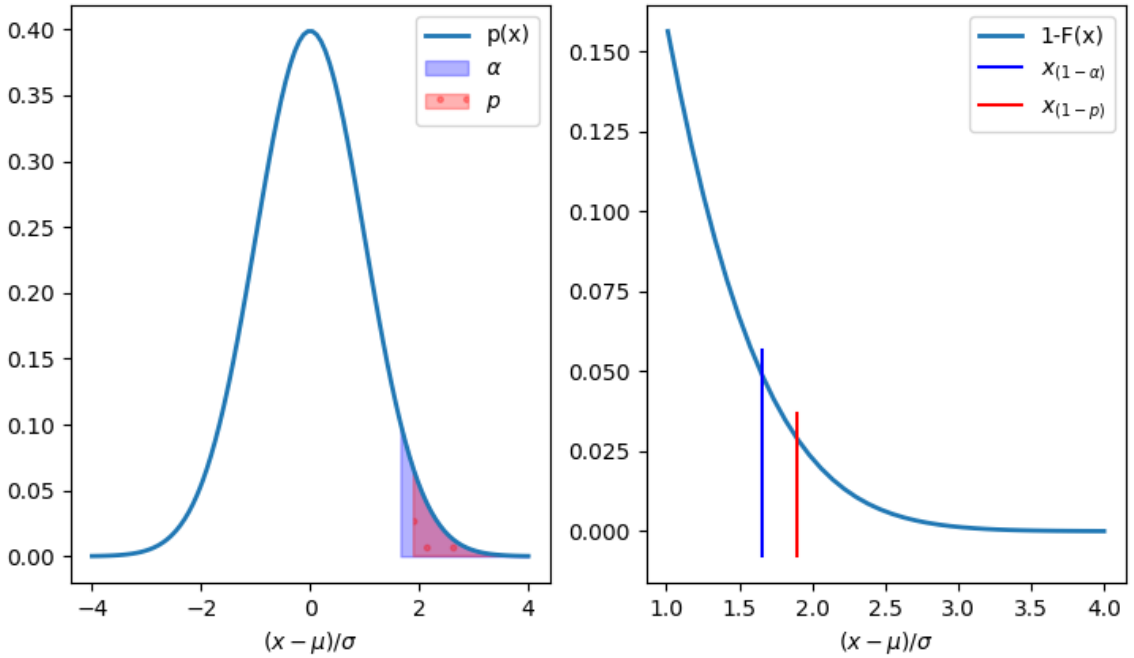


Abbildung 25: Links: Rechte Schwanzflächen der Standardnormalverteilung mit Flächeninhalt $\alpha = 0.05$ und $p = 0.03$. Die kleinere (rote) Fläche überdeckt die größere (blaue), ihr Durchschnitt erscheint magenta. Rechts: (x_α, α) und (x_p, p) als Punkte auf der Kurve $1 - F(x)$.

Beispiel: Wir betrachten die Verbesserung der Durchschnittsnoten $\bar{X}$ einer nach einer neuen Methode unterrichteten Schulklasse im Vergleich zum nationalen Durchschnitt μ_0. $\bar{X}$ sei unter der H_0: *" Es besteht kein Unterschied in den Leistungen"* normalverteilt mit $N(\mu_0, se^2)$. Wir wählen einen einseitigen Test mit 0.05-Signifikanzniveau und H_1: "Die neue Methode ist besser". Eine beobachtete Differenz von $x = 1.88 * se$ entspricht einem p-Wert von $1-F(1.88) = 0.03$, was kleiner als das vorab gewählte Signifikanz-Niveau 0.05 ist und somit zu einem signifikanten Test-Ergebnis führt (H_1 wird akzeptiert). Bezogen auf Abb. 25 bedeutet ein signifikantes Ergebnis, dass die rechts von x liegende Fläche der DF kleiner als 0.05 sein muß, bzw. das $(1-p)$-Quantil größer als das $(1-\alpha)$-Quantil sein muß.

Einseitige vs. zweiseitige p-Werte

Man beachte, dass sich im Falle eines zweiseitigen Tests die Wahrscheinlichkeit einer extremeren Beobachtung mit $F(1.88)+(1-F(1.88))$ ergibt: der p-Wert wäre somit mit 0.06 doppelt so groß wie im einseitigen Fall und das Ergebnis daher auf dem $\alpha = 0.05$-Niveau nicht signifikant. Allgemein folgt daraus: gibt eine Statistik-Routine einen zweiseitigen p-Wert zurück, so lässt sich, wenn die DF der H_0 symmetrisch ist, ein korrespondierender einseitiger p-Wert einfach durch Halbierung des zweiseitigen p-Werts berechnen.

Das Bayes-Theorem als Fundament der Bayes-Statistik

Das Bayes-Theorem (Bayes-Regel) erlaubt es, die bedingte Wahrscheinlichkeit $P(B|A)$ als Funktion der Randverteilungen $P(A), P(B)$ und der bedingten Wahrscheinlichkeit $P(A|B)$ auszudrücken:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}. \tag{124}$$

$P(B)$ . . . a priori Wahrscheinlichkeit (*prior*) von B

$P(B|A)$ . . . a posteriori Wahrscheinlichkeit (*posterior*) von B unter A

- Formal erhält man das Bayes-Theorem sehr einfach aus der Produktregel der Wahrscheinlichkeitsrechnung.

- Die Bayes-Regel kann auf zwei Arten gelesen werden:
  - Als Umrechnung einer bedingten Wahrscheinlichkeit in deren "Konverse" , z.B. Sensitivität in positiven Vorhersagewert: $P(+|H_1) \mapsto P(H_1|+)$.
  - Wenn wir B als Elementarereignis und $P(\omega = B) = p_\omega(B)$ als *a priori* Wahrscheinlichkeitsfunktion einer Hypothese (oder eines Parameters) ω und A als Merkmalsausprägung bzw. Menge von Merkmalsausprägungen (z.B. Stichprobe) auffassen, definiert die Bayes-Regel eine Transformation in die *a posteriori* Wahrscheinlichkeitsfunktion des Parameters:
    $$p_\omega(.) \rightarrow p_\omega(.|A). \tag{125}$$
    Dies werden wir im folgenden ausführlicher behandeln.

- Auf konzeptioneller Ebene unterscheidet sich die Bayessche Auffassung von Wahrscheinlichkeit von der bisher diskutierten frequentistischen, indem
  - alle interessierenden Größen – also nicht nur die Merkmale, sondern auch die mittels Inferenz zu bestimmenden Größen (Klassenzugehörigkeit, Verteilungsparameter) – als Zufallsvariablen betrachtet werden
  - (a posteriori) Wahrscheinlichkeiten gemäß obiger Formel aus bereits bekannten (a priori) Wahrscheinlichkeiten hergeleitet werden.

Die a priori Wahrscheinlichkeit reflektiert vor der Durchführung des Zufallsexperiments bzw. Beobachtung der Merkmalsausprägungen bereits vorhandenes Wissen (z.B. Prävalenz einer Krankheit). Die Notwendigkeit, solche *priors* bereitzustellen, wird von frequentistischer Seite oft als subjektiv, nicht-objektiv udglm. bezeichnet. Tatsächlich stellen a priori Wahrscheinlichkeiten eine hervorragende Möglichkeit dar, vorhandenes Hintergrundwissen konsistent in die Inferenz einfließen zu lassen.[8] Ist dieses Wissen sehr vage, so kann dies z.B. durch Annahme einer Gleichverteilung oder Wahl einer (unendlich) großen Varianz modelliert werden (*non-informative prior*). Die Inferenz erfolgt formal jedoch immer gemäß Gl. 124.

[8] Dazu gehört auch, Präferenzen für bestimmte Lösungen quantitativ formulieren zu können, z.B. bei Regressionproblemen die Größe der Regressionskoeffizienten oder die Kurvatur der Regressionsfunktion möglichst klein zu halten. In anderen Worten: der Bayes-Ansatz bietet einen natürlichen Rahmen für die Behandlung von Regularisierungsproblemen.

Einführung in die Bayes-Klassifizierung

Repräsentiert X ein diskretes Merkmal und ω die Klassenzugehörigkeit von Mustern, so gibt im Falle der beobachteten Merkmalsausprägung $X = x$

$$P(\omega = j|X = x) = \frac{P(X = x|\omega = j)P(\omega = j)}{P(X = x)} \tag{126}$$

die *a posteriori* Wahrscheinlichkeit an, dass das Muster zur Klasse j gehört. Die Regel gibt formal an, wie bereits vorhandenes *a priori* Wissen über die Klassenwahrscheinlichkeiten (Prävalenzen) mit der Information über eine beobachtete Merkmalsausprägung zu verknüpfen ist.[9]

[9] Wir schreiben im folgenden, wie in der Literatur üblich, oft kurz ω_j für $\omega = j$, um anzuzeigen, dass die Zufallsvariable ω den Wert j annimmt.

- **Transformation der Klassen-Wahrscheinlichkeitsfunktion (WF), Bayes-Inferenz**

  Gl. 126

$$p_\omega(j|x) = P(\omega_j|X = x) = \frac{P(X = x|\omega_j)p_\omega(j)}{P(X = x)}$$

  lässt sich auch als Transformation der *a priori* WF auf die *a posteriori* WF der Klassenzugehörigkeiten lesen

$$p_\omega(.) \rightarrow p_\omega(.|x), \tag{127}$$

  z.B. für die Verteilungen in Abb. 26 (in Felddarstellung)

$$p_\omega(.) = [0.67, 0.33] \rightarrow p_\omega(.|x) = [0.55, 0.45]. \tag{128}$$

Im Englischen bezeichnet man die *a priori* und *a posteriori* Verteilung meist kurz als **prior** bzw. **posterior**.

Der auch als **evidence** bekannte Nenner berechnet sich gemäß der erweiterten Summenregel (für c Klassen) mit

$$p(x) = \sum_{j=1}^{c} P(X = x|\omega_j)P(\omega_j). \tag{129}$$

Dieser fungiert als Normierungsfaktor, der sicherstellt, dass die linken Seiten in Gl. 126 zu 1 summieren und $P_\omega(.|x)$ somit formal die Anforderungen an eine Wahrscheinlichkeitsfunktion erfüllt.

Man bemerkt, dass $p(x)$ selbst eine Wahrscheinlichkeitsfunktion WF ist. Es handelt sich dabei um eine sogenannte **mixture distribution**, eine mit den korrespondierenden *a priori* Wahrscheinlichkeiten gewichtete Summe der WFen der einzelnen Klassen.

- **Bayes-Theorem für stetige Merkmale**

  Wir nehmen im folgenden eine stetige Merkmalsvariable X mit zugeordneter pdf $p(x)$ an. Eq. 126 wird zu

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}. \tag{130}$$

  $p(x|\omega_j)$ beschreibt für festes j und als Funktion von x die DF des Merkmals X für eine gegebene Klasse ω_j und besitzt alle Eigenschaften einer "normalen" Dichtefunktion.

- **Bayes-Entscheidungsregel (Bayes Decision Rule)**

  Gegeben die Beobachtung (Merkmalsausprägung) $X = x$, entscheide für die Klasse k, welche die größte *a posteriori* Wahrscheinlichkeit aufweist:

$$k = \arg\max_j P(\omega_j | X = x). \tag{131}$$

  Dies ist ein diskreter Spezialfall des Bayesschen Pendants zu Maximum Likelihood, der sogenannten **Maximum A Posteriori (MAP)** Regel: wähle jenen Wert für den gesuchten Parameter (hier: Klassenzugehörigkeit) mit maximaler a posteriori Wahrscheinlichkeit.

  In Abb. 26 würde ML bei einer Merkmalsausprägung $x = 10$ für die Klasse zwei entscheiden (links), MAP (mit *a priori* Klassenwahrscheinlichkeiten $P(\omega_1) = 2/3$ und $P(\omega_2) = 1/3$) hingegen für Klasse eins (rechts).

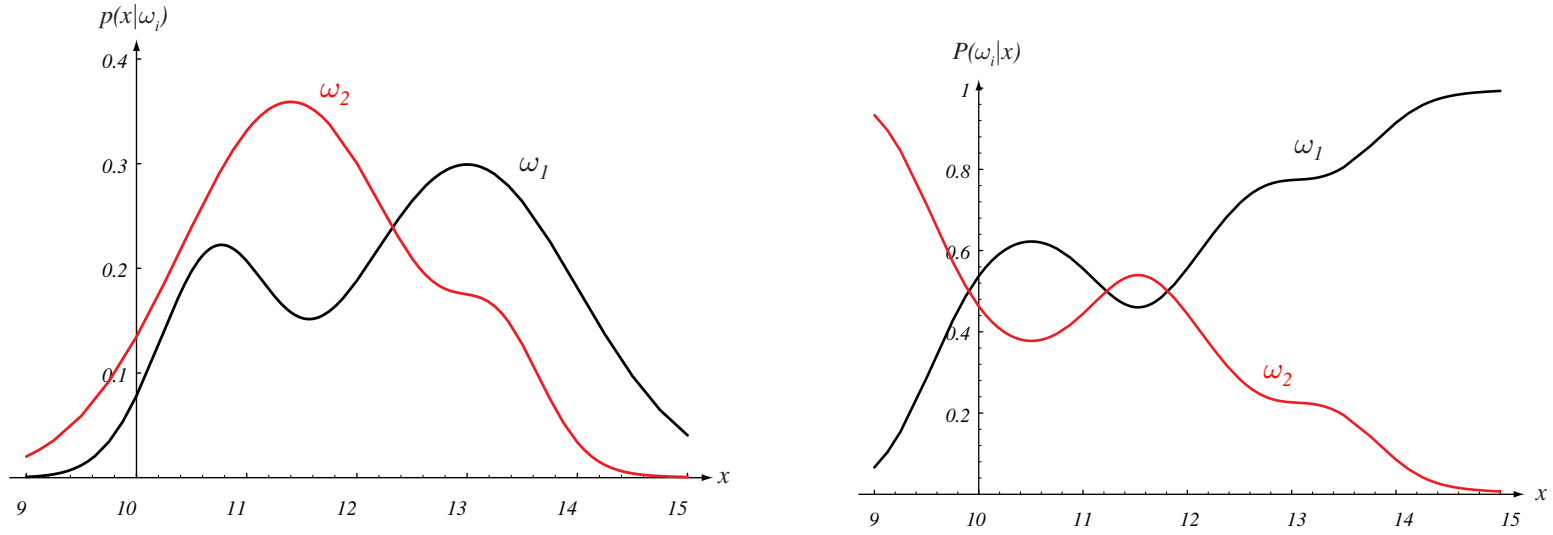


Abbildung 26: *DF bzw. likelihood der beiden Klassen* (links) und korrepondierende *posteriors* für $P(\omega_1) = 2/3$ und $P(\omega_2) = 1/3$ (rechts).
(Aus *Duda, Hart, Stork: Pattern Classification, 2nd ed.)*

Man bemerkt jedoch, dass die evidence $p(x)$ für alle Klassen identisch ist und daher keinen Einfluss auf das Verhältnis der *posteriors* hat. Für die Bestimmung der Klasse mit der größten *a posteriori* Wahrscheinlichkeit ist daher das Verhältnis der mit den korrespondierenden *priors* gewichteten *likelihoods* $p(x|\omega_i)P(\omega_i)$ hinreichend.

Im Fall der *Bayes rule* verschieben größere *priors* die Entscheidungsgrenze in Richtung der *a priori* weniger wahrscheinlichen Klasse.

Im Falle identischer *priors* $P(\omega_i) = P(\omega_j), 1 \leq i, j \leq c$ müssen nur die *likelihoods* berücksichtigt werden: ML und MAP liefern dann dasselbe Ergebnis.

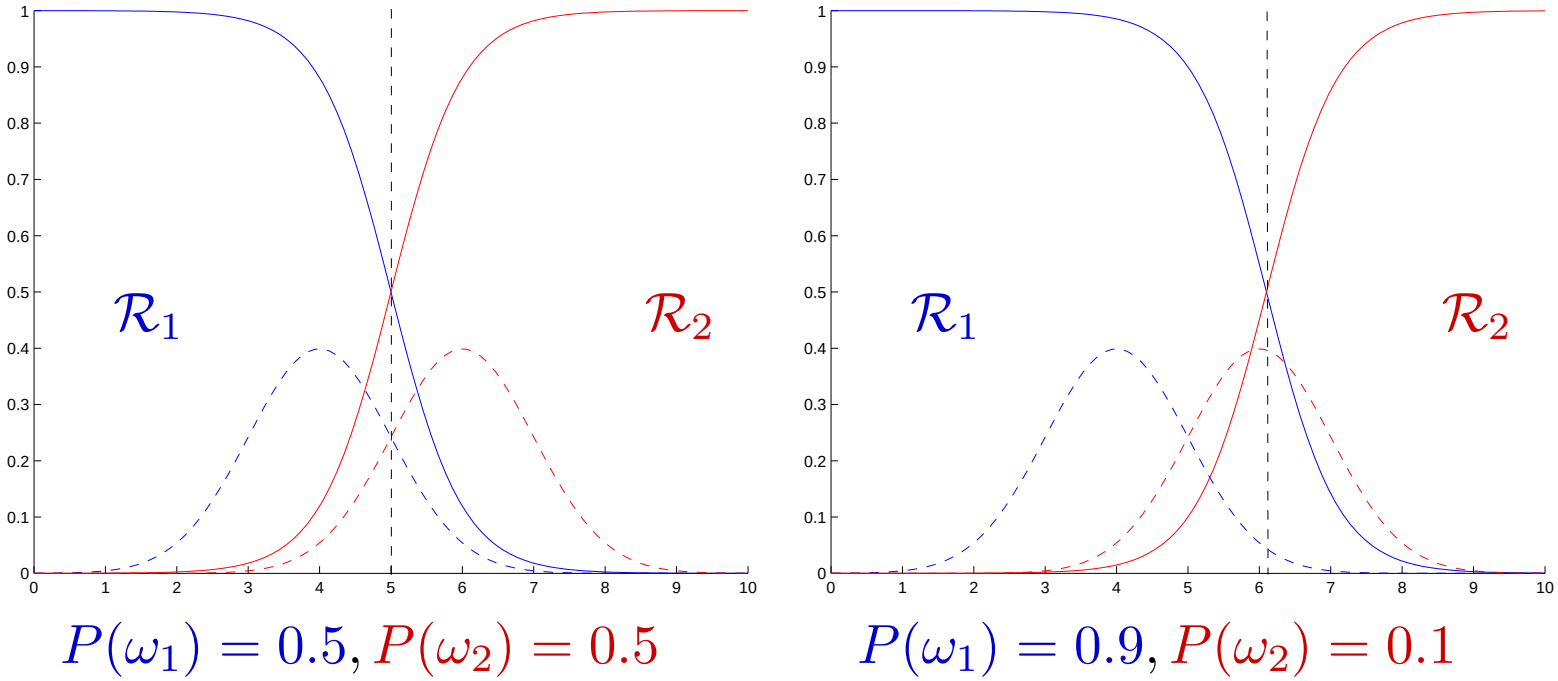


Abbildung 27: Bayes-optimale Entscheidungsgrenzen (schwarz gestrichelt) und korrespondierende Entscheidungsregionen für zwei Klassen ω_1 und ω_2 mit normalverteilten Merkmalen (Mittel $\mu_1 = 4, \mu_2 = 6$, Varianz $\sigma_1^2 = \sigma_2^2 = 1$). Die DF ist jeweils gestrichelt, der *posterior* durchgezogen dargestellt.

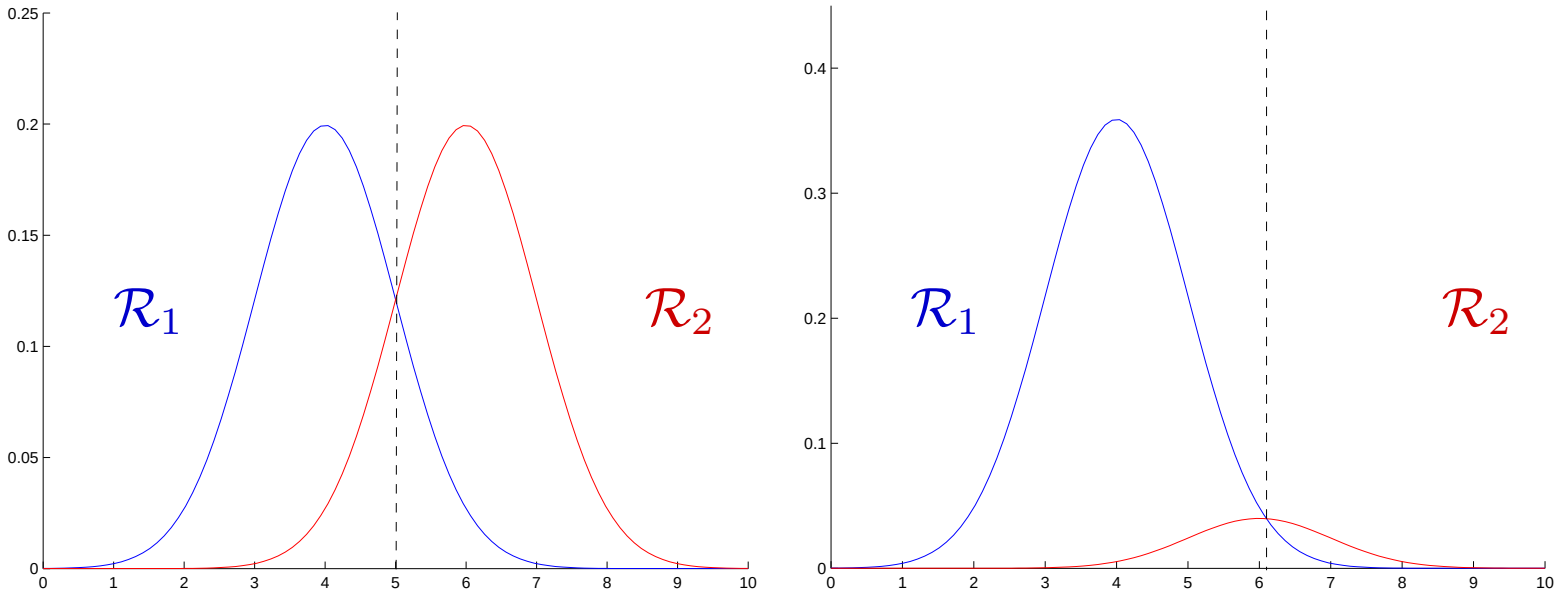


Abbildung 28: Bayes-optimale Entscheidungsgrenzen (schwarz gestrichelt) für die Klassen aus Abb. 27. Dargestellt ist der Verlauf der gewichteten DFen $p(x|\omega_1)P(\omega_1)$ und $p(x|\omega_2)P(\omega_2)$.

- **Likelihood Ratio**

  Die obigen Überlegungen führen für den Fall $c = 2$ zu folgender, äquivalenter Formulierung der *Bayes rule*:

  – Entscheide für ω_1, falls

$$\begin{aligned} P(\omega_1|x) &> P(\omega_2|x) \\ p(x|\omega_1)P(\omega_1) &> p(x|\omega_2)P(\omega_2) \\ \frac{p(x|\omega_1)}{p(x|\omega_2)} &> \frac{P(\omega_2)}{P(\omega_1)}. \end{aligned} \tag{132}$$

  Der Ausdruck $\frac{p(x|\omega_1)}{p(x|\omega_2)}$ wird als **likelihood ratio** bezeichnet, der Ausdruck $\frac{P(\omega_2)}{P(\omega_1)}$ als **threshold**. Übersteigt die *likelihood ratio* den *threshold*, entscheidet man für ω_1, sonst für ω_2.

- Beispiel: Test auf ESP(*Extrasensory Perception*)
  Experiment zu ESP von Soal, 1954. Beschrieben nach Jaynes, 2003.

  Mrs. Stewart führte eine Reihe von $N = 37100$ Experimenten durch, wobei jedes unter der H_0: " kein ESP " eine unabhängige Erfolgswahrscheinlichkeit von $\theta = 0.2$ aufwies. Tatsächlich war Mrs. Stewart in $r = 9420$ Fällen erfolgreich, dies entspricht einem relativen Anteil von $\hat{t} = 0.2536$ bzw. einer Differenz von fast 26 Standardfehlern der Binomialverteilung $B(37100, 0.2)$. Verwenden wir als H_1 " Mrs. Stewart ist telepathisch $\theta > 0.2$ mit spezifischem $\theta_1 = \hat{t}$, so haben wir für die likelihood ratio:

$$\frac{P(\mathcal{D}|H_0)}{P(\mathcal{D}|H_1)} = \frac{B(9420|37100, 0.2)}{B(9420|37100, 0.2536)} = \frac{3.15\ 10^{-139}}{0.00476} \tag{133}$$

  Hat Mrs. Stewart telepathische Kräfte?

Fehlerwahrscheinlichkeit

Laut Bayes-Theorem Eq. 130 ergibt sich im Falle eines binären Klassifikationsproblems für jede Merkmalsausprägung x die bedingte Wahrscheinlichkeit der Fehlklassifikation (**conditional error**) $P(error|x)$ zu

- $P(\omega_2|x)$, falls wir für ω_1 entscheiden
- $P(\omega_1|x)$, falls wir für ω_2 entscheiden.

Der mittlere Fehler $P(error)$, die **error rate** (Fehlerrate), berechnet sich gemäß Eq. 56 als

$$P(error) = \int_{-\infty}^{+\infty} P(error|x)p(x)dx. \tag{134}$$

- **Optimalität der Bayes Decision Rule**

  Die *Bayes Decision Rule* entscheidet für die Klasse ω_k mit der höchsten *a posteriori* Wahrscheinlichkeit

$$k = \arg\max_j P(\omega_j|x). \tag{135}$$

  Daher ergibt sich die bedingte Fehlerwahrscheinlichkeit $P(error|x)$ zu

$$\min[P(\omega_1|x), P(\omega_2|x)] = 1 - \max[P(\omega_1|x), P(\omega_2|x)]. \tag{136}$$

  Die *Bayes Rule* minimiert also den Integranden $P(error|x)$ in Eq. 134 für jede Merkmalsausprägung x, und folglich auch die mittlere Fehlerwahrscheinlichkeit $P(error)$.

  Die unter Verwendung der *Bayes rule* erzielte mittlere Fehlerwahrscheinlichkeit wird auch als *Bayes error rate* bezeichnet.

- Der allgemeine Fall: $c \geq 2$

  Entscheidet man sich im Punkt x für die Klasse ω_i, so ergibt sich die bedingte Fehlerwahrscheinlichkeit im allgemeinen Fall zu

$$P(error|x) = \sum_{j \neq i} P(\omega_j|x) = 1 - P(\omega_i|x), \tag{137}$$

  bzw. unter der *Bayes decision rule* zu

$$P(error|x) = 1 - \max_j P(\omega_j|x). \tag{138}$$

- Im Fall $c = 2$ lässt sich für eine gegebene Entscheidungsfunktion $\alpha(x)$ die Fehlerrate (*error rate*) Eq. 134 auch folgendermaßen formulieren

$$\begin{aligned} P(error) &= \int_{-\infty}^{+\infty} P(error|x)p(x)dx = \\ \int_{\mathcal{R}_1} P(\omega_2|x)p(x)dx &+ \int_{\mathcal{R}_2} P(\omega_1|x)p(x)dx = \qquad (139) \\ \int_{\mathcal{R}_1} P(\omega_2)p(x|\omega_2)dx &+ \int_{\mathcal{R}_2} P(\omega_1)p(x|\omega_1)dx = \qquad (140) \\ P(\omega_2)\varepsilon_2 &+ P(\omega_1)\varepsilon_1. \qquad (141) \end{aligned}$$

Hierbei gibt ε_j die Wahrscheinlichkeit an, dass ein Muster aus Klasse ω_j von $\alpha(x)$ falsch klassifiziert wird (d.h. in eine Entscheidungs-Region $\mathcal{R}_i$ mit $i \neq j$ fällt). Die Fehlerrate ergibt sich als mit den korrespondierenden *priors* gewichtetes Mittel der ε_i.

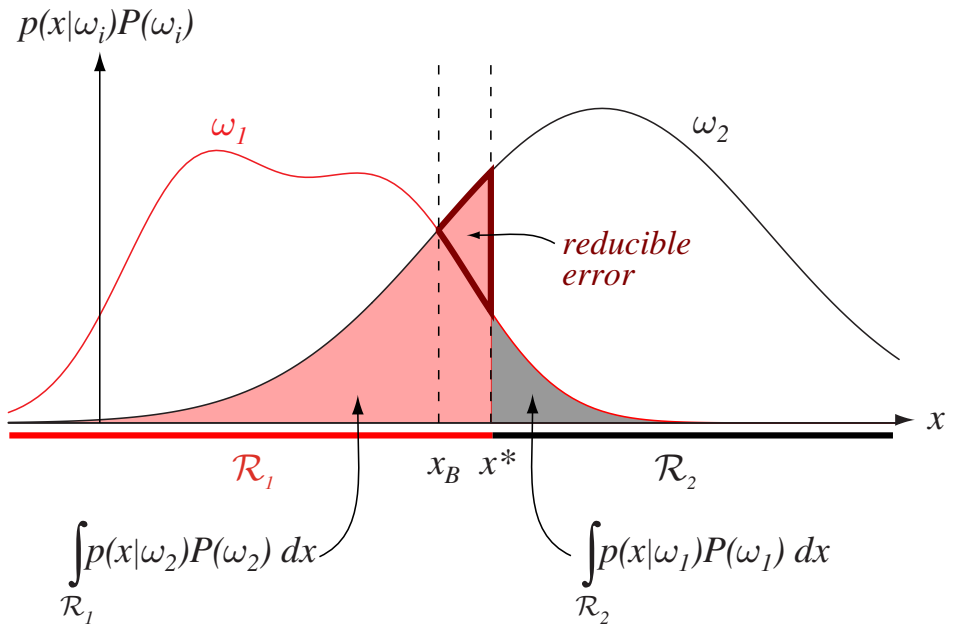


Abbildung 29: Die beiden Komponenten der Fehlerrate $P(\omega_1)\varepsilon_1$ (grau) und $P(\omega_2)\varepsilon_2$ (rosa) für zwei Entscheidungsgrenzen: die optimale Grenze x_B und eine nicht-optimale Grenze x^*. Die nichtoptimale Enscheidungsgrenze führt zu einer um den rot umrandeten Bereich (*reducible error*) größeren Fehlerrate. (Aus *Duda, Hart, Stork: Pattern Classification, 2nd ed.)*

Entscheidungstheorie *(Decision Theory)*

Entscheidungen sind mit Konsequenzen verbunden. Ob eine Entscheidung richtig oder falsch war, lässt sich i.a. erst im nachhinein beurteilen. Man kann aber versuchen, die Konsequenzen, die eine Entscheidung haben kann, quantitativ zu beschreiben, indem man diesen – vom wahren Zustand der Welt zum Zeitpunkt der Entscheidung abhängige – Konsequenzen Kosten zuweist.

Die Herleitung von geeigneten Kostenfunktionen und die Begründung von Entscheidungen im Rahmen eines quantitativen Rahmenwerks ist Gegenstand der **Entscheidungstheorie**. Wir diskutieren hier nur einige grundlegende, im Rahmen der Mustererkennung benötigte Konzepte, so dass im folgenden "Entscheidung" mit "Klassierung" gleichgesetzt werden kann.

