Erdemir, Burak:

Widely Applicabe and Bayesian Information Criterion

Bachelorarbeit, Wintersemester 2016

Fakultät für Mathematik, Informatik und Statistik

Ludwig-Maximilians-Universität München

https://doi.org/10.5282/ubm/epub.31977

Ludwig-Maximilian-Universität
Fakultät für
Mathematik, Informatik und Statistik

Institut für Statistik

Bachelorarbeit

über das Thema

Widely Applicabe and Bayesian Information Criterion

Autor: Burak Erdemir
burak.erdemir@hotmail.de

Prüfer und Betreuer: Prof. Dr. Volker Schmid

Abgabedatum: 29.11.2016

I Kurzfassung

In der Statistik unterscheiden wir zwischen regulären und singulären statistischen Modellen. In singulären statistischen Modellen ist die Abbildung $\theta \mapsto p(x|\theta)$ nicht injektiv und seine Fisher-Informationsmatrix nicht immer positiv definit. Die Folge davon wäre, dass die bisherigen konventionellen statistischen Ergebnisse, die auf die Fisher-Regularitäten beruhen, in solchen Fällen nicht mehr anwendbar sind. Das heißt die log-Likelihood kann nicht durch irgendeine quadratische Form der Parameter approximiert werden, was wiederum zur Folge hat, dass die asymptotische Normalität des Maximum-Likelihood Schätzers nicht gilt. In der Bayes-Schätzung kann die Posteriori-Verteilung nichtmal mehr asymptotisch durch eine Normalverteilung approximiert werden. Damit sind Informationskriterien, die auf solche Regularitäten basieren, wie zum Beipsiel das AIC, DIC und BIC, für singuläre statistische Modelle nicht anwendbar. Da aber die Modellbewertung und -vergleich zu den wichtigsten Bereichen einer statistischen Analyse gehört ist es wichtig auch für singuläre Modelle Informationskriterien zu konstruieren.
Zwei wichtige Informationskriterien, die sowohl in regulären als auch in singulären statistischen Modellen anwendbar sind, gelang Sumio Watanabe mithilfe der Singular Learning Theory. Die beiden Informationskriterien Widely Applicable Information Criterion (WAIC) und Widely Applicable Bayesian Information Criterion (WBIC) setzen keine Fisher-regulären Verteilungen voraus und sind damit für jedes statistische Modell einsetzbar.

II Inhaltsverzeichnis

III Abbildungsverzeichnis

1 Einführung

1.1 Motivation

Die Modellwahl und die Variablenselektion ist bei einer statistischen Analyse einer der wichtigsten Punkte, wenn es darum geht einen Datensatz zu analysieren. In vielen Anwendungsfällen haben wir die Möglichkeit aus einem gegebenen Modell-Set das beste bzw. wahre Modell auszuwählen. So muss man sich zum Beispiel zwischen einem komplexen oder sparsamen Modell entscheiden oder zwischen einer Normal- oder einer gemischten Normalverteilung.
Ein Ansatz einen Modell aus dem Modell-Set auszuwählen ist die Analyse seiner Prognosegüte für neue Daten. Um diese Prognosegüte zu beurteilen kann man den erwarteten Prognosefehler bei allen statistischen Modellen in drei Einzelterme zerlegen: In einen irreduziblen Prognosefehler, eine Varianz und einen Quadrierten Bias [1]. Man redet dabei in der Statistik vom Varianz-Bias Trade-Off. Je komplexer das Modell, desto geringer der quadrierte Bias und desto größer gleichzeitig die Varianz. Umgekehrt weisen einfachere Modelle einen größeren quadrierten Bias und im Ausgleich eine kleinere Varianz auf. Diese Bias-Varianz Trade-Off kann in fast allen statistischen Modellen beobachtet werden [1]. Das bedeutet, dass wir unter einem optimalen Modell folgendes verstehen: Erstens sollte das Modell für beobachtete und neue Daten die bestmögliche Prognose liefern; Zweitens sollte das Modell eine kleine Varianz und drittens ebenfalls einen kleinen Bias aufweisen.
Deshalb wurden in den letzten Jahrzehnten immer wieder neue Methoden und Erweiterungen von bestehenden Methoden entwickelt um genau das oben beschriebene Problem am besten zu lösen. Es gibt zwei unterschiedliche Ansätze solche Methoden zu entwicklen: Erstens versucht man das Modell mit der besten Prognosegüte, also mit dem geringsten Prognosefehler, zu finden oder zweitens versucht man das wahre Modell unter dem beobachtetem Modell-Set zu identifizieren. Man ist durch solche Ansätze in der Lage unterschiedliche - auch Bayesianische - Modelle zu untersuchen und miteinander zu vergleichen. Solche Ansätze liefern gute Werkzeuge zur Auswahl vernünftiger Modelle.
Die Methoden, die uns das beste bzw. das wahre Modell aus einer gegebenen Menge von Modellen aussuchen, werden allgemein Informationskriterien oder auch Modellwahlkriterien genannt. Das sind Maße, die die Prognosegenauigkeit eines Modells bewerten oder versuchen das wahre Modell zu finden. Die Definition der Informationskriterien basieren normalerweise auf der Devianz, also der log-Likelihood der Daten, gegeben einem Punktschätzer des gefitteten Modells und multipliziert mit -2, also $D = -2 \cdot \log p(y|\hat{\theta})$. Die in praktischen Anwendungen am häufigsten verwendeten Informationskriterien sind:

- Das Akaike Informationskriterium [AIC],
- Das Bayes Informationskriterium [BIC],
- Das Devianz Informationskriterium [DIC].

Aber das Problem dieser Informationskriterien ist, dass es nur für reguläre statistische Modelle gilt. Ein parametrisches Modell $p(y|\theta)$ wird regulär genannt, wenn die Abbildung $\theta \mapsto p(y|\theta)$

injektiv und seine Fisher-Informationsmatrix immer positiv definit ist [2]. Unter Standardbedingungen, d.h. wenn das statistische Modell regulär ist, dann konvergiert die Likelihood-Funktion gegen eine Normalverteilung und der Maximum-Likelihood-Schätzer erfüllt die asymptotische Normalität. Dieses asymptotische Ergebnis benutzen viele Informationskriterien als Ansatz.
Wenn nun aber das statistische Modell $p(y|\theta)$ nicht injektiv oder seine Fisher-Informationsmatrix nicht immer positiv definit ist, dann wird das Modell singulär genannt. Dadurch kann die Likelihood-Funktion nicht durch eine Normalverteilung approximiert werden, was zur Folge hat, dass Informationskriterien wie das AIC, BIC und DIC nicht für die statistische Modellevaluation verwendet werden können.
Ein weiteres Problem ist, dass man in praktischen Anwendungen eher mit singulären als mit regulären Modellen arbeiten muss. Viele statistische Modelle, wie die gemischte Normalverteilung, die gemischte Binomialverteilung oder Bayesianische Netzwerke, sind singulär. Im Allgemeinen kann man sagen, dass ein Modell singulär ist, wenn ein statistisches Modell hierarchische Strukturen oder versteckte Variablen enthält.
Das asymptotische Verhalten solcher Modelle kann mithilfe der Singualar Learning Theory, die von Sumio Watanabe eingeführt wurde, analysiert werden. Ein Ergebnis solcher Analysen sind die Informationskriterien: Widely Applicable Information Criterion (WAIC) und das Widely Applicable Bayesian Information Criterion (WBIC). Es sind Generalisierungen vom AIC und BIC, die keine regulären Voraussetzungen benötigen und die für jedes statistische Modell anwendbar sind.

1.2 Einleitung

Zwei Informationskriterien, die sowohl in regulären als auch in singulären statistischen Modellen verwendet werden können, werden in dieser Arbeit vorgestellt. Mithilfe der Singular Learning Theory entwickelte Sumio Watanabe die beiden Informationskriterien: das Widely Applicable Information Criterion (WAIC) und das Widely Applicable Bayesian Information Criterion. Beide Kriterien setzen keine Fisher-regulären Bedingungen voraus und können als die verallgemeinerte Versionen vom AIC und BIC verstanden werden, da in regulären statistischen Modellen das WAIC asymptotisch gleich dem AIC ist und das WBIC asymptotisch gleich dem BIC ist.
In den folgenden Kapitel werden wir zeigen, dass der Erwartungswert vom WAIC asymptotisch gleich dem vom Bayes Generalization Error ist und, dass WBIC asymptotisch äquivalent zur Zufallsvariable Bayes-Free-Energy $\mathcal{F}$ ist. Anhand dieser Beweise sind wir dann in der Lage auch singuläre statistische Modelle zu bewerten und zu vergleichen. Wenn man den prädiktiven Verlust schätzen möchte, verwendet man das WAIC und wenn man das wahre Modell identifizieren möchte, dann eher das WBIC. Beide Informationskriterien sind anwendbar auch wenn die Posteriori-Verteilung weit entfernt von einer Normalverteilung oder die Fisher-Informationsmatrix nicht immer positiv definit ist. WAIC und WBIC haben den Vorteil, dass beide mathematisch bewiesen worden sind.

Im Kapitel 2 wird zunächst die Grundlage im Bereich der Bayes-Statistik und der Modellwahl dargestellt, die für das Verständnis der späteren Kapiteln sehr wichtig sind. Im anschließenden

Kapitel werden die üblichen Informationskriterien AIC, DIC und BIC vorgestellt und schon ein bisschen über die Vor- und Nachteile der jeweiligen Informationskriterien diskutiert. Im Hauptkapitel dieser Arbeit (Kapitel 4) werden die beiden Informationskriterien WAIC und WBIC mithilfe der Singular Learning Theory definiert und dann im Kapitel 5 mit den üblichen Informationskriterien verglichen und auf unterschiedliche Modelle angewendet. Im Kapitel 7 wird das Fazit gezogen und ein Ausblick im Bereich der Modellwahl vorgestellt.

2 Grundlagen

In diesem Kapitel werden die Grundlagen für die nächsten Kapiteln geschaffen. Es wird kurz in die Bayes-Statistik eingeführt und anschließend werden wichtige Definitionen, Annahmen und Hinweise dargestellt, die zum Verständnis der Informationskriterien beitragen sollen. Es wird speziell in die Prognosegenauigkeit eines Modells eingegangen.

2.1 Bayes-Statistik

In diesem Abschnitt wird kurz in die Bayes-Statistik eingeführt, da die beiden Informationskriterien WAIC und WBIC einen vollständig bayesianischen Ansatz vertreten.
Der Unterschied zwischen der Bayes-Statistik und der Likelihood-basierten Inferenz ist, dass die unbekannten Parameter $\theta = (\theta_1, ..., \theta_p)'$ keine festen deterministischen Größen sind, sondern als zufällig angenommen werden und eine A Priori-Verteilung besitzen. Ein Bayesianisches Modell besteht damit aus zwei Teilen:

1. *Priori-Verteilung:* Das (subjektive) Vorwissen über die unbekannten Parameter wird durch die Spezifikation einer Wahrscheinlichkeitsverteilung für den unbekannten Parameter θ ausgedrückt. Diese Verteilung wird als Priori-Verteilung von θ bezeichnet. Die Wahrscheinlichkeitsfunktion (Dichte) der Priori-Verteilung wird mit $\varphi(\theta)$ gekennzeichnet.

