Universität Rostock

Traditio et Innovatio

Bernstein-von Mises Theorem for a Group Testing Problem

Dissertation
zur
Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Rostock

vorgelegt von
Jonathan Kunick, geb. am 1989-01-25 in Filderstadt
aus Rostock

Rostock, den 2025-03-25

https://doi.org/10.18453/rosdok_id00005404

Gutachter:
Prof. Dr. rer. nat. Alexander Meister, Universität Rostock, MNF
Prof. Ismaël Castillo, Sorbonne Université Paris

Jahr der Einreichung: 2024
Jahr der Verteidigung: 2024

Abstract

We propose a Bayesian model to investigate a group testing regression model of equally-sized subgroups. The group testing design was developed by Dorfman (1943) to reduce costs and increase efficiency when detecting illnesses in populations and can be used to detect contamination in samples, when prevalence is low enough. We develop a Laplace-Bernstein-von Mises (BvM) Theorem in the style of Le Cam (1986), where distributional convergence to a Gaussian holds in total variation distance almost surely, and we list sufficient conditions. We also deduce additional conditions to derive a strong BvM for the group testing posterior distribution.

Wir schlagen ein Bayessches Modell zur Untersuchung eines Blocktest-Regressionsmodells mit gleichgroßen Gruppen vor. Das Blocktestverfahren wurde von Dorfman (1943) entwickelt, um beim Nachweis von Krankheiten in der Bevölkerung eine Kostenreduzierung und Effizienzsteigerung zu erreichen, und kann zum Nachweis von Verunreinigungen in Proben genutzt werden, wenn die Prävalenz ausreichend klein ist. Wir entwickeln ein Laplace-Bernstein-von Mises (BvM) Theorem im Stile von Le Cam (1986), bei dem die Verteilungskonvergenz zu einer Normalverteilung im Totalvariationsabstand fast sicher gilt. Wir leiten zusätzliche Bedingungen her, unter denen ein starkes BvM Theorem für unsere Blocktest-a-posteriori Verteilung gilt.

Acknowledgements: I would like to thank my supervisor, Alexander Meister, who has provided me with the problem and who has offered me his professional advice. I would also like to thank Rafael Weißbach who recommended me for my current position at Uni Rostock. Moreover, I am grateful for my friends and family for supporting me, no matter what. Dedicated to the memory of Kurt Kunick.

felix, qui potuit rerum cognoscere causas

Contents

Glossary and Acronyms

a.a. *for* almost all *values with respect to some measure* μ / μ-almost-all. 3, 42, 45, 48, 61

a.e. *for every value in a measure space* $(\mathscr{S}, \mathfrak{S}, \mu)$*, possibly excluding those in* μ*-null sets /* μ*-almost-everywhere.* 2, 9, 12, 16

a.s. μ-almost-everywhere if μ is a probability measure. 11, 13, 14, 15, 16, 17, 19, 33, 36, 42, 43, 44, 45, 47, 48, 49, 50, 51, 52, 58, 61

$\mathfrak{B}(\Omega)$ Borel σ-field / algebra, σ-field generated by all open subsets of Ω in a metric setting, for a topological setting it is the σ-field generated by that particular topology. i

BvM Laplace-Bernstein-von Mises. 19, 29, 31, 32, 34, 35, 36, 37, 52, 61

CLT Central Limit Theorem. 18, 34

ciid conditionally independent and identically distributed. 17, 30, 39, 40, 41, 61

càdlàg 'continue à droite et limité à gauche'. 17

cdf cumulative distribution function. 17, 21, 39

CMT Continuous Mapping Theorem. 17, 50

DQM differentiable in quadratic mean. 30, 31

DCT Dominated Convergence Theorem. 7, 46, 49

Dirac measure $\delta_x(A)$ equals $\mathbb{1}_A(x)$. i

EM Expectation Maximisation algorithm. 26

HPD highest posterior density. 15

i.e. that is to say / in other words. 2, 3, 4, 7, 11, 13, 16, 20

iff if and only if. 1, 2, 8, 11, 13, 16, 17, 20, 23, 30, 34, 42

Identity Matrix $\mathbf{I}_d$ has entries $\delta_{j,k}$. i

iid independent and identically distributed. 16, 17, 18, 22, 23, 29, 34, 35, 41, 58

Kronecker delta $\delta_{l,m}$ is one if l and m are equal, zero otherwise. i

LAN Local asymptotic normality / locally asymptotically normal. 19, 30, 31, 36

LAMN Locally asymptotically mixed normal. 31

LAQ Locally asymptotically quadratic. 31

LPE Local polynomial regression estimator. 22, 23, 25

MCMC Markov Chain Monte Carlo. 21, 37

MLE maximum likelihood estimator. 22, 24, 41, 49, 52, 61

mb. measurable. 1, 2, 7, 11, 12, 16, 32, 36

MISE mean integrated squared error. 23, 24

MSE mean squared error. 8

MH Metropolis-Hastings algorithm. 26

NHANES National Health and Nutrition Examination Survey
https://www.cdc.gov/nchs/nhanes/index.htm. 23, 25, 26

ONS orthonormal system. 4, 58, 59

$\mathfrak{P}(\Omega)$ the power set is the collection of all subsets of a nonempty set Ω. i

SLLN strong law(s) of large numbers. 17, 47, 48

subset if $A \subset B$ and $B \subset A$ then the sets are equal. i

v.s. see above. 37

WLLN weak law of large numbers. 18

WLOG without loss of generality. 46

wrt with respect to. 2, 3, 7, 8, 9, 10, 14, 24, 30, 31, 41, 45, 61

1 Basics and Notation

In this section some notation will be introduced with the purpose of clarifying what exactly is meant with each symbol used in this dissertation. The desire is to make the formulas as rigorous as possible whilst also ensuring readability and to avoid confusion.

1.1 Measure and Probability Spaces, Metric Spaces, L^p-Spaces

Measure and probability theory are fundamental to understand Bayesian statistics. Therefore a short introduction to these topics will now be presented, as well as results used in the main section. Throughout, $\subset$ will denote a subset such that $A \subset B \ \wedge \ B \subset A \Rightarrow A = B$, a strict subset is denoted by $\subsetneq$. The introduction of measure integrals requires the following definitions.

Definition 1.1 (σ-algebra/field)
For a nonempty set $\mathscr{S}$ a family of subsets $\mathfrak{S} \subset \mathfrak{P}(\mathscr{S})$ is called a σ-algebra, if and only if (iff)

(i) $\emptyset \in \mathfrak{S}$,

(ii) $S \in \mathfrak{S} \Rightarrow S^c \in \mathfrak{S}$,

(iii) $S_n \in \mathfrak{S} \ \forall n \in \mathbb{N} \Rightarrow \cup_{n\in\mathbb{N}} S_n \in \mathfrak{S}$.

The σ means countable as clarified in condition (iii).

The pair $(\mathscr{S}, \mathfrak{S})$ is called a measurable (mb.) space, and a measure on it is a function $\mu : \mathfrak{S} \to [0, \infty]$ such that $\mu\emptyset = 0$ and for pairwise disjoint sets one has $\mu \uplus_{n\in\mathbb{N}} S_n = \sum_{n\in\mathbb{N}} \mu S_n$. In this case, we call the triplet $(\mathscr{S}, \mathfrak{S}, \mu)$ a measure space, and if $\mu\mathscr{S} = 1$ a probability space. For a family of subsets $\mathfrak{F} \subset \mathfrak{P}(\mathscr{S})$ we denote by $\sigma(\mathfrak{F})$ the smallest σ-field containing $\mathfrak{F}$.

Definition 1.2 (Borel σ-field, Borel set (Bogachev and Smolyanov, 2020))
The Borel σ-algebra $\mathfrak{B}(\mathbb{R}^d)$ of $\mathbb{R}^d$ is the σ-algebra generated by all open sets. For an arbitrary set $E \subset \mathbb{R}^d$ let $\mathfrak{B}(E) := \{E \cap B : B \in \mathfrak{B}(\mathbb{R}^d)\}$ denote the corresponding trace σ-algebra.