- **Loss Function** $L(\alpha(x), j)$

  Die *loss function* (kurz: *loss*) gibt die mit der Entscheidung $\alpha(x)$ verbundenen Kosten (*cost*) an, wenn die wahre Klassenzugehörigkeit durch $\omega = j$ gegeben ist. Meistens findet der sogenannte *0/1-loss* Anwendung

$$L(\alpha(x), j) = 1 - \delta_{\alpha(x),j} = \begin{cases} 1 \ if \ \alpha(x) \neq j \\ 0 \ if \ \alpha(x) = j. \end{cases} \tag{142}$$

- **Total Risk**

  Der durch die Entscheidungsfunktion α verursachte mittlere *loss* berechnet sich als

$$R(\alpha) = \sum_{j=1}^{c} \int L(\alpha(x), j)\, p(j, x)\, dx, \tag{143}$$

  wobei $p(j, x)$ die Verbunddichte bzg. der Mermalsvariable X und Klassenvariable ω darstellt. Für obige Größe gibt es in der Literatur verschiedene Bezeichnungen; in dieser LVA nennen wir sie **total risk**. $R(\alpha)$ lässt sich mittels der Produktregel auf unterschiedliche Weise formulieren.

$$R(\alpha) = \sum_{j=1}^{c} \int L(\alpha(x), j)\, p(x|j)\, dx\, P(\omega_j) = \sum_{j=1}^{c} R(j)\, P(\omega_j) \qquad (144)$$

Hier berechnet man zunächst für jede Klasse (Parameter) $\omega = j$ den Erwartungswert $R(j)$ über alle Merkmalsausprägungen x und mittelt dann über die a priori-Verteilung der Klassen:

$$R(j) = \int L(\alpha(x), j)\, p(x|\omega_j)\, dx. \qquad (145)$$

$R(j)$ – als Erwartungswert über X für gegebenen Parameter j – ist das (frequentistische) **risk**. Obige Berechnung von $R(\alpha)$ entspricht der Berechnung der Bayes-Fehlerrate gemäß (141).

$$R(\alpha) = \int \sum_{j=1}^{c} L(\alpha(x), j)\, P(\omega_j|x)\, p(x)\, dx = \int R(\alpha(x)|x)\; p(x)\, dx \tag{146}$$

Den für eine gegebene Merkmalsausprägung x erwarteten *loss* bzg. der a posteriori Klassenzugehörigkeit $P(\omega_j|x)$

$$R(\alpha(x)|x) = \mathcal{E}[L(\alpha(x), j)] = \sum_{j=1}^{c} L(\alpha(x), j) P(\omega_j|x). \tag{147}$$

bezeichnen wir im folgenden als **conditional risk**; auch dies ist allerdings keine in der Literatur übliche Bezeichnung.

Analog zur *Bayes rule* lässt sich das *total risk* R minimieren, indem man das *conditional risk* $R(\alpha(x)|x)$ in jedem Punkt x minimiert.

Um die optimale Entscheidung im Punkt x zu bestimmen, werten wir das *conditional risk* für alle Klassenzugehörigkeiten $1 \leq i \leq c$ aus:

$$R(i|x) = \sum_{j=1}^{c} L_{ij} P(\omega_j|x). \tag{148}$$

Für *0/1-loss* gilt $L_{ij} = 1 - \delta_{ij}$, sodass

$$R(i|x) = \sum_{j \neq i} P(\omega_j|x) = 1 - P(\omega_i|x). \tag{149}$$

Das *conditional risk* $R(i|x)$ unter *0/1-loss* (Eq. 149) ist also identisch mit dem *conditional error* $P(error|x)$ (Eq. 137).

$R(\alpha(x)|x)$ wird in jedem Punkt x minimal, wenn $\alpha(x)$ die *Bayes decision rule* implementiert, d.h. das Label der Klasse mit der größten *a posteriori*

Wahrscheinlichkeit zurückliefert

$$\alpha(x) = \arg\max_j P(\omega_j|x). \tag{150}$$

- **Asymmetrischer Loss**

  Der *0/1-loss* wird häufig auch als *symmetrical loss* bezeichnet. Eine asymmetrische *loss*-Funktion kann verwendet werden, um die Fehlklassifikation von verschiedenen Klassen unterschiedlich stark zu "bestrafen". Achtung: das *total risk* kann jedoch nur unter *0/1-loss* als Fehlerrate, d.h. als mittlere Fehlerwahrscheinlichkeit interpretiert wird.

Beispiel: Früherkennung von Krankheiten

Sei X ein Merkmal, welches verwendet wird, um gesunde (ω_1) von potentiell kranken (ω_2) Patienten zu unterscheiden; in diesem Fall ist es "kostspieliger", einen kranken Patienten als gesund zu klassieren als einen gesunden Patienten als krank.

Schreiben wir Eq. 148 für die beiden möglichen Entscheidungen $\alpha(x) = 1$ und $\alpha(x) = 2$ explizit aus, so erhalten wir

$$\begin{aligned} R(1|x) &= L_{11}P(\omega_1|x) + L_{12}P(\omega_2|x) \\ R(2|x) &= L_{21}P(\omega_1|x) + L_{22}P(\omega_2|x). \end{aligned} \tag{151}$$

In unserem Beispiel sollte klarerweise $L_{12} > L_{21}$ gelten.

Um das *conditional Risk* im Punkt x zu minimieren, entscheiden wir für ω_1, falls

$$\begin{aligned} R(2|x) &> R(1|x) \\ L_{21}P(\omega_1|x) + L_{22}P(\omega_2|x) &> L_{11}P(\omega_1|x) + L_{12}P(\omega_2|x) \\ (L_{21} - L_{11})P(\omega_1|x) &> (L_{12} - L_{22})P(\omega_2|x) \\ (L_{21} - L_{11})P(\omega_1)p(x|\omega_1) &> (L_{12} - L_{22})P(\omega_2)p(x|\omega_2). \quad (152) \end{aligned}$$

Man sieht, dass der *loss* effektiv die *priors* neu gewichtet und somit die Entscheidungsgrenze von der stärker gewichteteten Klasse weg verschiebt.

Um die Diskussion zu vereinfachen, nehmen im folgenden $L_{11} = L_{22} = 0$ an.

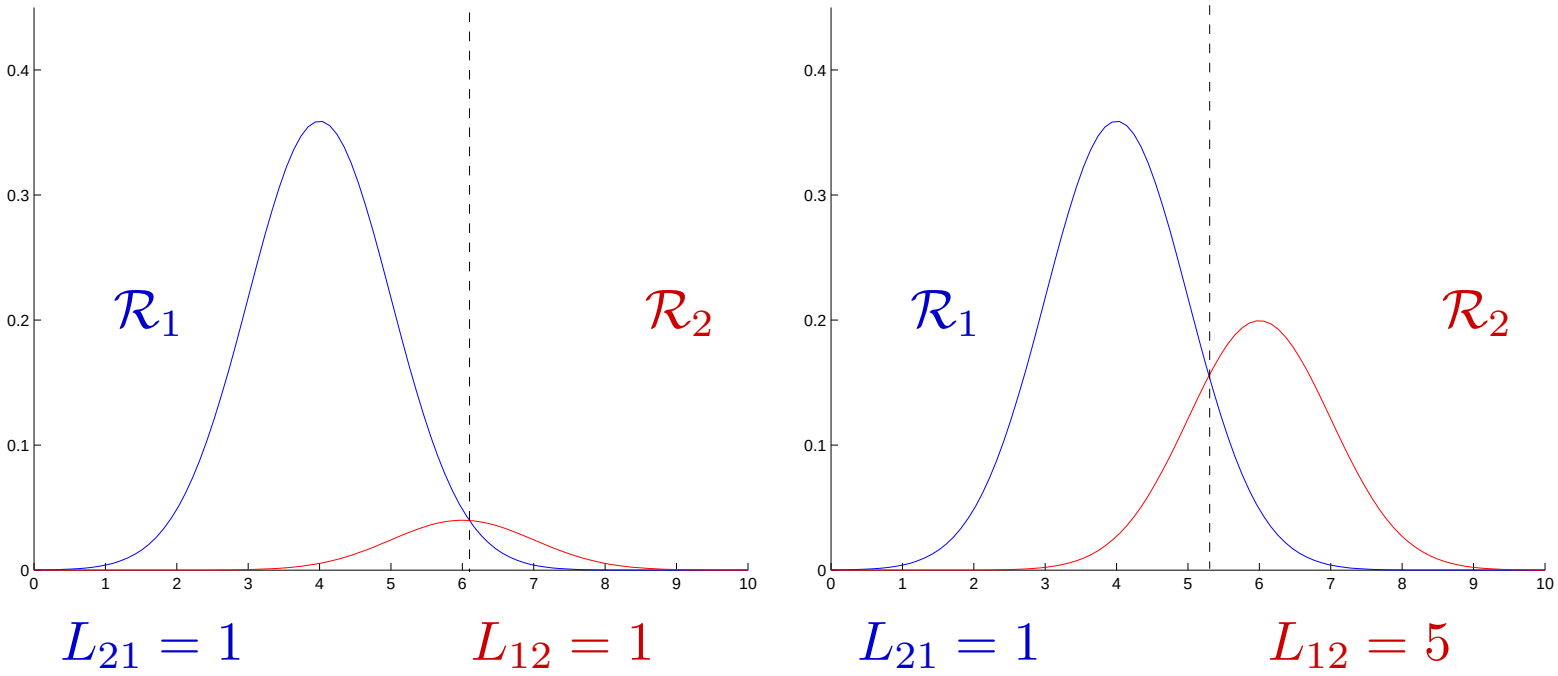


Abbildung 30: *Minimum risk decision boundaries* für die Klassen aus Abb. 27 mit *priors* $P(\omega_1) = 0.9$ und $P(\omega_2) = 0.1$. Dargestellt sind die Funktionen $p(x|\omega_1)L_{21}P(\omega_1)$ und $p(x|\omega_2)L_{12}P(\omega_2)$.
Für *0/1-loss* (links) sind *risk minimization* und *minimum error rate classification* äquivalent. Für $L_{12} > L_{21}$ (rechts) verschiebt sich die Entscheidungsgrenze in Richtung der Klasse ω_1.

Die Ungleichung Eq. 152 lässt sich analog zu Eq. 132 äquivalent als *likelihood ratio* formulieren

$$\begin{aligned}(L_{21} - L_{11})P(\omega_1)p(x|\omega_1) &> (L_{12} - L_{22})P(\omega_2)p(x|\omega_2) \\ \frac{p(x|\omega_1)}{p(x|\omega_2)} &> \frac{P(\omega_2)}{P(\omega_1)}\frac{(L_{12} - L_{22})}{(L_{21} - L_{11})}. \end{aligned} \tag{153}$$

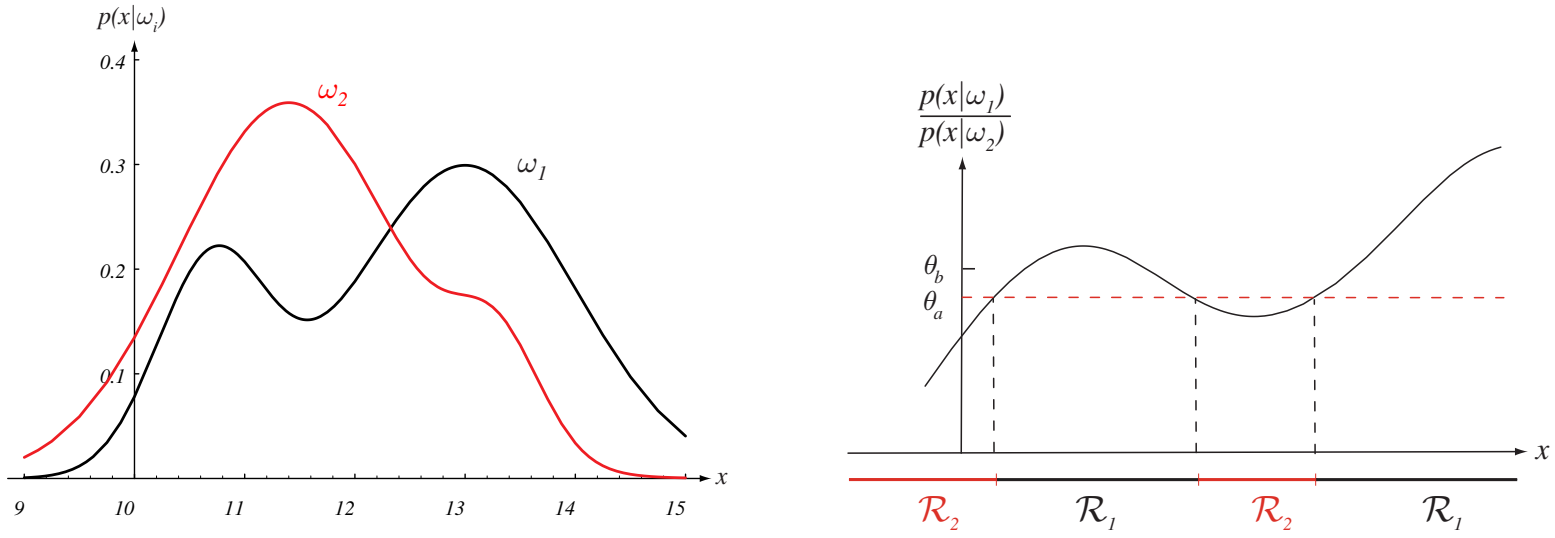


Abbildung 31: *Class conditional pdfs* (links) und korrespondierende *likelihood ratio* (rechts). Für *0/1-loss* und *priors* $P(\omega_1) = 2/3$ und $P(\omega_2) = 1/3$ erhält man den *threshold* θ_a. Ein asymmetrischer *loss* mit $L_{12} > L_{21}$ erhöht den *threshold* (θ_b) und verkleinert somit die Entscheidungsregion für ω_1. (Aus *Duda, Hart, Stork: Pattern Classification, 2nd ed.)*

Parameterschätzung II: Bayes-Ansatz

- Wir wenden das Bayes Paradigma nun auf die Schätzung von stetigen Parametern an. Ein interessierender Parameter θ habe eine *a priori* Verteilung mit DF (Prior) $p(\theta)$. Diese kann ihrerseits wieder von Parametern abhängen, welche als **Hyperparameter** bezeichnet werden. Die DF der Merkmals- oder Datenverteilung $p(x|\theta)$ (die *likelihood*) hängt vom gesuchten Parameter ab. Eine Beobachtung x_1 wird gemäß der Bayes-Regel mit dem Prior verknüpft, und ergibt die *a posteriori* Verteilung mit DF (Posterior)

$$p(\theta|x_1) = \frac{p(x_1|\theta)p(\theta)}{p(x_1)} \tag{154}$$

  Der Posterior dient nun als Prior für den nächsten Inferenzschritt mit der

– als iid angenommenen – Beobachtung x_2:

$$p(\theta|x_1, x_2) = \frac{p(x_2|\theta)p(\theta|x_1)}{\int p(x_2|\theta)p(\theta|x_1)d\theta} = \frac{p(x_2|\theta)p(x_1|\theta)p(\theta)}{\int p(x_2|\theta)p(x_1|\theta)p(\theta)d\theta} \quad (155)$$

Man erhält also – ausgehend von der *a priori* Verteilung – eine Sequenz von *a posteriori* Verteilungen. Bezeichne $\mathcal{D}$ die gesamte Stichprobe, so erhält man den finalen Posterior der Sequenz auch durch einmalige Anwendung der Bayes-Regel bzg. der *likelihood* $p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}. \quad (156)$$

Wir illustrieren die grundlegenden Ideen am Beispiel der Bernoulli- und der Normalverteilung.

- **Bernoulli-Verteilung**

  Die likelihood-Funktion des Parameters $\theta = P(X = 1)$ der Bernoulli-Verteilung hat bekanntlich die Form

$$l(\theta) = p(x_1, ..x_N|\theta) = \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{(1-x_i)} = \theta^k(1-\theta)^{(N-k)}, \quad (157)$$

  wobei $x_i \in \{0, 1\}$ den Ausfall des i-ten Versuchs bezeichnet. Die *a priori*-Verteilung $p(\theta)$ kann prinzipiell jede beliebige Form annehmen, die Berechnung der *a posteriori*-Verteilung $p(\theta|x)$ vereinfacht sich jedoch beträchtlich, wenn wir, gegeben die likelihood $p(x|\theta)$, die Verteilung von $p(\theta)$ so wählen, dass sie sich unter der Bayes-Inferenz reproduziert, d.h., $p(\theta)$ und $p(\theta|x)$ derselben parametrischen Familie von Dichtefunktionen angehören; man spricht dann von einem *conjugate prior*.

Im Falle der Bernoulli-Verteilung hat die **Beta-Verteilung** diese Eigenschaft:

$$\text{Beta}(\theta|a,b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}, \tag{158}$$

wobei

$$B(a,b) = \int_0^1 x^{a-1}x^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{159}$$

die Beta-Funktion und $\Gamma(x)$ die Gamma-Funktion bezeichnet.

Die Form der DF hängt von den Hyperparametern a, b ab. Abb. 32 zeigt den Verlauf der DF für verschiedene Werte dieser Parameter. Die Funktionsgraphen sind ident mit jenen der Bernoulli-*likelihood* in Abb. 17, allerdings ist die Parametrisierung eine andere, und die Fläche unter den Kurven beträgt nun, wie es sich für eine DF gehört, 1.

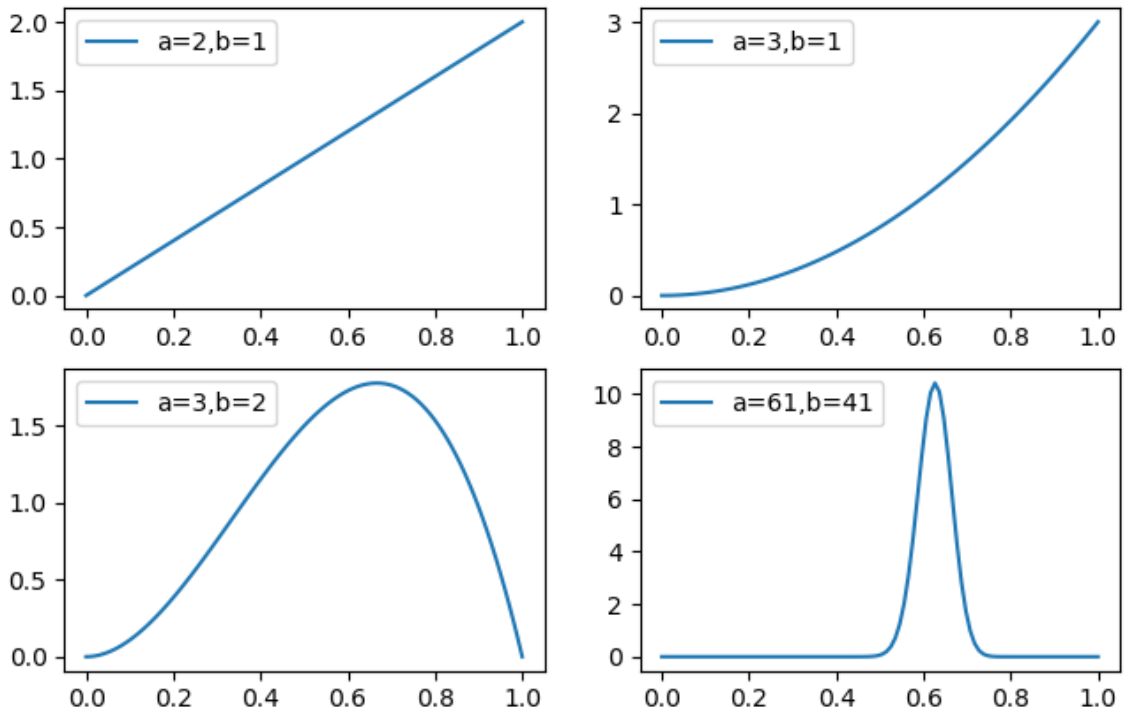


Abbildung 32: Beta-Verteilung Beta$(\theta|a,b)$ für verschiedene Werte der Hyperparameter a, b.

Es gilt

$$\mathcal{E}[\text{Beta}(\theta|a,b)] = \frac{a}{a+b} \quad (160)$$

$$Modus[\text{Beta}(\theta|a,b)] = \frac{a-1}{a+b-2} \quad (161)$$

$$Var[\text{Beta}(\theta|a,b)] = \frac{ab}{(a+b)^2(a+b+1)} \leq \frac{1}{a+b+1}, \quad (162)$$

wobei Modus *(mode)* das Maximum der DF bezeichnet.

Die Parameter a und b lassen sich als $k_0 = a-1$ günstige und $(N_0-k_0) = b-1$ ungünstige Ausfälle in einer virtuellen Stichprobe vom Umfang $N_0 = a+b-2$ auffassen. Die DF wird bei gleichbleibendem Verhältnis $a:b$ mit zunehmender Größe der virtuellen Stichprobe immer schmäler.

Die *a posteriori* DF erhalten wir gemäß der Bayes-Regel mit

$$p(\theta|x_1,..x_N) = \frac{1}{NF}\, p(x_1,..x_N|\theta)\ \text{Beta}(\theta|a,b) \tag{163}$$

$$= \frac{1}{NF}\, \theta^{k+a-1}(1-\theta)^{N-k+b-1}. \tag{164}$$

Dies ist offensichtlich wieder eine Beta-DF, mit $NF = \frac{\Gamma(k+a)\Gamma(N-k+b)}{\Gamma(N+a+b)}$, welche die Verteilung der mit der virtuellen *a priori* Stichprobe „gepoolten“ beobachteten Stichprobe beschreibt.

Je größer $a+b$ im Verhältnis zu N, desto mehr Einfluß hat die *a priori*-DF auf Position und Breite der *a posteri*-DF.

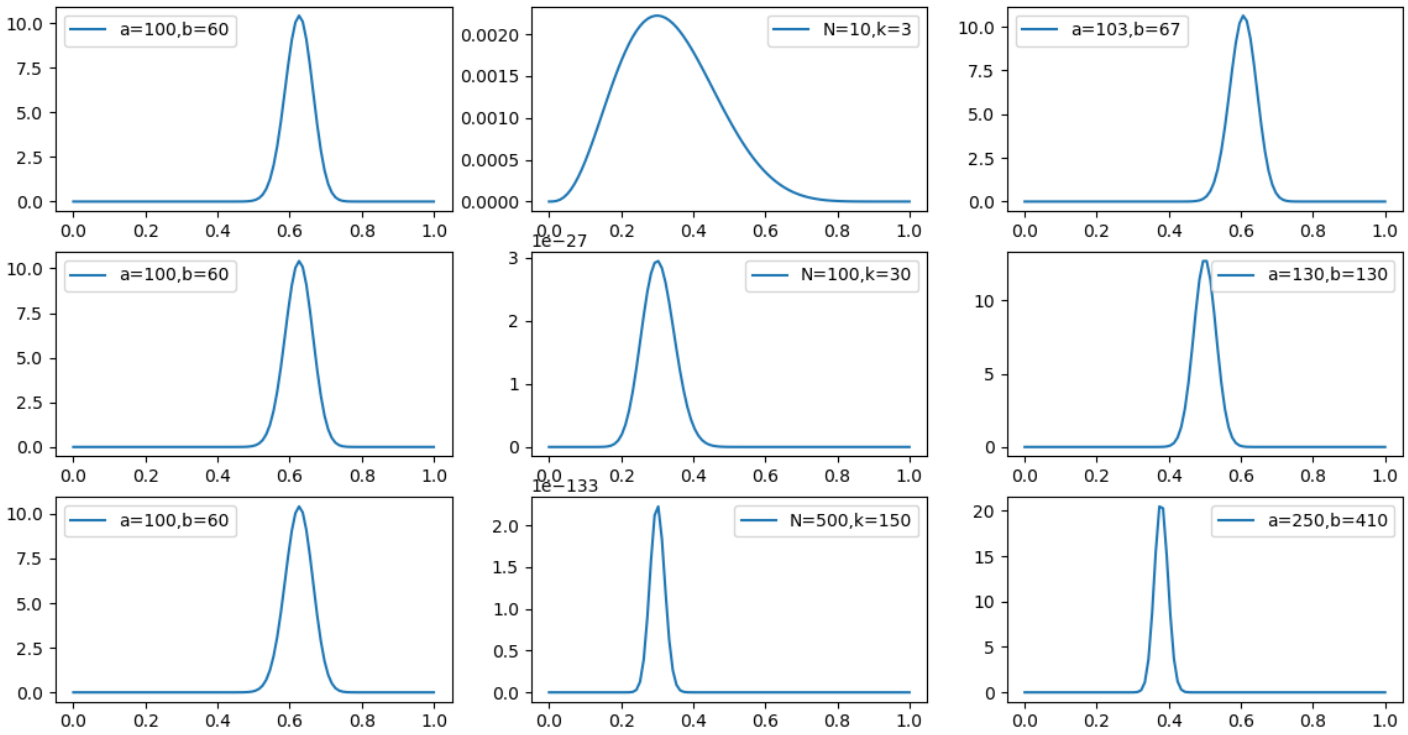


Abbildung 33: Beta-Prior (links), *likelihood* (Mitte) und Beta-Posterior (rechts) für unterschiedliche Stichprobengrößen. Siehe Text.

Abb. 33 veranschaulicht das Prinzip. In der linken Spalte ist die *a priori* Verteilung von θ mit Mittel $\frac{a}{a+b} = \frac{2}{3} = 0.67$ dargestellt, in der mittleren Spalte die *likelihoods* für verschiedene Stichprobengrößen (jedoch mit konstantem Verhältnis $\frac{k}{N} = 0.3$), und in der rechten Spalte die resultierende *a posteriori* Verteilung.

In der ersten Zeile ist die Stichprobengröße $N = 10$ klein gegenüber der virtuellen Stichprobengröße $a + b - 2 = 158$ des Priors, und somit die Varianz der Stichprobe groß gegenüber jener es Priors; der Posterior unterschiedet sich daher kaum vom Prior. In der zweiten Zeile sind die Stichproben ungefähr gleich groß, der Posterior erscheint als Kompromiss mit Mittel 0.5 (und verringerter Varianz). In der untersten Zeile ist die Stichprobe deutlich größer als die virtuelle Stichprobengröße des Priors. Somit dominiert erstere den Posterior: das Mittel befindet sich nahe am Modus der *likelihood*, und die Varianz ist gegenüber dem Prior deutlich verringert.

- **Mittel einer Normalverteilung**

  Wir nehmen als *a priori* Verteilung des Mittels der Einfachheit halber ebenfalls eine Normalverteilung mit Mittel μ_0 und Varianz σ_0^2 an (dies sind die Hyperparameter):

$$p(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) \tag{165}$$

  Im folgenden bezeichnen Größen mit Subskript wie z.B. μ_0 konstante Werte.

Die Merkmale kommen voraussetzungsgemäß aus einer Normalverteilung – der Datenverteilung – mit dem zu bestimmenden Mittel μ und als bekannt angenommener Varianz σ_x^2

$$p(x|\mu) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{(x-\mu)^2}{2\sigma_x^2}\right) \tag{166}$$

Nach Beobachten einer Merkmalsausprägung x_1 kombinieren wird nun Gl. 165 und Gl. 166 gemäß der Bayes-Regel und erhalten

$$\begin{aligned} p(\mu|x_1) &= p(x_1|\mu)p(\mu)/p(x_1) \\ &= \frac{1}{2\pi\sigma_x\sigma_0 p(x_1)} \exp\left(-\frac{(x-\mu)^2}{2\sigma_x^2} - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right). \end{aligned} \tag{167}$$

Der Nenner im obigen Ausdruck – die *evidence* $p(x_1)$ – ist durch

$$p(x_1) = \int_{-\infty}^{+\infty} p(x_1|\mu)p(\mu)\ d\mu \tag{168}$$

gegeben.

Integrale der obigen Form treten in der Bayes-Statistik häufig auf. In einigen Spezialfällen können diese analytisch bestimmt werden, in der Praxis werden sie jedoch meist numerisch ausgewertet bzw. der Posterior mittels spezieller stochastischer Verfahren wie *Markov Chain Monte Carlo* bestimmt.

In unserem Fall ist $-\log p(\mu|x_1)$ gemäß Gl. 167 proportional einer quadratischen Funktion von μ, d.h. der Posterior $p(\mu|x_1)$ ist wiederum eine Normalverteilung; in etwas schlampiger Notation (welche μ nun als Zufallsvariable auffaßt) $\mu|x_1 \sim N(\mu_1, \sigma_1^2)$. Mittel μ_1 und Varianz σ_1^2 des

posteriors erhält man durch Koeffizientenvergleich der in μ quadratischen und linearen Terme des Exponenten (*Vervollständigung des Quadrats)*

$$\begin{aligned} \frac{(\mu-\mu_1)^2}{\sigma_1^2} &= \frac{(x-\mu)^2}{\sigma_x^2} + \frac{(\mu-\mu_0)^2}{\sigma_0^2} \qquad (169) \\ \frac{1}{\sigma_1^2}\mu_1^2 - 2\frac{\mu_1}{\sigma_1^2}\mu &= \left(\frac{1}{\sigma_x^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{x}{\sigma_x^2}+\frac{\mu_0}{\sigma_0^2}\right)\mu \end{aligned}$$

und somit

$$\begin{aligned} \frac{1}{\sigma_1^2} &= \left(\frac{1}{\sigma_x^2}+\frac{1}{\sigma_0^2}\right) \qquad (170) \\ \mu_1 &= \left(\frac{1}{\sigma_x^2}+\frac{1}{\sigma_0^2}\right)^{-1}\left(\frac{x_1}{\sigma_x^2}+\frac{\mu_0}{\sigma_0^2}\right) \qquad (171) \end{aligned}$$

Das Mittel μ_1 der *a posteriori* Verteilung ergibt sich als mit den inversen

Varianzen (*precisions*) gewichtetes Mittel aus dem *a priori* Mittel μ_0 und der Beobachtung x_1.

Das obige Schema lässt sich iterativ auf weitere Beobachtungen anwenden, wenn man annimmt, dass die Beobachtungen bedingt unabhängig (iid) bzg. des Parameters sind. Der aktuelle Posterior fungiert dabei als Prior für den nächsten Inferenz-Schritt, z.B.

$$p(\mu|x_1, x_2) = p(x_2|\mu)p(\mu|x_1) / \int p(x_2|\mu)p(\mu|x_1)d\mu \tag{172}$$

Fassen wir $\mu|x_1, x_2$ als Zufallsvariable auf, so gilt $\mu|x_1, x_2 \sim N(\mu_2, \sigma_2^2)$

mit

$$\frac{1}{\sigma_2^2} = \left(\frac{1}{\sigma_x^2} + \frac{1}{\sigma_1^2}\right) = \left(\frac{1}{\sigma_x^2} + \frac{1}{\sigma_x^2} + \frac{1}{\sigma_0^2}\right) = \left(\frac{2}{\sigma_x^2} + \frac{1}{\sigma_0^2}\right) \quad (173)$$

$$\mu_2 = \left(\frac{2}{\sigma_x^2} + \frac{1}{\sigma_0^2}\right)^{-1} \left(\frac{x_1}{\sigma_x^2} + \frac{x_2}{\sigma_x^2} + \frac{\mu_0}{\sigma_0^2}\right), \quad (174)$$

oder allgemein nach N Beobachtungen

$$\frac{1}{\sigma_N^2} = \left(\frac{N}{\sigma_x^2} + \frac{1}{\sigma_0^2}\right) \quad (175)$$

$$\mu_N = \left(\frac{N}{\sigma_x^2} + \frac{1}{\sigma_0^2}\right)^{-1} \left(\frac{\sum_{i=1}^{N} x_i}{\sigma_x^2} + \frac{\mu_0}{\sigma_0^2}\right), \quad (176)$$

Dasselbe Ergebnis hätte man in einem einmaligen Inferenz-Schritt un-

ter Verwendung der Verbund-Daten-DF $p(\mathcal{D}|\mu) = p(x_1, \ldots, x_N|\mu) = \prod p(x_i|\mu)$ (iid-Annahme) erhalten.

Das Ergebnis der Bayes-Inferenz ist im Unterschied zum frequentistischen Ansatz keine Punktschätzung der Parameters (hier des Mittels), sondern eine Verteilung bzw. DF des Parameters, die *a posteriori* Verteilung oder kurz Posterior. Diese reflektiert die Gesamtheit unserer *a priori* vorhandenen und der in den Daten enthaltenen Information.

