Inhaltsverzeichnis

1 Einleitung

Die statistische Analyse komplexer Datensätze erfordert oftmals Modellierungen, die über eine einfache parametrische Verteilung hinausreichen. Eine Klasse von Verfahren, die in einem solchen Kontext häufig Verwendung finden, sind sogenannte Mixture Modelle. Diese setzen sich aus einer endlichen oder unendlichen Anzahl parametrischer Verteilungen zusammen, die verschiedene Eigenschaften der Daten charakterisieren können. Die Anzahl der als Komponenten bezeichneten Verteilungen, welche das Mixture Modell umfassen soll, besitzt erheblichen Einfluss auf die Qualität der Modellierung : eine zu kleine Anzahl führt dazu, dass die Komplexität der Daten nicht adäquat abgebildet werden kann, während eine zu große Zahl eine Überanpassung an die Daten nach sich zieht. Der in der vorliegenden Arbeit betrachtete Ansatz verwendet eine nichtparametrische bayessche Formulierung, in der ein Mixture Modell mit einer unendlichen Anzahl an Komponenten zum Einsatz kommt. Trotz der unendlichen Komplexität des Modells führt die Wahl einer geeigneten Prioriverteilung über die Parameter in der Posterioriverteilung zu einer Konzentration der Wahrscheinlichkeitsmasse auf einige wenige Komponenten, so dass die Bestimmung der Komponentenanzahl implizit aus den Daten heraus erfolgt. Somit wird eine Festlegung auf eine fixe Komponentenanzahl im Vorfeld der Modellierung vermieden, über die in der praktischen Anwendung oftmals keine präzisen a priori Informationen vorliegen. Eine Verteilung, die sich durch gute theoretische Eigenschaften sowie praktische Handhabbarkeit auszeichnet und daher in der nichtparametrischen bayeschen Modellierung breite Anwendung findet, ist der Dirichlet Prozess. Entsprechende Modelle werden in der Literatur als Dirichlet Prozess Mixture (DPM) Modelle [8, 2] bezeichnet.
Da keine analytisch handhabbare Lösung der Posterioriverteilung eines DPM Modells existiert, werden Verfahren zur Approximation der Verteilung benötigt. Hierzu kann auf eine Vielzahl an Verfahren aus der Klasse der Markov Chain Monte Carlo (MCMC) Methoden zurückgegriffen werden, wobei der überwiegende Teil dieser Algorithmen auf Variationen des Gibbs Samplers basiert [7, 15, 16, 17]. Obwohl MCMC Verfahren theoretisch eine beliebig genaue Approximation der wahren Verteilung ermöglichen, benötigen sie in der Anwendung auf multivariaten Datensätzen mit einer hohen Anzahl an Beobachtungen oftmals lange Laufzeiten und schränken die praktische Anwendbarkeit des DPM Modells damit ein. Eine deterministische Alternative zu den MCMC Verfahren stellen variationale Verfahren dar, welche die Berechnung der Posterioriverteilung als Optimierungsproblem formulieren und diese auf einer Familie einfacherer aber mathematisch handhabbarer Verteilungen bestmöglich approximieren. Die Konstruktion einer variationalen Approximation für das DPM Modell wurde von Blei und Jordan [6] entwickelt, wobei anhand einer empirischen Untersuchung ein gegenüber den MCMC Methoden verbessertes Laufzeitverhalten bei vergleichbarer Qualität der gewonnenen Lösung gezeigt werden konnte.
Der Fokus der vorliegenden Arbeit liegt in der Darstellung und Implementierung eines Dirichlet Prozess Mixture Modells mit normalverteilten Komponenten, wobei eine variationale Approximation der Posterioriverteilung zum Einsatz

kommt. Als konkrete Anwendungen werden Dichteschätzung sowie Clustering betrachtet. Ein empirischer Performancevergleich zeigt, dass DPM Modelle mit Kerndichteschätzern konkurrieren können und diesen bei der Anwendung auf höherdimensionale Datensätze überlegen sind. Die Implementierung des DPM Modells erfolgt mittels der Statistiksoftware R.

Die Arbeit gliedert sich wie folgt. Nach einer kurzen Darstellung der Grundlagen bayesscher Statistik und der Mixture Modelle in Kapitel 2 widmet sich Kapitel 3 dem Dirichlet Prozess, wobei zunächst dessen Eigenschaften anhand unterschiedlicher Konstruktionen betrachtet und anschließend die Anwendung als Prioriverteilung in einem DPM Modell aufgezeigt wird. Kapitel 4 präsentiert zunächst einen allgemeinen Überblick über variationale Verfahren zur Approximation von Verteilungen sowie im Anschluss eine konkrete Herleitung eines Algorithmus zur variationalen Approximation eines DPM Modelles mit normalverteilten Komponenten. Kapitel 5 liefert schließlich eine empirische Untersuchung des Modells anhand verschiedener Datensätze.

2 Grundlagen

In diesem Kapitel werden nach einer grundlegenden Ausführung zur bayesschen Statistik Mixture Modelle vorgestellt und in einen Zusammenhang mit bayesscher Modellierung gestellt. Weiterführende Details sowohl zum Bayesschen Ansatz als auch zur Analyse von Mixture Modellen finden sich in [3, 9].

2.1 Bayessche Statistik

Verfolgt man einen bayesschen Modellansatz, so werden die Parameter des Modells als Zufallsvariablen behandelt, die bestimmten Wahrscheinlichkeitsverteilungen folgen. Diese wird in einer Prioriverteilung $p(\theta)$ kodiert und mit der Likelihoodfunktion $f(\theta \mid \mathbf{x}) = p(\mathbf{x} \mid \theta)$, also der Funktion der Parameter bedingt auf die beobachteten Daten $\mathbf{x} = \{x_1, \ldots, x_N\}$, zu einer Posterioriverteilung verknüpft. Die statistische Inferenz erfolgt dann auf der Posterioriverteilung von θ, gegeben die Daten. Nach dem Satz von Bayes ergibt sich diese aus

$$p(\theta \mid \mathbf{x}) = \frac{p(\theta)\, p(\mathbf{x} \mid \theta)}{\int_\Theta p(\theta)\, p(\mathbf{x} \mid \theta)\, \mathrm{d}\theta}. \tag{2.1}$$

Bei der Wahl geeigneter Prioriverteilungen spielen neben der Sicherstellung sinnvoller Modellierungsannahmen Interpretierbarkeit und Berechenbarkeit eine Rolle. Eine Klasse von Verteilungen, die aufgrund ihrer analytischen Handhabbarkeit häufig Verwendung finden, sind sogenannte *konjugierte Prioriverteilungen*. Eine Familie von Prioriverteilungen $p(\theta \mid \lambda) \in \mathcal{F}$ einer Verteilungsfamilie $\mathcal{F}$ wird als konjugiert zur Verteilung $p(\mathbf{x} \mid \theta)$ bezeichnet, falls für alle Beobachtungen $\mathbf{x}$ und Parametervektoren λ die Posterioriverteilung ebenfalls zur

Verteilungsfamilie $\mathcal{F}$ gehört:

$$p(\theta \mid \mathbf{x}, \lambda) \propto p(\theta \mid \lambda)\, p(\mathbf{x} \mid \theta) \propto p\left(\theta \mid \bar{\lambda}\right), \tag{2.2}$$

wobei $\bar{\lambda}$ den aktualisierten Parametervektor bezeichnet, der die Posterioriverteilung eindeutig bestimmt.
Die Posterioriverteilung können wir dazu verwenden, Vorhersagen für einen neue Beobachtung x^* oder allgemeiner für beliebige Funktionen, deren Verteilung von den Modellparametern abhängt, zu treffen. Für N Beobachtungen $\mathbf{x} = \{x_1, \ldots, x_N\}$ erhalten wir als *Vorhersagewahrscheinlichkeit*:

$$p(x^* \mid \mathbf{x}) = \int_\Theta p(x^* \mid \theta)\, p(\theta \mid \mathbf{x})\, \mathrm{d}\theta \tag{2.3}$$

Zur Approximation von Wahrscheinlichkeitsverteilungen wird ein Distanzmaß benötigt, dass die Diskrepanz zwischen Verteilungen quantifiziert. Die *relative Entropie* oder *Kullback-Leibler (KL) Divergenz* zwischen zwei Verteilungen $p(x)$ und $q(x)$ definiert sich als

$$\mathrm{KL}(q \| p) = -\int q(x) \ln\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x \tag{2.4}$$

Eine wichtige Eigenschaft der KL Divergenz folgt mittels der *Jensenschen Ungleichung*, welche den Erwartungswert konvexer Funktionen beschränkt:

$$E[f(x)] \geq f(E[x]), \quad f : \mathcal{X} \to \mathbb{R} \quad \text{konvex} \tag{2.5}$$

Eine Anwendung der Jensenschen Ungleichung auf den Logarithmus von (2.4), der die Eigenschaft der Konkavität erfüllt, zeigt, dass $\mathrm{KL}(q \| p) \geq 0$, wobei $\mathrm{KL}(q \| p) = 0$ genau dann folgt, wenn gilt $p(x) = q(x)$. Bezüglich der einzelnen Beobachtungen $x_1, \ldots, x_n$ treffen wir die Annahme, dass die Reihenfolge ihres Auftretens innerhalb der Beobachtungssequenz $\mathbf{x}$ vernachlässigbar sein soll. Diese Eigenschaft bezeichnet man als *Austauschbarkeit* oder *exchangeability*. Wir stellen die Austauschbarkeit der Verteilung sicher, indem wir die Unabhängigkeit der Datenpunkte für gegebenen Parametervektor θ annehmen:

$$p(x_1, \ldots, x_N \mid \theta) = \prod_{i=1}^{N} p(x_i \mid \theta). \tag{2.6}$$

Die Annahme von Austauschbarkeit impliziert also einen zugrunde liegenden Parameter, eine Priorverteilung über den Parameter sowie die Annahme, dass die Beobachtungen für gegebenen Parameter unabhängig identisch verteilt sind. In Erweiterung der Definition bezeichnet man eine Sequenz von Beobachtungen $\{x_i\}_{i=1}^{\infty}$ als *unendlich austauschbar*, falls jede endliche Teilmenge austauschbar ist. Folgender Satz zeigt, dass sich austauschbare Sequenzen von Zufallsvariablen immer mittels einer Prioriverteilung über einem latenten Parameterraum darstellen lassen.

Satz von De Finetti 2.7. *Für jede unendlich austauschbare Sequenz von Zufallsvariablen* $\{x_i\}_{i=1}^{\infty}, x_i \in \mathcal{X}$*, existiert ein Raum* Θ *mit zugehöriger Verteilungsfunktion* $P(\theta)$*, so dass die gemeinsame Verteilung beliebiger* N *Beobachtungen folgende Mixture Repräsentation besitzt:*

$$p(x_1, \ldots, x_N) = \int_{\Theta} \left(\prod_{i=1}^{N} p(x_i \mid \theta) \right) dP(\theta) \tag{2.8}$$

Beweis. siehe [11]. □

2.2 Mixture Modelle

Mixture Modelle bezeichnen Wahrscheinlichkeitsverteilungen, die sich aus der konvexen Kombination unterschiedlicher als Komponenten bezeichneter Verteilungen zusammensetzen. Unsere Betrachtung beschränkt sich dabei auf Mixture Modelle, in denen sämtliche Komponenten $p(x \mid \phi)$ einer parametrischen Verteilungsfamilie mit unbekanntem Parameter ϕ entstammen. Für ein Mixture Modell mit K Komponenten folgt die Dichte einer Beobachtung x aus der gewichteten Summe der Dichten der K Komponenten :

$$p(x \mid \pi, \phi) = \sum_{k=1}^{K} \pi_k p(x \mid \phi_k) \tag{2.9}$$

Die π_k, $k = 1, \ldots, K$, bezeichnen die Gewichtskoeffizienten, wobei deren Summe auf 1 normalisiert wird.
Im Folgenden sei nun ein Datensatz mit N Beobachtungen $x_1, \ldots, x_N$ gegeben. Da im Mixture Modell nach Voraussetzung jeder der N Datenpunkte genau einer Komponente entstammt, führen wir N Indikatorvariablen $z_i, i = 1, \ldots, N$ ein, wobei z_i die mit der i-ten Beobachtung x_i assoziierte Komponente spezifiziert:

$$z_{ik} = \begin{cases} 1 & \text{falls i-te Beobachtung aus k-ter Komponente gezogen wurde} \\ 0 & \text{sonst} \end{cases}$$

Daraus folgt für die bedingte Verteilung der Indikatorariablen

$$p(z_{ik} \mid \pi) = \pi_k \quad i = 1, \ldots, N. \tag{2.10}$$

Unter Verwendung der Indikatorvariablen und gleichzeitiger Marginalisierung über alle möglichen Zuordnungen gelangen wir für die Dichte einer Beobachtung x_i wiederum zur Darstellung (2.9):

$$p(x_i \mid \pi, \phi) = \sum_{k=1}^{K} p(z_{ik} \mid \pi)\, p(x_i \mid z_{ik}, \phi) = \sum_{k=1}^{K} \pi_k p(x_i \mid \phi_k). \tag{2.11}$$

Wir können die Modellierung alternativ durch folgenden generativen Prozess beschreiben. Dazu definieren wir eine Verteilung

$$G \sim \sum_{k=1}^{K} \pi_k \delta_{\phi_k}, \tag{2.12}$$

wobei δ_{ϕ_k} ein Diracmaß an der Position ϕ_k bezeichnet. Um nun eine Beobachtungssequenz $x_1, \ldots, x_N$ aus dem Mixture Modell zu erzeugen, ziehen wir für jeden Datanpunkt zunächst eine Realisation θ_i, $i = 1, \ldots, N$, aus der Verteilung G, wobei jede Realisation gemäß Gl.(2.12) genau einem Parameter ϕ_k, $k = 1, \ldots, K$ entspricht. Anschließend erzeugen wir die Beobachtung x_i aus der durch θ_i spezifizierten Komponente. Formal wird dieser Prozess folgendermaßen dargestellt:

$$\begin{aligned} x_i \mid \theta_i &\sim p\left(x_i \mid \theta_i\right) \\ \theta_i &\sim G \end{aligned} \tag{2.13}$$

Die Datenpunkte x_i sind somit unabhängig, identisch verteilt bedingt auf einen gegebenen Parameter θ_i, während die Parameter θ_i selbst unabhängig, identisch verteilt bezüglich G sind. Die Verteilung G kann in diesem Zusammenhang als diskrete K-dimensionale Approximation der unendlich dimensionalen Prioriverteilung $P\left(\theta\right)$ im Satz von De Finetti, Gl.(2.8), interpretiert werden.