2. *Beobachtungsmodell:* Im sogenannten Beobachtungsmodell wird die bedingte Verteilung der Stichprobenvariablen $Y = (Y_1, ..., Y_n)'$, bei gegebenem unbekanntem Parameter dargestellt. Die Dichte dieser Verteilung ist proportional zur Likelihood $L(\theta)$ und wird im Folgenden mit $p(y|\theta)$ bezeichnet.

"*Basierend auf der Priori-Verteilung und dem Beobachtungsmodell können wir unter Zuhilfenahme des Satzes von Bayes die bedingte Verteilung von θ bei gegebenen Beobachtungen $y = (y_1, ..., y_n)'$ bestimmen*"[1]. Wir erhalten

$$p(\theta|y) = \frac{p(y|\theta) \cdot \varphi(\theta)}{\int p(y|\theta) \cdot \varphi(\theta) d\theta} = c \cdot p(y|\theta) \cdot \varphi(\theta),$$

mit der Normierungskonstante $c = \left[\int p(y|\theta) \cdot p(\theta) d\theta\right]^{-1}$. Diese Verteilung wird als Posteriori-Verteilung bezeichnet. Von der Posteriori-Verteilung gehen im Bayes-Ansatz sämtliche Inferenzschlüsse bezüglich der unbekannten Parameter θ aus.

Punktschätzer

Die üblichen Punktschätzer der Bayesianischen Inferenz sind Posteriori-Erwartungswert, Posteriori-Median und Posteriori-Modus.
Der Posteriori-Erwartungswert als Punktschätzer ist gegeben durch

$$\hat{\theta} = \mathbb{E}(\theta|y) = \int \theta \cdot p(\theta|y) d\theta = c \cdot \int \theta \cdot p(y|\theta) d\theta.$$

Mithilfe simulationsbasierter Methoden, insbesondere MCMC-Verfahren, kann man diesen Posteriori-Erwartungswert berechnen [1].
Ein weiterer möglicher Punktschätzer $\hat{\theta}$ für θ ist der Posteriori-Modus, also das globale Maximum der Posteriori-Verteilung:

$$\hat{\theta} = \arg\max_{\theta} \left(p(y|\theta) \cdot p(\theta) \right).$$

Der Vorteil des Posteriori-Modus im Vergleich zum Posteriori-Erwartungswert besteht darin, dass die Normierungskonstante zur Berechnung nicht benötigt wird.
Der letzte Punktschätzer ist der Median. Da der Median robust gegenüber Ausreißern ist, wird er häufig als Punktschätzer bevorzugt.

2.2 Maße für die Prognosegenauigkeit

Ein Weg einen Modell zu bewerten ist die Genauigkeit seiner Prognosen zu bewerten. In diesem Abschnitt werden Maße für die Prognosegenauigkeit vorgestellt, die als Fundament für die Informationskriterien AIC, DIC und BIC dienen, die im Kapitel 3 vorgestellt werden.
Maße für die Prognosegenauigkeit sind in der statistischen Analyse von essentieller Bedeutung, da wir durch angemessene Schätzer in der Lage sind nicht nur ein Modell zu bewerten, sondern mehrere Modelle miteinander vergleichen können um anschließend das Modell mit der besten Prädiktion auszuwählen.
Betrachtet werden Daten $y_1, ..., y_n$, die unabhängig und identisch verteilt sind und mit gegebenem Parameter θ modelliert werden. Wir erhalten dann die Likelihood durch:

$$p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta).$$

Bei einem Regressionsmodell würde man mit folgender Verteilung arbeiten:

$$p(y|\theta, x) = \prod_{i=1}^{n} p(y_i|\theta, x_i).$$

Um aber die Einfachheit in den folgenden Definitionen zu bewahren unterdrücken wir in unserer Notation jede Abhängigkeit auf x. Sie können jederzeit durch Hinzunahme von x mit modelliert werden.
Maße für die Pronosegenauigkeit misst so gut wie möglich den Verlust der Prognose zukünftiger Daten mit dem Modell. Die große Schwierigkeit, die sich bei der Konstruktion solcher Maße ergibt, ist, dass wir aus den beobachteten Daten die Prognosegenauigkeit für die zukünftigen Daten schätzen müssen. Dabei unterscheiden wir zwischen dem within-sample Prognosegenauigkeit - Prognosegenauigkeit für die beobachteten Daten - und dem out-of-sample Prognosegenauigkeit - Prognosegenauigkeit für zukünftige Daten. Natürlich steht die Schätzung der out-of-sample Prognosegenaugkeit im Vordergrund dieser Arbeit.

2.2.1 Mean-Squared-Error (MSE)

Eine Möglichkeit die Progosegenauigkeit eines Modells zu bewerten ist der Mean Squared Error (MSE). Das Fit eines Modells an neue erhobene Daten kann numerisch durch den MSE

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbb{E}(y_i|\theta))^2$$

oder durch eine gewichtete Version wie z.B.

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \mathbb{E}(y_i|\theta))^2}{Var(y_i|\theta)}$$

zusammengefasst werden [3]. Der Vorteil von solch einem Maß ist die einfache Berechnung, Implementation und Interpretation. Dabei wird das Modell mit dem niedrigsten MSE, also mit dem niedrigsten Prognosefehler, bevorzugt. Der Nachteil dieser Maße sind, dass sie für Modelle, die nicht einer Normalverteilung folgen, weniger angemessen und stabil sind.

2.2.2 Die log-Likelihood

Die log-Likelihood (log predictive density) ist einer der wichtigsten Größen in der statistischen Modellevaluation. Die Informationskriterien, wie das AIC, DIC und BIC basieren auf der log-Likelihood der Daten und Addieren jeweils mit eigener Begründung einen zusätzlichen Term, welches den Bias des Schätzers repräsentieren soll.
Die log-Likelihood $\log p(y|\theta)$ ist eine allgemeinere Zusammenfassung (als der MSE) über den prädiktiven Fit. Die log-Likelihood ist proportional zum Mean Squared Error, wenn das Modell normalverteilt mit konstanter Varianz ist [3].
Wie schon erwähnt spielt die log-Likelihood eine wichtige Rolle beim Vergleich und Bewertung von statistischen Modellen aufgrund seiner Verbindung zur Kullback-Leibler Informationsmaß [4]. Mit steigender Stichprobengröße wird das Modell mit der niedrigsten Kullback-Leibler Distanz - und daher die höchste erwartete log-Likelihood - die höchste Posteriori-Wahrscheinlichkeit besitzen [3]. Deshalb scheint es sinnvoll zu sein die erwartete log-Likelihood als ein Maß für die Modellanpassung zu verwenden.
Bei Bayesianischen Modellen kann die Wahl der Priori bei spärlichen Daten sehr wichtig sein. Denn eine schlechte Wahl der Priori-Verteilung führt zu schwachen Inferenzen und zu schlechten Prädiktionen.
In regulären statistischen Modellen nähert sich die A Posteriori-Verteilung $p(\theta|y)$ mit steigender Stichprobengröße n einer Normalverteilung an. In dieser asymptotischen Grenze wird die Posteriori durch die Likelihood dominiert - die Priori trägt nur einen Faktor bei, während die Likelihood n Faktoren beiträgt, nämlich einen für jeden Datenpunkt - also nähert sich die Likelihood-Funktion ebenfalls der selben Normalverteilung an [3].

Da die Stichprobengröße gegen unendlich geht, $n \to \infty$, können wir die Posteriori-Verteilung als $\theta|y \to N(\theta_0, V_0/n)$ darstellen und erhalten für die log-Likelihood

$$\log p(y|\theta) = c(y) - \frac{1}{2}\left(d\log(2\pi) + \log|V_0/n| + (\theta-\theta_0)^T(V_0/n)^{-1}(\theta-\theta_0)\right), \tag{1}$$

wobei $c(y)$ eine Konstante ist, die nur von den Daten y und der Modellklasse abhängt, aber nicht von den Parametern θ [3].
Die entsprechende multivariate Normalverteilung für θ erzeugt eine Posteriori-Verteilung für die log-Likelihood, das als eine Konstante (identisch zu $c(y) - \frac{1}{2}(d\log(2\pi) + \log|V_0/n|)$ minus $\frac{1}{2}$ mal eine χ^2_d-Zufallsvariable endet, wo d die Dimension von θ ist, also die Anzahl an Parameter im Modell. Das Maximum dieser Verteilung der log-Likelihood wird erreicht, wenn θ gleich dem Maximum-Likelihood Schätzer ist und ihr Posteriori-Erwartungswert ist um $\frac{k}{2}$ kleiner [3]. Für die eigentliche Posteriori-Verteilung ist dieses asymptotische Ergebnis nur eine Approximation, aber ist als ein Orientierungswert für das Interpretieren der log-Likelihood als ein Maß für die Anpassung hilfreich.
Nun gilt die oben beschriebene Asymptotik nur für reguläre statistische Modelle. Mit singulären Modellen, wie z.B. gemischte Normal- und Binomialmodelle oder überparametrisierte komplexe Modelle, haben wir keine Eindeutigkeit des ML-Schätzers mehr, die Fisher-Informationsmatrix ist nicht positiv definit, Plug-In Schätzer sind nicht Repräsentative der Posteriori und die Verteilung der Devianz konvergiert nicht gegen eine χ^2-Verteilung. Das asymptotische Verhalten solcher Modelle kann mithilfe der Singular Learning Theory [5] analysiert werden.

2.2.3 Prognosegenauigkeit für einen einzigen Datenpunkt

Viele Informationskriterien, wie später das WAIC und DIC, bewerten die Prognosegenauigkeit eines Modells, indem sie die Prognosegenauigkeit einzelner Datenpunkte bewerten. Solche Informationskriterien haben das Vorteil stabiler und auch präziser zu sein, weil sie jeden einzelnen Datenpunkt für die Berechnung der Prognosegenauigkeit verwenden.
Wir bezeichnen im folgenden f als das wahre Modell, y als die beobachteten Daten und $\tilde{y}$ als die zukünftigen Daten. Das out-of-sample Vorhersagefit für einen neuen Datenpunkt $\tilde{y_i}$ durch Benutzung des logarithmischen Scores ist dann

$$\log p_{post}(\tilde{y_i}) = \log \mathbb{E}_{post}\left(p(\tilde{y_i}|\theta)\right) = \int p(\tilde{y_i}|\theta)p_{post}(\theta)d\theta.$$

Dabei ist im oberen Ausdruck $p_{post}(\tilde{y_i})$ die Likelihood für $\tilde{y_i}$, die durch die Posteriori-Verteilung $p_{post}(\theta) = p(\theta|y)$ erzeugt wurde. Wir benutzen p_{post} und $\mathbb{E}_{post}$ um jede Wahrscheinlichkeit oder Erwartung über die Posteriori-Verteilung darzustellen.
Da aber die zukünftigen Daten $\tilde{y_i}$ selbst unbekannt sind, müssen wir die erwartete out-of-sampple

log-Likelihood für einen neuen Datenpunkt $\tilde{y_i}$ definieren und natürlich auch schätzen:

$$\begin{aligned} \text{elpd} &= \text{expected log predictive density for a new data point} \\ &= \mathbb{E}_f \left(\log p_{post}(\tilde{y_i}) \right) \\ &= \int \left(\log p_{post}(\tilde{y_i}) \right) f(\tilde{y_i}) d\tilde{y}. \end{aligned} \tag{2}$$