Nehmen wir an, dass wir keinerlei Information über die *a priori* Verteilung des Mittelwerts haben; dies können wir quantitativ anzeigen, indem wir die Varianz σ_0^2 der *a priori* Verteilung auf $+\infty$ setzen. Dann werden aber Gl. 175, 176 zu

$$\sigma_N^2 = \frac{\sigma_x^2}{N} \tag{177}$$

$$\mu_N = \frac{\sum_{i=1}^{N} x_i}{N} \tag{178}$$

Das sind aber genau die Werte, die wir zuvor mit ML erhalten haben!

Der Posterior $N(\mu_N, \sigma_N^2)$ ist das Bayes-Pendant der frequentistischen Schätzfunktion. Er hat dieselbe Form wie das Stichprobenmittel Gl. 93, allerdings ist sein Mittel durch den Schätzwert bzw. Hyperparameter μ_N, und nicht durch das "wahre", aber unbekannte Populationsmittel gegeben.

Für den gegebenen Posterior lassen sich nun direkt Intervalle angeben, in denen das Mittel mit einer gewissen Wahrscheinlichkeit liegt. Diese sogenannten **credible intervals CI** entsprechen der menschlichen Intuition wesentlich besser als die klassischen Konfidenzintervalle: so ist die Wahrscheinlichkeitsdichte des Posteriors im konkreten Beispiel in der Nachbarschaft von μ_N (des frequentistischen Schätzwerts) am größten.

Multivariate stetige Verteilungen $p \geq 2$

- **Rand-DF/VF**

  Die **Rand-Dichtefunktion (DF)** *(marginal pdf)* der i-ten Variable (Komponente) von $\vec{X} = (X_1, \ldots, X_P)^T$ erhält man durch Integration der *joint pdf* über alle anderen Variablen

$$p_i(x_i) = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} p(x_1, \ldots, x_p) dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_p. \quad (179)$$

Die *marginal pdf* einer Menge von Variablen S erhält man durch Integration der *joint pdf* über die restlichen Variablen $\{X_1, \ldots, X_p\} - S$. Z.B. ergibt sich die *marginal pdf* von $S = \{X_1, \ldots, X_r\}$ zu

$$p_{1\ldots r}(x_1, \ldots, x_r) = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} p(x_1, \ldots, x_r, x_{r+1}, \ldots, x_p) dx_{r+1} \ldots dx_p. \quad (180)$$

Die **Rand-VF** *(marginal cdf)* der i-ten Komponente erhält man durch Integration über die Rand-DF der i-ten Komponente

$$F_i(x_i) = F(+\infty, \ldots, +\infty, x_i, +\infty, \ldots, +\infty) = \int_{-\infty}^{x_i} p_i(x_i') dx_i'. \quad (181)$$

(Analog für eine Menge von Variablen.)

- **Unabhängigkeit**

  $X_1, \ldots, X_p$ sind wechselseitig **unabhängig** (*mutually independent*) g.d.w. die DF (VF) in das Produkt der Rand-DFen (VFen) faktorisiert:

$$F(x_1, \ldots, x_p) = F_1(x_1) \ldots F_p(x_p) = \prod_i F_i(x_i), \tag{182}$$

$$p(x_1, \ldots, x_p) = p_1(x_1) \ldots p_p(x_p) = \prod_i p_i(x_i). \tag{183}$$

- **Erwartung und Momente**

  Die Erwartung $\mathcal{E}[]$ einer reellwertigen Funktion einer multivariaten Zufallsvariablen X $h : \mathbf{R}^p \rightarrow \mathbf{R}$ ist definiert als

$$\begin{aligned} \mathcal{E}[h(\vec{X})] &= \int_{-\infty}^{+\infty} h(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \\ &= \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} h(x_1, \ldots, x_p) p(x_1, \ldots, x_p) dx_1 \ldots dx_p. \end{aligned} \tag{184}$$

  Für

$$h(X_1, \ldots, X_p) = \prod_{i=1}^{p} X_i^{l_i}, \quad l_i \in \mathbf{N}, \sum_{i=1}^{p} l_i = k, \tag{185}$$

  erhält man die **Momente k-ter Ordnung** (*k-th order moments*) von $\vec{X}$.

Speziell erhält man für $k = 1$ die p Momente erster Ordnung μ_i

$$\begin{aligned}\mu_i &= \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} x_1^0 \dots x_{i-1}^0 x_i^1 x_{i-1}^0 \dots x_p^0 \, p(x_1, \dots, x_p) dx_1 \dots dx_p \\ &= \int_{-\infty}^{+\infty} x_i p_i(x_i) dx_i = \int_{-\infty}^{+\infty} x p_i(x) dx = \mathcal{E}[X_i]. \end{aligned} \tag{186}$$

Wie man leicht sieht, ist Eq. 186 äquivalent zu Eq. 65, dem Mittelwert im univariaten Fall; μ_i ist also das Mittel von X_i.

Die μ_i sind die Komponenten des **Mittelwertvektors** von $\vec{X}$, $\boldsymbol{\mu}$

$$\boldsymbol{\mu} = \mathcal{E}[\vec{X}] = (\mu_1, \dots, \mu_p)^T = (\mathcal{E}[x_1], \dots, \mathcal{E}[x_p])^T. \tag{187}$$

$\boldsymbol{\mu}$ beschreibt als *Ortparameter* das Zentrum (den Schwerpunkt) der Verteilung von $\vec{X}$.

Die zentralen (d.h. mittelwertbereinigten) Momente zweiter Ordnung σ_{ij} bezeichnet man als **Varianz** von X_i $(i = j)$

$$\begin{aligned} \sigma_{ii} = \sigma_i^2 &= \int_{-\infty}^{+\infty} (x_i - \mu_i)^2 p_i(x_i) dx_i \\ &= \mathcal{E}[(X_i - \mu_i)(X_i - \mu_i)] \end{aligned} \tag{188}$$

bzw. als **Kovarianz** $(i \neq j)$ von X_i und X_j

$$\begin{aligned} \sigma_{ij} &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x_i - \mu_i)^1 (x_j - \mu_j)^1 p_{ij}(x_i, x_j) dx_i dx_j \\ &= \mathcal{E}[(X_i - \mu_i)(X_j - \mu_j)] \end{aligned} \tag{189}$$

(vergleiche Eq. 66 und Eq. 71). Die Matrix

$$\begin{aligned} Var[\vec{X}] &= \mathbf{\Sigma} = (\sigma_{ij}) = \begin{pmatrix} \sigma_{11} & \dots & \sigma_{1p} \\ \dots & \dots & \dots \\ \sigma_{p1} & \dots & \sigma_{pp} \end{pmatrix} \\ &= \mathcal{E}[(\vec{X} - \boldsymbol{\mu})(\vec{X} - \boldsymbol{\mu})^T] \end{aligned} \tag{190}$$

bezeichnet man als **Kovarianzmatrix** von $\vec{X}$.

Matrizen werden im folgenden durch fette Großbuchstaben bezeichnet.

Analog zum bivariaten Fall (Eq. 71) läßt sich $\mathbf{\Sigma}$ unter Verwendung der Linearität des Erwartungsoperators, Eq. 68, folgendermaßen schreiben

$$\begin{aligned}
\mathbf{\Sigma} &= \mathcal{E}[(\vec{X} - \boldsymbol{\mu})(\vec{X} - \boldsymbol{\mu})^T] \\
&= \mathcal{E}[\vec{X}\vec{X}^T] - \mathcal{E}[\vec{X}]\boldsymbol{\mu}^T - \boldsymbol{\mu}\mathcal{E}[\vec{X}]^T + \boldsymbol{\mu}\boldsymbol{\mu}^T \\
&= \mathcal{E}[\vec{X}\vec{X}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T, \qquad (191)
\end{aligned}$$

wobei $\mathcal{E}[\vec{X}\vec{X}^T]$ die (nicht mittelwertbereinigten) Momente 2-ter Ordnung enthält.

In der Herleitung von Eq. 191 wurde von folgendem Lemma Gebrauch gemacht, welches wir im folgenden noch häufiger benötigen werden.

Lemma 1. *Sei $\mathcal{A} = (a_{ij})$ eine $p \times q$ Zufallsmatrix, d.h. eine Matrix deren Elemente a_{ij} Zufallsvariablen darstellen. Seien weiters $\mathbf{F} \in \boldsymbol{R}^{n \times p}, \mathbf{G} \in \boldsymbol{R}^{q \times m}, \mathbf{H} \in \boldsymbol{R}^{n \times m}$ reelle Matrizen. Es gilt*

$$\mathcal{E}[\mathbf{F}\mathcal{A}\mathbf{G} + \mathbf{H}] = \mathbf{F}\mathcal{E}[\mathcal{A}]\mathbf{G} + \mathbf{H}. \tag{192}$$

Als Spezialfall erhält man

$$\mathcal{E}[\boldsymbol{\mu}\vec{X}^T] = \boldsymbol{\mu}\mathcal{E}[\vec{X}^T]. \tag{193}$$

- **Schätzung des Mittels**

  Gegeben seien N p-dimensionale Beobachtungen $\mathbf{x}_i$ (Realisierungen von N iid verteilten Zufallsvektoren $\vec{X}_i \in \mathbf{R}^p$), welche wir (als Spaltenvektoren) in der *sample matrix* $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N) \in \mathbf{R}^{p \times N}$ zusammenfassen.

  Der (erwartungstreue) Schätzer des Mittelwerts ergibt sich, analog zum univariaten Fall, als

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{i=1}^{N} \vec{X}_i, \tag{194}$$

  d.h. der Schätzer für die i-te Komponente ist durch Eq. 65 gegeben. Man beachte, dass $\hat{\boldsymbol{\mu}}$ wiederum ein Zufallsvektor ist.

Der konkrete Wert des Schätzers für gegebene *sample matrix* $\mathbf{X}$ berechnet sich daher wie folgt

$$\hat{\mathbf{m}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i. \tag{195}$$

- **Schätzung der Kovarianz-Matrix**

  Ein erwartungstreuer Schätzer der Kovarianz ist durch

$$\hat{\mathbf{\Sigma}} = (\hat{\sigma}_{ij}) \quad = \quad \frac{1}{N-1} \sum_{i=1}^{N} (\vec{X}_i - \hat{\boldsymbol{\mu}})(\vec{X}_i - \hat{\boldsymbol{\mu}})^T \tag{196}$$

  gegeben. Alle Komponenten $\hat{\sigma}_{ij}$ sind wiederum Zufallsvariablen (und $\hat{\mathbf{\Sigma}}$ somit eine Zufallsmatrix). Auch hier muss, wie im univariaten Fall (siehe Eq. 105), durch $N-1$ und nicht durch N dividiert werden, um die Erwartungstreue von $\hat{\mathbf{\Sigma}}$ zu gewährleisten.

Bezeichne im folgenden $\tilde{\mathbf{X}}$ die mittelwertbereinigten (*mean normalized*) *samples*

$$\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_N) = ((\mathbf{x}_1 - \hat{\mathbf{m}}), \ldots, (\mathbf{x}_N - \hat{\mathbf{m}})). \tag{197}$$

Die Realisierung von $\hat{\mathbf{\Sigma}}$ für gegebene *sample matrix* $\mathbf{X}$ (bzw. $\tilde{\mathbf{X}}$) berechnet sich wie folgt

$$\begin{aligned} \hat{\mathbf{C}} = (\hat{s}_{ij}) &= \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_i - \hat{\mathbf{m}})(\mathbf{x}_i - \hat{\mathbf{m}})^T \\ &= \frac{1}{N-1} \sum_{i=1}^{N} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i{}^T \qquad (198) \\ &= \frac{1}{N-1} \tilde{\mathbf{X}} \tilde{\mathbf{X}}^T. \qquad (199) \end{aligned}$$

Die analytisch äquivalente Formulierung

$$\hat{\mathbf{C}} = \frac{1}{N-1}(\mathbf{X}\mathbf{X}^T - N\hat{\mathbf{m}}\hat{\mathbf{m}}^T) \tag{200}$$

sollte aus numerischen Gründen (Akkumulation von Rundungsfehlern) vermieden wenden.

Kovarianzmatrix und multivariate Normalverteilung

- Analog zum univariaten Fall ist eine multivariate Normalverteilung durch Angabe ihrer ersten beiden Momente – des Mittelwertvektors und der Kovarianzmatrix – vollständig festgelegt. Eigenschaften und Bedeutung der Kovarianzmatrix lassen sich daher am anschaulichsten anhand der multivariaten Normalverteilung demonstrieren.

- **Symmetrie und Positive Definitheit**

  Kovarianzmatrizen sind symmetrisch, d.h $\mathbf{\Sigma} = \mathbf{\Sigma}^T$.

  Darüberhinaus ist $\mathbf{\Sigma}$ **positiv semi-definit**

$$\mathbf{x}^T \mathbf{\Sigma} \mathbf{x} \geq 0 \quad \forall \mathbf{x} \in \mathbf{R}^p. \tag{201}$$

  Für nicht degenerierte Verteilungen, d.h. Verteilungen, die in allen Richtungen eine positive Varianz aufweisen, ist $\mathbf{\Sigma}$ darüberhinaus **positiv definit**

$$\mathbf{x}^T \mathbf{\Sigma} \mathbf{x} > 0 \quad \forall (\mathbf{x} \neq \mathbf{0}) \in \mathbf{R}^p, \tag{202}$$

  Positive Definitheit impliziert Invertierbarkeit (aber nicht *vice versa*).

- Die Dichtefunktion (joint pdf) eines normalverteilten Zufallsvektors $\vec{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ mit Mittelwert $\boldsymbol{\mu}$ und Kovarianzmatrix $\boldsymbol{\Sigma}$ ist wie folgt definiert

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{p\frac{1}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \tag{203}$$

  wobei $|\boldsymbol{\Sigma}|$ die Determinante von $\boldsymbol{\Sigma}$ bezeichnet.

  Der Exponent in Eq. 203 hängt vom Wert der quadratischen Form

$$(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) =< \mathbf{x}-\boldsymbol{\mu}, \mathbf{x}-\boldsymbol{\mu} >_{\boldsymbol{\Sigma}^{-1}} = d^2(\mathbf{x}) \tag{204}$$

  ab. $\boldsymbol{\Sigma}^{-1}$ ist, wie auch $\boldsymbol{\Sigma}$, symmetrisch und positiv semi-definit.

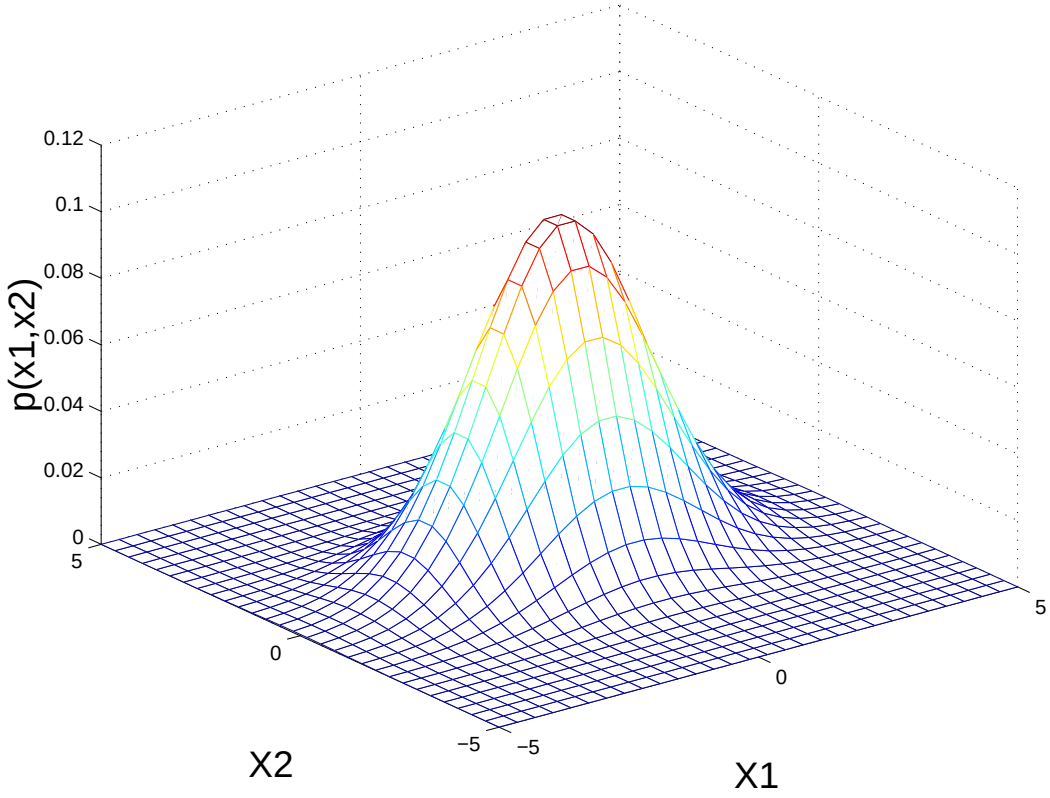


Abbildung 34: Beispiel für die Dichtefunktion einer bivariaten Normalverteilung.

- **Mahalanobis-Distanz**

  Die an der gemeinsamen Kovarianzmatrix standardisierte Distanz (Metrik) zweier Punkte $\mathbf{x}, \mathbf{y}$:

$$\sqrt{(\mathbf{x}-\mathbf{y})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\mathbf{y})} = d(\mathbf{x}, \mathbf{y}) \tag{205}$$

  bezeichnet man als deren **Mahalanobis-Distanz**. Ist das zweite Argument das Mittel der Verteilung, kann es weggelassen werden. Die Menge aller Punkte $\{\mathbf{x} : d^2(\mathbf{x}) = c^2\}$, für welche die Mahalanobis-Distanz vom Mittel einer Normalverteilung gleich einer Konstanten c ist, ist für $p = 2$ durch eine Ellipse, für $p \geq 3$ durch ein Hyper-Ellipsoid im $\mathbf{R}^p$ mit Mittelpunkt $\boldsymbol{\mu}$ gegeben. Für alle auf einem solchen Hyper-Ellipsoid liegenden Punkte liefert die DF $p(\mathbf{x})$ denselben Wert.

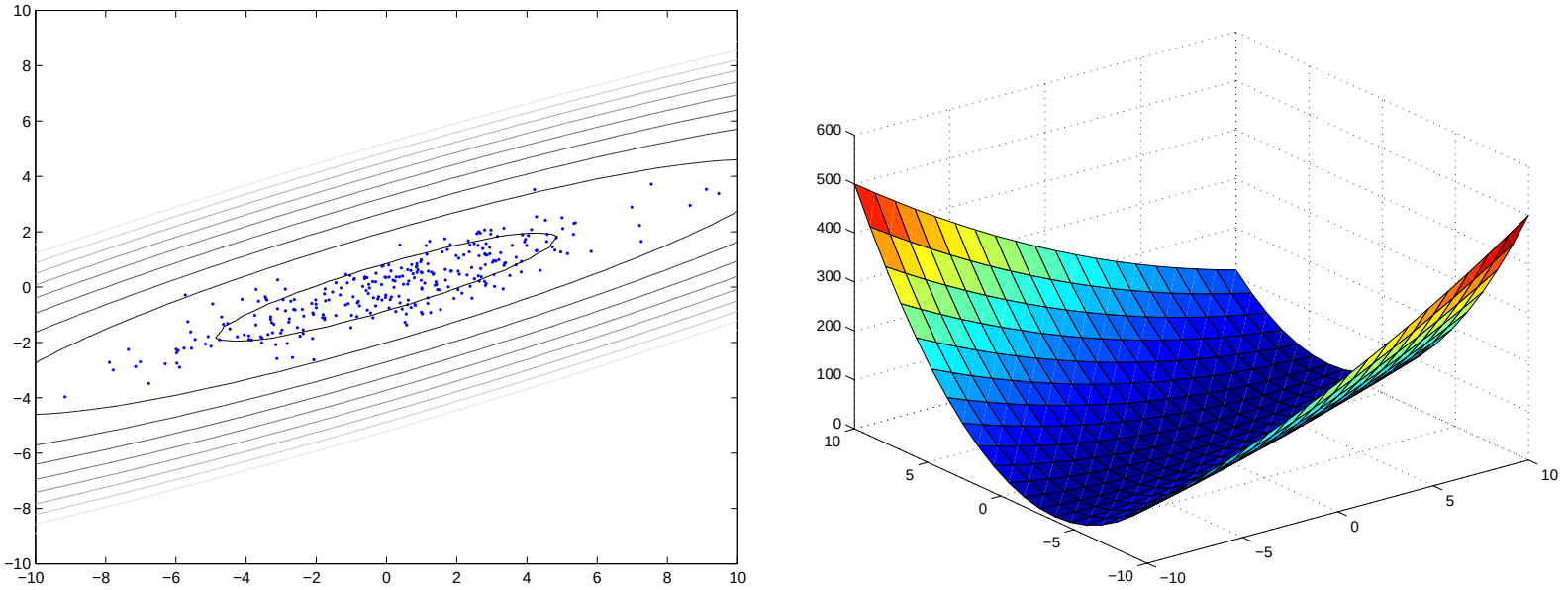


Abbildung 35: **Mahalanobis-Distanz**
der bivariaten Normalverteilung $\boldsymbol{\mu} = \mathbf{0}, \sigma_X^2 = 12, \sigma_Y^2 = 2, \rho_{XY} = 0.9$
Links: Konturplot, jede Ellipse entspricht einem konstanten Wert c^2 für $d^2(\mathbf{x})$.
Rechts: Darstellung der Mahalanobis-Distanz als Fläche über (x, y). Die Konturlinien erhält man als Schnittkurven der Fläche mit zur $x - y$-Ebene parallelen Ebenen.

Nehmen wir zunächst an, dass für eine bivariate Normalverteilung

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & 0 \\ 0 & \sigma_Y^2 \end{pmatrix} \tag{206}$$

eine Diagonalmatrix ist (d.h. $\sigma_{ij} = 0$ für $i \neq j$) und somit die Komponenten X_i wechselseitig dekorreliert sind. In diesem Fall gilt

$$\mathbf{\Sigma}^{-1} = \begin{pmatrix} \sigma_1^{-2} & 0 \\ 0 & \sigma_2^{-2} \end{pmatrix} \tag{207}$$

und somit

$$d^2(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \sum_{i=1}^{p} \frac{(x_i - \mu_i)^2}{\sigma_i^2} = c^2, \tag{208}$$

d.h. wir erhalten im bivariaten Fall $(p = 2)$ tatsächlich die Gleichung einer Ellipse, bzw. für $p \geq 3$ eines Hyper-Ellipsoids mit Achsenlängen $c\sqrt{\sigma_{ii}}$ und Mittelpunkt $\boldsymbol{\mu}$.

- **Kovarianz und Korrelation**

  Die Kovarianz-Matrix beschreibt sowohl die Breite der Rand-Verteilungen der einzelnen Komponenten X_i (Varianz $\sigma_{ii} = \sigma_i^2$ in der Diagonale) als auch den (linearen) Zusammenhang zwischen diesen (Kovarianz σ_{ij}). Ist die Kovarianz zwischen zwei Komponenten 0, so bezeichnet man diese als **dekorreliert**. Korrelation entsteht durch Rotieren der Achsen des Hyper-Ellipsoids gegenüber den Achsen des Koordinatensystems des Merkmalsraums.

  Dies wird in Abb. 36, wieder am Beispiel einer bivariaten Normalverteilung, illustriert. Für eine solche sind die Iso-Linien konstanter DF durch Ellipsen gegeben. Die innere Ellipse deckt 40% der Wahrscheinlichkeitsmasse ab, in diesem Fall entspricht der Schnittpunkt der Ellipse mit einer Halbachse ungefähr der Standardabweichung entlang dieser Halbachse, $b = \sigma_{min}$ für die Neben- bzw. $a = \sigma_{maj}$ für die Hauptachse. Die nächste umschließende Ellipse enthält weitere 10% usw.

Die Orientierung (Winkel zwischen Haupt- und x-Achse) und Form (Elongation) der Ellipsen (Iso-Linien) sind durch die Kovarianz-Matrix festgelegt. Für $p \geq 3$ sind die Iso-Flächen konstanter DF durch (Hyper-)Ellipsoide gegeben.

In Abb. 36 wird die Ellipsenschar (d.h. die Kovarianzmatrix) ausgehend von der Startposition links im Gegenuhrzeigersinn gedreht, mit jeder Drehung wird die Standardabweichung σ_{min} entlang der Nebenachse um 0.1 reduziert, d.h. die Elongation wird größer. Ursprünglich sind die Variablen dekorreliert (die Kovarianz ist 0). Gelb entspricht einer positiven Kovarianz, die grünen Ellipsen dekorrelierten Variablen, wobei die Haupt- und Nebenachse im Vergleich zur Startkonfiguration nun jedoch vertauscht sind. Rechts wird die Kovarianz schließlich negativ.

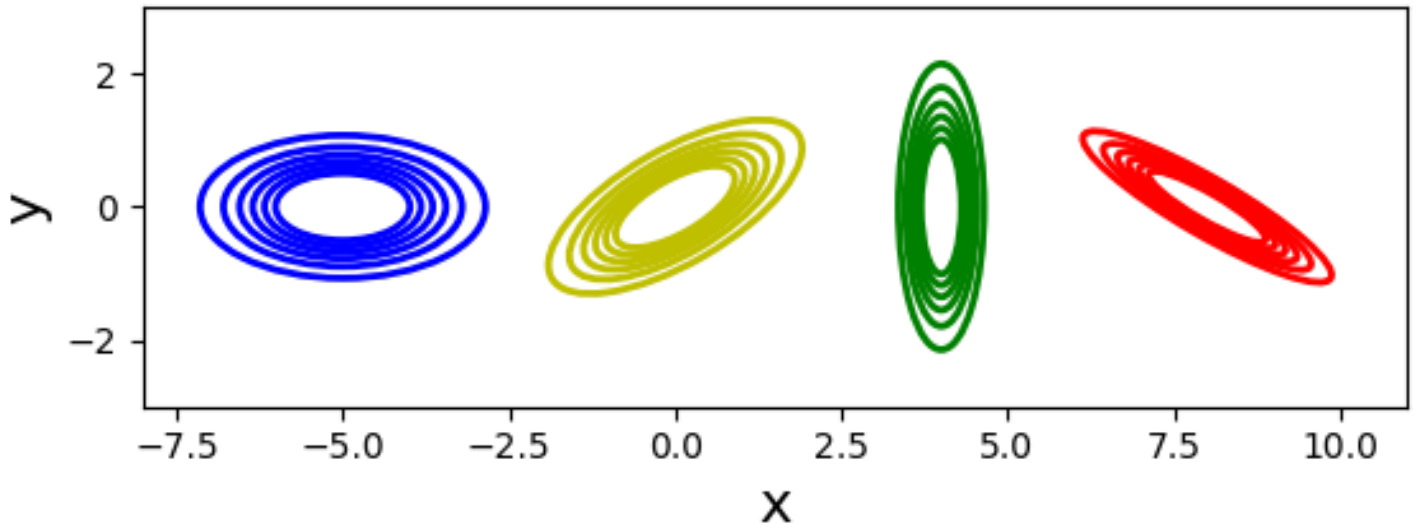


Abbildung 36: Kovarianz-Ellipsen mit unterschiedlichen Orientierungen und Elongationen. Links: $\sigma_{maj} = \sigma_x = 1, \sigma_{min} = \sigma_y = 0.5$. Von links nach rechts: Rotation erfolgt im Gegenuhrzeigersinn bei gleichzeitiger Verkleinerung von σ_{min} um je 0.1 (siehe Text).

Es muss zwischen den Standardabweichungen entlang der Koordinatenachsen σ_x, σ_y (also den Standardabweichungen der Randverteilungen) und jenen entlang der Achsen der Ellipse $\sigma_{maj}, \sigma_{min}$ unterschieden werden. Bei einem Winkel von 0^o fällt die Hauptachse mit der x-Achse zusammen, und es gilt $\sigma_x = \sigma_{maj}$. Mit zunehmendem Winkel wird die von der Ellipse auf die x-Achse projizierte Standardabweichung geringer, bei einer Drehung von 90^0 ist $\sigma_x = \sigma_{min}$, um schließlich bei 180^o wieder das Maximum $\sigma_x = \sigma_{max}$ anzunehmen. Als Funktion des Drehwinkels der Kovarianzmatrix verhalten sich die Projektionen der Standardabweichung auf die Koordinatenachsen cos- bzw. sin-förmig.

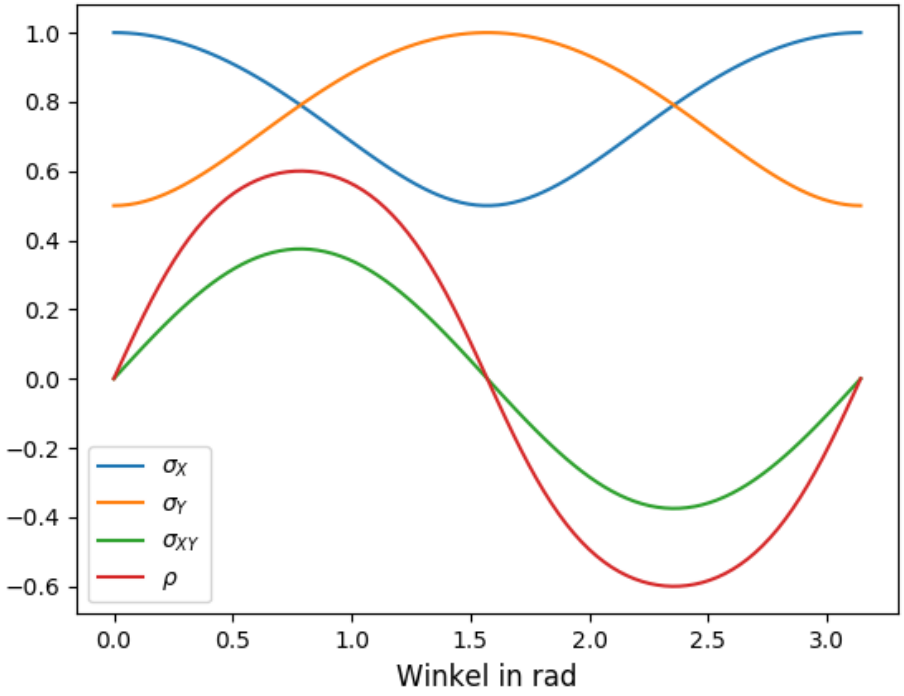


Abbildung 37: Varianzen/Kovarianzen als Funktion des Rotationswinkels für die erste (blaue) Kovarianzmatrix in Abb. 36. Oben: Auf die x und y-Achse projizierte Standardabweichung. Unten: Kovarianz und Korrelation.

In Abb. 37 oben ist die Standardabweichung der Randverteilungen für die erste der in Abb. 36 dargestellte Ellipsen bzw. Kovarianzmatrizen als Funktion des Rotationswinkels dargestellt. Ursprünglich betragen $\sigma_x = 1$ und $\sigma_y = 0.5$. Bei einem Winkel von 45^o werden sie gleich, bei 90^0 schließlich sind sie gegenüber der Ausgangsposition (0^o) vertauscht.

In Abb. 37 unten sind sowohl die Kovarianz als auch die Korrelation als Funktion des Winkels dargestellt. Man sieht, dass diese – für gegebene Elongation – ihr Maximum bei 45^0 Grad aufweisen.