Definition 1.3 (Measurable Function (Bogachev and Smolyanov, 2020))

(a) *If $(\mathscr{S}, \mathfrak{S})$ is a mb. space, and $f : \mathscr{S} \to \mathbb{R}^1$, we call f $\mathfrak{S}$-measurable, iff $\forall c \in \mathbb{R}^1 : \{s : f(s) < c\} \in \mathfrak{S}$.*

(b) *Let $(\mathscr{S}, \mathfrak{S})$ and $(\mathscr{T}, \mathfrak{T})$ be mb. spaces. A mapping $f : \mathscr{S} \to \mathscr{T}$ is called $(\mathfrak{S}, \mathfrak{T})$-measurable, iff $f^{-1}(\mathfrak{T}) \subset \mathfrak{S} :\Leftrightarrow \forall T \in \mathfrak{T} : f^{-1}(T) \in \mathfrak{S}$.*

Remark 1.4. For further definitions and theorems on measurability, in particular the definition of the *Lebesgue* measure λ and complete measures, we refer to Bogachev (2007), Bogachev and Smolyanov (2020), and Kallenberg (2021).

To introduce (Lebesgue) integrals as in Bogachev and Smolyanov (2020), consider a measure space $(\mathscr{S}, \mathfrak{S}, \mu)$. A function f is called simple, if there are constants $c_j,\ j \in J \subset \mathbb{N}$ and pairwise disjoint sets $S_j \in \mathfrak{S},\ j \in \{1, \dots, k\}$ such that $f = \sum_{j=1}^k c_j \mathbb{1}_{S_j}$. Here $\mathbb{1}_{S_j}(s)$ defines the indicator function, which takes value one if $s \in S_j$ and zero otherwise, in other words it indicates if $s \in S_j$. Let $f^+ := \max(f, 0)$ be the positive and $f^- := \max(-f, 0)$ the negative part of f.

Definition 1.5 (Lebesgue Integral)
The (Lebesgue) integral *of a nonnegative simple function is the (possibly infinite) number*

$$\mu f := \int f d\mu := \sum_{j=1}^{k} c_j \mu(S_j).$$

The integral *of a μ-mb. function f which is nonnegative μ-almost everywhere (a.e.) is defined as*

$$\int f\, d\mu := \int_{\mathscr{S}} f(s)\, \mu(ds) := \sup\left\{ \int \varphi\, d\mu : \varphi \text{ is simple and nonnegative}, \varphi \leq f\ \mu - a.e. \right\}.$$

If this number is finite, f is called (Lebesgue) integrable *(with respect to (wrt) μ); more generally a signed function f is integrable if both f^+ and f^- are integrable. In that case $\int f d\mu := \int f^+ d\mu - \int f^- d\mu$. The set of μ-integrable functions is denoted as $\mathcal{L}^1 = \mathcal{L}^1(\mathscr{S}, \mathfrak{S}, \mu)$; compare Bogachev (2007). The integral over a set $S \in \mathfrak{S}$ is defined as*

$$\int_S f d\mu := \int \mathbb{1}_S f d\mu.$$

Proofs of the following theorem can be found in Bogachev (2007) and Bogachev and Smolyanov (2020).

Theorem 1.6 (Properties of the Integral)
Let $(\mathscr{S}, \mathfrak{S}, \mu)$ be a measure space, let f, g be mb. functions. The integral satisfies the following properties:

(a) *f is μ-integrable iff $|f|$ is μ-integrable, moreover $|\int f d\mu| \leq \int |f|\, d\mu$. (triangle inequality)*

(b) *If f is μ-integrable and $|g| \leq |f|$ μ-a.e., then g is μ-integrable and $\int |g|\, d\mu \leq \int |f|\, d\mu$. This also implies that if $g \leq f$ μ-a.e. then $\int g d\mu \leq \int f d\mu$. (monotonicity)*

(c) *If α, β are real numbers, and both f, g are μ-integrable: $\int (\alpha f + \beta g) d\mu = \alpha \int f d\mu + \beta \int g d\mu$. (linearity)*

(d) *For $A, B \in \mathfrak{S} : A \cap B = \emptyset : \int_{A \uplus B} f d\mu = \int_A f d\mu + \int_B f d\mu$.*

The next theorems are often of use when evaluating integrals.

Theorem 1.7 (Beppo Levi Monotone Convergence Theorem (Bogachev and Smolyanov, 2020))
Let $(f_n)_{n \in \mathbb{N}}$ be an isotone sequence of μ-integrable functions, id est (i.e.) $\forall n \in \mathbb{N} : f_n \leq f_{n+1}$ μ-a.e.. Suppose that

$$\sup_{n \in \mathbb{N}} \int f_n d\mu < \infty.$$

Then $f := \lim_{n \to \infty} f_n$ is finite μ-a.e., integrable, and

$$\int f d\mu = \lim_{n \to \infty} \int f_n d\mu.$$

Theorem 1.8 (Lebesgue Dominated Convergence Theorem (Bogachev and Smolyanov, 2020))
Suppose that μ-integrable functions f_n converge a.e. to f. If there exists a μ-integrable function g such that

$$\forall n \in \mathbb{N} : |f_n| \leq g \quad \mu - a.e.$$

then f is integrable and

$$\int f d\mu = \lim_{n \to \infty} \int f_n d\mu,$$
$$\lim_{n \to \infty} \int |f - f_n|\, d\mu = 0.$$

A measure μ on $(\mathscr{X}, \mathfrak{X})$ is called σ-finite, if there exists sequence $(\mathcal{X}_n)_{n \in \mathbb{N}} \in \mathfrak{X}$ such that $\mathcal{X}_n \uparrow \mathscr{X}$ and $\forall n \in \mathbb{N} : \mu(\mathcal{X}_n) < \infty$. The following two theorems are useful for the evaluation of integrals with respect to product measures.

Theorem 1.9 (Fubini Theorem (Bogachev, 2007))
Let μ and ν be σ-finite measures on the spaces $(\mathscr{X}, \mathfrak{X})$ and $(\mathscr{Y}, \mathfrak{Y})$. Suppose that a function on $\mathscr{X} \times \mathscr{Y}$ is integrable wrt to the product measure $\mu \otimes \nu$. Then, the function $y \mapsto f(x, y)$ is integrable wrt ν for μ-almost all (a.a.) x, and the function $x \mapsto f(x, y)$ is integrable wrt μ for ν-a.a. y, and the functions

$$x \mapsto \int_{\mathscr{Y}} f(x, y)\nu(dy), \quad y \mapsto \int_{\mathscr{X}} f(x, y)\mu(dx)$$

are integrable on the corresponding spaces, and one has

$$\int_{\mathscr{X}\times\mathscr{Y}} f d(\mu \otimes \nu) = \int_{\mathscr{Y}} \int_{\mathscr{X}} f(x, y)\mu(dx)\nu(dy) = \int_{\mathscr{X}} \int_{\mathscr{Y}} f(x, y)\nu(dy)\mu(dx).$$

Theorem 1.10 (Tonelli Theorem (Bogachev, 2007))
Let f be a nonnegative $\mu\otimes\nu$-measurable function on $\mathscr{X} \times \mathscr{Y}$, where μ and ν are σ-finite measures. Then $f \in \mathcal{L}^1(\mathscr{X} \times \mathscr{Y}, \mathfrak{X} \otimes \mathfrak{Y}, \mu \otimes \nu)$ provided that

$$\int_{\mathscr{Y}} \int_{\mathscr{X}} f(x, y)\mu(dx)\nu(dy) < \infty.$$

Let $(\mathscr{X}, \mathfrak{X})$ be a measure space, $\mu : \mathfrak{X} \to [0, \infty]$ a measure and let f be a μ-integrable function. Then a σ-additive set function (possibly signed) is obtained by

$$\nu(A) := \int_A f d\mu. \tag{1.1}$$

In this situation denote ν by $f \cdot \mu$ and f by $d\nu/d\mu$, which is called the Radon-Nikodym derivative or density of ν wrt μ (Bogachev and Smolyanov, 2020).

Definition 1.11
For two countably additive measures μ and ν on $(\mathscr{X}, \mathfrak{X})$ we say

(i) *ν is absolutely continuous wrt (or dominated by) μ if for every set $\mathcal{X} \in \mathfrak{X}$, $\mu\mathcal{X} = 0$ implies $\nu\mathcal{X} = 0$. This is denoted by $\nu \ll \mu$.*

(ii) *The measures are called mutually singular, if there exists a set $\mathcal{X} \in \mathfrak{X}$ such that $\mu\mathcal{X} = 0$ and $\nu(\mathscr{X} \setminus \mathcal{X}) = 0$. This is denoted by $\nu \perp \mu$.*

The next theorem provides a fundamental tool for the development of the results in section 4.1.

Theorem 1.12 (Radon-Nikodym Theorem (Bogachev and Smolyanov, 2020))
Take μ and ν to be finite measures on $(\mathscr{X}, \mathfrak{X})$. The measure ν is absolutely continuous with respect to the measure μ precisely when there exists a μ-integrable function f such that ν is given by (1.1).