Oft wird diese Größe als auch die durchschnittliche log-Likelihood bezeichnet. Ein weiteres Problem welches nun auftritt ist, dass wir in jeder Anwendung zwar eine Posteriori-Verteilung p_{post} haben, aber im Allgemeinen die wahre Datenverteilung f nicht kennen. Ein üblicher Weg die erwartete out-of-sample log-Likelihood trotzdem zu schätzen wäre in die Formel ein Schätzer für f einzusetzen.
Um die Vergleichbarkeit mit dem vorliegenden Datensatz zu bewahren, kann man einen Maß für die Prognosegenauigkeit für die n Datenpunkte, einzeln genommen, definieren [3]

$$\begin{aligned} \text{elppd} &= \text{expected log pointwise predictive density for a new dataset} \\ &= \sum_{i=1}^{n} \mathbb{E}_f \left(\log p_{post}(\tilde{y_i}) \right). \end{aligned} \tag{3}$$

Der Vorteil der Benutzung eines 'punktweisen' Maßes, statt das Arbeiten mit der gemeinsamen posterior predictive distribution $p_{post}(\tilde{y})$ ist in der Verbindung der punktweisen Berechnung zur Cross-Validation, welches einige ziemlich allgemeine Ansätze zur Approximation der out-of-sample Fits, durch Benutzen verfügbarer Daten, erlaubt [3].
Manchmal ist es trotzdem hilfreich die Prognosegenauigkeit, gegeben einem Punktschätzer $\hat{\theta}(y)$, zu betrachten:

$$\text{erwartete log-Likelihood, gegeben Punktschätzer } \hat{\theta} : \mathbb{E}_f \left(\log p(\tilde{y}|\hat{\theta}) \right). \tag{4}$$

Für Modelle mit unabhängigen Daten, gegebenen Parametern, gibt es keinen Unterschied zwischen der gemeinsamen und der punktweisen Prädiktion, gegeben einem Punktschätzer, da $p(\tilde{y}|\hat{\theta}) = \prod_{i=1}^{n} p(\tilde{y_i}|\hat{\theta})$ [3].

2.2.4 Prognosegenauigkeit eines gefitteten Modells

In diesem Abschnitt wird nun das Fundament für die wichtigsten Informationskriterien, wie das DIC und WAIC, gelegt. Der Ansatz der hier vorgestellt wird, wird ebenfalls von den beiden Informationskriterien DIC und WAIC verwendet.
In der Praxis ist der Parameter θ nicht bekannt, deshalb können wir auch die log-Likelihood $\log p(y|\theta)$ nicht kennen. Deshalb würden wir gerne mit der Posteriori-Verteilung $p_{post}(\theta) = p(\theta|y)$ arbeiten und die Prognosegenauigkeit des gefitteten Modells an die Daten durch einen punkt-

weisen Maß *lppd* zusammenfassen:

$$\begin{aligned} \text{lppd} &= \text{log pointwise predictive density} \\ &= \log \prod_{i=1}^{n} p_{post}(y_i) \\ &= \sum_{i=1}^{n} \log \int p(y_i|\theta) p_{post}(\theta) d\theta. \end{aligned} \tag{5}$$

Um diese Likelihood in der Praxis zu berechnen, können wir die Erwartung abschätzen, indem wir Ziehungen aus der $p_{post}(\theta)$ benutzen, welches wir mit θ^s, $s = 1, ..., S$, bezeichnen [3]

$$\begin{aligned} \text{computed lppd} &= \text{computed log pointwise predictive density} \\ &= \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta^s) \right). \end{aligned} \tag{6}$$

Wir nehmen normalerweise an, dass die Anzahl der Simulationsziehungen S groß genug ist um die Posteriori-Verteilung komplett zu erfassen, daher werden wir uns auf den theoretischen Wert (5) und auf die Berechnung (6) abwechselnd beziehen.
Wie wir im nächsten Abschnitt diskutieren werden ist die *lppd* der beobachteten Daten y eine Überschätzung der *elppd* für zukünftige Daten (3). Daher ist der Plan mit (6) anzufangen und dann einige Versionen von Bias-Korrekturen anzuwenden um einen vernünftigen Schätzer für (3) zu erhalten.

2.2.5 Schätzen der out-of-sample Prognosegenauigkeit

Es sind mehrere Methoden verfügbar um die erwartete Prognosegenauigkeit zu schätzen ohne auf die out-of-sample Daten zu warten. Wir können solche Formeln wie (2) nicht direkt berechnen, weil wir die wahre Verteilung f nicht kennen. Stattdessen können wir verschiedene Approximationen betrachten. Wir listen hier im Folgenden einige vernünftig-scheinende Approximationen auf. Jede dieser Methoden hat Nachteile, was uns zeigt, dass jedes Maß für die Prognosegenauigkeit, das wir berechnen, nur approximativ sein wird.

Within-Sample Prognosegenauigkeit: Ein naiver Schätzer der expected log predictive density für neue Daten ist der log predictive density für die beobachteten Daten. Wie bereits oben diskutiert werden wir mit der Bayesianischen punktweisen Formel (5) arbeiten, indem wir die Simulationen (6) verwenden. Diese Zusammenfassung ist schnell und einfach zu verstehen, aber ist im Allgemeinen eine Überschätzung von (3), weil es an den Daten ausgewertet wurde, indem das Modell gefittet wurde [3].

Adjusted within-sample Prognosegenauigkeit: Wenn man bedenkt, dass lppd ein verzerrter Schätzer von elppd ist, ist der nächste logische Schritt diesen Bias zu korrigieren. Formeln

wie AIC, DIC und WAIC, die unten noch ausführlich diskutiert werden, geben approximativ unverzerrte Schätzer für elppd an, indem sie mit dem lppd starten und dann ein Korrektur für die Anzahl an Parametern subtrahieren. Diese Anpassungen sind in der Lage vernünftige bzw. akzeptable Antworten in vielen Fällen zu geben, aber haben das allgemeine Problem bestenfalls nur in Erwartung korrekt zu sein. Dennoch sind es nützliche Kriterien, die uns bei der Modellwahl in praktischen Anwednungen immer wieder helfen.

Cross-Validation: Um die out-of-sample Prognosefehler zu erfassen kann man versuchen das Modell an sogenannte Trainingsdaten zu fitten und dann anschließend diese Prognosegenauigkeit an einem holdout-set (Validierungsdatensatz) bewerten [3]. Cross-Validation meidet das Problem des Overfittings, aber bleibt an die vorliegenden Daten gebunden. Dadurch kann es bestenfalls nur in Erwartung korrekt sein. Außerdem kann Cross-Validation rechnerisch sehr aufwändig sein: um einen stabilen Schätzer zu bekommen benötigt man normalerweise viele Datenpartitionen und Fits. Im Extremfall erfordert leave-one-out cross-validation (LOO-CV) n Fits [3].

3 Informationskriterien

In diesem Kapitel kommen wir nun zu den Informationskriterien, die versuchen für die vorliegenden Daten das beste bzw. wahre Modell aus einem gegebenen Modell-Set auszuwählen. Dabei versuchen sie die Prognosegenauigkeit für zukünftige Daten aus den beobachteten Daten zu schätzen. Der Vergleich und die Bewertung von unterschiedlichen Modellen ist das Hauptziel der Informationskriterien. Die hier vorgestellten Kriterien basieren normalerweise auf der Devianz - die log-Likelihood der Daten, gegeben einem Punktschätzer $\hat{\theta}$ des gefitteten Modells, multipliziert mit -2, also $D = -2 \log p(y|\hat{\theta})$ - und beziehen sich auf die asymptotische Normalität der log-Likelihood bzw. auf die χ^2-Verteilung der Devianz.
Es wird in diesem Kapitel, die am meisten verwendeten Informationskriterien dargestellt: das AIC, DIC und BIC. Während das AIC und DIC versuchen anhand der Prognosegenauigkeit einen Modell auszuwählen, kann man anhand von BIC das wahre Modell identifizieren.
Sei $\hat{\theta}$ ein Punktschätzer und $p_{post}(\theta)$ die Posteriori-Verteilung der gebenen Daten y. Verständlicherweise ist dann die out-of-sample Prädiktionen weniger präzise als durch die within-sample Prognosegenauigkeit angedeutet. Anders ausgedrückt, die Genauigkeit der Prädiktionen von einem gefitteten Modell für zukünftige Daten wird in Erwartung niedriger sein, als die Genauigkeit der Prädiktionen vom gleichen Modell für die beobachteten Daten [3].
Wir sind an der Genauigkeit der Prädiktionen eines Modells aus zwei Gründen interessiert: erstens, um die Performanz eines Modells zu messen und zweitens, um mehrere Modelle miteinander zu vergleichen. Unser Ziel Hauptziel ist eine Möglichkeit zu finden unterschiedliche Modelle zu vergleichen.
Wenn unterschiedliche Modelle die selbe Anzahl an Parametern haben, die auf gleiche Weise geschätzt wurden, könnte man ganz einfach deren log-Likelihoods direkt vergleichen. Dabei entscheidet man sich immer für das Modell mit der höchsten log-Likelihood. Aber wenn man Modelle von unterschiedlicher Größe vergleicht, ist es wichtig an die log-Likelihood einige Korrekturen durchzuführen. Die Korrektur ist für die natürliche Fähigkeit eines komplexeren Modells Daten besser zu fitten. Das heisst das Modelle, die alle Einflussvariablen mit ins Modell aufnimmt, werden besser an die beobachteten Daten gefittet als zukünftige. Für diese natürliche Fähigkeit soll an dem log-Likelihood eine eine Korrektur durchgeführt werden.

3.1 Akaike's Informationskriterium [AIC]

In vielen statistischen Literaturen über die Prognosegenauigkeit von Modellen wird die Inferenz für θ nicht durch eine Posteriori Verteilung p_{post} zusammengefasst, sondern durch einen Punktschätzer $\hat{\theta}$, normalerweise der Maximum-Likelihood Schätzer [3]. Die out-of-sample Prognosegenauigkeit ist dann nicht durch (2) gegeben, sondern durch

$$\mathsf{elpd}_{\hat{\theta}} = \mathbb{E}_f \left(\log p(\tilde{y}|\hat{\theta}(y)) \right),$$

die in (4) definiert wurde , wo beide y und $\tilde{y}$ zufällig sind [3]. Es gibt keinen direkten Weg (4) zu berechnen; stattdessen ist der Standardansatz die log-Posteriori Dichte der beobachteten Daten y, gegeben einem Punktschätzer $\hat{\theta}$, zu benutzen und den Bias angemessen korrigieren.

Sei d die Anzahl der Parameter, die im Modell geschätzt wurden. Die einfachste Bias-Korrektur beruht auf der asymptotischen Posteriori Normalverteilung. In diesem Grenzwert ist das Subtrahieren von d von der log-Likelihood, gegeben dem Maximum-Likelihood Schätzer, eine Korrektur für das Overfitting:

$$\widehat{\mathsf{elpd}}_{AIC} = \log p(y|\hat{\theta}_{mle}) - k. \tag{7}$$

Akaike hat für seine Definition vom AIC die Form der Devianz genommen; also ist AIC der obere Ausdruck multipliziert mit -2

$$AIC = -2 \log p(y|\hat{\theta}_{mle}) + 2k.$$

Es macht Sinn die Devianz für gefittete Parameter anzupassen, da wir dadurch die Vergleichbarkeit mit anderen Modellen bewahren. Aber sobald wir Lineare Modelle mit flachen Prioris verlassen, können wir nicht einfach als Bias-Korrektur d hinzufügen. Informative Prioris und hierarchische Strukturen tendieren dazu den Betrag des Overfittings im Vergleich zum KQ- oder ML-Schätzung zu reduzieren.