- **Varianz einer Linearkombination von Zufallsvariablen**

  Angenommen, wir sind an der Varianz der Linearkombination von p Zufallsvariablen $\vec{X} = (X_1, \ldots, X_p) \in \mathbf{R}^p$ mit dem Koeffizientenvektor $\mathbf{w} \in \mathbf{R}^p$ interessiert. Die transformierte Variable Y erhält man als Linearkombinationen der X_i mit Koeffizienten w_i.

$$Y = \mathbf{w}^T \vec{X} = \sum w_i X_i. \tag{209}$$

  Sei $\mathcal{E}[\vec{X}] = \mathbf{0}$ und somit $\mathcal{E}[Y] = \mathbf{0}$. Es gilt

$$\begin{aligned} Var(Y) &= \mathcal{E}[Y^2] = \mathcal{E}[YY^T] = \mathcal{E}[\mathbf{w}^T \vec{X} \vec{X}^T \mathbf{w}] \\ &= \mathbf{w}^T \mathcal{E}[\vec{X}\vec{X}^T]\mathbf{w} = \mathbf{w}^T \mathbf{\Sigma} \mathbf{w}. \end{aligned} \tag{210}$$

Bezeichne $I = \{i_1, \ldots, i_k\}$ eine Teilmenge von $\{1, \ldots, p\}$, und sei $\mathbf{w}_I \in \mathbf{R}^p$ definiert als

$$w_{I_i} = \begin{cases} 1 & \text{falls } i \in I \\ 0 & \text{sonst.} \end{cases}$$

Dann liefert Eq. 210 die Varianz der Summe der k Komponenten $\{X_{i_1}, \ldots, X_{i_k}\}$ von $\vec{X}$. So erhält man z.B. für $p = 5$ und $\mathbf{w}_I = (1, 1, 0, 0, 0)^T$

$$Var(X_1 + X_2) = \mathbf{w}_I^T \mathbf{\Sigma} \mathbf{w}_I = \sigma_{11} + 2\sigma_{12} + \sigma_{22} \tag{211}$$

(vergleiche Eq. 70). Ist die Kovarianz σ_{12} zwischen der ersten und zweiten Komponente 0, so ergibt sich die Varianz der Summe $X_1 + X_2$ als Summe der Einzelvarianzen.

- **Varianz unter Projektion**

  Ein Spezialfall ist die Ermittlung der Varianz des Zufallsvektors $\vec{X} \in \mathbb{R}^p$ entlang der Richtung $\mathbf{w} \in \mathbb{R}^p$, oder, anders formuliert, der Varianz der Projektion $Y = \mathbf{w}^T \vec{X}$ unter der Nebenbedingung $\|\mathbf{w}\| = 1$.

$$Var(Y) = \frac{\mathbf{w}^T \mathbf{\Sigma} \mathbf{w}}{\|\mathbf{w}\| \|\mathbf{w}\|} = \frac{\mathbf{w}^T \mathbf{\Sigma} \mathbf{w}}{\mathbf{w}^T \mathbf{w}}. \tag{212}$$

  Man sieht, dass sich die Varianz der Projektion Y als Quotient zweier (symmetrischer) quadratischer Formen auffassen läßt.

Seien allgemein $\mathbf{A}, \mathbf{B}$ symmetrische Matrizen und $\mathbf{B}$ darüberhinaus positiv definit. Der Quotient der durch $\mathbf{A}, \mathbf{B}$ induzierten quadratischen Formen

$$r(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{A} \mathbf{w}}{\mathbf{w}^T \mathbf{B} \mathbf{w}} \tag{213}$$

wird als **Rayleigh Quotient** bezeichnet.

Sei nun

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{maj}^2 & 0 \\ 0 & \sigma_{min}^2 \end{pmatrix} \tag{214}$$

$$\mathbf{w} = \begin{pmatrix} \cos(\alpha) \\ \sin(\alpha) \end{pmatrix}. \tag{215}$$

Ist die Ellipse gegenüber der x-Achse um $-\alpha$ gedreht, so erscheint die x-Achse gegenüber der Hauptachse der Ellipse um α gedreht. Die Varianz der Projektion der Verteilung auf die x-Achse erhält man gemäß Gl 212 mit

$$\sigma_{xx} = \sigma_x^2 = \cos^2(\alpha)\sigma_{maj}^2 + \sin^2(\alpha)\sigma_{min}^2. \tag{216}$$

Man sieht, dass die Varianz der Projektion der Verteilung auf die x-Achse in Abhängigkeit vom Winkel Werte zwischen σ_{maj}^2 und σ_{min}^2 annimmt. Diese Funktion wurde bereits in Abb. 37 oben (blaue Kurve) dargestellt.

- **Mittelwert und Kovarianz unter affiner Transformation**
  (*Uncertainty Propagation*)

  Lemma 2. *Sei $\vec{X} \in \boldsymbol{R}^p$ eine p-dimensionale Zufallsvariable mit Mittelwert $\boldsymbol{\mu}$ und Kovarianzmatrix $\boldsymbol{\Sigma}$. Dann berechnen sich Mittelwert und Varianz der unter der affinen Transformation*

$$\vec{Y} = \mathbf{F}\vec{X} + \mathbf{H}, \tag{217}$$

  $\mathbf{F} \in \boldsymbol{R}^{q \times p}, \mathbf{H} \in \boldsymbol{R}^q, q \leq p$, erhaltenen Zufallsvariablen $\vec{Y}$ wie folgt

$$\mathcal{E}[\vec{Y}] = \mathbf{F}\boldsymbol{\mu} + \mathbf{H} \tag{218}$$

$$Var[\vec{Y}] = \mathbf{F}\boldsymbol{\Sigma}\mathbf{F}^T. \tag{219}$$

  Eq. 218 folgt direkt aus Lemma 1, Eq. 219 erhält man durch Einsetzen von Eq. 217 und Eq. 218 in $\mathcal{E}[(\vec{Y} - \mathcal{E}[\vec{Y}])(\vec{Y} - \mathcal{E}[\vec{Y}])^T]$.

Die zuvor behandelte Projektion ist klarerweise ein Spezialfall des obigen Lemmas (mit $q = 1$).

Alle gültigen (positiv semi-definiten) Kovarianzmatrizen lassen sich erzeugen, indem man in Gl. 219 für $\mathbf{\Sigma}$ eine Diagonalmatrix mit nicht-negativen Elementen (entspricht dekorrelierten Zufallsvariablen bzw. Ellipsen in Hauptform) und für $\mathbf{F}$ eine Drehmatrix, also $\det \mathbf{F} = 1$, wählt. Da für Drehmatrizen $\mathbf{F}^T = \mathbf{F}^{-1}$ gilt, wird Gl. 219 in diesem Fall zu einer **Ähnlichkeitstransformation**.

- Ist $\vec{X}$ normalverteilt mit $\vec{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, so ist die Verteilung der transformierten Variablen $\vec{Y} = \mathbf{F}\vec{X} + \mathbf{H}$ durch $\vec{Y} \sim N(\mathbf{F}\boldsymbol{\mu} + \mathbf{H}, \mathbf{F}\boldsymbol{\Sigma}\mathbf{F}^T)$ gegeben (dieses Ergebnis folgt nicht trivial aus Lemma 2).

  Weiters sind die Randverteilungen und bedingten Verteilungen einer multivariat normal verteilten Zufallsvariablen wiederum multivariat normal.

- **Korrelation**

  Die Kovarianz

$$Cov(X,Y) = \sigma_{XY} = \mathcal{E}[(X-\mu_x)(Y-\mu_y)] \tag{220}$$

  ist ein Maß für den linearen Zusammenhang zwischen X und Y. Allerdings hängt die Kovarianz auch von der Varianz (Skalierung) der Variablen ab

$$\begin{aligned} Var(\alpha X) &= \mathcal{E}[(\alpha(X-\mu_x))^2] = \alpha^2 Var(X) && (221)\\ Cov(\alpha X, Y) &= \mathcal{E}[(\alpha(X-\mu_x))(Y-\mu_y)] = \alpha Cov(X,Y). && (222) \end{aligned}$$

Ein skalierungsunabhängiges Maß für den linearen Zusammenhang ist durch die Korrelation

$$Corr(X,Y) = Cov(X,Y)/\sqrt{Var(X)Var(Y)} \quad (223)$$

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad (224)$$

gegeben, welche man aus der Kovarianz durch Division durch das Produkt der Standardabweichungen der betreffenden Variablen erhält.

Für den **Korrelationskoeffizienten** ρ_{XY} gilt

$$-1 \leq \rho_{XY} \leq 1, \quad (225)$$

wobei im Fall $|\rho_{XY}| = 1$ ein perfekter (deterministischer) linearer Zusammenhang zwischen X und Y besteht. Im Fall $\rho_{XY} = 0$ besteht keinerlei linearer Zusammenhang zwischen den Variablen (sie sind *dekorreliert*).

Aus der Definition des Korrelationskoeffizienten Eq. 224 folgt

$$\sigma_{XY} = \rho_{XY}\,\sigma_X\,\sigma_Y. \tag{226}$$

Daher muss die Kovarianz stets im Intervall $[-\sigma_X\,\sigma_Y, \sigma_X\,\sigma_Y]$ liegen.

Für Z-standardisierte Variablen $Z_1 = (X - \mu_X)/\sigma_X, Z_2 = (Y - \mu_Y)/\sigma_Y$ ($Var(Z_1) = Var(Z_2) = 1$) erhält man

$$Corr(Z_1, Z_2) = Cov(Z_1, Z_2)/(1 * 1), \tag{227}$$

d.h. die Kovarianz ist gleich der Korrelation. Weiter ist der Korrelationskoeffizient unter Z-Normalisierung (Skalierung der Achsen) invariant

$$\begin{aligned} Corr(Z_1, Z_2) &= \mathcal{E}[(X - \mu_X)/\sigma_X\,(Y - \mu_Y)/\sigma_Y] && (228)\\ &= \sigma_{XY}/(\sigma_X\sigma_Y) = Corr(X, Y). && (229) \end{aligned}$$

In Abb. 38 ist der Korrelationskoeffizient zwischen X und Y für die 4 unterschiedlich elongierten Ellipsenscharen/Kovarianzmatrizen in Abb. 36, wieder als Funktion des Winkels zwischen Haupt- und x-Achse, dargestellt. Zusätzlich ist der Verlauf dieser Kurve für eine fast in eine Gerade degenerierte Ellipse (in magenta) dargestellt. Je größer die Elongation, desto mehr nähert sich der absolute Korrelationskoeffizient dem größtmöglichen Wert 1 an. Im Limit (magenta) wird die Kurve stückweise konstant mit $\rho = 1$ bzw. $\rho = -1$, mit Unstetigkeitsstellen bei 0^o und 90^o (hier wird $\rho = 0$).

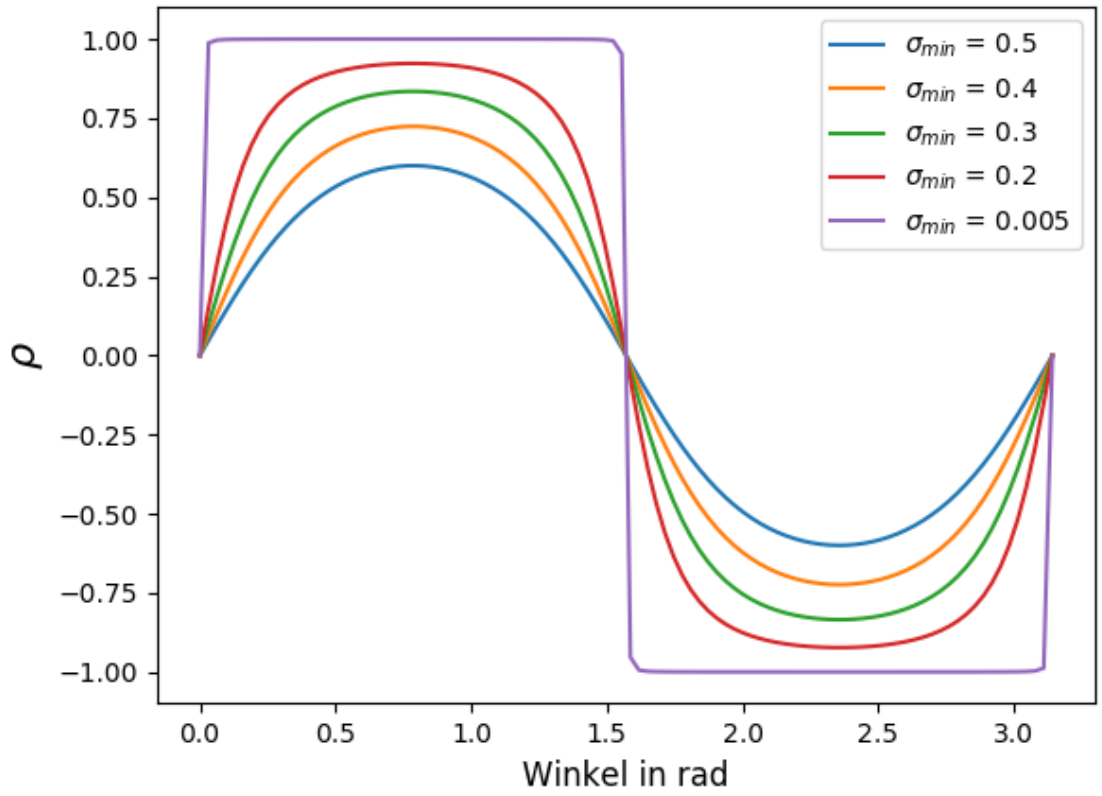


Abbildung 38: Korrelation als Funktion des Rotationswinkels für die 4 Matrizen in Abb. 36 sowie für eine fast degenerierte Ellipse (magenta).

Abb. 39 illustriert das bisher gesagte noch einmal. Eine bivariate Normalverteilung mit Kovarianzmatrix $\mathbf{\Sigma}$ hat eine elliptische Form, wobei die Hauptachse in Richtung der größten Varianz

$$\mathbf{w}^* = \arg\max \frac{\mathbf{w}^T \mathbf{\Sigma} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \tag{230}$$

liegt (die Nebenachse liegt in Richtung der minimalen Varianz). Für $\rho = 0$ fallen die Achsen der Ellipse mit den Koordinatenachsen x_i zusammen.

Werden die Variablen Z-standardisiert, so liegt die Hauptachse der Ellipse auf der ersten ($\rho > 0$) bzw. auf der zweiten ($\rho < 0$) Mediane. Das Verhältnis der Achsen der Ellipse hängt vom Absolutbetrag des Korrelationskoeffizienten ρ ab: je größer $|\rho|$, desto elongierter, je kleiner $|\rho|$, desto kreisförmiger die Ellipse. Für $\rho = 0$ erhält man einen perfekten Kreis (d.h., es gibt keine "ausgezeichnete" Hauptachse mehr).

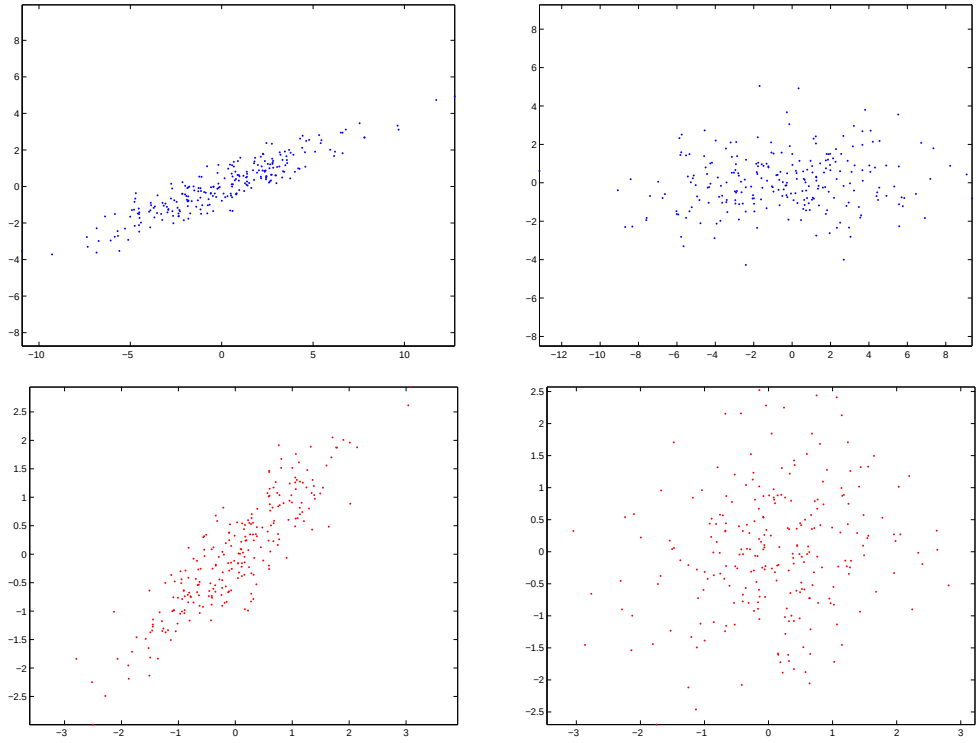


Abbildung 39: **Kovarianz vs. Korrelation am Beispiel einer bivariaten Normalverteilung.**
Oben: $\sigma_X^2 = 12, \sigma_Y^2 = 2$. Unten: Z-standardisierte Variablen ($\sigma_X^2 = \sigma_Y^2 = 1$). Links: $\rho_{XY} = 0.9$. Rechts: $\rho_{XY} = 0.1$

- **Schätzung des Korrelationskoeffizienten**

  Ein erwartungstreuer Schätzer des Korrelatioskoeffizienten (Stichproben-Korrelationskoeffizient bzw. *sample correlation coefficient*) ist durch

$$\hat{\rho} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \, \hat{\sigma}_Y} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2 \sum_{i=1}^{N}(Y_i - \bar{Y})^2}} \tag{231}$$

  gegeben. Einen konkreten Schätzwert erhält man, wie gehabt, durch Ersetzen der Zufallsvariablen X_i, Y_i durch die Elemente einer gegebenen Stichprobe

$$r = \frac{\hat{s}_{XY}}{\hat{s}_X \, \hat{s}_Y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{232}$$

Wir haben an dieser Stelle die vorherrschende Konvention übernommen, den Schätzwert des Korrelationskoeffizienten mit r (ohne Dach) zu bezeichnen.

Im Falle einer bivariaten Normalverteilung von X, Y mit $\rho = 0$ ist die Statistik

$$T = \hat{\rho} \sqrt{\frac{N-2}{1-\hat{\rho}^2}} \tag{233}$$

Student-t verteilt mit $N-2$ Freiheitsgraden. Mittels dieser Statistik kann also ein Hypothesentest $H_1 : |\rho| > 0$ unter der Null-Hypothese $H_0 : \rho = 0$ konstruiert werden. Liegt der beobachtete Wert (die Realisierung von T) z.B. außerhalb des Intervalls $[t_{0.025;N-2}, t_{0.975;N-2}]$, so wird die H_0:

$$X \textit{ und } Y \textit{ sind unkorreliert}$$

auf dem 5%-Niveau abgelehnt ($t_{\alpha;N-2}$ bezeichne hier das α-Quantil der Student-t Verteilung mit $N-2$ Freiheitsgraden).

Eigenwertzerlegung und Hauptachsentransformation

- Wir haben gesehen, dass ursprünglich dekorrelierte Zufallsvariablen durch Drehung der Verteilung gegenüber den Achsen des Merkmalsraums korreliert werden: geometrisch entspricht dies der Drehung einer in Hauptlage befindlichen Schar von Kovarianz-Ellipsen, vektoralgebraisch einer Ähnlichkeitstransformation einer Diagonalmatrix $\mathbf{\Lambda}$:

$$\mathbf{\Sigma} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^T, \quad \det(\mathbf{E}) = 1. \tag{234}$$

  Wir zeigen im folgenden, wie sich $\mathbf{\Lambda}$ und $\mathbf{E}$ für eine gegebene Kovarianzmatrix $\mathbf{\Sigma}$ bestimmen lassen.

- **Eigenwert-Zerlegung** *(Eigenvalue Decomposition EVD)*

  Sei $\mathbf{A} \in \mathbf{R}^{p \times p}$ eine quadratische Matrix. Gilt für ein $\mathbf{e} \in \mathbf{C}^p, \mathbf{e} \neq \mathbf{0}$ und einen Skalar $\lambda \in \mathbf{C}$

  $$\mathbf{A}\mathbf{e} = \lambda \mathbf{e}, \tag{235}$$

  so nennen wir $\mathbf{e}$ einen **Eigenvektor** von $\mathbf{A}$ mit korrespondierendem **Eigenwert** $\lambda = \lambda(\mathbf{e})$. Man beachte, dass mit $\mathbf{e}$ auch jedes Vielfache $\alpha\mathbf{e}, \alpha \in \mathbf{R}$ ein Eigenvektor von $\mathbf{A}$ mit Eigenwert λ ist, d.h., ein Eigenvektor legt einen eindimensionalen Unterraum fest.

Die Eigenwerte erhält man z.B. als Lösung der Gleichung

$$p_{\mathbf{A}}(\lambda) = |\mathbf{A} - \lambda\mathbf{I}| = \prod_{i}^{p}(\lambda - \lambda_i) = 0, \tag{236}$$

d.h. als Nullstellen des *charakteristischen Polynoms* $p_{\mathbf{A}}(\lambda)$ von $\mathbf{A}$.

$p_{\mathbf{A}}(\lambda)$ ist ein Polynom p-ter Ordnung in λ, und hat somit p (möglicherweise komplexe) Lösungen. Somit verfügt jede $p \times p$-Matrix über p Eigenwert/Eigenvektor-Paare $(\lambda_i, \mathbf{e}_i)$.

Spezialfälle:

- **0-Eigenwerte**: treten im Fall singulärer Matrizen für Eigenvektoren im Kern der Matrix ($\{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{0}\}$) auf.
- **Multiple Eigenwerte**, d.h. $\lambda_i = \lambda_j, i \neq j$, es tritt also mindestens ein Eigenwert mit Vielfachheit > 1 auf. Eine Linearkombination von Eigenvektoren $\mathbf{e}_{m_i}, \mathbf{e}_{m_j}$, welche über denselben Eigenwert λ_m mit Vielfachheit m verfügen, ist wiederum ein Eigenvektor von $\mathbf{A}$:

$$\mathbf{A}(\alpha_{m_i}\mathbf{e}_{m_i} + \alpha_{m_j}\mathbf{e}_{m_j}) = \lambda_m(\alpha_{m_i}\mathbf{e}_{m_i} + \alpha_{m_j}\mathbf{e}_{m_j}), \tag{237}$$

d.h. sie spannen einen maximal m-dimensionalen Unterraum des $\mathbf{R}^p$ auf.

Beispiel

Die Matrix $\mathbf{A}_d = \text{diag}([1, 1, 0])$ (d.h, die 3×3-Einheitsmatrix $\mathbf{I}_3$, deren drittes Diagonalelement 0 gesetzt wurde) ist singulär, mit dem vom kanonischen Vektor $\mathbf{e}_3 = (0, 0, 1)^T$ aufgespannten Subraum als Kern; $\mathbf{e}_3$ erfüllt die Eigenvektor-Gleichung (235) mit Eigenwert 0

$$\mathbf{A}_d \, \mathbf{e}_3 = \mathbf{0} = 0 \, \mathbf{e}_3.$$

Wie man sich leicht überzeugt, sind die kanonischen Basisvektoren der Ebene $\mathbf{e}_1 = (1, 0, 0)^T$ und $\mathbf{e}_2 = (0, 1, 0)^T$ beide Eigenvektoren von $\mathbf{A}_d$ mit Eigenwert 1. Somit ist jeder Punkt der Ebene $(x, y, 0), x \, y \, \# \, 0$ ebenfalls ein Eigenvektor von $\mathbf{A}_d$ mit Eigenwert 1!

Fassen wir nun die p Eigenvektoren von $\mathbf{A}$ in der Eigenvektormatrix $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_p)$ und die zugehörigen Eigenwerte in der Diagonalmatrix $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$ zusammen, so lässt sich (235) für alle p Eigenvektoren simultan als

$$\mathbf{AE} = \mathbf{E\Lambda} \tag{238}$$

formulieren. Sind die Eigenvektoren darüberhinaus linear unabhängig, so ist $\mathbf{E}$ invertierbar und wir erhalten mit

$$\mathbf{A} = \mathbf{E\Lambda E}^{-1} \tag{239}$$

die **Eigenwertzerlegung** (*eigenvalue decomposition, EVD*, auch *spectral factorization*) von $\mathbf{A}$.

Im Fall einer symmetrischen, reellen Matrix $\mathbf{A}$ gelten folgende Aussagen

- $\mathbf{A}$ hat ausschließlich reelle Eigenwerte und Eigenvektoren.
- Zu verschiedenen Eigenwerten gehörende Eigenvektoren sind orthogonal. Auch im Fall von Eigenwerten mit Vielfachheit > 1 (oder 0-Eigenwerten) lassen sich stets p wechselseitig orthogonale Eigenvektoren finden.

Normalisieren wir die Eigenvektoren weiters auf Einheitslänge, so ist $\mathbf{E}$ eine **Orthonormalmatrix** (mit $|\mathbf{E}| = \pm 1$). Da die Inverse einer Orthonormalmatrix durch ihre Transponierte gegeben ist, d.h. $\mathbf{E}^{-1} = \mathbf{E}^T$, erhalten wir für Eq. 239

$$\mathbf{A} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^T = \sum_{i=1}^{p} \lambda_i \mathbf{e}_i \mathbf{e}_i^T. \tag{240}$$

Man bemerkt, dass die Eigenwertdekomposition Eq. 239 nicht eindeutig ist, da wir die Eigenvektor/Eigenwert-Paare (Zeilen von $\mathbf{E}$ bzw. $\mathbf{\Lambda}$) beliebig permutieren können.

Wir gehen im folgenden davon aus, dass die Eigenwerte absteigend sortiert sind, d.h. $\lambda_1 \geq \lambda_2 \ldots \lambda_{p-1} \geq \lambda_p$. Unter dieser Konvention wird $\mathbf{e}_1$ ($\mathbf{e}_p$) auch als größter (kleinster) Eigenvektor bezeichnet.

- **Invertierung einer reellen symmetrischen Matrix**

  Die Inverse einer symmetrischen Matrix $\mathbf{A}$ mit Eigenwertzerlegung

$$\mathbf{A} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^T \tag{241}$$

  ist durch

$$\mathbf{A}^{-1} = \mathbf{E}\mathbf{\Lambda}^{-1}\mathbf{E}^T = \mathbf{E}\ \text{diag}(\lambda_1^{-1}, \ldots, \lambda_p^{-1})\ \mathbf{E}^T \tag{242}$$

  gegeben, lässt sich also durch Invertieren der Eigenwerte berechnen. $\mathbf{A}^{-1}$ besitzt somit dieselben Eigenvektoren wie $\mathbf{A}$, jedoch mit reziproken Eigenwerten.

  Insbesondere ist die Inverse einer symmetrischen Matrix wiederum symmetrisch.

- **Beziehung zwischen Rayleigh Quotient und EVD**
  Eine notwendige Bedingung dafür, dass der Rayleigh Quotient

$$r(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{A} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \tag{243}$$

  im Punkt $\mathbf{w}$ ein Extremum annimmt, ist durch

$$\nabla r(\mathbf{w}) = (\frac{dr(\mathbf{w})}{d\mathbf{w}})^T = (\partial r(\mathbf{w})/\partial w_1, \ldots, \partial r(\mathbf{w})/\partial w_p)^T = \mathbf{0} \tag{244}$$

  gegeben, wobei $\nabla r(\mathbf{w}) \in \mathbf{R}^p$ den **Gradienten** von r bezeichnet (der Gradient ist die Transponierte der Funktionalmatrix bzw. der ersten Ableitung von r nach $\mathbf{w}$).

  Die Extremstellen $\mathbf{w}^*$, welche Eq. 244 erfüllen, werden im Englischen auch *stationary points* genannt.

Lemma 3. *Die Extremstellen $\mathbf{w}^*$ (Extremwerte $r(\mathbf{w}^*)$) des Rayleigh-Quotienten Eq. 243 sind durch die Eigenvektoren $\mathbf{e}$ (Eigenwerte $\lambda(\mathbf{e})$) von $\mathbf{A}$ gegeben, können also als Lösungen des korrespondierenden symmetrischen Eigenwertproblems erhalten werden.*

- **Diagonalisierung der Kovarianzmatrix**

  Betrachten wir nun die EVD der (symmetrischen!) Kovarianzmatrix $\mathbf{\Sigma}$ von $\vec{X}$. Aus Eq. 240 folgt, dass

$$\mathbf{E}^T \mathbf{\Sigma} \mathbf{E} = \mathbf{\Lambda}. \tag{245}$$

  Man sieht, dass die durch

$$\vec{Y} = \mathbf{E}^T \vec{X} \tag{246}$$

$$\vec{Y} = \mathbf{E}^T (\vec{X} - \boldsymbol{\mu}_x) \tag{247}$$

  (sprich: durch Projektion auf die Eigenvektoren) gegebenen affinen Abbildungen die Kovarianzmatrix diagonalisieren[10].

[10] Eine allfällige Mittelwertnormalisierung hat auf die Kovarianzmatrix von $\vec{Y}$ natürlich keinen Einfluß, so daß beide Abbildungsvorschriften als Diagonalisierung bezeichnet werden. Wir werden im folgenden jedoch von Gl. 247 ausgehen.

Der i-te Eigenwert λ_i entspricht der Varianz der Projektion auf den i-ten Eigenvektor $Y_i = \mathbf{e}_i^T \vec{X}$, d.h. $\lambda_i = Var(Y_i)$. Weiters sind die Komponenten Y_i dekorreliert, da $Cov(Y_i, Y_j) = \lambda_{ij} = 0$ für $i \neq j$.

Die Eigenvektoren erklären sukzessive (absteigend vom größten zum kleinsten) maximale Varianz (siehe auch vorhergehendes Lemma).

Die Eigenvektoren $\mathbf{e}_i$ entsprechen den Achsen der Ellipsoide konstanter pdf (iso-Linien bzw. iso-Flächen) von $\vec{X}$ und $\vec{Y}$, wobei die Achsenlängen proportional zu den Quadratwurzeln der Eigenwerte $\sqrt{\lambda_i}$ (Standardabweichungen $\sqrt{\sigma_{ii}}$ von $\vec{Y}$) sind.

Geometrisch kann Eq. 247 als Transformation des ursprünglichen Koordinatensystem $\mathcal{C}_x$ aufgefasst werden, wobei

- der Ursprung des neuen Systems $\mathcal{C}_y$ (relativ zu $\mathcal{C}_x$) durch $\boldsymbol{\mu}_x$ gegeben ist und
- die Achsen des neuen System (relativ zu $\mathcal{C}_x$) durch die Eigenvektoren (Achsen des Ellipsoide konstanter pdf) gegeben sind.

Fig. 40 auf der nächsten Seite veranschaulicht diesen Prozess anhand einer bivariaten Normalverteilung mit Kovarianzmatrix $\begin{pmatrix} 12 & 4.41 \\ 4.41 & 2 \end{pmatrix}$. Die Kovarianzmatrix der diagonalisierten Verteilung ist durch $diag(13.66, 0.33)$ gegeben.

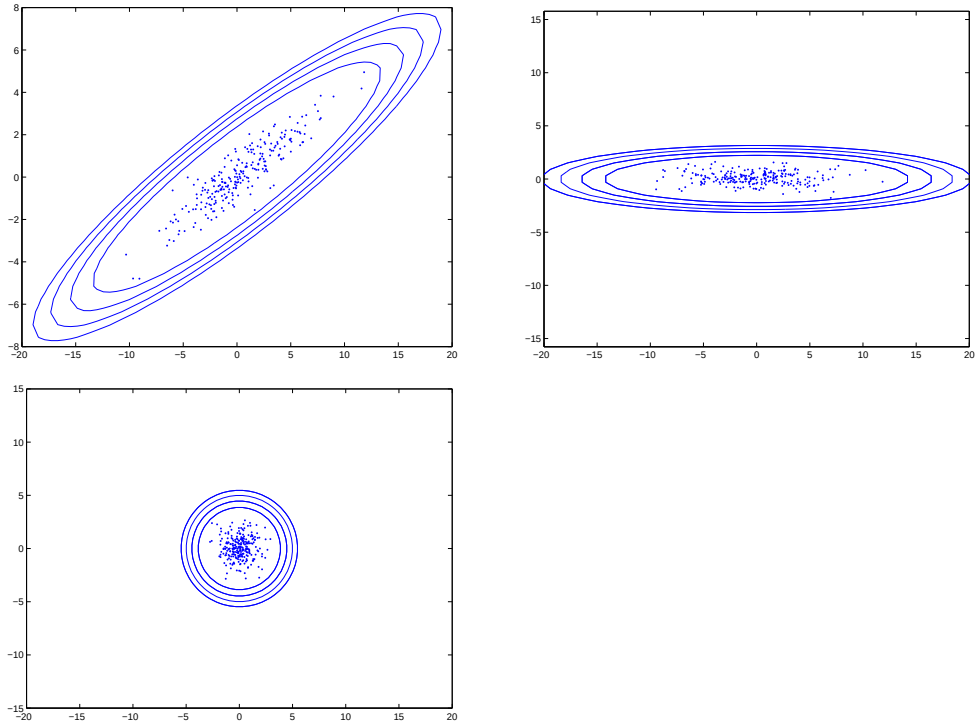


Abbildung 40: **KLT und Whitening**
Von links oben nach rechts unten: Ursprüngliche Verteilung, diagonalisierte Verteilung (die Achsen der Ellipsoide koinzidieren mit den Achsen des Koordinatensystems), *whitened distribution* mit Kovarianzmatrix $diag(1, 1)$.