Remark 1.13. Klenke (2020) relaxes the theorem for countable, i.e. σ-finite measures.

For the following two definitions let $\mathscr{X} \neq \emptyset$ be a linear/vector space over the field of real numbers (Bogachev and Smolyanov, 2020).

Definition 1.14 (Scalar Product)
$\mathscr{X}$ is called Euclidean, if it is equipped with a scalar/inner product, i.e. a function $\langle \cdot, \cdot \rangle : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ satisfying the axioms

(i) $\forall x \in \mathscr{X} : \langle x, x\rangle \geq 0, \qquad \langle x, x\rangle = 0 \Leftrightarrow x = 0,$

(ii) $\forall x, y \in \mathscr{X} : \langle x, y\rangle = \langle y, x\rangle,$

(iii) $\forall x, y, z \in \mathscr{X}, \alpha, \beta \in \mathbb{R} : \langle \alpha x + \beta y, z\rangle = \alpha \langle x, z\rangle + \beta \langle y, z\rangle.$

Two vectors are called orthogonal $x \perp y :\Leftrightarrow \langle x, y\rangle = 0$.

Examples are $\langle x, y\rangle = \sum_{j=1}^{n} x_j y_j$ in $\mathbb{R}^n$, or if $f, g \in L^2(\mathscr{S}, \mathfrak{S}, \mu)$, as defined later in Definition 1.39, $\langle f, g\rangle = \int f g d\mu$.

Definition 1.15 (Norm)
$\mathscr{X}$ is called normed, if it is equipped with a norm, i.e. a function $\|\cdot\| : \mathscr{X} \to [0, \infty)$ satisfying the axioms

(i) $\|x\| = 0 \Leftrightarrow x = 0,$

(ii) $\forall x \in \mathscr{X}, \alpha \in \mathbb{R} : \|\alpha x\| = |\alpha| \, \|x\|,$

(iii) $\forall x, y \in \mathscr{X} : \|x + y\| \leq \|x\| + \|y\|.$

For example, any Euclidean space can be normed using $\|x\| := \sqrt{\langle x, x\rangle}$; a norm on $\mathbb{R}^1$ is $|\cdot|$. A system of mutually orthogonal vectors of unit length in a Euclidean space $\mathscr{X}$ is called an orthonormal system (ONS).

Definition 1.16 (Metric)
A set $\mathscr{X}$ equipped with a function $d : \mathscr{X} \times \mathscr{X} \to [0, \infty)$, if

(i) $d(x, y) = 0 \Leftrightarrow x = y,$

(ii) $\forall x, y \in \mathscr{X} : d(x, y) = d(y, x),$

(iii) $\forall x, y, z \in \mathscr{X} : d(x, z) \leq d(x, y) + d(y, z).$

Metrics represent distances. Any normed spaced can be made a metric space defining $d(x, y) := \|x - y\|$. A metric essential for the results in section 4 is the total variation distance.

Definition 1.17 (Total Variation Distance (Tsybakov, 2009))
Let $(\Omega, \mathfrak{A})$ be a measurable space, and let Q, P be probability measures (p-measures) on $\mathfrak{A}$. Define

$$d_{TV}(P, Q) := \sup_{A \in \mathfrak{A}} |P(A) - Q(A)| .$$

If $Q \ll \mu$ and $P \ll \mu$ for some σ-finite measure μ, then the Radon-Nikodym Theorem 1.12 implies the existence of the densities $p := \frac{dP}{d\mu}$ and $q := \frac{dQ}{d\mu}$ and the total variation distance can be calculated as

$$d_{TV}(P, Q) = \sup_{A \in \mathfrak{A}} \left| \int_A (p - q) d\mu \right| .$$

The total variation distance fulfils all metric axioms and takes values in $[0, 1]$.

Proof. Definiteness from left to right is obvious, from right to left consider that if the supremum of the absolute value is zero, the measures must coincide on every set. Symmetry is induced by the absolute value. For the triangle inequality, exploit the definiteness and consider $\sup_{A \in \mathfrak{A}} |P_1 A - P_2 A| = \sup_{A \in \mathfrak{A}} |P_1 A - P_3 A + P_3 A - P_2 A|$. ■

Lemma 1.18 (Scheffé Lemma (Tsybakov, 2009))
In the setting of definition 1.17, the following equalities hold:

$$d_{TV}(P,Q) \overset{(a)}{=} \frac{1}{2}\int |p-q|\,d\mu \overset{(b)}{=} 1 - \int \min(dP, dQ) := 1 - \int \min(p,q)d\mu.$$

Proof. Remember that both positive and negative part of a function are nonnegative. Since p and q are μ-p-densities we follow:

$$\begin{aligned} 1 = \int p d\mu = \int q d\mu \Rightarrow 0 = \int (p-q)d\mu = \int (p-q)^+ d\mu - \int (p-q)^- d\mu \\ \Leftrightarrow \int (p-q)^+ d\mu = \int (p-q)^- d\mu. \end{aligned} \tag{1.2}$$

Using a similar argument for the absolute function values one follows exploiting (1.2):

$$\int |p-q|\,d\mu = \int (p-q)^+ d\mu + \int (p-q)^- d\mu = 2\int (p-q)^+. \tag{1.3}$$

Using (1.2) and (1.3) one follows:

$$\begin{aligned} \forall A \in \mathfrak{A} : PA - QA = \int \mathbb{1}_A p d\mu - \int \mathbb{1}_A q d\mu = \int \mathbb{1}_A (p-q) d\mu \\ = \int \mathbb{1}_A (p-q)^+ d\mu - \int \mathbb{1}_A (p-q)^- d\mu \leq \int \mathbb{1}_A (p-q)^+ d\mu \leq \frac{1}{2}\int |p-q|\,d\mu, \end{aligned}$$

since $\mathbb{1}_A \leq 1$. Analogously, $\forall A \in \mathfrak{A} : QA - PA \leq \frac{1}{2}\int |p-q|\,d\mu$. Thus

$$\forall A \in \mathfrak{A} : |PA - QA| \leq \frac{1}{2}\int |p-q|\,d\mu \Rightarrow \sup_{A\in\mathfrak{A}} |PA - QA| = \frac{1}{2}\int |p-q|\,d\mu.$$

For proof of the second part consider the set

$$A := \cup_{B\in\mathfrak{A}} \{\omega \in B : p(\omega) \geq q(\omega)\}.$$

Then we have

$$\begin{aligned} \text{on } A : |p-q| = p-q, \quad \min(p,q) = q, \\ \text{on } A^c : |p-q| = q-p, \quad \min(p,q) = p, \end{aligned}$$

thus, exploiting that P and Q are p-measures:

$$\begin{aligned} \frac{1}{2}\int |p-q|\,d\mu = \frac{1}{2}\underbrace{\int \mathbb{1}_A p d\mu}_{1-PA^c} - \frac{1}{2}\int \mathbb{1}_A q d\mu + \frac{1}{2}\underbrace{\int \mathbb{1}_{A^c} q d\mu}_{1-QA} - \frac{1}{2}\int \mathbb{1}_{A^c} p d\mu \\ = 1 - \int \mathbb{1}_{A^c} p d\mu - \int \mathbb{1}_A q d\mu \end{aligned}$$

and with this the second claim. ■

Another measure of distance often used in statistics, is the *relative information* or *Kullback Leibler divergence* as defined in Bretagnolle and Huber (1978). It is nonnegative but does not represent a metric because of its asymmetry. It is defined in the setting of definition 1.17 as

$$\mathcal{K}(P,Q) := \int \left(-\ln\left(\frac{dP}{dQ}\right)\right) dP. \tag{1.4}$$

Remark 1.19. This can be defined using the definition below as

$$\mathcal{K}(P,Q) = \mathbb{E}_P\left(-\ln\left(\frac{dP}{dQ}\right)\right).$$

Theorem 1.20 ((First) Pinsker Inequality (Tsybakov, 2009))
Let Q, P be probability measures on the same measurable space $(\Omega, \mathfrak{A})$, then

$$d_{TV}(P,Q) \leq \sqrt{\frac{1}{2}\mathcal{K}(P,Q)}.$$

Corollary 1.21
Theorem 1.20 implies that a sequence $(P_n)_{n\in\mathbb{N}}$ of probability measures satisfying $\lim_{n\to\infty}\mathcal{K}(P_n,P)=0$ converges to P in the total variation distance.