3.2 Devianz Informationskriterium [DIC]

DIC kann als die Bayesianische Version von AIC betrachtet werden, das die Formel (7) verwendet und zwei Veränderungen durchführt: erstens wir der Maximum-Likelihood Schätzer $\hat{\theta}$ mit dem Posteriori-Erwartungswert

$$\hat{\theta}_{Bayes} = \mathbb{E}(\theta|y)$$

vertauscht und d mit einer Daten-basierten Bias-Korrektur ersetzt. Der neue Maß für die Vorhersagegenauigkeit ist

$$\widehat{\mathsf{elpd}}_{DIC} = \log p(y|\hat{\theta}_{Bayes}) - p_{DIC}, \tag{8}$$

wobei p_{DIC} die effektive Anzahl an Parametern ist, die definiert ist durch

$$p_{DIC} = 2\left(\log p(y|\hat{\theta}_{Bayes}) - \mathbb{E}_{post}(\log p(y|\theta))\right), \tag{9}$$

wo die Erwartung im zweiten Term ein Mittelwert von θ über seine Posteriori-Verteilung ist [3]. Gleichung (9) wird berechnet, indem man Simulationen θ^s, $s = 1, ..., S$, wie folgt verwendet:

$$\mathsf{computed}\ \mathsf{p}_{DIC} = 2\left(\log p(y|\hat{\theta}_{Bayes}) - \frac{1}{S}\sum_{s=1}^{S} \log p(y|\theta^s)\right). \tag{10}$$

Der Posteriori-Erwartungswert von θ wird das Maximum der log-Likelihood erzeugen, wenn es zufällig dasselbe ist wie der Modus ($\theta_{EW} = \theta_{Mod}$), und negative p_{DIC} kann erzeugt werden, wenn der Posteriori-Erwartungswert weit entfernt vom Modus ist [3].

3.3 Bayesianische Informationskriterium [BIC]

Obwohl wir jezt fast die ganze Arbeit über die Prognosegenauigkeit gesprochen haben, gibt es einen weiteren Ansatz um ein gutes Modell auszuwählen. Undzwar anhand der marginalen Likelihood versucht das BIC und WBIC das wahre Modell unter dem gegebenen Modell-Set zu finden.

Während es für AIC und DIC auch einen Likelihood-basierten Ansatz für die Berechnung gibt, ist das Bayesianische Informationskriterium ein komplett bayesianisch fundiertes Informationskriterium. Entwickelt wurde das BIC von Gideon Schwarz im Jahr 1978 und ist wie folgt definiert:

$$BIC = -2 \cdot ln(L(\hat{\theta}|y)) + \frac{d}{2} \cdot ln(n).$$

Der erste Unterschied zum AIC ist: Der hintere Term, der beim AIC zuvor den (asymptotischen) Bias der Schätzung der erwarteten K-L-I bildete, ist nun vom Stichprobenumfang n abhängig [4]. Das BIC bestraft die Anzahl der Parameter stärker als das AIC. Deswegen favorisiert das BIC eher sparsamere Modelle und das AIC eher komplexere Modelle. Die Interpretation ist aber analog zum AIC: Das Modell mit dem geringsten BIC-Wert ist zu bevorzugen. Da das BIC nicht so interessant wie das AIC und DIC ist wird es hier ein bisschen vernachlässigt.

4 Widely Applicable and Bayesian Information Criterion

Wie in der Einleitung schon erwähnt sind die üblichen Informationskriterien, wie das AIC, DIC und BIC in der Praxis oft nicht anwendbar, da man im Alltag eher mit singulären statistischen Modellen konfrontiert ist als mit regulären Modellen. Alle vorgeführten Informationskriterien setzen die Fisher-Regularitäten voraus und somit auch die Eindeutigkeit des Maximum-Likelihood-Schätzers. Ein Beispiel, wo die üblichen Informationskriterien nicht hinzugezogen werden sollten wäre eine gemischte Normalverteilung.

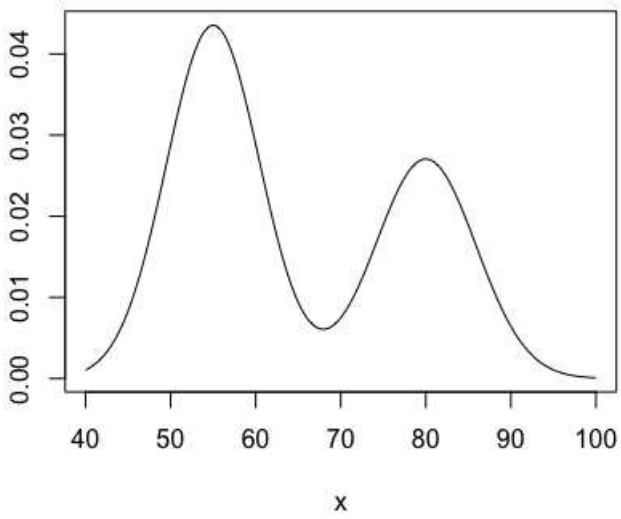

Abbildung 1: Dichte einer gemischten Normalverteilung

An der Abbildung 1 ist zu erkennen, dass die gemischte Normalverteilung eine bimodale Verteilung ist. Deswegen ist hier die Anwendung des DIC mit Punktschätzer dem Posteriori-Modus nicht anwendbar. Aber auch der Posteriori-Erwartungswert und der Posteriori-Median wäre in dieser Anwendung keine richtige Wahl. Aber auch das AIC und BIC, welches als Punktschätzer den Maximum-Likelihood Schätzer verwenden, sind nicht anwendbar, da die Eindeutigkeit des ML-Schätzers nicht gewährleistet ist. Zusätzlich zu den Problem kommt hinzu, dass die Fisher-Informationsmatrix nicht positiv definit ist, was zur Folge hat, dass die konventionellen statistischen Ergebnisse im Bereich der Modellevaluation, die die Fisher-regularitäten voraussetzen, in solchen Fällen nicht ohne schlechtes Gewissen einsetzbar sind.
Aus diesem Grund wurde von Sumio Watanabe die 'Singular Learning Theory' eingeführt, welches sich genau mit solchen Singularitäten (Problemen) beschäftigt. Er stellte zwei neue Informationskriterien vor und bewies, dass beide Kriterien sowohl in regulären als auch in singulären statistischen Modellen funktioniert.
Das erste Informationskriterium ist das Widely Applicable Bayesian Information Criterion [WBIC], welches versucht die Bayes-Free-Energy $\mathcal{F}$ zu schätzen um das wahre Modell zu identifizieren.

Die Größe $\mathcal{F}$ spielt in der Bewertung und Vergleich von statistischen Modellen eine große Rollen. Auf die Definition und Bedeutung wird später ausführlicher eingegangen.
Das zweite Informationskriterium ist das Widely Applicable Information Criterion [WAIC], welches versucht den Generalization Error G, also den prädiktiven Verlust, wenn auch nur in Erwartung, zu schätzen. Das WAIC erlaubt uns den das Modell mit der niedrigsten Prognosefehler und damit mit der höchsten Prognosegenauigkeit auszuwählen.
Wie wir im Laufe dieses Kapitels sehen werden, sind beide Informationskriterien sind in regulären und auch in singulären Modellen einsetzbar. Während WBIC in regulären Modellen asymptotisch gleich dem BIC ist, ist WAIC in regulären Fällen asymptotisch gleich dem AIC und DIC. Die Intention der jeweiligen Informationskriterien bleibt im regulären Fall erhalten. Sobald wir aber die regulären Modelle verlassen, sind die Maße, wie das AIC, DIC und BIC nicht mehr korrekt.

Bevor wir die beiden Informationskriterien ausführlich bereden, wird zunächst eine Grundlage für die Definitionen und Theoreme gegeben, auf die die beiden Informationskriterien beruhen.

4.1 Grundlagen

Dieser Abschnitt führt in die wichtigsten Definitionen der statistischen Modellevaluation ein. Die Definitionen, die in diesem Abschnitt vorgestellt werden, bauen das Fundament für die beiden Informationskriterien WAIC und WBIC auf.
Sei $\mathbb{R}^N$ ein N-dimensionaler euklidischer Raum. Ein statistisches Modell wird durch eine Wahrscheinlichkeitsdichtefunktion

$$p(x|\theta), \quad x \in \mathbb{R}^N,$$

für einen gegebenen Parameter $\theta \in \Theta \subset \mathbb{R}^d$ repräsentiert. Dabei stellt Θ die Menge aller Parameter dar. Eine A Priori-Wahrscheinlichkeitsdichtefunktion wird durch

$$\varphi(\theta), \quad \theta \in \Theta$$

dargestellt. Die Trainingsdaten $X_1, X_2, ..., X_n$, $X_i \in \mathbb{R}^N$, seien unabhängig und identisch verteilt mit der selben Wahrscheinlichkeitsdichtefunktion wie

$$q(x), \quad x \in \mathbb{R}^N.$$

Hier wird $q(x)$ als die wahre Verteilung bezeichnet. Die Notation X^n wird benutzt um alle Trainingsdaten darzustellen:

$$X^n = (X_1, X_2, ..., X_n)\,.$$

Weiter definieren wir die Entropie und die empirische Entropie einer wahren Verteilung wie folgt:

$$S = -\int q(x) \log q(x) dx$$
$$S_n = -\frac{1}{n} \sum_{i=1}^{n} \log q(X_i).$$

Die Entropie ist ein Maß für die Information einer Dichte (Shannoninformation).
Die Posteriori-Verteilung in diesem Kontext ist definiert durch

$$P(\theta|X^n) = \frac{1}{Z_n} \varphi(\theta) \prod_{i=1}^{n} p(X_i|\theta),$$

wobei Z_n die Normalisierungskonstante ist.
Die log-Verlustfunktion einer wahren Verteilung $q(x)$ und die empirische log-Verlustfunktion (bzw. minus log-Likelihood) sind definiert durch

$$L(\theta) = -\int q(x) \log p(x|\theta) dx$$
$$L_n(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \log p(X_i|\theta).$$