Eine wichtige Fragestellung betrifft die Wahl der zu verwendenden Komponentenanzahl K. Wird diese a priori fixiert, so spricht man von endlichen Mixture Modellen. Falls keine hinreichenden Prioriinformation über K vorliegen, so können etwa verschiedene Werte $K_1, \ldots, K_n$ eingesetzt und in einem anschließenden Modellselektionsschritt die optimale Anzahl bestimmt werden. Die Selektion kann mittels verschiedener Methoden wie etwa Kreuzvalidierung, AIC oder BIC vorgenommen werden [4]. Eine elegante Alternative zu den parametrischen Modellen besteht darin, eine abzählbar unendliche Zahl an Komponenten zuzulassen und die adäquate Anzahl automatisch ohne Rückgriff auf Modellselektionsverfahren zu bestimmen:

$$p\left(x \mid \pi, \phi\right) = \sum_{k=1}^{\infty} \pi_k p\left(x \mid \phi_k\right) \tag{2.14}$$

Da mit einer unendlichen Anzahl an Komponenten eine unendliche Anzahl an Parametern einhergeht, benötigen wir eine Prioriverteilung über den Raum der durch die Parameter induzierten Wahrscheinlichkeitsmaße. Modelle, die eine unendliche Parameteranzahl beinhalten, bezeichnet man als *nichtparametrisch*. Eine Verteilung, die eine solche nichtparametrische Modellierung ermöglicht und aufgrund ihrer Handhabbarkeit breite Verwendung findet, ist der Dirichlet Prozess. Beim Dirichlet Prozess handelt es sich um eine Verteilung über Verteilungen, so dass dieser mit der Prioriverteilung über die unendlich vielen Parameterwerte identifiziert werden kann. Das resultierende sogenannte Dirichlet Prozess Mixture Modell kann wie folgt formuliert werden:

$$\begin{aligned} x_i \mid \theta_i &\sim p\left(x_i \mid \theta_i\right) \\ \theta_i &\sim G \\ G &\sim \mathrm{DP}\left(\alpha, G_0\right) \end{aligned} \tag{2.15}$$

Obwohl das Modell nun durch die unendliche Zahl an Parametern über unbegrenzte Komplexität verfügt, agiert die Dirichlet Prozess Prioriverteilung in

einem solchen bayesschen Kontext als Regularisierungsterm, der die Komplexiät kontrolliert und ein Overfitting an die Daten verhindert [19].

3 Dirichlet Prozess Mixture Modelle

Im Folgenden werden wir zunächst auf den Dirichlet Prozess eingehen, indem wir neben einer formalen Definition einige grundlegenden Eigenschaften anhand verschiedender äquivalenter Konstruktionen beleuchten. Anschließend betrachten wir als konkrete Anwendung den Einsatz eines Dirichlet Prozesses als Prioriverteilung eines Mixture Modells mit einer unendlichen Anzahl an Komponenten.

3.1 Dirichlet Prozess

Zur Modellierung der in der nichtparametrischen bayesschen Analyse auftretenden unendlich dimensionalen Räume werden stochastische Prozese verwendet, deren Definition oftmals implizit über die Randverteilungen erfolgt. Im Fall des Dirichlet Prozesses erreichen wir dies mittels Spezifikation der Verteilungen auf endlichen Partitionen des Parameterraumes, welche den Prozess eindeutig bestimmen.

Definition 3.1. *Sei $(A_1, \ldots, A_r)$ eine endliche Partition eines messbaren Raumes Ω:*

$$\bigcup_{k=1}^{r} A_k = \Theta \qquad A_k \cap A_l = \varnothing \tag{3.2}$$

*Sei weiterhin $(\Omega, \mathcal{B})$ ein messbarer Raum, G_0 ein Wahrscheinlichkeitsmaß auf diesem Raum und α_0 eine positive reelle Konstante. Dann bezeichnet man die Verteilung eines Wahrscheinlichkeitsmaßes G auf $(\Omega, \mathcal{B})$ als **Dirichlet Prozess**, falls für jede endliche Partition $(A_1, \ldots, A_r)$ von Ω der Zufallsvektor $(G(A_1), \ldots, G(A_r))$ dirichletverteilt ist:*

$$(G(A_1), \ldots, G(A_r)) \sim Dir(\alpha G_0(A_1), \ldots, \alpha G_0(A_r)). \tag{3.3}$$

Die zugrundeliegende Verteilung G_0 bezeichnet man als Basismaß und α_0 als Konzentrationsparameter. Für beliebige Basisverteilungen G_0 und Konzentrationsparameter α existiert ein eindeutiger Prozess, der mit $DP(\alpha, G_0)$ bezeichnet wird.

Beweis. Um die Existenz der in Gl.(3.3) gegebenen Randverteilungen sicherzustellen, muss sich bei Zusammenlegung mehrerer Zellen einer Partition zu einer einzigen Zelle deren Wahrscheinlichkeit aus der Summe der Einzelwahrscheinlichkeiten der involvierten Zellen ergeben. Die Additivitätseigenschaft der Dirichletverteilung liefert eine Möglichkeit, diese Konsistenzeigenschaft zu gewährleisten. (siehe Appendix A, speziell Gl.(A.16)). Technische Details finden sich in [8]. □

Für eine beliebige meßbare Menge $A \subset \Omega$ folgt $\left(G(A), G\left(\bar{A}\right)\right)$ nach obiger Definition des Dirichlet Prozesses einer Betaverteilung (siehe Gl.(A.7)) mit Parametern $\alpha G_0(A)$ und $\alpha G_0\left(\bar{A}\right)$, wobei $\bar{A}$ das Komplement von A bezeichnet. Wir erhalten also für Erwartungswert und Varianz:

$$E[G(A)] = G_0(A) \tag{3.4}$$

$$V[G(A)] = \frac{G_0(A)(1 - G_0(A))}{(\alpha + 1)} \tag{3.5}$$

G_0 nimmt somit die Rolle des Erwartungswertes des Prozesses $DP(\alpha, G_0)$ ein, während α als inverse Varianz interpretiert werden kann. Mit zunehmendem α konzentriert der Dirichlet Prozess mehr Wahrscheinlickeitsmasse um den Mittelwert und konvergiert schließlich im Limes $\alpha \to \infty$ punktweise gegen das Basismaß G_0.

Um den Dirichlet Prozess sinnvoll in eine bayessche Modellierung integrieren zu können, benötigen wir eine Darstellung der Posterioriverteilung für eine gegebene Beobachtungssequenz $\theta_1, \ldots, \theta_N$. Betrachten wir zunächst eine Situation mit einer einzelnen Beobachtung. Zunächst ziehen wir eine Beobachtung $G \sim DP(\alpha, G_0)$ aus einem Dirichlet Prozess und anschließend aus dieser Verteilung eine Beobachtung $\theta \sim G$. Für beliebige fixe Partition $G(A_1), \ldots, G(A_r)$ impliziert Gl.(3.3) eine korrespondierende Dirichletverteilung. Da die Zuordnung der Beobachtung zu einer Zelle einer Multinomialverteilung mit Parametern proportional zu $G(A_1), \ldots, G(A_r)$ folgt, erhalten wir aus der Konjugiertheit zwischen Multinomial- und Dirichletverteilung als Posterioriverteilung wiederum eine Dirichletverteilung (siehe Gl.(A.18)):

$$(G(A_1), \ldots, G(A_r)) \mid \theta \in A_j \sim \\ \operatorname{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_j) + 1, \ldots, \alpha G_0(A_r)). \tag{3.6}$$

Die Beobachtung θ wirkt sich also nur auf den Parameter der die Beobachtung enthaltende Zelle A_j der Partition aus. Da dies für alle beliebigen Partitionen und damit für beliebig kleine Zellen A_j gilt, folgt daraus, dass die Posteriorverteilung ein Dirac-Punktmaß δ_θ am Beobachtungspunkt besitzt [14]. Als Basismaß der Posterioriverteilung erhalten wir folglich $\alpha G_0 + \delta_\theta / (\alpha + 1)$ sowie als Konzentrationsparameter $\alpha + 1$. Durch wiederholte Anwendung auf jede Beobachtung einer Sequenz gelangen wir nun zu folgendem Satz.

Satz 3.7. *Sei $G \sim DP(\alpha, G_0)$ eine Zufallsvariable, deren Verteilung einem DP folgt. Für N unabhängige Beobachtungen $\theta_1, \ldots, \theta_N \sim G$ folgt das Wahrscheinlichkeitsmaß des Posteriors ebenfalls einem DP:*

$$(G(A_1), \ldots, G(A_r)) \sim DP\left(\alpha + N, \frac{1}{\alpha + N}\left(\alpha G_0 + \sum_{i=1}^{N} \delta_{\theta_i}\right)\right). \tag{3.8}$$

Der Dirichlet Prozess definiert also eine konjugierte Verteilung über Verteilungen auf beliebigen messbaren Räumen und ermöglicht damit die praktische

Anwendbarkeit als Prioriverteilung in der nichtparametrischen bayesschen Modellierung. Ein weiterer bemerkenswerter Punkt ist die Form des Basismaßes im Dirichlet Prozess der Posterioriverteilung als gewichtetes Mittel zwischen der Basisverteilung G_0 und der empirischen Verteilung $\frac{1}{N}\sum_{i=1}^{N}\delta_{\theta_i}$. Der Paramterwert α bestimmt folglich den Anteil, mit dem die Prioriverteeilung im Verhältnis zu den Beobachtungen in das Basismaß einfließt. In einer bayesschen Interpretation kann α auch als Indikator dafür aufgefasst werden, wie stark wir unserer Prioriverteilung vertrauen [19].

Wie wir im Folgenden sehen werden, können wir mittels Marginalisierung über den Dirichlet Prozess zu einer Darstellung gelangen, die keinen direkten Bezug auf den Dirichlet Prozess sondern lediglich auf Realisierungen des Prozesses nimmt. Diese Prozedur ermöglicht die Erzeugung von Beobachtungen aus dem Dirichlet Prozess, ohne diesen explizit repräsentieren zu müssen [5, 19].

3.1.1 Pólya-Urnen Konstruktion

Als Ausgangssituation betrachten wir eine leere Urne, wobei diese mit Kugeln gefüllt werden soll, deren Farbe jeweils einen Parameterwert aus Θ eindeutig repräsentiert. Im ersten Schritt ziehen wir eine Variable $\theta_1 \sim G_0$ und legen eine Kugel in einer den gezogenen Wert repräsentierenden Farbe in die Urne. In der $(n+1)$-ten Iteration wählen wir entweder mit einer Wahrscheinlichkeit $\alpha/(\alpha+n)$ eine neue Farbe aus (sprich wir ziehen einen neuen Wert aus Θ) oder wir nehmen mit der komplementären Wahrscheinlichkeit $n/(\alpha+n)$ zufällig eine Kugel aus der Urne heraus, wobei wir die gewählte Kugel anschließend durch zwei Kugeln der gleichen Farbe ersetzen. Die Wahrscheinlichkeit, eine Kugel einer bereits aufgetretenen Farbe zu ziehen, verhält sich also proportional zur Anzahl Kugeln dieser Farbe, die sich im gegenwärtigen Iterationsschritt in der Urne befinden. Dies impliziert, dass verschiedene Beobachtungen auch für stetige Funktionen G_0 mit strikt positiver Wahrscheinlichkeit identische Werte aufweisen können. Formal ergibt sich also im $(n+1)$-ten Iterationsschritt für ein messbares $A \in \Theta$ folgende Vorhersagedichte:

$$\begin{aligned} P(\theta_{n+1} \in A \mid \theta_1, \ldots, \theta_n) &= E(G(A) \mid \theta_1, \ldots, \theta_n) \\ &= \frac{\alpha}{\alpha+n} G_0(A) + \frac{1}{\alpha+n}\sum_{i=1}^{n}\delta_{\theta_i}(A), \end{aligned} \tag{3.9}$$

wobei die zweite Gleichheit aus dem Basismaß der Posterioriverteilung für gegebene n Beobachtungen in Verbindung mit Gl.(3.4) folgt. Die betrachtete Wahrscheinlichkeit hängt also lediglich von den bisher aufgetretenen Beobachtungen θ_i aus G sowie von G_0, jedoch nicht mehr von dem Dirichlet Prozess selbst ab. Wir schreiben:

$$p(\theta_1 = \theta) = G_0(\theta) \tag{3.10}$$

$$p(\theta_{n+1} = \theta \mid \theta_1, \ldots, \theta_n) = \frac{1}{\alpha+n}\left(\alpha G_0(\theta) + \sum_{i=1}^{n}\delta_{\theta_i}(\theta)\right) \tag{3.11}$$

Um nun zur gemeinsamen Verteilung der θ_i zu gelangen, erhalten wir unter Anwendung der Bayesschen Regel für $N \geq 1$:

$$p(\theta_1, \ldots, \theta_N) = p(\theta_1) \prod_{i=2}^{N} p(\theta_i \mid \theta_1, \ldots, \theta_{i-1}). \tag{3.12}$$

Nehmen die N unabhängig identisch verteilten Beobachtungen $\theta_1, \ldots, \theta_N \sim G$ insgesamt $K \leq N$ verschiedene Werte $\phi_1, \ldots, \phi_K$ an, so bezeichnen wir mit $n_j := \#\{\theta_i = \phi_j\}$ die Anzahl, mit der die jeweiligen Werte $j = 1, \ldots, K$ insgesamt auftreten und erhalten:

$$\begin{aligned}
\prod_{i=1}^{N} p(\theta_i \mid \theta_1, \ldots, \theta_{i-1}) &= \frac{\alpha G_0(\theta_1)}{\alpha} \prod_{i=2}^{N} \frac{\alpha G_0(\theta_i) + \sum_{j=1}^{i-1} \delta_{\theta_j}(\theta_i)}{(\alpha + i - 1)} \\
&= \frac{\left(\alpha G_0\left(\theta_{(1)}\right)\right)^{[n_1]} \left(\alpha G_0\left(\theta_{(2)}\right)\right)^{[n_2]} \cdots \left(\alpha G_0\left(\theta_{(K)}\right)\right)^{[n_K]}}{\alpha^{[n]}} \\
&= p\left(x_{\tau(1)}, \ldots, x_{\tau(N)}\right)
\end{aligned} \tag{3.13}$$

für beliebige Permutation $\tau(\cdot)$, wobei $m^{[n]} := m(m+1)\cdots(m+n-1)$.
Die Positionen der einzelnen Beobachtungen innerhalb der Sequenz üben also keinen Einfluss auf die gemeinsame Verteilung aus. Somit handelt es sich um eine austauschbare Sequenz von Zufallsvariablen, für die nach dem Satz von De Finetti eine Mixture Verteilung G existiert, bezüglich der die Beobachtungen θ_i gezogen wurden (siehe Gl.(2.8)):

$$p(\theta_1, \ldots, \theta_N) = \int_{\Theta} \prod_{i=1}^{N} G(\theta_i)\, dP(G) \tag{3.14}$$

Da nach Konstruktion der Pólya-Urne ein Dirichlet Prozess als Prioriverteilung Verwendung findet, entspricht die Verteilung $P(G)$ genau $DP(\alpha, G_0)$ und stellt damit die Existenz des Dirichlet Prozesses sicher [5, 19].

