Klassifikation von Knochenumrissen mit Hilfe spezieller Abstandsmaße

Elias Korbinian Huber

München 2018

Erstgutachter: Prof. Dr. Sonja Greven

Zweitgutachter: Lisa Steyer

Klassifikation von Knochenumrissen mit Hilfe spezieller Abstandsmaße

Bachelorthesis
am Institut für Statistik
Fakultät für Mathematik, Informatik und Statistik
Ludwig–Maximilians–Universität München

Autor:	Elias Korbinian Huber
Betreuung:	Prof. Dr. Sonja Greven
	Lisa Steyer

München, den 06.07.2018

Zusammenfassung

In dieser Arbeit werden zwei distanzbasierte Klassifikationsmethoden, die k-Nächste-Nachbarn-Methode sowie eine kernbasierte Methode, auf Daten, die die Umrisse von Sprungbeinknochen (med. *Talus*) von Schafen verschiedener Domestikationsstatus darstellen, angewandt. Dazu werden verschiedene Möglichkeiten vorgestellt, den Abstand beziehungsweise den Unterschied der *shapes* zweier Objekte zu quantifizieren. Aufgrund der besonderen Struktur der Daten wird dabei sowohl auf Procrustes-Distanzmaße zwischen Landmark-Konfigurationen als auch auf Distanzmaße für durch Pseudo-Landmarks diskretisierte Kurven eingegangen. Es stellt sich dabei heraus, dass eine im R-Paket **fdasrvf** enthaltene Implementierung der auf der *Square Root Velocity* Repräsentation basierenden elastischen Distanz fehlerhaft ist. Als Alternative stellen wir eine Möglichkeit vor, Kurven als Funktionen zu betrachten und darauf basierend Abstandsmaße zu definieren.
Zudem werden zwei konzeptionell unterschiedliche Methoden präsentiert, die Information mehrerer Distanzmaße zu aggregieren. Während die Linearkombination zweier Semi-Metriken als Semi-Metrik wiederum Grundlage für distanzbasierte Klassifikationsmethoden ist, werden durch die Ensemble-Methode die Prognosen verschiedener Klassifikationsmodelle gewichtet kombiniert. Die Stärke dieser Ensemble-Methode liegt auch in der interpretierbaren Gewichtsschätzung, die vor allem bei einer vergleichenden Anwendung der Methode mit Grundmodellen, die auf „traditionellen“ linearen Maßen der Knochen basieren, zu tragen kommt.
Es zeigt sich, dass die diskretisierten Kurven entlang der oberen Kante der Dorsalansicht der Talusknochen diskriminative Kraft besitzen und darauf basierende Distanzen insbesondere in Kombination mit Procrustes-Distanzen zur Klassifikation geeignete Semi-Metriken sind. Während auf einem auf zwei Rassen reduzierten Teil des Datensatzes eine beinahe perfekte Klassifikation gelingt, werden auf dem kompletten Datensatz ungefähr ein Viertel der vorhandenen Daten falsch klassifiziert.

Inhaltsverzeichnis

1 Einleitung

Miss alles, was sich messen lässt, und mach alles messbar, was sich nicht messen lässt.

Dieses meist Archimedes (287 - 212 v.Chr.) zugeordnete Zitat [1] zeigt die Faszination der Menschen, Objekte und Zusammenhänge zu quantifizieren und damit vergleichbar zu machen. Während man intuitiv beim Begriff „messen“ eher an Längen, Gewichte oder Geschwindigkeiten denkt und in der Messung dieser Größen keine besonderen Herausforderungen oder Probleme sieht, gibt es auch Objekte und Eigenschaften dieser, deren geeignete Messung nicht so klar ist und die vielleicht auf den ersten Blick sogar „nicht messbar“ scheinen. Ein Beispiel dafür ist die Ähnlichkeit der Formen (englisch *shape*) zweier Objekte.

Die Morphometrie befasst sich mit der quantitativen Analyse des *shape* eines Objektes, beziehungsweise der Variation der *shapes* zwischen mehreren Objekten [21, S.299]. Dabei kann man die Ansätze in zwei grundsätzliche Herangehensweisen gliedern. In der traditionellen Morphometrie (auch: multivariate Morphometrie) werden multivariate statistische Methoden auf eine Menge von Variablen – meist Längenmaße zwischen spezifischen Stellen, manchmal auch Verhältnisse oder Winkel zwischen diesen Längen – angewandt [22, S.129]. Auch wenn es einfach erscheint, diese Variablen festzuhalten, zeigt sich ein wesentliches Problem der traditionellen Morphometrie: Die geometrischen Beziehungen zwischen den linearen Distanzen werden häufig nicht festgehalten, wodurch teilweise wesentliche Information über den *shape* verloren geht [2, S.6]. Es ist demnach nicht möglich, den (abstrakt repräsentierten) *shape* des Objektes aus den festgehaltenen Variablen zu reproduzieren [22, S.129]. Daran ansetzend wird bei den Methoden der geometrischen Morphometrie die geometrische Information der Objekte in Koordinatenbasierten Daten festgehalten [5, S.1395].

Eine der typischsten Anwendungen morphometrischer Methoden findet sich in der Vermessung von Knochen, welche sowohl für Biologen als auch für Archäologen interessant ist. Die Fragestellung dieser Arbeit ist, ob es anhand von Daten, die den Umriss von Sprungbeinknochen (medizinisch *talus*) von Schafen beschreiben, möglich ist, den Domestizierungsstatus dieser Tiere zu bestimmen. Dabei bauen wir auf einer vorhergehenden Analyse von Pöllath et. al. [17] eines kleineren, verwandten Datensatzes auf. In dieser Arbeit sollen distanzbasierte Klassifikationsmethoden benutzt werden, die auf der Prämisse beruhen, dass sich in den Kovariablen nähere Objekte eher der gleichen Klasse entstammen. Es müssen also Distanzmaße definiert werden, die die Unähnlichkeit zweier Knochenumrisse quantifizieren.

Der weitere Verlauf der Arbeit gliedert sich wie folgt: In Kapitel 2 werden die verwendeten Klassifikationsmethoden vorgestellt. In Kapitel 3 wird auf die besondere Struktur der vorliegenden Daten eingegangen. Darauf basierend werden in Kapitel 4 verschiedene Abstandsmaße zur Messung der Unähnlichkeit zweier Knochenumrisse eingeführt. Die Güte der verschiedenen Klassifikationsmethoden und Distanzmaße bei der Klassifikation der Knochen wird in Kapitel 5 verglichen. Abschließend werden die Ergebnisse der Arbeit zusammengefasst und ein kurzer Ausblick auf mögliche weitere Schritte gegeben.

2 Verwendete Klassifikationsmethoden

2.1 Allgemeines

Die Aufgabenstellung bei Klassifikationsproblemen ist es, anhand von bekannten Kovariablen eines Objekts die Zugehörigkeit dessen zu einer von G Klassen vorherzusagen. Genauer bezeichnet überwachte Klassifikation die Situation, dass im gegebenen Datensatz $\mathcal{D}$ n Beobachtungen $(x_i, y_i), i = 1, \ldots, n$ aus $\mathcal{X} \times \mathcal{G},\ \mathcal{G} = \{1, .., G\}$ vorliegen, wobei $\mathcal{X}$ zunächst ein allgemeiner Kovariablenraum sein soll. Es wird angenommen, dass $(x_i, y_i), i = 1, \ldots, n$ zufällige, unabhängige Realisierungen der mehrdimensionalen Zufallsvariable $(X, Y) \sim \mathbb{P}_{\mathcal{X}\mathcal{G}}$ sind. Nun soll für ein neues, ebenfalls $\mathbb{P}_{\mathcal{X}\mathcal{G}}$ entstammendes Datum (x, y) anhand dem bekannten $x \in \mathcal{X}$ die entsprechende unbekannte Klasse (auch: *label*) $y \in \mathcal{G}$ bestimmt werden, wobei die n gegebenen Beobachtungen zum Training verwendet werden dürfen (und sollen).
Wir suchen zunächst eine möglichst gute Prädiktionsfunktion $f : \mathcal{X} \to \mathcal{G}$, wobei die Güte einer Vorhersage $f(x) = \hat{y}_x$ anhand einer sogenannten Verlustfunktion $\mathcal{L} : \mathcal{G} \times \mathcal{G} \to \mathbb{R}_0^+$ und die Güte einer konkreten, auf den Trainingsdaten $\mathcal{D}_{Train}$ geschätzten vorhersagenden Funktion $\hat{f}_{\mathcal{D}_{Train}}$ als Erwartungswert einer Verlustfunktion, dem *Generalization Error*

$$GE(\hat{f}_{\mathcal{D}_{Train}}) = \mathbb{E}[\mathcal{L}(\hat{f}_{\mathcal{D}_{Train}}(X), Y)|\mathcal{D}_{Train}],$$

bestimmt werden kann. Dabei bezieht sich der Erwartungswert auf die Zufälligkeit der Testdaten als Ausprägungen von (X, Y). Wir sind jedoch vor allem interessiert an der Güte eines Lernalgorithmus $\mathcal{A}$, der auf einen Trainingsdatensatz $\mathcal{D}_{Train}$ der Größe $|\mathcal{D}_{Train}| = n$ angewandt wird, woraus sich eine Vorhersagefunktion $\hat{f}_{\mathcal{D}_{Train}}$ ergibt. Der *Expected Generalization Error*

$$EGE_n(\mathcal{A}) = \mathbb{E}_{|\mathcal{D}_{Train}|=n}[GE(\hat{f}_{\mathcal{D}_{Train}})] = \mathbb{E}_{|\mathcal{D}_{Train}|=n}[\mathcal{L}(\hat{f}_{\mathcal{D}_{Train}}(X), Y)]$$

berücksichtigt nun auch die Zufälligkeit der Ziehung der n Trainingsdaten aus $\mathbb{P}_{\mathcal{X}\mathcal{G}}$.
Eine häufig bei Klassifikationsproblemen verwendete Verlustfunktion ist der sogenannte *0-1 Loss*

$$\mathcal{L}(\hat{y}_x, y) = \begin{cases} 0 & \text{wenn } \hat{y}_x = y \\ 1 & \text{wenn } \hat{y}_x \neq y \end{cases},$$

der sämtliche Fehlklassifikationen, unabhängig von der Ausprägung der tatsächlichen oder der geschätzten Klasse, mit dem Wert 1 bestraft [14, S.20]. Man kann zeigen, dass die Minimierung des entsprechenden erwarteten Generalisierungsfehlers zu der als Bayes-Regel bekannten Entscheidungsfunktion

$$\hat{y}_x = f(x) = \arg\max_{g \in \mathcal{G}} \mathbb{P}(Y = g|X = x)$$

führt,vgl. [14, S. 20f.]. Dieses wenig verwunderliche Ergebnis besagt, dass die in Bezug auf den obig vorgestellten *0-1 Loss* beste Klassifikation eines $x \in \mathcal{X}$ der Modus der auf X bedingten, diskreten Wahrscheinlichkeitsverteilung $\mathbb{P}(Y|X = x)$ ist. Diese Verteilung ist allerdings unbekannt.

Im Folgenden werden zwei verwandte Methoden vorgestellt, diese bedingten Klassenwahrscheinlichkeiten zu schätzen.

2.2 k-Nächste-Nachbarn-Schätzer

Der erste Schritt dieser Approximation besteht darin, die bedingte Wahrscheinlichkeit $\mathbb{P}(Y = g|X = x)$, kurz $p_g(x)$, durch die bedingte relative Häufigkeit in den beobachteten Daten $\mathcal{D}$ zu schätzen:

$$\hat{p}_g(x) = \frac{1}{|\{x_i = x\}_{i=1,..,n}|} \sum_{i:x_i=x} \mathbf{1}_{[y_i=g]}.$$

Sobald jedoch stetige Größen unter den Kovariablen sind, gilt

$$\mathbb{P}(|\{x_i = x\}_{i=1,..,n}| = 0) \overset{\text{f.s.}}{=} 1,$$

weshalb die punktweise Bedingung $x_i = x$ zu $x_i \in N_k(x)$ „gelockert“ wird, wobei $N_k(x)$ die Nachbarschaft von x mit den k nächsten Datenpunkten $x_{(1)}, \ldots, x_{(k)}$ (im Abstand von x aufsteigend geordnet) bezeichnet:

$$x_i \in N_k(x) \Leftrightarrow d(x_i, x) \leq d(x_k, x).$$

Da die Wahl der Abstandsfunktion $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_0^+$ an zentraler Stelle in dieser Arbeit steht und in Kapitel 4 genau behandelt wird, soll hier auf diese nicht genauer eingegangen werden. Als „k-nächste-Nachbarn“-Schätzer (kurz: kNN-Schätzer) für $p_g(x)$ ergibt sich

$$\hat{p}_f(x) = \frac{1}{k} \sum_{i:x_i \in N_k(x)} \mathbf{1}_{[y_i=g]}.$$

Es ist offensichtlich, dass $\hat{p}_g(x)$ die Eigenschaften

$$\begin{aligned} &\text{i) } 0 \leq \hat{p}_g(x) \ \forall y \in \mathcal{G} \text{ und} \\ &\text{ii) } \sum_{g \in \mathcal{G}} \hat{p}_g(x) = 1 \end{aligned}$$

besitzt und damit ein wohldefinierter Schätzer einer Wahrscheinlichkeitsverteilung ist. Mit diesem Schätzer lässt sich die Entscheidungsregel folgendermaßen schreiben:

$$\hat{y}_x = f(x) = \arg\max_{g \in \mathcal{G}} \hat{p}_g(x) = \arg\max_{g \in \mathcal{G}} \sum_{i:x_i \in N_k(x)} \mathbf{1}_{[y_i=g]}.$$

Es wird also die Klasse prognostiziert, die unter den k nächsten Nachbarn am häufigsten vertreten ist.

2.3 Kernbasierte Schätzer

Bei dem kNN-Schätzer lässt sich Folgendes beobachten: Der Datenpunkt $x_{(k)}$, der im Kovariablenraum am k-weitesten von x entfernt ist, besitzt denselben Einfluss auf die Prognose $\hat{y}_x$ wie der näheste Datenpunkt $x_{(1)}$. $x_{(k+1)}$ hingegen beeinflusst den Schätzer überhaupt nicht. Hier setzen kernbasierte Schätzer an. Den Datenpunkten werden Gewichte zugeteilt, die mit zunehmendem Abstand vom zu klassifizierenden Datum stetig abnehmen, im Gegensatz zu der effektiven 0/1-Gewichtung des kNN-Schätzers [14, S.17]. Diese flexiblere Gewichtung wird erreicht durch Berechnung der Werte $\Delta_1, \ldots, \Delta_n$ über

$$\Delta_i = \Delta_i(x, h, K) = \frac{1}{h} K(\frac{d(x_i, x)}{h}).$$

Dabei ist h die sogenannte Bandweite und $K(\cdot)$ eine Kernfunktion. Theoretisch kann jede Dichte und sogar jede nichtnegative Funktion als Kernfunktion genutzt werden [10, S.38]. Wir beschränken uns jedoch auf die folgenden klassischen asymmetrischen Dichten mit positivem Träger ($\{v \in \mathbb{R} : K(v) > 0\} \subset \mathbb{R}_0^+$), die beispielsweise von Ferraty und Vieu [10, S.40] vorgestellt und verwendet werden:

i) Asymmetrischer Box-Kern :	$K_B(u) = \mathbf{1}_{[0,1]}(u)$
ii) Asymmetrischer Dreieckskern :	$K_D(u) = 2(1-u)\mathbf{1}_{[0,1]}(u)$
iii) Asymmetrischer quadratischer Kern :	$K_Q(u) = \frac{3}{2}(1-u^2)\mathbf{1}_{[0,1]}(u)$
iv) Asymmetrischer Gauss-Kern :	$K_G(u) = \frac{2}{\sqrt{2\pi}}\exp(-\frac{u^2}{2})\mathbf{1}_{[0,\infty)}(u)$

In Abbildung 1 sind diese vier Kerne dargestellt. Es ist zu erkennen, dass die Kerne i)-iii) einen kompakten Träger besitzen, denn für diese Kerne gilt $u > 1 \Rightarrow K(u) = 0$, was bedeutet, dass Datenpunkte x_i mit großem Abstand von x (genauer $d(x_i, x) > h$) mit $\Delta_i = 0$ gewichtet werden und damit die Schätzung $\hat{y}_x$ nicht beeinflussen. Bei der Verwendung des asymmetrischen Gauss-Kerns hingegen hat jedes Datum x_i auf die Prognose $\hat{y}_x$ Einfluss, wenn auch teilweise nur sehr wenig. Während mit dem asymmetrischen Box-Kern jedes x_i im Bereich der Bandweite h identisch gewichtet wird, nimmt unter Verwendung der Kerne ii)-iv) der Einfluss der Trainingsbeobachtungen mit zunehmendem Abstand vom zu klassifizierenden Datum ab. Man kann diese Verringerung des Einflusses quantifizieren, indem man Quotienten der Form

$$Q(u) = K(u)/K(0), u > 0$$

bildet und diese für die verschiedenen Kernfunktionen vergleicht. Im Folgenden wollen wir dies für den asymmetrischen Dreieckskern K_D, den asymmetrischen quadratischen Kern K_Q und den asymmetrischen Gauss-Kern K_G an den Werten $u_1 = 0.2$ und $u_2 = 0.8$ durchführen. Für u_1 ergeben sich die Werte $Q_D(u_1) = 0.8$, $Q_Q(u_1) = 0.96$ und $Q_G(u_1) \approx 0.98$. Demnach ist bei Verwendung des asymmetrischen Dreieckskerns der Einfluss eines Datums x_j mit $d(x_j, x) = 0.2h$ um den Faktor 0.8 geringer als der eines Datums x_k mit $d(x_k, x) = 0$. Bei quadratischem Kern und Gauss-Kern haben diese beiden Daten jedoch beinahe das gleiche Gewicht. An den Werten

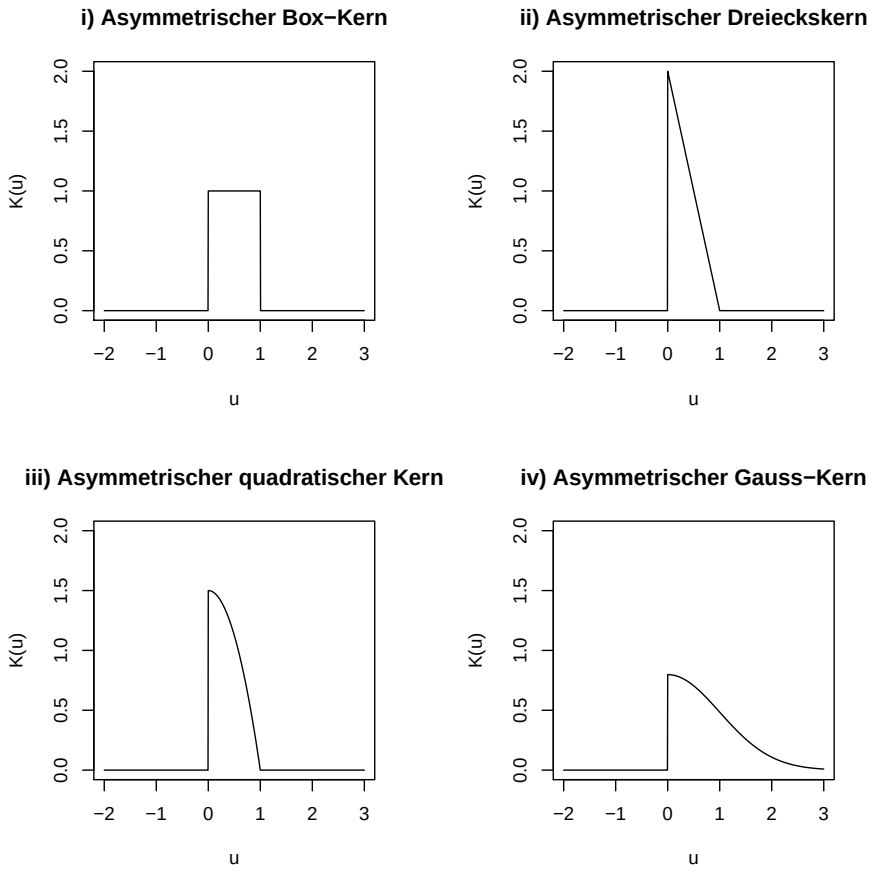

Abbildung 1: Klassische asymmetrische Kerne

$Q_D(u_2) = 0.2$, $Q_Q(u_2) = 0.36$ und $Q_G(u_2) \approx 0.73$ erkennt man eine ähnliche Struktur. Der asymmetrische Dreieckskern gewichtet ein Datum x_l mit $d(x_l, x) = 0.8h$ im Verhältnis zu x_k geringer als die Kerne K_Q und K_G. Bei gleicher Bandbreite h fällt demnach bei steigender Distanz das Gewicht unter Verwendung des asymmetrischen Dreieckskerns schneller ab als beim asymmetrischen quadratischen Kern und dem asymmetrischen Gauss-Kern, welcher den flachsten Verlauf besitzt.

Nun, da wir uns mit Kernfunktionen und der abstandsbasierten Gewichtung auf Basis dieser beschäftigt haben, können wir eine Schätzung für $p_g(x)$ unter Verwendung dieser Gewichte formulieren. Der kernbasierte Schätzer ergibt sich zu

$$\hat{p}_g(x) = \frac{\sum_{i=1}^{n} \mathbf{1}_{[y_i=g]} \Delta_i}{\sum_{i=1}^{n} \Delta_i} = \frac{\sum_{i=1}^{n} \mathbf{1}_{[y_i=g]} \frac{1}{h} K(\frac{d(x_i,x)}{h})}{\sum_{i=1}^{n} \frac{1}{h} K(\frac{d(x_i,x)}{h})} = \frac{\sum_{i=1}^{n} \mathbf{1}_{[y_i=g]} K(\frac{d(x_i,x)}{h})}{\sum_{i=1}^{n} K(\frac{d(x_i,x)}{h})}.$$

Der Nenner dient hierbei zur Normierung. So wird sichergestellt, dass die zwei bereits bei der Beschreibung des kNN-Schätzers erwähnten Eigenschaften

$$\text{i) } 0 \leq \hat{p}_g(x) \ \forall y \in \mathcal{G} \text{ und}$$
$$\text{ii) } \sum_{g \in \mathcal{G}} \hat{p}_g(x) = 1$$

gelten und somit die geschätzten Wahrscheinlichkeiten eine wohldefinierte diskrete Wahrscheinlichkeitsverteilung ergeben [10, S.115]. Man sieht in der Formel des kernbasierten Schätzers, dass sich sowohl der Vorfaktor h^{-1} als auch normierende Faktoren in den Kernfunktionen kürzen. Das Zentrale an der kernbasierten Gewichtung und Schätzung ist demnach die Form des Kerns, das Abstandsmaß und die gewählte Bandbreite, mit deren Bedeutung wir uns nun zum Abschluss dieses Kapitels auseinander setzen wollen.

2.4 Auswirkung der Lokalisationsparameter

Zu Beginn dieses Absatzes soll auf eine weitere logische Verknüpfung von kNN-Schätzern und kernbasierten Schätzern hingewiesen werden. Wie schon bemerkt, gilt unter Verwendung von Kernen mit kompakten Trägern, dass Daten, deren Distanz vom zu prognostizierendem Datum größer als h ist, keinerlei Einfluss auf diese Prognose haben. Wählt man nun $h_k = d(x_{(k)}, x)$, wobei $x_{(k)}$ das am k-weitesten von x entfernte Trainingsdatum ist, so fließen genau k Datenpunkte in die Prognose $\hat{p}_g(x)$ mit ein. Während die kernbasierte Schätzung unter Verwendung des asymmetrischen Box-Kerns und h_k damit genau dem kNN-Schätzer entspricht, werden mit dem asymmetrischen Dreieckskern und dem asymmetrischen quadratischen Kern die k nächsten Daten noch gemäß ihrem Abstand von x gewichtet. Da also ein kleines (respektive großes) k beim kNN-Schätzer und ein kleines (resp. großes) h bei kernbasierten Schätzern Ähnliches zu bewirken scheinen, soll im folgenden allgemein von kleinen (resp. großen) Lokalisationsparametern gesprochen werden.
Wie wirkt sich diese Lokalität auf den erwarteten Generalisierungsfehler EGE aus? Um diese Frage zu beantworten, betrachten wir eine Bias-Varianz-Zerlegung des EGE. In der wohl bekanntesten Anwendungsform, einem Regressionsproblem $Y = f(X) + \epsilon, \mathbb{E}(\epsilon) = 0, Var(\epsilon) = \sigma^2$ mit quadratischer Verlustfunktion $\mathcal{L}(\hat{y}_x, y) = (\hat{y} - y)^2$ kann gezeigt werden (vgl. [14, S.223]), dass sich der EGE an einem „Input“ $X = x$ zerlegen lässt zu

$$EGE(x) = \sigma^2 + (\mathbb{E}[\hat{f}(x)] - f(x))^2 + \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2] = \sigma^2 + Bias(\hat{f}(x)) + Var(\hat{f}(x)).$$

Diese Darstellung ist sehr geläufig und auch leicht verständlich. Domingos [7, S.4] führt allerdings in Vorbereitung auf eine Verallgemeinerung der Bias-Varianz-Zerlegung eine Darstellung der drei Komponenten in Bezug auf die verwendete Verlustfunktion $\mathcal{L}$ ein: Der erste Term, σ^2, ist die Varianz der Zielvariable um den wahren, unterliegenden Funktionswert $f(x)$. Er kann auch gesehen werden als der erwartete Verlust $\mathbb{E}[\mathcal{L}(f^*(x), y_x)]$ der optimalen Vorhersage $f^*(x) = f(x)$. Da dieser Term nicht beeinflusst werden kann, wird er auch als *Noise* bezeichnet. Der zweite Term ist der Bias, der allgemein der Verlust der erwarteten Vorhersage $\mathbb{E}[\hat{f}(x)]$ in Bezug auf die optimale Vorhersage $f^*(x)$ ist. Die dritte Komponente ist die Varianz der $\hat{f}(x)$, die auch als erwarteter Verlust $\mathbb{E}[\mathcal{L}(\hat{f}(x), \mathbb{E}[\hat{f}(x)])]$ von den einzelnen Vorhersagen $\hat{f}(x)$ verglichen mit der erwarteten Vorhersage $\mathbb{E}[\hat{f}(x)]$ ausdrückbar ist.
Mit Hilfe dieser Definitionen zeigt Domingos [7, S.8], dass auch bei Klassifikationsproblemen und der Verwendung des *0-1 Loss* der EGE einer additiven Kombination aus Noise, Bias und

Varianz entspricht. Die optimale Vorhersage $f^*(x)$ entspricht dabei der schon zu Beginn des Kapitels vorgestellten Bayes-Regel, und der entsprechende *Noise* ist

$$\mathbb{E}[\mathcal{L}(f^*(x), y_x)] = 1 - \max_{y \in \mathcal{G}} \mathbb{P}(y|\mathcal{X} = x),$$

die sogenannte Bayes-Fehlerrate.

Betrachten wir nun unsere beiden lokalen Klassifikationsmethoden, den kNN-Schätzer und den kernbasierten Schätzer. Für einen kleinen Lokalisationsparameter k beziehungsweise h entspricht eine konkrete Schätzung $\hat{f}(x)$ dem Modus von wenigen, aus einer meist zu $\mathbb{P}(y|\mathcal{X} = x)$ ähnlichen Verteilung gezogenen Labels. Der Erwartungswert dieser Schätzungen ist demnach meist der Modus der bedingten Wahrscheinlichkeitsverteilung $\mathbb{P}(Y|\mathcal{X} = x)$, was der Definition der Bayes-Regel entspricht. Demnach ist der Bias in diesem Fall sehr gering. Da wie erwähnt die einzelnen konkreten Schätzungen jedoch nur von einer geringen Anzahl von gezogenen Labels abhängen, kann es häufiger vorkommen, dass sich einzelne Schätzungen von der erwarteten Vorhersage unterscheiden. Die Varianz ist also größer. Das Gegenteil zeigt sich beim anderen Extrem, der Verwendung eines sehr großen Lokalisationsparameters. Hier fließen nun (fast) alle n aus gezogenen Trainingsdaten in die Schätzung $\hat{f}(x)$ mit ein, wodurch sowohl diese als auch ihr Erwartungswert meist dem Modus der a-priori-Verteilung $\mathbb{P}(y)$ der Zielgröße y entsprechen. Folglich ist die Varianz gering. Da es aber vorkommt, dass für manche x der Modus der a-priori-Verteilung ungleich dem Modus der auf dieses x bedingten Verteilung $\mathbb{P}(Y|\mathcal{X} = x)$ ist, ist der Bias hier größer. Generell lässt sich also sagen, dass mit wachsendem Lokalisationsparameter der Bias zu- und die Varianz abnimmt (vgl. [14, S.38]). Dieses Zusammenspiel nennt sich *Bias-Variance-Tradeoff*. Eine allgemeine Möglichkeit, den *Estimated Generalization Error EGE* zu schätzen und dabei das empirisch „beste“ k beziehungsweise h zu bestimmen, wird in Unterkapitel 5.1 vorgestellt und in 5.3 auf kNN-Schätzer und kernbasierte Schätzer angewandt, denen verschiedene Abstandsmaße zu Grunde liegen. Eine weitere Möglichkeit, die geeignetsten Lokalisationsparameter zu bestimmen, bietet die im folgenden Abschnitt präsentierte Ensemble-Methode.