Proof. Exploiting the properties of the total variation distance and the Kullback-Leibler divergence leads to the claim. ■

Theorem 1.22 (Bretagnolle-Huber Bound (Bretagnolle and Huber, 1978; Canonne, 2022))
In the setting of Theorem 1.20 we have $d_{TV}(P,Q) \leq \sqrt{1-\exp(-\mathcal{K}(P,Q))}$.

Remark 1.23. Tsybakov (2009) only features a weaker version of this bound.

A metric space is called *complete* if every Cauchy sequence (Bogachev and Smolyanov, 2020) converges to a limit in this space. It is called *separable* if it contains a dense countable subset (Dudley, 2002).

Definition 1.24
We have seen that a scalar product can induce a norm which can induce a metric, justifying these definitions.

(a) *A* Polish space *is a complete and separable metric space.*

(b) *A complete normed space is called a* Banach space.

(c) *A complete Euclidean space is called a* Hilbert space.

Metrics can also be used to define concepts of continuity.

Definition 1.25 (Hölder Continuity)
Let $(\mathscr{X}, \|\cdot\|_{\mathscr{X}})$ and $(\mathscr{Y}, \|\cdot\|_{\mathscr{Y}})$ be normed spaces. A function $f : \mathscr{X} \to \mathscr{Y}$ is called Hölder continuous, if there exist constants $R \in \mathbb{R}$, $\alpha \in (0,1]$ such that

$$\forall s,t \in \mathscr{X} : \|f(s)-f(t)\|_{\mathscr{Y}} \leq R\,\|s-t\|_{\mathscr{X}}^{\alpha}.$$

In Banach spaces it is possible to extend the notion of the Lebesgue integral. Let $(\mathscr{B}, \|\cdot\|_{\mathscr{B}})$ be a Banach space, and let $(\mathscr{S}, \mathfrak{S}, \mu)$ be a measure space. The integral is then defined in analogy to the Lebesgue integral with one technical difference. Again, a simple function $f : \mathscr{S} \to \mathscr{B}$ is defined such that it can be represented as $f = \sum_{j=1}^{k} b_j \mathbb{1}_{S_j}$, $\forall j \in \{1,\dots,k\} : b_j \in \mathscr{B}$, $S_j \in \mathfrak{S}$.

Definition 1.26 (Bochner Integral (Hsing and Eubank, 2015))
The Bochner integral is defined in two steps via the Lebesgue integral.

(i) *Let f be a simple function as above. We call f μ-integrable, if $\forall j \in J : \mu(S_j) < \infty$. In that case*

$$\int f\,d\mu := \sum_{j=1}^{k} b_j \mu(S_j).$$

(ii) *A measurable function f is called μ-integrable, if there exists a sequence of simple functions $(f_n)_{n\in\mathbb{N}}$ such that the Lebesgue integrals of $\|f - f_n\|_{\mathscr{B}}$ converge to zero:*

$$\lim_{n\to\infty} \int \|f - f_n\|_{\mathscr{B}}\, d\mu = 0.$$

In this case we define

$$\int f d\mu := \lim_{n\to\infty} \int f_n d\mu.$$

An interesting property is that the Dominated Convergence Theorem (DCT) of $\mathscr{B}$-valued functions can be formulated wrt to a Lebesgue integrable majorant.

Remark 1.27. We will call a function from a measure space into a Banach space integrable, if the integral $\int \|f\|\, d\mu$ is finite; if $\mathscr{B} \subset \mathbb{R}^{m\times n}$ the integration can be calculated componentwise.

Now, let $(\Omega, \mathfrak{A}, \mathbb{P})$ be a probability space and let $(\mathscr{X}, \mathfrak{X})$ a mb. space and take $X : \Omega \to \mathscr{X}$ to be a $(\mathfrak{A}, \mathfrak{X})$-measurable map, i.e. $X^{-1}(\mathfrak{X}) := \cup_{\mathcal{X}\in\mathfrak{X}}\{\omega \in \Omega : X(\omega) \in \mathcal{X}\} \subset \mathfrak{A}$, then we call X a random variable, and the *law of* X is defined as the corresponding image measure $\mathfrak{L}(X) := \mathbb{P}_X := \mathbb{P} \circ X^{-1}$. Sometimes the notation $\mathbb{P}[X \in \mathcal{X}] := \mathfrak{L}(X)(\mathcal{X})$ will be used for some $\mathcal{X} \in \mathfrak{X}$, especially when only the induced probability space $(\mathscr{X}, \mathfrak{X}, \mathfrak{L}(X))$ is of interest.

Theorem 1.28 (Integration under Mappings and with Change of Variables (Bogachev, 2007))
Presented here are a general version and the special case in $\mathbb{R}^n$ where μ is the n-dimensional Lebesgue-Borel measure $\lambda^{(n)}$.

(i) *Let $(\mathscr{A}, \mathfrak{A}, \mu)$ be a measure space, $(\mathscr{B}, \mathfrak{B})$ a mb. space, and let $f : \mathscr{A} \to \mathscr{B}$ be a $(\mathfrak{A}, \mathfrak{B})$-mb. A $\mathfrak{B}$-mb. function g on $\mathscr{B}$ is integrable wrt the measure $\mu \circ f^{-1}$ precisely when the function $g \circ f$ is integrable wrt μ. Additionally, the equality*

$$\int_{\mathscr{B}} g\, d(\mu \circ f^{-1}) = \int_{\mathscr{A}} g \circ f\, d\mu$$

holds true.

(ii) *Let $\mathscr{O} \subset \mathbb{R}^n$ be an open set, and $f : \mathscr{O} \to \mathbb{R}^n$ a continuously differentiable mapping and denote its Jacobian as $\mathcal{J}_f(x) := \det f'(x) = \det Df(x)$. If f is injective on $\mathscr{O}$, then, for any measurable set $O \subset \mathscr{O}$ and any Borel function $g \in L^1(\mathbb{R}^n)$, one has the equality*

$$\int_O g \circ f(x)\, |\mathcal{J}_f(x)|\, dx = \int_{f(O)} g(y) dy.$$

Remark 1.29. When transforming integrals it may be more convenient to denote the sets with indicators:

$$\int_{\mathscr{B}} g\, d(\mu \circ f^{-1}) = \int \mathbb{1}_{\mathscr{B}} g\, d(\mu \circ f^{-1}) = \int (\mathbb{1}_{\mathscr{B}} g) \circ f\, d\mu = \int \mathbb{1}_{\mathscr{A}} g \circ f\, d\mu.$$

Remark 1.30. As is common when dealing with the Lebesgue measure, abbreviate $\lambda^{(n)}(d\mathbf{x}) =: d\mathbf{x}$.

Remark 1.31. Bogachev (2007) states the formula for nonnegative measures, which should be considered when working with signed measures.

Using a change of measure comes naturally when calculating expectations. For greater clarification the corresponding image measure will *often* be denoted at the bottom of the operator.

Definition 1.32 (Expectation, Expected Value)
As in the above setting, let $(\Omega, \mathfrak{A}, \mathbb{P})$ be a probability space and $X \sim \mathfrak{L}(X)$ a random variable on this space such that $(\mathscr{X}, \mathfrak{X}, \mathfrak{L}(X))$ is the induced probability space. We say the expectation of X exists, if $\mathbb{E}|X| < \infty$. In this case the number

$$\mathbb{E}X := \mathbb{E}_{\mathfrak{L}(X)}X := \mathbb{P}X := \int_\Omega X d\mathbb{P} = \int_{\mathscr{X}} id_{\mathscr{X}} d\mathfrak{L}(X)$$

the expectation or the expected value of X. Note that the last equality can be seen by setting g in Theorem 1.28 (i) *to be the identity operator. If μ is a σ-finite measure, $\mathfrak{L}(X) \ll \mu$, and $f := \frac{d\mathfrak{L}(X)}{d\mu}$, the Radon-Nikodym Theorem 1.12 implies $\mathbb{E}X = \int_{\mathscr{X}} id_{\mathscr{X}} f d\mu$. Furthermore, if g is $\mathfrak{X}$-measurable, the expectation of $g(X)$ can be calculated as*

$$\mathbb{E}_{\mathfrak{L}(X)}g(X) = \int_\Omega g \circ X d\mathbb{P} = \int_{\mathscr{X}} g(x)\mathfrak{L}(X)(dx),$$

if it exists. We denote $\mathbb{E}(X; A) := \mathbb{E}\mathbb{1}_A X$.