- **Totale Varianz**

  Die **totale Varianz** ist als Summe der Diagonalelemente der Kovarianzmatrix, Spur($\mathbf{\Sigma}$), definiert. Die Spur *(trace)* einer Matrix ist invariant gegenüber zyklischen Permutationen. Wir haben somit:

$$\text{Spur}(\mathbf{\Lambda}) = \text{Spur}(\mathbf{E\Sigma E}^T) = \text{Spur}(\mathbf{E}^T\mathbf{E\Sigma}) = \text{Spur}(\mathbf{\Sigma}). \qquad (248)$$

  Im speziellen erhält man die totale Varianz als Summe der Eigenwerte. In Abhängigkeit von der Drehung (Eigenvektormatrix) wird die totale Varianz unterschiedlich auf die Koordinaten-Achsen aufgeteilt, bleibt aber in Summe gleich.

- **Invarianz der Mahalanobis-Distanz**

  Unter Verwendung der Identitäten $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$ und $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ erhalten wir

$$\begin{aligned} d^2(\mathbf{y}) &= (\mathbf{y}-\boldsymbol{\mu}_y)^T\boldsymbol{\Lambda}^{-1}(\mathbf{y}-\boldsymbol{\mu}_y) \\ &= (\mathbf{E}^T(\mathbf{x}-\boldsymbol{\mu}_x))^T(\mathbf{E}^T\boldsymbol{\Sigma}\mathbf{E})^{-1}\mathbf{E}^T(\mathbf{x}-\boldsymbol{\mu}_x) \\ &= (\mathbf{x}-\boldsymbol{\mu}_x)^T\mathbf{E}\mathbf{E}^{-1}\boldsymbol{\Sigma}^{-1}(\mathbf{E}^T)^{-1}\mathbf{E}^T(\mathbf{x}-\boldsymbol{\mu}_x) \\ &= (\mathbf{x}-\boldsymbol{\mu}_x)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_x) = d^2(\mathbf{x}), \end{aligned} \tag{249}$$

  d.h. $d^2(\mathbf{x})$ ist unter $\mathbf{E}^T$ (allgemein: unter jeder invertierbaren linearen Transformation) invariant.

- **Karhunen-Loeve Transformation**

  Aus vektoralgebraischer Sicht entspricht die Transformation

$$\mathbf{y} \quad = \quad \mathbf{E}^T(\mathbf{x} - \boldsymbol{\mu}_x) = \mathbf{E}^T\tilde{\mathbf{x}} \tag{250}$$

  einem Basiswechsel von der kanonischen Basis zur Basis $\mathbf{E}$ (bzg. der mittelwert-normalisierten Koordinaten $\tilde{\mathbf{x}}$). Man spricht in diesem Zusammenhang von der

  – (diskreten) **Karhunen-Loeve Transformation** (KLT),
  – **Hauptachsentransformation** oder
  – **Principal Components Analysis** (PCA.)

  Achtung: die absteigende Sortierung der Eigenwerte/Eigenvektoren ist hier wesentlich.

Für einen Punkt $\mathbf{y}$ ist dessen Repräsentation bzg. der kanonischen Basis (Urbild) durch die inverse Transformation

$$\tilde{\mathbf{x}} \quad = \quad \mathbf{E}\mathbf{y} = \sum_{i=1}^{p} \mathbf{e}_i y_i \tag{251}$$

gegeben. Eq. 251 ist die **Rekonstruktion** (*Karhunen-Loeve expansion*) von $\tilde{\mathbf{x}}$, wobei sich die Koeffizienten der Linearkombination gemäß Eq. 250 berechnen.

Lassen wir in Gl. 251 jene Richtungen, welche den $p-k$ kleinsten Eigenwerten (d.h. Varianzen) entsprechen, weg, so erhalten wir eine eine unvollständige Rekonstruktion

$$\tilde{\mathbf{x}}^{(k)} = \mathbf{E}_{[k]}\mathbf{y}_{[k]} = \sum_{i=1}^{k} \mathbf{e}_i y_i, \in \mathbf{R}^p, k < p \tag{252}$$

des Originalvektors $\tilde{\mathbf{x}}$. Wie man leicht zeigt, entspricht der Erwartungswert des Rekonstruktionsfehlers der Summe der Eigenwerte der weggelassenen Richtungen

$$\mathcal{E}[\|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}^{(k)}\|^2] = \sum_{i=k+1}^{p} \lambda_i \tag{253}$$

Es lässt sich zeigen, dass die KLT-Basis von allen Basen der Dimension $k < p$ den mittleren Rekonstruktionsfehler im obigen Sinne minimiert, also bzg. der Minimierung des Rekonstruktionsfehlers optimal ist.

Diesen Zusammenhang kann man für Datenkomprimierung bzw. Dimensionalitätsreduktion verwenden. Statt der Originaldaten speichert man nur die Projektionen auf die k größten Eigenvektoren. Die Originale lassen sich dann aus den Projektionen und den Eigenvektoren rekonstruieren.

- **Whitening**

  Skaliert man die Basisvektoren $\mathbf{e}_i$ der KLT mit $\lambda_i^{-\frac{1}{2}}$, so liefert die resultierende Transformation

$$\vec{Y}^w = (\mathbf{E}\mathbf{\Lambda}^{-\frac{1}{2}})^T(\vec{X} - \boldsymbol{\mu}_x) = \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{E}^T(\vec{X} - \boldsymbol{\mu}_x) \tag{254}$$

  einen Zufallsvektor mit dekorrelierten und Z-normalisierten Variablen ($Var(Y_i) = 1$ für $1 \leq i \leq p$). Die resultierende Verteilung ist kreisförmig; man spricht auch von **Whitening**.

  Genauer wird $\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{E}^T$ (manchmal jedoch auch $\mathbf{E}\mathbf{\Lambda}^{-\frac{1}{2}}$) als **Whitening-Transformation** bezeichnet.

Betrachten wir das Quadrat der Länge eines gemäß Gl. 254 transformierten Vektors $\mathbf{x}$:

$$\begin{aligned} \|\mathbf{y}\|^2 &= \mathbf{y}^T\mathbf{y} \\ &= (\mathbf{x}-\boldsymbol{\mu}_x)^T\mathbf{E}\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{E}^T(\mathbf{x}-\boldsymbol{\mu}_x) \\ &= (\mathbf{x}-\boldsymbol{\mu}_x)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_x), \end{aligned} \tag{255}$$

wobei wir beim Übergang von der vorletzten zu letzten Zeile Gl. 242 benutzt haben. Die euklidische Norm eines durch Whitening erhaltenen Vektors entspricht also der Mahalanobis-Distanz seines Urbilds; dies gilt klarerweise auch für Distanzen zwischen zwei beliebigen Punkten im Bild- bzw. Urbildraum.

Um unsere Eingangsgrößen zu standardisieren, können wir also entweder explizit ein Whitening durchführen, oder wir verwenden bei Distanzberechnungen statt der euklidischen Metrik die Mahalanobis-Distanz.

- **Verteilung der Mahalanobis-Distanz**

  Die Summe der Quadrate von p standardnormalverteilten Größen ist Chi-Quadrat verteilt mit p Freiheitsgraden. Ersetzen wir in Gl. 255 die Variablen durch Zufallsvariablen, ist unmittelbar ersichtlich, daß die Mahalanobis-Distanz eine solche Summe darstellt (es gilt $Y_i \sim N(0,1)$). Somit ist diese ebenfalls Chi-Quadrat verteilt mit p Freiheitsgraden.

- **Beispiele für Anwendungen der KLT**

  - Zufallszahlengenerator: Mittels der inversen Whitening-Transformation lassen sich aus Vektoren von je p $N(0,1)$ verteilten samples $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ verteilte samples generieren.
  - Schätzung der Orientierung einer Punktwolke, Extraktion von Ebenen in 3D-Punktwolken.
  - Merkmalsberechnung in Bildern, z.B. Elongation (definiert als $\frac{\lambda_1}{\lambda_2}$, Kanten- und Eck-Detektion.
  - Komprimierung: Bilder eines Objekts lassen sich als Linearkombination einiger weniger Bilder (*eigenimages*) darstellen.

- **Exkurs: KLT und Faktoranalyse**

  Es besteht ein enger Zusammenhang zwischen der KLT und der sogenannten Faktoranalyse, welche vor allem in der Psychologie eingesetzt wird. Unterschiede bestehen hauptsächlich in den Grundannahmen bezüglich der Kovarianzstruktur der Fehler, worauf hier aber nicht näher eingegangen werden soll. Ziel ist es, Vektoren von p korrelierten Variablen als Linearkombination von $k < p$ sogenannten Faktoren darzustellen.

  - Unter den **Faktoren** versteht man die mit den Wurzeln ihrer korrespondierenden Eigenwerte (d.h. Standardabweichungen) skalierten Eigenvektoren $\mathbf{E}\mathbf{\Lambda}^{\frac{1}{2}}$.
  - Die Elemente der Faktoren bezeichnet man als **Faktorladungen**. Die Faktorladungen sind die Kovarianzen zwischen den X_i und den gemäß Gl. 254 erhaltenen Y_j^w.

– Die Elemente y_i der Ausprägungen der transformierten Größe Gl. 254

$$\mathbf{y}^w = (y_1, ..., y_k)^T = (\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{E}^T)_{[k]}(\mathbf{x} - \boldsymbol{\mu}_x) \tag{256}$$

bezeichnet man als **Faktorwerte**.

Regression 1: Die bedingte multivariate Normalverteilung

- Wir betrachten die Zerlegung eines multivariat normalverteilten Zufallsvektors in 2 Teilvektoren:

$$\vec{X} \in \mathbf{R}^{p+q} = \begin{pmatrix} \vec{X}^{(1)} \\ \vec{X}^{(2)} \end{pmatrix}, \vec{X}^{(2)} \in \mathbf{R}^{p}, \vec{X}^{(1)} \in \mathbf{R}^{q}, \tag{257}$$

mit

$$Cov(\vec{X}) = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}, \mathbf{\Sigma}_{11} \in \mathbf{R}^{q \times q}, \mathbf{\Sigma}_{22} \in \mathbf{R}^{p \times p}, \mathbf{\Sigma}_{12} \in \mathbf{R}^{q \times p} \tag{258}$$

Wir bestimmen im folgenden eine affine Transformation

$$\vec{Y}^{(1)} = \vec{X}^{(1)} + \mathbf{B}\vec{X}^{(2)} \quad (259)$$

$$\vec{Y}^{(2)} = \vec{X}^{(2)}, \quad (260)$$

welche die Komponenten in $\vec{Y}^{(1)}$ und $\vec{Y}^{(2)}$ dekorreliert, d.h.

$$\mathcal{E}\left[\left(\vec{Y}^{(1)} - \mathcal{E}[\vec{Y}^{(1)}]\right)\left(\vec{Y}^{(2)} - \mathcal{E}[\vec{Y}^{(2)}]\right)^T\right] = \mathbf{0}$$

Die Komponenten der obigen Matrix werden als Kreuz-Kovarianzen oder, nicht ganz korrekt, als **Kreuz-Korrelationen** *(cross correlations)* bezeichnet. Im Falle einer Normalverteilung folgt aus der obigen Bedingung, dass alle Komponenten des Teil-Vektors $\vec{Y}^{(1)}$ von allen Komponenten des Teil-Vektors $\vec{Y}^{(2)}$ unabhängig sind.

Wie lässt sich die obige Abbildung nun bestimmen?

$$
\begin{aligned}
\mathbf{0} &= \mathcal{E}\left[\left(\vec{Y}^{(1)} - \mathcal{E}[\vec{Y}^{(1)}]\right)\left(\vec{Y}^{(2)} - \mathcal{E}[\vec{Y}^{(2)}]\right)^T\right] \\
&= \mathcal{E}\left[\left(\vec{X}^{(1)} + \mathbf{B}\vec{X}^{(2)} - \mathcal{E}[\vec{X}^{(1)} + \mathbf{B}\vec{X}^{(2)}]\right)\left(\vec{X}^{(2)} - \mathcal{E}[\vec{X}^{(2)}]\right)^T\right] \\
&= \mathcal{E}\left[\left((\vec{X}^{(1)} - \mathcal{E}[\vec{X}^{(1)}]) + \mathbf{B}(\vec{X}^{(2)} - \mathcal{E}[\vec{X}^{(2)}])\right)\left(\vec{X}^{(2)} - \mathcal{E}[\vec{X}^{(2)}]\right)^T\right] \\
&= \mathbf{\Sigma_{12}} + \mathbf{B}\mathbf{\Sigma}_{22} \qquad (261)
\end{aligned}
$$

Somit muss

$$\mathbf{B} = -\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1} \qquad (262)$$

gelten.

Die gesuchte Abbildung lässt sich kompakt als Matrix-Vektor Produkt schreiben

$$\begin{pmatrix} \vec{Y}^{(1)} \\ \vec{Y}^{(2)} \end{pmatrix} = \vec{Y} = \begin{pmatrix} \mathbf{I} & -\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \vec{X}. \tag{263}$$

Die Kovarianzmatrix von $\vec{Y}$ berechnet sich gemäß Lemma 2 mit

$$Cov(\vec{Y}) = \begin{pmatrix} \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{22} \end{pmatrix}. \tag{264}$$

Nachdem die Komponenten von $\vec{X}^{(2)}$ und des Differenzvektors $\vec{Y}^{(1)} = \vec{X}^{(1)} - \left(-\mathbf{B}\vec{X}^{(2)}\right)$ wechselseitig dekorreliert sind, enthält $\vec{Y}^{(1)}$ keinerlei Anteile von $\vec{X}^{(1)}$ mehr, die sich linear aus $\vec{X}^{(2)}$ vorhersagen lassen. Im Umkehrschluss gilt, dass $-\mathbf{B}\vec{X}^{(2)}$ die bestmögliche lineare Vorhersage von $\vec{X}^{(1)}$ gegeben $\vec{X}^{(2)}$ darstellt.

Aus dem obigen erhalten wir unter der Normalverteilungsannahme und Verwendung des Transformationssatzes für Verteilungen folgendes Resultat (ohne Beweis):

Lemma 4. *Sei $\vec{X}$ multivariat normal verteilt mit*

$$\vec{X} \in \boldsymbol{R}^{p+q} = \begin{pmatrix} \vec{X}^{(1)} \\ \vec{X}^{(2)} \end{pmatrix}, \tag{265}$$

Dann ist die bedingte Verteilung von $\vec{X}^{(1)}$ unter $\vec{X}^{(2)} = \mathbf{x}^{(2)}$ wiederum normal, mit

$$\mathcal{E}[\vec{X}^{(1)}|\mathbf{x}^{(2)}] = \boldsymbol{\mu}^{(1)} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}) \tag{266}$$

$$Cov(\vec{X}^{(1)}|\mathbf{x}^{(2)}) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \tag{267}$$

$\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}$ ist die Matrix der **Regressionskoeffizienten** der Regression von $\vec{X}^{(1)}$ auf $\mathbf{x}^{(2)}$.

Kenntnis der Regressionskoeffizienten erlaubt es uns also, das Mittel der bedingten Verteilung von $\vec{X}^{(1)}$ als lineare Funktion der Realisierung $\mathbf{x}^{(2)}$ von $\vec{X}^{(2)}$ zu bestimmen. Die Kovarianzmatrix der bedingten Verteilungen ist um $\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$ "kleiner" als $\mathbf{\Sigma}_{11}$ und ist konstant; im speziellen hängt sie nicht von $\mathbf{x}^{(2)}$ ab.

- **Beispiel: Univariate Regression**

  In Abb. 41 ist eine bivariate Normalverteilung der Größen X und Y, mit $\sigma_x = \sigma_y = 2$ für unterschiedliche Werte des Korrelationskoeffizienten $\rho = 0.95$ (links) sowie $\rho = 0.2$ (rechts) dargestellt. Die Kovarianz-Matrizen unterscheiden sich nur in den Off-Diagonalen $\sigma_{xy} = \rho\sigma_x\sigma_y$, die Rand-DFen (grün dargestellt) sind in beiden Fällen identisch.

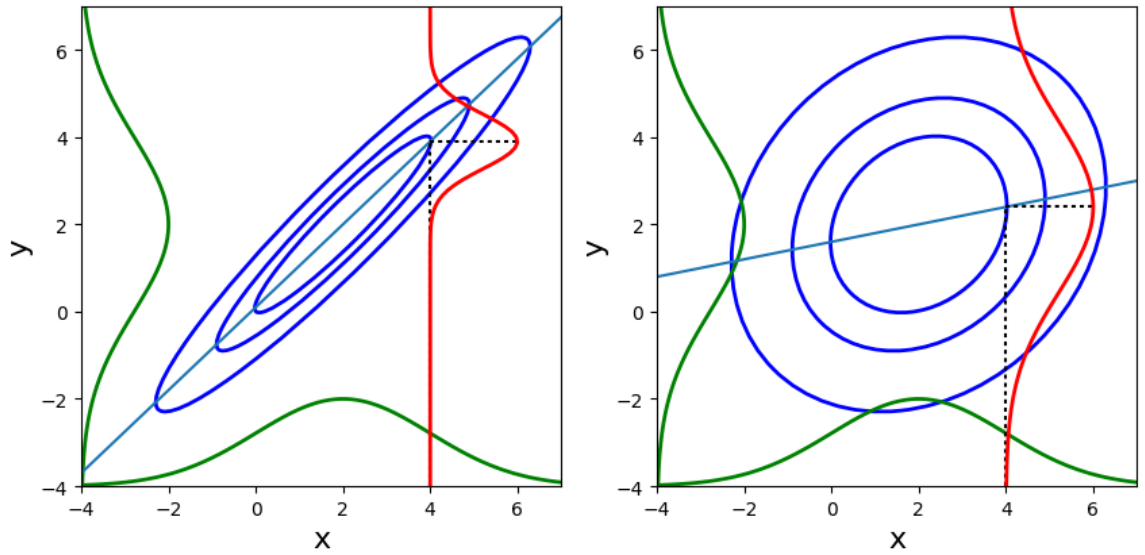


Abbildung 41: Bivariate Normalverteilung mit $\mu_x = 2$, $\mu_y = 2$,$\sigma_x = 2$, $\sigma_y = 2$, und $\rho = 0.95$ (links) und $\rho = 0.2$ (rechts). Die Iso-Ellipsen und die Regressionsgerade sind blau dargestellt, die Rand-Dichtefunktionen grün. Letztere hängen nicht von ρ ab. Die bedingte DF $Y|(x = 4)$ ist rot dargestellt.

Die Steigung der Regessionsgeraden erhalten wir gemäß Gl. 266 mit

$$\frac{\sigma_{xy}}{\sigma_x^2} = \rho \frac{\sigma_y}{\sigma_x} \tag{268}$$

Da wir $\sigma_x = \sigma_y$ angenommen haben, ist die Steigung der Regressionsgeraden somit unmittelbar durch ρ gegeben. Im Fall perfekter positiver Korrelation hat diese eine Steigung von 45 Grad, und fällt mit abnehmender Korrelation auf 0 Grad. Man bemerkt, dass die Richtung der größten projizierten Varianz, d.h. der erste Eigenvektor der Kovarianzmatrix, für positives ρ stets durch die erste Mediane gegeben ist. Der Anteil der auf diese Richtung projizierten Gesamtvarianz wir jedoch mit abnehmendem ρ kleiner.

- **Varianzaufklärung**

  Die Regressionsfunktion liefert das bedingte Mittel von Y im Punkt $X = x$. Die Varianz der bedingten Verteilung $Y|(X = x)$ (in der Abbildung beispielhaft für $x = 4$ dargestellt) erhält man gemäß Gl. 267 mit

$$\sigma_y^2 - \frac{\sigma_{xy}^2}{\sigma_x^2} = \sigma_y^2 - \rho^2\sigma_y^2 = \sigma_y^2(1 - \rho^2) \tag{269}$$

  ρ^2 (meist R^2 geschrieben) ist somit der Anteil der durch die Regression erklärten Varianz von Y, $(1 - \rho^2)$ der nicht erklärte Varianzanteil . Diese werden oft mit Signal- und Rauschvarianz identifiziert.

- **Lineare Regression vs. PCA**

  Lineare Regression entspricht also der Ermittlung des bedingten Mittels der unbekannten bzw. abhängigen Größe(n) $Y|\mathbf{x}$ für bekannte Werte der unabhängigen Größen $\mathbf{x}$ (Prädiktoren). $\mathbf{x}$ wird als Konstante behandelt, anhand derer die Regressionskoeffizienten entweder aus der bekannten Kovarianzmatrix (s.o.) oder durch Minimierung des quadratischen Fehlers aus einer Stichprobe geschätzt werden (s.u.). Die im Regressionsansatz zu minimierenden **Residuen** sind in Abb. 42 grün strichliert dargestellt: sie verlaufen parallel zur Ordinate und normal zu den Prädiktoren.

  PCA hingegen behandelt alle Dimensionen bzw. Variablen als gleichwertig. Das Ziel ist hier nicht, eine Teilmenge der Variablen (bei gegebenen Werten der restlichen Variablen) vorherzusagen, sondern die Verteilung aller Variablen möglichst kompakt zu beschreiben.

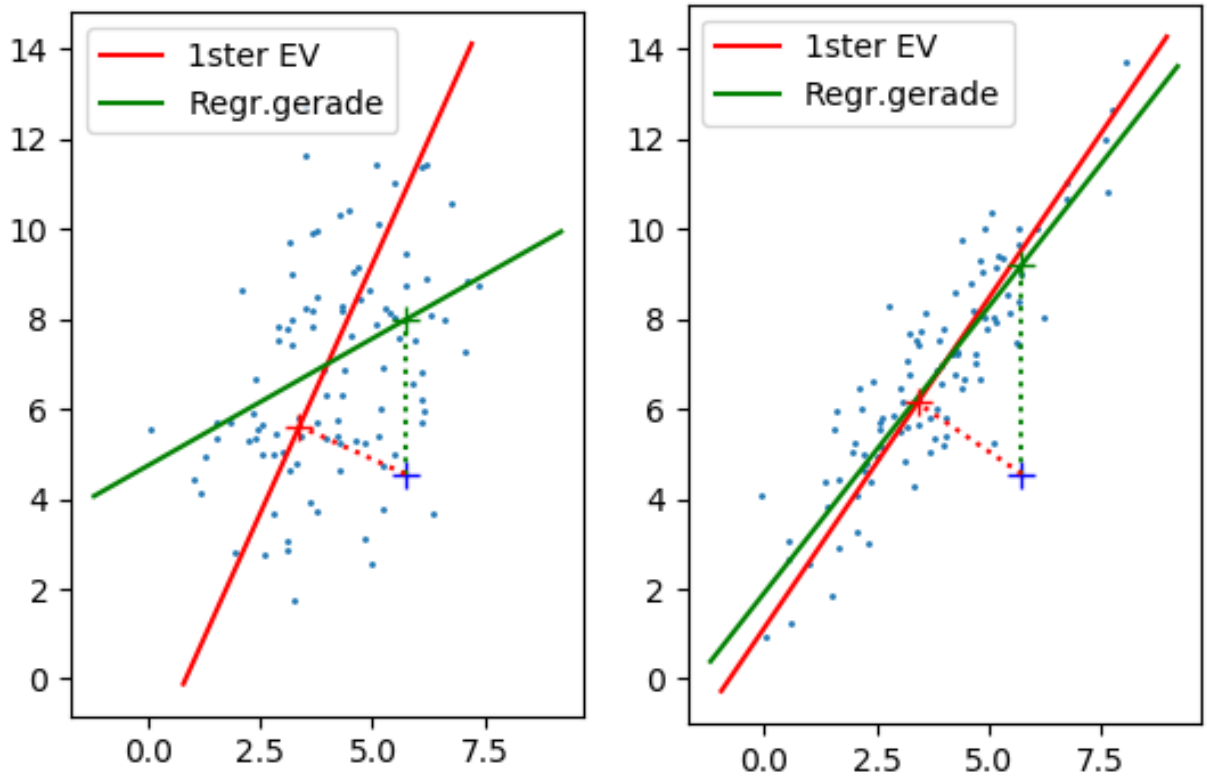


Abbildung 42: Scatterplots (je 100 Punkte) einer bivariaten Normalverteilung mit $\mu_x = 4$, $\mu_y = 7$,$\sigma_x^2 = 3$, $\sigma_y^2 = 6$, und $\rho = 0.4$ (links) und $\rho = 0.9$ (rechts). Dargestellt sind weiters der erste Eigenvektor (rot), die Regressionsgerade (grün) sowie die Residuen für einen gegebenen Punkt (blaues Kreuz).

Große Eigenvektoren entsprechen Richtungen großer Varianz (Erwartungswert der quadratischen Abweichung entlang dieser Richtung), kleine Eigenvektoren (EV) Richtungen kleiner Varianz. Im 2D-Fall minimiert die Projektion auf den größeren EV bzw. das Weglassen des kleineren EVs den quadratischen Fehler der linearen Rekonstruktion; die Residuen verlaufen hier normal zum ersten Eigenvektor.

Da PCA den Rekonstruktionsfehler minimiert, eignet sie sich daher i.a. besser für geometrische Rekonstruktion als ein linearer Regressionsansatz. Ungünstige Verteilung der Daten (z.B Orientierung entlang der y-Achse) führt bei linearer Regression zu extremen Werten für die Koeffizienten (Steigung von 90 Grad!) und daher zu numerischen Problemen und instabilen Ergebnissen.

Bayes-Klassifizierung für normalverteilte Merkmale

- **Diskriminanten-Funktionen**

  Gemäß der MAP-Regel entscheiden wir uns bei gegebenem Merkmalsvektor $\mathbf{x} \in \mathbf{R}^p$ für die Klasse ω_k mit der größten *a posteriori* Wahrscheinlichkeit

$$\alpha(\mathbf{x}) = k = \arg\max_j P(\omega_j|\mathbf{x}), 1 \leq j \leq c. \tag{270}$$

  Die Entscheidungsfunktion $\alpha(\mathbf{x})$ lässt sich allgemeiner durch sogenannte **Diskriminanten-Funktionen** $g_j(\mathbf{x})$ ausdrücken

$$\alpha(\mathbf{x}) = k = \arg\max_j g_j(\mathbf{x}). \tag{271}$$

Die **Entscheidungsgrenze** zwischen den Klassen ω_j und ω_k ist durch die Gleichung

$$g_j(\mathbf{x}) = g_k(\mathbf{x}) \tag{272}$$

gegeben. Berechnen sich die $g_j(\mathbf{x})$ als streng monoton wachsende Funktion der *posteriors*

$$g_j(\mathbf{x}) = f(P(\omega_j|\mathbf{x})), \text{wobei} \tag{273}$$

$$x > y \Rightarrow f(x) > f(y), \tag{274}$$

so ist die Entscheidungsregel Eq. 271 wiederum optimal, z.B. für

$$\begin{aligned} g_j(\mathbf{x}) &= P(\omega_j|\mathbf{x})p(\mathbf{x}) = \frac{P(\omega_j)p(\mathbf{x}|\omega_j)}{p(\mathbf{x})}p(\mathbf{x}) \\ &= P(\omega_j)p(\mathbf{x}|\omega_j). \end{aligned} \tag{275}$$

Sind im speziellen die Merkmale für alle Klassen normalverteilt, d.h. $(\vec{X}|\omega_j) \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ mit DF

$$p(\mathbf{x}|\omega_j) = \frac{1}{(2\pi)^{p\frac{1}{2}}|\boldsymbol{\Sigma}_j|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu_j)^T \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\mu_j)}, \tag{276}$$

so erhält man durch Logarithmieren der *posteriors* die folgenden (opti-

malen) Diskriminantenfunktionen

$$
\begin{aligned}
g_j(\mathbf{x}) &= \ln \frac{P(\omega_j)p(\mathbf{x}|\omega_j)}{p(\mathbf{x})} && (277)\\
&= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T {\boldsymbol{\Sigma}_j}^{-1}(\mathbf{x}-\boldsymbol{\mu}_j) \\
&\quad -\frac{1}{2}\ln|\boldsymbol{\Sigma}_j| + \ln P(\omega_j) \\
&\quad -\frac{p}{2}\ln 2\pi - \ln p(\mathbf{x}). && (278)
\end{aligned}
$$

Man bemerkt, dass die beiden Terme in der letzten Zeile

$$-\frac{p}{2}\ln 2\pi - \ln p(\mathbf{x})$$

nicht von ω_j abhängen und daher beim Vergleich der g_j nicht

berücksichtigt werden müssen.

Die g_j sind im Falle normalverteilter Merkmale somit quadratische Funktionen in $\mathbf{x}$

$$g_j(\mathbf{x}) = -\frac{1}{2}d_j^2(\mathbf{x}) + (-\frac{1}{2}\ln|\mathbf{\Sigma}_j| + \ln P(\omega_j)), \tag{279}$$

wobei $d_j^2(\mathbf{x})$ die Mahalanobis-Distanz der Klasse ω_j bezeichnet.

Wir betrachten im folgenden zwei Spezialfälle, die zu linearen Diskriminantenfunktionen bzw. Entscheidungsgrenzen führen.

- **Naive Bayes $\boldsymbol{\Sigma}_j = \mathbf{I}\sigma$**

  Die Merkmale $X_{ij} = (X_i|\omega_j)$ sind also innerhalb jeder Klasse ω_j dekorreliert ($Cov(X_{ij}, X_{kj}) = 0$ für $i \neq k$) und somit unabhängig. Weiters weisen alle Komponenten dieselbe Varianz auf, d.h. $Var(X_{ij}) = \sigma^2$ für $1 \leq i \leq p, 1 \leq j \leq c$.

  Die g_j berechnen sich als affine Funktion der Mahalanobis-Distanz $d_j^2(\mathbf{x})$

$$g_j(\mathbf{x}) = (-\frac{1}{2})\frac{1}{\sigma^2}(\mathbf{x} - \boldsymbol{\mu}_j)^T(\mathbf{x} - \boldsymbol{\mu}_j) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_j| + \ln P(\omega_j) \quad (280)$$

$$= -\frac{1}{2\sigma^2}(\mathbf{x}^T\mathbf{x} - 2\boldsymbol{\mu}_j^T\mathbf{x} + \boldsymbol{\mu}_j^T\boldsymbol{\mu}_j) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_j| + \ln P(\omega_j) \quad (281)$$

  Nachdem die Terme $-\frac{1}{2}\ln|\boldsymbol{\Sigma}_j|$ und $-\frac{1}{2\sigma^2}\mathbf{x}^T\mathbf{x}$ für alle Klassen gleich sind, können diese weggelassen werden.