2.5 Ensemble-Methode

In diesem Unterkapitel soll eine Möglichkeit vorgestellt werden, Klassifikationsmodelle zu aggregieren. Dabei lehnen wir uns stark an das von Fuchs et. al. [12] vorgestellte *Nearest Neighbor Ensemble*, welches ursprünglich zur Klassifikation eindimensionaler funktionaler Daten eingeführt wurde. Die Struktur der Daten spielt dabei jedoch nur eine untergeordnete Rolle, so das es ohne Weiteres möglich ist, diese Ensemble-Methode auf Daten der vorliegenden Struktur, also Objekte beziehungsweise Kurven im zweidimensionalen Raum, anzuwenden.

Bei der Ensemble-Methode werden Linearkombinationen der geschätzten Posteriori-Wahrscheinlichkeiten $\hat{p}_g^l, l = 1, \ldots, L$ für Klasse $g \in \mathcal{G}$ von L verschiedenen Modellen gebildet. Dabei ist das l-te Modell für unsere Anwendung ein kNN-Modell mit einer Nachbarschaft der Größe k_l und verwendetem Distanzmaß d_l. Theoretisch ist es möglich, dieses kNN-Ensemble um kernbasierte Modelle zu ergänzen. Darauf wollen wir aufgrund des erhöhten komputationalen Aufwandes je-

doch verzichten. Die aggregierte Schätzung der Posteriori-Wahrscheinlichkeit von Klasse g hat die Form

$$\hat{p}_g = \sum_{l=1}^{L} \hat{c}_l \hat{p}_g^l.$$

Dabei ist $\hat{c}_l$ das geschätzte Gewicht des l-ten Grundmodells. Diese Gewichte unterliegen den Restriktionen

$$c_l \geq 0 \ \forall l \text{ und } \sum_{l=1}^{L} c_l = 1,$$

wodurch gesichert ist, dass für die Schätzungen der Klassenwahrscheinlichkeiten $0 \leq \hat{p}_g \leq 1 \ \forall g$ gilt [12, S.188]. Da die Gewichte eines Modells für jede Schätzung gleich sind, gilt zudem $\sum_{g \in G} \hat{p}_g = 1$ [13, S.37f.]. Zusätzlich führt diese Lasso-artige Penalisierung bei der Schätzung der Gewichtskoeffizienten dazu, dass auch Gewichte $c_l = 0$ geschätzt werden können, womit das entsprechende Modell nicht im Ensemble enthalten ist. Es findet also eine automatische Modellselektion statt. Nun stellt sich noch die Frage, wie die Gewichte auf den vorhandenen n Daten geschätzt werden. Fuchs et. al. minimieren dazu den *Brier Score* (eingeführt in [6])

$$Q = \sum_{i=1}^{n} \sum_{g=1}^{G} (z_{ig} - \hat{p}_g(x_i))^2$$

mit

$$z_{ig} = \begin{cases} 0 & \textit{wenn } y_i \neq g \\ 1 & \textit{wenn } y_i = g \end{cases},$$

der über $\hat{p}_g$ von den Gewichten abhängt [12, S.189]. Mit Matrizen kann der Brier Score in Abhängigkeit vom Gewichtsvektor $c = (c_1 \ \dots \ c_L)^T$ durch

$$Q(c) = (z - Pc)^T (z - Pc)$$

mit dem Vektor $z = (z_1 \ \dots z_n)^T$ mit $z_i = (z_{i1} \ \dots z_{iG})^T$ und der Matrix $P = (P_1^T \ \dots P_N^T)^T$ mit

$$P_i = \begin{pmatrix} \hat{p}_1^1(x_i) & \dots & \hat{p}_1^L(x_i) \\ \vdots & \dots & \vdots \\ \hat{p}_g^1(x_i) & \dots & \hat{p}_g^L(x_i) \end{pmatrix}$$

dargestellt werden, wobei die Schätzungen $p_g^l(x_i)$ des l-ten Grundmodells für das Datum x_i auf den restlichen Trainingsdaten (typischerweise entsprechend der *Leave-one-out Cross Validation*, welche in Abschnitt 5.1 vorgestellt wird) durchgeführt werden [12, S.189]. Das Minimieren von $Q(c)$ führt zum geschätzten Koeffizientenvektor

$$\hat{c} = \arg\min_c Q(c).$$

Durch diese Gewichte lässt sich Information über die Klassifikationsstärke der Grundmodelle gewinnen, da sie widerspiegeln, welche Modelle am wichtigsten zur möglichst erfolgreichen Klassifikation sind [12, S.187]. Doch welche Basismodelle wählt man aus? Wie bereits erwähnt, spielt neben dem Lokalisationsparameter auch die Wahl des Abstandsmaßes sowohl bei Modellen nach der kNN-Methode als auch bei kernbasierten Modellen eine zentrale Rolle. Damit hat sie über die aufgenommenen Basismodelle auch großen Einfluss auf die Ensemble-Methode. Die Wahl des Abstandsmaßes hängt sehr stark von der Anwendung ab. Um genaue Anforderungen an diese Distanzmaße stellen zu können, müssen wir uns deshalb zuerst mit der besonderen Struktur der vorliegenden Daten beschäftigen. Dies soll im folgenden Kapitel geschehen.

Bezeichnung (Rasse/Ausgrabung)	**Domestikations-status**	**Anzahl Umrisse**	**Anzahl Lineare Maße**
Karakulschaf	spätes Hausschaf	40	40
Marschschaf	spätes Hausschaf	23	23
Tall Munbaqa	spätes Hausschaf	46	50
Soayschaf	frühes Hausschaf	41	65
Güvercinkayası	frühes Hausschaf	29	30
Elburs-Gebirge	Wildschaf	41	40
Göbekli Tepe	Wildschaf	36	22
Gusir Höyük	Wildschaf	17	17

Tabelle 1: Übersicht über die Schafe, deren Knochen in den beiden Datensätzen vorhanden sind.

3 Knochenumrisse als *shape*-Objekte

3.1 Vorstellung der Daten

Für diese und vorhergehende Analysen wurden die Umrisse vom Sprungbeinknochen (medizinisch *Talus*, veraltet auch *Astragalus*) von Schafen verschiedener Rassen festgehalten. Genauer sind im vorliegenden Datensatz die zweidimensionalen Dorsalansichten (zum Fußrücken hin) von 273 Knochen vorhanden. Dabei teilt sich der Datensatz wie in Tabelle 1 aufgeführt in acht Rassen beziehungsweise archäologische Ausgrabungsstätten auf. Der Haltungsstatus der Schafe als Zielvariable, nach der klassifiziert werden soll, hat drei Ausprägungen: Die Schafe aus dem Elburs-Gebirge und aus den südostanatolischen Ausgrabungsstätten Göbekli Tepe und Gusir Höyük bilden zusammen die Population der Wildschafe (insgesamt also 94 Beobachtungseinheiten), die verwilderten Soayschafe und die Schafe aus Güvercinkayası werden zu frühen (primitiven) Hausschafen (70) zusammengefasst, und die Gruppe der durch Züchtung bereits verbesserten, spät domestizierten Schafe (109) ergibt sich aus den Karakulschafen, den aus den norddeutschen Marschen stammenden Marsch-Schafen und den Schafen aus Tall Munbaqa. Die Einteilung mancher Rassen beziehungsweise Schafgruppen (insbesondere die Schafe aus den archäologischen Fundorten Tall Munbaqa und Güvercinkayası) ist dabei nicht komplett sicher, da die Knochen selbst die einzige Informationsquelle über die Schafe sind. Zudem ist auch der Domestizierungsprozess fortlaufend, wonach zu vermuten ist, dass die Klassen nicht unbedingt komplett homogen sind.

Zusätzlich steht ein weiterer Datensatz zur Verfügung, in dem lineare Messungen an der Dorsalansicht von insgesamt 287 Knochen enthalten sind. Diese Maße sind in Abbildung 2 aufgezeichnet. Dabei sind GLm, GLl und Bd traditionelle Maße nach von den Driesch [26, S.89]. M1, M2, und M3 sind Maße entlang der oberen Seite des Knochens, die Pöllath et. al. [17] als Alternativen beziehungsweise Ergänzungen zu den traditionellen Maßen vorgeschlagen haben. Eine Klassifikation mit Hilfe dieser Maße soll als Vergleich zu der primären Analyse anhand der Umrissdaten dienen.

Neben dieser Analyse auf dem gesamten Datensatz wollen wir auch die Klassifikation auf einem eingeschränkten Teil davon durchführen, um die Ergebnisse der verwendeten Methoden besser mit einem vorhergehenden Klassifikationsversuch vergleichen zu können. Pöllath et. al.

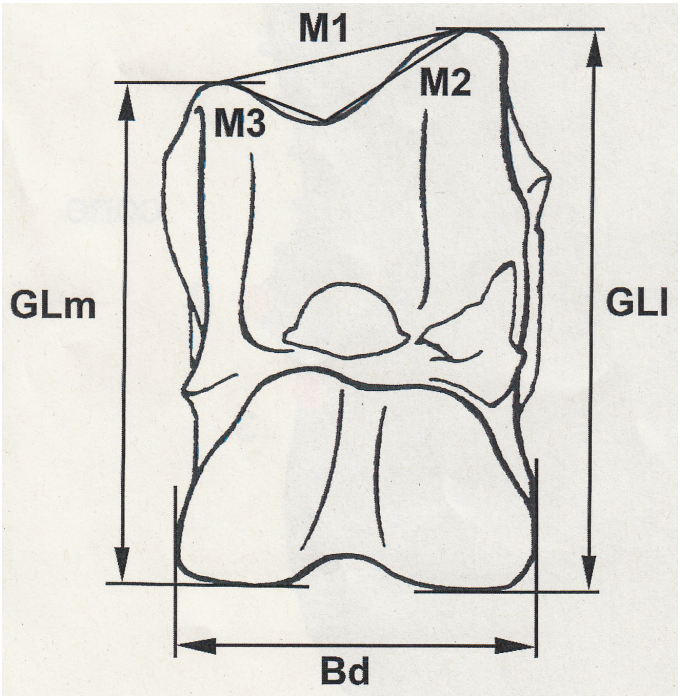

Abbildung 2: Lineare Maße an der Dorsalansicht eines Talusknochen. Aus [17].

[17] haben eine binäre Klassifikation auf einem Datensatz basierend auf Knochen der vollständig domestizierten Karakulschafe und den in Göbekli Tepe und Gusir Höyük gefundenen Knochen von Wildschafen durchgeführt. Wir werden uns folglich auf die entsprechenden im Datensatz vorhandenen Knochenumrisse beschränken. Da Pöllath et. al. zusätzlich zu der Dorsalansicht noch Daten aus der Medialansicht benutzt haben und dementsprechend einige Knochen mit beschädigter Medialansicht nicht verwenden konnten, unterscheiden sich die Datensätze etwas. In diesem Datensatz sind entsprechend der Übersichtstabelle Umrissdaten von 40 Knochen von Karakulschafen und insgesamt 53 Knochen aus den beiden ostanatolischen Ausgrabungsstätten sowie lineare Maße von 40 Karakulschafen und 39 Wildschafen vorhanden, während Pöllath et. al. bei Verwendung der Knochenumrisse 41 Knochen beider Gruppen und für die Klassifikation anhand der linearen Maße 40 Wildschafe und 41 Karakulschafe zur Verfügung standen. Beim Vergleich der geschätzten Klassifikationsergebnisse der Methoden sollten diese Unterschiede in den verwendeten Ansichten und den vorhandenen Untersuchungseinheiten mit bedacht werden. Es stellt sich die Frage, wie man den Umriss von Objekten, der an sich eine durchgehende Linie ist und damit unendlich-dimensional ist, in endlicher Größe, also diskret, beschreiben kann. Dafür wurden für diesen Datensatz zwei zusammenhängende, aber doch etwas unterschiedliche Methoden angewandt. Das wohl traditionellere Verfahren ist, an den Objekten, also in diesem Fall den Talusknochen, *Landmarks* zu identifizieren. Landmarks sind über alle Objekte hinweg vergleichbare korrespondierende Punkte, das heißt sie sind an jedem Objekt vorhanden [8, S.3]. Wie in Abbildung 3 zu sehen ist, wurden dafür 11 Punkte (rot) ausgewählt, die entweder besondere biologische Bedeutung haben oder charakteristische mathematische Merkmale, beispielsweise Extremstellen, sind. Zudem wurde jedoch auch noch die in der dargestellten Ansicht obere (zwischen den Landmarks 1 und 3) und untere Kante (zwischen den Landmarks 6 und 8) durch äquidistante *Pseudo-Landmarks* (blaue Punkte) in feinerer „Auflösung“, genauer durch 14

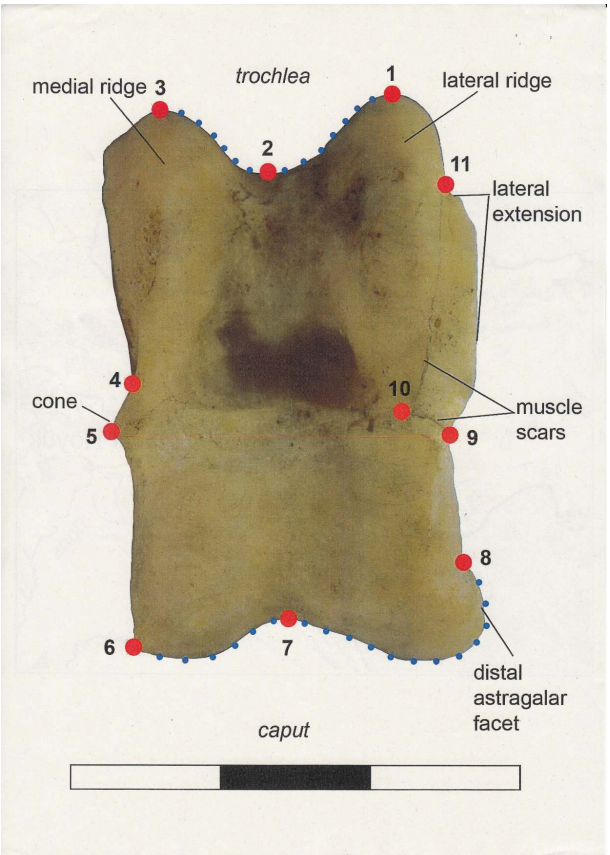

Abbildung 3: Landmarks (rot) und Pseudo-Landmarks (blau) an der Dorsalansicht eines Talusknochen. Aus [17].

(oben) beziehungsweise 18 (unten) Punkte, diskretisiert festgehalten. Pseudo-Landmarks müssen im Gegensatz zu normalen Landmarks nicht über verschiedene Objekte oder Populationen hinweg korrespondieren, da sie nur konstruierte Punkte an einem Objekt sind [8, S.4]. Hat man nun die Landmarks und Pseudo-landmarks ausgewählt und identifiziert, so kann man die x- und y-Werte der insgesamt 43 Punkte eines untergelegten Koordinatensystems als Kovariablen betrachten.

3.2 Der Begriff des *shape*

Für die Einführung der folgenden Begrifflichkeiten wollen wir uns auf die „echten“ Landmarks beschränken. Sie können jedoch mit gewisser Vorsicht auch auf die Pseudo-Landmarks beziehungsweise deren Ursprung, die unter diesen liegende zweidimensionale Kurve, angewandt werden. Da wir uns für die geometrische Form der Umrisse der Knochen interessieren, wollen wir nicht die einzelnen Landmarks oder gar die einzelnen Kovariablen separat analysieren, sondern die Menge aller Landmarks, genannt Konfiguration. Die Konfigurationsmatrix X ist die $k \times m$ Matrix der kartesischen Koordinaten der k Landmarks in m Dimensionen [8, S.33]. Da an den Talusknochen 11 zweidimensionale Landmarks identifiziert wurden, enthält der Datensatz 273 Konfigurationsmatrizen der Größe 11×2. Der Konfigurationsraum ist der Raum aller Konfigurationsmatrizen $\mathbb{M}^{k \times m}$ oder äquivalent $\mathbb{R}^{km}$, wobei eventuell einige spezielle Konfigurationen, etwa mit zwei identischen Landmarks, entfernt sind [8, S.33f.]. In diesen Konfigurationen steckt jedoch noch Information, die für die Analyse und den Vergleich von Umrissen nicht von Bedeu-

tung ist und dies sogar erschwert. Wir sind interessiert am *shape* eines Objektes, was nach der Definition von Dryden & Mardia [8, S.65] sämtliche gegenüber Lokation, Rotation und isotropischer Skalierung invariante geometrische Information über eine Konfigurationsmatrix X ist. Der shape einer Konfiguration X kann demnach dargestellt werden durch eine Menge von Elementen aus dem Konfigurationsraum, genauer

$$[X] = \{\beta XO + 1_k\delta^T : \beta \in \mathbb{R}^+, O \in SO(m), \delta \in \mathbb{R}^m\},$$

wobei 1_k ein k-dimensionaler Vektor, bei dem jeder Eintrag 1 ist, β ein beliebiger Skalierungsfaktor, O eine beliebige Rotationsmatrix und δ ein beliebiger Translationsvektor ist [8, S.62]. $SO(m)$ ist die Menge aller $m \times m$-dimensionaler Rotationsmatrizen O, welche die Eigenschaften $O^TO = I_m$ (Orthogonalität) und $det(O) = 1$ erfüllen. Für m = 2 kann die Rotationsmatrix durch einen einzelnen Winkel $\Theta, 0 \leq \Theta \leq 2\pi$ parametrisiert werden:

$$O = \begin{pmatrix} \cos\Theta & \sin\Theta \\ -\sin\Theta & \cos\Theta \end{pmatrix}.$$

Θ bezeichnet hierbei den Winkel in Radiant, um den die ursprüngliche Konfiguration X gegen den Uhrzeigersinn gedreht wird. Jedes Element der Menge $[X]$ besitzt den gleichen *shape* wie die Konfiguration X. In Abbildung 4 sind diese drei umrisserhaltende euklidischen Ähnlichkeitstransformationen am Beispiel der Konfiguration eines einzelnen Knochens dargestellt. Es ist offensichtlich, dass das Verschieben eines Objektes (A) keine Auswirkung auf dessen Umriss hat. Folglich sollte für unsere Interessen das ursprüngliche schwarze Objekt und das verschobene rote Objekt als identisch betrachtet werden. Dies ergibt Sinn, da es nicht von Interesse ist, wo im zweidimensionalen Koordinatensystem der Umriss eines Knochens festgehalten wurde. Dasselbe gilt auch bei Rotation (B) und Skalierung (C) des Objektes. Das Ignorieren beziehungsweise Entfernen der Größe eines Objektes mag weniger intuitiv als die anderen Operationen erscheinen, da man erwartet, dass in der Größe eines Objektes viel Information steckt. Die geometrische Form des Umrisses bleibt jedoch auch nach Skalierung erhalten. Zudem wird angeführt [17], dass die Annahme, dass die Größe beziehungsweise die Größenreduktion ein guter Marker für Domestizierung ist, in den letzten Jahren in Frage gestellt wurde. Einige Forscher raten zur Vorsicht, da das Größenprofil einer Population durch Faktoren wie den Klimawandel oder einen übermäßig großen Anteil an Jungtieren und weiblichen Tieren in der Stichprobe beeinflusst werden kann, während andere Forscher die Abnahme der Größe als verspäteten Effekt der Domestizierung sehen, wodurch die frühesten Stufen der Domestikation nicht durch größenbasierte Messungen festgehalten werden können [17]. Es ist also von großem Interesse, ob anhand des *shape* der Talusknochen eine Klassifizierung dieser nach dem Haltungsstatus möglich ist.

3.3 Deskriptive Darstellung der Knochenumrisse

Abschließend soll die Menge der Knochenumrisse und eventuelle Unterschiede zwischen den Klassen deskriptiv dargestellt werden. Um die Konfigurationen sinnvoll in Bezug auf ihre *shapes* vergleichen zu können, müssen diese einheitlich skaliert, verschoben und rotiert werden. Letz-

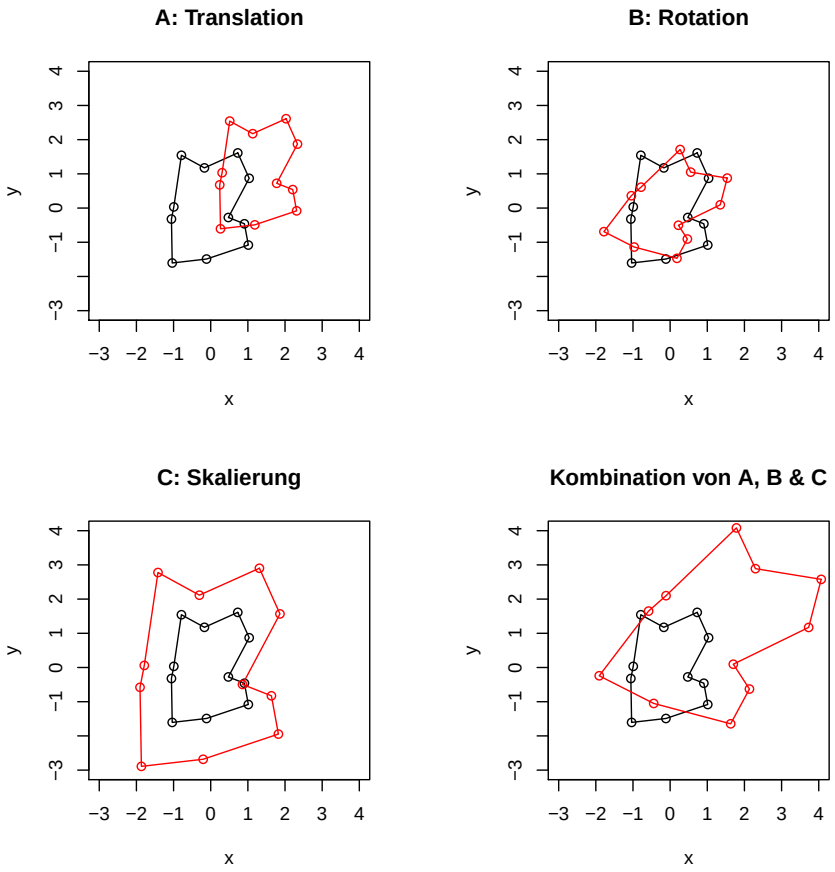

Abbildung 4: Umrisserhaltende Operationen

teres kann für eine erste rein deskriptive Analyse vernachlässigt werden, da die Knochen bei der Messung der Landmarks auf einheitliche Art und Weise ausgerichtet wurden. Eine Möglichkeit, die Translation und Skalierung einer Konfigurationsmatrix einheitlich zu entfernen, ist die Zentrierung und Standardisierung. Um eine Konfigurationsmatrix $X \in \mathbb{M}^{k \times m}$ zu zentrieren, verschieben wir diese so, dass ihr Schwerpunkt (auch: Zentroid) im Ursprung liegt. Der Schwerpunkt ist ein m-dimensionaler Vektor $(\bar{X}_{\cdot 1}\ \bar{X}_{\cdot 2}\ \ldots\ \bar{X}_{\cdot m})^T$ mit

$$\bar{X}_{\cdot j} = \frac{1}{k} \sum_{i=1}^{k} X_{ij},$$

also ein Vektor mit den jeweiligen arithmetischen Mittelwerten in den m Dimensionen. In der zentrierten Konfigurationsmatrix

$$X_Z = \begin{pmatrix} (X_{11} - \bar{X}_{\cdot 1}) & (X_{12} - \bar{X}_{\cdot 2}) & \ldots & (X_{1m} - \bar{X}_{\cdot m}) \\ (X_{21} - \bar{X}_{\cdot 1}) & (X_{22} - \bar{X}_{\cdot 2}) & \ldots & (X_{2m} - \bar{X}_{\cdot m}) \\ \ldots & \ldots & \ldots & \ldots \\ (X_{k1} - \bar{X}_{\cdot 1}) & (X_{k2} - \bar{X}_{\cdot 2}) & \ldots & (X_{km} - \bar{X}_{\cdot m}) \end{pmatrix}$$

ist demnach der Durchschnittswert in jeder Dimension (jeder Spalte) gleich null [27, S.77].
X_Z kann ebenfalls berechnet werden über $X_Z = CX$, wobei

$$C = I_k - \frac{1}{k} 1_k 1_k^T = \begin{pmatrix} \frac{k-1}{k} & -\frac{1}{k} & \dots & -\frac{1}{k} \\ -\frac{1}{k} & \frac{k-1}{k} & \dots & -\frac{1}{k} \\ \dots & \dots & \ddots & \dots \\ -\frac{1}{k} & -\frac{1}{k} & \dots & \frac{k-1}{k} \end{pmatrix}$$

die sogenannte Zentrierungsmatrix ist.
Das Standardisieren einer Konfiguration erfolgt durch Dividieren durch die Zentroidengröße

$$S(X) = ||X_Z|| = ||CX|| = \sqrt{\sum_{i=1}^{k} \sum_{j=1}^{m} (X_{ij} - X_{\cdot j})^2}$$

ein Skalar, der sich als Wurzel der Summe aller quadrierten Einträge von X_Z ergibt [8, S.34].
$||\cdot||$ bezeichnet dabei die Frobenius-Norm, die aus der euklidischen Norm abgeleitete Matrixnorm.
In Abbildung 5 sind die Landmarks aufgeteilt nach den verschiedenen Haltungsstatus deskriptiv dargestellt. Die grelleren Punkte repräsentieren dabei jeweils die klassenspezifischen Mittelwerte der 11 Landmarks, und die farbigen Ellipsen geben das 90%-Quantil der bivariaten Normalverteilung mit der aus den jeweiligen Landmarks geschätzten Varianz an. Sowohl in den Lagen der Mittelwerte der Landmarks als auch bei den Quantilen zeigen sich kaum Unterschiede. Die größten Differenzen lassen sich noch an den oberen Landmarks (in Abbildung 3 mit den Nummern 2 und 3) erkennen: Die Furche scheint bei den Knochen der späten Hausschafe etwas tiefer zu sein und das Landmark links davon bei diesen Schafen weiter außen zu liegen als bei den Wildschafen und den primitiveren Hausschafen.
In ähnlicher Manier können wir die an den Pseudo-Landmarks festgehaltenen Kurven betrachten. Da in vorhergehenden Ergebnissen in [17] die obere Kurve entlang der Sprungbeinrolle für aussagekräftiger als die untere Kurve zwischen dem sechsten und achten Landmark befunden wurde, wird diese deskriptive Analyse sowie die weiteren kurvenbezogenen Analysen und Methoden nur auf die obere Kurve angewandt. In Vorbereitung auf die deskriptive Darstellung wird die Lokation dieser oberen Kurven entfernt, indem der Schwerpunkt der drei an dieser Kurve liegenden Landmarks auf den Ursprung gesetzt wird. Auch die Zentroidengröße zur Skalierung der Kurven wird allein an diesen drei Punkten berechnet. Damit können wir die Kurven vereinheitlichen, ohne die nicht korrespondierenden Pseudo-Landmarks verwenden zu müssen.
Auch diese Kurven untersuchen wir wieder getrennt nach den drei Domestizierungsniveaus in Abbildung 6. Dabei repräsentieren die fettgedruckten Linien die interpolierte Mittelwertskurve der jeweiligen Klasse. Die eingefärbten Flächen entsprechen den interpolierten 10%- und 90%-Quantilen der $y - Werte$ der jeweiligen Pseudo-Landmarks. Hier wird der eben festgestellte Eindruck bestätigt: Die Mittelwertskurve der späten Hausschafe liegt in der Furche beinahe genauso tief wie die 10%-Quantile der anderen beiden Populationen. Zudem ist zu erkennen, dass die mittlere Kurve der Wildschafe im linken Teil niedriger als die der domestizierten Schafe liegt und ebenfalls fast deren 10%-Quantilskurve unterschreitet.