The expectation of an indicator is the corresponding probability, since for $A \in \mathfrak{A}$: $\mathbb{E}\mathbb{1}_A = \mathbb{E}_{\mathfrak{L}(\mathbb{1}_A)}\mathbb{1}_A = 0 \cdot \mathbb{P}A^c + 1 \cdot \mathbb{P}A$. Apart from the expectation as a 'measure' of location, the variance is introduced as a 'measure' of dispersion.

Definition 1.33 (Variance, Covariance)
If $\mathbf{X}$ is a $\mathbb{R}^n$-valued random variable satisfying $\mathbb{E}\langle \mathbf{X}, \mathbf{X}\rangle < \infty$ the covariance (matrix) is defined as

$$\mathbb{V}\mathbf{X} := \mathrm{Cov}(\mathbf{X}, \mathbf{X}) := \mathbb{E}(\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{X} - \mathbb{E}\mathbf{X})^\top,$$

where the expectance of a matrix is taken element-wise by convention following Definition 1.26. If $\mathbf{X}$ and $\mathbf{Y}$ are of the same dimension, we write

$$\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) := \mathbb{E}(\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{Y} - \mathbb{E}\mathbf{Y})^\top$$

Similarly to the expectation, the image measure will be denoted when further clarification is deemed necessary or helpful, e.g. $\mathbb{V}_{\mathfrak{L}(X)}X$. In the case $n = 1$ $\mathbb{V}X = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}\left(X^2 + (\mathbb{E}X)^2 - 2(\mathbb{E}X)X\right)$ which, exploiting measurability, can be computed as $\mathbb{E}X^2 - (\mathbb{E}X)^2$.

Closely related to this is the *bias-variance decomposition*; in statistics, a random variable $\hat{\theta} := T \circ X$ will be used to estimate a parameter θ and to evaluate the quality of the estimator one could use the bias $\mathbb{B}\hat{\theta} := \mathbb{E}_{\mathbb{P}^\theta_X}\hat{\theta} - \theta$ (where the expectation is taken wrt the measure $\mathbb{P}^\theta_X$, formally introduced in and below Theorem 1.44) and the mean squared error (MSE) (wrt $\mathbb{P}^\theta_X$)

$$\mathrm{MSE}\left(\hat{\theta}, \theta\right) := \mathbb{E}_{\mathbb{P}^\theta_X}\left|\hat{\theta} - \theta\right|^2 = \mathbb{E}_{\mathbb{P}^\theta_X}\hat{\theta}^2 - 2\theta\mathbb{E}_{\mathbb{P}^\theta_X}\hat{\theta} + \theta^2 \pm \left(\mathbb{E}_{\mathbb{P}^\theta_X}\hat{\theta}\right)^2 = \mathbb{V}_{\mathbb{P}^\theta_X}\hat{\theta} + (\mathbb{B}\hat{\theta})^2.$$

A distribution required for section 4 is the multivariate Gaussian or normal distribution.

Lemma 1.34 (Finite Dimensional Gaussian Distribution (Bogachev, 2015))
Let $\mathbb{R}^d \ni t \mapsto \mathbb{E}\exp(i\langle X, t\rangle)$ denote the Fourier transform. A measure γ on $\mathbb{R}^d$ is called Gaussian if its Fourier transform has the form $\exp\left(i\langle t, \mu\rangle - \frac{1}{2}\langle Ct, t\rangle\right)$ where $\mu \in \mathbb{R}^d$ and $C \in \mathbb{R}^{d\times d}$ with nonnegative entries. The measure has a Lebesgue-Borel density iff C is regular and symmetric. In this case the density is

$$x \mapsto \frac{1}{\sqrt{(2\pi)^d \det C}} \exp\left(-\frac{1}{2}\left\langle C^{-1}(x - \mu), x - \mu\right\rangle\right).$$

In this case we write $\mathbf{X} \sim \gamma =: \mathcal{N}_d(\mu, C)$ and we have $\mathbb{E}\mathbf{X} = \mu$ and $\mathbb{V}\mathbf{X} = C$.

An infinite dimensional example of a Gaussian distribution is the *Wiener process*.

Definition 1.35 (Wiener Measure, Wiener Process (Bogachev, 2015; Henze, 2024))
Let $\mathscr{C} := \mathscr{C}[0,1]$ denote the space of continuous real-valued functions on $[0,1]$. The image measure $\mathbb{W}$ on $(\mathscr{C}, \mathfrak{B}(\mathscr{C}))$ of a stochastic process $W : \mathscr{C} \to \mathscr{C}, x \mapsto x$ is called a Wiener measure if

(i) $\mathbb{W}(W(0) = 0) = 1$.

(ii) *If $W(t)$ denotes the projection $\pi_t \circ W$ we have $\forall t \in (0,1] : W(t) \sim \mathcal{N}(0,t)$ given $\mathbb{W}$.*

(iii) *For $k \geq 0$ and each choice of $t_0, \dots, t_k \in [0,1] : 0 \leq t_0 \leq \cdots \leq t_k \leq 1$ the increments $W(t_1) - W(t_0), \dots, W(t_k) - W(t_{k-1})$ are independent (Definition 1.41) wrt $\mathbb{W}$.*

The Fisher information formalises how much the likelihood changes given some data wrt small changes in the parameter value; intuitively this hints at how well this parameter may be estimated.

Definition 1.36 (Fisher Information)
For each θ in the parameter space Θ let $(\mathscr{X}, \mathfrak{X}, \mathbb{P}^\theta)$ be a probability space, with the image measure $\mathbb{P}^\theta \ll \mu$ for some random variable wrt some σ-finite measure μ. Moreover let the Radon-Nikodym derivative $f(\cdot|\theta) := d\mathbb{P}^\theta / d\mu$ be the corresponding (probability-)density. Then the Fisher information of θ is defined as $\mathcal{I}(\theta) := \mathbb{V}_{\mathbb{P}^\theta} \nabla_\theta \ln f(\cdot|\theta)$.

A useful, although sometimes imprecise, tool to bound probabilities is the Chebyshev inequality (compare Bogachev (2007)).

Theorem 1.37 (Chebyshev inequality)
We present a general version and a special case more common in statistics.

(a) *Let $(\mathscr{S}, \mathfrak{S}, \mu)$ be a measure space and take $f : \mathscr{S} \to \mathscr{T}$ to be a μ-integrable function. Then*

$$\forall R > 0 : \mu(|f| \geq R) := \mu(\{s \in \mathscr{S} : |f(s)| \geq R\} \leq \frac{1}{R} \int |f| \, d\mu.$$

(b) *Let $(\Omega, \mathfrak{A}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathscr{X}$ a random variable satisfying $\mathbb{E}X^2 < \infty$. Then*

$$\forall \varepsilon > 0 : \mathbb{P}(|X - \mathbb{E}X| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} \mathbb{V}X.$$

Proof. (a) Since the inequality $R \cdot \mathbb{1}_{\{s \in \mathscr{S} : |f(s)| \geq R\}} \leq |f(s)|$ holds pointwise on $\mathscr{S}$ the proposition follows from the monotonicity and the linearity of measure integrals.

(b) Setting $\mu = \mathbb{P}, f = (X - \mathbb{E}X)^2, R = \varepsilon^2$ leads to the claim. ■

Corollary 1.38
In the setting of Theorem 1.37: $\int |f| \, d\mu = 0 \Rightarrow f = 0 \ \mu-a.e.$.

Proof. $\forall \varepsilon > 0 : 0 \leq \mu(|f| \geq \varepsilon) \leq \frac{0}{\varepsilon}$. ■

Definition 1.39 (L^p Spaces)
For any measure space $(\mathscr{S}, \mathfrak{S}, \mu)$ and constant $p \in (0, \infty)$, let $L^p := L^p(\mathscr{S}, \mathfrak{S}, \mu)$ be the equivalence class of measurable functions $f : \mathscr{S} \to \mathbb{R}$ satisfying

$$\|f\|_p := (\mu |f|^p)^{1/p} := \left(\int |f|^p \, d\mu \right)^{1/p} < \infty.$$

We also abbreviate this by $L^p(\mathscr{S})$ or $L^p(\mu)$ when the rest is clear from the context.