Die Darstellung der Vorhersagewahrscheinlichkeit in Gl.(3.9) impliziert, dass für jedes messbare $A \in \Theta$ und endliches α im Limes $N \to \infty$ mit Wahrscheinlichkeit 1 gilt:

$$\begin{aligned}
E(G(A) \mid \theta_1, \ldots, \theta_N) &= \frac{1}{\alpha + N} \left(\alpha G_0(A) + \sum_{i=1}^{N} \delta_{\theta_i}(A) \right) \\
&\overset{N \to \infty}{\to} \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}(A) := F,
\end{aligned} \tag{3.15}$$

wobei die ϕ_k die eindeutigen Werte der θ_i und $\pi_k = \lim_{n\to\infty} \frac{n_k}{n}$ die Häufigkeit, mit der die jeweiligen ϕ_k auftreten, bezeichnen. Der Einfluss des kontinuierlichen Teils αG_0 schwindet mit zunehmender Länge der Beobachtungssequenz

$\theta_1, \ldots, \theta_N$ relativ zur empirischen Verteilung, so dass im Limit lediglich eine gewichtete Summe von Punktmaßen verbleibt.

Diese Darstellung illustriert zwei wichtige Eigenschaften des Dirichlet Prozesses: der Prozess ist diskret und er besitzt eine Clustering-Eigenschaft. Das folgende Modell betrachtet die durch den Dirichlet Prozess induzierte Partitionierung der Beobachtung in verschiedene Klassen.

3.1.2 China-Restaurant Prozess

Als Ausgangspunkt betrachten wir analog zur Situation in der Pólya-Urnen Konstruktion eine Menge von N Beobachtungen aus einem Dirichlet Prozess, wobei wir annehmen, dass den Beobachtungen insgesamt K verschiedenen Parameterwerte zugeordnet wurden. Um nun eine Aussage über die Zuordnungswahrscheinlichkeiten einer neuen Beobachtung $N+1$ machen zu können, identifizieren wir zunächst die K beobachteten Parameterwerte mit den Clusterbezeichnungen $1, \ldots, K$ und ordnen jede Beobachtung dem mit ihrem Parameterwert assoziierten Cluster zu. Dazu führen wir N Indikatorvariablen $c_i \in \{1, \ldots, K\}$ ein, wobei c_i die Zuordnung der i-ten Beobachtung kodiert. In Analogie zu Gl.(3.9) können wir für die Vorhersagedichte der Clusterzuordnung schreiben:

$$p\left(\theta_{N+1} = k \mid c_1, \ldots, c_n\right) = \frac{1}{\alpha + N}\left(\alpha\delta_{\bar{i}}\left(k\right) + \sum_{i=1}^{K} n_i \delta_i\left(k\right)\right) \quad k = 1, \ldots, K, \tag{3.16}$$

wobei $\bar{i} := K + 1$ das Label eines Clusters bezeichnet, dem keine Beobachtungen zugeordnet sind.
Diese Verteilung über die Clusterpartitionen ist in der Literatur als *China-Restaurant Prozess* oder *chinese restaurant process* bekannt [19]. Nach dieser Metapher korrespondiert jedes Cluster mit einem Tisch in einem Restaurant, wobei eine unendliche Anzahl an Tischen zur Verfügung stehen, die selbst wieder eine unendliche Sitzkapazizät aufweisen. Kunden, die das Lokal betreten, haben entweder die Möglichkeit sich einem bereits besetzten Tisch anzuschließen oder einen neuen Tisch zu eröffnen. Für erstere Möglichkeit entscheiden sie sich mit einer zu der Anzahl der sich gegenwärtig am Tisch befindlichen Personen proportionalen Wahrscheinlichkeit, für einen neuen Tisch hingegen mit einer Wahrscheinlichkeit proportional zu α. Die Vorliebe der Besucher für große Gruppen, in der englischen Literatur auch als *rich gets richer phenomenon* bezeichnet [19], führt zu einer Verklumpung der Daten, wobei der Grad der Verklumpung durch α bestimmt wird. Große Gruppen implizieren, dass die Anzahl besetzter Tische weit unterhalb der Anzahl der Restaurantbesucher liegt. Wir können die erwartete Anzahl M besetzter Tische für N Besucher konkret bestimmen. Nach Eintreffen des i-ten Besuchers erhöht sich die Anzahl besetzter Tische mit Wahrscheinlichkeit $\alpha/\left(\alpha + i - 1\right)$ um 1. Also folgt für den Erwartungswert

$$E\left(M \mid N\right) = \sum_{i=1}^{N} \frac{\alpha}{\alpha + i - 1} \approx \alpha \ln\left(\frac{N + \alpha}{\alpha}\right). \tag{3.17}$$

Obwohl die Anzahl der Cluster $M \to \infty$ mit Wahrscheinlichkeit 1 für $N \to \infty$, wächst deren erwartete Anzahl also lediglich logarithmisch in der Zahl der Beobachtungen [2]. Für einen Datensatz mit einigen hundert Beobachtungen favorisiert der Dirichlet Prozess somit eine einstellige Anzahl repräsentierter Komponenten [7]. Der China-Restaurant Prozess induziert eine austauschbare Verteilung über die Clusterzuordnung, so dass sich die gemeinsame Verteilung der Clusterzuordnungen invariant gegenüber der Reihenfolge verhält, in der die Beobachtungen den Clustern zugeordnet werden. Diese Eigenschaft lässt sich via Gl.(3.16) analog zur Herleitung der Austauschbarkeitseigenschaft im Pólya-Urnen Modell (3.13) zeigen [11].

Neben den Modellierungen mittels Pólya-Urne und China-Restaurant Prozess, die eine implizite Darstellung des Dirichlet Prozesses basierend auf Beobachtungen aus der nichtparametrischen Verteilung ermöglichen, betrachten wir im Folgenden noch eine explizite Charakterisierung des Prozesses selbst. Die Darstellung führt uns zu einer Approximation des Prozesses durch Beschränkung auf eine endliche Anzahl an Parametern.

3.1.3 Stick-Breaking Konstruktion

Wie bereits im Zusammenhang mit der Darstellung des Dirichlet Prozess über eine Pólya-Urne gesehen (siehe Gl.(3.15)), lässt sich der Prozess immer als eine gewichtete Summe von abzählbar unendlich vielen Punktmaßen darstellen [8]. Eine explizite Konstruktion erfordert also die Bestimmung der Gewichtswerte der einzelnen Komponenten sowie der mit den Komponenten assoziierten Parameter. Dies führt uns zu folgendem Satz:

Stick-Breaking Konstruktion 3.18. *Seien* $\pi(\boldsymbol{v}) = \{\pi_j(\boldsymbol{v})\}_{j=1}^{\infty}$ *sowie* $\phi = \{\phi_j\}_{j=1}^{\infty}$ *unabhängige Sequenzen von unabhängig, identisch verteilten Zufallsvariablen, die mittels folgender Konstruktion generiert werden:*

$$\begin{aligned} v_j &\sim Beta(1, \alpha) \\ \phi_j &\sim G_0 \end{aligned}$$

Definiere nun folgendes Wahrscheinlichkeitsmaß:

$$G(\phi) = \sum_{j=1}^{\infty} \pi_j(\boldsymbol{v})\, \delta_{\phi_j}(\phi), \tag{3.19}$$

wobei

$$\pi_j(\boldsymbol{v}) = v_j \prod_{k=1}^{j-1} (1 - v_k) = v_j \left(1 - \sum_{k=1}^{j-1} \pi_k(\boldsymbol{v})\right). \tag{3.20}$$

Dann folgt G der Verteilung eines Dirichlet Prozesses: $G \sim DP(\alpha, G_0)$. *Umgekehrt gilt: Beobachtungen eines Dirichletprozesses sind mit Wahrscheinlichkeit* 1 *diskret und besitzen eine Darstellung gemäß Gl.(3.19).*

Beweis. Ziehe zunächst eine Realisierung $\phi \sim G$ aus dem Dirichlet Prozess G und betrachte die Partition $(\phi, \Omega \setminus \phi)$ des Raumes Ω. Der Posterioriprozess $G \mid \phi \sim DP\left(\alpha + 1, \frac{G_0 + \delta_\phi}{\alpha + 1}\right)$ impliziert nach Definition

$$\begin{aligned}(G(\phi), G(\Omega \setminus \phi)) \mid \phi &\sim \operatorname{Dir}(\alpha G(\phi) + \delta_\phi(\phi), \alpha G(\Omega \setminus \phi) + \delta_\phi(\Omega \setminus \phi)) \\ &= \operatorname{Dir}(1, \alpha) \\ \Leftrightarrow G(\phi) \mid \phi &\sim \operatorname{Beta}(1, \alpha).\end{aligned}$$

Die letzte Zeile folgt, da die Dirichletverteilung für jede Partition in 2 Teilmengen einer Betaverteilung entspricht.
Entsprechend setzt sich G aus einer Kombination eines Punktmaßes an der Position ϕ sowie eines Wahrscheinlichkeitsmaßes G_1 auf dem komplementären Raum $\Omega \setminus \phi$ zusammen, so dass gilt:

$$G = v\delta_\phi + (1 - v) G_1, \quad v \sim \operatorname{Beta}(1, \alpha). \tag{3.21}$$

Da nach einer grundlegenden Eigenschaft der Dirichletverteilung das um das Punktmaß ϕ reduzierte Maß G_1 mit dem Dirichlet Prozess G auf dem vollständigen Raum Ω übereinstimmmt, also $G_1 \sim DP(\alpha, G_0)$ gilt [11], folgt mittels eines Induktionsargumentes

$$\begin{aligned}G &= v_1\delta_{\phi_1} + (1 - v_1) G_1 \\ G &= v_1\delta_{\phi_1} + (1 - v_1)(v_2\delta_{\phi_2} + (1 - v_2) G_2) \\ &\vdots \\ G &= \sum_{j=1}^{\infty} \pi_j(\mathbf{v}) \delta_{\phi_j}\end{aligned}$$

mit $G_i \sim DP(\alpha, G_0), i \in \mathbb{N}$. Weiterhin erhält man mit $v_j \in [0, 1]$:

$$1 - \sum_{j=1}^{K} \pi_j(\mathbf{v}) = \prod_{j=1}^{K} (1 - v_j) \overset{K \to \infty}{\to} 0 \tag{3.22}$$

mit Wahrscheinlichkeit 1, so dass Gl.(3.19) der Definition eines Wahrscheinlichkeitsmaßes genügt. Für weitere Details siehe [18, 20]. □

Der Name legt bereits eine Analogie der Konstruktion zum Zerbrechen eines Stabes nahe. Nach dieser Metapher beginnen wir mit einem Stab der Länge 1, wobei dieser an einer zufälligen Stelle zerbrochen wird. Das abgebrochene Stück wird abgelegt und das verbliebene Stück wiederum an einer zufälligen Stelle zerbrochen. Dieser Vorgang wird nun sukzessive wiederholt. Die Länge des abgebrochenen Stückes der j-ten Iteration definiert dann das Komponentengewicht π_j der j-ten Komponente. Im Gegensatz zur Konstruktion mittels Pólya-Urne oder China-Restaurant Prozess erfordert die Charakterisierung von G eine Bestimmung der unendlich vielen Parameter ϕ_k und π_k. Zur Generierung von Zufallszahlen aus einem Dirichlet Prozess approximieren wir nun den

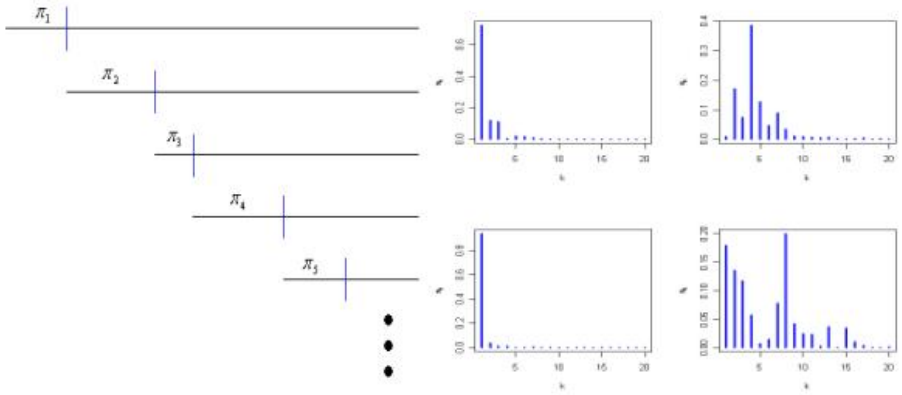

Abbildung 1: Links: Graphische Darstellung des Stick-Breaking Prozesses. Der oberste Stab besitzt die Länge 1, $v_1 \sim \text{Beta}(1, \alpha)$ entspricht dem Gewicht der ersten Komponente. Jedes weitere Gewicht v_k ergibt sich aus einer zufälligen Bruchstelle des verbleibenden Stabes der Länge π_k. Rechts: Darstellung von je 2 zufälligen Realisationen v_k aus einem Stick-Breaking Prozess mit $V \sim \text{Beta}(1, 0.5)$ in der linken Spalte sowie $V \sim \text{Beta}(1, 5)$ in der rechten Spalte. Wie sich gut erkennen lässt, setzt sich der Prozess mit höherem α aus einer größeren Menge signifikanter Gewichten zusammen.

Prozess durch eine endlichdimensionale Summe, indem wir nur die ersten K Bruchstellen berücksichtigen:

$$G_K = \sum_{i=1}^{K} \pi_i \delta_{\phi_i}(\cdot) \tag{3.23}$$

Dieser endliche Prozess wird als *trunkierter Stick-Breaking Prozess* [12] bezeichnet. Durch die Wahl eines entsprechenden Wertes K kann der Stick-Breaking Prozess somit beliebig gut approximiert werden. Ishwaran und James [13] haben gezeigt, dass eine Approximation mittels trunkiertem Stick-Breaking Prozess G_K mit Wahrscheinlichkeit 1 gegen einen Dirichlet Prozess konvergiert, wobei die Variation zwischen Prozess und Approximation von der Ordnung $\exp(-(K-1)/\alpha)$ ist. Da die Bruchstellen $v_k \sim \text{Beta}(1, \alpha)$ durch betaverteilte Zufallsvariablen mit Erwartungswert $E(v_k) = 1/(1+\alpha)$ bestimmt werden, erreichen die kumulierten Gewichtswerte für kleine Parameterwerte α typischerweise bereits nach wenigen Iterationen einen Wert nahe Eins und konzentrieren Zufallszahlen aus dem Dirichlet Prozess auf wenige Komponenten. Wählt man hingegen einen hohen Wert für α, so erzeugt $\text{Beta}(1, \alpha)$ approximativ gleichverteilte Gewichtswerte.

3.2 Unendliche Gaußsche Mixture Modelle

Nach der Charakterisierung des Dirichlet Prozesses kehren wir nun nochmals zu den Ausführungen über Mixture Modelle in Kapitel 2 zurück. Verwenden wir für $DP(\alpha, G_0)$ die Stick-Breaking Konstruktion, so können wir das Dirichlet Prozess Mixture Modell aus Gl.(2.15) wie folgt formuliere:

$$\begin{aligned} x_i \mid \theta_i &\sim p(x_k \mid \theta_i) \\ \theta_i &\sim G \\ G &\sim \textstyle\sum_{j=1}^{\infty} \pi_j \delta_{\phi_j} \end{aligned} \tag{3.24}$$

Dies entspricht der Spezifikation eines Mixture Modells in Gl.(3.19) mit dem Unterschied, dass der Verteilung G nun eine unendliche Anzahl an Komponenten zugrunde liegen (vergleiche dazu mit Gl.(2.12)). Entsprechend gelangen wir zu einem Mixture Modell der folgenden Form:

$$p(x_n \mid \pi, \phi) = \sum_{k=1}^{\infty} \pi_k p(x_n \mid \phi_k). \tag{3.25}$$

Wir konkretisieren das Modell nun durch die Annahme normalverteilter Mixture-Komponenten. Die gewählte Spezifikation der weiteren Modellparameter bedient sich konjugierter Verteilungen und stellt eine Standardkonfiguration in der Anwendung von Dirichlet Prozess Mixture Modellen mit normalverteilten Komponenten dar [7, 17].