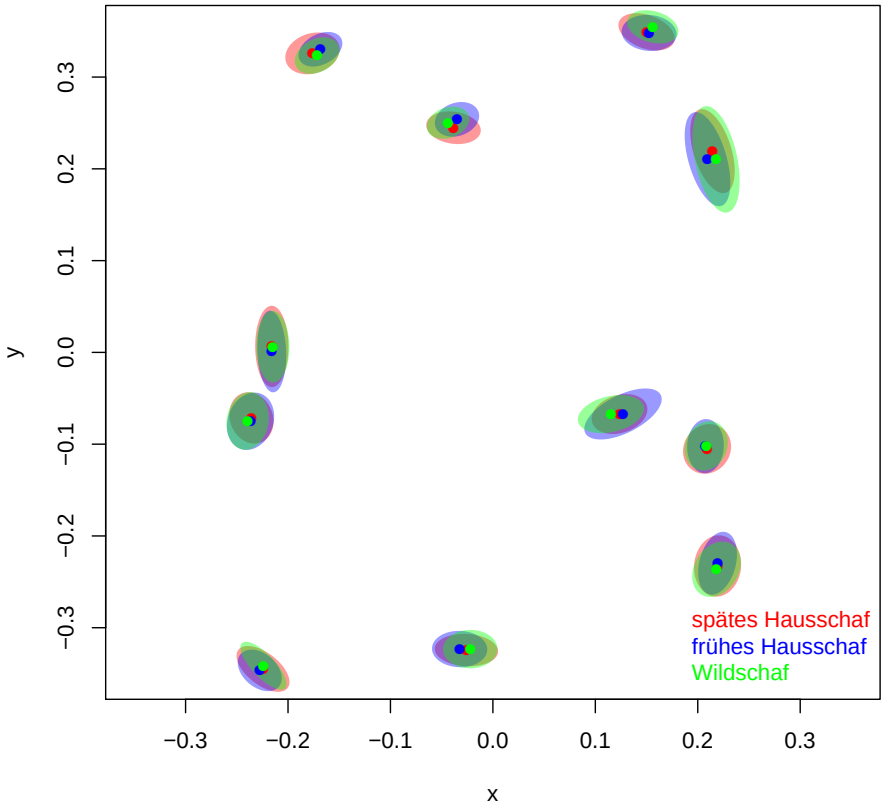

Abbildung 5: Deskriptive Darstellung der Landmarks. Die farbigen Punkte entsprechen dem Mittelwert der jeweiligen Landmarks der entsprechenden Klasse. Die Farbflächen markieren das 90%-Quantil der anhand den entsprechenden Landmark-Koordinaten geschätzten bivariaten Normalverteilung.

Da wir nun die besondere Struktur der Daten aufgezeigt sowie den Begriff des *shape* eingeführt haben, können wir uns im folgenden Kapitel mit der Frage beschäftigen, welche Distanzmaße geeignet sind, um Ähnlichkeiten beziehungsweise Unähnlichkeiten zwischen zwei Knochenumrissen festzustellen.

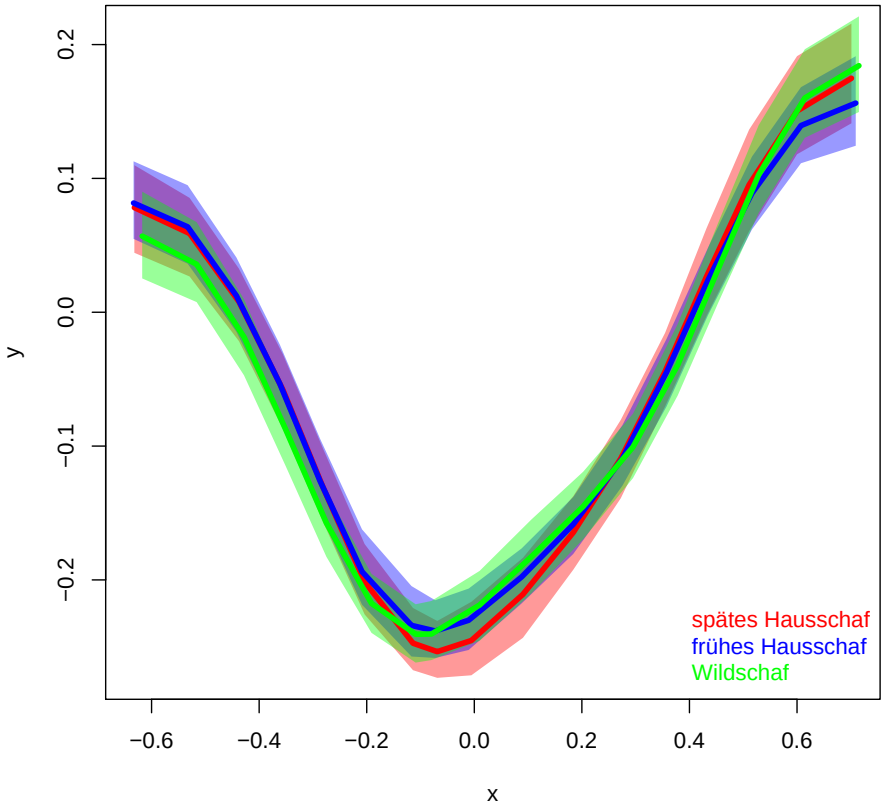

Abbildung 6: Deskriptive Darstellung der oberen Kurve. Die fetten Linien sind die interpolierten Mittelwerte der Landmarks und Pseudo-Landmarks entlang der Kurve. Die Farbflächen markieren den Bereich zwischen den entsprechenden 10%- und 90%- Quantilen.

4 Spezielle Abstandsmaße

4.1 Anforderungen

Einführend wollen wir gewöhnlicherweise von Abstandsmaßen geforderte Eigenschaften vorstellen. Demnach sollen Distanzfunktionen Metriken sein, also Abbildungen $d : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ (beispielsweise: $\mathcal{X} = \mathbb{R}^n$), die für alle $x, y, z \in \mathcal{X}$ folgende Eigenschaften erfüllen (vgl. [11, S.1f.]):

i) $d(x, y) = d(y, x)$	Symmetrie,
ii) $d(x, y) \leq d(x, z) + d(z, y)$	Dreiecksungleichung,
iii) $d(x, y) = 0 \Leftrightarrow x = y$	und (aus i)-iii) folgend)
iv) $d(x, y) \geq 0$	iii) & iv): Positive Definitheit.

Die dritte Eigenschaft besagt, dass der Abstand zweier Elemente aus $\mathcal{X}$ genau dann gleich null ist, wenn diese beiden Elemente identisch sind. Wir müssen jedoch feststellen, dass diese Eigenschaft zumindest für Distanzmaße in unserem Konfigurationsraum nicht unserem Ziel entspricht. Zwei verschiedene Konfigurationen sollen auch den Abstand null haben und damit in gewisser Weise als identisch betrachtet werden, wenn sie denselben *shape* besitzen. Daher sind als Abstandsmaß

auf dem Konfigurationsraum für unsere Interessen Semi-Metriken besser geeignet. Diese erfüllen anstatt der Eigenschaft iii) die abgeschwächte Form

$$\text{iii}^*) \; x = y \Rightarrow d(x, y) = 0.$$

Semi-Metriken erlauben also, dass auch der Abstand zwischen zwei nicht identischen Objekten null ist.

4.2 Procrustes-Distanzen

Pre-shape Raum Zuerst soll eine Methode zur Distanzmessung zwischen den *shapes* zweier Objekte, die jeweils durch korrespondierende Landmarks festgehalten sind, vorgestellt werden. Die Procrustes-Distanzen, wie in [8, S.69ff.] vorgestellt, sind hierfür bewährte Metriken zur Bestimmung des Abstandes zweier Umrisse im Raum der *shapes* beziehungsweise Semi-Metriken im Konfigurationsraum. Um deren Definition verständlich zu machen, wird zuerst der Begriff der *pre-shapes* eingeführt. Der *pre-shape* Z einer Konfigurationsmatrix X enthält keine Information mehr über deren Lokation und deren Größe. Eine Möglichkeit, den *pre-shape* einer Konfigurationsmatrix zu berechnen, ist mit den in Kapitel 3 definierten Größen durch

$$Z = \frac{X_Z}{S(X)} = \frac{CX}{||CX||}, Z \in \mathbb{M}^{k \times m}$$

gegeben, womit der *pre-shape* offensichtlich invariant gegenüber Translation und isotropischer Skalierung von X ist. Demnach ergibt sich aus zwei Konfigurationen, die sich nur in Lokation und/oder Zentroidengröße unterscheiden, der selbe *pre-shape*. Unterscheiden die Konfigurationen sich jedoch durch Rotation und/oder durch unterschiedliche *shapes*, so ist auch ihr *pre-shape* verschieden [27, S.80]. Der *pre-shape* Raum $S_m^k \equiv S^{km-m-1}$, der Raum aller zu Konfigurationsmatrizen aus $\mathbb{M}^{k \times m}$ gehörenden *pre-shapes*, ist demnach der Bahnenraum aller Elemente des Konfigurationsraum unter der Translations- und Rotationsoperation.
S_m^k besitzt die Dimension $km - m - 1$, da von den ursprünglichen km Einträgen in der Matrix Z m Einträge durch die Bedingungen $\bar{Z}_{\cdot j} = 0, j = 1, \ldots, m$ „gebunden“ werden und eine Dimension für $S(Z) = 1$ verwendet wird. Folglich ist der *pre-shape* Raum S_m^k eine Einheitssphäre in $\mathbb{R}^{km-m}$ [8, S.64]. Wie schon bemerkt, repräsentiert ein Punkt auf dieser Oberfläche eine Menge von Konfigurationen, die dieselbe Rotation und denselben *shape* haben, sich in Lokation und/oder Skalierung jedoch unterscheiden. Die Menge aller Punkte im *pre-shape* Raum, die durch Rotation eines einzelnen *pre-shape* Z „erreicht“ werden können, bilden eine Faser (engl. *fiber*) von S_m^k [27, S.81].
Wie man an der Repräsentation eines *shape* durch die Menge von Konfigurationsmatrizen

$$[X] = \{ZO : O \in SO(m)\}, \; Z \coloneqq pre - shape \text{ von } X$$

erkennt, entspricht eine Faser des *pre-shape* Raums genau einem *shape* (vergleiche Abbildung 7). Der Raum der *shapes* Σ_m^k ist demnach der Bahnenraum der Rotationsoperation auf dem *pre-shape* Raum S_m^k.

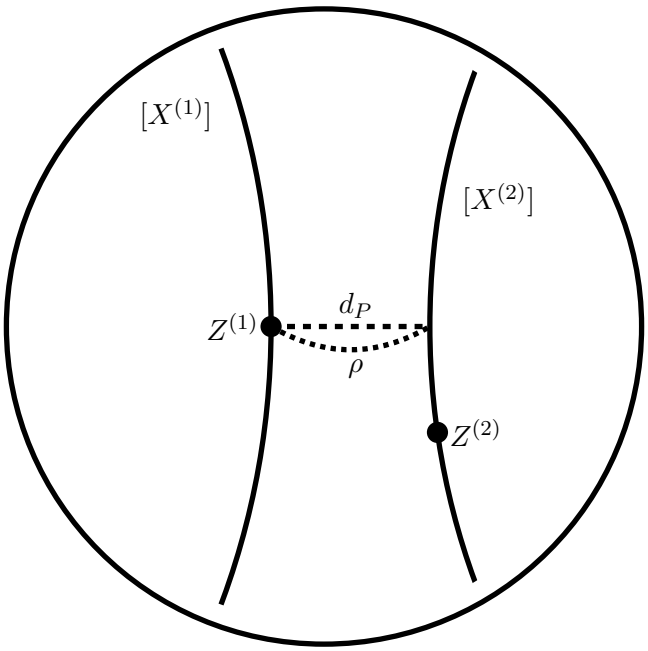

Abbildung 7: Schematische Darstellung zweier Fasern $[X^{(1)}]$ und $[X^{(2)}]$ des *pre-shape* Raums, welche den *shapes* der Konfigurationen $X^{(1)}$ und $X^{(2)}$ entsprechen. Deren *pre-shapes* $Z^{(1)}$ und $Z^{(2)}$ sind Punkte im *pre-shape* Raum, die auf den jeweiligen Fasern liegen. ρ bezeichnet die Länge der kürzesten Orthodrome zwischen den Fasern, d_P die Länge der kürzesten Sehne. Nach [8, S.73].

Procrustes-Distanzmaße Aufgrund dieser Äquivalenz von Fasern und *shapes* empfiehlt sich als Maß der Distanz zwischen zwei *shapes* die Länge der kürzesten Strecke zwischen den entsprechenden Fasern im *pre-shape* Raum. Da dieser eine Sphäre ist, ist die gesuchte Strecke eine Orthodrome, also ein Teil eines Großkreises auf der Sphäre. Die Länge dieser Strecke wird als Riemannsche Distanz ρ bezeichnet. Da der *pre-shape* Raum S_m^k eine Einheitskugel darstellt, entspricht die Riemannsche Distanz ρ, welche die Bogenlänge zwischen den beiden Punkten auf den Fasern der Kugel beschreibt, an denen diese sich am nächsten sind, dem Winkel zwischen den beiden Radien, die den Nullpunkt mit den erwähnten Punkten verbinden [27, S.83]. Diesen Winkel können wir anhand des Skalarprodukts zwischen den Richtungsvektoren der beiden Punkte berechnen. Als Formel für die Riemannsche Distanz zwischen zwei Konfigurationen $X^{(1)}, X^{(2)}$ mit *pre-shapes* $Z^{(1)}, Z^{(2)}$ ergibt sich damit:

$$\rho(X^{(1)}, X^{(2)}) = \inf_{O \in SO(m)} \cos^{-1}(\langle Z^{(1)}, Z^{(2)}O \rangle).$$

Der kürzeste Bogen auf der Oberfläche der Sphäre ist jedoch nicht die kürzeste Entfernung zweier Punkte, die auf dieser Oberfläche liegen. Die Länge der kürzesten direkten Verbindung der Fasern, also der kürzesten Sehne durch das Innere der Kugel, ist die partielle Procrustes-Distanz d_P. Die partielle Procrustes-Distanz zwischen den *shapes* zweier Konfigurationen $X^{(1)}, X^{(2)}$ entspricht dem geringsten Abstand der beliebig rotierten *pre-shapes* $Z^{(1)}, Z^{(2)}$ und kann dementsprechend über

$$d_P(X^{(1)}, X^{(2)}) = \inf_{O \in SO(m)} ||Z^{(2)} - Z^{(1)}O||$$

berechnet werden. Dies ist zwar der geringste Abstand zweier Fasern im *pre-shape* Raum, doch nicht der geringste Abstand aller Konfigurationen mit den entsprechenden, zu vergleichenden *shapes*. Um diese volle Procrustes-Distanz d_F zu bestimmen, optimieren wir zusätzlich über mögliche Skalierungen einer der beiden *pre-shapes*. Demnach wird die volle Procrustes-Distanz zwischen den *shapes* zweier Konfigurationen $X^{(1)}, X^{(2)}$ über

$$d_P(X^{(1)}, X^{(2)}) = \inf_{O \in SO(m), \beta \in \mathbb{R}^+} ||Z^{(2)} - \beta Z^{(1)} O||$$

berechnet. Dies bedeutet, dass für eine der beiden Konfigurationen die Einschränkung $S(Z^{(2)}) = 1$ aufgehoben wird und sie damit, um die Distanz zu minimieren, entlang dem Radius, der den entsprechenden (rotierten) *pre-shape* $Z^{(2)}$ und den Ursprung der Einheitskugel verbindet, „auf und ab wandern“ kann. An dem Punkt, an dem diese Distanz minimal ist, steht dieser Radius senkrecht zur Verbindungslinie des Punktes mit dem Referenz-*pre-shape* $Z^{(1)}$ auf der Sphäre [27, S.85].
Um die Distanzen und Zusammenhänge zwischen ihnen besser zu verstehen, ist in Abbildung 8 ein Teil eines Querschnitts der Einheitskugel, deren Oberfläche der *pre-shape* Raum ist, skizziert. Aufgrund der in der Skizze eingezeichneten rechtwinkligen Dreiecke ergeben sich $d_P = 2\sin(\rho/2)$

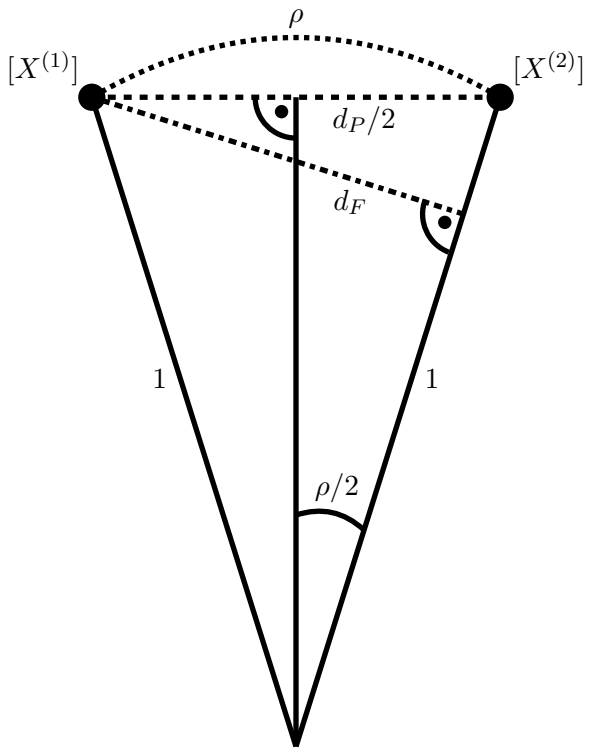

Abbildung 8: Querschnitt der Einheitskugel, deren Oberfläche der *pre-shape* Raum ist. ρ, d_P und d_F bezeichnen die Riemannsche Distanz, die partielle Procrustes-Distanz und die volle Procrustes-Distanz zwischen den Konfigurationen $X^{(1)}$ und $X^{(1)}$. Nach [8, S.78].

und $d_F = \sin\rho$. Es lässt sich zeigen, dass $0 \leq \rho \leq \pi/2$ gilt (vgl. [8, S.75f.]) und daraus folgend $0 \leq d_P \leq \sqrt{2}$ sowie $0 \leq d_F \leq 1$. Dryden & Mardia führen zudem an, dass für nah beieinander liegende *shapes* der Unterschied zwischen den verschiedenen Distanzen nur sehr gering ist, und es folglich bei der Analyse vieler Datensätze kaum Auswirkungen hat, welche der Procrustes-

Distanzen man verwendet [8, S.77]. Basierend auf den deskriptiven Betrachtungen in Abschnitt 3.3 können wir annehmen, dass dies auch auf die vorliegenden Daten zutrifft. Eine Umsetzung der Procrustes-Distanzen in der Programmiersprache R [18] ist in dem R-Paket **shapes** [9] als Funktion `procdist` vorhanden.

Abschließend soll durch simulierte Daten eine Situation dargestellt werden, in der die Verwendung eines Procrustes-Distanzmaß zu sehr guter Klassifikation führt. Dabei werden für jede von zwei Klassen Objekte generiert, indem die einzelnen Koordinatenpunkte einer klasseneigenen „Grundkonfiguration“ durch einen zufälligen Fehler verzerrt werden und diese neue Konfiguration zufällig gestreckt, verschoben und rotiert wird. In der linken Grafik in Abbildung 9 sind drei Objekte beider Klassen, markiert durch die jeweiligen Farben Rot und Blau, zu sehen, anhand welcher die beiden mit schwarzen Linien gezeichneten Objekte (eines von jeder Klasse) klassifiziert werden sollen. Beim Betrachten dieser Grafik scheint es unklar, wie das Klassifikati-

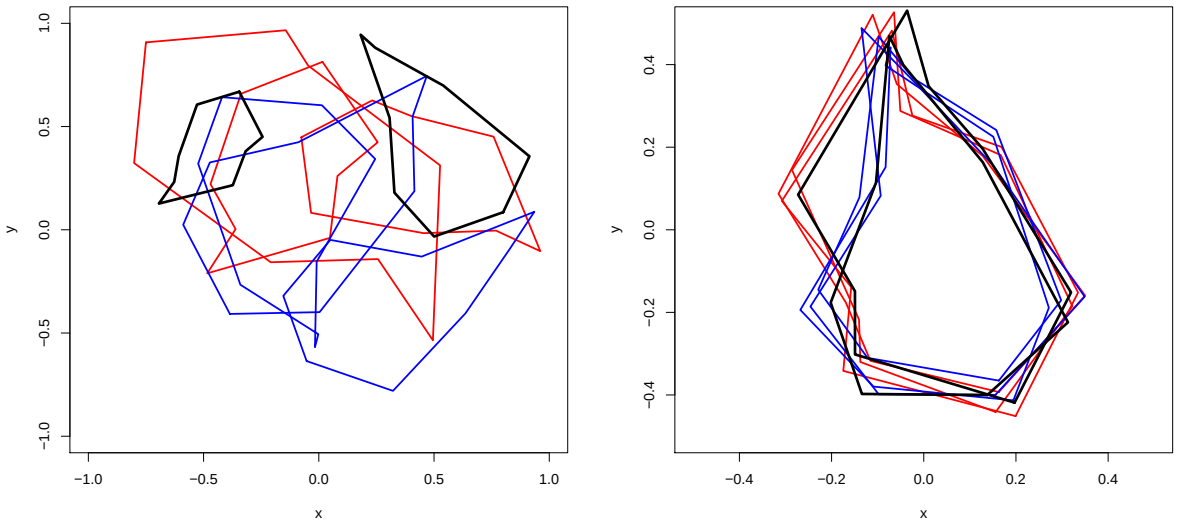

Abbildung 9: Simulation einer Situation, in der die Verwendung einer Procrustes-Distanz zu einer erfolgreichen Klassifikation führt. Im linken Panel ist die Ausgangssituation mit je drei Objekten zweier Klassen (Rot und Blau) sowie zwei zu klassifizierenden Objekten (Schwarz) zu sehen. In der rechten Grafik sind die Objekte einheitlich skaliert, verschoben und klassenweise durch Rotation angeglichen.

onsergebnis bei Verwendung klassischer multivariater Abstandsmaße ausfallen würde. Im rechten Panel sehen wir jedoch, dass nach Entfernung von Skalierung, Translation und Rotation die Objekte der jeweiligen Klassen sich stark ähneln und die Prädiktionen eindeutig scheinen, was durch den entsprechenden Ausschnitt der Distanzmatrix (Tabelle 2) der Objekte bei Verwendung der vollen Procrustes-Distanz d_F bestätigt wird. Die Distanz des ersten zu klassifizierenden Objekts zu den drei roten Objekten ist geringer als die Distanz zu jedem Objekt der blauen Klasse. Zum zweiten Testdatum hingegen sind die drei blauen Konfigurationen näher als die drei roten.

	Testdatum 1	Testdatum 2
Rot 1	0.101	0.275
Rot 2	0.093	0.222
Rot 3	0.129	0.246
Blau 1	0.197	0.131
Blau 2	0.287	0.131
Blau 3	0.250	0.106

Tabelle 2: Ausschnitt der Distanzmatrix bei Verwendung der vollen Procrustes-Distanz und simulierten Konfigurationen.

4.3 Elastische Distanz basierend auf der SRV Repräsentation

4.3.1 Theorie

Distanzmaß für eindimensionale Funktionen Im letzten Abschnitt haben wir typische Distanzmaße zum Vergleich der *shapes* von Konfigurationen korrespondierender Landmarks vorgestellt. Wie in Kapitel 3 erwähnt, enthält der Datensatz jedoch auch nicht notwendigerweise korrespondierende Pseudo-Landmarks, die eine diskrete Approximation einer Kurve $f : D \to \mathbb{R}^m$, $D = [D_1, D_2]$ darstellen. In diesem Abschnitt soll eine Methode zur Messung des Abstands zwischen den *shapes* zweier Kurven präsentiert werden, genauer die von Srivastava et. al. [23] vorgestellte elastische (Semi-)Metrik auf Basis der *square-root velocity* (SRV) Darstellung der Kurven. Um dieses Konzept einzuführen, betrachten wir zuerst das Pendant dieser Metrik für eindimensionale Funktionen, also Abbildungen $f : D \to \mathbb{R}$ (typisch: $D = [0, 1]$).
Der Term „elastisch“ stammt daher, dass wir Strecken und Stauchen der Funktion entlang der horizontalen Achse erlauben. Diese Reparametrisierung (auch: *warping*) der Funktion f kann ausgedrückt werden als Verknüpfung $f \circ \gamma, \gamma \in \Gamma$ mit

$$\Gamma = \{\gamma : D \to D |\ \gamma(D_1) = D_1, \gamma(D_2) = D_2, \gamma \text{ ist ein Diffeomorphismus}\}.$$

Ein häufig verwendetes Vorgehen ist die Definition des elastischen Abstandes als

$$\inf_{\gamma \in \Gamma} ||f_1 - (f_2 \circ \gamma)||,$$

wobei $||f|| = (\int_D f(t)^2 dt)^{1/2}$ die $\mathbb{L}^2$-Norm bezeichnet, was dem Minimieren des Abstandes der Funktionen durch Reparametrisierung von f_2 entspricht. Diese Methodik ist jedoch problematisch, da sie nicht symmetrisch ist. Das Ergebnis der vorgestellten Formel unterscheidet sich zum Minimum des Abstandes unter Reparametrisierung von f_1[24, S.4]. Stattdessen schlagen Srivastava et. al. [24, S.4f.] eine Herangehensweise vor, bei der die Metrik nicht für Elemente aus dem Funktionenraum F, sondern für Elemente aus dem den Bahnenraum F/Γ definiert ist. Wir wollen also Äquivalenzklassen vergleichen, wobei Funktionen als äquivalent betrachtet werden, wenn sie durch Reparametrisierung perfekt aneinander angeglichen werden können.
Das Abstandsmaß soll invariant gegenüber identischer Reparametrisierung der Funktionen sein, also $d(f_1, f_2) = d(f_1 \circ \gamma, f_2 \circ \gamma)$ für alle $f_1, f_2 \in F$ und $\gamma \in \Gamma$ erfüllen, wie etwa die Fisher-Rao Riemannsche Metrik d_{FR}, welche in dieser Arbeit nicht genauer beschrieben werden soll. Für

eine genaueren Einblick auf dieses Distanzmaß sei hier auf [24, S.6] verwiesen. Die Berechnung von d_{FR} kann jedoch durch die Repräsentation von f durch die *square-root velocity function* (SRVF)

$$q : D \to \mathbb{R}$$
$$q(t) = Q(f'(t)) = f'(t)/\sqrt{|f'(t)|}$$

vereinfacht werden, wobei

$$Q : \mathbb{R} \to \mathbb{R}$$
$$Q(x) = \begin{cases} \frac{x}{\sqrt{|x|}} & \text{wenn } |x| \neq 0 \\ 0 & \text{sonst} \end{cases}$$

eine stetige Abbildung ist [24, S.5]. Es kann gezeigt werden, dass unter der Bedingung, dass f absolut stetig ist, die entsprechende SRVF quadratintegrierbar ist und damit $\mathbb{L}^2(D, \mathbb{R})$ (kurz $\mathbb{L}^2$) die Menge aller SRVFs ist [24, S.5]. Ein Hauptgrund für die Verwendung der SRVF-Transformation ist, dass die Fisher-Rao Riemannsche Metrik d_{FR} zweier Funktionen der $\mathbb{L}^2$-Distanz ihrer SRVFs entspricht [16, S.12]:

$$d_{FR}(f_1, f_2) = ||q_1 - q_2||.$$

Folglich ist die Operation der Reparametrisierung isometrisch unter der $\mathbb{L}^2$-Distanz der SRVFs. Es gilt also für alle SRVFs $q_1, q_2 \in \mathbb{L}^2$ und $\gamma \in \Gamma$ $||q_1 - q_2|| = ||(q_1, \gamma) - (q_2, \gamma)||$, wobei (q, γ) die SRVF von $f \circ \gamma$ darstellt:

$$\begin{aligned}(q, \gamma) &= Q((f \circ \gamma)'(t)) = Q(f'(\gamma(t))\gamma'(t)) = \\ &= \frac{f'(\gamma(t))\gamma'(t)}{\sqrt{|f'(\gamma(t))\gamma'(t)|}} = \frac{f'(\gamma(t))}{\sqrt{|f'(\gamma(t))|}}\frac{\gamma'(t)}{\sqrt{\gamma'(t)}} = q(\gamma(t))\sqrt{\gamma'(t)}.\end{aligned}$$

Die Bahn $[q] = \{(q, \gamma), \gamma \in \Gamma\}$ einer SRVF $q \in \mathbb{L}^2$ enthält alle SRVFs, die durch Reparametrisierung der ursprünglichen Funktion f gebildet werden können, und der Bahnenraum Σ ist die Menge aller Bahnen. Für zwei Funktionen $f_1, f_2 \in F$ und ihren korrespondierenden SRVFs $q_1, q_2 \in \mathbb{L}^2$ ist die elastische Distanz d im Bahnenraum Σ definiert als die unter Reparametrisierung minimale Distanz im SRVF-Raum $\mathbb{L}^2$:

$$d([q_1], [q_2]) = \inf_{\gamma \in \Gamma} ||q_1 - (q_2, \gamma)|| = \inf_{\gamma \in \Gamma} ||(q_1, \gamma) - q_2||.$$

Man sieht, dass diese Distanz für zwei SRVFs aus der gleichen Bahn $q_1, q_2 \in [q]$ null beträgt und damit im $\mathbb{L}^2$-Raum nur eine Semi-Metrik ist. Es kann jedoch gezeigt werden, dass die Distanz für zwei SRVFs aus unterschiedlichen Bahnen positiv ist und d damit auf dem Bahnenraum Σ eine „echte“ Metrik ist [24, S.7].