Wir erhalten somit die äquivalente **lineare Diskriminantenfunktion**

$$g_j(\mathbf{x}) = \begin{array}{ccc} \frac{1}{\sigma^2}\boldsymbol{\mu}_j^T\mathbf{x} & + & -\frac{1}{2\sigma^2}\boldsymbol{\mu}_j^T\boldsymbol{\mu}_j + \ln P(\omega_j) \\ \mathbf{w}_j^T\mathbf{x} & + & b_j, \end{array} \tag{282}$$

welche für jede Klasse ω_j eine Ebene im $\mathbf{R}^{p+1}$ festlegt. Die Entscheidungsgrenzen $g_j(\mathbf{x}) = g_k(\mathbf{x})$ ergeben sich als Schnittmenge je zweier solcher Ebenen, d.h. als $(p-1)$-dimensionale Hyperebenen im $\mathbf{R}^p$

$$\mathbf{w}^T(\mathbf{x} - \mathbf{b}) \quad = \quad \mathbf{0}, \tag{283}$$

wobei

$$\mathbf{w} \quad = \quad \boldsymbol{\mu}_j - \boldsymbol{\mu}_k \tag{284}$$

$$\mathbf{b} \quad = \quad \frac{1}{2}(\boldsymbol{\mu}_j + \boldsymbol{\mu}_k) - \frac{\sigma^2}{\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_k\|^2} \ln \frac{P(\omega_j)}{P(\omega_k)}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k). \tag{285}$$

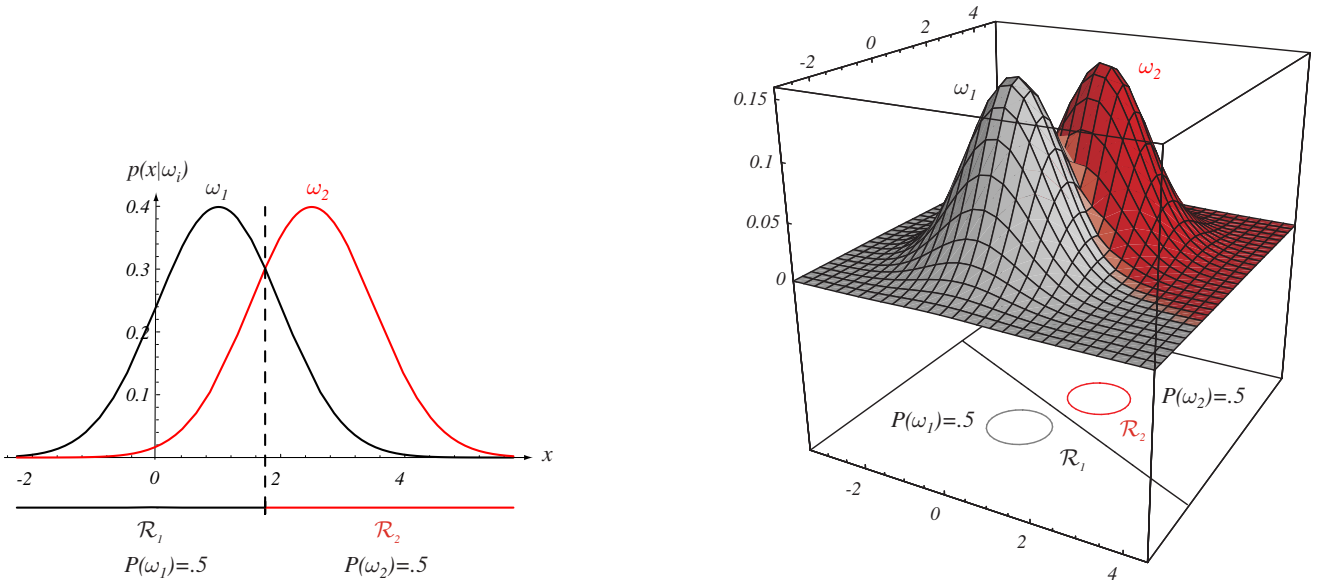


Abbildung 43: Entscheidungsgrenzen für zwei univariate (links) bzw. bivariate (rechts) Normalverteilungen mit $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}\sigma$. Die Entscheidungsgrenzen sind linear und normal zur Verbindungsstrecke zwischen den beiden Klassenmitteln. Für gleiche priors verläuft die Entscheidungsgrenze durch $(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$, ansonsten wird sie von der a priori wahrscheinlicheren Klasse wegverschoben.
(Aus *Duda, Hart, Stork: Pattern Classification, 2nd ed.*)

- **Linear Discriminant Analysis (LDA) $\boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}$**

  Alle Klassen haben dieselbe Kovarianzmatrix (naive Bayes ist also ein Spezialfall von LDA).

  Schreiben wir in Eq. 279 die Mahalonobis-Distanz $d_j^2(\mathbf{x})$ aus und lassen wir den von ω_j unabhängigen Term $-\frac{1}{2}|\Sigma|$ weg, so erhalten wir

$$g_j(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_j) + \ln P(\omega_j). \tag{286}$$

  $d_j^2(\mathbf{x})$ zerfällt in einen quadratischen und einen affinen Anteil

$$d_j^2(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\Sigma}^{-1}\mathbf{x} - 2\boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j, \tag{287}$$

  wobei der quadratische Anteil wiederum nicht von ω_j abhängt und somit weggelassen werden kann.

Die äquivalente lineare Diskriminantenfunktion ist - analog zum Fall $\mathbf{\Sigma}_j = \mathbf{I}\sigma$ - durch

$$g_j(\mathbf{x}) = \begin{array}{ccc} \boldsymbol{\mu}_j^T \mathbf{\Sigma}^{-1}\mathbf{x} & + & -\frac{1}{2}\boldsymbol{\mu}_j^T \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_j + \ln P(\omega_j) \\ \mathbf{w}_j^T \mathbf{x} & + & b_j, \end{array} \tag{288}$$

gegeben, die Entscheidungsgrenzen $g_j(\mathbf{x}) = g_k(\mathbf{x})$ durch

$$\mathbf{w}^T(\mathbf{x} - \mathbf{b}) \quad = \quad \mathbf{0}, \tag{289}$$

wobei

$$\mathbf{w} \quad = \quad \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k) \tag{290}$$

$$\mathbf{b} \quad = \quad \frac{1}{2}(\boldsymbol{\mu}_j + \boldsymbol{\mu}_k) - \frac{(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k)}{(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k)} \ln \frac{P(\omega_j)}{P(\omega_k)}. \tag{291}$$

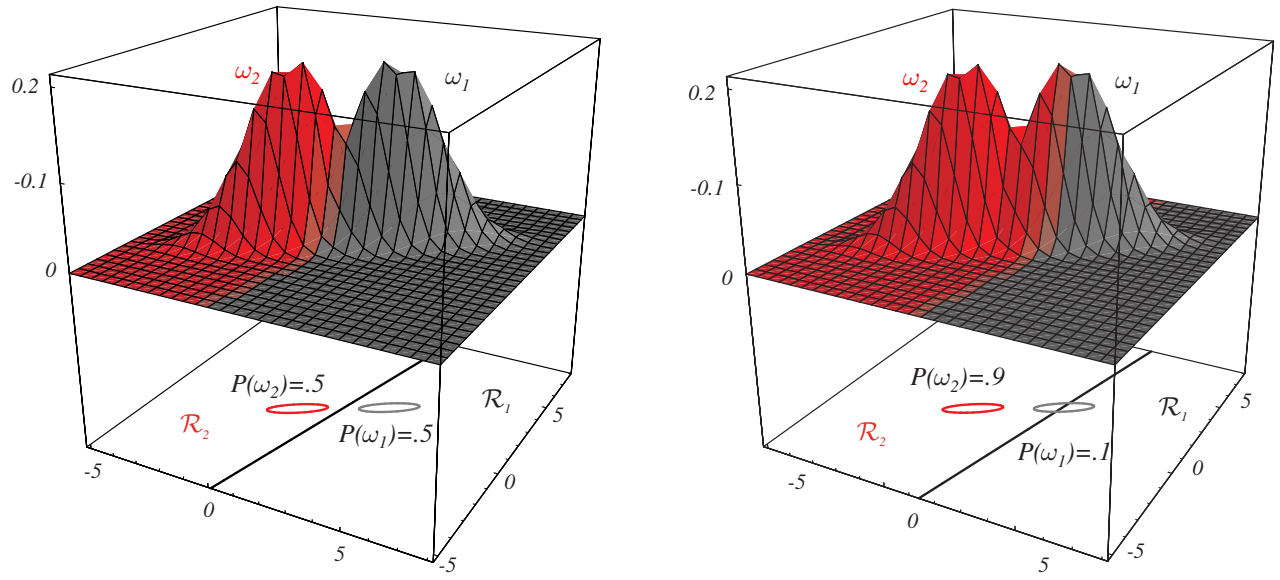


Abbildung 44: Entscheidungsgrenzen für zwei bivariate Normalverteilungen mit $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. Die Entscheidungsgrenzen sind wieder linear, jedoch i.a. nicht normal zur Verbindungsstrecke zwischen den beiden Klassenmitteln. Für gleiche priors verläuft die Entscheidungsgrenze durch $(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)/2$, ansonsten wird sie von der a priori wahrscheinlicheren Klasse wegverschoben.
(Aus *Duda, Hart, Stork: Pattern Classification, 2nd ed.*)

- **Quadratic Discriminant Analysis (QDA)** $\boldsymbol{\Sigma}_j$ beliebig

  Im allgemeinen Fall berechnen sich die Diskriminantenfunktionen gemäß Eq. 279

$$g_j(\mathbf{x}) = -\frac{1}{2}d_j^2(\mathbf{x}) + (-\frac{1}{2}\ln|\boldsymbol{\Sigma}_j| + \ln P(\omega_j)) \tag{292}$$

  Die Entscheidungsgrenzen sind durch (Hyper-) Quadriken gegeben, wobei die korrespondierenden Entscheidungsregionen nicht einfach zusammenhängend sein müssen.

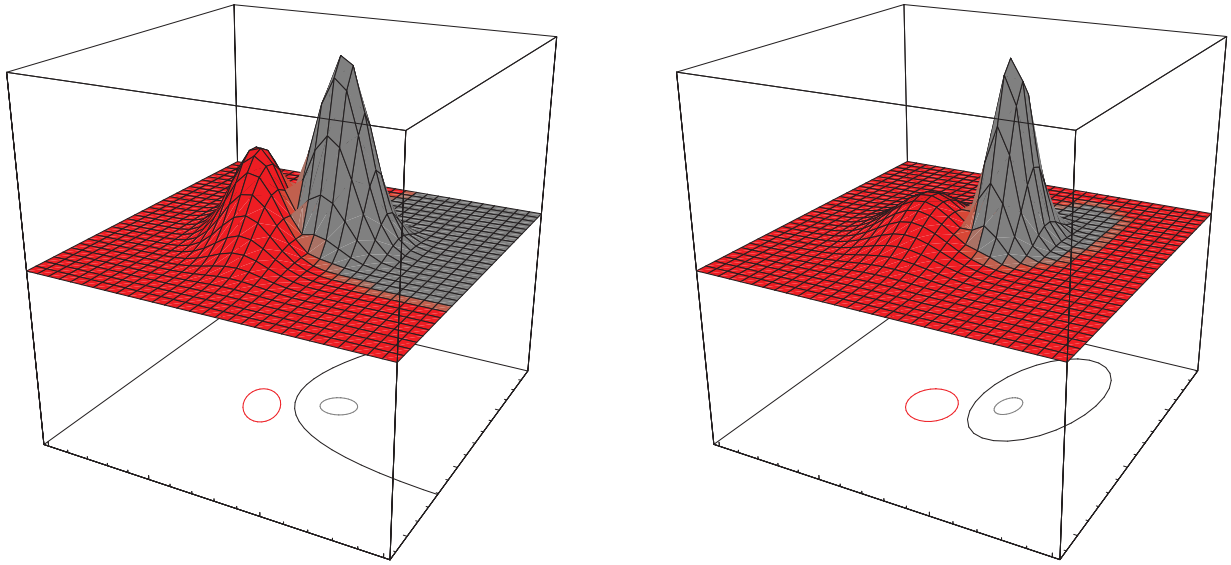


Abbildung 45: Entscheidungsgrenzen für zwei bivariate Normalverteilungen mit $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$. Die Entscheidungsgrenzen sind i.a. nicht linear, sondern durch Quadriken gegeben. Die Entscheidungsregionen müssen in diesem Fall nicht einfach zusammenhängend sein.
(Aus *Duda, Hart, Stork: Pattern Classification, 2nd ed.*)

Minimax Kriterium

Die optimale Bayes Entscheidungsgrenze hängt sowohl von den class conditional pdfs $p(\mathbf{x}|\omega_i)$ als auch von den priors $P(\omega_i)$ ab. Die für gegebene *priors* $P(\omega_i)$ gefundene Entscheidunsgrenze ist jedoch nicht (mehr) optimal, falls die beim Training verwendeten *priors* nicht korrekt waren bzw. diese sich nachträglich ändern. In diesem Fall wird die tatsächliche Fehlerrate über der Bayes-Fehlerrate liegen.

Wir betrachten im folgenden wieder den Fall $c = 2$. Für feste Entscheidungsgrenzen (-Regionen) ist die Fehlerrate $P(error)$ eine lineare Funktion in $P(\omega_1)$ und nimmt entweder für $P(\omega_1) = 0$ oder $P(\omega_1) = 1$ das Maximum an. Das Minimax-Kriterium wählt jene Entscheidungsgrenze, für welche dieses Maximum minimal wird und begrenzt somit den "Schaden" (die Fehlerrate) im ungünstigsten (*worst-case*) Fall.

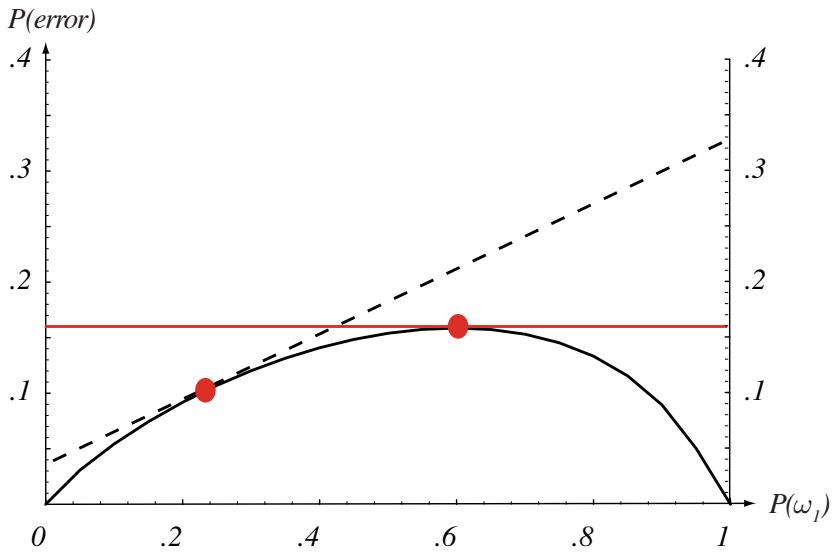


Abbildung 46: Die konvexe Kurve gibt den Verlauf des Bayes-Risk (bzw. der Fehlerrate) als Funktion der priors wieder. Ändern sich die priors nachträglich, so ändert sich das Risk ebenfalls, und zwar als lineare Funktion von $P(\omega_1)$. Für den Punkt links nimmt diese Funktion ihr Maximum (3.3) für $P(\omega_1) = 1$ an. Wird die Entscheidungsgrenze nach dem Minimax-Kriterium gewählt (rechter Punkt), so wird der Anstieg der Geraden 0, d.h. das Risk bleibt auch bei nachträglicher Änderung der priors konstant.
(Aus *Duda, Hart, Stork: Pattern Classification, 2nd ed.*)

Das Minimax-Risk R_{mm} (welches den Mininmax-Fehler als Spezialfall enthält) ist wie folgt definiert

$$\begin{aligned} R_{mm} &= \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2) d\mathbf{x} \\ &= \lambda_{11} + (\lambda_{21} - \lambda_{11}) \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1) d\mathbf{x}. \end{aligned} \tag{293}$$

Die Entscheidungsgrenze ist also dadurch definiert, dass die Beiträge der beiden Klassen zum Risk jeweils gleich groß sind. Man bemerkt, dass das Minimax-Risk nicht von den *priors* abhängt (die Steigung der Fehlergeraden ist 0).

Lineare Regression II: Funktionsapproximation mittels der Methode der kleinste Quadrate *(Linear Least Squares)*

- **Überblick**

  Die Regression (Funktionsapproximation) ist mit dem Problem befasst, den Wert einer

  – abhängigen Variablen (*output, response or target variable*) $y = f(\mathbf{x}) \in \mathbf{R}$ anhand einer
  – unabhängigen Variable (*input, predictor or explanatory variable*) $\mathbf{x} \in \mathbf{R}^p$

  vorherzusagen, wobei die zugrundeliegende Funktion f meist als stetig (*continuous*) oder sogar als einmal oder mehrfach stetig differenzierbar

(*smooth*) vorausgesetzt wird. Das "klassische" Regressions-Problem kann wie folgt formuliert werden:

Gegeben sei ein Familie parametrisierter Funktionen $f(\mathbf{x}, \mathbf{w})$ mit Parametervektor $\mathbf{w}$, z.B. die affinen (linearen) Funktionen

$$f(\mathbf{x}, \mathbf{w}) = w_2 x_2 + w_1 x_1 + w_0. \tag{294}$$

Da der Wert von y an der Stelle $\mathbf{x}$ von $\mathbf{w}$ abhängt, wird für $f(\mathbf{x}, \mathbf{w})$ oft auch $f(\mathbf{x}|\mathbf{w})$ geschrieben.

Der Zusammenhang zwischen $\mathbf{x}$ und y sei durch

$$y(\mathbf{x}) = f(\mathbf{w}^*, \mathbf{x}) + \epsilon \tag{295}$$

gegeben, wobei $\mathbf{w}^*$ den wahren Wert des Parametervektors und ϵ zufälliges Rauschen (noise) mit Mittel 0 bezeichne. Die Werte $y(\mathbf{x})$

setzen sich also aus einer deterministischen Komponente $f(\mathbf{x}, \mathbf{w}^*)$ und einer stochastischen (zufälligen) Komponente ϵ zusammen.

Anders formuliert, stellt $y(\mathbf{x})$ eine von $\mathbf{x}$ abhängige Zufallsvariable $Y|\mathbf{x}$ mit DF $p(y|\mathbf{x})$ dar. Eq. 295 wird somit zu

$$Y|\mathbf{x} = f(\mathbf{w}^*, \mathbf{x}) + \epsilon \tag{296}$$

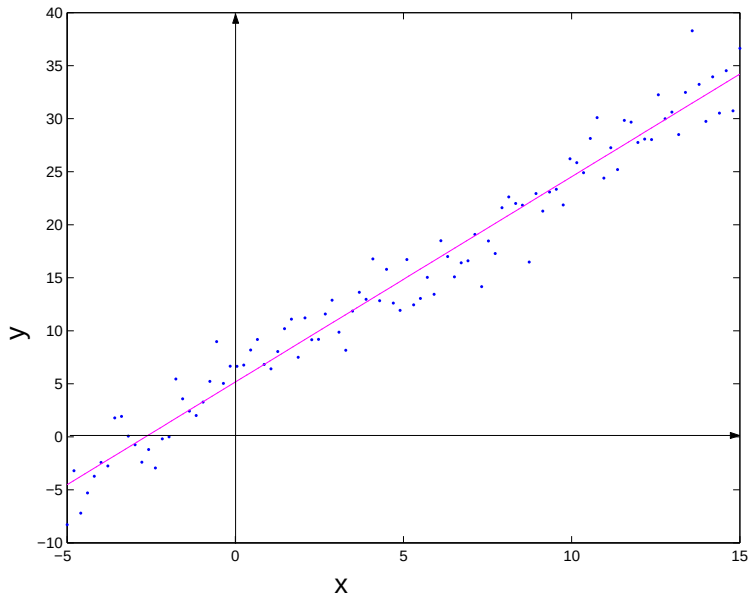


Abbildung 47: Beispiel eines linearen Modells mit additivem Gaußschem Rauschen. Für jeden Wert von x sind die Werte von y normalverteilt - $Y|x$ - mit Mittel (deterministischer Komponente) $\mathcal{E}[Y|x] = f(x, \mathbf{w}^*) = w_0 + w_1 * x$.

Man beachte, dass

$$\mathcal{E}[Y|\mathbf{x}] = \mathcal{E}[f(\mathbf{w}^*, \mathbf{x}) + \epsilon] = \mathcal{E}[f(\mathbf{w}^*, \mathbf{x})] + \mathcal{E}[\epsilon] = f(\mathbf{w}^*, \mathbf{x}), \quad (297)$$

d.h., das Mittel von Y an der Stelle $\mathbf{x}$ ist durch die deterministische Komponente $f(\mathbf{w}^*, \mathbf{x})$ gegeben.

Ziel ist es nun, einen Parametervektor $\mathbf{w}$ zu finden, welcher die mittlere "Diskrepanz" zwischen $Y|\mathbf{x}$ und der Vorhersage $f(\mathbf{x}, \mathbf{w})$ minimiert. Ein häufig verwendetes Maß für die Abweichung im Punkt $\mathbf{x}$ - bei gegebenem (gemessenem) y - ist der quadratische Fehler (*squared loss*, L_2-*loss*)

$$L(y, f(\mathbf{w}, \mathbf{x})) = (y - f(\mathbf{w}, \mathbf{x}))^2. \quad (298)$$

Da y allerdings eine - i.a. von $\mathbf{x}$ abhängige! - Zufallsvariable $Y|\mathbf{x}$ mit Dichtefunktion $p(y|\mathbf{x})$ darstellt, müssen wir den mittleren Fehler im Punkt $\mathbf{x}$ - das *conditional risk* - minimieren:

$$R(\mathbf{w}|\mathbf{x}) = \int (y - f(\mathbf{w}, \mathbf{x}))^2 p(y|\mathbf{x}) dy. \tag{299}$$

Um ein globales Fehlermaß zu erhalten, fassen wir auch $\mathbf{x}$ als Zufallsvariable auf und berechnen schließlich den Mittelwert von $R(\mathbf{w}|\mathbf{x})$ bzg. $\mathbf{x}$, das sogenannte *total risk*

$$R(\mathbf{w}) = \int \int (y - f(\mathbf{w}, \mathbf{x}))^2 p(y|\mathbf{x}) p(\mathbf{x}) dy d\mathbf{x}. \tag{300}$$

Unter den oben genannten Voraussetzungen lässt sich leicht zeigen, dass das *total risk* Eq. 300 durch Wahl von $\mathbf{w} = \mathbf{w}^*$ minimal wird, wobei

der Residualfehler durch die - von $\mathbf{w}$ unabhängige - Rausch-Varianz $Var(\epsilon) = \epsilon^2$ gegeben ist.

Die Bestimmung des optimalen Parametervektors bezeichnet man als Funktions-Approximation. Unter der Annahme eines linearen Modells für die deterministische Komponente von y, d.h. $f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T\mathbf{x}$ erhalten wir den wichtigen Spezialfall der **linearen Regression**.

- **Lineare Regression** (*ordinary linear least squares*, OLS)

  Sei $\mathcal{S}_{Tr} = \{\mathbf{X}, \mathbf{y}\}$ ein Trainingsset, wobei $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N) \in \mathbf{R}^{(d+1)\times N}$ die Spaltenmatrix homogenisierter Merkmalsvektoren und $\mathbf{y} = (y_1, \ldots, y_N) \in \mathbf{R}^N$ den Vektor korrespondierender (verrauschter!) Ausgabewerte bezeichne.

  Eine Schätzung des *total risk* Eq. 300 ist durch

$$\begin{aligned} mse(\mathbf{w}) &= \frac{1}{N}\sum_{i=1}^{N}(y_i - \mathbf{x}_i^T\mathbf{w})^2 = \frac{1}{N}\|\mathbf{y} - \mathbf{X}^T\mathbf{w}\|^2 \\ &= \frac{1}{N}(\mathbf{y}^T - \mathbf{w}^T\mathbf{X})(\mathbf{y} - \mathbf{X}^T\mathbf{w}) \end{aligned} \tag{301}$$

  gegeben. Man spricht in diesem Zusammenhang auch vom *empirical risk* bzw. im speziellen Fall einer quadratischen *loss*-Funktion (wie in Eq. 301) vom *mean squared error (mse)*.

Ist die gesuchte Funktion - wie im vorliegenden Fall der linearen Regression - linear in den Parametern $\mathbf{w}$, so hat die Kostenfunktion Eq. 301 (*mse*) folgende Eigenschaften. Sie

- ist glatt (hat eine stetige erste Ableitung)
- ist nicht-negativ und wird 0 g.d.w. $y_i = \mathbf{w}^T \mathbf{x}_i$ für alle $1 \leq i \leq N$, und
- ist eine quadratische ($\Rightarrow$ und somit konvexe!) Funktion der Parameter $\mathbf{w}$. Somit ist garantiert, dass es keine lokalen Minima gibt.
- Der Gradient (s.u.) von Eq. 301 bzg. $\mathbf{w}$ ist eine lineare Funktion des Parameter-Vektors $\mathbf{w}$.

- **Pseudoinverse**

  Unser Ziel ist es, das durch Gl. 301 gegebene *empirical risk* zu minimieren. Multiplizieren wir Gl. 301 aus und setzen wir den Gradienten gleich $\mathbf{0}$ (notwendige - und im Fall einer konvexen Funktion auch hinreichende - Bedingung für ein Minimum), so erhalten wir

$$\frac{1}{N}\nabla_{\mathbf{w}}(\mathbf{w}^T\mathbf{X}\mathbf{X}^T\mathbf{w} - 2\mathbf{y}^T\mathbf{X}^T\mathbf{w} + \mathbf{y}^T\mathbf{y}) = \mathbf{0} \quad (302)$$

$$\mathbf{X}\mathbf{X}^T\mathbf{w} = \mathbf{X}\mathbf{y}. \quad (303)$$

  Nachdem die Kostenfunktion Eq. 301 konvex ist, liefert uns jede Lösung $\mathbf{w}^*$ der sogenannten *normal equations* Eq. 303 ein globales Minimum von Eq. 301. Ist $\mathbf{X}\mathbf{X}^T$ invertierbar, so erhalten wir schließlich

$$\mathbf{w}^* = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{y}. \quad (304)$$

Eq. 304 gibt uns also die Lösung des *ordinary linear least squares (OLS)* Problems in geschlossener (nicht-iterativer) Form.

Der Ausdruck $(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}$ wird als **Pseudoinverse** oder auch als **Moore-Penrose-Inverse** von $\mathbf{X}^T$ bezeichnet.

- **Normalisierung**

  Wir haben bisher angenommen, dass die Trainingsvektoren (Spalten $\mathbf{x}_i$ von $\mathbf{X}$) in homogenen Koordinaten vorliegen. Dadurch wird der Verschiebungsanteil der affinen Transformation automatisch bestimmt, und die Regressionsebene geht immer durch den Mittelpunkt der Trainingsdaten $(\bar{\mathbf{x}}, \bar{y})$. Es ist jedoch aus numerischen Gründen sinnvoll, zumindest die Prädiktorgrößen $\mathbf{x}$ vor Berechnung der Regression komponentenweise zu standardisieren (d.h. auf Mittel 0 und Standardabweichung 1 zu bringen). Dieselbe Standardisierung muß dann natürlich ebenfalls auf die Test- bzw. Eingabedaten angewandt werden.

- **Lineare Regression als Parameterschätzung**

  Wir können die oben gefundene Lösung des quadratischen Minimierungsproblems auch als Schätzung $\hat{\mathbf{w}}$ des wahren Parametervektors $\mathbf{w}$ auffassen. Die korrespondierende Schätzfunktion (Statistik)

$$\vec{W} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\vec{Y} \tag{305}$$

  erhalten wir, wenn wir in Gleichung Eq. 304 den Vektor der beobachteten Größen $\mathbf{y}$ durch den Zufallsvektor

$$\vec{Y} = [Y_1, .., Y_N] = [Y|\mathbf{x}_1, ..., Y|\mathbf{x}_N] = \mathcal{E}[\vec{Y}] + \vec{\epsilon} \tag{306}$$

  ersetzen, welcher die Verteilung des Fehlers um die bedingten Erwartungswerte

$$\mathcal{E}[Y_i] = \mathbf{x}_i^T\mathbf{w} \tag{307}$$

beschreibt.

Bezeichne $\mathbf{\Sigma}_\epsilon$ die Kovarianz-Marix der Meßfehler, dann erhalten wir gemäß Lemma 2 die Kovarianzmatrix des Fehlers von $\vec{W}$ mit

$$\mathbf{\Sigma}_{\vec{W}} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{\Sigma}_\epsilon\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1} \tag{308}$$

Im Fall dass die Fehlerkomponenten ϵ_i unabhängig sind und der gleichen Verteilung folgen (iid), haben wir $\mathbf{\Sigma}_\epsilon = \mathbf{I}\sigma_\epsilon^2$, und Eq. 308 vereinfacht sich zu

$$\mathbf{\Sigma}_{\vec{W}} = (\mathbf{X}\mathbf{X}^T)^{-1}\sigma_\epsilon^2 \tag{309}$$

Unter den obigen Voraussetzungen ist der Schätzer Eq. 305 auch erwartungstreu, da

$$\begin{aligned} \mathcal{E}[\vec{W}] &= \mathcal{E}[(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\vec{Y}] = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathcal{E}[\vec{Y}] && (310)\\ &= (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{X}^T\mathbf{w} = \mathbf{w} && (311) \end{aligned}$$

- **Merkmalsselektion und Regularisierung**

  Man könnte erwarten, dass die Fähigkeit, das bedingte Mittel der Target-Variable als Funktion der Prädiktorvariablen (Merkmale) vorherzusagen, (z.B. gemessen durch den MSE auf einem Testset) um so besser wird, je mehr Merkmale man verwendet. Dies ist jedoch nicht der Fall; tatsächlich führt eine zu große Dimensionalität des Merkmalsraums zu einer Verschlechterung der Generalisierungsfähigkeit (*curse of dimensionality*). Andererseits wird eine zu kleine Anzahl an Merkmalen ebenfalls zu schlechten Ergebnissen führen. Der resultierende Zielkonflikt erinnert an das weiter oben am Beispiel der kNN-Klassifizierung besprochene *bias-variance* Dilemma.

  Methoden, welche versuchen, eine möglichst kleine Menge an Merkmalen auszuwählen, mit der man gerade noch in der Lage ist, den zugrundeliegenden funktionalen Zusammenhang herzustellen, fallen in die Kategorie **Merkmalsselektion**. Eine einfache Strategie besteht z.B darin, mit dem

konstanten Term zu beginnen und dann in jedem Schritt jenes Merkmal hinzuzufügen, welches den Testfehler am stärksten verringert *(stepwise forward selection)*. Das so erhaltene Ergebnis muß jedoch nicht optimal sein.

Ein anderer Ansatz besteht darin, alle Merkmale zu verwenden, und *overfitting* durch eine Modifikation der Zielfunktion zu verhindern, dergestalt, dass unerwünschte Lösungen zu einem großen Trainingsfehler führen; diesen Ansatz bezeichnet man als **Regularisierung**. Im Bereich der linearen Regression kann zu diesem Zweck die sogenannte *ridge penalty* $\lambda \mathbf{w}^T \mathbf{w}$ zum mse addiert werden.

$$mse_{reg}(\mathbf{w}, \lambda) = \frac{1}{N}(\mathbf{y}^T - \mathbf{w}^T \mathbf{X})(\mathbf{y} - \mathbf{X}^T \mathbf{w}) + \lambda \mathbf{w}^T \mathbf{w}, \lambda \geq 0 \quad (312)$$

Absolut große Werte der Regressionskoeffizienten w_i weisen auf *overfitting* und hohe Varianz des Schätzers hin. Indem man $mse_{reg}(\mathbf{w}, \lambda)$

anstelle des reinen least squares Kriteriums minimiert, werden Lösungen mit großen Koeffizienten bestraft. Der Regularisierungsparameter λ steuert, wie stark die Koeffizienten geschrumpft werden: für $\lambda = 0$ erhält man das ursprüngliche *least squares* Kriterium (kein Schrumpfen), für $\lambda = +\infty$ ist nur mehr der 0-Vektor als Lösung zulässig. Geeignete Werte für λ können z.B. mittels Kreuzvalidierung bestimmt werden.