Zur Modellierung des Dirichlet Prozesses verwenden wir die Stick-Breaking Konstruktion, wobei $\pi(\mathbf{v})$ den unendlich dimensionalen Vektor der Gewichtswerte und $\{\theta_1, \theta_2, \ldots\}$ die entsprechenden Punktmaße der mit den Komponenten assoziierten Parameter bezeichnet. Wir definieren zur Modellierung wiederum Indikatorvariablen

$$z_{nk} = \begin{cases} 1 & \text{falls i-te Beobachtung aus k-ter Komponente gezogen wurde} \\ 0 & \text{sonst} \end{cases}$$

und erhalten daraus folgende bedingte Verteilung der Indikatorvariablen für gegebene Gewichtswerte $\pi(\mathbf{v})$:

$$p(\mathbf{Z} \mid \pi(\mathbf{v})) = \prod_{n=1}^{N} \prod_{k=1}^{\infty} \pi_k(\mathbf{v})^{z_{nk}}. \tag{3.26}$$

Die Verteilung der n-ten Indikatorvariable z_n bezeichnen wir mit $p(z_n \mid \pi(\mathbf{v})) \sim$ Diskret $(\pi(\mathbf{v}))$.
Zur Spezifikation der normalverteilten Komponenten benötigen wir jeweils Mittelwertvektor sowie Kovarianzmatrix, entsprechend setzt sich der Vektor $\theta = \{\theta_1, \theta_2, \ldots\}$ aus einer abzählbar unendlichen Anzahl an Parametern zusammen, wobei $\theta_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ den Mittelwertvektor sowie die Kovarianzmatrix der k-ten Komponente bezeichnet. Die gemeinsame Verteilung der N Beobachtungen

ergibt folglich:

$$p\left(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta}\right) = \prod_{n=1}^{N} \prod_{k=1}^{\infty} \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\theta}_k\right)^{z_{nk}} \tag{3.27}$$

In einer bayesschen Modellierung wird eine Prioriverteilung über die Parameter $\theta_k \sim G_0$ benötigt. Wir wählen eine bedingte Normal-Wishartverteilung, wobei wir aus Gründen übersichtlicher Notation im Folgenden statt der Kovarianzmatrix Σ deren Inverse $\Lambda := \Sigma^{-1}$, die als *Präzisionsmatrix* bezeichnet wird, verwenden:

$$\begin{aligned} G_0 &= p\left(\boldsymbol{\mu}, \boldsymbol{\Lambda}\right) = p\left(\boldsymbol{\mu} \mid \boldsymbol{\Lambda}\right) p\left(\boldsymbol{\Lambda}\right) \\ &= \mathcal{N}\left(\boldsymbol{\mu} \mid \boldsymbol{m}_0, \left(\beta_0 \boldsymbol{\Lambda}\right)^{-1}\right) \mathcal{W}\left(\boldsymbol{\Lambda} \mid \mathbf{W}_0, \nu_0\right) \end{aligned} \tag{3.28}$$

Da die Prioriverteilung eine zur Normalverteilung konjugierte Form besitzt [9], folgt die Posterioriverteilung wiederum einer Normal-Wishartverteilung. Entsprechend wird die Posterioriverteilung durch die gleiche Anzahl Parameter wie die Prioriverteilung bestimmt, so dass eine effiziente Berechenbarkeit der Posterioriverteilung gewährleistet wird.
Fügen wir die Spezifikationen zusammen, gelangen wir zu folgendem Modell:

$$\begin{aligned} V_k \mid \alpha &\sim \text{Beta}\left(1, \alpha\right) \\ \theta_k \mid G_0 &\sim G_0 \\ z_i \mid \{v_1, v_2, \ldots\} &\sim \text{Diskret}\left(\pi\left(\mathbf{v}\right)\right) \\ x_i \mid z_i = k, \theta_k &\sim \mathcal{N}\left(x_i \mid \theta_k\right) \\ i &= 1, \ldots, N \\ k &= 1, 2, \ldots \end{aligned} \tag{3.29}$$

Betrachten wir das Modell als generativen Prozess zur Erzeugung von Beobachtungen, so ziehen wir zunächst eine unendliche Anzahl an Gewichten V_k sowie Komponentenparameter θ_k. Anschließend erzeugen wir den Datensatz, indem wir für jede Beobachtung zunächst die Komponente auswählen, aus der diese entstammen soll und daraufhin eine Beobachtung aus der Normalverteilung der entsprechenden Komponente $z_n = k$ ziehen.

3.3 Variationale Approximation in Dirichlet Prozess Mixture Modellen

Zwar ist die Posterioriverteilung des Modells (3.29) prinzipiell analytisch lösbar, dies erfordert jedoch eine Summation über alle möglichen Zuordnungskonfigurationen der Beobachtungen zu den Komponenten und entzieht sich somit schon für kleine Datensätze einer praktischen Handhabung [2, 15]. Somit muss zu deren Auswertung auf approximative Verfahren zurückgegriffen werden, wobei wir zwischen stochastischen und deterministischen Verfahren unterscheiden.
Im ersteren Fall finden meist MCMC-Verfahren Verwendung. MCMC-Methoden

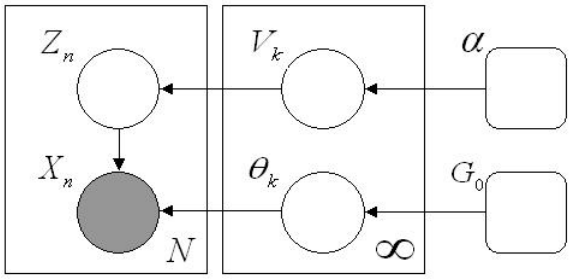

Abbildung 2: Graphische Darstellung des Modells. Die beobachteten Variablen sind grau schattiert, die Parameter, deren Posterioriverteilung wir bestimmen wollen, werden durch die ungefüllten Kreise repräsentiert. Die großen Rechtecke bezeichnen ein mehrfaches Auftreten der umschlossenen Variablen, wobei deren konkrete Anzahl durch den Wert in der rechten unteren Ecke gegeben wird.

generieren eine Markovkette, deren stationäre Verteilung der gesuchten Posterioriverteilung entspricht. Nach erfolgter Konvergenz können Beobachtungen aus der Markovkette generiert und zur Bildung empirischer Statistiken wie etwa verschienenen Momenten oder der Verteilungsfunktion verwendet werden. Die einfachste Methode aus der Klasse der MCMC-Verfahren ist der Gibbs-Sampler, der eine Markovkette durch iteratives Sampling aus den auf die Daten sowie die zuletzt erzeugten Werte aller übrigen Parameter bedingten Verteilungen erzeugt. Zur Lösung eines Dirichlet Prozess Mixture Modells existieren verschiedene Gibbs-Sampling Algorithmen, die in [16] überblicksartig vorgestellt werden. Obwohl theoretisch bei unendlicher Rechenkapazität eine exakte Lösung der Posterioriverteilung gefunden werden kann, wird der praktische Nutzen von MCMC Verfahren vor allem im Kontext großer und multivariater Datensätze durch die oftmals langen Laufzeiten eingeschränkt.
Alternativ dazu besteht die Möglichkeit, die Approximation als Optimierungsproblem zu formulieren, indem die Posterioriverteilung durch eine simplere aber gleichzeitig handhabbare Verteilung approximiert und als Schranke der wahren Verteilung aufgefasst wird. Die zur Approximation in Betracht kommenden Verteilungen werden über eine Familie zulässiger Verteilungen spezifizert. Die beste Lösung wird dann von demjenigen Mitglied der Verteilungsfamilie erzielt, dessen Schranke am dichtesten am Wert der wahren Verteilung liegt.

Wir betrachten im Folgenden die Approximation mittels variationaler Methoden konkreter und wenden diese im Anschluß auf unser Modell (3.29) an. Der Überblick zu variationalen Verfahren folgt den Ausführungen von [4].

3.3.1 Variationale Methoden

Als Ausgangspunkt betrachten wir ein bayessches Modell mit N vorliegenden Beobachtungen $\mathbf{X} = \{x_1, \ldots, x_N\}$. Die Menge aller im Modell auftretenden Parameter bezeichnen wir mit $\mathbf{Z} = \{z_1, \ldots, z_M\}$. Gesucht wird nun im Folgenden eine Verteilung $q(\mathbf{Z})$, welche die Posterioriverteilung $p(\mathbf{Z} \mid \mathbf{X})$ des Modells bestmöglichst approximiert, wobei wir als Ähnlichkeitsmaß der beiden Verteilungen die Kullback-Leibler Divergenz (2.4) verwenden. Die Verteilung $q(\mathbf{Z})$ wird in einem solchen Zusammenhang als *variationale Verteilung* bezeichnet.
Um die optimale variationalen Verteilung zu bestimmen, schreiben wir zunächst die logarithmierte Randverteilung über die Beobachtungen

$$\ln p(\mathbf{X}) = \ln \int p(\mathbf{X}, \mathbf{Z}) \, \mathrm{d}\mathbf{Z} \tag{3.30}$$

in Abhängigkeit der Verteilung $q(\mathbf{Z})$ und erhalten nach einigen elementaren Umformungen:

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q \| p), \tag{3.31}$$

wobei wir folgende Definitionen verwenden:

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left(\frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right) \mathrm{d}\mathbf{Z} \tag{3.32}$$

$$\mathrm{KL}(q \| p) = -\int q(\mathbf{Z}) \ln \left(\frac{p(\mathbf{Z} \mid \mathbf{X})}{q(\mathbf{Z})} \right) \mathrm{d}\mathbf{Z} \tag{3.33}$$

Der Term $\mathrm{KL}(q \| p)$ bezeichnet die Kullback-Leibler Divergenz zwischen den Verteilungen $p(\mathbf{Z} \mid \mathbf{X})$ und $q(\mathbf{Z})$ und entspricht damit genau dem zu minimierenden Ausdruck, wobei die Optimierung bezüglich der variationalen Verteilung erfolgt. Da die linke Seite der Gleichung (3.31) unabhängig von $q(\mathbf{Z})$ ist, können wir statt einer Minimierung der Kullback-Leibler Divergenz äquivalent eine Maximierung des ersten Terms der rechten Seite $\mathcal{L}(q)$ betrachten. Wegen $\mathrm{KL}(q \| p) \geq 0$ folgt $\ln p(\mathbf{X}) \geq \mathcal{L}(q)$, so dass $\mathcal{L}(q)$ die logarithmierte Randverteilung der Beobachtungen nach unten beschränkt.
Falls wir die Familie der variationalen Verteilungen, die zur Lösung der Optimierung in Betracht kommt, keinerlei Einschränkungen unterziehen, erhalten wir unmittelbar aus der Kullback-Leibler Divergenz als Lösung die wahre Posterioriverteilung $q(\mathbf{Z}) = p(\mathbf{Z} \mid \mathbf{X})$. Da sich im Fall des DPM Modells die wahre Posterioriverteilung analytisch nicht handhaben lässt, beschränken wir die Optimierung im Folgenden auf eine Familie variationaler Verteilungen, deren Mitglieder analytisch lösbar sind und die zugleich eine möglichst gute Approximation der wahren Verteilung ermöglicht. Wir betrachten dazu konkret diejenige Verteilungsfamilie, die zwischen den Verteilungen der einzelnen Parameter faktorisiert:

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i) \tag{3.34}$$

Die funktionale Form der einzelnen Verteilungen $q_i(\mathbf{Z}_i)$ unterliegt hingegen keinerlei Einschränkungen. In der Literatur bezeichnet man diese faktorisierte Form von Verteilungen als *Mean-Field Gleichungen*.

Zur Lösung der Optimierung bezüglich der faktorisierten Verteilungsfamilie maximieren wir nun die Unterschranke $\mathcal{L}(q)$ über sämtliche Verteilungen $q(\mathbf{Z})$. Die Unabhängigkeit der einzelnen Faktoren ermöglicht eine separate Optimierung bezüglich jedes einzelnen Faktors $q_i(\mathbf{Z}_i)$. Wir setzen dazu (3.34) in die Unterschranke (3.32) ein und betrachten die resultierende Gleichung als Funktion des j-ten Faktors $q_j(\mathbf{Z}_j)$:

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_{i=1}^{M} q_i(\mathbf{Z}_i) \left(\ln p(\mathbf{X}, \mathbf{Z}) - \sum_{i=1}^{M} \ln q_i(\mathbf{Z}_i) \right) \mathrm{d}\mathbf{Z} \\
&\propto \int q_j(\mathbf{Z}_j) \left(\int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i(\mathbf{Z}_i) \, \mathrm{d}\mathbf{Z}_i \right) \mathrm{d}\mathbf{Z}_j - \int q_j(\mathbf{Z}_j) \ln q_j(\mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j \\
&= \int q_j(\mathbf{Z}_j) \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j - \int q_j(\mathbf{Z}_j) \ln q_j(\mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j \\
&= \int q_j(\mathbf{Z}_j) \ln \left(\frac{\tilde{p}(\mathbf{X}, \mathbf{Z}_j)}{q_j(\mathbf{Z}_j)} \right) \mathrm{d}\mathbf{Z}_j, \qquad (3.35)
\end{aligned}$$

wobei die Verteilung $\ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) = E_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})]$ den Erwartungswert bezüglich der Verteilung q über alle Variablen $\mathbf{Z}_i, i \neq j$ bezeichnet. Der Ausdruck in der letzten Zeile entspricht genau der negativen Kullback-Leibler Divergenz zwischen der variationalen Verteilung $q_j(\mathbf{Z}_j)$ und $\tilde{p}(\mathbf{X}, \mathbf{Z}_j)$. Somit fällt das Maximum der Unterschranke $\mathcal{L}(q)$ bezüglich der Verteilung $q_j(\mathbf{Z}_j)$ mit dem Minimum der Kullback-Leibler Distanz zusammen, wobei das Minimum für $q_j(\mathbf{Z}_j) = \tilde{p}(\mathbf{X}, \mathbf{Z}_j)$ erreicht wird. Also ergibt sich die optimale Lösung der variationalen Verteilung proportional zu $\tilde{p}(\mathbf{X}, \mathbf{Z}_j)$:

$$\ln q_j^*(\mathbf{Z}_j) \propto E_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] \qquad (3.36)$$

In einem letzten Schritt führen wir eine Normalisierung $\int q_j^*(\mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j = 1$ durch und erhalten schließlich als Verteilung der optimalen variationalen Verteilung des j-ten Faktors

$$q_j^*(\mathbf{Z}_j) = \frac{\exp\left(E_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right)}{\int \exp\left(E_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right) \mathrm{d}\mathbf{Z}_j} \qquad (3.37)$$

Die M Gleichungen (3.37), $j = 1, \ldots, M$, definieren gemeinsam das Optimum der Unterschranke bezüglich der faktorisierten variationalen Verteilung $q(\mathbf{Z})$. Da die Lösung des j-ten Faktors $q_j^*(\mathbf{Z}_j)$ jeweils vom Erwartungswert bezüglich aller übrigen Faktoren $\mathbf{Z}, i \neq j$ abhängt und die Gleichungen somit gekoppelt sind, kann keine direkte Optimierung vorgenommen werden. Aufgrund der Konvexität der Schranke (3.35) kann jedoch gezeigt werden, dass der Algorithmus

gegen ein lokales Optimum konvergiert.