Auch für dieses Distanzmaß können wir die Funktionsweise anhand der Klassifikation simulierter Daten aufzeigen. Ähnlich wie im vorangegangen Abschnitt kreieren wir für zwei Klassen

Funktionen, indem wir eine klasseneigene glatte Ausgangsfunktion „verschmutzen“, sie zufällig vertikal verschieben und reparametrisieren. Im linken Panel der Abbildung 10 sind jeweils drei Funktionen der durch die Farben Rot und Blau gekennzeichneten Klassen eingezeichnet sowie zwei zu klassifizierende Funktionen. In der rechten Grafik sind die Funktionen klassenintern an-

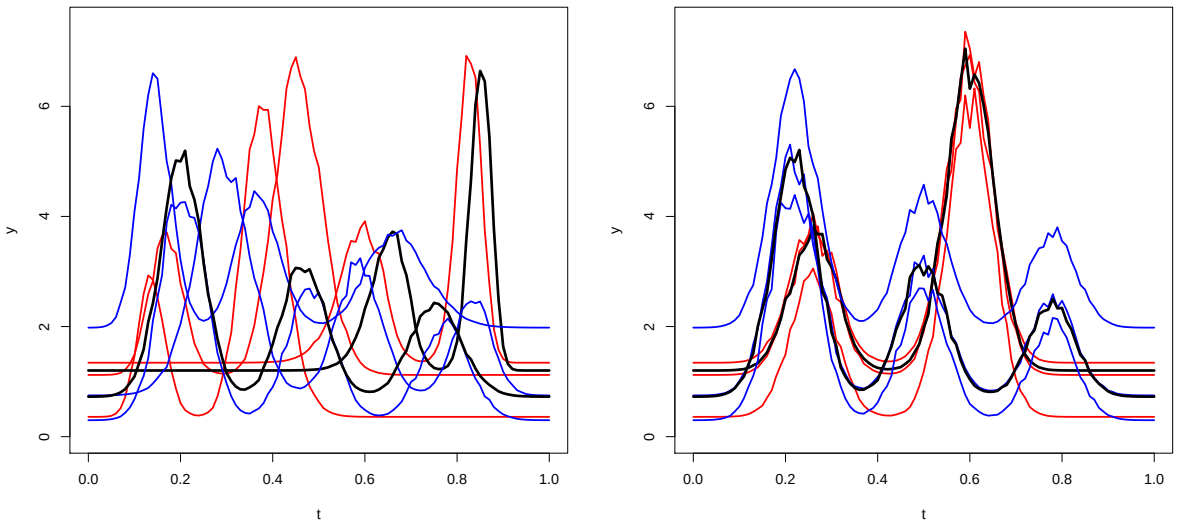

Abbildung 10: Simulation einer Situation, in der die Verwendung einer funktionalen elastischen Distanz zu einer erfolgreichen Klassifikation führt. Im linken Panel ist die Ausgangssituation mit je drei Funktionen zweier Klassen (Rot und Blau) sowie zwei zu klassifizierenden Objekten (Schwarz) zu sehen. In der rechten Grafik sind die Funktionen klassenweise durch Reparametrisierung angeglichen.

einander durch Reparametrisierung angepasst. Abermals hilft eine bestimme Operation dabei, Daten erfolgreich zu klassifizieren, was durch einen Blick auf die entsprechende Distanzmatrix in Tabelle 3 bestätigt wird. Die drei nächsten Nachbarn der ersten zu klassifizierenden Funktion

	Testdatum 1	**Testdatum 2**
Rot 1	0.365	2.200
Rot 2	0.664	2.273
Rot 3	0.460	2.289
Blau 1	2.137	0.511
Blau 2	2.428	0.461
Blau 3	2.185	0.548

Tabelle 3: Ausschnitt der Distanzmatrix bei Verwendung der elastischen Distanz und simulierten Funktionen.

gehören allesamt der roten Klasse an, bei dem zweiten Testdatum stellt sich die Situation genau anders herum dar.

Distanzmaß für Kurven In diesem Abschnitt soll gezeigt werden, wie mit Hilfe der SRV-Darstellung von Kurven eine Metrik, die invariant gegenüber den euklidischen Ähnlichkeitstransformationen Translation, Skalierung und Rotationen sowie der Reparametrisierung ist, definiert

werden kann. Ähnlich wie im eindimensionalen Fall repräsentieren wir eine Kurve $f : D \to \mathbb{R}^m$ (generell: $D = [0, 1]$, sowie bei unseren Daten: $m = 2$) durch ihre SRV-Darstellung

$$q : D \to \mathbb{R}^m$$
$$q(t) = Q(f'(t)) = f'(t) / \sqrt{||f'(t)||_{Eukl}},$$

wobei

$$Q : \mathbb{R}^m \to \mathbb{R}^m$$
$$Q(x) = \begin{cases} \frac{x}{\sqrt{||x||_{Eukl}}} & \text{wenn } ||x||_{Eukl} \neq 0 \\ 0 & \text{sonst} \end{cases}$$

eine stetige Abbildung ist und $|| \cdot ||_{Eukl}$ die euklidische Norm in $\mathbb{R}^m$ kennzeichnet [23, S.2].
Da $q \in \mathbb{L}^2(D, \mathbb{R}^m)$ nur von der ersten Ableitung von f abhängt, ist die SRV-Darstellung unabhängig von möglicher Verschiebung von f. Um Skalierung zu entfernen, skalieren wir sämtliche Kurven auf Einheitslänge, was der Standardisierung der SRVFs $||q|| = (\int_D ||q(t)||^2_{Eukl} dt)^{1/2} = 1$ entspricht [23, S.2]. Demnach liegen die SRV-Repräsentationen von auf Einheitslänge skalierten Kurven allesamt auf der Einheitssphäre S in $\mathbb{L}^2(D, \mathbb{R}^m)$. Da ein Element von S sämtliche Kurven repräsentiert, die den gleichen *shape* haben und gleich rotiert und parametrisiert sind, sich aber in Lokation und Skalierung unterscheiden können, kann S als *pre-shape* Raum gesehen werden. Die Distanz d_S zwischen zwei Elementen dieser Sphäre $q_1, q_2 \in S$ kann berechnet werden als die minimale Länge eines Pfades $\alpha : [0, 1] \to S$ mit $\alpha(0) = q_1$ und $\alpha(1) = q_2$. Die Länge eines Pfades ist definiert als $L(\alpha) = \int_0^1 \langle \alpha'(\tau), \alpha'(\tau) \rangle^{1/2} d\tau$, und damit ist die Distanz in S

$$d_S(q_1, q_2) = \inf_{\alpha:[0,1]\to S | \alpha(0)=q_1, \alpha(1)=q_2} L(\alpha).$$

Einer ähnlichen Vorgehensweise wie in 4.2 folgend, bilden wir den Bahnenraum Σ der Rotations- und Reparametrisierungsoperation auf S, wobei die Bahn einer einzelnen SRVF $q \in S$ als

$$[q] = \{O^T (q \circ \gamma) \sqrt{\gamma'} | (\gamma, O) \in \Gamma \times SO(m)\}$$

gegeben ist, wobei sowohl Γ als auch $SO(m)$ exakt den in diesem Kapitel definierten Mengen entsprechen. SRV-Darstellungen, die Elemente der selben Bahn sind, besitzen denselben *shape*, aber können beliebig verschoben, skaliert, gedreht und reparametrisiert sein. Die Distanz der *shapes* zweier Funktionen $f_1, f_2 \in F$ mit SRV-Repräsentationen $q_1, q_2 \in \mathbb{L}^2(D, \mathbb{R}^m)$ ist gegeben durch

$$d_{elast}(f_1, f_2) = d_\Sigma([q_1], [q_2]) = \inf_{(\gamma,O)\in\Gamma\times SO(m)} d_S(q_1, O^T (q_2 \circ \gamma) \sqrt{\gamma'}.$$

Nach Srivastava et. al. ist diese Distanz positiv, wenn q_1 und q_2 nicht auf derselben Bahn liegen. Demnach ist die Distanz im Bahnenraum Σ eine Metrik [23, S.4].
Abschließend soll der Effekt der Reparametrisierung auf zweidimensionalen Kurven veranschau-

licht werden. Wie bei den eindimensionalen Funktionen werden durch die Reparametrisierung die Funktionswerte, also in diesem Fall zweidimensionale Vektoren $(x\ y)^T$, nicht verändert, sondern nur die Zuordnung $t \mapsto (x\ y)^T$. Folglich ist die Reparametrisierung bei glatten Kurven in der üblichen Darstellung mit einer Dimension x als horizontale Achse und der anderen Dimension y als vertikale Achse nicht zu erkennen. Betrachtet man die Kurven aber in einer diskretisierten Form, also beispielsweise die Pseudo-Landmarks in dem vorliegenden Datensatz, so kann man den Effekt nachvollziehen. Vergleicht man wie in Abbildung 11 links zwei unterschiedliche diskretisierte Darstellungen der exakt gleichen Kurve, so sollte der Abstand zwischen ihnen Null betragen. Mit der Reparametrisierung ist es möglich, die diskreten Punkte einer Kurve entlang

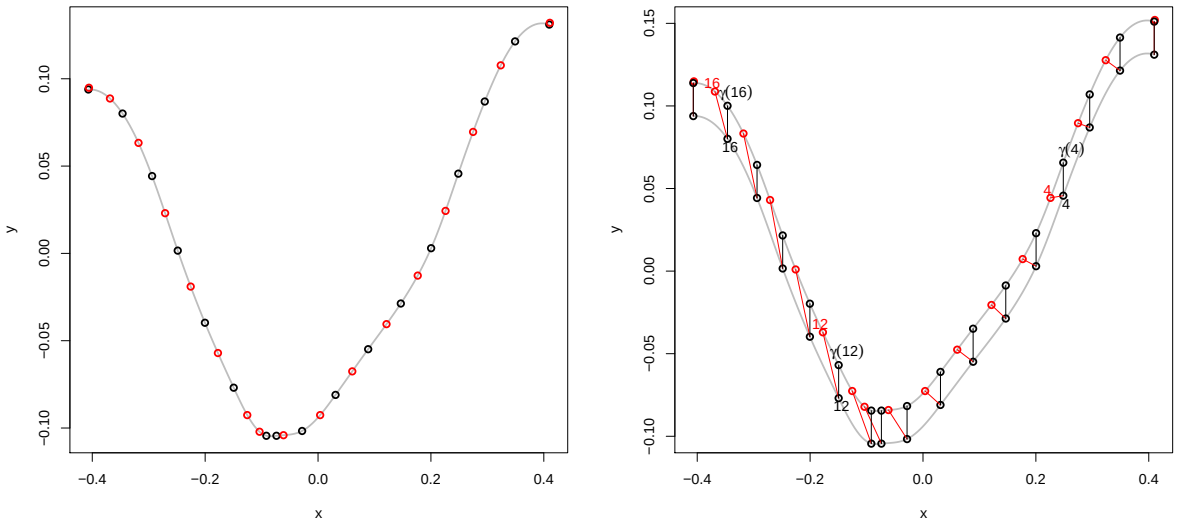

Abbildung 11: Darstellung der Reparametrisierung auf Kurven. Links: Aufgrund der Reparametrisierung können die diskreten Pseudo-Landmarks entlang der Kurve verschoben werden. Damit ist es bei der Verwendung der elastischen Distanz zumindest in der Theorie egal, ob die graue Kurve durch die roten oder die schwarzen Punkte festgehalten wird. Rechts: Verdeutlichung des Angleichens der beiden diskretisierten Kurven durch Reparametrisierung. Zur besseren Darstellung ist dazu die anzugleichende Kurve (rote Punkte) um 0.02 Einheiten nach oben verschoben worden.

dieser Kurve zu verschieben, um sie so gut wie möglich (in diesem Fall exakt) aneinander anzupassen. Dies ist im rechten Panel in Abbildung 11 noch einmal verdeutlicht dargestellt. Dazu wird die anzupassende Kurve, die durch die roten Punkte diskretisiert ist, um 0,02 Einheiten entlang der y-Achse verschoben. Im Umkehrschluss bedeutet dies, dass zwei Punkte auf den beiden Kurven mit einem vertikalen Abstand von 0,02 an sich identisch sind. Man erkennt, dass durch die Reparametrisierungsfunktion γ die Ausprägungen der Kurve an gleichen Trägerwerten t genau übereinander geschoben werden, so dass also zum Beispiel $f_{rot}(\gamma(4)) = f_{schwarz}(4)$ und allgemein $f_{rot}(\gamma(t)) = f_{schwarz}(t)\ \forall t \in D$ gilt.

4.3.2 Probleme in der Implementierung

Eine Implementierung der elastischen SRV-Distanz für Kurven ist in dem R-Paket **fdasrvf** [25] gegeben. Wir müssen jedoch feststellen, dass die numerische Approximation in der entspre-

chenden Funktion `calc_shape_dist` zumindest für die Anwendung auf den vorliegenden Daten der Knochenumrisse zu ungenau ist, um verwertbare Ergebnisse zu liefern. Dies lässt sich an folgenden Kennzahlen aufzeigen. In der Diagonale der entsprechenden (273×273) Distanzmatrix, in der der Eintrag $d_{ij}, i = 1, \ldots, 273,\ j = 1, \ldots, 273$ die elastische Distanz der Pseudo-Landmark-Konfiguration der oberen Kurve des i-ten Knochens zur entsprechenden Konfiguration der oberen Kurve des j-ten Knochens ist, erwartet man theoretisch nach der Definition der Semi-Metriken nur die Werte 0. Dies ist bei Verwendung der Funktion `calc_shape_dist` nicht der Fall, was verschmerzbar wäre, wenn die Distanzen auf der Diagonale im Vergleich zu den restlichen Werten der Distanzmatrix zumindest verhältnismäßig klein wären. Doch auch das trifft nicht zu. Betrachtet man die in jeder Spalte $j, j = 1, \ldots, 273$ zugeordneten Rangwerte der Zellen $d_{ij}, i = 1, \ldots, 273$, zum Beispiel mit der R-Funktion `apply(Distanzmatrix, 2, rank)`, so sollten sämtliche Diagonalenwerte der resultierenden Matrix 1 sein. Wendet man dieses Verfahren jedoch auf die auf der Funktion `calc_shape_dist` basierende Distanzmatrix an, ergibt sich als Mittelwert der Ränge der Diagonalen 119.8. Dies bedeutet, dass durchschnittlich der berechnete Abstand einer Pseudo-Landmark-Konfiguration zu sich selbst größer ist als zu 118 anderen Konfigurationen (von 272).

Ein weiteres Problem ist, dass die aus der Anwendung von `calc_shape_dist` auf die Daten der diskreten oberen Kurve resultierende Distanzmatrix nicht symmetrisch ist, das heißt $d_{ij} \neq d_{ji}$ zumindest für die meisten Datenpaare gilt. Auch hier wären verhältnismäßig geringe Unterschiede, obwohl der Definition der (Semi-)Metriken widersprechend, vertretbar. Dies ist jedoch nicht der Fall, wie Abbildung 12 zeigt. Der linke Boxplot stellt die $\frac{273*272}{2} = 37128$ absoluten Differenzen $|d_{ij} - d_{ji}|, i = 1, \ldots, 273, j > i$ dar. Als Vergleich dazu sind in dem mittleren und linken Boxplot die 37128 Werte $|d_{ik} - d_{jk}|, i = 1, \ldots, 273, j > i$ für die beiden nach Belieben ausgewählten Distanzmatrixspalten $k = 1$ und $k = 200$ zusammengefasst. Man erkennt, dass sich die Schwankungen zwischen den theoretisch identischen Zellen der Distanzmatrix in der gleichen Größenordnung bewegen wie die Unterschiede zwischen zwei nicht verwandten Distanzen.

Es stellt sich noch die Frage, ob und wie stark diese Schwächen der `calc_shape_dist` Funktion auch auf anderen Daten zu beobachten sind. Dazu betrachten wir den im Paket **fdasrvf** enthaltenen Datensatz `MPEG7`, der jeweils 20 ähnliche Kurven aus 65 Klassen enthält. Wir beschränken uns für die folgende Analyse auf die 20 Kurven der ersten enthaltene Klasse, die jeweils durch 100 zweidimensionale Punkte diskretisiert sind. Da ein Grund für die beobachteten Ungenauigkeiten der Funktion auf den Knochenumrissdaten die vergleichsweise geringe Anzahl an Datenpunkten entlang der oberen Kurve sein könnte, dünnen wir die 20 Kurven schrittweise aus. Wir betrachten analog zu den Untersuchungen auf dem Schafknochendatensatz die durchschnittliche Rangposition der Kurve i nach aufsteigender Anordnung der Kurven entsprechend ihrer Distanz zu ebenjener i-ten Kurve (Diag. Rang). Zudem berechnen wir den Mittelwert der absoluten Distanzdifferenzen $|d_{ij} - d_{ji}|, i = 1, \ldots, 20, j > i$ (MW Sym.) und die Mittelwerte von $|d_{ik} - d_{jk}|, i = 1, \ldots, 20, j > i$ für $k = 1$ (MW 1) und $k = 20$ (MW 20). Die Ergebnisse sind in Tabelle 4 zusammengefasst. Es ist zu erkennen, dass die Situation mit den vergleichsweise weit von null entfernten Distanzen der Objekte zu sich selbst an diesem Datensatz nicht auftritt. Die Unterschiede zwischen den Distanzen von Objekt i zu Objekt j und vice

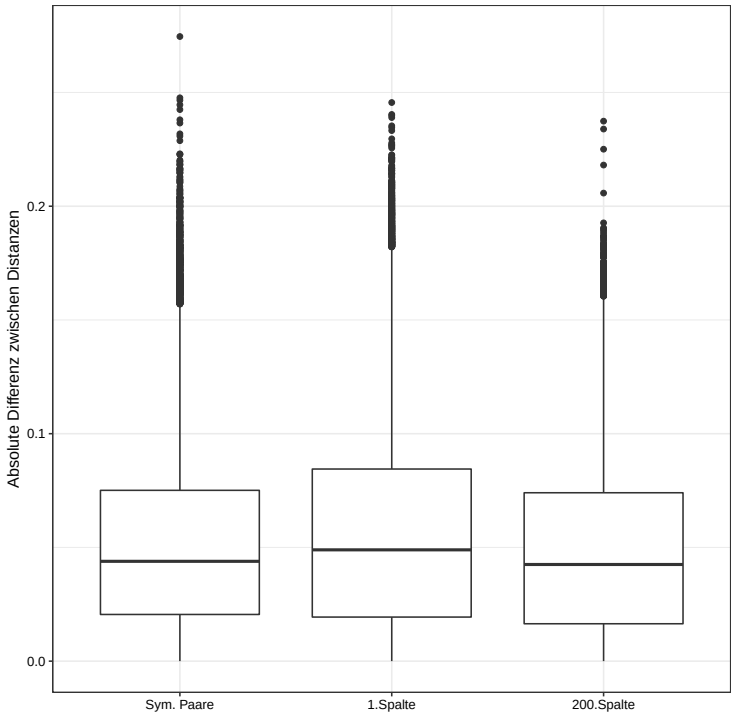

Abbildung 12: Darstellung der absoluten Differenzen zwischen ausgewählten Werten der Distanzmatrix bei Verwendung von `calc_shape_dist`. Im linken Boxplot sind die Unterschiede zwischen den theoretisch identischen Distanzen von Objekt i zu Objekt j und von Objekt j zu Objekt i zusammengefasst. Zum Vergleich sind mittig und links die Unterschiede zwischen Objekten in der angegeben Spalte der Distanzmatrix zu sehen.

Anzahl Datenpunkte	**Diag. Rang**	**MW Sym.**	**MW 1**	**MW 20**
100	1	0.05	0.43	0.27
50	1	0.10	0.36	0.26
30	1	0.26	0.27	0.35
15	1	0.40	0.73	0.33

Tabelle 4: Kennwerte von `calc_shape_dist` auf dem Datensatz `MPEG7`

versa werden jedoch mit weniger Datenpunkten immer größer. Bei einer Darstellung der Kurven durch 15 Punkte, was in etwa der Anzahl der Pseudo-Landmarks entlang der oberen Kurve der Schafknochen entspricht, liegen diese absoluten Differenzen in einem ähnlichen Wertebereich wie die entsprechenden Differenzen theoretisch nicht identischer Distanzen.

4.4 Abstandsmaße durch Betrachtung der Kurven als Funktionen

Bookstein-Koordinaten In diesem Unterkapitel soll eine Alternative für die eben vorgestellte elastische Distanz auf den durch die Pseudo-Landmarks diskretisierten zweidimensionalen Kurven vorgestellt werden. Da diese Kurven im zweidimensionalen kartesischen Koordinatensystem horizontal liegen und für jede Kurve für jedes x zwischen der x-Koordinate von Landmark 3 und der x-Koordinate von Landmark 1 genau ein zugehöriger y-Wert identifiziert werden kann, können die Kurven (also die obere und untere Seite der Knochen nach der Betrachtung in Abbildung 3) als stetige, injektive Abbildungen $x \mapsto y$ gesehen werden. Da auf dieser funktionalen

Betrachtungsweise basierende Abstandsmaße genauso wie die anderen in dieser Arbeit vorgestellten Abstandsmaße Ähnlichkeiten beziehungsweise Unähnlichkeiten zwischen den Umrissen zweier Knochen und allgemein den *shapes* zweier Objekte messen sollen, müssen diese ebenfalls invariant gegenüber den euklidischen Ähnlichkeitstransformationen Translation, isotropischer Skalierung und Rotation sein. Um sich bei der Wahl der (Semi-)Metriken auf den funktionalen Daten nicht einschränken zu müssen, ist es wünschenswert, diese Invarianz durch geeignete Darstellung der Kurven zu erreichen, bevor man diese als eindimensionale Abbildungen betrachtet. Dafür bieten sich die Bookstein-Koordinaten an. Die Idee hinter dieser Darstellung ist, zwei Landmarks auf fixe Positionen zu platzieren. Dazu muss das gesamte Objekt geeignet verschoben, gestreckt und gedreht werden. Für eine Konfiguration der Größe $k \times 2$, von der das erste Landmark auf die Position $(-\frac{1}{2}, 0)$ und das zweite Landmark an die Stelle $(\frac{1}{2}, 0)$ gesetzt werden soll, berechnen sich die übrigen $k-2$ Bookstein Koordinaten $(x_j^B, y_j^B)^T, j = 3, \ldots, k$ durch

$$x_j^B = \frac{\{(x_2 - x_1)(x_j - x_1) + (y_2 - y_1)(y_j - y_1)\}}{D_{12}^2} - \frac{1}{2},$$
$$y_j^B = \frac{\{(x_2 - x_1)(y_j - y_1) - (y_2 - y_1)(x_j - x_1)\}}{D_{12}^2},$$

wobei $D_{12}^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2$ die quadrierte euklidische Distanz zwischen dem ersten und dem zweiten Landmark ist [8, S.41]. In Anwendung auf die Kurve entlang der oberen Kante der Talusknochen setzen wir das Landmark 3 als den Punkt der Kurve, der am weitesten links außen liegt, auf $(-\frac{1}{2}, 0)$ und den äußerst rechten Punkt, also Landmark 1, auf $(\frac{1}{2}, 0)$. Bei der Berechnung der Bookstein-transformierten Koordinaten der (Pseudo-)Landmarks der Kurve entspricht damit Landmark 3 dem Landmark und Index 1 aus der vorgestellten Formel und Landmark 1 dem Landmark und Index 2 der Formel. Die Bookstein-Koordinaten eines Objekts und einer „Kopie“ dieses Objektes, die beliebig verschoben, rotiert und isotropisch skaliert ist, sind identisch. Demnach sind auf den Bookstein-Koordinaten basierende Abstandsmaße geeignet, um Unterschiede zwischen *shapes* zu messen. Ein weiterer Vorteil der Darstellung der Kurven per Bookstein-Koordinaten in Bezug auf die geplante Betrachtung der Kurven als Funktionen $f(x) = y$ ist, dass damit der Definitionsbereich für jede der Funktionen einheitlich der Bereich $[-\frac{1}{2}, \frac{1}{2}]$ ist.