At this point, it should be noted that $\|\cdot\|_p$ only represents a norm on L^p because on $\mathcal{L}^p$ the definiteness does not hold as suggested by Corollary 1.38 (compare Hsing and Eubank (2015)). Another norm used in later sections is the supremum or uniform norm as defined in Bogachev and Smolyanov (2020) where $\mathscr{B}(\Omega)$ is the set of bounded real functions on $\Omega \neq \emptyset$:

$$\|f\|_\infty : \mathscr{B}(\Omega) \to [0, \infty], \quad f \mapsto \sup\{|f(t)| : t \in \Omega\}.$$

The space $L^p(\mathscr{S}, \mathfrak{S}, \mu)$ is a Banach space, and for $p = 2$ it is a Hilbert space, if one sets $\langle f, g \rangle := \int f g d\mu$ (Chung and Williams, 1990).

Lemma 1.40 (Jensen-Hölder Inequality (Kallenberg, 2021))
For any integrable vector X with values in $\mathbb{R}^d$ and convex function $f : \mathbb{R}^d \to \mathbb{R}$, we have $\mathbb{E}f(X) \geq f(\mathbb{E}X)$.

Another important concept, especially in the Bayesian context, is that of stochastic independence and will now be formalised as in (Klenke, 2020) or (Kallenberg, 2021):

Definition 1.41 (Stochastic Independence)
Let I be an index set, and let $(\Omega, \mathfrak{A}, \mathbb{P})$ be a probability space, and for each $i \in I$ assume $X_i : (\Omega, \mathfrak{A}) \to (\mathscr{X}_i, \mathfrak{X}_i)$, and $h_i : (\mathscr{X}_i, \mathfrak{X}_i) \to (\mathscr{H}_i, \mathfrak{H}_i)$ to be measurable mappings.

(a) *Events $(A_i)_{i \in I}$ are called independent (wrt $\mathbb{P}$)*

$$:\Leftrightarrow \forall (i_1, \ldots, i_n) \in I^n : i_1 < \cdots < i_n : \mathbb{P}\left[\cap_{k \in \{1,\ldots,n\}} A_{i_k} \right] = \prod_{k \in \{1,\ldots,n\}} \mathbb{P} A_{i_k}.$$

(b) *Families of events, e.g. σ-algebras, $(\mathfrak{F}_i)_{i \in I}$ are called independent if for any choice of $A_i \in \mathfrak{F}_i$ the $(A_i)_{i \in I}$ are independent.*

(c) *Random variables $(X_i)_{i \in I}$ are called independent if the generated σ-fields $(\sigma(X_i))_{i \in I}$ are independent.*

Pairwise independence, for instance between two random variables X, Y, will be denoted as $X \perp\!\!\!\perp Y$.

Lemma 1.42
In the setting of Definition 1.41 $(h(X_i))_{i \in I}$ are independent random variables.

Proof.

$$\begin{aligned} \sigma(h_i \circ X_i) &= \cup_{\mathcal{H} \in \mathfrak{H}} \{\omega \in \Omega : h_i \circ X_i(\omega) \in \mathcal{H}\} = \cup_{\mathcal{H} \in \mathfrak{H}} (h_i \circ X_i)^{-1}(\mathcal{H}) \\ &= \cup_{\mathcal{H} \in \mathfrak{H}} X_i^{-1}(h_i^{-1}(\mathcal{H})) \subset \cup_{\mathcal{X} \in \mathfrak{X}} X_i^{-1}(\mathcal{X}) = \sigma(X_i) \end{aligned}$$

■

1.2 Conditioning, Bayesian Statistics

In Bayesian statistics the task is to update some prior information or assumption on a random variable, in particular the parameter of interest, given some data sample. The following lemma motivates that this can be done iteratively, i.e. the probability of an event A of interest can be calculated in one go or sequentially as data, e.g. $B = \{X_1 = x_1\}$, $C = \{X_2 = x_2\}$, is available one at a time.

Lemma 1.43 (Iterative Conditioning for Events)
In the most basic setting let $(\Omega, \mathfrak{A}, \mathbb{P})$ be a probability space and let $A, B, C \in \mathfrak{A}$ be events satisfying $\mathbb{P}B > 0$ and $\mathbb{P}[B \cap C] > 0$. The conditional probability of A given B is defined as:

$$\mathbb{P}^B A := \mathbb{P}[A|B] := \frac{\mathbb{P}[A \cap B]}{\mathbb{P}B}.$$

Then it follows that $\mathbb{P}^B$ is a probability measure on $(\Omega, \mathfrak{A})$ and $\mathbb{P}[A|B \cap C] = \mathbb{P}^B[A|C]$.

Proof. $\mathbb{P}^B$ inherits the countable additivity from $\mathbb{P}$ given pairwise disjoint sets $A_n \in \mathfrak{A} : \uplus_{n=1}^{\infty} A_n \in \mathfrak{A}$, exploiting the distributivity of intersections and unions, and the fact that the $(A_n \cap B)$ remain pairwise disjoint:

$$\mathbb{P}^B[\uplus_{n=1}^{\infty} A_n] = \frac{\mathbb{P}[\uplus_{n=1}^{\infty}(A_n \cap B)]}{\mathbb{P}B} = \frac{\sum_{n=1}^{\infty} \mathbb{P}(A_n \cap B)}{\mathbb{P}B} = \sum_{n=1}^{\infty} \frac{\mathbb{P}(A_n \cap B)}{\mathbb{P}B} = \sum_{n=1}^{\infty} \mathbb{P}^B A_n.$$

Furthermore $\mathbb{P}^B\emptyset = 0$ and $\mathbb{P}^B\Omega = \frac{\mathbb{P}[\Omega \cap B]}{\mathbb{P}B} = 1$. By definition it follows that

$$\mathbb{P}^B[A|C] = \frac{\mathbb{P}^B[A \cap C]}{\mathbb{P}^B C} = \left(\frac{\mathbb{P}[A \cap B \cap C]}{\mathbb{P}B}\right) / \left(\frac{\mathbb{P}[B \cap C]}{\mathbb{P}B}\right) = \frac{\mathbb{P}[A \cap (B \cap C)]}{\mathbb{P}[B \cap C]}.$$

■

The following theorem defines what a conditional expectation is.

Theorem 1.44 (Conditional Expectation, Kolmogorov (Kallenberg, 2021))
Let $(\Omega, \mathfrak{A}, \mathbb{P})$ be a probability space. For any σ-field $\mathfrak{B} \subset \mathfrak{A}$, there exists an almost surely (a.s.) unique linear operator $\mathbb{E}^{\mathfrak{B}} : L^1(\Omega, \mathfrak{A}, \mathbb{P}) \to L^1(\Omega, \mathfrak{B}, \mathbb{P})$, such that

(i) $\forall B \in \mathfrak{B} : \mathbb{E}\mathbb{1}_B \mathbb{E}^{\mathfrak{B}} X = \mathbb{E}\mathbb{1}_B X, \quad X \in L^1$.

These operators have further properties *whenever the corresponding expressions exist for the absolute values:*

(ii) $X \geq 0 \Rightarrow \mathbb{E}^{\mathfrak{B}} X \geq 0$ *a.s. (positivity),*

(iii) $\mathbb{E}\left|\mathbb{E}^{\mathfrak{B}} X\right| \leq \mathbb{E}|X|$ *(contractivity),*

(iv) $0 \leq X_n \uparrow X \Rightarrow \mathbb{E}^{\mathfrak{B}} X_n \uparrow \mathbb{E}^{\mathfrak{B}} X$, *a.s. (monotone convergence),*

(v) *if X is $\mathfrak{B}$-mb., then $\mathbb{E}^{\mathfrak{B}} XY = X\mathbb{E}^{\mathfrak{B}} Y$ a.s. (pull out),*

(vi) $\mathbb{E}\left(X\mathbb{E}^{\mathfrak{B}} Y\right) = \mathbb{E}\left(Y\mathbb{E}^{\mathfrak{B}} X\right) = \mathbb{E}\left(\mathbb{E}^{\mathfrak{B}} Y\right)\left(\mathbb{E}^{\mathfrak{B}} X\right)$ *(self-adjointness),*

(vii) $\mathfrak{C} \subset \mathfrak{B} \Rightarrow \mathbb{E}^{\mathfrak{C}}\mathbb{E}^{\mathfrak{B}} X = \mathbb{E}^{\mathfrak{C}} X$ *a.s. (tower).*

In particular $\mathbb{E}^{\mathfrak{B}} X = X$ a.s. iff X is $\mathfrak{B}$-measurable, and $X \perp\!\!\!\perp \mathfrak{B} \Rightarrow \mathbb{E}^{\mathfrak{B}} X = \mathbb{E}X$.