Eine Verlustfunktion ordent jeder Entscheidung einen Verlust zu. Weiter kann man für die durchschnittliche log-Verlustfunktion $L(\theta)$ folgendes schreiben:

$$L(\theta) = S + D(q||p_\theta).$$

Dabei ist $D(q||p_\theta)$ die Kullback-Leiber Distanz:

$$D(q||p_\theta) = \int q(x) \log \frac{q(x)}{p(x|\theta)} dx.$$

Die Kullback-Leibler Distanz spielt eine zentrale Rolle in der Bewertung und Vergleich von Modellen, das es als die Distanz zwischen der wahren Verteilung $q(x)$ und dem statistischen Modell $p(x|\theta)$ verstanden werden kann. Viele Informationskriterien, wie das AIC, versuchen diese Distanz zu verringern. Da die Kullback-Leibler-Distanz stets positiv ist, also $D(q||p_\theta) \geq 0$, folgt $L(\theta) \geq S$. Daraus kann man ableiten, dass $L(\theta) = S$ nur genau dann gilt, wenn $p(x|\theta) = q(x)$. Diese Beziehung wird uns später bei der Definiion von regulären und singulären statistischen Modellen hilfreich sein.
Wir nehmen an dieser Stelle an, dass es einen Parameter θ_0 im Inneren von Θ gibt, welches die durchschnittliche log-Verlustfunktion $L(\theta)$ minimiert:

$$L(\theta_0) = \min_{\theta \in \Theta} L(\theta).$$

Im regulären Fall wäre θ_0 der Maximum-Likelihood Schätzer. Aber zu beachten ist, dass solch ein θ_0 im Allgemeinen nicht eindeutig ist, weil die Abbildung $\theta \mapsto p(x|\theta)$ in singulären statistischen Modellen allgemein nicht injektiv ist.
Weiter nehmen wir an, dass für ein beliebiges θ, welches die Gleichung $L(\theta) = L(\theta_0)$ erfüllt, $p(x|\theta)$ die selbe Wahrscheinlichkeitsdichtefunktion ist. Sei $p_0(x)$ solch eine eindeutige Wahrscheinlichkeitsdichtefunktion, dann besteht im Allgemeinen die Menge

$$\Theta_0 = \{\theta \in \Theta : p(x|\theta) = p_0(x)\}$$

nicht aus einem einzigen Element. In regulären Fällen besteht diese Menge genau aus einem einzigen Element, undzwar aus dem ML-Schätzer. Sollte diese Menge nicht aus einem einzigen Parameter bestehen, was in der Praxis sehr häufig vorkommt, dann ist der ML-Schätzer nicht eindeutig und wir hätten eine Singularität vorliegen.
Eine weiter wichtige Funktion ist die log density ratio Funktion

$$f(x, \theta) = \log \frac{p_0(x)}{p(x|\theta)} \Leftrightarrow p(x|\theta) = p_0(x) \exp(-f(x, \theta)),$$

mit der wir in der Lage sind die durchschnittliche log-Likelihood Ratio $K(\theta)$ und die empirische log-Likelihood Ratio Funktion $K_n(\theta)$ zu definieren:

$$K(\theta) = \int q(x) f(x, \theta) dx$$
$$K_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} f(X_i, \theta).$$

Die beiden Funktionen spielen bei der Schätzung der Bayes-Free-Energy $\mathcal{F}$ und somit bei der Berechnung Vom WBIC eine wichtige Rolle.
Die Beziehung zwischen den log-Verlustfunktionen und den log-Likelihood Ratios sind wie folgt darstellbar:

$$L(\theta) = L(\theta_0) + K(\theta)$$
$$L_n(\theta) = L_n(\theta_0) + K_n(\theta).$$

Der Erwartungswert über alle Mengen der Trainingsdaten $X_1, X_2, ..., X_n$ wird durch $\mathbb{E}[]$ gekennzeichnet:

$$\mathbb{E}[L_n(\theta)] = L(\theta)$$
$$\mathbb{E}[K_n(\theta)] = K(\theta).$$

Das Hauptziel dieser Arbeit ist zu zeigen, dass die Bayes-Free-Energy $\mathcal{F}$ durch das Widely

Applicable Bayesian Information Criterion (WBIC) approximierbar ist

$$\mathcal{F} = WBIC + O_p(\sqrt{\log n})$$

und, dass der Erwartungswert vom Generalization Error G approximativ mit dem Erwartungswert vom WAIC übereinstimmt

$$\mathbb{E}[G] = \mathbb{E}[WAIC] + O_p(\frac{1}{n^2}).$$

4.2 Reguläre und Singuläre statistische Modelle

Bevor die Theorie der beiden neuen Informationskriterien WAIC und WBIC besprochen wird, sind wir nun in der Lage den Unterschied zwischen regulären und singulären Modellen formal zu definieren. Weiter unterscheidet man zwischen Realisierbarkeit und Unrealisierbarkeit eines statistischen Modells.

Definition: Regulär und Singulär

(1) Wenn es einen Parameter θ_0 gibt, sodass die wahre Verteilung $q(x)$ dem statistischen Modell $p(x|\theta_0)$ entspricht, also $q(x) = p(x|\theta_0)$, dann sagt man, dass eine wahre Verteilung $q(x)$ realisierbar ist durch ein statistisches Modell $p(x|\theta_0)$. Andernfalls wird es unrealisierbar genannt [5].
(2) Wenn die Menge der optimalen Parameter Θ_0 aus einem einzigen Element θ_0 besteht und, wenn die Hesse-Matrix

$$J_{ij}(\theta) = \frac{\partial^2}{\partial\theta_i\partial\theta_j}L(\theta)$$

für $\theta = \theta_0$ positiv definit ist, dann sagt man, dass $q(x)$ regulär für $p(x|\theta)$ ist. Andernfalls wird $q(x)$ singulär für das statistische Modell $p(x|\theta)$ genannt [5].

Bemerkung:

(1) Die Matrix $J(\theta)$ ist gleich der Hesse-Matrix von $K(\theta)$ und $J(\theta_0)$ ist gleich der Fisher-Informationsmatrix, wenn die wahre Verteilung durch ein statistisches Modell realisierbar ist.

(2) Wenn $q(x)$ durch $p(x|\theta)$ realisierbar ist, dann ist $K(\theta)$ die Kullback-Leibler-Distanz von $q(x)$ und $p(x|\theta)$.

(3) All diese Bemerkungen gelten nicht, wenn $q(x)$ durch $p(x|\theta)$ nicht realisierbar ist.

4.3 Das Widely Applcable Bayesian Information Criterion [WBIC]

Nun haben wir die nötigen Grundlagen um die Absichten und die Theorie der Widely Applicable Bayesian Information Criterion (WBIC) und Widely Applicable Information Criterion (WAIC)

darzustellen.

Das Widely Applicable Bayesian Information Criterion (WBIC) versucht die Bayes-Free-Energy

$$\mathcal{F} = -\log \int \prod_{i=1}^{n} p(X_i|\theta)\varphi(\theta)d\theta$$

zu schätzen. Der Wert von $\mathcal{F}$ kann als das $-\log$(Die marginale Likelihood von einem Model und einem Priori) verstanden werden und spielt deshalb in der Bewertung von statistischen Modellen eine enorm wichtige Rolle. Ein Modell oder eine Priori wird oft durch Maximierung der Bayes Marginal Likelihood optimiert, welches äquivalent zur Minimierung der Bayes-Free-Energy $\mathcal{F}$ ist [5]. Um in der Lage zu sein diese Größe zu approximieren brauchen wir die Singular Learning Theory, die von Sumio Watanabe eingeführt worden ist.
Die Singular Learning Theory baut auf vier grundlegende Bedingungen auf, die uns die Singularitäten in statistischen Modellen erlauben werden. Mihilfe dieser Annahmen bekommen wir die wichtigen Eigenschaften eines Informationskriteriums, damit es auch in singulären Modellen funktioniert. Das heißt wir betrachten den hauptsächlich den Fall, dass die Menge der wahren Parameter $\{\theta \in \Theta : q(x) = p(x|\theta)\}$ mehr als aus einem Element besteht und, dass die Fisher-Informationsmatrix nicht positiv definit ist.

Grundlegende Bedingungen:

(1) Die Menge der Parameter Θ ist eine kompakte Menge in $\mathbb{R}^d$, dessen Inneres nicht die leere Menge $\emptyset$ ist. Seine Grenzen werden durch mehrere analytische Funktionen $\pi_1(\theta), \pi_2(\theta), ..., \pi_k(\theta)$ definiert:

$$\Theta = \left\{\theta \in \mathbb{R}^d : \pi_1(\theta) \geq 0, \pi_2(\theta) \geq 0, ..., \pi_k(\theta) \geq 0\right\}.$$

(2) Die Priori-Verteilung erfüllt $\varphi(\theta) = \varphi_1(\theta)\varphi_2(\theta)$, wo $\varphi_1(\theta) \geq 0$ eine analytische Funktion ist und $\varphi_2(\theta) > 0$ eine C^∞-Class Funktion ist.

(3) Sei $s \geq 6$ und

$$L^s(q) = \left\{ f(x) : \|f\|_s \equiv \left(\int |f(x)|^s q(x)dx \right)^{1/s} < \infty \right\}$$

ein Banach-Raum. Dann gibt es eine offene Menge $\Theta' \supset \Theta$, sodass die Abbildung $\Theta' \ni \theta \mapsto f(x, \theta)$ eine $L^s(q)$-wertige analytische Funktion ist.

(4) Die Menge Θ_ϵ ist definiert durch

$$\Theta_\epsilon = \{\theta \in \Theta : K(\theta) \leq \epsilon\}.$$

Es wird angenommen, dass es konstanten $\epsilon, c > 0$ gibt, sodass gilt:

$$(\forall\theta \in \Theta_\epsilon) : \mathbb{E}_X\left[f(X,\theta)\right] \geq c \cdot \mathbb{E}_X\left[f(X,\theta)^2\right]. \tag{11}$$

Bemerkung zu den grundlegenden Bedingungen (1)-(4):
(1) Diese Bedingungen erlauben, dass die Menge der optimalen Parameter

$$\Theta_0 = \{\theta \in \Theta : p(x|\theta) = p(x|\theta_0)\} = \{\theta \in \Theta : K(\theta) = 0\}$$

Singularitäten beinhalten können - d.h. Θ_0 besteht mehr aus einem Element - und, dass die Hesse-Matrix $J(\theta)$ bei $\theta \in \Theta_0$ nicht positiv definit ist. Das heißt, dass wir uns nicht auf die positiv Definitheit der Fisher-Informationsmatrix und die Eindeutigkeit des optimalen Parameters stützen, wenn wir die beiden Informationskriterien definieren.
(2) Die Bedingung, dass Θ kompakt ist, ist notwendig um den Fall $|\theta| = \infty$ zu umgehen. Denn wenn Θ nicht kompakt ist und Θ_0, $|\theta| = \infty$ enthält, dann existiert der Maximum-Likelihood Schätzer allgemein nicht. Aber da wir die Vergleichbarkeit der beiden Informationskriterien WAIC und WBIC bewahren wollen wird Θ als kompakt angenommen.
(3) Die grundlegende Bedingung 2 ist für den Beweis der Asymptotik notwedig. Beim WBIC kommt man zu einer Abweichung von $O_p(\sqrt{\log n})$ zwischen der Bayes-Free-Energy $\mathcal{F}$ und WBIC; beim WAIC kommt man zu einer Abweichung von $O_p(\frac{1}{n^2})$ zwischen den beiden Erwartungswerten vom Generalization Error G und WAIC.
(4) Bedingung 3 wird für den Beweis von WAIC gebraucht um die Existenz der asymptotischen Erweiterung des Bayes Generalization Errors zu gewährleisten.
Durch diese Annahmen lassen sich eine Reihe von Theoreme und Sätze ableiten und beweisen. Ein Ergebnis dieser Annahmen ist die allgemeine Darstellung der Bayes-Free-Energy $\mathcal{F}$ in regulären und auch in singulären statistischen Modellen.

Theorem 1 Ängenommen die grundlegenden Konditionen (1)-(4) sind erfüllt. Dann gilt die folgende Gleichung

$$\mathcal{F} = nL_n(\theta_0) + \lambda \log n - (m-1)\log\log n + R_n,$$

wo θ_0 der Parameter ist, welches die Kullback-Leibler Distanz zwischen der wahren Verteilung $q(x)$ und dem statistischen Modell $p(x|\theta)$ minimiert; wo λ eine rationale Zahl ist, welches als das Real Log Canonical Threshold (RLCT) bekannt ist und m seine Multiplizität; und wo R_n eine Sequenz von Zufallsvariablen ist, welches für $n \to \infty$ gegen eine Zufallsvariable konvergiert"[5].