In Abbildung 13 sind die nach der Bookstein-Transformation geglätteten Kurven beziehungsweise Funktionen deskriptiv dargestellt. Die Mittelwertfunktion $\bar{f}_g(x)$ der Klasse g berechnet sich als punktweises Mittel der Funktionen in der Klasse [20, S.22]:

$$\bar{f}_g(x) = \frac{1}{n_g} \sum_{i=1}^{n_g} f_{gi}(x), \ g = 1, \ldots, G,$$

wobei n_g die Anzahl der Beobachtungen in Klasse g und f_{gi} die i-te funktionale Beobachtung in Klasse g ist. Hier ist $G = 3$ mit den drei Klassen „Wildschaf“, „spätes Hausschaf“ und „frühes Hausschaf“. Zusätzlich zu den drei fett gedruckten Mittelwertfunktionen sind auch die Bereiche zwischen der 10%- und 90%-Quantilsfunktion in der jeweiligen Farbe markiert. Auch diese empirischen Quantilfunktionen berechnen sich wieder punktweise, sodass für jedes $x \in [-\frac{1}{2}, \frac{1}{2}]$ 10%

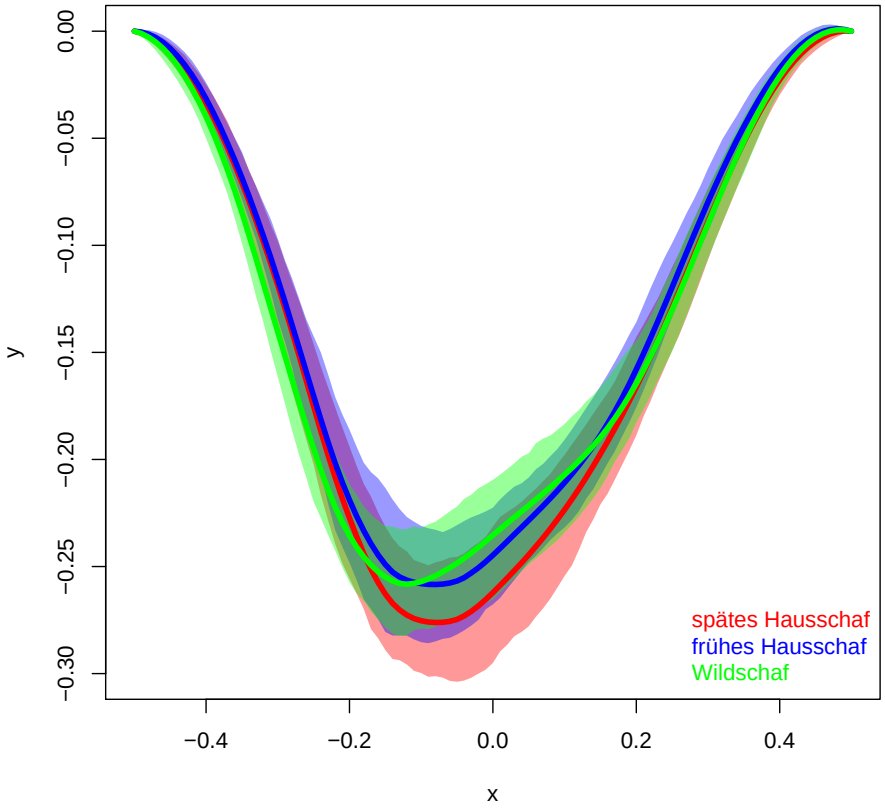

Abbildung 13: Deskriptive Darstellung der geglätteten Bookstein-transformierten oberen Kurve. Die fetten Linien stellen die Mittelwertfunktionen dar. Die Farbflächen markieren den Bereich zwischen den 10%- und 90%- Quantilen der y-Werte.

der Funktionswerte der jeweiligen Klasse unter der entsprechenden 10%-Quantilsfunktion und über der entsprechenden 90%-Quantilsfunktion liegen. Auffallend bei der Betrachtung dieser Darstellung ist die geringe Varianz der Funktionen zu Beginn und am Ende des Definitionsbereichs. Sie folgt aus der Grundidee der Bookstein-Darstellung, da jede Funktion durch die Punkte $(-\frac{1}{2}, 0)$ sowie $(\frac{1}{2}, 0)$ läuft. Demnach ist die größte Variabilität im Bereich um $x = 0$. Inhaltlich lassen sich ähnliche Unterschiede zwischen den drei Domestizierungsstatus erkennen wie sie auch schon in der ersten Betrachtung der oberen Kurven der Talusknochen in Abbildung 6 bemerkbar waren: Die Minima der Funktionen der fortgeschritten domestizierten Schafe sind geringer (also extremer) als bei den anderen beiden Klassen. Auch liegt die Mittelwertfunktion der Gruppe „Wildschaf“ auf der linken Seite, also für $x < -0.2$, tiefer (kleinere y-Werte) als die beiden anderen Mittelwertfunktionen. Sie verläuft in diesem Bereich fast auf Niveau der 10% Quantilfunktionen der beiden domestizierten Klassen.

Distanzmaße nach der $\mathbb{L}^2$-Norm Da wir nun eine geeignete Darstellung der Kurven für deren Betrachtung als funktionale Daten vorgestellt haben, stehen uns sämtliche Abstandsmaße

aus der funktionalen Datenanalyse zur Anwendung zur Verfügung. In dieser Arbeit wird jedoch nur das wohl klassischste Distanzmaß, die aus der $\mathbb{L}^2$-Norm $||\cdot||$ folgende Metrik

$$||(f_1 - f_2)|| = \sqrt{\int_D (f_1(x) - f_2(x))^2 \, dx}$$

(hier $D = [-\frac{1}{2}, \frac{1}{2}]$) beziehungsweise Abwandlungen davon als Maß des Abstandes zwischen zwei Funktionen f_1 und f_2 genauer betrachtet. Auf der Menge der dargestellten, als geglättete Funktionen betrachteten, Bookstein-transformierten Kurven ist dieses Distanzmaß eine Metrik, im ursprünglichen Konfigurationsraum ist es wie die anderen gegenüber den euklidischen Ähnlichkeitstransformationen invarianten Abstandsmaße nur eine Semi-Metrik. Die erwähnten abgewandelten Distanzmaße sind im Stile der von Fuchs et. al. [12] als *shortEucl* bezeichneten Semi-Metrik definiert:

$$d^{shortEucl}_{a,D_{small}}(f_1, f_2) = \sqrt{\int_{D_{small}} (f_1^a(x) - f_2^a(x))^2 \, dx},$$

wobei $D_{small} \subseteq D$ einen Teil des Definitionsbereichs und f^a die a-te Ableitung von f bezeichnet. Zur Implementierung dieses Abstandsmaßes in R benutzen wir das Paket **shapes** [9] zur vorhergehenden Bookstein-Transformation und das Paket **fda** [19] zur Bildung der Ableitungen der Funktionen.

4.5 Linearkombination von Abstandsmaßen

Wir haben in den vorherigen Teilen dieses Kapitels verschiedene Distanzmaße vorgestellt, die entweder Unähnlichkeiten zwischen den *shapes* zweier Konfigurationen von Landmarks oder zwischen den ebenfalls gegenüber euklidischen Ähnlichkeitstransformationen invarianten Verläufen zweier durch Pseudo-Landmarks diskretisierten Kurven messen. Damit beziehen sich die Abstandsmaße auf unterschiedliche Merkmale der Sprungbeinknochen. Es stellt sich die Frage, ob man diese verschiedenen Informationen kombinieren kann, um Unähnlichkeiten zwischen zwei Knochen besser quantifizieren zu können und damit die Grundlage für eine erfolgreichere Klassifikation zu legen. Eine Möglichkeit zur gewichteten Verknüpfung zweier Abstandsmaße d_1 und d_2 ist eine Linearkombination der Distanzen in der Form $d^{LK}_{1,2} = wd_1 + (1-w)d_2$, wobei $w \in [0, 1]$ der Gewichtungsfaktor des ersten Abstandsmaßes ist. Ruft man sich die Eigenschaften einer Semi-Metrik

i) $d(x, y) = d(y, x)$

ii) $d(x, y) \leq d(x, z) + d(z, y)$

iii) $d(x, x) = 0$

in Erinnerung, so lässt sich zeigen, dass auch die Linearkombination d_{LK} zweier Semi-Metriken eine Semi-Metrik ist:

$$\begin{aligned}
\text{i*)}\ d_{1,2}^{LK}(x,y) &= wd_1(x,y) + (1-w)d_2(x,y) \overset{\text{i)}}{=} \\
&= wd_1(y,x) + (1-w)d_2(y,x) = d_{1,2}^{LK}(y,x)
\end{aligned}$$

$$\begin{aligned}
\text{ii*)}\ d_{1,2}^{LK}(x,y) &= wd_1(x,y) + (1-w)d_2(x,y) \overset{\text{ii)}}{\leq} \\
&\leq w(d_1(x,z) + d_1(z,y)) + (1-w)(d_2(x,z) + (d_2(z,y)) = \\
&= wd_1(x,z) + (1-w)d_2(x,z) + wd_1(z,y) + (1-w)d_2(z,y) = \\
&= d_{1,2}^{LK}(x,z) + d_{1,2}^{LK}(z,y)
\end{aligned}$$

$$\text{iii*)}\ d_{1,2}^{LK}(x,x) = wd_1(x,x) + (1-w)d_2(x,x) \overset{\text{iii)}}{=} 0.$$

In dieser Art der Linearkombination haben wir jedoch das Problem, dass sich die Gewichtungsfaktoren nur schlecht interpretieren lassen. Dies liegt an den unterschiedlichen Wertebereichen der Distanzmaße. Beispielsweise beträgt die mittlere volle Procrustes-Distanz d_P zwischen den Landmark-Konfigurationen zweier unterschiedlicher Knochen 0.073 mit Standardabweichung 0.018, das arithmetische Mittel der $\mathbb{L}^2$-Distanzen zwischen den als Funktionen gesehenen oberen Kurven zweier unterschiedlicher Knochen ist jedoch 0.230 mit Standardabweichung 0.104. Eine Lösungsmöglichkeit ist, die Distanzen zu normieren. Dabei wird der geringste beobachtete Wert auf 0 gesetzt (ist bereits erfüllt aufgrund der Definition von (Semi-)Metriken), und der maximale beobachtete Wert auf 1, wobei der verhältnismäßige Abstand aller Werte erhalten bleibt. Die entsprechende Formel zur Berechnung der normierten Werte ist

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \overset{x_{min}=0}{=} \frac{x}{x_{max}},$$

es werden also alle Distanzwerte durch die maximale Distanz bei Verwendung des Distanzmaßes dividiert. Da der maximale Wert bei Verwendung der $\mathbb{L}^2$-Distanz 0.996 ist, werden die Werte dieses Abstandsmaßes kaum verändert (Arithmetisches Mittel: 0.231, Standardabweichung: 0.105). Die standardisierten Procustes-Distanzwerte haben nun allerdings den Mittelwert 0.388 und die Standardabweichung 0.094. Auch wenn diese Werte sich nicht perfekt entsprechen, so sind die Distanzen durch die Normalisierung doch vergleichbarer und damit der Gewichtsfaktor der Linearkombination besser interpretierbar geworden. Wir verwenden also die Linearkombination der normierten Distanzen $d_{1,norm}$ und $d_{2,norm}$:

$$d_{1,2}^{LK} = wd_{1,norm} + (1-w)d_{2,norm} \text{ mit } w \in [0,1].$$

Die in diesem Kapitel vorgestellte Distanzmaße sollen im folgenden Kapitel auf die beiden Datensätze zur Klassifikation anhand des kNN-Schätzers und eines kernbasierten Schätzers angewandt werden. Um die verschiedenen Modelle beziehungsweise Algorithmen vergleichen zu können,

müssen wir uns erst damit auseinandersetzen, wie man die Klassifikationsgüte auf neuen Daten ohne Label möglichst genau simulieren beziehungsweise schätzen kann.

5 Datenbasierter Vergleich der Klassifikationsmodelle

5.1 Modellevaluation

Einfache Performanzschätzung Um verschiedene Möglichkeiten, die Performanz eines Lernalgorithmus $\mathcal{A}$ zu schätzen, vergleichen zu können, ist es wichtig, sich noch einmal vor Augen zu führen, an welchem Vorhersagefehler wir interessiert sind. Wie bereits in Kapitel 2 erläutert, wollen wir den *Expected Generalization Error* $EGE_n(\mathcal{A})$, also den erwarteten Fehler der Prognose eines aus der gemeinsamen Verteilung $\mathbb{P}_{\mathcal{XG}}$ von Kovariablen und Label stammenden Datums auf n ebenfalls zufällig aus $\mathbb{P}_{\mathcal{XG}}$ gezogenen Trainingsdaten, minimieren. Zur Schätzung des EGE scheint es naheliegend, die gewünschte Größe des Trainingsdatensatzes $|\mathcal{D}_{Train}| = n$ einzuhalten und den Lernalgorithmus auf sämtliche vorliegende Daten anzuwenden. An dieser Stelle sei erwähnt, dass in diesem Fall die Schätzung des EGE der Schätzung des GE entspricht, da wir eine einzige Prognosefunktion $\hat{f}_{\mathcal{D}_{Train}}$ auf Basis des kompletten vorliegenden Datensatzes bestimmen. Um die Güte dieser Funktion empirisch zu schätzen, berechnen wir den durchschnittlichen Verlust auf den n Daten, welcher hier aufgrund des verwendeten *0-1-Loss* der Fehlklassifikationsrate entspricht. Dieser *Training Error*

$$\hat{GE}_{\mathcal{D}_{Train}}(\hat{f}_{\mathcal{D}_{Train}}) = \frac{1}{n}\sum_{i=1}^{n} L(\hat{f}_{\mathcal{D}_{Train}}(x_i), y_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[\hat{f}_{\mathcal{D}_{Train}}(x_i)\neq y_i]}$$

ist jedoch zu optimistisch, das heißt er unterschätzt den wahren Generalisierungsfehler, da die gleichen Daten zur Modellanpassung und -evaluation verwendet werden [14, S.228]. Insbesondere wird bei hoher Modellkomplexität (in unserem Fall kleine Lokalisationsparameter) *Overfitting*, also übermäßige Anpassung an die Trainingsdaten, nicht bestraft, da keine neuen, nicht zum Training verwendeten Daten zur Schätzung des Fehlers verwendet werden. Im extremsten Fall, der „k-nächste-Nachbarn"-Klassifikation mit $k = 1$, würden wir jedes Datum korrekt klassifizieren, da der nächste Nachbar in den Trainingsdaten das Datum selbst ist, und damit einen Generalisierungsfehler von 0 schätzen.
Demnach müssen wir zur verlässlichen Beurteilung eines Lernalgorithmus eine Situation simulieren, in der wir neue, noch nicht beobachtete Daten zur Evaluierung eines auf anderen, „bekannten" Daten angewandten Algorithmus verwenden. Die einfachste Möglichkeit, dies zu erreichen, ist das sogenannte *Hold-out Splitting*. Dabei werden die n vorhandenen Daten nach einer bestimmten Rate (meist zwischen 50% und 80%) in Trainings- und Testdaten aufgeteilt, siehe Abbildung 14 oben. Die Güte der auf den Trainingsdaten bestimmten Prognosefunktion $\hat{f}_{\mathcal{D}_{Train}}$ kann damit durch das arithmetische Mittel der Verluste der Testdaten $\mathcal{D}_{Test}$ bestimmt werden:

$$\hat{GE}_{\mathcal{D}_{Test}}(\hat{f}_{\mathcal{D}_{Train}}) = \frac{1}{|\mathcal{D}_{Test}|}\sum_{(x,y)\in\mathcal{D}_{Test}} L(\hat{f}_{\mathcal{D}_{Train}}(x), y) = \frac{1}{|\mathcal{D}_{Test}|}\sum_{(x,y)\in\mathcal{D}_{Test}} \mathbf{1}_{[\hat{f}_{\mathcal{D}_{Train}}(x)\neq y]}.$$

Da unser tatsächliches Ziel ist, die Güte des auf n Trainingsdaten angepassten Modells zu bestimmen, überschätzt dieser *Test Error* den Generalisierungsfehler unter der meist zutreffenden Annahme, dass auf einem kleineren Datensatz weniger gelernt werden kann. Dieser pessimistische Bias ist demnach umso geringer, je mehr Daten zum Training verwendet werden, d.h.

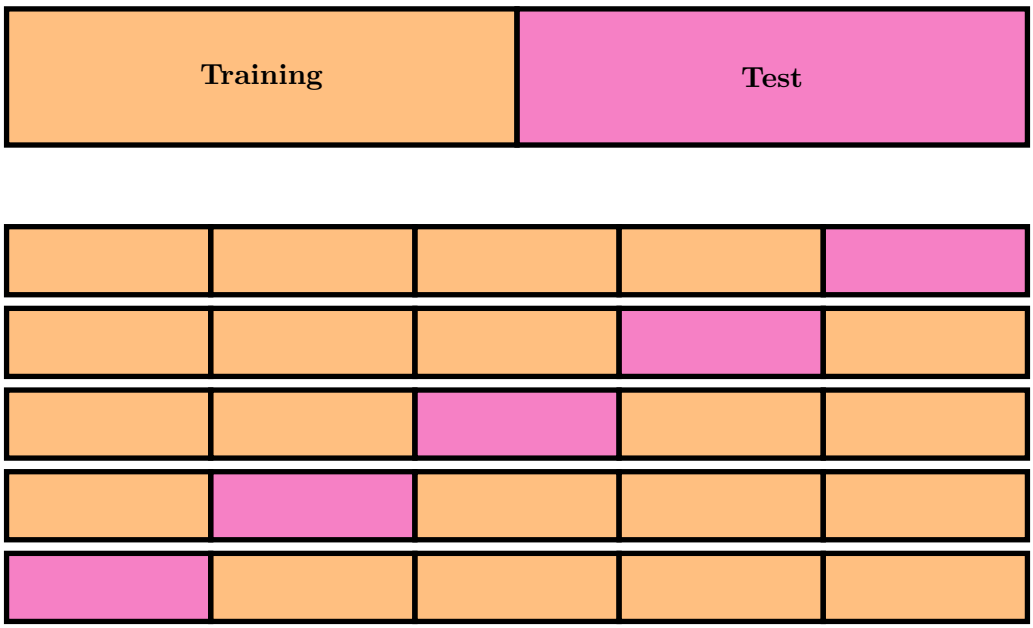

Abbildung 14: Methoden zur Evaluation eines Lernalgorithmus mit fest vorgegebenem Modell. Oben: Aufteilung der Daten in ein Trainingsdatensatz und ein Testdatensatz (*Hold-out set*). Unten: Kreuzvalidierung mit 5 Schichten (*5-Fold CV*).

je größer die Aufteilungsrate ist. Andererseits nimmt mit wachsendem Trainingsdatensatz und damit schrumpfendem Testdatensatz die Varianz der Schätzung zu, da diese von der zufälligen Ziehung immer weniger Testdaten aus den n vorhandenen Daten abhängt (vgl. [15, S.178]). Es tritt also auch hier ein *Bias-Variance-Tradeoff* ein.
Eine Möglichkeit, damit umzugehen, bietet das *Resampling*, wobei wir uns hier auf die wohl bekannteste Variante, die Kreuzvalidierung, beschränken wollen. Bei der *k-Fold Cross Validation (k-Fold CV)* werden die n Daten zufällig in k Schichten eingeteilt. In einem Schritt des k-mal wiederholten Prozesses setzt sich der Trainingsdatensatz aus $k-1$ Schichten zusammen, es gilt also $|\mathcal{D}^j_{Train}| = \frac{k-1}{k}n, j = 1, \ldots, k$. Das auf diesen Daten geschätzte Modell $\hat{f}_{\mathcal{D}^j_{Train}}$ kann nun auf den in diesem Schritt noch nicht verwendeten Daten der überbleibenden Schicht evaluiert werden. Diese wiederholte, geschichtete Aufteilung in Trainings- und Testset ist im unteren Teil von Abbildung 14 visualisiert. Man erhält also für $j = 1, .., k$ jeweils eine Schätzung des Generalisierungsfehlers beziehungsweise der Fehlklassifikationsrate:

$$\hat{GE}_{\mathcal{D}^j_{Test}}(\hat{f}_{\mathcal{D}^j_{Train}}) = \frac{1}{|\mathcal{D}^j_{Test}|} \sum_{(x,y)\in\mathcal{D}^j_{Test}} L(\hat{f}_{\mathcal{D}^j_{Train}}(x), y) = \frac{1}{|\mathcal{D}^j_{Test}|} \sum_{(x,y)\in\mathcal{D}^j_{Test}} \mathbf{1}_{[\hat{f}_{\mathcal{D}^j_{Train}}(x)\neq y]}.$$

Der erwartete Generalisierungsfehler des zugrunde liegenden Lernalgorithmus $\mathcal{A}$ kann damit als Mittel dieser k Werte geschätzt werden:

$$\hat{EGE}_n(\mathcal{A}) = \frac{1}{k}\sum_{j=1}^{k} \hat{GE}_{\mathcal{D}^j_{Test}}(\hat{f}_{\mathcal{D}^j_{Train}})$$

Der Vorteil dieses Verfahrens ist, dass die einzelnen Trainingsdatensätze meist größer als bei einem einfachen *Hold-out Splitting* sind und damit der pessimistische Bias geringer ist, und gleichzeitig die Varianz aufgrund der Bildung des arithmetischen Mittels der k geschätzten Ge-

neralisierungsfehler relativ gering gehalten wird (vgl. [15, S.182]).
Ein Spezialfall der *k-Fold CV* ist die *Leave-One-Out Cross-Validation (LOOCV)*, bei der $k = n$ gesetzt wird. Damit wird in jedem Schritt genau ein Datum auf einem Trainingsset der Größe $n - 1$ getestet. Folglich ist bei der LOOCV fast kein Bias zu beobachten. Ein weiterer Vorteil der LOOCV ist, dass die Ergebnisse deterministisch sind, da die Einteilung in einelementige Schichten eindeutig ist. Ein Nachteil im Vergleich zu beispielsweise 5-Fold CV oder 10-Fold CV liegt darin, dass sich die k Trainingssets sehr stark überlappen; sie sind beinahe identisch. Damit tendiert die LOOCV Schätzung zu einer höheren Varianz [15, S.183f.]. Zudem ist LOOCV bei Verfahren, in denen die Modellberechnung aufwändig ist, deutlich rechenaufwändiger, da n verschiedene Modelle berechnet werden müssen.

Performanzschätzung mit Hyperparametertuning Sämtliche bis hierhin in diesem Kapitel vorgestellte Methoden dienen dazu, die Güte eines komplett vorgegebenen Modells, wie zum Beispiel einer kNN-Schätzung unter Verwendung von genau 15 nächsten Nachbarn, zu schätzen. Üblicherweise will man aber den Algorithmus nicht im Voraus so stark einschränken. Es ist wünschenswerter, wenn die sogenannten Hyperparameter, wie eben die Lokalisationsparameter oder auch die Gewichtung in der Linearkombination, datenbasiert bestimmt werden. Das Ziel ist nun also, den erwarteten Generalisierungsfehler eines Algorithmus, der das beste Modell auf geeignete Art und Weise bestimmt, zu schätzen. Doch wie wählt man das optimale Modell aus? Auch hier muss man wieder eine Aufteilung der Daten vornehmen, um die Güte der verschiedenen Modelle auf den Trainingsdaten bestimmen zu können. Würde man eine der obig vorgestellten Methoden, beispielsweise *Hold-out Splitting*, für verschiedene Modelle anwenden und das Modell mit dem besten Ergebnis (im Beispiel dem kleinsten *Test Error*) als bestes Modell und demnach dieses beste Ergebnis als Schätzung des erwarteten Generalisierungsfehler betrachten, unterschätzt man den wahren Fehler [14, S.222]. Dies lässt sich an einem einfachen Beispiel nach Bischl et. al. [4, S.257f.] verdeutlichen. Wir betrachten ein binäres Klassifikationsproblem mit gleichen Klassengrößen. Zur Klassifikation verwenden wir einen Algorithmus, der vollkommen zufällig eine der beiden Klassen auswählt und einen Hyperparameter besitzt, der jedoch keinen Einfluss hat. Der wahre Testfehler ist demnach unabhängig vom Hyperparameter 50%, genauso wie der erwartete Generalisierungsfehler eines Algorithmus, der den „optimalen“ Hyperparameter auswählt. Führen wir nun für eine Reihe von Hyperparameterwerten die einfache Unterteilung in Trainings- und Testdatensatz durch, so erreichen wir willkürlich für einen zufälligen Wert einen Testfehler niedriger als 50%. Dieser Wert ist also keine geeignete Schätzung für die Vorhersagegüte auf neuen, unbeobachteten Daten.
Eine recht simple Lösung dieses Problems ist eine Dreiteilung der Daten in ein Trainingsset, auf dem die einzelnen Modelle angepasst werden, ein Validierungsset, das zur Bewertung und zum Vergleich der einzelnen Modelle verwendet wird, und ein Testset, anhand dessen der Generalisierungsfehler des auf Trainings - und Validierungsdatensatz trainierten besten Modells geschätzt wird (vgl. [14, S.222]). In der oberen Grafik in Abbildung 15 ist diese Aufteilung dargestellt. Dieses Verfahren hat jedoch ähnliche Schwächen wie das *Hold-out Splitting*: Aufgrund der zufälligen Einteilung besitzt es eine hohe Varianz und durch die zu geringe Anzahl an Trainingsdaten

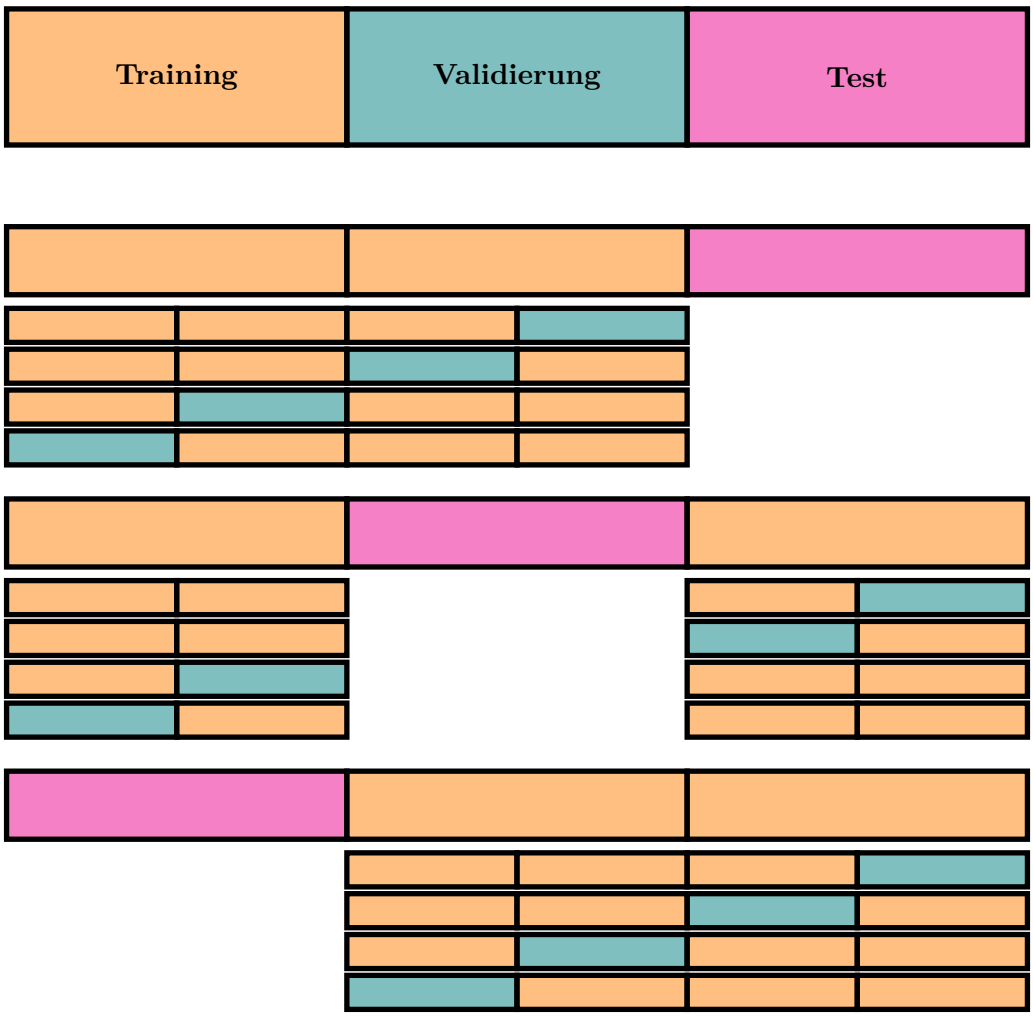

Abbildung 15: Methoden zur Evaluation eines Lernalgorithmus mit Modellselektion. Oben: Aufteilung der Daten in ein Test- und Trainingsset, wobei letzteres zur Modellselektion unterteilt wird in einen Trainings- und einen Validierungsteil. Unten: *Nested Cross Validation* mit 3 äußeren und 4 inneren Schichten.

einen pessimistischen Bias. Auch die Schätzung der Güte der einzelnen Modelle unterliegt einer hohen Varianz, da sie ebenfalls von einer einfachen Aufteilung abhängig ist.
Um die Auswahl des optimalen Modells stabiler zu gestalten, scheint es auch für diesen Schritt angebracht, eine *Resampling*-Methode, wie etwa Kreuzvalidierung, anzuwenden. Gleichzeitig will man aber auch die letztendliche Schätzung des erwarteten Generalisierungsfehlers nicht von einem einzelnen Split abhängig machen und auch in diesem Schritt eine wiederholte Aufteilung mit anschließender Ergebnisaggregation verwenden. Da die Modellauswahl Teil des Trainings des Lernalgorithmus (in Bezug auf das beiseite gelegte Testset) ist, muss sie für jede „äußere“ Aufteilung in Trainings- und Testdaten wiederholt werden [4, S.258]. Dies führt zum *Nested Resampling*, einem rechenintensiven Verfahren, bei dem zwei Kreuzvalidierungsverfahren (beziehungsweise allgemein *Resampling*-Verfahren) ineinander verschachtelt durchgeführt werden, wie in Abbildung 15 unten verdeutlicht. In einer äußeren Schleife wird jeweils ein „bestes“ Modell auf einer Schicht evaluiert. Zur Bestimmung des optimalen Modells wird jedoch für jede äußere Iteration auf den Trainingsdaten wiederum eine Kreuzvalidierung durchgeführt, mit welcher die Performanz der verschiedenen Modelle geschätzt wird (vgl. [4, S.258]).