Das Bestimmen der Lösung von Gl 312 bezeichnet man als **ridge regression**. Die Lösung ist, wie man sich leicht überzeugt, durch

$$\mathbf{w}^* = (\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbf{y}. \tag{313}$$

gegeben. Wenn mehrere Merkmalsdimensionen "fast" linear abhängig sind (und $\mathbf{X}\mathbf{X}^T$ somit eine große Konditionszahl aufweist, d.h. numerisch singulär ist), wird durch Addition von $\lambda\mathbf{I}$ die Konditionszahl verkleinert. Das Anbringen der *ridge penalty* verbessert also auch die numerischen Eigenschaften des Problems.

- **Hat-Matrix, Orthogonale Projektion und PCA-Regression**

  Multiplizieren wir die Transponierte der Designmatrix von rechts mit dem OLS-Lösungsvektor, so erhalten wir die lineare Prädiktion (Rekonstruktion) der Beobachtungen y_i

$$\hat{\mathbf{y}} = \mathbf{X}^T \mathbf{w}^* = \mathbf{X}^T (\mathbf{X}\mathbf{X}^T)^{-1} \mathbf{X}\mathbf{y}. \tag{314}$$

  Nachdem die obige Matrix jedem y_i den Schätzwert $\hat{y}_i$ zuordnet, bezeichnet man sie im Englischen auch als **hat matrix**. Tatsächlich ist $\hat{\mathbf{y}}$ nichts anderes als die Projektion von $\mathbf{y}$ auf den Zeilenraum von $\mathbf{X}$.

Eine besonders erhellende und nützliche Darstellung dieses Sachverhaltes erhält man, wenn man $\mathbf{X}^T$ durch seine Singulärwertzerlegung *(singular value decomposition, SVD)* ersetzt[11]

$$\mathbf{X}^T = \mathbf{U}\mathbf{D}\mathbf{V}^T, \tag{315}$$

wobei $\mathbf{U}$ eine Orthogonalbasis des Spaltenraums von $\mathbf{X}^T$ ist. Man erhält dann

$$\hat{\mathbf{y}} = \mathbf{U}\mathbf{U}^T\mathbf{y} \tag{316}$$

Der Vektor der Regressionskoeffizienten bzg. der Orthogonalbasis $\mathbf{U}$ ist einfach durch die Koordinaten von $\hat{\mathbf{y}}$ bzg. der Basis $\mathbf{U}$ gegeben: $\mathbf{w_u} = \mathbf{U}^T\mathbf{y}$. Die Spalten von $\mathbf{U}$ sind darüberhinaus die Eigenvektoren der Autokovarianzmatrix $\mathbf{X}^T\mathbf{X}$, mit zugeordeten Eigenwerten d_i^2 (den Diagonalelementen der Diagonalmatrix $\mathbf{D}$).

[11] Auf der Homepage zur LVA findet sich ein Link zu einem Dokument, welches die SVD und ihre Anwendung auf OLS-Probleme ausführlich diskutiert.

Ein Ansatz zur Merkmalsselektion bzw. Dimensionalitätsreduktion besteht darin, nur Eigenvektoren mit großen Eigenwerten in der obigen Rekonstruktion zu verwenden *(PCA regression)*.

Beispiel: PCA-Regression auf einer Gauß-RBF-Basis

Um die Anwendbarkeit des OLS-Modells auch auf nicht-lineare Probleme zu erweitern, können die ursprünglichen Merkmale durch nicht-lineare Funktionen ihrer selbst ersetzt oder ergänzt werden. Bekannte Vertreter solcher **Basis-Funktionen** sind Polynome, Splines und radiale Basis-Funktionen (RBF).

Wir betrachten im folgenden eine Menge von 64 Merkmalspunkten im Intervall $[-5, 5]$, wobei jedem Punkt x_i eine Gauß-DF mit seinem Wert als Mittel $\mu_i = x_i$ und Varianz $\sigma_i^2 = 1$ zugewiesen wird; dies ist in Abb. 48-LO für eine Teilmenge der Basisfunktionen dargestellt. Die i-te Zeile der Designmatrix $\mathbf{X}$ enhält somit die Funktionswerte der i-ten Basisfunktion für alle N Merkmalsausprägungen. Unter den obigen Annahmen ist $\mathbf{X}$ quadratisch und invertierbar. Abb. 48-RO zeigt die 1ste, 4te, 7te und 10te Eigenfunktion (Eigenvektor von $\mathbf{X}^T\mathbf{X}$) der RBF-Basis.

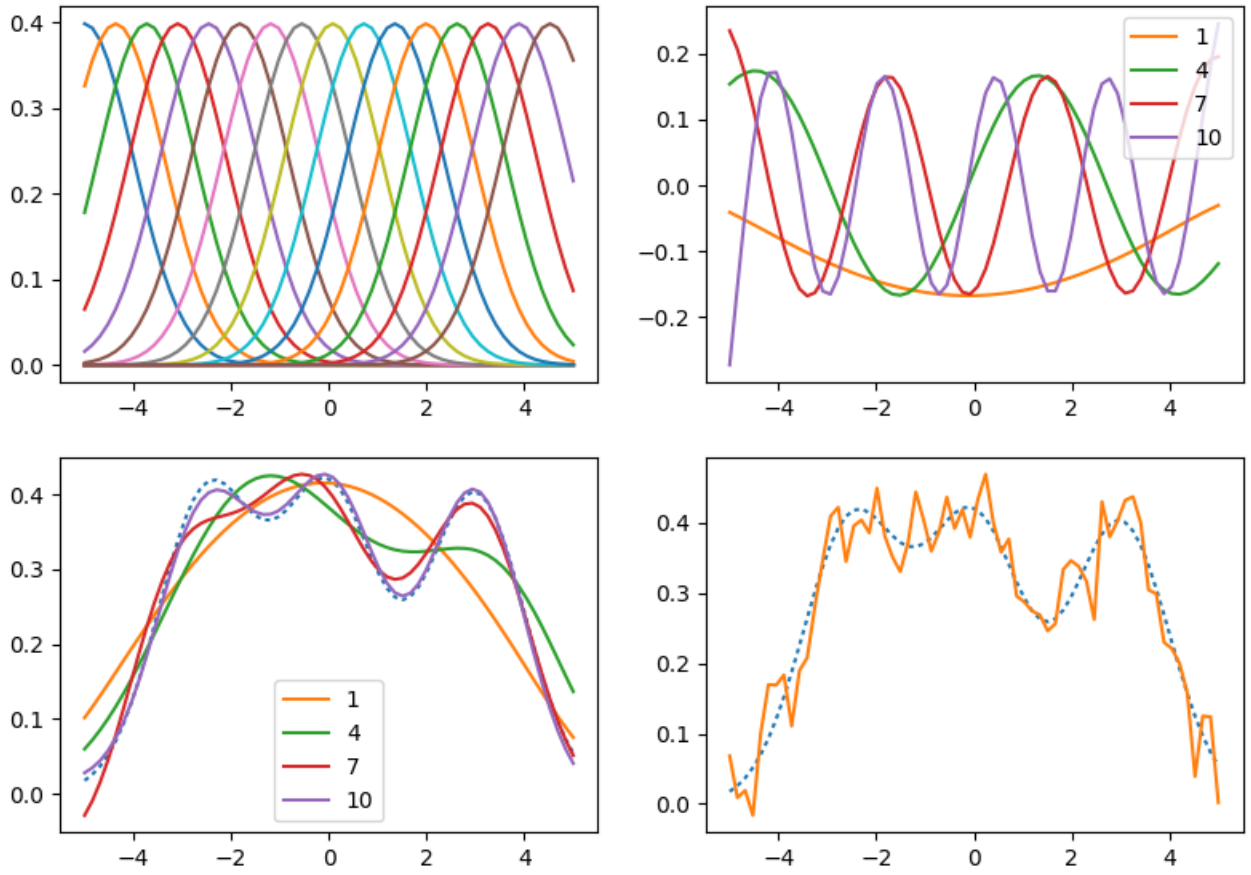


Abbildung 48: RBF-Basis (oben) und PCA-Regression eines verrauschten Signals (unten). Für Details siehe Text.

In der unteren Zeile ist die Regression eines verrauschten Signals auf die RBF-Basis dargestellt. Abb. 48-RU zeigt sowohl das ungestörte Signal $\mathbf{y}_t$ (blau gestrichtelt, eine Linearkombination dreier Gaußkurven) und die verrauschte Version $\mathbf{y}_n$ (orange). Lineare Regression mit allen Basisvektoren repliziert aufgrund der Invertierbarkeit von $\mathbf{X}$ das verrauschte Signal, d.h. $\mathbf{y}_n$ und $\hat{\mathbf{y}}$ sind identisch.

Abb. 48-LU zeigt die Ergebnisse der PCA-Regression unter Verwendung der 1, 4, 7 und 10 größten Eigenvektoren. Dies entspricht der partiellen Rekonstruktion des Signals bezüglich der RO gezeigten Eigenbasis. Das Ergebnis der PCA-Regression mit 10 Eigenvektoren ist dem unverrauschten Signal nicht nur visuell sehr ähnlich: tatsächlich wird der RMSE zwischen $\hat{\mathbf{y}}$ und $\mathbf{y}_t$ für 10 Eigenvektoren minimal, und bei Verwendung aller 64 Eigenvektoren mehr als doppelt so groß. Dieses Beispiel demonstriert also, wie Merkmalsselektion die Sensitivität der Regression bzg. Störungen der beobachteten Größe y verringern kann.

- ***ridge regression* als stetige Version der PCA-Regression**

Die weiter oben diskutierte *ridge regression* lässt sich als korrespondierende Regularisierungsstrategie auffassen, welche zwar alle Eigenvektoren verwendet, aber Eigenvektoren (Richtungen), deren Eigenwert (Varianz) im Verhältnis zum Regularisierungparameter λ_r klein ist, weniger stark gewichtet. Unter Berücksichtigung der *ridge penalty* erhält man die

Rekonstruktion

$$
\begin{aligned}
\hat{\mathbf{y}} &= \mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda_r\mathbf{I})^{-1}\mathbf{X}\mathbf{y} && (317)\\
&= \mathbf{U}\mathbf{D}\mathbf{V}^T(\mathbf{V}\mathbf{D}^2\mathbf{V}^T + \lambda_r\mathbf{I})^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{y} && (318)\\
&= \mathbf{U}\mathbf{D}\mathbf{V}^T\mathbf{V}(\mathbf{D}^2 + \lambda_r\mathbf{I})^{-1}\mathbf{V}^T\mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{y} && (319)\\
&= \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda_r\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^T\mathbf{y} && (320)\\
&= \sum_{i=1}^{p}\mathbf{u}_i\frac{d_i^2}{d_i^2 + \lambda_r}\mathbf{u}_i^T\mathbf{y}. && (321)
\end{aligned}
$$

Die Spur der *hat matrix* $dof(\lambda_r) = \sum_{i=1}^{p}\frac{d_i^2}{d_i^2+\lambda_r}$ lässt sich als effektive Anzahl der Freiheitsgrade *(effektive DOF)* des Schätzers interpretieren.

Lineare Regression III: Bayes-Ansatz

- In diesem Abschnitt erweitern wir die Bayes-Schätzung des Mittels einer Normalverteilung in 2 Richtungen: erstens betrachten wir nun multivariate Verteilungen, und zweitens suchen wir nicht das globale Mittel der Verteilung von Y, sondern einen Parametervektor $\mathbf{w}$, welcher es uns erlaubt, das bedingte Mittel von $Y|X = \mathbf{x}$ als lineare Funktion des Prädiktors (Merkmalsvektors) $\mathbf{x}$ zu berechnen:

$$\mathcal{E}[Y|\mathbf{x}] = \mathbf{w}^T\mathbf{x} \tag{322}$$

  Im Bayes-Ansatz wird der mittels Inferenz zu bestimmende Parametervektor $\mathbf{w}$ als Zufallsgröße behandelt.

Nehmen wir an, dass die N Beobachtungen Y_i multivariat normalverteilt sind, so lautet unser Modell für die Daten-DF[12]

$$\vec{Y}|\mathbf{w}, \mathbf{X} \sim N(\mathbf{X}^T\mathbf{w}, \mathbf{\Sigma}_D) \tag{323}$$

bzw.

$$-2\log p(\mathbf{y}|\mathbf{w}) = \left(\mathbf{y} - \mathbf{X}^T\mathbf{w}\right)^T \ \mathbf{\Sigma}_D^{-1}\left(\mathbf{y} - \mathbf{X}^T\mathbf{w}\right) + c_1 \tag{324}$$

Wir haben in Gl. 323 die Abhängigkeit der Beobachtungen $\vec{Y}$ sowohl von der Matrix der Merkmals-Vektoren $\mathbf{X}$ als auch vom Parametervektor $\mathbf{w}$ explizit gemacht. Da $\mathbf{X}$ jedoch bekannt und keine mittels Inferenz zu ermittelnde Größe ist, werden wir es im folgenden als Hintergrundwissen behandeln und nicht mehr explizit als bedingende Größe anführen.

[12] Im Ausdruck $\vec{Y}|\mathbf{w}, \mathbf{X}$ bindet , stärker als $|$.

Wir nehmen als *a priori* Verteilung des Gewichtsvektors ebenfalls eine Normalverteilung mit Mittel $\mathbf{w}_0$ und Kovarianzmatrix $\mathbf{\Sigma}_0$ an:

$$\vec{W} \sim N(\mathbf{w}_0, \mathbf{\Sigma}_0) \tag{325}$$

bzw.

$$-2 \log p(\mathbf{w}) = (\mathbf{w} - \mathbf{w}_0)^T \ \mathbf{\Sigma}_0^{-1} (\mathbf{w} - \mathbf{w}_0) + c_2 \tag{326}$$

Im folgenden bezeichnen Größen mit Subskript wie z.B. $\mathbf{w}_0$ wiederum Hyper-Parameter, also konstante Werte.

Wie im univariaten Fall sehen wir, daß die *a posteriori* DF

$$p(\mathbf{w}|\mathbf{y}) = p(\mathbf{w})p(\mathbf{y}|\mathbf{w})/p(\mathbf{y}) \tag{327}$$

proportional dem Produkt der *a priori* DF und der Daten-DF und $-2\log p(\mathbf{w}|\mathbf{y})$ somit eine quadratische Funktion in $\mathbf{x}$ ist. Damit ist aber

$$\vec{W}|\mathbf{y} \sim N(\mathbf{w}_1, \mathbf{\Sigma}_1) \tag{328}$$

multivariat normalverteilt. Gleichsetzen der Exponenten

$$\begin{aligned} (\mathbf{w}-\mathbf{w}_1)^T\ \mathbf{\Sigma}_1^{-1}(\mathbf{w}-\mathbf{w}_1) &= (\mathbf{w}-\mathbf{w}_0)^T\ \mathbf{\Sigma}_0^{-1}(\mathbf{w}-\mathbf{w}_0) \\ &+ \left(\mathbf{y}-\mathbf{X}^T\mathbf{w}\right)^T\ \mathbf{\Sigma}_D^{-1}\left(\mathbf{y}-\mathbf{X}^T\mathbf{w}\right) \end{aligned} \tag{329}$$

und Koeffizientenvergleich der quadratischen

$$\mathbf{w}^T \boldsymbol{\Sigma}_1^{-1} \mathbf{w}^T = \mathbf{w}^T \left(\boldsymbol{\Sigma}_0^{-1} + \mathbf{X} \boldsymbol{\Sigma}_D^{-1} \mathbf{X}^T \right) \mathbf{w}^T \tag{330}$$

und linearen Terme in $\mathbf{w}$

$$-2\mathbf{w}^T \boldsymbol{\Sigma}_1^{-1} \mathbf{w}_1 = -2\mathbf{w}^T \left(\boldsymbol{\Sigma}_0^{-1} \mathbf{w}_0 + \mathbf{X} \boldsymbol{\Sigma}_D^{-1} \mathbf{y} \right) \tag{331}$$

liefert schließlich

$$\boldsymbol{\Sigma}_1 = \left(\boldsymbol{\Sigma}_0^{-1} + \mathbf{X} \boldsymbol{\Sigma}_D^{-1} \mathbf{X}^T \right)^{-1} \tag{332}$$

$$\mathbf{w}_1 = \left(\boldsymbol{\Sigma}_0^{-1} + \mathbf{X} \boldsymbol{\Sigma}_D^{-1} \mathbf{X}^T \right)^{-1} \left(\boldsymbol{\Sigma}_0^{-1} \mathbf{w}_0 + \mathbf{X} \boldsymbol{\Sigma}_D^{-1} \mathbf{y} \right). \tag{333}$$

Anhang A: Multivariate stetige Verteilungen

- **Joint pdf und Joint cdf**
  Die multivariate Verteilungsfunktion (*joint cdf*) ist wie folgt definiert:

$$F(\mathbf{x}) = P(\vec{X} \leq \mathbf{x}) = P(X_1 \leq x_1, \ldots, X_p \leq x_p). \tag{334}$$

  $F(\mathbf{x})$ ergibt sich, analog zum skalaren Fall, als p-faches Integral über eine nicht-negative multivariate Dichtefunktion (*joint pdf*)

$$F(\mathbf{x}) = \int_{-\infty}^{\mathbf{x}} p(\mathbf{x}')\, d\mathbf{x}' = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_p} p(x_1', \ldots, x_p')\, dx_1' \ldots dx_p'. \tag{335}$$

- Eigenschaften der joint pdf und joint cdf
  - $F(\mathbf{x})$ ist monoton wachsend in allen Koordinaten
  - $\lim_{x_i \to -\infty} F(\mathbf{x}) = 0$, d.h. $F(\mathbf{x})$ wird 0 wenn nur eines der x_i gegen $-\infty$ geht
  - $\lim_{x_1,\ldots,x_p \to +\infty} F(\mathbf{x}) = 1$, d.h. $F(\mathbf{x})$ wird 1 wenn alle x_i gegen $+\infty$ gehen
  - $p(\mathbf{x}) \geq 0 \quad \forall \mathbf{x} \in \mathbf{R}^p$
  - $p(\mathbf{x}) = \partial^p F(\mathbf{x}) / \partial x_1 \ldots \partial x_p$

- **Randverteilung** (*marginal distribution*)
  Seien X, Y zwei stetige Zufallsvariablen mit pdf $p(x, y)$ und cdf $F(x, y)$.

  Die *Randverteilung der Dichtefunktion* (**marginal pdf**) bzg. X ergibt sich durch Integration über alle möglichen Ausprägungen von Y

$$p_X(x) = \int_{-\infty}^{+\infty} p(x, y')dy' \tag{336}$$

  Die *Randverteilung der Verteilungsfunktion* (**marginal cdf**) bzg. X erhält man als Integral über die *marginal pdf*

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} \int_{-\infty}^{+\infty} p(x', y')dy'dx' \\ &= \int_{-\infty}^{x} p_X(x')dx' = F(x, +\infty). \end{aligned} \tag{337}$$

Die *marginal pdf* $p_Y(y)$ und *marginal cdf* $F_Y(y)$ bzg. Y berechnen sich analog.

In der Praxis wird oft kurz $p(x)$ für $p_X(x)$ bzw. $F(x)$ für $F_X(x)$ geschrieben (analog für Y).

- **Beispiel: Rechtecksverteilung**
  Gleichverteilung im Bereich $B = B_1 \times B_2 = [a_1, b_1] \times [a_2, b_2]$. Die *joint pdf* ist innerhalb von B konstant:

$$p(x, y) = \frac{1}{(b_1 - a_1)(b_2 - a_2)} \tag{338}$$

  für $(x, y) \in B$, 0 sonst.

  Die *joint cdf* berechnet sich wie folgt:

  $F(x, y) =$

  - 0, falls $x < a_1$ oder $y < a_2$
  - $(x - a_1)/(b1 - a1)$, falls $x \in B_1, y > b_2$ (Randverteilung von x)
  - $(y - a_2)/(b_2 - a_2)$, falls $y \in B_2, x > b_1$ (Randverteilung von y)
  - $(x - a_1)(y - a_2)/\left((b_1 - a_1)(b_2 - a_2)\right)$, falls $(x, y) \in B$
  - 1, falls $x > b_1$ und $y > b_2$.

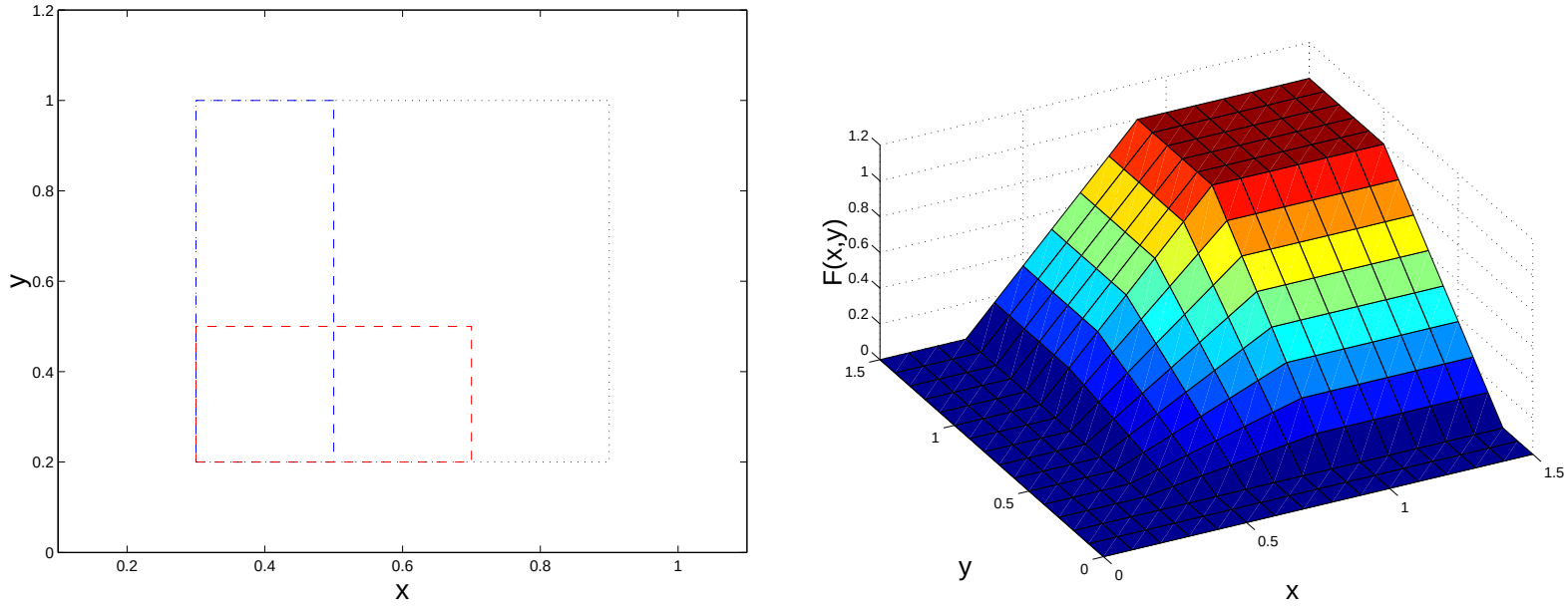


Abbildung 49: Rechtecksverteilung im Bereich $[0.3, 0.9] \times [0.2, 1]$.
Links: Die Dichtefunktion (*joint pdf*) $p(x, y)$ ist innerhalb der schwarz gepunkteten Umrandung konstant und positiv mit $1/(0.6 * 0.8)$. Die Werte der Verteilungsfunktion (*joint cdf*) $F(0.7, 0.5)$ und $F(0.5, 1) = F(0.5, +\infty)$ ergeben sich als Gebietsintegrale $(x-0.3)(y-0.2)/(0.6*0.8)$ über die jeweils gestrichelt umrandeten Bereiche.
Rechts: Verteilungsfunktion $F(x, y)$.

- **Unabhängigkeit**
  X und Y sind unabhängig (*independent)*, wenn

$$F(x, y) = F_X(x)F_Y(y) = F(x)F(y), \tag{339}$$

  d.h., wenn die *joint cdf* gleich dem Produkt der *marginal cdfs* ist ($F(x, y)$ faktorisiert in $F_X(x)$ und $F_Y(y)$).

  Im Falle der Unabhängigkeit gilt ebenfalls

$$p(x, y) = p_X(x)p_Y(y) = p(x)p(y). \tag{340}$$

- **Bedingte Verteilung**
  Die bedingte Verteilung der Dichtefunktion (*conditional pdf*) von X unter $Y = y$ erhält man als

$$p(x|y) = \frac{p(x,y)}{p_Y(y)}, \tag{341}$$

  die korrespondierende *conditional cdf* als

$$F(x|y) = \int_{-\infty}^{x} p(x'|y)dx'. \tag{342}$$

  Ebenso wie im diskreten Fall gilt für unabhängige Zufallsvariablen X, Y, dass

$$p(x|y) = \frac{p_X(x)p_Y(y)}{p_Y(y)} = p_X(x) = p(x). \tag{343}$$

Anhang B: Receiver Operating Characteristics - ROC

ROC sind ein aus der Signalverarbeitung kommender Ansatz zur Beschreibung bzw. Behandlung von Testproblemen. Sie haben ihren Ursprung in der Radartechnologie, wo sie ursprünglich für den Zweck konzipiert wurden, ein Signal – ein von einem Objekt reflektiertes Radarecho – vom Hintergrundrauschen zu unterscheiden.

Wir treffen im folgenden die Annahme, dass sowohl das Signal als auch das Rauschen normalverteilt mit gleicher Varianz sind. Bezeichne im folgenden ω_1 das Rauschen und ω_2 das wahre Signal, und seien die Verteilungen durch $N(\mu_i, \sigma)$ gegeben, wobei wir weiters $\mu_2 > \mu_1$ annehmen.

Rauschen und Signal werden umso leichter zu unterscheiden sein, je größer die Differenz ihrer Mittelwerte relativ zur Standardabweichung ist; die (von der Entscheidungsgrenze x^* unabhängige) Kenngröße

$$d' = \frac{|\mu_1 - \mu_2|}{\sigma} \tag{344}$$

wird auch **discriminability** genannt.

Bei der Klassierung des Signals können vier verschiedene Ereignisse eintreten

- $X > x^*|\omega_2$: **hit (tp)**,
- $X < x^*|\omega_2$: **miss (fn)**,
- $X < x^*|\omega_1$: **true negative (tn)**
- $X > x^*|\omega_1$: **false alarm (fp)**

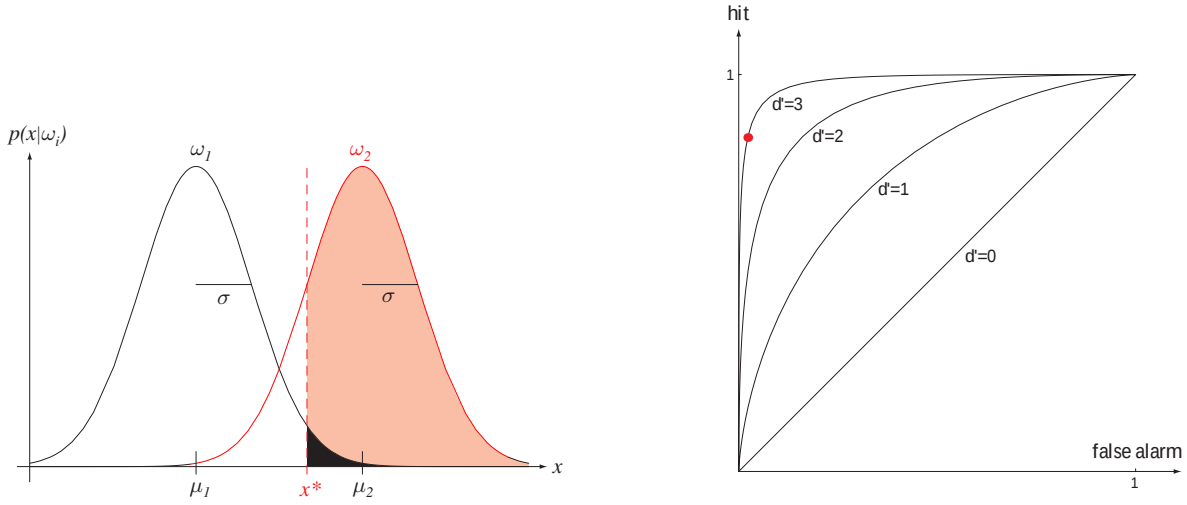


Abbildung 50:
Links: Verteilung des Rauschens und des Nutz-Signals. Dargestellt sind außerdem die Wahrscheinlichkeiten $P(hit)$ (rosa) sowie $P(false\ alarm)$ (schwarz).
Rechts: ROC-curves. Je größer d', desto schneller konvergiert die Kurve (als Funktion von $P(false\ alarm)$ betrachtet) gegen 1.
(Aus *Duda, Hart, Stork: Pattern Classification, 2nd ed.*)

Von Bedeutung ist hier insbesondere das Verhältnis von $P(hit)$ zu $P(false\ alarm)$. Wünschenswert ist natürlich eine große *hit-rate* bei gleichzeitig möglichst geringer Wahrscheinlichkeit für einen *false alarm*. Dieser Zusammenhang wird i.a. durch sogenannte *ROC-curves* dargestellt. Jede *ROC-curve* ist durch die *discriminability* des Systems eindeutig festgelegt (je größer, desto schneller steigt die Kurve anfangs an). Jeder Punkt auf einer solchen Kurve entspricht einer Entscheidungsgrenze x^*.

Achtung: im allgemeinen Fall (keine Normalverteilungen oder ungleiche Varianz) sind die *ROC-curves* nicht symmetrisch.

Appendix C: Lineare Algebra

- **Rechenregeln für Determinanten**
  - $|\mathbf{A}| = 0$ g.d.w. $\mathbf{A}$ singulär
  - $|\mathbf{A}| = \prod_i a_{ii}$ falls $\mathbf{A} = (a_{ij})$ eine Diagonalmatrix ist (speziell gilt $|\mathbf{I}| = 1$)
  - $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$
  - $|\mathbf{A}^{-1}| = |\mathbf{A}|^{-1}$
  - $|\mathbf{A}| > 0 (\geq 0)$, für positiv definites (positiv semi-definites) $\mathbf{A}$.

- **Rechenregeln für Matrix-Produkte**

  Für zwei Matrizen $\mathbf{A} \in \mathrm{R}^{p \times q}, \mathbf{B} \in \mathrm{R}^{q \times r}$ gilt, daß

$$(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T. \tag{345}$$

Für zwei Matrizen $\mathbf{A}, \mathbf{B} \in \mathbf{R}^{p \times p}$ gilt, daß

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}. \tag{346}$$

Appendix D: Gradienten

Der **Gradient** einer Funktion $f : \mathbf{R}^p \to \mathbf{R}$

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = \nabla f = (\frac{\partial f}{\partial w_1}, \ldots, \frac{\partial f}{\partial w_p})^T = (\frac{df}{d\mathbf{w}})^T \tag{347}$$

(sprich: nabla f) bzg. $\mathbf{w}$ ist definiert als Transponierte der ersten Ableitung nach $\mathbf{w}$; er zeigt (als Vektor) in die Richtung des steilsten Anstiegs (bei linearer Fortsetzung) von f. Folglich zeigt $-\nabla f$ in die Richtung des steilsten Abfalls von f; $-\nabla f$ wird auch als Richtung des **steepest descent** bezeichnet. Das "Verschwinden" des Gradienten $\nabla_{\mathbf{w}} f(\mathbf{w})|_{\mathbf{w}=\mathbf{w}^*} = \mathbf{0}$ an der Stelle $\mathbf{w} = \mathbf{w}^*$ ist eine notwendige Voraussetzung dafür, dass f an der Stelle $\mathbf{w}^*$ ein Extremum annimmt.