Diese Darstellung der Bayes-Free-Energy gilt auch wenn eine wahre Verteilung $q(x)$ unrealisierbar ist durch ein statistisches Modell $p(x|\theta)$.
Das RLCT ist eine bekannte Größe in der algebraischen Geometrie und Analysis. Wenn eine

wahre Verteilung, ein statistisches Modell und ein A Priori-Verteilung festgesetzt sind, dann gibt es eine algebraisch geometrische Prozedur, welches uns erlaubt ein RLCT für diesen speziellen Fall zu finden. Das RLCT zeigt allgemein das theoretische Verhalten der Bayes-Free-Energy $\mathcal{F}$ und dem Generalization Loss G, welches wir später für das WAIC brauchen werden. Denn mithilfe des Real Log Canonical Thresholds ist der quantitative Unterschied zwischen singulären und regulären Modellen deutlich zu erkennen. In praktischen Anwendungen kennen wir allerdings nicht die wahre Verteilung, daher ist das RLCT und m unbekannt. Deshalb können wir Theorem 1 in solchen Fällen nicht direkt anwenden. Das Ziel wäre dann eine neue Methode zu finden um $\mathcal{F}$ auch ohne jede Information über eine wahre Verteilung schätzen zu können.
Da das RLCT und seine Multiplizität eine eigene große Theorie darstellen und die Darstellung den Rahmen dieser Arbeit sprengen würde, wird es im Folgenden als bekannt vorausgesetzt.

Eine Möglichkeit $\mathcal{F}$ zu schätzen ohne die wahre Verteilung zu kennen, ist die Verwendung des Widely Applicable Bayesian Information Criterion:

$$WBIC = \mathbb{E}_\theta^\beta \left[nL_n(\theta) \right], \quad \beta = \frac{1}{\log n},$$

wobei $\mathbb{E}_\theta^\beta$ [] den Erwartungswert über die Posteriori-Verteilung auf Θ zeigt. Dieser Erwartungswert ist für eine beliebig integrierbare Funktion $G(\theta)$ wie folgt definiert

$$\mathbb{E}_\theta^\beta[G(\theta)] = \frac{\int G(\theta) \prod_{i=1}^n p(X_i|\theta)^\beta \varphi(\theta) d\theta}{\int \prod_{i=1}^n p(X_i|\theta)^\beta \varphi(\theta) d\theta}.$$

In dieser Definition wird $\beta > 0$ die inverse temperature genannt [5]. Das Ziel, welches wir jetzt endlich zeigen wollen ist

$$\mathcal{F} = WBIC + O_p(\sqrt{\log n}).$$

Das erste Theorem, welches uns helfen wird die obere Gleichung zu bestätigen ist Theorem 2, denn es behauptet, dass es ein eindeutiges β^* gibt, welches die Gleichung

$$\mathcal{F} = \mathbb{E}_\Theta^{\beta^*} \left[nL_n(\theta) \right]$$

erfüllt.

Theorem 2 Angenommen $L_n(\theta)$ ist keine konstante Funktion von θ. Dann gelten die folgenden Sätze:
(1) Der Wert $\mathbb{E}_\theta^\beta \left[nL_n(\theta) \right]$ ist eine fallende Funktion von β.
(2) Es existiert ein eindeutiges β^* $(0 < \beta^* < 1)$, welches die folgende Gleichung erfüllt [5]

$$\mathcal{F} = \mathbb{E}_\Theta^{\beta^*} \left[nL_n(\theta) \right]. \tag{12}$$

Anzumerken ist, dass $L_n(\theta)$ in gewöhnlichen statistischen Modellen nie konstant ist. Der eindeutige Parameter β^*, das die Gleichung (12) erfüllt wird der optimale inverse temperature genannt. Als zweites wird Theorem 3 vorgestellt. Die Gleichung in Theorem 3 ist das Hauptergebnis der Arbeit von Watanabe in der Singular Learning Theory, denn es zeigt, dass das WBIC die selben asymptotischen Eigenschaften wie die Bayes-Free-Energy hat auch wenn das statistische Modell singulär ist.

Theorem 3 Angenommen die grundlegenden Bedingungen (1)-(4) sind erfüllt und

$$\beta = \frac{\beta_0}{\log n},$$

wobei β_0 eine Konstante ist. Dann gibt es eine Zufallsvariable U_n, sodass

$$\mathbb{E}_\theta^\beta \left[nL_n(\theta) \right] = nL_n(\theta_0) + \frac{\lambda \log n}{\beta_0} + U_n \cdot \sqrt{\frac{\lambda \log n}{2\beta_0}} + O_p(1),$$

wo λ das RLCT ist und $\{U_n\}$ eine Sequenz von Zufallsvariablen ist, dessen Erwartungswert 0 ist, also $\mathbb{E}\left[U_n\right] = 0$ und es konvergiert für $n \to \infty$ gegen eine normalverteilte Zufallsvariable [5]. Außerdem gilt, wenn eine wahre Verteilung $q(x)$ realisierbar durch ein statistisches Modell $p(x|\theta)$ ist, dann ist $\mathbb{E}\left[(U_n)^2\right] < 1$.
Theorem 3 mit $\beta_0 = 1$ zeigt die asymptotische Äquivalenz der Bayes-Free-Energy und WBIC [5]

$$WBIC = nL_n(\theta_0) + \lambda \log n + U_n \cdot \sqrt{\frac{\lambda \log n}{2}} + O_p(1).$$

Die ersten beiden Terme von WBIC ähneln den ersten beiden Terme der Bayes-Free-Energy.
Aus Theorem 4 und seinen Beweisen kommen zwei wichtige Korollare hervor:

Korollar 1: Wenn die Parität eines statistischen Modells ungerade ist, $\mathcal{Q}(K(\theta), \varphi(\theta)) = 1$, dann ist $U_n = 0$ [5]. Da die Parität eines statistischen Modells zu sehr in die Theorie der algebraischen Geometrie eingeht ist wichtig für den Leser zu wissen, dass die Parität eines regulären Modells immer ungerade ist.

Korollar 2: Sei β^* der optimale inverse temperature. Dann gilt

$$\beta^* = \frac{1}{\log n}\left(1 + \frac{U_n}{\sqrt{2\lambda \log n}} + o_p\left(\frac{1}{\sqrt{\log n}}\right)\right).$$

Und zuletzt wollen wir noch die Gleichheit zwischen dem üblichen Informationskriterium BIC und dem neuen Informationskriterium WBIC in regulären statistischen Modellen zeigen.

Lemma 1 Angenommen die grundlegenden Konditionen (1)-(4) sind erfüllt und eine wahre Verteilung $q(x)$ ist regulär für einen statistischen Modell $p(x|\theta)$. Wenn θ_0 im Inneren von Θ enthalten ist und wenn $\varphi(\theta_0) > 0$, dann gilt [5]

$$\lambda = \frac{d(AnzahlParameter)}{2}, \quad m = 1,$$

und

$$\mathcal{Q}(K(\theta), \varphi(\theta)) = 1.$$

Durch Einsetzen der Ergebnisse in Lemma 1 in die Formel von WBIC erhalten wir das folgende Theorem:

Theorem 4 Wenn eine wahre Verteilung $q(x)$ regulär für ein statistisches Modell $p(x|\theta)$ ist, dann gilt

$$WBIC = nL_n(\hat{\theta}) + \frac{d}{2}\log n + O_p(1). \tag{13}$$

Dabei ist $\hat{\theta}$ der Maximum-Likelihood Schätzer und d die Anzahl der Parameter im Modell. Dieses Theorem zeigt, dass der Unterschied von WBIC und BIC kleiner als ein konstanter Term ist, wenn eine wahre Verteilung regulär für ein statistisches Modell ist [5]. Dieses Theorem gilt ebenfalls wenn eine wahre Verteilung $q(x)$ unrealisierbar ist durch $p(x|\theta)$.
Da nun in regulären statistischen Modellen für WBIC die Gleichung (13) gilt, und die rechte Seite der Gleichung dem BIC

$$BIC = nL_n(\hat{\theta}) + \frac{d}{2}\log n$$

entspricht und zusätzlich WBIC auch in singulären statistischen Modellen anwendbar ist, sagt man, dass WBIC eine Generalisierung des BIC ist. Ein weiterer wichtiger Vorteil des WBIC in praktischer Hinsicht ist, dass der Rechenaufwand für die numerische Berechnung des WBIC weitaus kleiner ist als die für die Bayes-Free-Energy.

4.4 Das Widely Applicable Information Criterion [WAIC]

In diesem Abschnitt kommen wir zum zweiten wichtigen Informationskriterium: dem Widely Applicable Information Criterion (WAIC). Das WAIC vertritt einen komplett unterschiedlichen Ansatz als das WBIC. Denn es versucht nicht die Bayes-Free-Energy, und somit das wahre Modell zu identifizieren, sondern versucht den durchschnittlichen Prognosefehler zu schätzen um somit das Modell mit der besten Prädiktion auszuwählen. In regulären statistischen Modellen könnte man für diese Absicht den AIC oder DIC verwenden. Aber in singulären Modellen ist die Verwendeung von AIC und DIC, wie in der Einleitung schon erläutert, nicht anwendbar. Deshalb bietet das WAIC in singulären Fällen ein gutes Werkzeug für den Vergleich unterschiedlicher

Modelle.
Das Ziel in diesem Abschnitt ist eine Methode vorzustellen, das den Generalization Error aus dem Training Error ohne jegliche Information über die wahre Verteilung und unabhängig von Singularitäten schätzt.
Eine Möglichkeit ans Ziel zu kommen ist der Widely Applicable Information Criterion. Es werden durchgehend die folgenden vier Fehler beobachtet und versucht die Beziehungen unter diesen Fehlern zu zeigen: (1) der Bayes Generalization Error B_g, (2) der Bayes Training Error B_t, (3) der Gibbs Generalization Error G_g und (4) der Gibbs Traing Error G_t [2]:

(1) Bayes Generalization Error

$$B_g = \mathbb{E}_X \left[\log \frac{q(X)}{\mathbb{E}_\theta[p(X|\theta]} \right].$$

(2) Bayes Training Error

$$B_t = \frac{1}{n} \sum_{j=1}^{n} \log \frac{q(X_j)}{\mathbb{E}_\theta[p(X_j|\theta)]}.$$

(3) Gibbs Generalization Error

$$G_g = \mathbb{E}_\theta \left[\mathbb{E}_X \left[\log \frac{q(X)}{p(X|\theta)} \right] \right].$$

(4) Gibbs Training Error

$$G_t = \mathbb{E}_\theta \left[\frac{1}{n} \sum_{j=1}^{n} \log \frac{q(X_j)}{p(X_j|\theta)} \right].$$

Da diese vier Fehler messbare Funktionen auf X^n sind, sind sie ebenfalls Zufallsvariablen [2], die wir im folgenden untersuchen möchten. Der Bayes Generalization Error ist gleich der Kullback-Leiber Distanz zwischen der wahren Verteilung $q(x)$ und der Bayes prädiktiven Verteilung $p(x|X^n) = \mathbb{E}_\theta [p(x|\theta)]$. Der Erwartungswert dieses Fehler setzt sich aus der Entropie der wahren Verteilung und der Kullback-Leibler Distanz zwischen der wahren Verteilung und der prädiktiven Verteilung zusammen (natürlich als Summe) [2]. Der Gibbs Generalization Error ist gleich dem Mittelwert der Kullback-Leibler Distanz zwischen der wahren Verteilung und der Gibbs Schätzung. In der Praxis ist der Bayes Generalization Error wichtiger, aber darauf wird später nochmal eingegangen. Beide Fehler zeigen die Genauigkeit der Bayes und Gibbs Schätzungen. Für statistische Modelle ist es sehr wichtig in der Lage zu sein diese Fehler aus den (gegebenen) Zufallsstichproben zu schätzen.
Auch das WAIC baut auf die grundlegenden mathematischen Bedingungen (1)-(4), die im Abschnitt 4.3 vorgestellt wurden. Basierend auf diesen Annahmen wurden folgende Ergebnisse

bewiesen [2]:

Theorem 5

(1) Es existieren Zufallsvariablen B_g^*, B_t^*, G_g^*, und G_t^*, sodass für $n \to \infty$, die folgenden Konvergenzen in Wahrscheinlichkeit gelten:

$$nB_g \to B_g^*, \quad nB_t \to B_t^*, \quad nG_g \to G_g^*, \quad nG_t \to G_t^*.$$

(2) Für $n \to \infty$ gilt die Konvergenz in Wahrscheinlichkeit

$$n(B_g - B_t - G_g + G_t) \to 0.$$

(3) Die Erwartungswerte der vier Fehler konvergieren wie folgt,

$$\begin{aligned} \mathbb{E}\left[nB_g\right] &\to \mathbb{E}\left[B_g^*\right], \quad \mathbb{E}\left[nB_t\right] \to \mathbb{E}\left[B_t^*\right] \\ \mathbb{E}\left[nG_g\right] &\to \mathbb{E}\left[G_g^*\right], \quad \mathbb{E}\left[nG_t\right] \to \mathbb{E}\left[G_t^*\right]. \end{aligned}$$

Das folgende Theorem ist das Hauptergebnis der zweiten Arbeit von Watanabe [2]. Der Beweis für das nächste Theorem ist genau so wichtig wie das Theorem 3 im Abschnitt 4.3, welches als Grundlage für das WBIC gedient hat. So dient das folgende Theorem als Grundlage für das WAIC:

Theorem 6 Es gelten die folgenden Gleichung

$$\begin{aligned} \mathbb{E}\left[B_g^*\right] - \mathbb{E}\left[B_t^*\right] &= 2\beta\left(\mathbb{E}\left[G_t^*\right] - \mathbb{E}\left[B_t^*\right]\right), \\ \mathbb{E}\left[G_g^*\right] - \mathbb{E}\left[G_t^*\right] &= 2\beta\left(\mathbb{E}\left[G_t^*\right] - \mathbb{E}\left[B_t^*\right]\right). \end{aligned}$$

Theorem 6 behauptet also, dass die erhöhten Fehler vom Training zur Prädiktion (zwischen Training Error und Generalization Error) in Proportion zu der Differenz zwischen dem Bayes und Gibbs Training Errors sind [2]. Theorem 6 gilt für jede wahre Verteilung, für jedes statistische Modell, für jede A Priori-Verteilung und für jede Singularität. Ein wichtiger Vorteil gegenüber dem WBIC ist, dass Theorem 6 auch gilt wenn die wahre Verteilung nicht in dem parametrischen Modell enthalten ist. Eine weitere wichtige Eigenschaft ist die Invarianz zwischen dem Generalization und Training Error [2]:

$$\mathbb{E}\left[G_g^*\right] - \mathbb{E}\left[B_g^*\right] = \mathbb{E}\left[G_t^*\right] - \mathbb{E}\left[B_t^*\right].$$

Aus Theorem 6 sind wir nun in der Lage die beiden Generalization Errors zu schätzen.

Korollar 3 Die zwei Generalization Errors können durch die zwei Training Errors geschätzt werden [2],

$$\begin{pmatrix} \mathbb{E}\left[B_g^*\right] \\ \mathbb{E}\left[G_g^*\right] \end{pmatrix} = \begin{pmatrix} 1-2\beta & 2\beta \\ -2\beta & 1+2\beta \end{pmatrix} \begin{pmatrix} \mathbb{E}\left[B_t^*\right] \\ \mathbb{E}\left[G_t^*\right] \end{pmatrix}.$$

Wenn das statistische Modell nun regulär ist, also wenn die Menge der wahren Parameter $\Theta_0 = \{\theta \in \Theta : q(x) = p(x|\theta)\}$ nicht aus einem einzigen Punkt besteht und wenn die Fisher-Informationsmatrix immer positiv definit ist, gelten die folgenden Generalization und Training Errors:

$$\begin{aligned} \mathbb{E}\left[B_g^*\right] &= \frac{d}{2}, \quad \mathbb{E}\left[G_g^*\right] = (1+\frac{1}{\beta})\frac{d}{2}, \\ \mathbb{E}\left[B_t^*\right] &= -\frac{d}{2}, \quad \mathbb{E}\left[G_t^*\right] = (-1+\frac{1}{\beta})\frac{d}{2}. \end{aligned}$$

Somit hat Sumio Watanabe die allgemeingültige Beziehung unter diesen vier Fehlern aufgedeckt. Basierend auf Korollar 3 entwickelte Sumio Watanabe das neue Informationskriterium, welches für reguläre und singuläre Modelle einsetzbar ist. Bevor wir die beiden Definitionen von WAIC angeben, definieren wir zunächst den Bayes Generalization Loss, den Bayes Training Loss, den Gibbs Generalization Loss und den Gibbs Training Loss durch

$$\begin{aligned} BL_g &= -\mathbb{E}_X\left[\log \mathbb{E}_\theta\left[p(X|\theta)\right]\right], \\ BL_t &= -\frac{1}{n}\sum_{j=1}^{n} \log \mathbb{E}_\theta\left[p(X_j|\theta)\right], \\ GL_g &= -\mathbb{E}_\theta \mathbb{E}_X\left[\log p(X|\theta)\right], \\ GL_t &= -\mathbb{E}_\theta\left[\frac{1}{n}\sum_{j=1}^{n} \log p(X_j|\theta)\right]. \end{aligned}$$

Beide Training Losses BL_t und GL_t können numerisch anhand von Trainingsdaten X^n und einem statistischen Modell $p(x|\theta)$ ohne jede Kenntnis der wahren Verteilung $q(x)$ berechnet werden. Durch Kombinieren der Entropie der wahren Verteilung S mit Korollar 1 [2]

$$S = -\int q(x)\log q(x)dx = -\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\log q(X_i)\right],$$

erhalten wir die Gleichung

$$\begin{aligned} \mathbb{E}\left[BL_g\right] &= \mathbb{E}\left[BL_t\right] + 2\beta(\mathbb{E}\left[GL_t\right] - \mathbb{E}\left[BL_t\right]) + o(\frac{1}{n}) \\ \mathbb{E}\left[GL_g\right] &= \mathbb{E}\left[GL_t\right] + 2\beta(\mathbb{E}\left[GL_t\right] - \mathbb{E}\left[BL_t\right]) + o(\frac{1}{n}). \end{aligned}$$

Basierend auf den theoretischen Ergebnissen, die bis jetzt vorgeführt wurden, sind wir in der Lage die zwei Versionen des WAIC zu definieren [2]:

$$WAIC_1 = BL_t + 2\beta(GL_t - BL_t),$$
$$WAIC_2 = GL_t + 2\beta(GL_t - BL_t).$$

Viel wichtiger sind die Beziehungen der Erwartungswerte, auf die sich die Praxis beruht:

$$\mathbb{E}\left[BL_g\right] = \mathbb{E}\left[WAIC_1\right] + o(\frac{1}{n}),$$
$$\mathbb{E}\left[GL_g\right] = \mathbb{E}\left[WAIC_2\right] + o(\frac{1}{n}).$$

Die beiden Erwartungswerte gelten auch wenn eine wahre Verteilung unrealisierbar ist durch oder singulär für ein statistisches Modell $p(x|\theta)$ ist. Da die beiden Informationskriterien den allgemeinen Fehler eines statistischen Modells schätzen, stellen sie präzise Kennziffern für die Modellwahl zur Verfügung.
Wie beim WBIC wollen wir hier zum Schluss WAIC mit den üblichen Informationskriterien vergleichen. Wenn eine wahre Verteilung $q(x)$ realisierbar ist durch und regulär für ein statistisches Modell $p(x|\theta)$ ist, dann gilt für den Bayes Generalization Loss

$$\mathbb{E}[BL_g] = S + \frac{d}{2n} + o(\frac{1}{n}).$$

In solch einem Fall gilt für AIC und DIC ebenfalls

$$\mathbb{E}[AIC] = S + \frac{d}{2n} + o(\frac{1}{n}),$$
$$\mathbb{E}[DIC] = S + \frac{d}{2n} + o(\frac{1}{n}),$$

da wenn ein Modell regulär ist und die wahre Verteilung im parametrischen Modell enthalten ist, dann ist $\lambda = d/2$ [2].

5 Vergleich und Anwendungsbeispiel

In diesem Kapitel werden die beiden neuen Informationskriterien WAIC und WBIC mit den üblichen Informationskriterien verglichen. Es wurden im Laufe dieser Arbeit viele Zusammenhänge zwischen den Informationskriterien gezeigt und diskutiert. Deswegen soll dieses Kapitel einen Überblick über die wichtigsten Eigenschaften und Beziehungen der hier besprochenen Informationskriterien darstellen. Zusätzlich zum Vergleich wird versucht ein Anwendungsbeispiel zu zeigen, welches die Beziehungen zwischen den Kriterien bestätigen sollen. Da die vorherigen Kapitel sehr theorielastig war, wird hier nochmal kurz das Ziel der beiden Informationskriterien WAIC und WBIC gezeigt.

Wie schon durch die ganze Arbeit erwähnt ist der Hauptunterschied zwischen den üblichen Informationskriterien (AIC, DIC, BIC) und den beiden neuen Informationkriterien (WAIC, WBIC), dass WAIC und WBIC in regulären und auch in singulären statistischen Modellen anwendbar sind, während die üblichen Kriterien nur in regulären Fällen gelten.

Wenn man ins Detail geht, kann man das WAIC mit dem AIC und DIC vergleichen, da alle drei Informationskriterien versuchen den allgemeinen Prognosefehler eines Modells zu schätzen. Das WBIC kann man mit dem BIC vergleichen, da beide Kriterien versuchen die Bayes marginale Likelihood zu maximieren und somit das wahre Modell finden.

Wir unterscheiden bei jedem Vergleich zwischen einem regulären und einem singulären Fall.

5.1 WAIC vs. AIC

In regulären statistischen Modellen favorisieren beide Kriterien das Modell mit den besten Prädiktionen. Beide sind im regulären Fall asymptotisch ähnlich wie wir anhand eines Beispiels auch zeigen werden. Dazu betrachten wir zunächste ein generalisiertes lineares Modell und gehen anschließend in eine gemischte Normalverteilung über. Das Ziel dieser Analyse ist nur zu beobachten, wann und wo die beiden Informationskriterien überinstimmen.