Da sowohl beim kNN-Schätzer als auch bei den kernbasierten Schätzern ein Hyperparameter vorhanden ist, den wir nicht vorgeben wollen, verwenden wir in der folgenden Auswertung diese eben vorgestellte verschachtelte Kreuzvalidierung. Als Balance zwischen möglichst niedrigem Bias und möglichst niedriger Varianz verwenden wir 10 äußere Schichten, was eine standardmäßige Wahl ist, wie auch James et. al. schreiben [15, S.184]. Im Vergleich zu LOOCV bietet die 10-Fold CV zudem den Vorteil, dass es bedeutend weniger innere Schleifen gibt und damit der Rechenaufwand deutlich geringer ist. Zur Modellwahl benutzen wir in der inneren Schleife LOOCV, da wir hier eine Schätzung mit möglichst wenig Bias erzielen wollen, während für uns die Varianz der Schätzungen in der inneren Schleife nicht weiter interessant ist. Dies liegt daran, dass die Daten im inneren Schritt als Trainingsdaten des äußeren Splits gegeben sind und wir daher auf genau diesen Daten die Güte der Modelle möglichst genau schätzen wollen, während Abweichungen der Ergebnisse bei Schätzung auf anderen Daten nicht von Bedeutung sind. Auch der Rechenaufwand erhöht sich im Vergleich zu weniger inneren Schichten kaum, da sowohl bei der kNN-Methode als auch bei kernbasierten Schätzern keine Modelle beziehungsweise Modellparameter berechnet werden müssen.
Zur Schätzung der Performanz der Ensemble-Methode verwenden wir eine Kreuzvalidierung mit 5 äußeren Schichten, da die Rechenzeit hier – zumindest bei der benutzten Implementierung über das R-Paket **mlr** [3] – sehr schnell sehr hoch wird und wir durch die Zeiteinsparung durch Verwendung von weniger äußeren Schichten dafür etwas mehr Basismodelle mit aufnehmen können. Bei dieser Methode gibt es keine Hyperparameter, unter denen der optimale Wert gefunden werden muss. Die Schätzungen der Posteriori-Wahrscheinlichkeiten der jeweiligen Trainingsdaten werden jedoch ebenfalls mittels LOOCV durchgeführt.

5.2 Betrachtete Modelle und Vorgehensweise

Zu Beginn dieses Absatzes wollen wir auf die untersuchten Distanzmaße eingehen. Wie bereits in dem entsprechenden Kapitel 4.2 erwähnt, macht es bei der Analyse dieses Datensatzes sehr wahrscheinlich keine Unterschiede, welche der drei vorgestellten Procrustes-Distanzen man verwendet. Wir verwenden demnach nur eine davon, nämlich die volle Procrustes-Distanz d_F. Auch wenn wir die Procrustes-Distanzen ursprünglich als Methoden zur Distanzbestimmung zwischen zwei Konfigurationen mit korrespondierenden Landmarks eingeführt haben, so wenden wir sie auf verschiedene Mengen von Datenpunkten an: Nur auf die Landmarks (Kennzeichnung LM), auf die Landmarks und die Pseudo-Landmarks der oberen Kurve (LM+UC), sowie nur auf die Pseudo-Landmarks der oberen Kurve (UC). Wir betrachten also für die volle Procrustes-Distanz die Pseudo-Landmarks, die die obere Kurve diskretisieren, als normale Landmarks, die jeweils miteinander korrespondieren.
Für die euklidische Distanz zwischen den als Funktionen betrachteten Kurven $d^{shortEucl}_{a,D_{small}}$ betrachten wir die acht Kombinationen (also das Kartesischen Produkt der Mengen) von $D_{small} \in \{[-\frac{1}{2},\frac{1}{2}],[-\frac{1}{2},0],[-\frac{1}{4},\frac{1}{4}],[0,\frac{1}{2}]\}$ und $a \in \{0,1\}$. Wir wenden das aus der $\mathbb{L}^2$-Norm folgende Distanzmaß also auf die Funktionen selbst oder ihre ersten Ableitungen, entweder auf dem ganzen Definitionsbereich, der linken Hälfte des Definitionsbereichs, dem mittleren Bereich oder der rechten Hälfte des Definitionsbereichs an.

Für jedes dieser elf Distanzmaße wird in einem ersten Schritt jeweils ein kNN-Schätzer und ein kernbasierter Schätzer betrachtet. Wir benutzen nur einen der in Abschnitt 2.3 vorgestellten Kerne, den quadratischen Kern, um die Anzahl der Modelle zu reduzieren. Zudem haben sich in kleineren vorhergehenden Analysen wenig Unterschiede bei Verwendung verschiedener Kerne gezeigt. Da wir das im vorherigen Unterkapitel vorgestellte *Nested Cross Validation* Verfahren anwenden wollen, müssen wir die möglichen Werte der Hyperparameter festlegen, unter denen für jede äußere Schicht jeweils der optimale Wert ausgewählt wird. Als Anzahl nächster Nachbarn in der kNN-Methode benutzen wir für jedes Modell die ganzen Zahlen im Bereich von 1 bis 35. Die mögliche Bandweiten beim quadratischen Kern liegen aufgrund der unterschiedlichen Größenordnungen der resultierenden Distanzen von Modell zu Modell in unterschiedlichen Intervallen, um Rechenzeit zu sparen und trotzdem ein feines Raster an Werten zur Verfügung zu stellen. Die jeweiligen Bereiche sind basierend auf den Bandweiten mit den geringsten Fehlklassifikationsraten gewählt. In Tabelle 5 sind die Mengen der möglichen Bandweiten für die elf Abstandsmaße aufgeführt.

Distanz	**Intervall**	**Schrittweite**
Procrustes LM	[0.03, 0.08]	0.002
Procrustes LM+UC	[0.03, 0.08]	0.002
Procrustes UC	[0.03, 0.08]	0.002
$\mathbb{L}^2$ komplett	[0.1, 0.3]	0.01
$\mathbb{L}^2$ 1.Abl. komplett	[1.8, 2.3]	0.02
$\mathbb{L}^2$ links	[0.05, 0.25]	0.01
$\mathbb{L}^2$ mittig	[0.05, 0.25]	0.01
$\mathbb{L}^2$ rechts	[0.05, 0.25]	0.01
$\mathbb{L}^2$ 1.Abl. links	[1.2, 1.7]	0.02
$\mathbb{L}^2$ 1.Abl. mittig	[1.2, 1.7]	0.02
$\mathbb{L}^2$ 1.Abl. rechts	[1.2, 1.7]	0.02

Tabelle 5: Übersicht über die möglichen Werte der Bandweite bei Verwendung des quadratischen Kerns und den angegebenen Distanzmaßen.

In einem zweiten Schritt sollen Linearkombinationen dieser Distanzmaße betrachtet werden. Dabei beschränken wir uns auf Kombinationen von Procrustes-Distanzen mit jenen $\mathbb{L}^2$-Distanzen, die im ersten Schritt als einzelne Distanzen am besten abgeschnitten haben. Auch hierbei wenden wir wieder verschränkte Kreuzvalidierung an. Die Modelle haben nun allerdings zwei Hyperparameter, den Lokalisationsparameter der Schätzmethode und der Gewichtsfaktor w der Linearkombination. Bei Verwendung des kNN-Schätzers werden dabei für jede äußere Kreuzvalidierungsschicht auf den Trainingsdaten alle Kreuzprodukt-Kombinationen von $k \in \{1, 2, \ldots, 34, 35\}$ und $w \in \{0, 0.1, \ldots, 0.9, 1\}$ validiert und die jeweils beste Kombination ausgewählt. Bei Verwendung des quadratischen Kerns erleichtert sich die Auswahl des passendes Bereiches für die Bandweite aufgrund der vorhergehenden Normierung der Distanzen. Die möglichen Hyperparameterkombinationen setzen sich damit aus $h \in \{0.1, 0.12, \ldots, 0.38, 0.4\}$ und $w \in \{0, 0.1, \ldots, 0.9, 1\}$ zusammen.

Zudem wird die in Abschnitt 2.5 vorgestellte Ensemble-Methode angewandt. Theoretisch würden wir dabei gerne sehr viele Basismodelle zur Verfügung stellen und die Modellselektion weitest-

gehend der Methode selbst überlassen. Praktisch müssen auch hier wieder aufgrund des hohen komputationalen Aufwandes, vor allem bei Verwendung der $\mathbb{L}^2$-Distanz in den Grundmodellen, Abstriche gemacht werden und eine Vorselektion der Modelle durchgeführt werden. Wir werden uns daher auf die besten kNN-Modelle aus dem ersten Schritt und die jeweils am häufigsten verwendeten Hyperparameter k beschränken. Da die Berechnung auf den gesamten Datensatz aufgrund der höheren Datenanzahl ungefähr zehn mal so lange dauert als auf dem auf Karakulschafe und archäologische ostanatolische Wildschafe eingeschränktem Datensatz, werden wir die Ensemble-Methode nur auf letzteren Datensatz anwenden.
Als Vergleich zu diesen auf den Knochenumrissen basierenden Modellen soll eine Klassifikation anhand der linearen Maße aus Abbildung 2 durchgeführt werden. Diese sind zwar invariant gegenüber Translation und Rotation, aber nicht gegenüber isotropischer Skalierung. Um also einen fairen Vergleich mit den in Kapitel 4 vorgestellten *shape*-Distanzen ziehen zu können, müssen wir auch die linearen Maße größenbereinigen. Dazu wenden wir die selbe Methode wie Pöllath et. al. [17] an, in dem wir die *log shape ratio* (LSR) der Messungen an jedem Knochen bilden. Im folgenden soll x_i der Vektor der p Maße (hier: $p = 6$) an Untersuchungseinheit (Knochen) i sein. Der entsprechende LSR-Vektor x_i^{LSR} berechnet sich mit

$$x_{ik}^{LSR} = log(\frac{x_{ik}}{\sqrt[p]{\prod_{j=1}^{p} x_{ij}}}) = log(x_{ik}) - \frac{1}{p}\sum_{j=1}^{p} log(x_{ij}).$$

Als Semi-Metriken zwischen zwei Untersuchungseinheiten definieren wir die absoluten Differenzen zwischen den einzelnen Messungen nach LSR-Transformation. Auf Basis dieser sechs Abstandsmaße bilden wir ein kNN-Ensemble. Da diese Distanzen wesentlich schneller zu berechnen sind als die *shape*-Distanzen, können wir die jeweiligen Modelle mit einer größeren Anzahl an Hyperparametern als Grundmodelle verwenden.

5.3 Ergebnisse

5.3.1 Eingeschränkter Datensatz

In diesem Abschnitt sollen die Ergebnisse der binären Klassifikation auf dem eingeschränkten Datensatz bestehend aus den Knochen von Karakulschafen und Wildschafknochen aus den archäologischen Grabungen in Ostanatolien vorgestellt werden. In den Abbildungen 16 und 17 sind die mittleren Fehlklassifikationsraten für die äußeren Schichten des *nested CV*-Verfahrens für die kNN-Methode sowie die kernbasierte Methode mit asymmetrischem quadratischem Kern als Boxplot dargestellt, wobei anstatt dem Median jeweils der Mittelwert der zehn Fehlklassifikationsraten der Schichten, also die geschätzte mittlere Fehlklassifikationsrate, durch die rote Linie gekennzeichnet ist. Man kann sehen, dass viele der Modelle beziehungsweise Distanzmaße zu einer meist gelungenen Klassifikation führen. Die Procrustes-Distanz auf den Landmarks scheint deutlich weniger zur Trennung der beiden Gruppen geeignet zu sein als bei zusätzlicher Aufnahme der Pseudo-Landmarks entlang der oberen Kurve oder bei deren alleiniger Betrachtung. Bei Verwendung des kNN-Schätzers und des Procrustes-Distanzmaß auf den (Pseudo-)Landmarks der oberen Kurve wurden nur vier von 93 Knochen falsch zugeordnet, was einer Fehlklassifika-

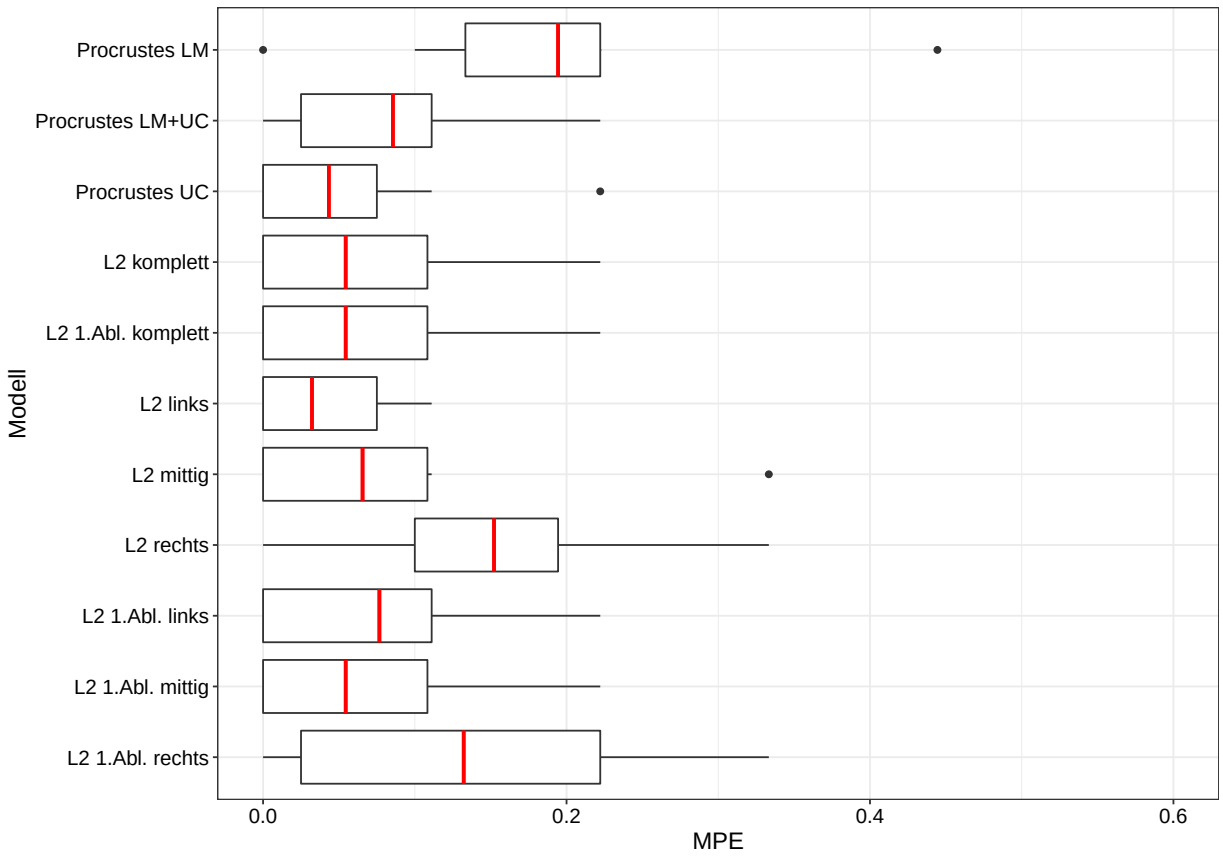

Abbildung 16: Ergebnisse der *nested CV* bei Verwendung der kNN-Methode und den angegebenen Distanzmaßen auf dem eingeschränkten Datensatz.

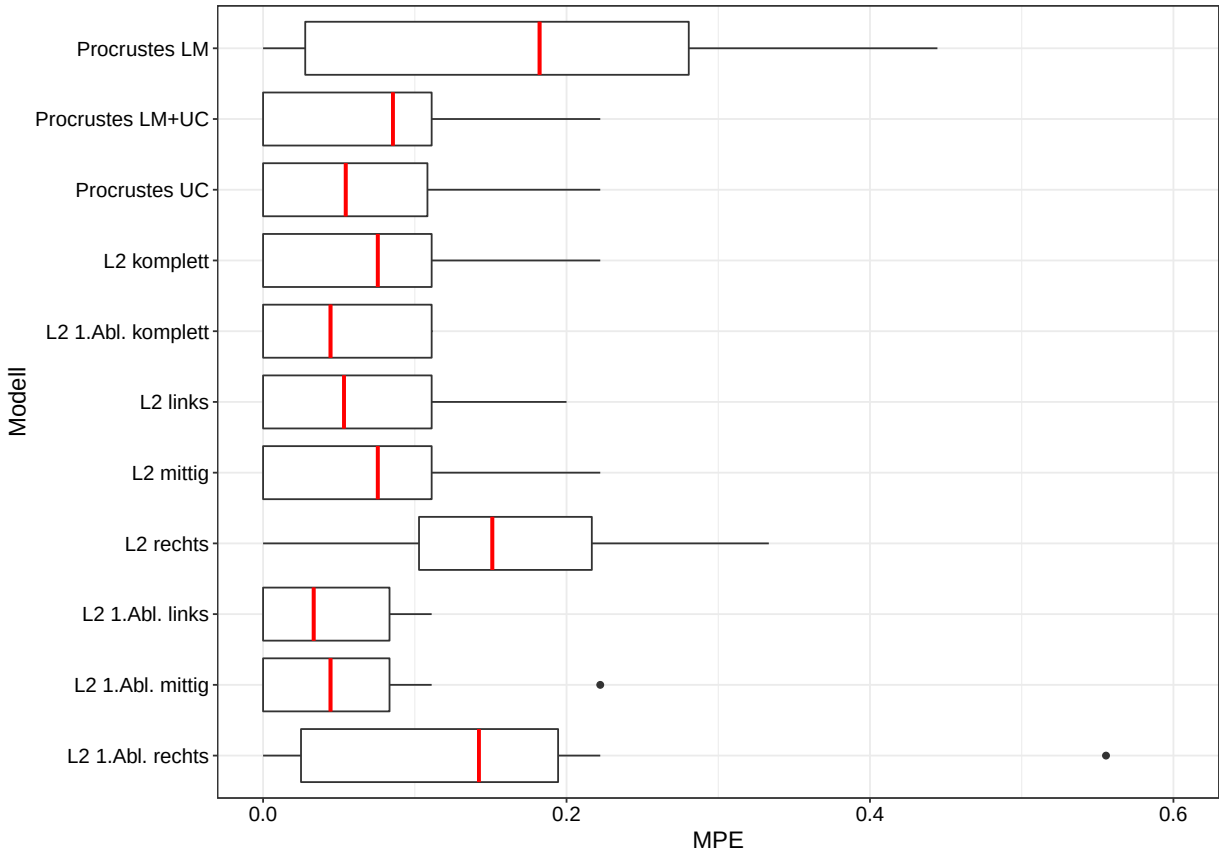

Abbildung 17: Ergebnisse der *nested CV* bei Verwendung der kernbasierten Methode mit quadratischem Kern und den angegebenen Distanzmaßen auf dem eingeschränkten Datensatz.

tionsrate von 4.3% entspricht. Diese vier Knochen stammen allesamt von einem Karakulschaf, wie ein Blick auf die Konfusionsmatrix in Tabelle 6 zeigt. Von den Knochen der 53 Wildschafe

		Wahre Klasse	
		Karakulschaf	**arch. Wildschaf**
Prädiktion	**Karakulschaf**	36	0
	arch. Wildschaf	4	53

Tabelle 6: Konfusionsmatrix bei Verwendung des kNN-Schätzers der Procrustes-Distanz auf Landmarks und Pseudo-Landmarks der oberen Kurve auf dem eingeschränkten Datensatz.

wurde hingegen kein einziger fälschlicherweise als Karakul eingeordnet. Für sechs der zehn äußeren Schichten wurden hierbei in der jeweiligen inneren LOOCV fünf nächste Nachbarn als der optimale Hyperparameterwert bestimmt. Von den auf der $\mathbb{L}^2$-Norm basierenden Semi-Metriken sticht bei der kNN-Methode insbesondere die auf das linke Teilintervall der Funktion beschränkte euklidische Distanz hervor. Dabei werden nur drei der 93 Knochen (3.2%), wieder alle von Karakulschafen stammend, falsch klassifiziert. Als optimale Anzahl nächster Nachbarn wurde in acht der zehn äußeren Schichten der Wert $k = 3$ bestimmt. Generell bestätigt sich der Eindruck aus den deskriptiven Darstellungen, dass insbesondere im linken und mittleren Teil der oberen Kurve Unterschiede zwischen den Domestikationsstatus bestehen.
Bei Verwendung der kernbasierten Methode, hier mit quadratischem Kern, zeigt sich ein ähnliches Bild: Die meisten den Verlauf der oberen Kurve berücksichtigende Abstandsmaße sind gut für diese binäre Klassifikation geeignet, wobei die $\mathbb{L}^2$-Distanz zwischen den Funktionen beziehungsweise ihrer ersten Ableitung nur auf das rechte Teilintervall eingeschränkt zu höheren Fehlklassifikationsraten als bei Einschränkung auf die anderen Intervalle führt. Die nach dieser Analyse für den Einsatz zusammen mit dem quadratischen Kern am besten geeignete Semi-Metrik ist die $\mathbb{L}^2$-Distanz zwischen den ersten Ableitungen der als Funktionen gesehenen oberen Kurven, eingeschränkt auf den linken Teilbereich: Auch hier ergibt sich wieder eine mittlere Fehlklassifikationsrate von 3.2%. Ein Unterschied macht sich jedoch bei Betrachtung der Konfusionsmatrix (Tabelle 7) erkennbar.

		Wahre Klasse	
		Karakulschaf	**arch. Wildschaf**
Prädiktion	**Karakulschaf**	39	2
	arch. Wildschaf	1	51

Tabelle 7: Konfusionsmatrix bei Verwendung des kernbasierten Schätzers mit quadratischem Kern und der $\mathbb{L}^2$-Distanz zwischen den linken Teilbereichen der ersten Ableitungen der als Funktionen betrachteten oberen Kurven auf dem eingeschränkten Datensatz.

Nur einer der drei falsch klassifizierten Knochen stammt von einem Karakulschaf, die beiden anderen stammen demnach von den archäologischen Wildschafen. Der Wert $h = 1.54$ wurde dabei in sechs der zehn äußeren Schichten als optimale Bandweite ausgewählt. Alles in allem sind wir beinahe in der Lage, die von Pöllath et. al. [17] mittels *Generalized Procrustes Analysis* und linearer Diskriminanzanalyse erzielte Klassifikationsrate von 97.6% zu erreichen. Dabei scheint unter Verwendung der $\mathbb{L}^2$-Distanz schon ein Teil der oberen Kurve (also circa acht

Pseudo-Landmarks) zusammen mit den Landmarks 1 und 3 der dorsalen Knochenansicht für die vorhergehende Bookstein-Transformation als Datenpunkte auszureichen.
Auch wenn die Ergebnisse unter Verwendung der einzelnen Distanzen schon recht zufriedenstellend sind, wollen wir im Folgenden untersuchen, ob eine Linearkombination einer Procrustes-Distanz und einer $\mathbb{L}^2$-Distanz auf dem funktionalen Verlauf der oberen Kurve zu noch niedrigeren Klassifikationsraten führt. Wie bereits angekündigt, schließen wir dabei drei schlechter abschneidende $\mathbb{L}^2$-Distanzen aus, nämlich die euklidische Distanz zwischen den Funktionen im mittleren Abschnitt sowie zwischen den Funktionen und ihren ersten Ableitungen im rechten Abschnitt. Die Resultate bei Verwendung der kNN-Methode sind in Abbildung 18 und bei Verwendung der kernbasierten Methode mit quadratischem Kern in Abbildung 19 dargestellt. Wir sehen, dass mit jeder einzelnen Kombination eine Fehlklassifikationsrate von unter 10% erreicht wird. An dieser Stelle sei angemerkt, dass die Unterschiede in den mittleren Fehlklassifikationsraten zumindest zum Teil auch auf die Zufälligkeit in der Bestimmung der zehn äußeren Schichten zurückzuführen sind und es sein kann, dass sich die Ordnung zweier Modelle bei einer anderen Aufteilung eventuell umdrehen wird. Nichtsdestotrotz ist bei der benutzten Aufteilung die Kombination der Procrustes-Distanz auf allen Landmarks und den Pseudo-Landmarks der oberen Kurve zusammen und der $\mathbb{L}^2$-Distanz auf dem mittleren Teil der ersten Ableitung der als Funktionen betrachteten oberen Kurve bei Verwendung des quadratischen Kerns besonders auffällig. Hier wird von beiden Gruppen nur jeweils ein Knochen falsch klassifiziert, was einer gesamten Fehlklassifikationsrate von ungefähr 2.1% entspricht. Das optimale Gewicht w als Hyperparameter wurde dabei auf sechs der zehn äußeren Aufteilungen als 0.5 bestimmt. Die beiden Metriken wirken also etwa gleich stark auf die Linearkombination.
Für das kNN-Ensemble sollen die selben acht Distanzmaße, also die Procrustes-Distanz auf Landmarks, Pseudo-Landmarks der oberen Kurve und beiden Mengen zusammen, sowie die $\mathbb{L}^2$-Distanzen $d^{shortEucl}_{0,\text{komplett}}, d^{shortEucl}_{0,\text{links}}, d^{shortEucl}_{1,\text{komplett}}, d^{shortEucl}_{1,\text{links}}$ und $d^{shortEucl}_{1,\text{mittig}}$, verwendet werden. Da bei den einzelnen Modellen fast nur einstellige Werte als optimale Anzahl nächster Nachbarn ausgewählt wurden, setzen wir die Menge der Hyperparameter auf $k \in \{3, 7, 11\}$. Insgesamt werden also 24 Basismodelle für das Ensemble erstellt. Die resultierende mittlere Fehlklassifikationsrate liegt bei etwa 4.3%, es werden also 4 der 93 Knochen falsch klassifiziert. Von den 24 Modellen fließen in jeder der fünf Schichten durchschnittlich 5.2 Modelle mit positivem Gewicht in den Ensemble-Schätzer ein. 13 Modelle wurden in mindestens einer der fünf Schichten mit einem Gewicht größer null gewichtet. In Abbildung 20 sind die Gewichte dieser Modelle dargestellt. Dabei stehen die schwarzen Punkte für die Gewichte auf einer einzelnen Schicht, und die roten Punkte stellen das über die fünf Schichten gemittelte Gewicht des jeweiligen Modells dar. Es fällt auf, dass (fast) kein Modell auf Basis der Procrustes-Distanz auf den Landmarks und der $\mathbb{L}^2$-Distanz auf den nicht eingeschränkten sowie auf das linke Teilintervall beschränkten ersten Ableitungen der oberen Kurve mit in das Ensemble-Modell aufgenommen wird. Die Modelle mit der Procrustes-Distanz auf den (Pseudo-)Landmarks der oberen Kurve sowie der $\mathbb{L}^2$-Distanz eingeschränkt auf das linke Teilintervall werden am höchsten gewichtet, was auch mit den in Abbildung 16 dargestellten Ergebnissen der kNN-Modelle mit den einzelnen Distanzen übereinstimmt.