Im allgemeinen Fall einer vektorwertigen Funktion $\mathbf{f} : \mathbf{R}^p \rightarrow \mathbf{R}^q$ erhält man den Gradienten als Transponierte der Jacobi-Matrix $(\partial f_i / \partial w_j)_{1 \leq i \leq q, 1 \leq j \leq p}$.

Beispiel

Sei $\mathbf{w} \in \mathbf{R}^2$ und $f_1(\mathbf{w}) = \sin(w_1)\cos(w_2)$ sowie $f_2(\mathbf{w}) = 3w_1^2 w_2 + 2w_1$. Bezeichne weiters $f_{ij} = \frac{\partial f_i}{\partial w_j}$ die partielle Ableitung von f_i nach w_j. Es gilt

$$\begin{aligned} \nabla_{\mathbf{w}} f_1(\mathbf{w}) &= \begin{pmatrix} f_{11} \\ f_{12} \end{pmatrix} = \begin{pmatrix} \cos(w_1)\cos(w_2) \\ -\sin(w_1)\sin(w_2) \end{pmatrix} \\ \nabla_{\mathbf{w}} f_2(\mathbf{w}) &= \begin{pmatrix} f_{21} \\ f_{22} \end{pmatrix} = \begin{pmatrix} 6w_1 w_2 + 2 \\ 3w_1^2 \end{pmatrix}. \end{aligned}$$

Fassen nun wir f_1, f_2 als Komponenten der vektorwertigen Funktion $\mathbf{f}(\mathbf{w}) = \begin{pmatrix} f_1(\mathbf{w}) \\ f_2(\mathbf{w}) \end{pmatrix} : \mathbf{R}^2 \to \mathbf{R}^2$ auf, so ist der Gradient von $\mathbf{f}$ durch

$$\nabla_{\mathbf{w}}\mathbf{f} = (f_{ij})^T = (\nabla f_1 \nabla f_2) = \begin{pmatrix} f_{11} & f_{21} \\ f_{12} & f_{22} \end{pmatrix}$$

gegeben.

Der Gradient einer affinen Funktion ist durch

$$\nabla_{\mathbf{w}}(\mathbf{A}\mathbf{w} + \mathbf{b}) = \mathbf{A}^T, \quad \mathbf{w} \in \mathrm{R}^q, \mathbf{b} \in \mathrm{R}^p, \mathbf{A} \in \mathrm{R}^{p \times q} \tag{348}$$

gegeben.

Der Gradient einer **symmetrischen** quadratischen Form mit Koeffizienten-Matrix $\mathbf{A} = \mathbf{A}^T$ ist durch

$$\nabla_{\mathbf{w}}(\mathbf{w}^T\mathbf{A}\mathbf{w}) = 2\mathbf{A}\mathbf{w}, \quad \mathbf{w} \in \mathrm{R}^p, \mathbf{A} \in \mathrm{R}^{p \times p} \tag{349}$$

gegeben. Man beachte, dass Matrizen der Gestalt $\mathbf{C} = \mathbf{A}\mathbf{A}^T$ immer symmetrisch sind, d.h., $\mathbf{C} = \mathbf{C}^T$.

Appendix E: Das Perceptron

- Das Perceptron stellt einen Spezialfall eines binären, linearen Klassifikators dar. Lineare Modelle sind schnell und einfach zu trainieren und auszuwerten.

  Wir gehen im folgenden von d-dimensionalen Merkmalsvektoren $\mathbf{x} \in \mathbf{R}^d$ und zwei Klassen ω_1, ω_2 aus. Ziel ist es, eine Abbildung $g : \mathbf{R}^d \rightarrow \mathbf{R}$ zu finden, welche die Klassenzugehörigkeit wie folgt kodiert

$$g(\mathbf{x}) > 0 \quad \text{falls } \mathbf{x} \in \omega_1 \tag{350}$$

$$g(\mathbf{x}) < 0 \quad \text{falls } \mathbf{x} \in \omega_2 \tag{351}$$

  wobei der Absolutbetrag von g das "Vertrauen" in die vorhergesagte

Klassenzugehörigkeit von $\mathbf{x}$ widerspiegelt. g wird auch als **Diskriminantenfunktion** (*discriminant function*) bezeichnet.

Im speziellen Fall einer linearen Diskriminantenfunktion hat g die folgende Form

$$g(\mathbf{x}) = \sum_{i=1}^{d} w_i x_i - \theta = \mathbf{w}^T \mathbf{x} - \theta, \tag{352}$$

wobei

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_d \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ \dots \\ w_d \end{pmatrix}. \tag{353}$$

$\mathbf{w} \in \mathbf{R}^d$ wird oft als Gewichtsvektor und $\theta \in \mathbf{R}$ als *bias* oder *threshold* bezeichnet. Die Aufgabe besteht nun darin, geeignete Werte für $\mathbf{w}$ und θ zu finden.

Das Perceptron wurde gegen Ende der 1950er von Rosenblatt als Modell eines künstlichen neuralen Netzwerks entwickelt. Die Architektur des Perceptrons entspricht einer linearen Diskriminantenfunktion mit nachgeschalteter Signum-Funktion. Wenn wir mit $o(\mathbf{x})$ die Ausgabe des Perceptrons bezeichnen, so haben wir

$$o = o(\mathbf{x}) = \text{sgn}(\mathbf{w}^T\mathbf{x} - \theta) = \begin{cases} 1 & \text{falls } \mathbf{w}^T\mathbf{x} \geq \theta \\ -1 & \text{falls } \mathbf{w}^T\mathbf{x} < \theta \end{cases} \tag{354}$$

Das Ziel ist nun, den Gewichtsvektor $\mathbf{w}$ und bias θ zu bestimmen, sodass:

$$o(\mathbf{x}) = l_1 = 1 \quad (\Leftrightarrow \mathbf{w}^T\mathbf{x} \geq \theta) \text{ falls } \mathbf{x} \in \omega_1 \tag{355}$$

$$o(\mathbf{x}) = l_2 = -1 \quad (\Leftrightarrow \mathbf{w}^T\mathbf{x} < \theta) \text{ falls } \mathbf{x} \in \omega_2, \tag{356}$$

d.h. $\mathcal{C} = \{1, -1\}$.

- **Geometrische Interpretation**

  Für $\mathbf{w}, \mathbf{x} \in \mathbf{R}^d$, legt

$$\sum_{i=1}^{d} w_i x_i = \mathbf{w}^T \mathbf{x} = \theta, \tag{357}$$

  eine in den $\mathbf{R}^d$ eingebettete $(d-1)$-dimensionale Hyperebene (*hyperplane*) (im Fall $d = 2$ eine Gerade) mit Normalvektor $\mathbf{w}$ fest. Im Fall $\theta = 0$ geht die Hyperebene durch den Ursprung, andrenfalls ist sie entlang $\mathbf{w}$ um den Betrag $\theta / \|\mathbf{w}\|$ vom Ursprung verschoben.

  Das innere Produkt $\mathbf{w}^T \mathbf{x}$ kann alternativ als

$$\mathbf{w}^T \mathbf{x} = \cos(\mathbf{w}, \mathbf{x}) \|\mathbf{x}\| \|\mathbf{w}\| \tag{358}$$

  geschrieben werden, und entspricht daher der Projektion von $\mathbf{x}$ auf $\mathbf{w}$ ($\cos(\mathbf{w}, \mathbf{x}) \|\mathbf{x}\|$) mal der Norm von $\mathbf{w}$, $\|\mathbf{w}\|$.

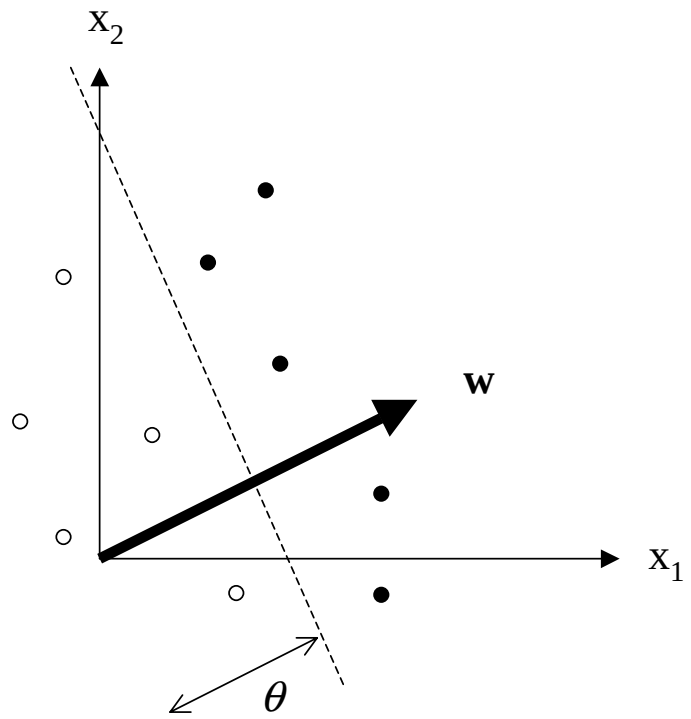


Abbildung 51: Die gestrichelte Gerade $\mathbf{w}^T\mathbf{x} = \theta$ ist durch ihren Normalvektor $\mathbf{w}$ und ihre Distanz vom Ursprung $\theta/\|\mathbf{w}\|$, gemessen entlang $\mathbf{w}$, festgelegt (hier für den Fall $\|\mathbf{w}\| = 1$). Für schwarze Punkte ($\in \omega_1$) gilt, $\mathbf{w}^T\mathbf{x} > \theta$, wohingegen für die weißen Punkte ($\in \omega_2$) $\mathbf{w}^T\mathbf{x} < \theta$ gilt.

- **Kanonische Repräsentation** (*Canonical Representation*)

  Wenn wir $\mathbf{w}$ und θ mit demselben positiven Faktor $\alpha \in \mathbf{R}^+$ multiplizieren, bleiben die Entscheidungsregionen unverändert:

$$\mathbf{w}^T\mathbf{x} = \theta \Leftrightarrow (\alpha\mathbf{w})^T\mathbf{x} = \alpha\theta \quad (\forall \mathbf{x} \in \mathbf{R}^d) \tag{359}$$

$$\mathbf{w}^T\mathbf{x} \geq \theta \Leftrightarrow (\alpha\mathbf{w})^T\mathbf{x} \geq \alpha\theta \quad (\forall \mathbf{x} \in \mathbf{R}^d) \tag{360}$$

  Setzen wir speziell $\alpha = 1/\|\mathbf{w}\|$, so erhalten wir die sogenannte **kanonische Repräsentation** der Hyperebene $\mathbf{w}_c = \frac{\mathbf{w}}{\|\mathbf{w}\|}, \theta_c = \frac{\theta}{\|\mathbf{w}\|}$ mit auf Einheitlänge normiertem Normalvektor $\|\mathbf{w}_c\| = 1$. In diesem Fall

  – entspricht das innere Produkt $\mathbf{w}_c^T\mathbf{x}$ der Projektion von $\mathbf{x}$ auf $\mathbf{w}_c$ (siehe Eq. 358), und
  – gibt der Wert der Diskriminantenfunktion $g(\mathbf{x}) = \mathbf{w}_c^T\mathbf{x} - \theta_c$ den Abstand von $\mathbf{x}$ zur Entscheidungsebene an (parallel zu $\mathbf{w}_c$).

- Die Hyperebene $\mathbf{w}^T\mathbf{x} = \theta$ partitioniert $\mathbf{R}^d$ in zwei Halbräume: $\mathcal{R}_1 = \{\mathbf{x} : \mathbf{w}^T\mathbf{x} \geq \theta\}$ und $\mathcal{R}_2 = \{\mathbf{x} : \mathbf{w}^T\mathbf{x} < \theta\}$.

  Da wir eine Beobachtung $\mathbf{x}$ an ω_1 zuweisen falls $\mathbf{x} \in \mathcal{R}_1$ und an ω_2 falls $\mathbf{x} \in \mathcal{R}_2$, werden die $\mathcal{R}_i$ auch **Entscheidungsregionen** *decision regions* genannt; die separierende Hyperebene $\mathbf{w}^T\mathbf{x} = \theta$ wird auch **Entscheidungsgrenze** (*decision boundary*) genannt.

- **Lineare Separierbarkeit** (*linear separability*)

  Sei $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N) \in \mathbf{R}^{d \times N}$ eine Menge von N Merkmalsvektoren mit zugeordneten **Klassen-Labels** $\mathbf{y}^T = (y_1, \ldots, y_N), y_i \in \{1, -1\}$. Wir sagen dass $\mathbf{X}$ linear separierbar (bzg. $\mathbf{y}$) ist, falls es einen Gewichtsvektor $\mathbf{w}$ und bias θ gibt, sodass

$$o(\mathbf{x}_i) = \text{sgn}(\mathbf{w}^T\mathbf{x}_i - \theta) = y_i, \;\; 1 \leq i \leq N. \tag{361}$$

- **Homogene Koordinaten** (*Homogeneous Coordinates*)

  Der bias kann durch einen kleinen Kunstgriff in den Gewichtsvektor "hineingezogen" werden. Wir führen zu diesem Zweck zusätzliche Koordinaten $x_0 \equiv 1$ and $w_0 = -\theta$ ein.

$$ {}^a\mathbf{x} = (1, \mathbf{x}^T)^T = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \dots \\ x_d \end{pmatrix}, \qquad {}^a\mathbf{w} = (-\theta, \mathbf{w}^T)^T = \begin{pmatrix} -\theta \\ w_1 \\ w_2 \\ \dots \\ w_d \end{pmatrix} \tag{362} $$

  Wir haben somit

$$ g(\mathbf{x}) = {}^a\mathbf{w}^T {}^a\mathbf{x} = \sum_{i=0}^{d} w_i x_i = -\theta + \sum_{i=1}^{d} w_i x_i = \mathbf{w}^T\mathbf{x} - \theta. \tag{363} $$

Im speziellen ist g **linear** in ${}^a\mathbf{x}$ (und ebenso in ${}^a\mathbf{w}$):

$$\begin{aligned} g(\alpha_1 {}^a\mathbf{x}_1 + \alpha_2 {}^a\mathbf{x}_2) &= {}^a\mathbf{w}^T(\alpha_1 {}^a\mathbf{x}_1 + \alpha_2 {}^a\mathbf{x}_2) = \\ \alpha_1 {}^a\mathbf{w}^{Ta}\mathbf{x}_1 + \alpha_2 {}^a\mathbf{w}^{Ta}\mathbf{x}_2 &= \alpha_1 g({}^a\mathbf{x}_1) + \alpha_2 g({}^a\mathbf{x}_2). \end{aligned} \quad (364)$$

Man beachte, das g nicht linear - im obigen, strengen Sinn - in den nicht-homogenen Koordinaten $\mathbf{w}$ bzw. $\mathbf{x}$ ist.

Die Transformation in homogene Koordinaten vereinfacht unser ursprüngliches Problem, indem es dessen Dimensionalität um 1 (von d auf $d+1$) erhöht; Eq. 363 definiert nun eine d-dimensionale Hyperebene im $\mathsf{R}^{(d+1)}$, welche **durch den Ursprung geht**.

Wir werden im folgenden - falls nicht anders erwähnt - stets homogene Koordinaten annehmen und daher das Superscript ^{a} weglassen.

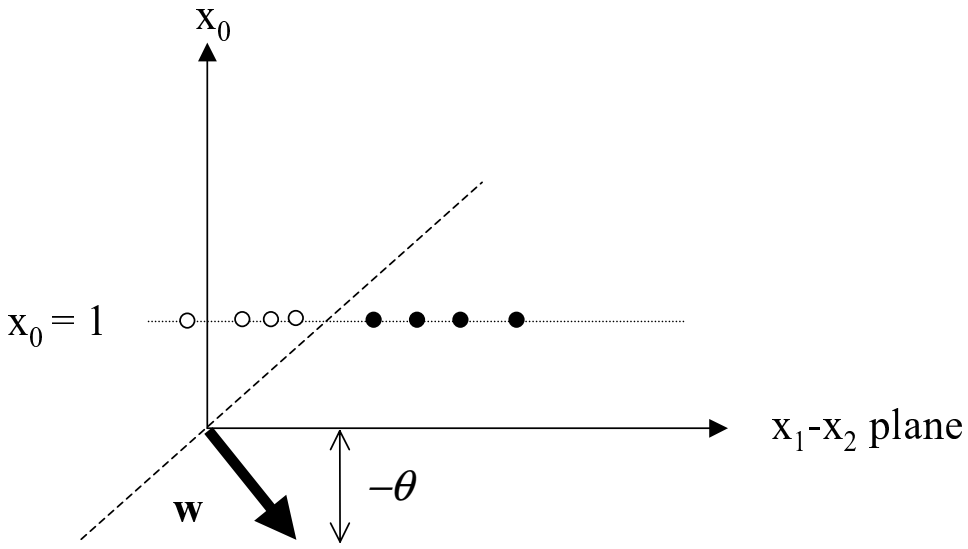


Abbildung 52: Beispiel für homogene Koordinaten im Fall $d = 2$. Ansicht parallel zur Entscheidungsebene). Die homogenen Merkmalsvektoren ($x_i \in \mathbf{R}^3$) liegen auf der $(x_0 = 1)$-Ebene. Die Hyperebene ist nun 2-dimensional, geht **durch den Ursprung** und liegt im $\mathbf{R}^3$. Die Entscheidungsgrenze für nicht-homogene Daten ist durch die Projektion der Schnittgeraden der Hyperebene mit der $(x_0 = 1)$-Ebene auf $(x_0 = 0)$ gegeben.

- **Training**

  Sei $\mathcal{S}_{Tr} = \{\mathbf{X}, \mathbf{y}\}$ eine Menge von N homogenen Merkmalsvektoren $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N) \in \mathbf{R}^{(d+1)\times N}$ und korrespondierenden Klassen-Labels $\mathbf{y}^T = (y_1, \ldots, y_N), y_i \in \{1, -1\}$. $\mathcal{S}_{Tr}$ ist das sogenannte **Trainingsset**.

  Wollten wir z.B. das binäre AND-Problem mittels eines Perceptrons lösen, so hätte unser Trainingsset folgende Form:

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \qquad \mathbf{y}^T = (-1, -1, -1, 1). \tag{365}$$

  Ziel: finde einen homogenen Gewichtsvektor $\mathbf{w}$, sodass

$$o(\mathbf{x}_i) = \text{sgn}(\mathbf{w}^T \mathbf{x}_i) = y_i, \quad 1 \leq i \leq N. \tag{366}$$

Idee: falls ein "positiver" Trainingsvektor $\mathbf{x}_j$ mit $y_j = 1$ falsch klassiert wurde ($\Rightarrow \mathbf{w}^T\mathbf{x}_j < 0$), so addiere ein Vielfaches von $\mathbf{x}_j$ zu $\mathbf{w}$: dadurch wird die Hyperebene auf den falsch klassierten Vektor hinbewegt. Man sieht dass

$$(\mathbf{w} + \gamma\mathbf{x}_j)^T\mathbf{x}_j = \mathbf{w}^T\mathbf{x}_j + \gamma\|\mathbf{x}_j\|^2 > \mathbf{w}^T\mathbf{x}_j,\ \gamma > 0. \tag{367}$$

Der positive Faktor γ wird auch **Lernrate** genannt.

Analog zum obigen Fall, sollte im Fall eines falsch klassierten "negativen" Trainingsvektors $\mathbf{x}_j$ die Hyperebene von diesem wegbewegt werden (indem wir Vielfaches von $\mathbf{x}_j$ von $\mathbf{w}$ subtrahieren).

In beiden Fällen ist es möglich, dass (abhängig vom Wert von γ und dem ursprünglichen $\mathbf{w}$), zuvor korrekt klassierte Vektoren durch die neue Hyperebene nun falsch klassiert werden.

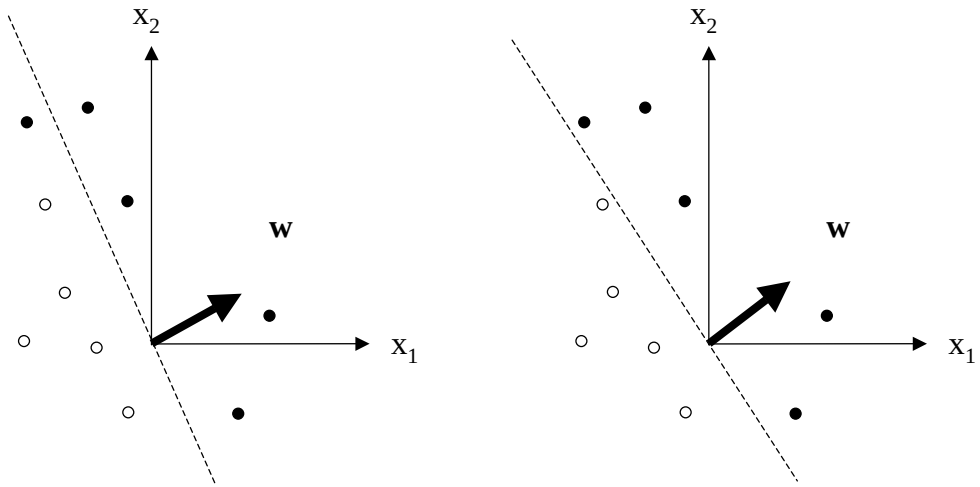


Abbildung 53: Perceptron Training: in der linken Abbildung wurde der obere linke "positive" Vektor $\mathbf{x}_j$ falsch klassiert. Indem wiederholt ein Vielfaches von $\mathbf{x}_j$, $\gamma\mathbf{x}_j, \gamma > 0$ zu $\mathbf{w}$ addiert wird, bewegt sich die Entscheidungsgrenze schließlich über $\mathbf{x}_j$ hinweg (wodurch $\mathbf{x}_j$ richtig klassiert wird). Dies ist in der rechten Abbildung dargestellt ($\gamma << 1$).

Wir können beide Fälle abdecken, indem wir beachten dass

$$\text{sgn}(\mathbf{w}^T\mathbf{x}_i) = y_i \Leftrightarrow \text{sgn}(\mathbf{w}^T\mathbf{x}_i)y_i = 1 \tag{368}$$

$$\Leftarrow (\mathbf{w}^T\mathbf{x}_i)y_i > 0 \Leftrightarrow \mathbf{w}^T(\mathbf{x}_i y_i) > 0. \tag{369}$$

Ausgehend von Eq. 369, welche eine etwas strengere Bedingung als Eq. 368 darstellt (da die Merkmalsvektoren nicht direkt auf der Entscheidungsebene liegen dürfen), suchen wir nun nach einem Gewichtsvektor welcher das modifizierte Trainingsset $\mathbf{x}_i y_i, 1 \leq i \leq N$ (mit ausschließlich positiven Klassen-Labels) in die positive Halb-Ebene abbildet.

Dies führt uns zum **Online Perceptron Training Algorithmus (online Variante):**

1. Initialize $\mathbf{w}, \gamma$
2. **do**
3. **for** $i = 1$ **to** N
4. **if** $\mathbf{w}^T(\mathbf{x}_i y_i) \leq 0$ (misclassified ith pattern)
5. $\mathbf{w} \leftarrow \mathbf{w} + \gamma \mathbf{x}_i y_i$
6. **end if**
7. **end for**
8. **until** all patterns correctly classified

Die Schritte 3. - 7. (Präsentation aller N Trainingsbeispiele) werden häufig als **Epoche** bezeichnet, der Schritt 5. als **Gewichts-Update**.

Werden die Gewichtskorrekturen (so wie oben) sofort angebracht, so spricht man von *incremental* oder *online learning*. Werden die Gewichtsänderungen hingegen akkumuliert und erst am Ende der Epoche (nach Präsentation aller Muster) angebracht, so spricht man von *batch learning*. Letztere Vorgehensweise führt im allgemeinen zu einem glatteren Verlauf der **Lernkurve** (Trainingsfehler als Funktion der Epoche).

Zwei wichtige Fragen

– Wie sollen $\mathbf{w}$, γ initialisiert werden?
– Terminiert der Algorithmus in einer endlichen Anzahl von Schritten?

Initialisierung

Sei $\mathbf{w} = \mathbf{0}$. In diesem Fall ist der mit dem obigen Algorithmus erhaltene Gewichtsvektor $\mathbf{w}_p$ eine Linearkombination der während des Trainings falsch klassierten Merkmalsvektoren:

$$\mathbf{w}_p = \sum_{i=1}^{N} \mathbf{x}_i(y_i \gamma k_i) = \gamma \sum_{i=1}^{N} \mathbf{x}_i(y_i k_i), \;\; k_i \in \mathbf{N}_0, \tag{370}$$

wobei k_i angibt, wie oft der i-te Merkmalsvektor falsch klassiert wurde.

Folglich ist γ lediglich ein Skalierungsfaktor und kann daher einfach $\gamma = 1$ gesetzt werden. Dies gilt allerdings nicht für andere Lernverfahren wie LMS.

Perceptron Konvergenz-Theorem

Der online Perceptron Algorithmus mit fixer Lernrate γ terminiert für jedes linear separierbare Trainingsset mit Lösung $\mathbf{w}_p$, d.h., falls eine separierende Hyperebene mit Normalvektor $\mathbf{w}^*$ existiert.

Der Algorithmus terminiert nicht im Falle eines nicht linear separierbaren Trainingssets (z.B. XOR-Problem).

Die Anzahl der Korrekturschritte (5.) ist nach oben beschränkt durch

$$\left(\frac{\max_j \|\mathbf{x}_j\| \|\mathbf{w}^*\|}{\min_i (\mathbf{w}^{*T} \mathbf{x}_i)} \right)^2, \quad 1 \leq i, j \leq N. \tag{371}$$

Die obige Formel ist jedoch nicht zur praktischen Berechnung der maximalen Anzahl der Iterationsschritte geeignet, da ja die Kenntnis einer Lösung $\mathbf{w}^*$ voraussetzt wird.

- **Margin**

  Eq. 371 steht in engem Zusammenhang mit der Größe

$$gm(\mathbf{x}_i) = \frac{\mathbf{w}^{*T}(\mathbf{x}_i y_i)}{\|\mathbf{w}^*_{(1:d)}\|}, \tag{372}$$

  welche den Abstand des i-ten Merkmalsvektors von der durch $\mathbf{w}^*$ festgelegten Hyperebene angibt und als geometrische *margin* (*geometric margin*) des Vektors $\mathbf{x}_i$ bzg. $\mathbf{w}^*$ bezeichnet wird. Man beachte, dass $gm(\mathbf{x}_i) > 0$ g.d.w. $\mathbf{x}_i$ korrekt klassiert wird.

  Der Vektor $\mathbf{x}_j$ mit minimaler geometrischer *margin* $gm(\mathbf{x}_j)$, also

$$j = \arg\min_i gm(\mathbf{x}_i), 1 \leq i \leq N, \tag{373}$$

legt die geometrische *margin* $gm(\mathbf{w}^*)$ der Hyperebene bzg. des Trainingssets $\{\mathbf{X}, \mathbf{y}\}$ fest: $gm(\mathbf{w}^*) = gm(\mathbf{x}_j)$.

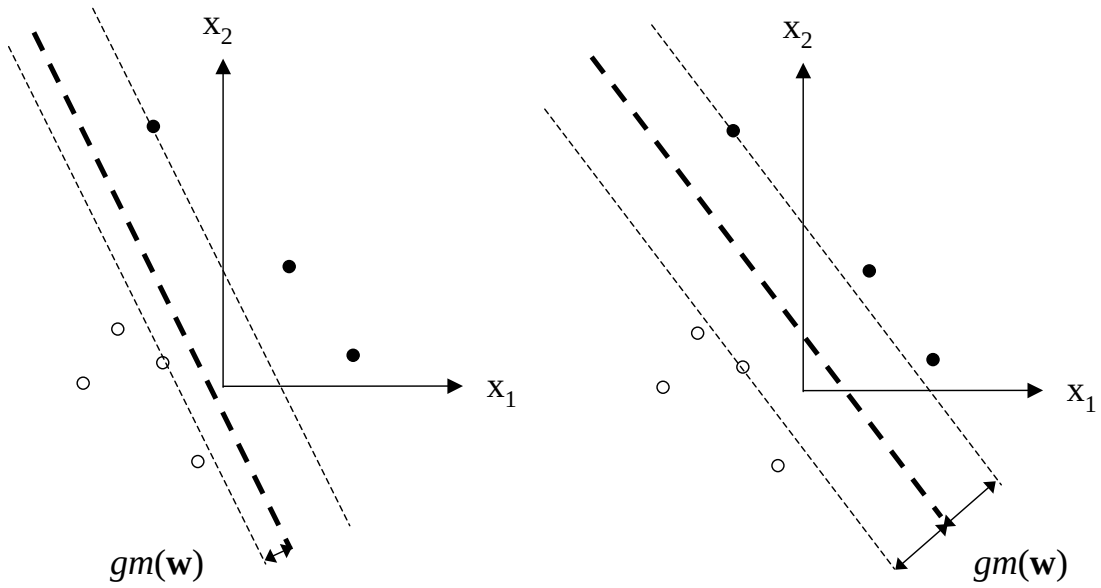


Abbildung 54: Links: eine Hyperebene (fett gestrichelt dargestellt), welche eine Menge von 7 Punkten separiert. Ebenfalls eingezeichnet sind die *margins* der der Hyperebene nächstgelegenen positiven bzw. negativen Beispiele. Die geometrische *margin* der Hyperebene $gm(\mathbf{w})$ ist das Minimum dieser beiden Werte.
Rechts: optimale separierende Hyperebene, welche $gm(\mathbf{w})$ maximiert.

Eq. 371 sagt somit aus, dass die Anzahl der Gewichts-Updates

- reziprok proportional zu $gm(\mathbf{w}^*)^2$ und
- direkt proportional zum Quadrat der Norm des längsten Merkmalsvektors (Radius der kleinsten Hyper-Kugel, welche alle Merkmalsvektoren in $\mathbf{X}$ enthält)

ist.

Für gegebenen Radius der Hyper-Kugel, welche alle Trainingsvektoren enthält, wird der "Schwierigkeitsgrad" des Lernproblems durch jene Vektoren bestimmt, welche am nächsten zur Hyperebene liegen (oder, anders formuliert, durch jene Vektoren, die fast "orthogonal" zu $\mathbf{w}^*$ liegen).

Die Generalisierungsfähigkeit des Perceptrons wird um so besser sein, je größer $gm(\mathbf{w}^*)$ ist; diese Idee - den minimalen Abstand der Trainings-Punkte von der Hyperebene respektive die *margin* $gm(\mathbf{w}^*)$ zu maximieren - liegt der *support vector machine* (SVM) zugrunde. Man spricht in diesem Zusammenhang auch von *maximum margin classifiers*. Siehe auch Fig. 54.

- **Verwandte Verfahren**

  Der Perceptron-Algorithmus in der hier präsentierten Form hat zwei wesentliche Nachteile, welche die Entwicklung leistungsfähigerer Verfahren motiviert haben:

  – Der Perceptron-Algorithmus terminiert nicht im Fall nicht linear separierbarer Daten. Der mit dem Perceptron verwandte *Ho-Kashyap*-Algorithmus erkennt diesen Fall und terminiert auch auf nicht linear separierbaren Daten.
  – Das Perceptron findet nicht unbedingt die optimale Lösung

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} gm(\mathbf{w}) \tag{374}$$

  mit maximaler *margin*. Die moderneren SVMs hingegen finden die optimale Lösung (hierzu muss in der SVM-Formulierung allerdings ein quadratisches Optimierungsproblem unter linearen Nebenbedingungen

gelöst werden). Es gibt auch verschiedene Erweiterungen der SVMs für nicht linear separierbare Daten (Schlupfvariablen, Kernelisierung). SVMs unterscheiden sich von den meisten im folgenden diskutierten Verfahren dadurch, dass sie "verteilungsfreie" Verfahren sind, also nicht auf einer Schätzung der zugrundeliegenden Dichtefunktion der Daten basieren; statt dessen minimieren sie das *worst-case risk*, also den schlimmsten anzunehmenden Fehler.