In der Praxis bietet der Einsatz variationaler Verfahren einige Vorteile gegenüber dem Gibbs Sampling. Neben dem im Vergleich zum Gibbs Sampling oftmals überlegenem Laufzeitverhalten bei der Lösung von Dirichlet Prozess Mixture Modellen kann die Konvergenz des Algorithmus durch Auswertung der Unterschranke konkret festgestellt werden. Die Überwachung der Konvergenz einer Markovkette auf die stationäre Verteilung erweist sich hingegen im Gibbs Sampling als schwierigeres Problem, für das in der Regel auf empirische Verfahren zurückgegriffen werden muss [1]. Demgegenüber besteht ein Nachteil variationaler Verfahren darin, dass die gefundene Lösung lediglich ein lokales Optimum bildet, die unter Umständen deutlich vom globalen Optimum abweichen kann. Die Qualität der Approximation bemisst sich problemabhängig danach, wie stark die Annahme faktorisierter Parameterverteilungungen von der Form der wahren Posterioriverteilung abweicht. Ein empirischer Vergleich in [6] zwischen der Approximation mittels variationaler Methoden sowie Gibbs-Sampling Algorithmen in Dirichlet Prozess Mixture Modellen zeigt, dass erstere Verfahren bei besserem Laufzeitverhalten mit dem Gibbs-Sampling vergleichbare Approximationen liefern.

Im Folgenden verwenden wir die vorgestellten Mean-Field Gleichungen zur Approximation der Posterioriverteilung des Dirichlet Prozess Mixture Modells (3.29).

3.3.2 Variationales Gaußsches Dirichlet Prozess Mixture Modell

Wir betrachten zunächst die gemeinsame Verteilung der Beobachtungen und Parameter, die vom Modell (3.29) induziert wird. Wir erhalten folgende Struktur:

$$p\left(\mathbf{X}, \mathbf{Z}, \mathbf{V}, \boldsymbol{\theta}\right) = p\left(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta}\right) p\left(\mathbf{Z} \mid \mathbf{V}\right) p\left(\boldsymbol{\theta}\right) p\left(\mathbf{V}\right) \tag{3.38}$$

Im Folgenden suchen wir die beste Approximation dieser Verteilung aus der Familie der Mean-Field Verteilungen (3.34):

$$q\left(\mathbf{Z}, \mathbf{V}, \boldsymbol{\theta}\right) = q\left(\mathbf{Z}\right) q\left(\mathbf{V}\right) q\left(\boldsymbol{\theta}\right) \tag{3.39}$$

Die variationale Verteilung beruht analog zu Modell (3.29) auf der Stick-Breaking Konstruktion des Dirichlet Prozesses, wir verwenden jedoch einen trunkierten Stick-Breaking Prozess (3.23), um eine endlich dimensionale Repräsentation und somit eine endliche Anzahl an Parametern zu erhalten. Wir fixieren dazu einen Wert K und setzen den Wert des K-ten Gewichts auf 1, $q\left(v_K = 1\right) = 1$. Diese Repräsentation impliziert, dass der verbleibende Stab nach der $(K-1)$-ten Unterteilung nicht mehr weiter zerbrochen wird, sondern als Gewicht der K-ten Komponente interpretiert wird. Dementsprechend haben alle Gewichte $\pi_k\left(\mathbf{v}\right)$ den Wert Null für $k > K$ (siehe 3.20). Wir können die Familie der Mean-Field Verteilungen damit genauer spezifizieren:

$$q\left(\mathbf{Z}, \mathbf{V}, \boldsymbol{\theta}\right) = \prod_{i=1}^{N} q\left(z_i\right) \prod_{k=1}^{K-1} q\left(v_k\right) \prod_{k=1}^{K} q\left(\theta_k\right) \tag{3.40}$$

Gemäß Gleichung (3.32) können wir die logarithmierte Randverteilung $\ln p(\mathbf{X}) \geq \mathcal{L}(q)$ über die Beobachtungen nach unten beschränken, wobei wir für die Schranke folgenden Ausdruck erhalten:

$$\begin{aligned}\mathcal{L}(q) &= \sum_{\mathbf{Z}} \iiint q(\mathbf{Z}, \mathbf{V}, \boldsymbol{\theta}) \ln\left(\frac{p(\mathbf{X}, \mathbf{Z}, \mathbf{V}, \boldsymbol{\theta})}{q(\mathbf{Z}, \mathbf{V}, \boldsymbol{\theta})}\right) \mathrm{d}\mathbf{V}\mathrm{d}\boldsymbol{\theta} \\ &= E_q[\ln p(\mathbf{X} \mid \mathbf{Z}, \mathbf{V}, \boldsymbol{\theta})] + E_q[\ln p(\mathbf{Z} \mid \mathbf{V})] + E_q[\ln p(\boldsymbol{\theta})] + E_q[\ln p(\mathbf{V})] \\ &\quad - E_q[\ln q(\mathbf{Z})] - E_q[\ln q(\mathbf{V})] - E_q[\ln q(\boldsymbol{\theta})] \end{aligned} \tag{3.41}$$

Die zweite Zeile folgt unter Berücksichtigung der Dekomposition von p und q gemäß Gleichung (3.38) respektive (3.39). Die Auswertung der Schranke erfolgt nun durch eine Optimierung bezüglich der variationalen Parameter, wobei sich diese gemäß Gleichung (3.37) berechnen. Wir betrachten zunächt den optimalen Faktor der Indikatorvariablen $q(\mathbf{Z})$. Der Logarithmus der optimalen variationalen Verteilung ergibt sich in unnormalisierter Form (3.36) als

$$\ln q^*(\mathbf{Z}) \propto E_{q(\mathbf{V},\boldsymbol{\theta})}[\ln p(\mathbf{X}, \mathbf{Z}, \mathbf{V}, \boldsymbol{\theta})] \tag{3.42}$$

Aufgrund der Struktur der Verteilung p gemäß Gleichung(3.38) und dem Umstand, dass alle funktional nicht von $\mathbf{Z}$ abhängigen Terme in die Normalisierungskonstante absorbiert werden können, schreiben wir

$$\ln q^*(\mathbf{Z}) \propto E_{q(\boldsymbol{\theta})}[\ln p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta})] + E_{q(\mathbf{V})}[\ln p(\mathbf{Z} \mid \mathbf{V})]. \tag{3.43}$$

Der erste Term der rechten Seite ergibt sich aus dem Erwartungswert von (3.27), den zweiten Term betrachten wir für eine einzelne Beobachtung Z_n und erhalten mittels Gleichung (3.20)

$$\begin{aligned} E_{q(\mathbf{V})}[\ln p(Z_n \mid \mathbf{V})] &= E_{q(\mathbf{V})}\left[\ln\left(\prod_{k=1}^{\infty}(1-v_k)^{\mathbf{1}[z_n>k]}\, v_k^{\mathbf{1}[z_n=k]}\right)\right] \\ &= \sum_{k=1}^{K}\sum_{i=k+1}^{K} z_{ni} E_{q(\mathbf{V})}[\ln(1-v_k)] + z_{nk} E_{q(\mathbf{V})}[\ln v_k], \end{aligned} \tag{3.44}$$

Der Übergang zur letzten Zeile folgt, da unter der variationalen Verteilung ein trunkierter Stick-Breaking Prozess mit K Komponenten Verwendung findet und dementsprechend nur die ersten K Bruchstellen $v_1, \ldots, v_K$ berücksichtigt werden.
Wir absorbieren wiederum die von $\mathbf{Z}$ unabhängigen Terme in die Normalisierungskonstante und gelangen zu

$$\ln q^*(\mathbf{Z}) \propto \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \ln \rho_{nk} \tag{3.45}$$

mit

$$\ln \rho_{nk} = \frac{1}{2} E_{q(\mathbf{\Lambda})} \left[\ln |\mathbf{\Lambda}_k| \right] - \frac{D}{2} \ln (2\pi) - \frac{1}{2} E_{q(\boldsymbol{\mu}, \mathbf{\Lambda})} \left[(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \mathbf{\Lambda}_k (\mathbf{x}_n - \boldsymbol{\mu}_k) \right] + E_{q(\mathbf{V})} \left[\ln v_k \right] + \sum_{i=1}^{K-1} E_{q(\mathbf{V})} \left[\ln (1 - v_i) \right] \tag{3.46}$$

Da die Verteilung der Indikatorvariable $q(z_{nk}), k = 1, \ldots, K$ die Wahrscheinlichkeit angibt, daß die Beobachtung x_n der k-ten Komponente entstammt, muss gelten $\sum_{k=1}^{K} q(\mathbf{Z}_n = k) = 1$. Zu beachten ist, dass im Fall der variationalen Verteilung im Gegensatz zur wahren Verteilung $p(z_{nk}), k = 1, 2, \ldots$ nur eine fixe endliche Anzahl K an Komponenten zur Auswahl stehen. Unter Anwendung der Exponentialfunktion folgt

$$q^*(\mathbf{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}} \tag{3.47}$$

mit Normalisierung

$$r_{nk} = \frac{\rho_{nk}}{\sum_{i=1}^{K} \rho_{ni}}. \tag{3.48}$$

Da es sich bei z_{nk} um binäre Variablen handelt, gilt $E_{q(\mathbf{Z})} [z_{nk}] = r_{nk}$.
Ein Vergleich mit der Prioriverteilung $p(\mathbf{Z} \mid \pi(\mathbf{v}))$ (3.26) zeigt, dass die optimale Approximation $q^*(\mathbf{Z})$ dieselbe Verteilungsform besitzt. Die Lösung $q^*(\mathbf{Z})$ hängt von den Erwartungswerten bezüglich weiterer variationaler Verteilungen ab und kann daher nicht direkt ausgewertet werden.
Wir definieren für den späteren Gebrauch folgende mit $E_{q(\mathbf{Z})} [z_{nk}]$ gewichteten empirische Momente:

$$N_k = \sum_{n=1}^{N} r_{nk} \tag{3.49}$$

$$\overline{\mathbf{x}}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \mathbf{x}_n \tag{3.50}$$

$$\mathbf{S}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\mathbf{x}_n - \overline{\mathbf{x}}_k) (\mathbf{x}_n - \overline{\mathbf{x}}_k)^T \tag{3.51}$$

Als nächste Verteilung betrachten wir die variationale Verteilung der Stabgewichte. Wir verfahren analog zur Vorgehensweise bei der Bestimmung von $q^*(\mathbf{Z})$ und erhalten:

$$\begin{aligned} \ln q^*(\mathbf{V}) &\propto E_{q(\mathbf{Z})} [\ln p(\mathbf{Z} \mid \mathbf{V})] + \ln p(\mathbf{V}) \\ &\propto \sum_{k=1}^{K-1} \sum_{n=1}^{N} \left\{ E_{q(\mathbf{Z})} [z_{nk}] \ln v_k + \sum_{j=k+1}^{K-1} E_{q(\mathbf{Z})} [z_{nj}] \ln (1 - v_k) \right\} \\ &\quad + (\alpha - 1) \sum_{k=1}^{K-1} \ln (1 - v_k), \end{aligned} \tag{3.52}$$

wobei der letzte Term der rechten Seite aus $V_i \sim \text{Beta}\,(1, \alpha)\,, i = 1, \ldots, K-1$ folgt. Wir definieren nun unter Verwendung von $E_{q(\mathbf{Z})}\,[z_{nk}] = r_{nk}$

$$\gamma_{k1} = 1 + \sum_{n=1}^{N} r_{nk} \tag{3.53}$$

$$\gamma_{k2} = \alpha + \sum_{n=1}^{N} \sum_{j=k+1}^{K-1} r_{nj} \tag{3.54}$$

und erhalten

$$\ln q^*\,(\mathbf{V}) \propto \sum_{k=1}^{K-1} (\gamma_{k1} - 1) \ln v_k + (\gamma_{k2} - 1) \ln\,(1 - v_k)\,, \tag{3.55}$$

Wie sich leicht erkennen lässt, folgt

$$q^*\,(\mathbf{V}) = \prod_{k=1}^{K-1} q^*\,(v_k) \tag{3.56}$$

einem Produkt aus Betaverteilungen mit Komponenten

$$q^*\,(v_k) \sim \text{Beta}\,(\gamma_{k1}, \gamma_{k2}) \qquad k = 1, \ldots, K-1 \tag{3.57}$$

Somit besitzen variationale Verteilung und Prioriverteilung $p\,(\mathbf{V})$ wiederum dieselbe funktionale Form, im Gegensatz zur Prioriverteilung weisen die einzelnen Verteilungen V_k im Fall der $q^*\,(\mathbf{V})$ allerdings unterschiedliche Parametrisierungen auf.
In einem letzten Schritt betrachten wir die variationale Verteilung über die Hyperparameter der Komponenten. Es folgt wiederum unter Anwendung der allgemeinen Optimierungsvorschrift (3.36)

$$\begin{aligned} q^*\,(\boldsymbol{\theta}) &= \prod_{k=1}^{K} q^*\,(\boldsymbol{\theta}_k) = \prod_{k=1}^{K} q^*\,(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) \\ &\propto \sum_{k=1}^{K} \ln p\,(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) + E_{q(\mathbf{Z})}\,[\ln p\,(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] \\ &\propto \sum_{k=1}^{K} \ln p\,(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) + \sum_{k=1}^{K} \sum_{n=1}^{N} E_{q(\mathbf{Z})}\,[z_{nk}] \ln \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right) \end{aligned} \tag{3.58}$$