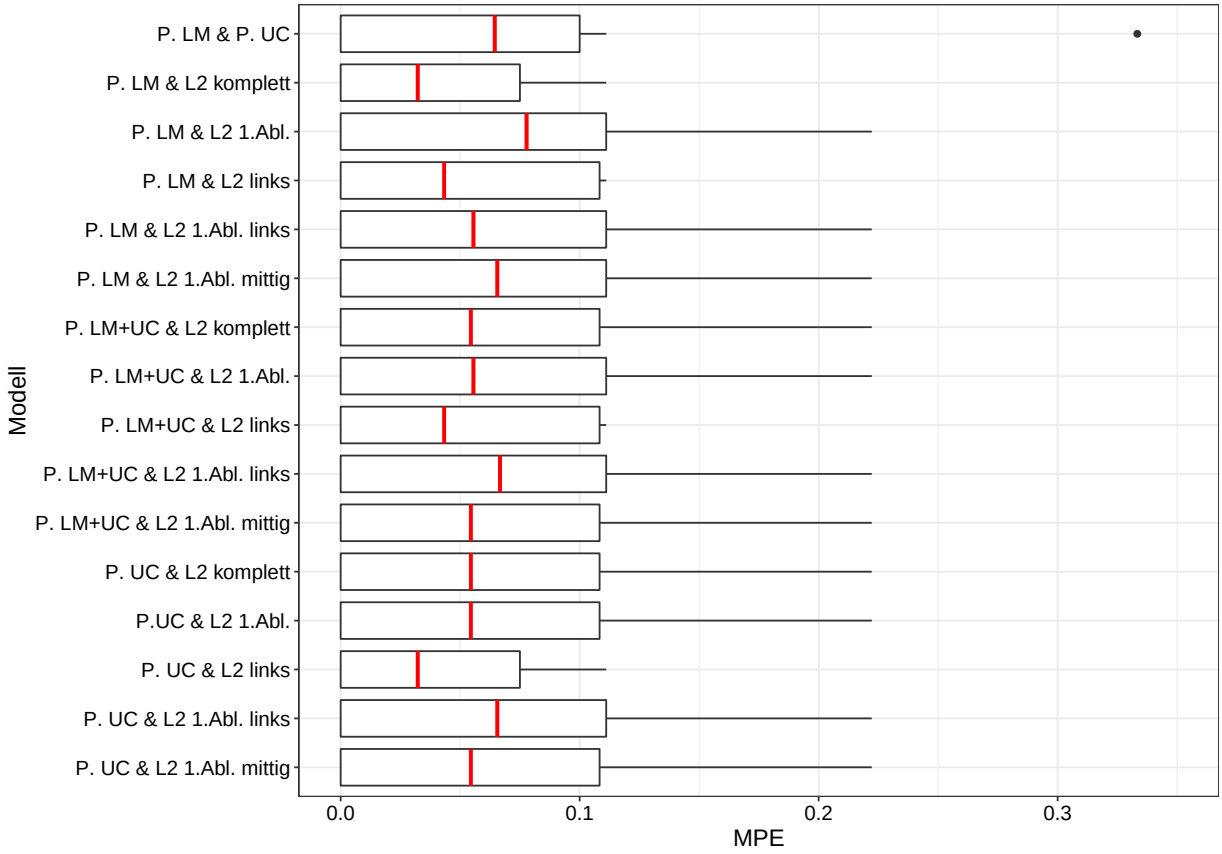

Abbildung 18: Ergebnisse der *nested CV* bei Verwendung der kNN-Methode und den angegebenen Distanzmaßen auf dem eingeschränkten Datensatz.

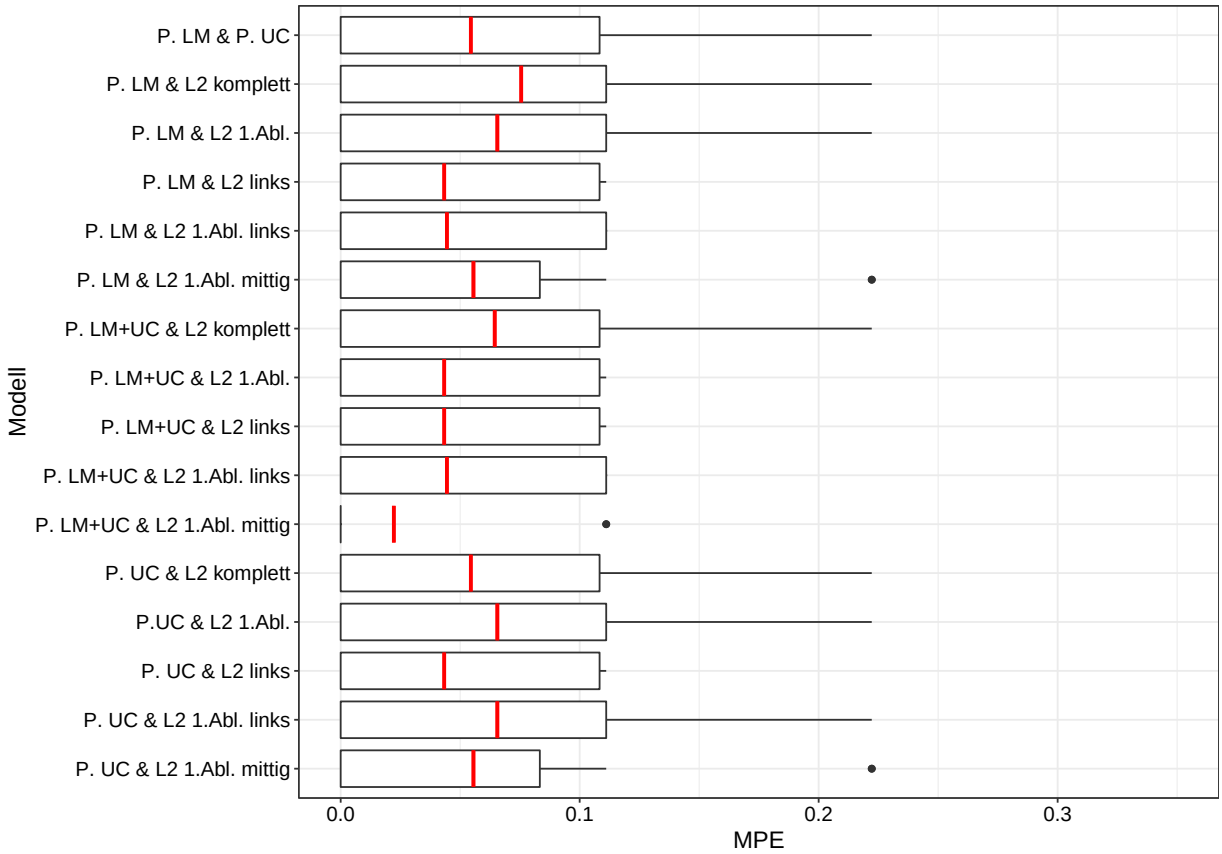

Abbildung 19: Ergebnisse der *nested CV* bei Verwendung der kernbasierten Methode mit quad. Kern und den angegebenen Distanzmaßen (eingeschränkter Datensatz).

Abschließend wollen wir das Ensemble auf Basis der größenbereinigten linearen Maße (79 Beobachtungen bei Einschränkung auf Karakulschafe und archäologische Wildschafe) betrachten. Dabei bilden die Kombinationen von den sechs univariaten absoluten Distanzen zwischen den jeweiligen linearen Maßen und dem Hyperparameter $k \in (1, 3, \ldots, 23, 25)$ 78 kNN-Basismodelle. Auf den fünf Kreuzvalidierungsschichten werden Fehlklassifikationsraten von 12.5%, 25%, 12.5%, 0% und ungefähr 6.7% erzielt. Die mittlere Fehlklassifikationsrate von circa 11% ist somit höher als bei Verwendung der auf die geometrische Form der Knochenumrisse bezogenen Distanzmaße sowie die von Pöllath et. al. erzielte Fehlklassifikationsrate von 2.5% unter Verwendung der linearen Maße. Es lohnt sich dennoch, die in der Ensemble-Methode geschätzten Gewichte genauer zu betrachten. Von den 78 Modellen bekamen je Schicht durchschnittlich nur sechs Modelle ein Gewicht größer als null zugeteilt. Über die fünf Schichten hinweg wurden 15 Modelle positiv gewichtet. Eine Übersicht über die Gewichte dieser 15 Modelle ist in Abbildung 21 gegeben. Es fällt auf, dass bis auf wenige niedrig gewichtete Ausnahmen nur Modelle mit Distanzen der von Pöllath et. al. definierten Maße M2 und M3 mit positivem Gewicht zum Ensemble-Modell beitragen. Dies steht in Einklang mit den von Pöllath et. al. berichteten Ergebnissen, die besagen, dass eine Hinzunahme der auf der oberen Kurve definierten Maße M1, M2 und M3 zu den „traditionellen“ Maßen von von den Driesch [26, S.89] eine bessere Klassifikation in die beiden Gruppen ermöglicht. Zudem entspricht dies auch der im ersten Teil der Analyse getroffenen Aussage, dass die obere Kurve der Knochen viel Information enthält und zumindest in dieser binären Klassifikation eine große diskriminative Kraft besitzt.

5.3.2 Gesamter Datensatz

Im Folgenden werden die Ergebnisse der simulierten Klassifikation der Knochen in die drei Klassen „Wildschaf“, „frühes Hausschaf“ und „spätes Hausschaf“ vorgestellt. Die Vorgehensweise ist dabei analog zu der Analyse der Klassifikationsergebnisse auf dem eingeschränkten Datensatz. Wir wollen also zuerst wieder die Ergebnisse der verschachtelten Kreuzvalidierung für kNN-Schätzer und kernbasiertem Schätzer mit den elf *shape*-Distanzen betrachten. Diese sind in den Abbildungen 22 und 23 zusammengefasst. Wir sehen, dass die Klassifikationsrate im Vergleich zur binären Klassifikation auf dem eingeschränkten Datensatz deutlich niedriger ist. Die durchschnittliche mittlere Fehlklassifikationsrate beträgt etwa 40%. Das noch beste Ergebnis bei Verwendung der kNN-Methode liefert die Procrustes-Distanz auf Landmarks und Pseudo-Landmarks der oberen Kurve mit einer mittleren Fehlklassifikationsrate von 30%. Um mögliche Strukturen in den Fehlklassifikationen zu erkennen, betrachten wir die Konfusionsmatrix (Tabelle 8). Der Anteil der korrekt zugeordneten Knochen (auch: *recall*) von frühen Hausschafen

		Wahre Klasse		
		frühes Hausschaf	**spätes Hausschaf**	**Wildschaf**
Prädiktion	**frühes Hausschaf**	48	12	10
	spätes Hausschaf	14	84	25
	Wildschaf	8	13	59

Tabelle 8: Konfusionsmatrix bei Verwendung der kNN-Methode und der Procrustes-Distanz auf Landmarks und Pseudo-Landmarks der oberen Kurve.

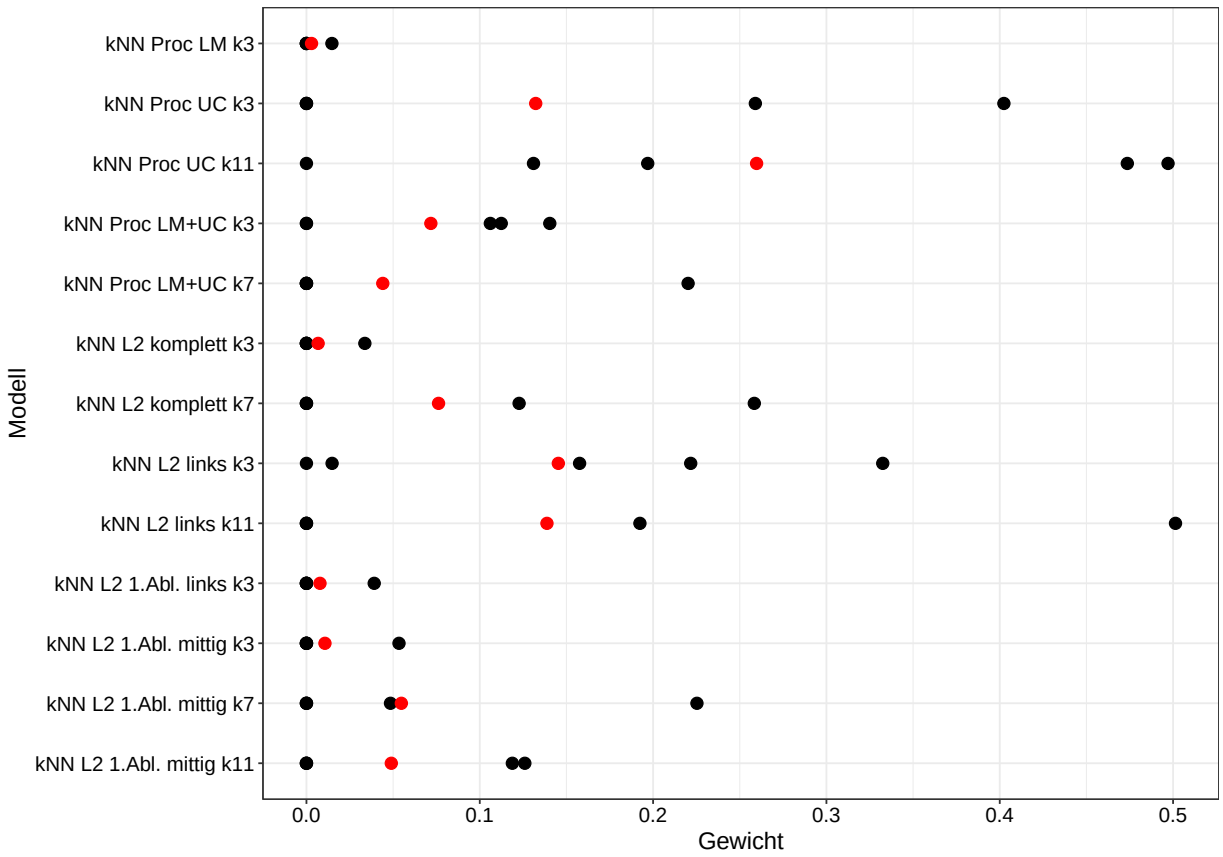

Abbildung 20: Modelle des kNN-Ensembles der *shape*-Distanzen auf dem eingeschränkten Datensatz mit einem positivem mittleren Gewicht.

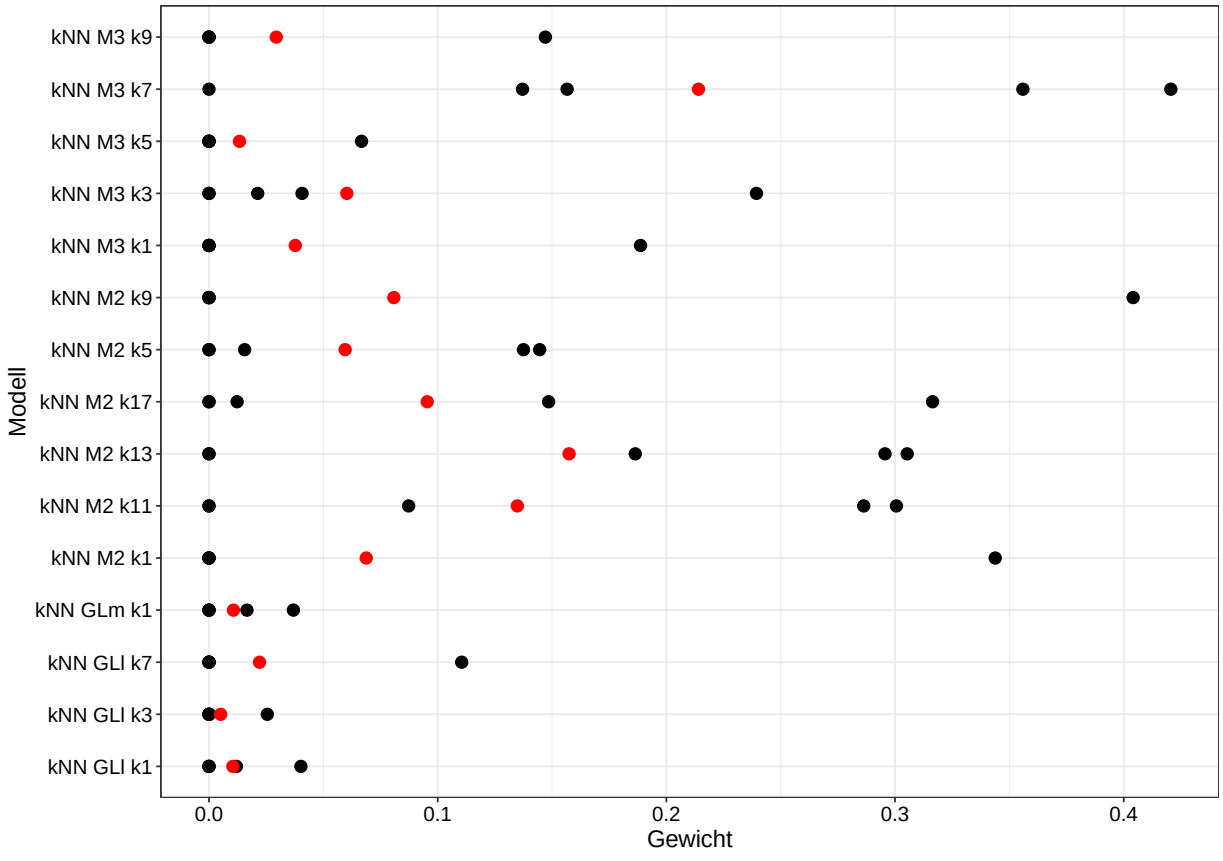

Abbildung 21: Modelle des kNN-Ensembles der linearen Maße auf dem eingeschränkten Datensatz mit einem positivem mittleren Gewicht.

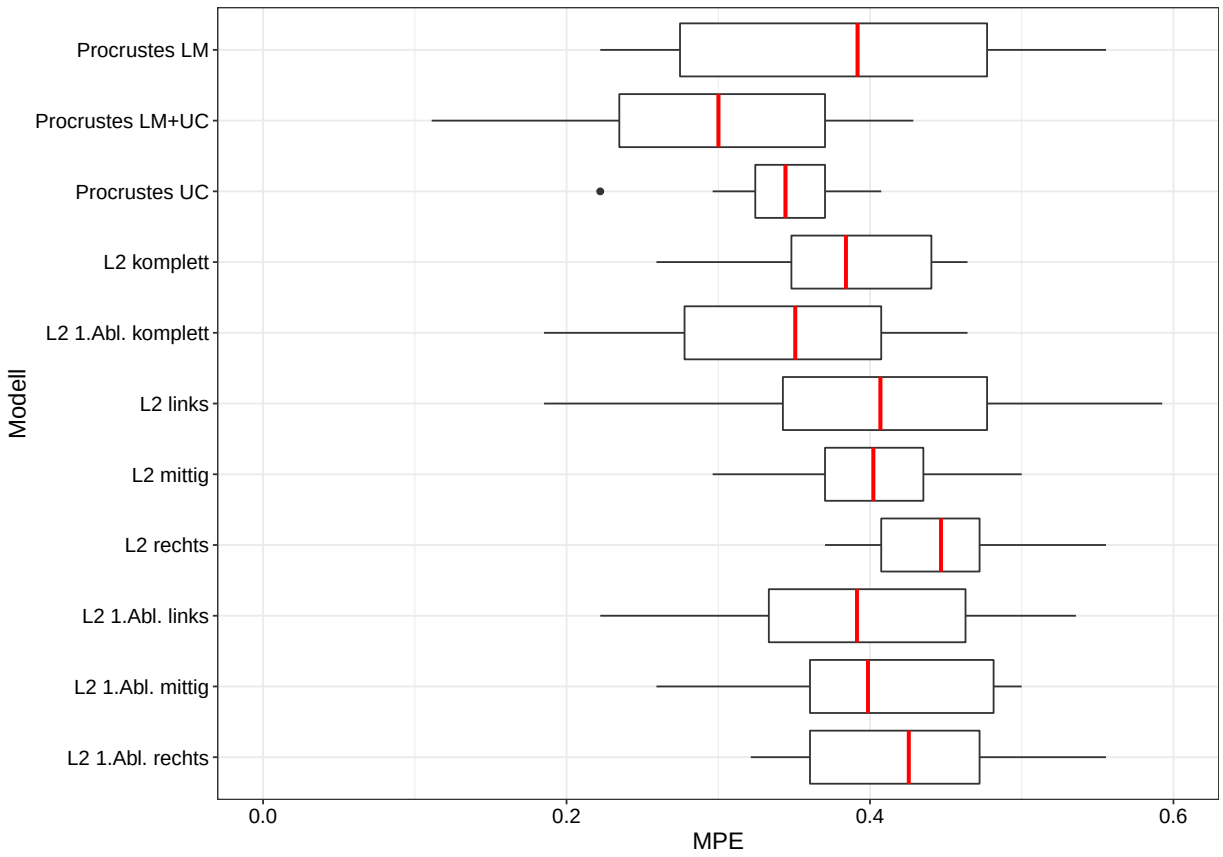

Abbildung 22: Ergebnisse der *nested CV* bei Verwendung der kNN-Methode und den angegebenen Distanzmaßen auf dem kompletten Datensatz.

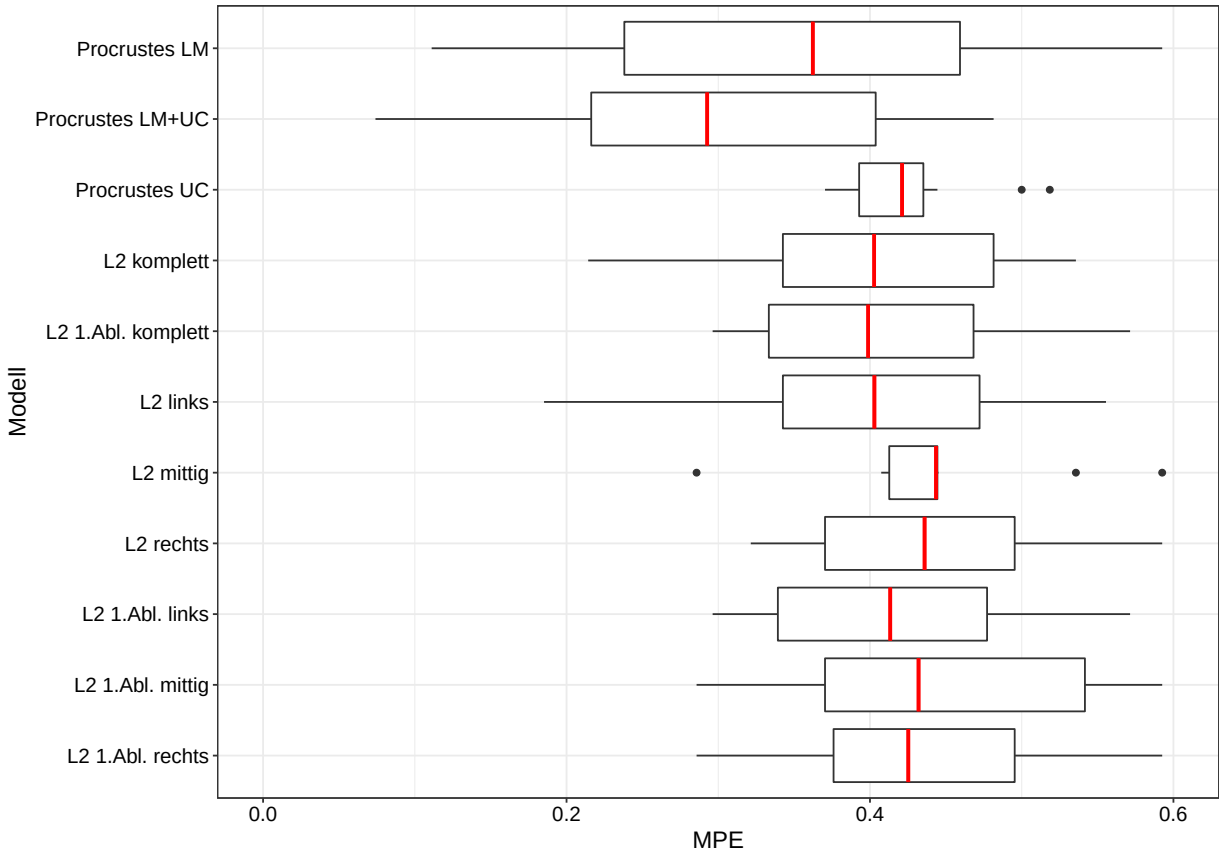

Abbildung 23: Ergebnisse der *nested CV* bei Verwendung der kernbasierten Methode mit quadratischem Kern und den angegebenen Distanzmaßen auf dem kompletten Datensatz.

beträgt $48/70 \approx 68.6\%$, der *recall* der Klasse „spätes Hausschaf“ ist $84/109 \approx 77.1\%$ und der *recall* der Knochen der Wildschafe ist $59/94 \approx 62.8\%$. Die Wildschafknochen lassen sich demnach unter Verwendung des kNN-Schätzers und der Procrustes-Distanz auf Landmarks und Pseudo-Landmarks der oberen Kurve verhältnismäßig am schlechtesten klassifizieren. Wirkliche Fehlerquellen lassen sich hier allerdings nicht erkennen. Zusätzlich lassen sich noch die Prädiktionen aufgeteilt nach den Schafrassen beziehungsweise den archäologischen Ausgrabungsstätten betrachten, wie in Tabelle 9 dargestellt. Hier sind größere Unterschiede zu erkennen. Während

		Prädiktion		
		frühes Hausschaf	**spätes Hausschaf**	**Wildschaf**
Bezeichnung	**Soayschaf**	**36**	4	1
	Güvercinkayası	**12**	10	7
	Karakulschaf	4	**35**	1
	Marschschaf	3	**18**	2
	Tall Munbaqa	5	**31**	10
	Elburs-Gebirge	7	13	**21**
	Göbekli Tepe	2	9	**25**
	Gusir Höyük	1	3	**13**

Tabelle 9: Vorhergesagte Klassen aufgeteilt nach den Rassen beziehungsweise Ausgrabungsstätten bei Verwendung der kNN-Methode und der Procrustes-Distanz auf den Landmarks und den Pseudo-Landmarks der oberen Kurve. Die fett gedruckten Anzahlen repräsentieren die jeweils wahre Klasse.

die Einordnung von Knochen mancher Rassen, insbesondere die der Soayschafe (ca. 87.8% korrekt klassifiziert) und der Karakulschafe (87.5%), sehr gut gelingt, gibt es andere Schafgruppen, deren Klassifikation eher schlecht funktioniert, vor allem die Knochen aus Güvercinkayası mit nur etwa 41.4% korrekt klassifizierten Daten. Ein Grund für dieses Ergebnis könnte, wie schon bei der Vorstellung der Daten erwähnt, die nicht mit vollständiger Sicherheit geklärte Klassenzugehörigkeit der Knochen aus Güvercinkayası sein.

Die Klassifikationsergebnisse bei Verwendung des kernbasierten Schätzers mit quadratischem Kern sind sehr ähnlich zu denen der kNN-Methode. Auch hier führt die Procrustes-Distanz auf Landmarks und oberer Kurve zu der besten Trennung in die drei Klassen mit einer mittleren Fehlklassifikationsrate von 29.3%. Ähnlich wie bei Verwendung der kNN-Methode streut die Fehlklassifikationsrate des Modells auf den einzelnen Schichten sehr, mit Werten von 7.4% bis 48.1% und einer geschätzten Standardabweichung von 13.7%. Auch die Konfusionsmatrix weist eine sehr ähnliche Struktur wie in Tabelle 8 auf.