Zunächst betrachten wir eine generalisierte Regression mit Possonverteilung als Response.

```
> library(blmeco)
> data(pondfrog1)

> mod1 <- glm(frog ~ ph + waterdepth + temp, data=pondfrog1, family=poisson)
> mod2 <- glm(frog ~    + waterdepth + temp, data=pondfrog1, family=poisson)
> mod3 <- glm(frog ~ ph +            + temp, data=pondfrog1, family=poisson)
> mod4 <- glm(frog ~ ph + waterdepth       , data=pondfrog1, family=poisson)
> mod5 <- glm(frog ~ ph                    , data=pondfrog1, family=poisson)
> mod6 <- glm(frog ~       waterdepth      , data=pondfrog1, family=poisson)
> mod7 <- glm(frog ~                 temp  , data=pondfrog1, family=poisson)
> mod8 <- glm(frog ~ 1                     , data=pondfrog1, family=poisson)

> AIC(mod1, mod2, mod3, mod4, mod5, mod6, mod7, mod8)
      df   AIC
mod1   4 804.1726
mod2   3 808.7593
mod3   3 2336.8679
```

```
18    mod4     3 5459.0026
19    mod5     2 6385.1750
20    mod6     2 5484.1144
21    mod7     2 2347.4718
22    mod8     1 6447.8024
23
24  > WAIC(mod1, nsim=1000)
25  $lppd
26  [1] -397.9597
27
28  $pwaic1
29  [1] 4.238523
30
31  $pwaic2
32  [1] 4.64778
33
34  $WAIC1
35  [1] 804.3964
36
37  $WAIC2
38  [1] 805.2149
39
40  > WAIC(mod2, nsim=1000)$WAIC2
41  [1] 809.2873
42  > WAIC(mod3, nsim=1000)$WAIC2
43  [1] 2388.375
44  > WAIC(mod4, nsim=1000)$WAIC2
45  [1] 5597.056
46  > WAIC(mod5, nsim=1000)$WAIC2
47  [1] 6501.268
48  > WAIC(mod6, nsim=1000)$WAIC2
49  [1] 5565.28
50  > WAIC(mod7, nsim=1000)$WAIC2
51  [1] 2379.661
52  > WAIC(mod8, nsim=1000)$WAIC2
53  [1] 6508.053
```

Wie man gut erkennen kann, sind alle jeweiligen acht Ergebnisse ähnlich zueinander. AIC und auch WAIC würden in dem Fall Modell 1 auswählen. Die generalisierten Modelle gehören zu den regulären Modellen, deswegen war es zu erwarten, dass beide Kriterien die (fast) gleichen Werte ausgeben.
Einen kleinen Test wird anhand einer gemischten Normalverteilung gezeigt, wo beide Informationskriterien sich nicht mehr ähneln, weil wir weiter weg von den regulären Modellen gehen. Wie in der Einleitung erwähnt sind gemischte Verteilugen keine regulären Modelle mehr, deshalb wird mit dem folgenden Modell die Werte für die beiden Kriterien berechnet.

```
1  > library(lme4)
2  > lmer.co2 <- lmer(CO2emission ~ Area + PopulationRural + Population + GDP +
```

```
3     Livestock + Employees.M.IND + continent + (1|country), data=wdi.past)
4  > AIC(lmer.co2)
5  [1] -1225.535
6  > WAIC(lmer.co2)
7  $lppd
8  [1] 1119.85
9
10 $pwaic1
11 [1] 94.56311
12
13 $pwaic2
14 [1] 121.4224
15
16 $WAIC1
17 [1] -2050.574
18
19 $WAIC2
20 [1] -1996.856
```

Anhand dieses Beispiels wird deutlich, dass das AIC gegenüber gemischte Verteilungen nicht stabil ist, denn die Abweichung zwischen $WAIC_2$ und AIC ist zum Vergleich von generalisierten Modellen stark gestiegen.

5.2 WAIC vs. DIC

Als nächstes wollen wir die Informationskriterien WAIC und DIC vergleichen. In der Theorie sind beide asymptotisch äquivalent, wenn wir im regulären Fall sind. Dieser Vergleich ist sehr interessant, denn von dem Standpunkt der numerischen Berechnung aus sehen beide Kriterien sehr ähnlich aus.
Ein Simulationsbeispiel stellt Sumio Watanabe in seiner Homepage dar. Dafür verwendete er eine gemischte Normalverteilung mit Mischungsparameter $a \in (0, 1)$ und einem dreidimensionalem Parameter $\theta = (b_1, b_2, b_3)$:

$$p(x|a, b_1, b_2, b_3) = (1 - a)N(0, 0, 0, 1) + aN(b_1, b_2, b_3, 1).$$

Für den regulären Fall nehmen wir an, dass die wahre Verteilung $p(x|0.5, 3, 3, 3)$ regulär für den obigen statistischen Modell ist.

Wie man an der Abbildung 2 sehen kann sind $DIC2$ und $WAIC$ im regulären Fall sehr ähnlich und schätzen den Generalization Error sehr exakt. Man kann sagen, dass im regulären Fall folgendes gilt:

$$WAIC = DIC1 + o_p(1) = DIC2 + o_p(1)$$

Wenn man die ganze Problematik für den Fall betrachtet, dass die wahre Verteilung $p(x|0.5, 0, 0, 0)$ singulär für den obigen statistischen Modell ist, dann erkennt man an der Abbildung 3, dass

Abbildung 2: Vergleich von WAIC, DIC1 und DIC2 im regulären Fall
Quelle: http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/dicwaic.html

die Informationskriterien sehr unterschiedlich sind. Dabei überschätzen die beiden Version vom DIC den Generalization Error, während WAIC als Schätzer sehr geeignet ist, weil es den Generalization Error im Mittel korrekt wiedergibt.

Das bedeutet für singuläre Modelle, dass die beiden Versionen vom DIC nicht anwendbar sind und für WAIC gilt:

$$\mathbb{E}[G] = \mathbb{E}[WAIC] + o(\frac{1}{n^2}).$$

5.3 WBIC vs. BIC

Als letztes kommt der Vergleich von WBIC und BIC. Dieser Vergleich wurde im ABschnitt 4.3 ausführlich besprochen. Im regulären Fall gilt, dass die beiden Schätzer BIC und WBIC asymptotisch gleich sind, denn es gilt dann:

$$WBIC = BIC + o_p(1).$$

Und auch hier ist können wir sagen, dass das BIC für singuläre Modelle nicht anwendbar ist, aber das WBIC schon. Die Ergebnisse würden genau die selben sein, die wir für die Vergleiche zwischen dem WAIC und AIC, DIC erhalten haben.

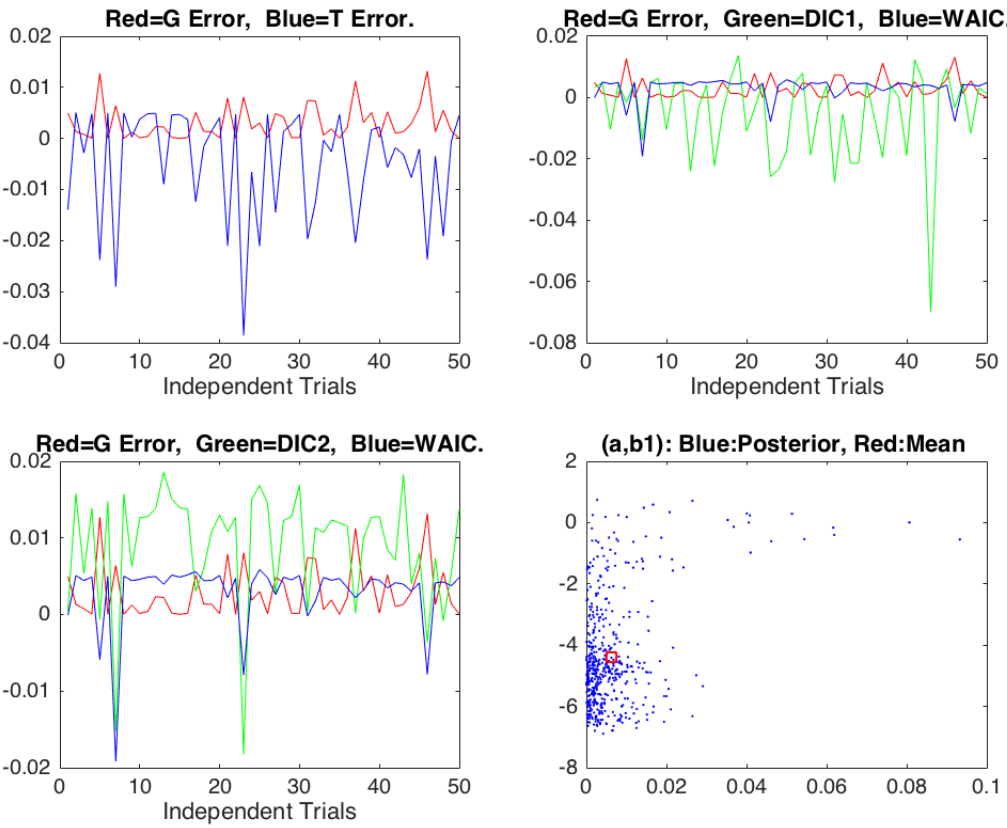

Abbildung 3: Vergleich von WAIC, DIC1 und DIC2 im singulären Fall

6 Fazit

Die Arbeit sollte dem Leser gezeigt haben, dass WAIC und WBIC sehr wichtige Ergebnisse der Singular Learning Theor sind. Denn anhand dieser beiden Informationskriterien sind wir in der Lage komplett unterschiedliche Modelle miteinander zu vergleichen. Früher, wo wir die Fisher-Regularitäten vorausgesetzt haben, brauchen wir das dank dem WAIC und WBIC nicht mehr tun. Beide haben im Bereich der Modellwahl die Informationskriterien 'revolutioniert'. Sie können für jede Art von Singularitäten eingesetzt werden und gelten auch wenn eine wahre Verteilung unrealisierbar ist durch ein statistisches Modell. Sie gelten sowohl in regulären, als auch in singulären Fällen. Beide Kriterien sind aufjedenfall vor den üblichen Informationskriterien zu bevorzugen, vorallem wenn wir komplexere Modelle untersuchen müssen.

6.1 Ausblick

Auch wenn es durch diese Arbeit so scheint, dass WBIC und WAIC die besten Informationskriterien seien gibt es im Bereich der Modellevaluation eine neue sehr beliebte Methode unter den Experten, die sich PSIS LOO-CV nennt. Da es die Darstellung dieser Methode den Rahmen dieser Arbeit sprengen würde, wäre es jedem Leser ratsam, das sich mit der Modellbewertung beschäftig, sich PSIS LOO-CV genauer anzusehen.

7 Quellenverzeichnis

[1] Ludwig Fahrmeir, Stefan L. Thomas Kneib K. Thomas Kneib: *Regression - Modelle, Methoden und Anwendungen*. Springer Verlag, 2007

[2] Watanabe, Sumio: Equations of States in SIngular Statistical Estimation. (2013), February

[3] Andrew Gelman, Aki V. Jessica Hwang H. Jessica Hwang: Understanding predictive information criteria for Bayesian models. (2013), August

[4] Kenneth P. Burnham, David R. A.: *Model Selection and Multimodel Inference*. Springer Verlag, 2002

[5] Watanabe, Sumio: A Widely Applicable Bayesian Information Criterion. (2013)

Erklärung

Hiermit erkläre ich, Burak Erdemir, dass ich meine Bachelorarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und die aus fremden Quellen direkt oder indirekt übernommenen Gedanken als solche kenntlich gemacht habe.

Die Arbeit habe ich bisher keinem anderen Prüfungsamt in gleicher oder vergleichbarer Form vorgelegt. Sie wurde bisher nicht veröffentlicht.

Datum: 29. November 2016 .....

(Unterschrift)