Aufgrund der Konjugiertheit zwischen Normalverteilung und Normal-Wishart Verteilung besitzt die optimale Lösung wiederum die funktionale Form einer Normal-Wishart Verteilung [9]:

$$q^*\,(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \mathcal{N}\left(\boldsymbol{\mu}_k \mid \boldsymbol{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}\right) \mathcal{W}\,(\boldsymbol{\Lambda}_k \mid \mathbf{W}_k, \nu_k) \tag{3.59}$$

mit Parametrisierung

$$\beta_k = \beta_0 + N_k \tag{3.60}$$

$$\mathbf{m}_k = \frac{1}{\beta_k}\left(\beta_0 \mathbf{m}_0 + N_k \overline{\mathbf{x}}_k\right) \tag{3.61}$$

$$\mathbf{W}_k^{-1} = \mathbf{W}_0^{-1} + N_k \mathbf{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k}\left(\overline{\mathbf{x}}_k - \mathbf{m}_0\right)\left(\overline{\mathbf{x}}_k - \mathbf{m}_0\right)^T \tag{3.62}$$

$$\nu_k = \nu_0 + N_k. \tag{3.63}$$

Nach Auswertung der Verteilungen $q^*(\boldsymbol{\theta})$ sowie $q^*(\mathbf{V})$ können wir nun die in der Verteilung $q^*(\mathbf{Z})$ involvierten Terme aus (3.46) bestimmen [4]:

$$E_{q(\boldsymbol{\mu},\boldsymbol{\Lambda})}\left[\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Lambda}_k \left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\right] = D\beta_k^{-1} + \nu_k \left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^T \mathbf{W}_k \left(\mathbf{x}_n - \boldsymbol{\mu}_k\right) \tag{3.64}$$

$$\ln \widetilde{\Lambda}_k \equiv E_{q(\boldsymbol{\Lambda})}\left[\ln |\boldsymbol{\Lambda}_k|\right] = \sum_{i=1}^{D} \psi\left(\frac{\nu_k + 1 - i}{2}\right) + D \ln 2 + \ln |\mathbf{W}_k| \tag{3.65}$$

$$\ln \widetilde{v}_{k1} \equiv E_{q(\mathbf{V})}\left[\ln V_k\right] = \psi\left(\gamma_{k1}\right) - \psi\left(\gamma_{k1} + \gamma_{k2}\right) \tag{3.66}$$

$$\ln \widetilde{v}_{k2} \equiv E_{q(\mathbf{V})}\left[\ln\left(1 - V_k\right)\right] = \psi\left(\gamma_{k2}\right) - \psi\left(\gamma_{k1} + \gamma_{k2}\right), \tag{3.67}$$

wobei wir $\widetilde{\Lambda}_k, \widetilde{v}_{k1}$ sowie $\widetilde{v}_{k2}$ definieren und $\psi(\cdot)$ die Digammafunktion (A.2) bezeichnet. Die Gleichung (3.64) folgt aus (A.20) in Verbindung mit (A.23), (3.65) sowie (3.66) und (3.67) folgen aus den Definitionen (A.24) respektive (A.9). Wir setzen (3.65),(3.66) und (3.67) in (3.46) ein und erhalten unter Berücksichtigung von (3.47)

$$r_{nk} \propto \widetilde{\Lambda}^{1/2} \exp\left\{-\frac{D}{2\beta_k} - \frac{\nu_k}{2}\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^T \mathbf{W}_k \left(\mathbf{x}_n - \boldsymbol{\mu}_k\right) + \ln \widetilde{v}_{k1} + \sum_{i=1}^{k-1} \ln \widetilde{v}_{i2}\right\}. \tag{3.68}$$

Nach der Spezifikation der variationalen Verteilung sind wir nun in der Lage, die zur Auswertung der Unterschranke $\mathcal{L}(q)$ (3.41) erforderlichen Erwartungswerte zu evaluieren. Wir betrachten die einzelnen Terme separat und erhalten unter der Bezeichnung q für die komplette Verteilung $q^*(\mathbf{Z}, \mathbf{V}, \boldsymbol{\theta})$ (vergleiche [4]):

$$\begin{aligned} E_q\left[\ln p\left(\mathbf{X} \mid \mathbf{Z}, \mathbf{V}, \boldsymbol{\Lambda}\right)\right] = &-\frac{ND \ln(2\pi)}{2} + \frac{1}{2}\sum_{k=1}^{K} N_k \left\{\ln \widetilde{\Lambda}_k - D\beta_k^{-1}\right. \\ &\left. - \nu_k \mathrm{Tr}\left(\mathbf{S}_k \mathbf{W}_k\right) - \nu_k \left(\overline{\mathbf{x}}_k - \mathbf{m}_k\right)^T \mathbf{W}_k \left(\overline{\mathbf{x}}_k - \mathbf{m}_k\right)\right\} \end{aligned} \tag{3.69}$$

$$E_q\left[\ln p\left(\mathbf{Z} \mid \mathbf{V}\right)\right] = \sum_{n=1}^{N}\sum_{k=1}^{K}\left\{ r_{nk} E_{q(\mathbf{V})}\left[\ln V_k\right] + \sum_{j=k+1}^{K} r_{nj} E_{q(\mathbf{V})}\left[\ln\left(1-V_k\right)\right]\right\} \tag{3.70}$$

$$E_q\left[\ln q\left(\mathbf{Z}\right)\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\ln r_{nk} \tag{3.71}$$

$$\begin{aligned} E_q\left[\ln p\left(\boldsymbol{\mu},\boldsymbol{\Lambda}\right)\right] = {} & K\ln\mathbf{B}\left(\mathbf{W}_0,\nu_0\right) + \frac{\nu_0 - D}{2}\sum_{k=1}^{K}\ln\widetilde{\Lambda}_k - \frac{1}{2}\sum_{k=1}^{K}\nu_k\mathbf{Tr}\left(\mathbf{W}_0^{-1}\mathbf{W}_k\right) \\ & + \frac{1}{2}\sum_{k=1}^{K}\left\{ D\ln\frac{\beta_0}{2\pi} - \frac{D\beta_0}{\beta_k} - \beta_0\nu_k\left(\mathbf{m}_k - \mathbf{m}_0\right)^T\mathbf{W}_k\left(\mathbf{m}_k - \mathbf{m}_0\right)\right\} \end{aligned} \tag{3.72}$$

$$E_q\left[\ln q\left(\boldsymbol{\mu},\boldsymbol{\Lambda}\right)\right] = \sum_{k=1}^{K}\left\{\frac{1}{2}\ln\widetilde{\Lambda}_k + \frac{D}{2}\ln\frac{\beta_k}{2\pi} - \frac{D}{2} - \mathbf{H}\left[q\left(\boldsymbol{\Lambda}_k\right)\right]\right\} \tag{3.73}$$

$$E_q\left[\ln p\left(\mathbf{V}\right)\right] = \left(\alpha - 1\right)\sum_{k=1}^{K-1} E_{q(\mathbf{V})}\left[\ln\left(1-V_k\right)\right] + \left(K-1\right)\ln\alpha \tag{3.74}$$

$$\begin{aligned} E_q\left[\ln q\left(\mathbf{V}\right)\right] = {} & \sum_{k=1}^{K-1}\left\{\left(\gamma_{k1}-1\right)E_{q(\mathbf{V})}\left[\ln v_k\right] + \left(\gamma_{k2}-1\right)E_{q(\mathbf{V})}\left[\ln\left(1-v_k\right)\right]\right. \\ & \left. + \ln\frac{\Gamma\left(\gamma_{k1}+\gamma_{k2}\right)}{\Gamma\left(\gamma_{k1}\right)\Gamma\left(\gamma_{k2}\right)}\right\}, \end{aligned} \tag{3.75}$$

wobei $\mathbf{B}\left(\mathbf{W},\nu\right)$ die Normalisierungskonstante (A.22) und $\mathbf{H}\left[q\left(\boldsymbol{\Lambda}\right)\right]$ die Entropie der Wishartverteilung (A.25) bezeichnen.
Ein Vergleich mit dem variationalen endlichen Mixture Modell mit normalverteilten Komponenten [4] zeigt, dass die Modelle eine große Übereinstimmung aufweisen. Der Unterschied liegt in der Verteilung der Gewichtskoeffizienten der einzelnen Komponenten, welche im endlichen Mixture Modell statt einem Produkt von Betaverteilungen einer Dirichletverteilung folgen.

3.3.3 Erweiterung des Modells

In der bisherigen Betrachtung wurde der Konzentrationsparameter α des Dirichlet Prozesses nicht in die Modellierung integriert, sondern als fixer Wert angenommen. Da α maßgeblichen Einfluss auf die Bestimmung der verwendeten Komponentenanzahl ausübt, in der praktischen Anwendung aber meist keine hinreichenden Informationen zu dessen Fixierung vorliegen [7], erweitern wir unser Modell, indem wir α als zusätzliche Zufallsvariable integrieren. Analog zu [6] verwenden wir eine zu den Stabgewichten konjugierte Verteilung:

$$\alpha \sim \mathrm{Gamma}\left(a,b\right), \tag{3.76}$$

wobei wir die Hyperparameter a und b so wählen, dass die Verteilung einen relativ flachen Verlauf hat und somit keine Präferenz über einen möglichen Wert

α wiedergibt. Verteilungen mit dieser Eigenschaft werden auch als *uninformative* Prioriverteilungen [9] bezeichnet.
Wir erhalten nach Erweiterung des Modells (3.29) um eine Verteilung über α folgende Struktur:

$$\begin{aligned} \alpha &\sim \text{Gamma}\,(a, b) \\ V_k \mid \alpha &\sim \text{Beta}\,(1, \alpha) \\ \theta_k \mid G_0 &\sim G_0 \\ x_k \mid \theta_k &\sim \mathcal{N}\,(x_k \mid \theta_k) \\ k &= 1, 2, \ldots \end{aligned} \tag{3.77}$$

Für die gemeinsame Verteilung folgt (siehe 3.38)

$$p\,(\mathbf{X}, \mathbf{Z}, \mathbf{V}, \boldsymbol{\theta}) = p\,(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta})\, p\,(\mathbf{Z} \mid \mathbf{V})\, p\,(\boldsymbol{\theta})\, p\,(\mathbf{V} \mid \alpha)\, p\,(\alpha) \tag{3.78}$$

Die gemeinsame variationale Verteilung aller Parameter faktorisiert sich dann entsprechend zu

$$q\,(\mathbf{Z}, \mathbf{V}, \boldsymbol{\theta}, \alpha) = q\,(\mathbf{Z})\, q\,(\mathbf{V})\, q\,(\boldsymbol{\theta})\, q\,(\alpha) \tag{3.79}$$

Weiterhin ergibt sich für den logarithmierten Erwartungswert von α unter der variationalen Verteilung gemäß (A.8) und (A.9)

$$E_q\left[\ln p\,(\alpha)\right] = (a-1)\, E_{q(\alpha)}\left[\ln \alpha\right] - b E_{q(\alpha)}\left[\alpha\right] + a \ln b - \ln \Gamma\,(a) \tag{3.80}$$

Als optimale variationale Verteilung erhalten wir schließlich

$$\begin{aligned} q^*\,(\alpha) &\propto E_{q(\mathbf{V})}\left[\ln p\,(\mathbf{V} \mid \alpha)\right] + \ln p\,(\alpha) \\ &\propto (K-1) \ln \alpha + (\alpha - 1) \sum_{k=1}^{K-1} E_{q(\mathbf{V})}\left[\ln\,(1 - v_k)\right] + (a-1) \ln \alpha - b\alpha \\ &\propto ((K + a - 1) - 1) \ln \alpha - \alpha \left(b - \sum_{k=1}^{K-1} E_{q(\mathbf{V})}\left[\ln\,(1 - v_k)\right] \right) \end{aligned} \tag{3.81}$$

Somit folgt aufgrund der Konjugiertheit zwischen trunkiertem Stick-Breaking Prozess und der Gammaverteilung α wiederum einer Gammverteilung mit entsprechend angepassten Parameterwerten

$$q^*\,(\alpha) = \text{Gamma}\,(a^*, b^*)\,, \tag{3.82}$$

wobei für die Parameterwerte gilt

$$a^* = K + a - 1 \tag{3.83}$$

$$b^* = b - \sum_{k=1}^{K-1} E_{q(\mathbf{V})}\left[\ln\,(1 - v_k)\right]. \tag{3.84}$$

Der Unterschied zur Modellierung mit fixiertem α ergibt sich nun dadurch, dass α durch seinen Erwartungswert $E_{q(\alpha)}\left[\alpha\right] = a^*/b^*$ (siehe A.4) sowie $E_{q(\alpha)}\left[\ln \alpha\right] =$

$\psi(a^*) - \ln(b^*)$ (siehe A.6) ersetzt wird . Konkret ändern sich die Gleichungen (3.74)

$$E_q[\ln p(\mathbf{V} \mid \alpha)] = \left(E_{q(\alpha)}[\alpha] - 1\right) \sum_{k=1}^{K-1} E_{q(\mathbf{V})}[\ln(1 - v_k)] + (K-1) E_{q(\alpha)}[\ln \alpha] \tag{3.85}$$

sowie (3.54)

$$\gamma_{k2} = E_{q(\alpha)}[\alpha] + \sum_{n=1}^{N} \sum_{j=k+1}^{K-1} r_{nj}. \tag{3.86}$$

Die Unterschranke passen wir schließlich wie folgt an

$$\begin{aligned}\mathcal{L}(q) = {} & E[\ln p(\mathbf{X} \mid \mathbf{Z}, \mathbf{V}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + E[\ln p(\mathbf{Z} \mid \mathbf{V})] + E[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda})] + E[\ln p(\mathbf{V} \mid \alpha)] \\ & + E[\ln p(\alpha)] - E[\ln q(\mathbf{Z})] - E[\ln q(\mathbf{V})] - E[\ln q(\boldsymbol{\mu}, \boldsymbol{\Lambda})] - E[\ln q(\alpha)]\end{aligned} \tag{3.87}$$

mit

$$E_{q(\alpha)}[\ln q(\alpha)] = (a^* - 1) E_{q(\alpha)}[\ln \alpha] - b^* E_{q(\alpha)}[\alpha] + a^* \ln b^* - \ln \Gamma(a^*) \tag{3.88}$$

Aufgrund der gegenseitigen Abhängigkeiten der Lösungen der variationalen Verteilungen besteht keine unmittelbare Möglichkeit der Auswertung. Wir bedienen uns daher zur Optimierung folgenden Iterationsschemas, wobei wir das um α erweiterte Modell betrachten:

Algorithmus: variationale Approximation

initialisiere die Verteilung der Indikatorvariablen $q(\mathbf{Z})$
setze $i = 0$ und $\mathcal{L}_0(q) = \delta$
do Iteration $(i+1)$
  $i = i + 1$
  evaluiere für $k = 1, \ldots, K-1$ Parameter der Verteilungen $q_i^*(v_k)$ (3.57):
    berechne γ_{k1}, γ_{k2} gemäß (3.53) respektive (3.86)
  evaluiere für $k = 1, \ldots, K$ Parameter der Verteilungen $q_i^*(\theta_k)$ (3.59):
    berechne $\beta_k, \mathbf{m}_k, \mathbf{W}_k^{-1}, \nu_k$ nach (3.60),(3.61),(3.62) respektive (3.63)
  evaluiere für $n = 1, \ldots, N$ Parameter der Verteilungen $q_i^*(z_n)$ (3.47):
    berechne die Momente (3.64) - (3.67)
    verwende diese zur Berechnung von r_{nk} für $k = 1, \ldots, K$ gemäß (3.68)
  evaluiere Unterschranke $\mathcal{L}_i(q)$ (3.87):
    berechne Terme gemäß (3.69) - (3.73),(3.85), (3.75), (3.80) sowie (3.88)
while $(\mathcal{L}_i(q) - \mathcal{L}_{i-1}(q)) > \epsilon$

Zur Initialisierung der Zuordnungswahrscheinlichkeit $q(\mathbf{Z})$ verwenden wir den K-means Algorithmus, wobei die Anzahl der zu verwendenden Cluster auf den Wert K des trunkierten Stick-Breaking Prozesses fixiert wird. Den Wert δ der initialen Unterschranke $\mathcal{L}_0(q)$ setzen wir auf einen hohen negativen Wert, so dass

kein unbeabsichtigter Abbruch der while-Schleife nach dem ersten Durchlauf erfolgt. Da der Algorithmus terminiert, sobald die Änderung der Unterschranke $\mathcal{L}(q)$ in aufeinanderfolgenden Iterationen ϵ unterschreitet, weisen wir der Variable ϵ einen sehr kleinen positiven Wert zu.
Die optimale approximative Verteilung nach erfolgtem Durchlauf des Algorithmus ergibt sich als $q^*(\mathbf{Z})\, q^*(\mathbf{V})\, q^*(\boldsymbol{\theta})\, q^*(\alpha)$, wobei sich die Parameter der Verteilungen aus den entsprechenden Werten der letzten Iteration des Algorithmus ergeben.