Es stellt sich die Frage, ob mit einer geeigneten Linearkombination zweier Distanzmaße als neue Semi-Metrik geringere Fehlklassifikationsraten erreicht werden können. Eine Übersicht über die Ergebnisse ist für die kNN-Methode in Abbildung 24 und für den kernbasierten Schätzer mit asymmetrischem quadratischem Kern in Abbildung 25 gegeben. Es fällt auf, dass die Kombinationen von Procrustes-Distanz auf den (Pseudo-)Landmarks der oberen Kurve und den euklidischen Distanzen auf den als Funktionen betrachteten oberen Kurven tendenziell schlechter abschneiden als die anderen Kombinationen. Demnach scheinen für diese Trennung in drei Klassen die Information aus der oberen Kurve nicht ausreichend zu sein. In Kombination mit

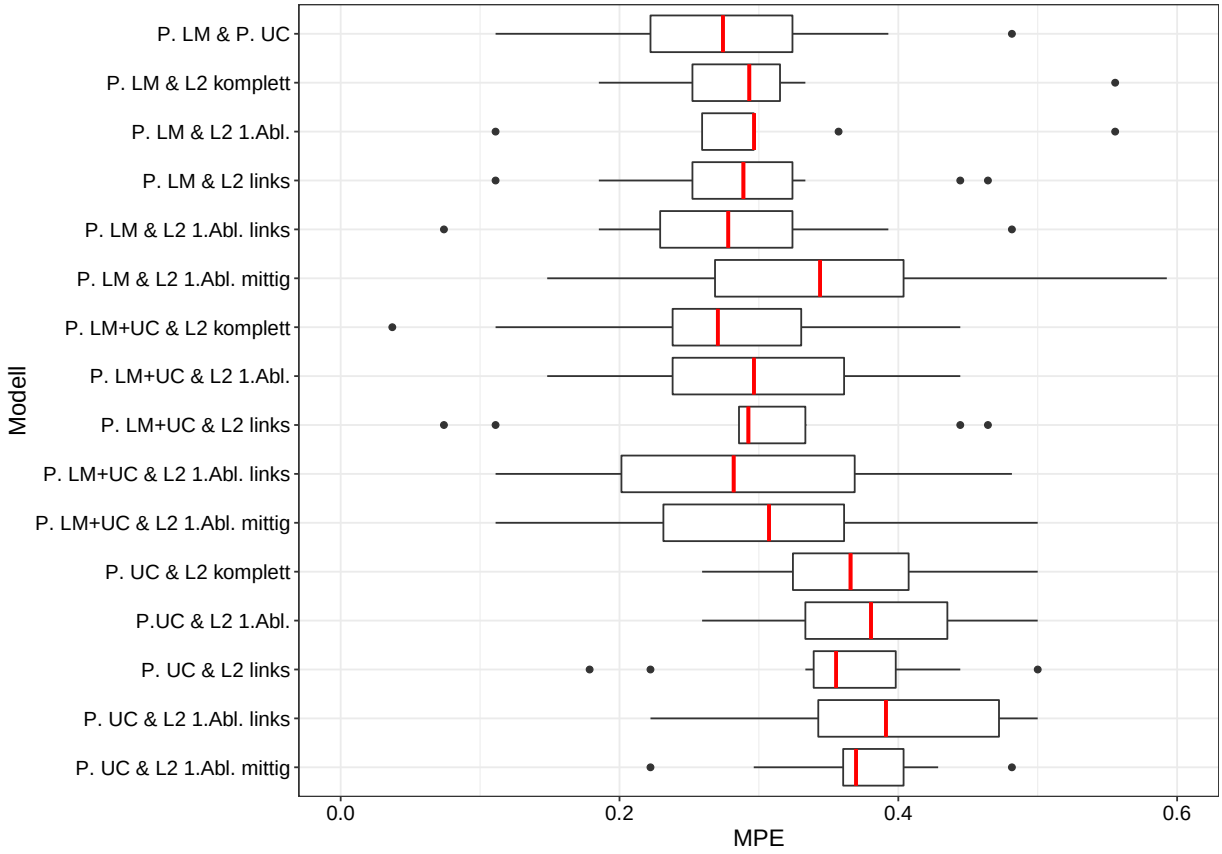

Abbildung 24: Ergebnisse der *nested CV* bei Verwendung der kNN-Methode und den angegebenen Distanzmaß-Kombinationen auf dem gesamten Datensatz.

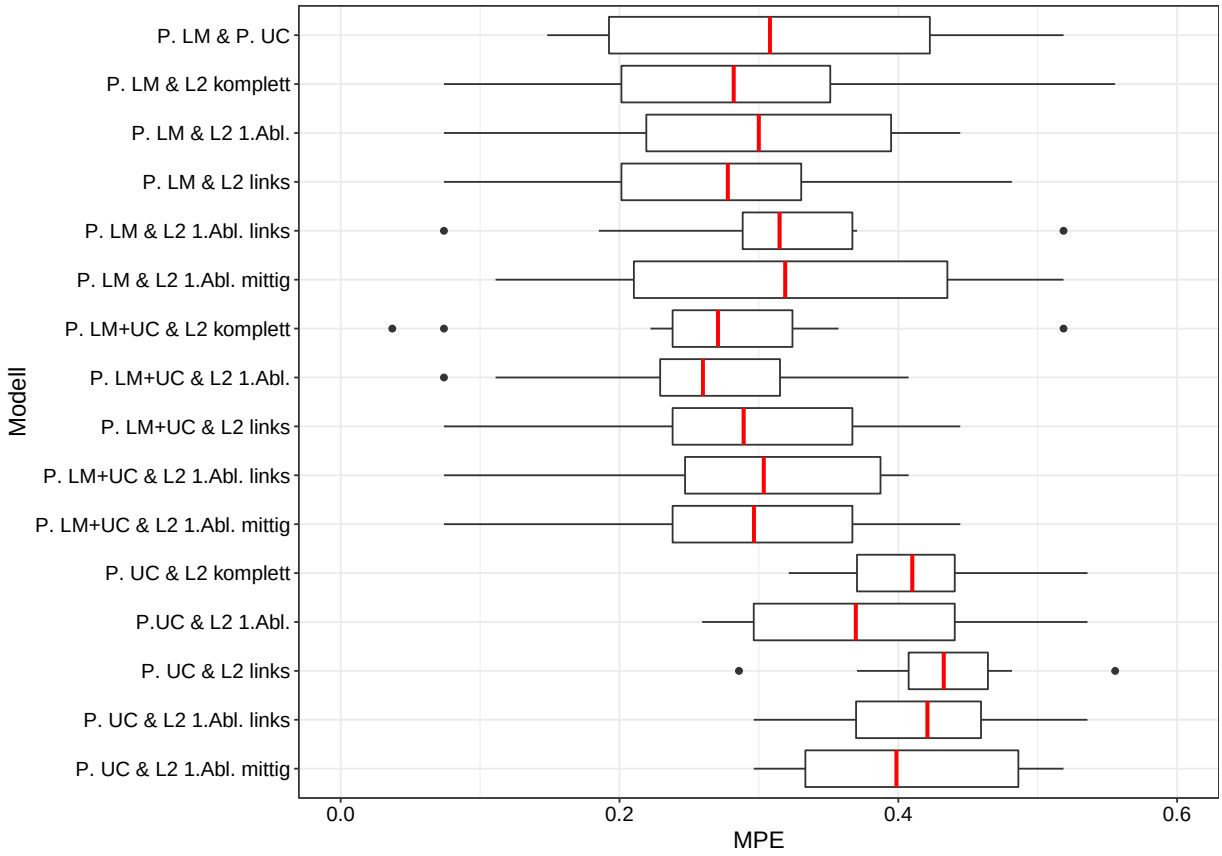

Abbildung 25: Ergebnisse der *nested CV* bei Verwendung der kernbasierten Methode mit quad. Kern und den angegebenen Distanzmaß-Kombinationen (gesamter Datensatz).

Distanzmaßen, die die restlichen Landmarks mit einbeziehen, lassen sich jedoch etwas bessere Klassifikationsergebnisse erzielen. Die Fehlklassifikationsrate bei Verwendung des kernbasierten Modells mit quadratischem Kern und der Linearkombination der Procrustes-Distanz auf Landmarks und oberer Kurve mit der $\mathbb{L}^2$-Distanz der ersten Ableitung der als Funktionen gesehenen oberen Kurven beträgt etwa 26%. Die entsprechende Konfusionsmatrix ist in Tabelle 10 dargestellt. Der größte Unterschied im Vergleich zur einzelnen Verwendung der Procrustes-Distanz auf

		Wahre Klasse		
		frühes Hausschaf	**spätes Hausschaf**	**Wildschaf**
Prädiktion	**frühes Hausschaf**	46	9	13
	spätes Hausschaf	14	87	12
	Wildschaf	8	13	69

Tabelle 10: Konfusionsmatrix bei Verwendung der kernbasierten Methode mit quadratischem Kern und einer Linearkombination der Procrustes-Distanz auf Landmarks und Pseudo-Landmarks der oberen Kurve und der $\mathbb{L}^2$-Distanz zwischen den kompletten ersten Ableitungen der als Funktionen betrachteten oberen Kurven.

Landmarks und Pseudo-Landmarks der oberen Kurve (Konfusionsmatrix in Tabelle 8) ist die verbesserte Klassifikation der Wildschafknochen, von denen jetzt etwa 73.4% korrekt zugeordnet werden. Die Dazunahme der $\mathbb{L}^2$-Distanz $d^{shortEucl}_{1,\text{komplett}}$ scheint sich dementsprechend vor allem auf die Einteilung der Wildschafknochen positiv auszuwirken. Auch bei den Prädiktionen nach Rasse beziehungsweise Ausgrabung bietet sich ein ähnliches Bild (Tabelle 11). Insbesondere fallen die

		Prädiktion		
		frühes Hausschaf	**spätes Hausschaf**	**Wildschaf**
Bezeichnung	**Soayschaf**	**34**	6	1
	Güvercinkayası	**12**	8	9
	Karakulschaf	3	**37**	1
	Marschschaf	3	**18**	2
	Tall Munbaqa	3	**32**	11
	Elburs-Gebirge	8	6	**27**
	Göbekli Tepe	4	4	**28**
	Gusir Höyük	1	2	**14**

Tabelle 11: Vorhergesagte Klassen aufgeteilt nach den Rassen beziehungsweise Ausgrabungsstätten bei Verwendung der kernbasierten Methode mit quadratischem Kern und einer Linearkombination der Procrustes-Distanz auf Landmarks und Pseudo-Landmarks der oberen Kurve und der $\mathbb{L}^2$-Distanz zwischen den kompletten ersten Ableitungen der als Funktionen betrachteten oberen Kurven. Die fett gedruckten Anzahlen repräsentieren die jeweils wahre Klasse.

Knochen aus Güvercinkayası wieder negativ auf mit nur circa 41.4% richtig klassifizierten Daten. Bei Betrachtung der bestimmten optimalen Gewichte zeigt sich, dass die Procrustes-Distanz in jeder der zehn Schichten mit einem Gewicht im Bereich von 0.6 bis 0.8 gewichtet wird, was die Annahme bestätigt, dass die Procrustes-Distanz mehr zum Modell beiträgt und die $\mathbb{L}^2$-Distanz eher zum „Fein-Tuning“ dient.

Auch hier wollen wir wieder vergleichend das Ensemble basierend auf den linearen Maßen betrachten. Als Basismodelle werden kNN-Modelle, jeweils mit einem der sechs Distanzmaße und

dem Hyperparameter k aus $1, 3, 5, 10, \ldots, 30, 35$, verwendet. Insgesamt stehen also 54 Basismodelle zur Verfügung. Dieses kNN-Ensemble schneidet mit einer Fehlklassifikationsrate von etwa 53% jedoch nicht gut ab. Auch die Modellselektion klappt nur bedingt. So werden je Schicht durchschnittlich 9.8 Modelle mit in den Ensemble-Schätzer aufgenommen. In Abbildung 26 sind die Gewichte der 22 Modelle dargestellt, die ein positives durchschnittliches Gewicht besitzen. Es fällt schwer, diese Gewichte inhaltlich zu interpretieren. kNN-Modelle mit der Distanz zwi-

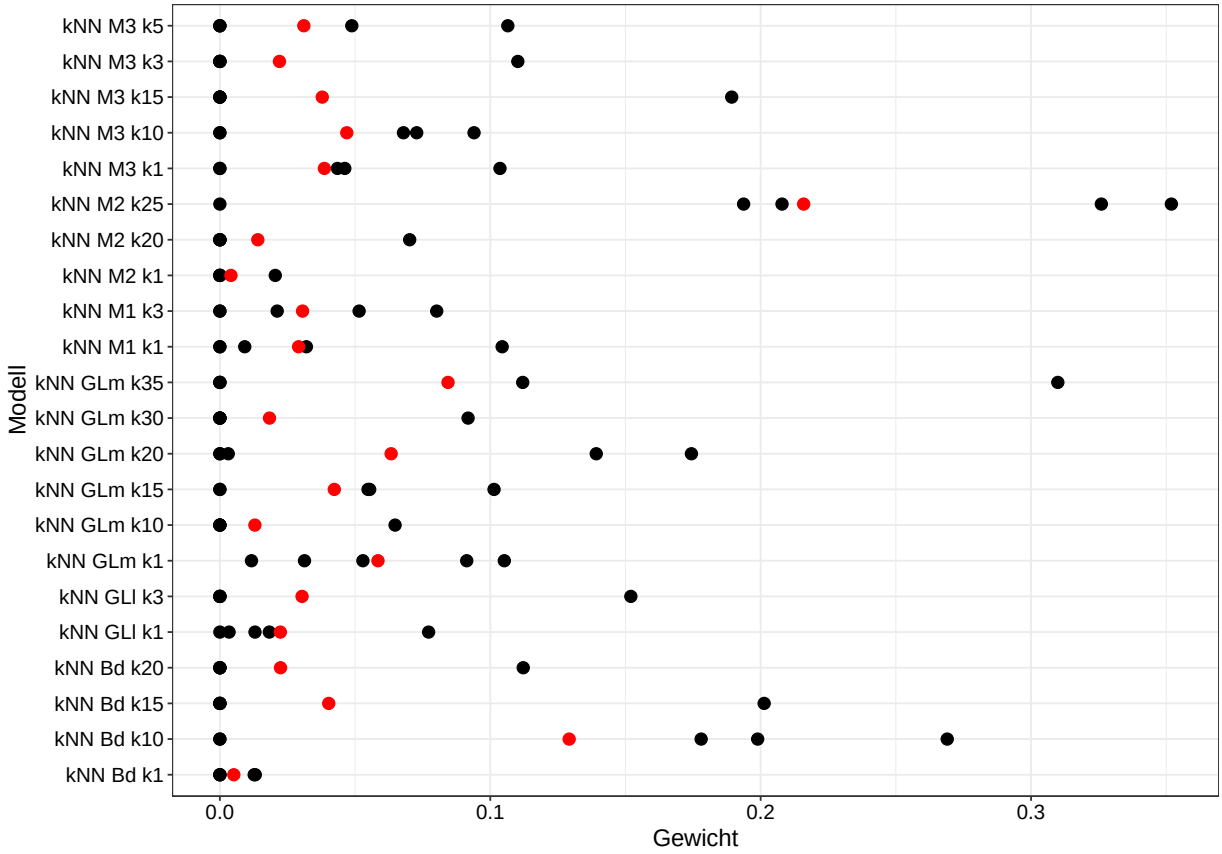

Abbildung 26: Modelle des kNN-Ensembles der linearen Maße auf dem gesamten Datensatz mit einem positivem mittleren Gewicht.

schen jedem der sechs linearen Maße sind mit in das Ensemble aufgenommen. Das nach diesen Gewichten am bedeutendste Modell ist das Modell mit der Distanz zwischen den M2-Maßen unter Verwendung 25 nächster Nachbarn. Doch auch dieses Modell ist in einer der fünf Schichten nicht im Ensemble-Schätzer enthalten.

Insgesamt lässt sich sagen, dass die in den linearen Messungen enthaltene Information nicht ausreicht, um die Schafknochen in die drei Gruppen „frühes Hausschaf“, „spätes Hausschaf“ und „Wildschaf“ zu trennen. Auch unter Verwendung der Distanzmaße, die die geometrische Anordnung der Landmarks berücksichtigen, klappt diese Klassifikation nicht fehlerfrei, aber dennoch bedeutend besser.

6 Fazit & Ausblick

In dieser Arbeit wurden mehrere Distanzmaße zur Quantifizierung der Unähnlichkeit des *shape* zweier Sprungbeinknochen von Schafen vorgestellt. Es hat sich gezeigt, dass bezüglich der Klassifikation nach dem Domestikationsstatus in der diskretisierten oberen Kurve der Dorsalansicht diskriminative Information enthalten ist und es sich definitiv lohnt, diese Daten zusammen mit den gewöhnlicheren Landmarkdaten zu erheben. Insbesondere der linke und mittlere Teil dieser Kurven eignet sich zur Klassifikation nach dem Domestizierungsstatus, wie wir sowohl an den Klassifikationsergebnissen der entsprechenden Modelle als auch an den deskriptiven Darstellungen der Kurven erkennen konnten. Wir mussten jedoch feststellen, dass die Implementation der auf der *Square Root Velocity* Transformation basierenden elastischen Distanz nach Srivastava et. al. [23] in der Funktion `calc_shape_dist` im R-Paket **fdasrvf** [25] fehlerhaft ist und diese damit nicht zur Klassifikation angewandt werden kann. Folglich ist es für weitergehende Analysen von Interesse, diese in der Theorie sehr vielversprechende Methode verbessert zu implementieren. Eine aufgrund der besonderen Lage der Kurven entlang der oberen Kante des Knochen mögliche Alternative bietet sich in der funktionalen Betrachtung der geglätteten Kurven. In dieser Arbeit wurde die auf der $\mathbb{L}^2$-Norm basierende funktionale euklidische Distanz sowohl auf die Kurven als auch auf ihre Ableitungen angewandt. Es lassen sich jedoch theoretisch sämtliche für die Analyse funktionaler Daten geeignete (Semi-)Metriken verwenden, so dass auch eine Anwendung dieser einen weiteren interessanten Ansatzpunkt für fortführende Analysen darstellt. Auch Linearkombinationen verschiedener Distanzen scheinen einen Mehrwert zu bieten. Die Kombination von Procrustes-Distanzmaßen mit $\mathbb{L}^2$-Distanzen auf den als Funktionen angenommenen oberen Kurven der Knochen hat sowohl auf einem auf zwei Schafgruppen beschränkten Datensatz als auch auf dem gesamten Datensatz zu einer verbesserten Klassifikationsrate geführt.
Bezüglich der Klassifikationsmethoden ließ sich in den Klassifikationsergebnissen kaum ein Unterschied zwischen dem k-Nächste-Nachbarn-Schätzer und dem kernbasierten Schätzer feststellen. Die Ensemble-Methode bietet durch die interpretierbare interne Gewichtung eine Möglichkeit zur Abschätzung der Bedeutung der benutzten Basismodelle, scheint aber – zumindest nach den aufgrund des hohen komputationen Aufwandes nur in geringem Ausmaß durchgeführten diesbezüglichen Analyse – nicht zu einer besseren Klassifikation zu führen. Auch die Verwendung weiterer Klassifikationsmethoden kann Grundlage für weitere Untersuchungen sein.
Während mit den verwendeten Methoden auf dem eingeschränkten Datensatz eine fast perfekte Trennung von Knochen von Karakul-Schafen und Schafknochen aus zwei ostanatolischen Grabungsstätten möglich ist, wurde mit keiner Methode auf dem kompletten Datensatz bei der Klassifikation nach den drei Domestizierungsstatus „Wildschaf“, „frühes Hausschaf“ und „spätes Hausschaf“ eine Fehlklassifikationsrate von unter 25% erreicht. Die Erhebung zusätzlicher Daten, insbesondere an Knochen weiterer Rassen beziehungsweise Herkunftsorte, kann neben den bereits vorgestellten weiteren Analysemöglichkeiten ein Schritt zur erfolgreicheren Klassifikation sein.

Inhalt des elektronischen Anhangs

Auf der beigelegten CD befinden sich die angefertigte Bachelorthesis im `.pdf`-Format sowie der Ordner *Analyse*. Dieser enthält folgende Unterordner:

- *Daten*: Enthält die zur Verfügung gestellten Knochenumrissdaten.
- *DistanzMatrizen*: Enthält die mit den verschiedenen Distanzmaßen berechneten Distanzmatrizen im `.RData`-Format.
- *EnsembleFiles*: Enthält die Ergebnisobjekte der in 5.3 dargestellten kNN-Ensembles im `.RData`-Format.
- *ErgebnisDataFrames*: Enthält die zur Erstellung der Grafiken in 5.3 nötigen `data.frames` mit den Fehlklassifikationsraten in den einzelnen Schichten im `.RData`-Format.
- *Plots*: Enthält Grafiken, darunter u.a. sämtliche in dieser Thesis verwendete Abbildungen im `.pdf`-Format.

Zudem befinden sich im Ordner *Analyse* verschiedene R-Skripte. Einige wichtige davon sind:

- `BasicFunctions.R`: Das „Herzstück“ der Analyse. In diesem Skript werden die meisten wichtigen Funktionen, z.B. Abstandsmaße, Kernfunktionen und Methoden zur Modellevaluation, definiert. Jedes andere Skript beginnt mit `source("BasicFunctions.R")`, sodass diese Funktionen zur Verfügung stehen.
- `CheckSRVFError.R`: Hier werden die Schwächen der Funktion `fdasrvf::calc_shape_dist` aufgezeigt und dargestellt.
- `DeskriptiveAnalyse.R`: Zur Erstellung der deskriptiven Abbildungen in 3.3.
- `ErgebnisseKlassifikationGesamterDatensatz.R`: Auswertung zu 5.3.2.
- `ErgebnisseKlassifikationKarakulArchWild.R`: Auswertung zu 5.3.1.
- `KurvenalsFunktionen.R`: Bookstein-Transformation, $\mathbb{L}^2$-Distanz, Ableitung von Funktionen. Basierend auf 4.4.
- `LaengenMessungen.R`: LSR-Transformation der Längenmaße, Durchführung der entsprechenden Ensemble-Methoden.
- `plotKerne.R`: Erstellung der Abbildung 1.
- `SimularityTransformations.R`: Erstellung der Abbildung 4.
- `SimulationKlassifikation.R`: Erstellung der Abbildungen 9, 10 und 11.

Abbildungsverzeichnis

Tabellenverzeichnis

Literatur

[1] https://www.aphorismen.de/zitat/119697. Aufgerufen am 29.06.2018

[2] Adams, Dean C. ; Rohlf, F. J. ; Slice, Dennis E.: Geometric morphometrics: Ten years of progress following the 'revolution'. In: *Italian Journal of Zoology* 71 (2004), S. 5 – 16. – Online erhältlich unter https://doi.org/10.1080/11250000409356545

[3] Bischl, Bernd ; Lang, Michel ; Kotthoff, Lars ; Schiffner, Julia ; Richter, Jakob ; Studerus, Erich ; Casalicchio, Giuseppe ; Jones, Zachary M.: mlr: Machine Learning in R. In: *Journal of Machine Learning Research* 17 (2016), 1 – 5. http://jmlr.org/papers/v17/15-066.html

[4] Bischl, Bernd ; Mersmann, Olaf ; Trautmann, Heike ; Weihs, Claus: Resampling Methods for Meta-Model Validation with Recommendations for Evolutionary Computation. In: *Evolutionary Computation* 20 (2012), S. 249 – 275. – Online erhältlich unter https://doi.org/10.1162/EVCO_a_00069

[5] Breno, Matteo ; Leirs, Herwig ; Van Dongen, Stefan: Traditional and geometric morphometrics for studying skull morphology during growth in *Mastomys natalensis* (Rodentia: Muridae). In: *Journal of Mammalogy* 92 (2011), S. 1395 – 1406. – Online erhältlich unter https://doi.org/10.1644/10-MAMM-A-331.1

[6] Brier, Glenn W.: Verification of forecasts expressed in terms of probability. In: *Monthly Weather Review* 78 (1950), S. 1 – 3. – Online erhältlich unter ftp://ftp.library.noaa.gov/docs.lib/htdocs/rescue/mwr/078/mwr-078-01-0001.pdf

[7] Domingos, Pedro: A Unified Bias-Variance Decomposition / Department of Computer Science and Engineering, University of Washington, Seattle, WA. 2000. – Forschungsbericht. – Online erhältlich unter ftp://ftp.cs.washington.edu/tr/2000/01/UW-CSE-00-01-02.pdf

[8] Dryden, Ian ; Mardia, Kanti: *Statistical Shape Analysis with Applications in R*. 2. Auflage. Wiley, 2016 (Wiley Series in Probability and Statistics). – ISBN 9780470699621

[9] Dryden, Ian L.: *shapes: Statistical Shape Analysis*, 2017. https://CRAN.R-project.org/package=shapes. – R package version 1.2.3

[10] Ferraty, Frédéric ; Vieu, Philippe: *Nonparametric Functional Data Analysis: Theory and Practice*. Springer, 2006 (Springer Series in Statistics). – ISBN 9780387303697

[11] Forster, Otto: *Analysis 2: Differentialrechnung im $\mathbb{R}^n$, gewöhnliche Differentialgleichungen*. 10. Auflage. Springer Spektrum, 2013 (Grundkurs Mathematik). – ISBN 9783658023577

[12] Fuchs, Karen ; Gertheiss, Jan ; Tutz, Gerhard: Nearest neighbor ensembles for functional data with interpretable feature selection. In: *Chemometrics and Intelligent Laboratory*

Systems 146 (2015), S. 186 – 197. – Online erhältlich unter https://www.sciencedirect.com/science/article/pii/S0169743915001100

[13] Gertheiss, Jan ; Tutz, Gerhard: Feature selection and weighting by nearest neighbor ensembles. In: *Chemometrics and Intelligent Laboratory Systems* 99 (2009), S. 30 – 38. – Online erhältlich unter http://www.sciencedirect.com/science/article/pii/S0169743909001464

[14] Hastie, Trevor ; Tibshirani, Robert ; Friedman, Jerome: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2. Auflage. Springer, 2009 (Springer Series in Statistics). – ISBN 9780387848587

[15] James, Gareth ; Witten, Daniela ; Hastie, Trevor ; Tibshirani, Robert: *An Introduction to Statistical Learning with Applications in R.* Springer, 2013 (Springer Series in Statistics). – ISBN 9781461471387

[16] Marron, J. S. ; Ramsay, James O. ; Sangalli, Laura M. ; Srivastava, Anuj: Functional Data Analysis of Amplitude and Phase Variation. In: *Statistical Science* 30 (2015), S. 468 – 484. – Online erhältlich unter https://arxiv.org/pdf/1512.03216.pdf

[17] Pöllath, Nadja ; Alibert, Paul ; Schafberg, Renate ; Peters, Joris: Striking new paths - Distinguishing ancient Ovis orientalis from its modern domestic descendant (Karakul breed) applying Geometric and Traditional Morphometric approaches to the astragalus. In: Çakirlar, C. (Hrsg.) ; Berton, R. (Hrsg.) ; Chahoud, J. (Hrsg.) ; Pilaar Birch, S. (Hrsg.): *Archaeozoology of the Near East XII. Proceedings of the Archaeozoology of Southwest Asia and Adjacent Regions Working Group Meeting in Groningen June 10-14 2015.* Barkhuis Publishers, 2018 (erscheinend)

[18] R Core Team: *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing, 2017. https://www.R-project.org/

[19] Ramsay, J. O. ; Wickham, Hadley ; Graves, Spencer ; Hooker, Giles: *fda: Functional Data Analysis*, 2017. https://CRAN.R-project.org/package=fda. – R package version 2.4.7

[20] Ramsay, J.O ; Silverman, B.W: *Functional Data Analysis.* 2.Auflage. Springer, 2005. – ISBN 9780387400808

[21] Rohlf, F. J.: Morphometrics. In: *Annual Review of Ecology and Systematics* 21 (1990), S. 299 – 316. – Online erhältlich unter https://www.researchgate.net/publication/270690558_Morphometrics

[22] Rohlf, F. J. ; Marcus, Leslie F.: A Revolution in Morphometrics. In: *Trends in ecology & evolution* 8 (1993), S. 129 – 132. – Online erhältlich unter https://www.researchgate.net/publication/49756524_A_Revolution_in_Morphometrics

[23] Srivastava, Anuj ; Klassen, Eric ; Joshi, Shantanu H. ; Jermyn, Ian H.: Shape Analysis of Elastic Curves in Euclidean Spaces. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33 (2011), S. 1415 – 1428. – Online erhältlich unter `https://pdfs.semanticscholar.org/5642/5f83e820b7711dd250fc7cc71c8e1bc177b4.pdf`

[24] Srivastava, Anuj ; Wu, Wei ; Kurtek, Sebastian ; Klassen, Eric ; Marron, J. S.: *Registration of Functional Data Using Fisher-Rao Metric*. 2011. – Online erhältlich unter `https://arxiv.org/abs/1103.3817`

[25] Tucker, J. D.: *fdasrvf: Elastic Functional Data Analysis*, 2017. `https://CRAN.R-project.org/package=fdasrvf`. – R package version 1.8.3

[26] von den Driesch, Angela: *A guide to the measurement of animal bones from archaeological sites: as developed by the Institut für Palaeoanatomie, Domestikationsforschung und Geschichte der Tiermedizin of the University of Munich*. Peabody Museum of Archaeology and Ethnology, 1976 (Peabody Museum Bulletins). – ISBN 9780873659505

[27] Zelditch, Miriam ; Swiderski, Donald L. ; Sheets, David ; Fink, William L.: *Geometric Morphometrics for Biologists: A Primer*. Elsevier Academic Press, 2004. – ISBN 0127784608

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorgelegte Bachelorthesis eigenständig und ohne fremde Hilfe verfasst, keine anderen als die angegebenen Quellen verwendet und die den benutzten Quellen entnommenen Passagen als solche kenntlich gemacht habe. Diese Arbeit habe ich bisher keinem anderen Prüfungsamt in gleicher oder vergleichbarer Form vorgelegt. Sie wurde bisher nicht veröffentlicht.

_____ _____

Ort, Datum Unterschrift