Let us introduce the following notation:

- $\mathbb{E}(X|\mathfrak{B}) := \mathbb{E}^{\mathfrak{B}}X$,
- $\mathbb{E}^{Y}X := \mathbb{E}(X|Y) := \mathbb{E}(X|\sigma(Y))$,
- $\mathbb{P}^{\mathfrak{B}}A := \mathbb{P}(A|\mathfrak{B}) := \mathbb{E}^{\mathfrak{B}}\mathbb{1}_A$,
- $\mathbb{P}_X^Y\mathcal{X} := \mathbb{P} \circ X^{-1}[\mathcal{X}|Y],\ \mathcal{X} \in \mathfrak{X}$.

Example 1.45. The conditional expectation defined by property (i) smoothes the random variable X over $\mathfrak{B}$. This phenomenon is exploited by the Rao-Blackwell Theorem. It can be illustrated as follows: Assume $\mathfrak{B} = \{\emptyset, B, B^c, \Omega\}, B \in \mathfrak{A}$. Then take all sets from this σ-field. For Ω we have $\mathbb{E}\mathbb{E}^{\mathfrak{B}}X = \mathbb{E}X$. On the other hand we have

$$\mathbb{E}X = \mathbb{E}\mathbb{1}_B X + \mathbb{E}\mathbb{1}_{B^c}X = \mathbb{E}\mathbb{1}_B\mathbb{E}^{\mathfrak{B}}X + \mathbb{E}\mathbb{1}_{B^c}\mathbb{E}^{\mathfrak{B}}X = \mathbb{E}(\mathbb{1}_B\mathbb{E}^{\mathfrak{B}}X + \mathbb{1}_{B^c}\mathbb{E}^{\mathfrak{B}}X),$$

and in summary we can see that

$$\mathbb{E}^{\mathfrak{B}}X = c_1\mathbb{1}_B + c_2\mathbb{1}_{B^c} \quad \mathbb{P}-a.s.,$$

where $c_1 = \frac{\mathbb{E}\mathbb{1}_B X}{\mathbb{P}B}$, $c_2 = \frac{\mathbb{E}\mathbb{1}_{B^c} X}{\mathbb{P}B^c}$. And if $X = \mathbb{1}_A$ we get $\mathbb{E}^{\mathfrak{B}}X = \mathbb{P}^{\mathfrak{B}}A = (\mathbb{P}^B A)\mathbb{1}_B + (\mathbb{P}^{B^c}A)\mathbb{1}_{B^c}$, which is consistent with lemma 1.43.

Remark 1.46. Lemma 1.43 can only be applied after the data is collected, e.g. $B = \{\omega \in \Omega : X_1(\omega) = x_1\}$, as the Doob-Dynkin factorisation lemma (Taraldsen, 2018; Henze, 2024) ($\mathbb{E}^Y X$ has a version such that it can be written as $\varphi \circ Y$ where φ is mb.) implies that if $X_1 \perp\!\!\!\perp X_2$:

$$\mathbb{E}(\mathbb{E}(\theta|X_1)|X_2) = \mathbb{E}\varphi(X_1) \equiv const. \neq \mathbb{E}(\theta|X_1, X_2) = \psi(X_1, X_2),$$

in general.

Definition 1.47 (Kernel, Regular Conditional Distribution (Kallenberg, 2021))
Let $(\mathscr{X}, \mathfrak{X})$ and $(\mathscr{Y}, \mathfrak{Y})$ be two measurable spaces. A (transition) kernel *is a function $\kappa : \mathscr{X} \times \mathfrak{Y} \to [0, \infty]$ such that*

(a) $\forall \mathcal{Y} \in \mathfrak{Y} : \kappa(\cdot, \mathcal{Y})$ *is $\mathfrak{X}$-mb., and*

(b) $\forall x \in \mathscr{X} : \kappa(x, \cdot)$ *is a measure on $\mathfrak{Y}$.*

If $\forall x \in \mathscr{X} : \kappa(x, \mathscr{Y}) = 1$, it is called a probability kernel. Let $(\Omega, \mathfrak{A}, \mathbb{P})$ be a probability space. A (regular) conditional distribution *is a probability kernel $\kappa = \mathfrak{L}(X|\mathfrak{B}) : \Omega \to \mathfrak{X}$, such that*

$$\forall \omega \in \Omega : \forall \mathcal{X} \in \mathfrak{X} : \kappa(\omega, \mathcal{X}) = \mathbb{P}(X \in \mathcal{X}|\mathfrak{B})(\omega),\ a.s.$$

More specifically, if X is a random variable in $(\mathscr{X}, \mathfrak{X})$ and Y a random variable in $(\mathscr{Y}, \mathfrak{Y})$:

$$\forall \mathcal{Y} \in \mathscr{Y} : \kappa(X, \mathcal{Y}) = \mathbb{P}[Y \in \mathcal{Y}|X]\ a.s.$$

Theorem 1.48 (Conditional Distributions, Disintegration (Kallenberg, 2021))
Let X, Y be random elements in $\mathscr{X}, \mathscr{Y}$, where $\mathscr{Y}$ is Borel. Then $\mathfrak{L}(X, Y) = \mathfrak{L}(X) \otimes \mu$ for a probability kernel $\mu : \mathscr{X} \to \mathscr{Y}$ that is unique $\mathfrak{L}(X)$-a.e. and satisfies

(i) $\mathfrak{L}(X|Y) = \mu(X, \cdot),\ a.s.,$

(ii) $\mathbb{E}\left(f(X, Y)|X\right) = \int f(X, y)\ \mu(X, dy) \quad a.s., f \geq 0.$

Kallenberg (2021) also states that (ii) can be extended to suitable real-valued functions f.

The notion of independence in Definition 1.41 can naturally be extended to the conditional setting introduced in Theorem 1.44.

Definition 1.49 (Conditional Independence (Schervish, 1995; Kallenberg, 2021))
σ-fields $(\mathfrak{A}_i)_{i\in I}$ are called conditionally independent given a σ-field $\mathfrak{B}$ iff

$$\forall(i_1,\dots,i_n)\in I^n : i_1<\dots<i_n : \mathbb{P}^{\mathfrak{B}}\left[\cap_{k\in\{1,\dots,n\}}A_k\right]=\prod_{k\in\{1,\dots,n\}}\mathbb{P}^{\mathfrak{B}}A_k,\quad A_k\in\mathfrak{A}_{i_k}.$$

Let this notion be extended to events and random variables. In that sense, in analogy to the unconditional case, we write $X \perp\!\!\!\perp_{\mathfrak{B}} Y$ if X and Y are conditionally independent given $\mathfrak{B}$.

Theorem 1.50 (Conditional Independence, Doob (Kallenberg, 2021))
Let $\mathfrak{F},\mathfrak{G},\mathfrak{H}$ be σ-fields and let $\mathfrak{F}\vee\mathfrak{G}:=\sigma(\mathfrak{F},\mathfrak{G})$. Then

$$\mathfrak{F}\underset{\mathfrak{G}}{\perp\!\!\!\perp}\mathfrak{H}\Leftrightarrow\mathbb{P}^{\mathfrak{F}\vee\mathfrak{G}}=\mathbb{P}^{\mathfrak{G}}\ a.s.\ on\ \mathfrak{H}.$$

Proof. We will give the proof from left to right. For the other direction compare Kallenberg (2021) or apply a monotone class argument or use algebraic induction.
"$\Leftarrow$": Take arbitrary but fixed sets $F\in\mathfrak{F},\ H\in\mathfrak{H}$, then (using the defining property, tower property, measurability and pull-out property, as well as the assumption)

$$\begin{aligned}\mathbb{P}^{\mathfrak{G}}(F\cap H)&\overset{\text{DEF}}{=}\mathbb{E}^{\mathfrak{G}}\mathbb{1}_{F\cap H}\\&\overset{\text{TOW}}{=}\mathbb{E}^{\mathfrak{G}}\mathbb{E}^{\mathfrak{F}\vee\mathfrak{G}}\mathbb{1}_{F\cap H}\\&\overset{\text{P-O}}{=}\mathbb{E}^{\mathfrak{G}}\mathbb{1}_F\mathbb{E}^{\mathfrak{F}\vee\mathfrak{G}}\mathbb{1}_H\\&\overset{\text{ASS}}{=}\mathbb{E}^{\mathfrak{G}}\mathbb{1}_F\mathbb{E}^{\mathfrak{G}}\mathbb{1}_H\\&\overset{\text{DEF}}{=}(\mathbb{P}^{\mathfrak{G}}F)(\mathbb{P}^{\mathfrak{G}}H).\end{aligned}$$

■

Bayesian Statistics Castillo (2014) writes "*Why Bayesian estimators? Often, priors have a natural probabilistic interpretation and insights from the construction of various stochastic processes in probability theory can be helpful. Additional smoothing parameters may themselves get a prior, thus leading to natural constructions of priors via hierarchies. ... There are other attractive aspects of the Bayesian approach, for instance the fact that there are natural priors corresponding to exchangeable data, as developed among others by the Italian school after de Finetti.*" "*Bayesian methods are a prominent tool in statistics, machine learning, and practical applications*" such as medical imaging, astro-statistics, genomics, etc. (Castillo, 2024).