Im Folgenden betrachten wir nun, wie die gefundene optimale Verteilung zur statistischen Analyse verwendet werden kann.

3.4 Dichteschätzung und Clustering

Die approximative Posterioriverteilung können wir zur Vorhersage der Dichte einer neuen Beobachtung $\hat{\mathbf{x}}$ verwenden. Wir ordnen zu diesem Zweck der Beobachtung eine Indikatorvariable $\hat{\mathbf{z}}$ zu und erhalten gemäß (2.3)

$$p(\hat{\mathbf{x}} \mid \mathbf{X}) = \sum_{k=1}^{\infty} \iiint \pi_k(\mathbf{V} \mid \alpha)\, p(\hat{\mathbf{x}} \mid \hat{\mathbf{z}} = k, \boldsymbol{\theta}_k)\, p(\mathbf{V}, \boldsymbol{\theta}, \alpha \mid \mathbf{X})\, \mathrm{d}\mathbf{V}\mathrm{d}\boldsymbol{\theta}\mathrm{d}\alpha. \tag{3.89}$$

Um diesen Ausdruck analytisch auswerten zu können, approximieren wir die Posterioridichte durch die variationale Verteilung

$$\begin{aligned} p(\hat{\mathbf{x}} \mid \mathbf{X}) &\approx \sum_{k=1}^{K} \iiint \pi_k(\mathbf{V} \mid \alpha)\, p(\hat{\mathbf{x}} \mid \hat{\mathbf{z}} = k, \boldsymbol{\theta}_k)\, q(\mathbf{V})\, q(\boldsymbol{\theta})\, q(\alpha)\, \mathrm{d}\mathbf{V}\mathrm{d}\boldsymbol{\theta}\mathrm{d}\alpha \\ &= \sum_{k=1}^{K} E_{q(\mathbf{V}\mid\alpha)}\left[\pi_k(\mathbf{V} \mid \alpha)\right] E_{q(\boldsymbol{\theta})}\left[p(\hat{\mathbf{x}} \mid \boldsymbol{\theta}_k)\right] \end{aligned} \tag{3.90}$$

Für den ersten Term der rechten Seite schreiben wir unter Verwendung von (3.57) für $k = 1, \ldots, K$:

$$\hat{\pi}_k \equiv E_{q(\mathbf{V},\alpha)}\left[\pi_k(v_k) \mid \alpha\right] = E_{q(\mathbf{V},\alpha)}\left[v_k \mid \alpha\right] \prod_{j=1}^{k-1} E_{q(\mathbf{V},\alpha)}\left[(1 - v_j) \mid \alpha\right] \tag{3.91}$$

mit Erwartungswerten (siehe A.8)

$$E_{q(\mathbf{V},\alpha)}\left[v_k \mid \alpha\right] = \frac{\gamma_{k1}}{\gamma_{k1} + \gamma_{k2}} \tag{3.92}$$

$$E_{q(\mathbf{V},\alpha)}\left[(1 - v_k) \mid \alpha\right] = \frac{\gamma_{k2}}{\gamma_{k1} + \gamma_{k2}}. \tag{3.93}$$

Der Erwartungswert des zweiten Termes in (3.90) für festes k, der eine Integration einer Normalverteilung bezüglich einer Normal-Wishart Verteilung beinhaltet, lässt sich ebenfalls analytisch lösen und führt zu einer Studentverteilung

$\mathrm{St}(\cdot)$ [4, 9]:

$$E_{q(\boldsymbol{\mu},\boldsymbol{\Lambda})}\left[p\left(\hat{\mathbf{x}} \mid \hat{\mathbf{z}}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\right)\right] = \iint \mathcal{N}\left(\hat{\mathbf{x}} \mid \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k\right) q\left(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k\right) \mathrm{d}\boldsymbol{\mu}_k \mathrm{d}\boldsymbol{\Lambda}_k$$
$$= \mathrm{St}\left(\hat{\mathbf{x}} \mid \mathbf{m}_k, \mathbf{L}_k, \nu_k + 1 - D\right), \tag{3.94}$$

wobei der Mittelwert $\mathbf{m}_k$ und die Freiheitsgrade ν_k aus (3.61) respektive (3.63) folgen und die Präzisionsmatrix durch folgenden Ausdruck gegeben wird:

$$\mathbf{L}_k = \frac{\left(\nu_k + 1 - D\right)\beta_k}{1 + \beta_k}\mathbf{W}_k. \tag{3.95}$$

Neben der Dichteschätzung über die Vorhersagedichte ermöglicht die durch die Indikatorvariablen $\mathbf{Z}$ induzierte Partitionierung ein Clustering der Beobachtungen in verschiedene Gruppen. Wir bedienen uns dazu der variationalen Verteilung der Indikatorvariablen $q^*(\mathbf{Z})$, wobei wir jede Beobachtung $\mathbf{x}_i$ derjenigen Komponente $\hat{k}$ zuordnen, für welche die mit der i-ten Beobachtung assoziierte Indikatorvariable $\mathbf{z}_i$ die höchste Zuordnungswahrscheinlichkeit aufweist:

$$\hat{k} \equiv \arg\max_k q^*\left(z_{ik}\right) \quad i = 1, \ldots, N \tag{3.96}$$

Da die Anzahl zu verwendender Komponenten als Bestandteil der Modellierung inferiert wird, impliziert dies für das Clustering, dass die Clusteranzahl ebenfalls automatisch bestimmt und nicht im Vorfeld fixiert werden muss.

4 Empirische Untersuchung

Im Folgenden betrachten wir die Ergebnisse empirischer Untersuchungen auf Basis des variationalen Dirichlet Prozess Mixture Modells. Die Spezifikation der im Modell vorhandenen fixen Parameter orientiert sich an [17] und bedient sich uninformativer Prioriverteilungen. Konkret fixieren wir den Erwartungswert der Mittelwerte der einzelnen Komponenten auf den empirischen Mittelwert der Beobachtungen $\mathbf{m}_0 = \hat{\boldsymbol{\mu}}$, die inverse Kovarianzmatrix auf ein Vielfaches der empirischen inversen Kovarianzmatrix $\boldsymbol{\Sigma} = D * \hat{\boldsymbol{\Sigma}}$, wobei der Multiplikator D der Dimension des Datensatzes entspricht sowie den Freiheitsgrad der Wishartverteilung $\nu_0 = D$ ebenfalls auf die Dimension des Datensatzes. Die Parameter a und b der Verteilung über den Konzentrationsparameter α des Dirichlet Prozesses erhalten die in [6] vorgeschlagenen Werte.

$\mathbf{m}_0$	$\boldsymbol{\Sigma}$	ν_0	a	b
$\hat{\boldsymbol{\mu}}$	$D * \hat{\boldsymbol{\Sigma}}$	D	1	1

Als Parameter des trunkierten Stick-Breaking Prozesses wählen wir in Analogie zu [6] $K = 20$, was eine Approximation des nichtparametrischen Modells durch ein Mixture Modell mit 20 Komponenten impliziert.

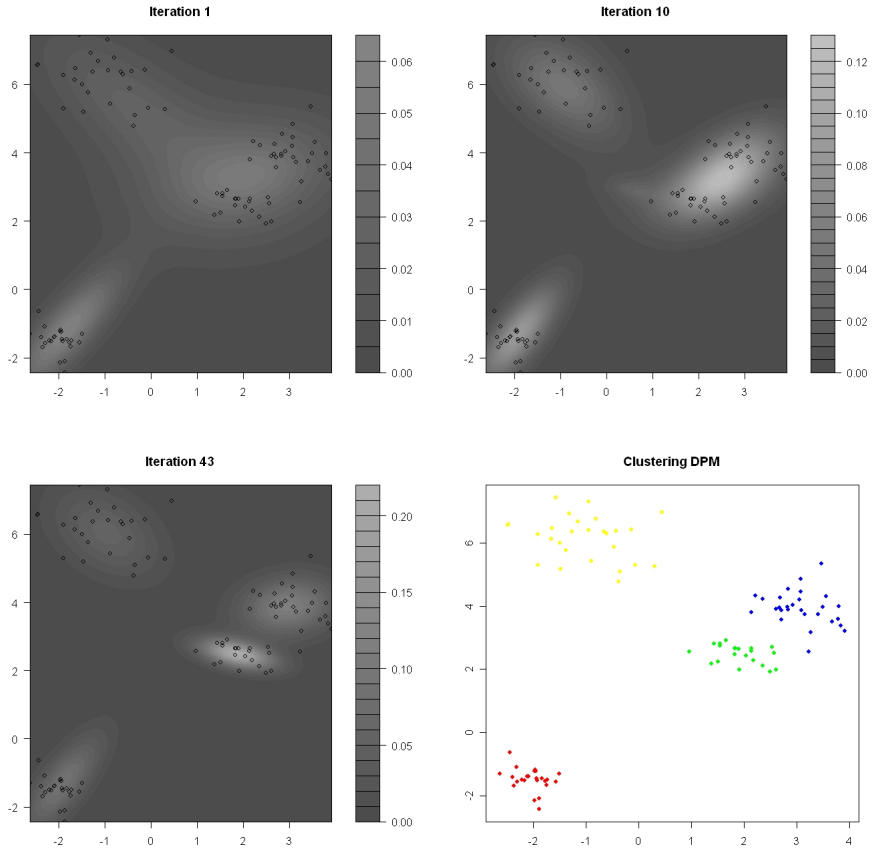

Abbildung 3: Ausgabe des Algorithmus nach Abschluss der 1., 5. sowie 43. und letzten Iteration basierend auf 100 zufällig aus der Verteilung (4.1) generierten Beobachtungen. Das Clustering erfolgt auf Basis der optimalen Verteilung.

Zunächst betrachten wir einen synthetischen Datensatz bestehend aus 100 Beobachtungen, der mittels eines Mixture Modells vier normalverteilter Komponenten mit diagonaler Kovarianzmatrix erzeugt wurde:

$$X \sim 0.2 * \mathcal{N}\left(\begin{pmatrix} -2 \\ -1.5 \end{pmatrix}, \begin{pmatrix} 0.2 & 0 \\ 0 & 0.2 \end{pmatrix}\right) + 0.3 * \mathcal{N}\left(\begin{pmatrix} -1 \\ 6 \end{pmatrix}, \begin{pmatrix} 0.4 & 0 \\ 0 & 0.4 \end{pmatrix}\right)$$
$$+ 0.2 * \mathcal{N}\left(\begin{pmatrix} 2 \\ 2.5 \end{pmatrix}, \begin{pmatrix} 0.2 & 0 \\ 0 & 0.2 \end{pmatrix}\right) + 0.3 * \mathcal{N}\left(\begin{pmatrix} 3 \\ 4 \end{pmatrix}, \begin{pmatrix} 0.3 & 0 \\ 0 & 0.3 \end{pmatrix}\right) \quad (4.1)$$

Die nachfolgende Abbildung zeigt die approximative Vorhersageverteilung nach der initialen, zehnten sowie der dreiundvierzigsten Iteration, die gleichzeitig die

letzte zur Konvergenz benötigte Iteration markiert und damit die optimale variationale Verteilung darstellt. Die graphische Darstellung wird erzeugt, indem nach Berechnung der optimalen variationalen Verteilung die approximative Vorhersagewahrscheinlichkeit (3.90) auf einem Grid mit 100×100 Gitterpunkten evaluiert wird. Nach dem ersten Durchlauf besitzt die Dichte einen relativ flachen Verlauf über den beobachteten Daten, während sich nach Abschluss der zehnten Iteration drei Modi in der Dichteschätzung abzeichnen. Die 43. Iteration liefert für die in der linken Bildhälfte befindlichen Beobachtungen eine nahezu unveränderte Schätzung, die Beobachtungen der rechten Bildhälfte werden jedoch durch zwei Komponenten modelliert, sodass sich insgesamt vier Modi in den Daten ausmachen lassen. Die vier Modi korrespondieren genau mit vier Komponenten der variationalen Verteilung, während die übrigen 16 Komponenten keinen signifikanten Einfluss ausüben. Somit inferiert der Algorithmus die korrekte Anzahl der den Daten zugrunde liegenden Komponenten. Das Clustering mittels Regel (3.96) führt zudem zu einer fehlerfreien Rekonstruktion der wahren Partition der Daten, das heisst die Anzahl sowie die Zuordnung der Beobachtungen zu den einzelnen Clustern entspricht bis auf eine abweichende Clusterbezeichnung genau der Partitionierung der Daten anhand ihrer tatsächlichen Herkunft.

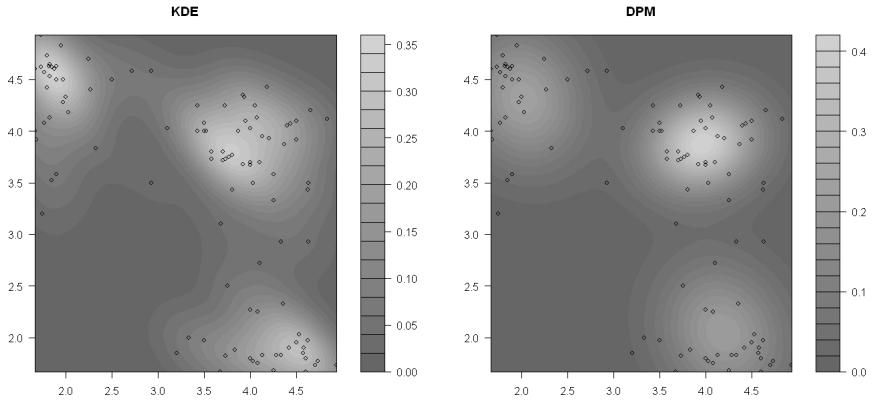

Abbildung 4: Die Abbildung illustriert die Dichteschätzung mittels des DPM Modelles und des Kerndichteschätzers auf Basis des Geysierdatensatzes. Beide Dichteschätzungen zeichnen sich durch drei Modi aus, die Schätzung des DPM Modells besitzt jedoch einen glatteren Verlauf, da hier nur drei Komponenten einen signifikanten Einfluss ausüben.

Im Folgenden vergleichen wir die Performance des Kerndichteschätzers (KDS) mit dem DPM Modell auf verschiedenen Datensätzen unterschiedlicher Dimensionalität. Als Gütekriterium betrachten wir die mittlere Leave-one-out Vorhersagedichte. Dazu wählen wir eine Beobachtung aus dem Datensatz aus, berech-

nen die Vorhersagedichte des um diese Beobachtung reduzierten Datensatzes und verwenden diese zur Dichteschätzung an der Position der ausgeschlossenen Beobachtung. Diesen Vorgang führen wir nun sukzessive für jede Beobachtung des Datensatzes aus und berechnen schließlich den Mittelwert der Dichteschätzungen. Nach Annahme entstammen sämtliche Beobachtungen derselben Verteilung, somit spricht ein höherer Wert des Gütekriteriums für eine bessere Schätzung der wahren zugrunde liegenden Verteilung. Da die variationale Approximation lediglich ein lokales Optimum in Abhängigkeit von der Startinitialisierung der Parameter garantiert, wurde das DPM Modell pro Datensatz mit 20 zufälligen Initialisierungen durchgeführt, wobei in der Tabelle jeweils der empirische Mittelwert und die empirische Standardabweichung wiedergegeben wird.

Datensatz	**Dim**	**Beob**	**DPM**	**KDS**
Synthetisch	2	100	$-2.7431(0.0356)$	-2.7723
Geysier	2	106	$-1.8675(0.0633)$	-1.7282
Krabben	4	200	$-5.5541(0.1647)$	-5.8279
Banknoten	6	200	$-2.5845(0.2094)$	-3.0317
Wein	13	178	$-16.2519(0.3315)$	-18.9788

Abbildung 5: Mittlere Leave-one-out Vorhersagedichte für verschiedene Datensätze. *Dim* und *Beob* bezeichnen die Dimension sowie die Anzahl der Beobachtungen des jeweiligen Datensatzes. Im Falle des DPM Modells repräsentiert der geklammerte Wert die Standardabweichung.

Im Fall des synthetischen Datensatzes, dessen Beobachtungen einem Mixture Modell normalverteilter Komponenten entstammen, liefert das DPM Modell eine geringfügig bessere Anpassung als das KDE Modell, während sich für den ebenfalls zweidimensionalen Geysier Datensatz ein umgekehrtes Bild ergibt. Die Anwendung des DPM Modells liefert im letzteren Fall eine Dichteschätzung, die sich durch einen im Vergleich zum KDE Modell glatteren Verlauf auszeichnet (vergleiche Abbildung 4). Die schlechtere Anpassung des DPM Modells deutet in diesem Fall darauf hin, dass dieser die Komplexität des Datensatzes tendenziell unterschätzt. Betrachtet man die weiteren Datensätze, so liefert die Dichteschätzung des DPM Modells eine höhere Anpassungsgüte. Eine mögliche Erklärung liegt darin begründet, dass die Prioriinformation im DPM Modell die Wahrscheinlichkeitsmasse auf Dichtefunktionen mit niedriger Komplexität konzentriert und damit die Wirkung einer Dimensionsreduktion erzielt. Dieser Effekt wirkt dem *curse of dimensionality* entgegen, nach dem hochdimensionale Dichteschätzungen nur für sehr große Datensätze sinnvoll durchgeführt werden können [4].
Die geringe Standardabweichung in den Lösungen des DPM Modells deutet darauf hin, dass sich der Algorithmus robust gegenüber den zufälligen Initialisierungen verhält.

A Wahrscheinlichkeitsverteilungen

A.1 Gammaverteilung und Gammafunktion

Die Gammafunktion definiert sich als

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} \exp\{-t\}\, \mathrm{d}t \tag{A.1}$$

und besitzt die Rekursionseigenschaft $\Gamma(\alpha) = (\alpha-1)\,\Gamma(\alpha-1)$, wobei im Fall natürlicher Zahlen $\alpha \in \mathbb{N}$ gilt: $\Gamma(\alpha) = (\alpha-1)!$ Die Digammafunktion bezeichnet die Ableitung der logarithmierten Gammafunktion:

$$\psi(\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha} \ln \Gamma(\alpha) \tag{A.2}$$

Folgt die Zufallsvariable $X \sim \mathcal{G}(a, b)$ einer Gammaverteilung mit Parametern a und b, so besitzt ihre Dichte die Form

$$p(x \mid a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp\{-bx\}, \quad x \geq 0, \tag{A.3}$$

mit Momenten

$$\mathrm{E}(x) = \frac{a}{b} \tag{A.4}$$

$$\mathrm{Var}(x) = \frac{a}{b^2} \tag{A.5}$$

$$\mathrm{E}(\ln x) = \psi(a) - \ln b. \tag{A.6}$$

A.2 Betaverteilung

Die Zufallsvariable $\pi \sim \mathrm{Beta}(\alpha_1, \alpha_2)$ wird als betaverteilt mit Parametern α_1, α_2 bezeichnet, falls für die Dichtefunktion gilt

$$p(\pi \mid \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\,\Gamma(\alpha_2)} \pi^{\alpha_1 - 1} (1-\pi)^{\alpha_2 - 1} \tag{A.7}$$

Es ergeben sich folgende Momente:

$$\mathrm{E}[\pi] = \frac{\alpha_1}{(\alpha_1 + \alpha_2)} \tag{A.8}$$

$$\mathrm{E}[\ln \pi_k] = \psi(\alpha_k) - \psi(\alpha_0) \quad k = 1, 2 \tag{A.9}$$

$$\mathrm{Var}[\pi] = \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2 (\alpha_1 + \alpha_2 + 1)}. \tag{A.10}$$

A.3 Dirichletverteilung

Folgt der Zufallsvektor $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K) \sim \mathrm{Dir}(\boldsymbol{\alpha})$ einer Dirichletverteilung mit Parametern $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, so besitzt die Wahrscheinlichkeitsdichte die

Form

$$p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_k \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \quad \alpha_k > 0, \tag{A.11}$$

wobei $\boldsymbol{\pi}$ über dem $K-$dimensionalen Simplex definiert ist:

$$\Delta_K = \left\{ (\pi_1, \ldots, \pi_K) : \pi_k \geq 0, \sum_{i=1}^{K} \pi_k = 1 \right\}. \tag{A.12}$$

Der erste Term der rechten Seite bildet die Normalisierungskonstante und setzt sich aus einem Quotient von Gammafunktionen zusammen. Erwartungswert und Varianz sind gegeben durch

$$\mathrm{E}[\pi_j] = \frac{\alpha_j}{\alpha_0} \qquad \mathrm{Var}[\pi_j] = \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)} \tag{A.13}$$

mit $\alpha_0 := \sum_{k=1}^{K} \pi_k$.
Für den Spezialfall $k = 2$ entspricht die Dirichletverteilung der Betaverteilung. Die Dirichletverteilung stellt nicht nur eine Verallgemeinerung der Betaverteilung dar, sondern besitzt zudem auch eine enge Verwandtschaft zur Gammaverteilung. Definiert man K unabhängige Zufallsvariablen $Z_1, \ldots, Z_K$ mit Verteilung $Z_k \sim \mathcal{G}(\alpha_k, 1), k = 1, \ldots, K$, so folgt der Vektor der normalisierten Zufallsvariablen einer Dirichletverteilung mit Parameter $\boldsymbol{\alpha}$:

$$(\pi_1, \ldots, \pi_k) := \left(\frac{Z_1}{\sum_k Z_k}, \ldots, \frac{Z_K}{\sum_k Z_k} \right) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K) \tag{A.14}$$

Aus dieser Darstellung lässt sich nun eine wichtige Eigenschaft der Dirichletverteilung herleiten: Jede Aggregation von Teilmengen dirichletverteilter Zufallsvariablen ist wiederum dirichletverteilt.

Satz A.15. *Sei $\pi \sim Dir(\boldsymbol{\alpha})$ verteilt und $r_1, \ldots, r_l$ natürliche Zahlen mit $1 \leq r_1 < \ldots < r_l = K$. Dann gilt:*

$$\left(\sum_{i=1}^{r_1} \pi_i, \sum_{i=r_1+1}^{r_2} \pi_i, \ldots, \sum_{i=r_{K-1}+1}^{r_K} \pi_i \right) \sim Dir\left(\sum_{i=1}^{r_1} \alpha_i, \sum_{i=r_1+1}^{r_2} \alpha_i, \ldots, \sum_{i=r_{K-1}+1}^{r_K} \alpha_i \right) \tag{A.16}$$

Beweis. siehe [11] □

A.4 Multinomialverteilung

Die Multinomialverteilung beschreibt die Verteilung über eine Unterteilung von N Beobachtungen in K Gruppen, wobei $n_k := \sum_{i=1}^{N} \delta_k(x_i)$ Beobachtungen auf

Gruppe k entfallen. Folgen die Beobachtungen $\mathbf{x} = (x_1, \ldots, x_N)$ einer Multinomialverteilung mit Parametervektor $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ und korrespondierender Dichtefunktion

$$p(\mathbf{x} \mid \boldsymbol{\pi}) = \frac{N!}{n_1! n_2! \cdots n_K!} \prod_{j=1}^{K} \pi_j^{n_j}, \tag{A.17}$$

so ist die Posterioriverteilung des Zufallsvektors $\boldsymbol{\pi}$ ebenfalls dirichletverteilt. Somit folgt, dass die Dirichletverteilung die konjugierte Verteilung zur Dirichletverteilung bildet. Formal schreiben wir mit $p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$:

$$\begin{aligned} p(\boldsymbol{\pi} \mid \mathbf{x}, \boldsymbol{\alpha}) &\propto p(\boldsymbol{\pi} \mid \boldsymbol{\alpha})\, p(\mathbf{x} \mid \boldsymbol{\pi}) \\ &\propto \prod_{k=1}^{K} \pi_k^{\alpha_k + n_k - 1} \propto \mathrm{Dir}(\alpha_1 + n_1, \ldots, \alpha_K + n_K) \end{aligned} \tag{A.18}$$

A.5 Multivariate Normalverteilung

Folgt $X \sim \mathcal{N}(\boldsymbol{\mu}, S^{-1})$ einer multivariaten Normalverteilung mit Mittelwertvektor $\boldsymbol{\mu}$ und Präzisionsmatrix S, so besitzt ihre Dichte folgende Gestalt:

$$p(\mathbf{x} \mid \boldsymbol{\mu}, S^{-1}) = \frac{1}{(2\pi)^{d/2} |S|^{-1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T S (\mathbf{x} - \boldsymbol{\mu}) \right\} \tag{A.19}$$

Für den Erwartungswert einer quadratischen Form ergibt sich

$$E\left[(\mathbf{x} - \mathbf{m})^T \mathbf{A} (\mathbf{x} - \mathbf{m}) \right] = (\boldsymbol{\mu} - \mathbf{m})^T \mathbf{A} (\boldsymbol{\mu} - \mathbf{m}) + \mathrm{Tr}\left(\mathbf{A}\mathbf{S}^{-1}\right) \tag{A.20}$$

A.6 Wishartverteilung

Die Wishartverteilung ist die multivariate Verallgeinerung der Gammverteilung. Folgt die Zufallsvariable $\Sigma \sim \mathcal{W}(\Delta, \nu)$ einer Wishartverteilung mit Parametern Δ und ν, gilt für ihre Wahrscheinlichkeitsdichte:

$$p(\Sigma \mid \Delta, \nu) = \mathrm{B}(\Delta, \nu) |\Sigma|^{(n-d-1)/2} \exp\left\{ -\frac{1}{2} \mathrm{tr}\left((\nu\Delta)^{-1} \Sigma \right) \right\}, \tag{A.21}$$

wobei $\mathrm{B}(\Delta, \nu)$ die Normalisierungskonstante bezeichnet:

$$\mathrm{B}(\Delta, \nu) = |\Delta|^{-\nu/2} \left(2^{\nu d/2} \pi^{d(d-1)/4} \prod_{i=1}^{d} \Gamma\left[\frac{1}{2} (\nu + 1 - i) \right] \right)^{-1} \tag{A.22}$$

Für den Erwartungswert ergibt sich

$$E(\Sigma) = \nu\Delta \tag{A.23}$$

sowie

$$E[\ln |\boldsymbol{\Lambda}_k|] = \sum_{i=1}^{D} \psi\left(\frac{\nu_k + 1 - i}{2} \right) + D \ln 2 + \ln |\boldsymbol{\Delta}_k| \tag{A.24}$$

Die Entropie der Wishartverteilung wird durch folgenden Ausdruck definiert:

$$H\left[\Delta\right] = -\ln \mathrm{B}\left(\nu, \Delta\right) - \frac{\nu - D - 1}{2} E\left[\ln|\Delta|\right] + \frac{\nu D}{2} \tag{A.25}$$

Literatur

[1] C. Andrieu, N. de Freitas, A. Doucet and M.I. Jordan. An Introduction to MCMC for machine learning. *Machine Learning*, 50:5-43, 2003.

[2] C.E. Antoniak. Mixture of Dirichlet Processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2(6):1152-1174, 1974.

[3] J.M. Bernardo and A.F.M. Smith. *Bayesian theory*. John Wiley, New York, 2000.

[4] C.M. Bishop. *Pattern recognition and Machine learning*. Springer-Verlag, New York, 2006.

[5] D. Blackwell and J.B. MacQueen. Ferguson distributions via Pólya urn schemes. *Annals of Statistics*, 1:(2):353-355, 1973.

[6] D.M. Blei and M.I. Jordan. Variational Inference for Dirichlet Process Mixtures. *Journal of Bayesian Analysis*, 1(1):121-144, 2006.

[7] M.D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577-588, June 1995.

[8] T.S. Ferguson. A Bayesian analysis of some nonpararametric problems. *Annals of Statistics*, 1(2):209-230, 1973.

[9] A. Gelman, J.B. Carlin, H.S. Stern and D.B. Rubin. *Bayesian Data Analysis, Second Edition*. Chapman & Hall /CRC, 2003.

[10] D. Görür. *Nonparametric bayesian discrete latent variable models for unsupervised learning*. PhD-Thesis, Technische Universität Berlin, 2007.

[11] J.K. Gosh and R.V. Ramamoorthi. *Bayesian nonparametrics*. Springer-Verlag, New York, 2003.

[12] H. Ishwaran and L.F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161-173, March 2001.

[13] H. Ishwaran and L.F. James. Approximate Dirichlet process computing in finite normal mixtures:Smoothing and prior information. *Journal of Computational and Graphical Statistics*, 11(3):1-26, 2002.

[14] M.I. Jordan. Dirichlet processes, Chinese restaurant processes and all that. Tutorial at *Neural Information Processing Systems*, 2005.

[15] S.N. MacEachern and P. Müller. Estimating mixtures of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223-238, 1998.

[16] R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249-265, 2000.

[17] C.E. Rasmussen. The infinite Gaussian mixture model. *Neural Information Processing Systems 12*, MIT Press, 2000.

[18] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639-650, 1994.

[19] Y.W. Teh. Dirichlet Processes. submitted to *Encyclopedia of Machine Learning*, 2007.

[20] Y.W. Teh. A Tutorial on Dirichlet processes and hierarchical Dirichlet processes, Tutorial at *Machine Learning Summer School*, Tübingen, 2007.

[21] Y.W. Teh, M.I. Jordan, M.J. Beal and D.M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566-1581, 2006.

[22] E.P. Xing, R. Sharan and M.I. Jordan. Bayesian Haplotype Inference via the Dirichlet Process. *Journal of Computational Biology*, 14(3):267-284, 2007.