The introduction of conditioning is crucial for a theoretical understanding of Bayesian statistics. The field is named after Thomas Bayes (1701–1761) whose notes were published posthumously by Richard Price. A simplified version of their statement can be formulated for events as in the setting of lemma 1.43 and $\mathbb{P}A>0$ as

$$\mathbb{P}^B A=\frac{\mathbb{P}^A B\mathbb{P}A}{\mathbb{P}B},$$

i.e. in essence an inversion of conditions.

Proof.

$$\mathbb{P}^B A\overset{\text{DEF}}{=}\frac{\mathbb{P}(A\cap B)}{\mathbb{P}B}=\frac{\frac{\mathbb{P}(A\cap B)}{\mathbb{P}A}\mathbb{P}A}{\mathbb{P}B}$$

■

Notable early contributions to the subject were also made by Pierre-Simon Laplace (1749–1827) as we will see in section 2.2. In Bayesian statistics the parameter space $(\Theta, \mathfrak{T})$ is equipped with a prior (probability) measure Π and the distribution of the data X given the parameter $\theta \in \Theta$ is defined by the kernel $\mathbb{P}_X^\theta$ (note that the X will often be suppressed in notation) such that $\forall\theta \in \Theta : \forall\mathcal{X} \in \mathfrak{X} : \mathbb{P}_X^\theta(\mathcal{X}) = \kappa(\theta, \mathcal{X})$, when $(\mathscr{X}, \mathfrak{X}, \mathbb{P}_X^\theta)$ is the probability space induced by $X|\theta$. In other words (X, θ) are defined as random variables on the product space $(\mathscr{X} \times \Theta, \mathfrak{X} \otimes \mathfrak{T})$ such that

$$\mathbb{P}(X \in \mathcal{X}, \theta \in \mathcal{T}) = \int \mathbb{1}_{\mathcal{T}}(\theta)\kappa(\theta, \mathcal{X})\Pi(d\theta).$$

A common notation for this is $\theta \sim \Pi,\ X|\theta \sim \mathbb{P}_X^\theta$. As suggested by Ghosal and van der Vaart (2017) we will from now on assume Θ to be the Borel subset of a Polish space. For a dominated collection of measures, a version of the posterior is then given by Bayes formula as found in Ghosal and van der Vaart (2017) and Castillo (2024) (also compare Durrett (2019) on regular conditional probabilities):

$$\forall\mathcal{T} \in \mathfrak{T} : \Pi^X(\mathcal{T}) = \Pi(\mathcal{T}|X) = \frac{\int_{\mathcal{T}} p(X|\theta)\Pi(d\theta)}{\int p(X|\theta)\Pi(d\theta)}, \quad p(\cdot|\theta) := \frac{d\mathbb{P}_X^\theta}{d\mu}. \tag{1.5}$$

For a dominated prior and $\pi := \frac{d\Pi}{d\nu}$ one has the handy proportionality rule for the posterior density (Schervish, 1995; Shao, 2003; Robert, 2007)

$$\pi(t|x) := \frac{\Pi(\cdot|X = x)}{d\nu}(t) = \frac{p(x|t)\pi(t)}{\int p(x|\theta)\pi(\theta)\nu(d\theta)} \propto p(x|t)\pi(t), \tag{1.6}$$

where the equation defines a version of the posterior's Radon-Nikodym derivative.

Theorem 1.51 (Bayes Theorem (Schervish, 1995))
Suppose that X was drawn from the experiment $(\mathscr{X}, \mathfrak{X}, \{\mathbb{P}_X^\theta : \theta \in \Theta\})$ with $\mathbb{P}_X^\theta \ll \mu$ for all $\theta \in \Theta$ such that X has the conditional density $p(\cdot|\theta) := \frac{d\mathbb{P}_X^\theta}{d\mu}$. Then $\Pi(\cdot|X) \ll \Pi$ a.s. wrt the marginal of X and the Radon-Nikodym derivative is

$$\frac{d\Pi(\cdot|X)}{d\Pi}(\theta|x) = \frac{p(x|\theta)}{\int_\Theta p(x|t)\Pi(dt)}$$

for those x such that the denominator is neither zero nor infinite. The prior predictive probability of the set of x values such that the denominator is zero or infinite is zero, hence the posterior can be defined arbitrarily for such x.

Schervish (1995) writes "*A major use of statistical inference is its application to decision making under uncertainty. When the costs and/or benefits of our actions depend on quantities we will not know until after we make our decisions, we need to be able to weigh the costs against the uncertainties intelligently.*" And Le Cam and Lo Yang (2000) begin their book with an introduction to statistical decision theory in the sense of Abraham Wald: A statistician may observe sample data x from a process governed by the parameter which is not observed, but a decision of the set D is to be made. Thus, the need of a cost function $L : \Theta \times D \to (-\infty, +\infty]$ and a randomised decision $\delta(X)$ are summarised by means of a risk function. Decisions are then made by imposing a paradigm, among which Bol (2002) lists Wald's minimax rule (pessimistic), Hurwicz's, minimin rule (optimistic), and the Bayes rule of averaging with respect to a prior. In Bayesian statistics one may be content with the posterior distribution quantifying all stochastic information given the data. If the goal is to derive a point estimate one tries to minimise the Bayes risk. Often the decision is to estimate a parameter such that we write $\delta(X) = \hat{\theta}$.

Definition 1.52 (Bayes Risk (Schervish, 1995))
The Bayesian risk associated with the model introduced above and the loss function L is

$$\int\int L(\delta \circ X(\omega), t)\kappa(\theta, d\omega)\Pi(d\theta) = \mathbb{E}_\Pi \mathbb{E}^\theta_{\mathbb{P}_X} L(\delta \circ X, \theta).$$

Typical loss functions may be derived from metrics such as

$$L_2(\delta, \theta) := (\delta - \theta)^2, \tag{1.7}$$
$$L_c(\delta, \theta) := c(\delta - \theta)\mathbb{1}_{\{\delta \geq \theta\}} + (1 - c)(\theta - \delta)\mathbb{1}_{\{\delta < \theta\}}, \quad c \in (0, 1). \tag{1.8}$$

Lemma 1.53 (Bayes Decisions (Schervish, 1995))
The Bayes decision on a point estimate for the loss in (1.7) is the posterior mean. For the loss in (1.8) the optimal decision is the $(1 - c)$-posterior quantile.

Proof. To prove the first part of lemma 1.53 we restrict ourselves to a dominated experiment. The risk function then becomes

$$\begin{aligned}\int\int (\delta(x) - \theta)^2 p(x|\theta)\mu(dx)\pi(\theta)\nu(d\theta) &= \int\int (\delta(x) - \theta)^2 p(x|\theta)\pi(\theta)\nu(d\theta)\mu(dx) \\ &= \int\int (\delta(x) - \theta)^2 \pi(\theta|x)\nu(d\theta)g(x)\mu(dx) \\ &= \int \mathbb{E}_\Pi \left((\delta(x) - \theta)^2 | X = x\right) g(x)\mu(dx),\end{aligned}$$

where the second equation holds thanks to Tonelli, and for the third we have set $g(x) := \int p(x|\theta)\pi(\theta)\nu(d\theta)$ as in (1.6) (also compare disintegration of the joint measure in Le Cam and Lo Yang (2000), Chapter 8.3). This means that we want to choose $\delta(x)$ such that it minimises the posterior variance for the observation $X = x$, which is achieved by the posterior mean $\delta(X) = \mathbb{E}^X_\Pi \theta$ as long as $\mathbb{E}^X_\Pi \theta^2 < \infty$ according to the Hilbert space projection argument in Kallenberg (2021). For proof of the second part we refer to Schervish (1995) or Robert (2007). ■

The posterior can also be used to construct *credible regions* or the *predictive distribution*.

An α-(posterior) credible region is a random set $C = C_{\alpha,n}$ such that in the situation above $\Pi(C|\mathbf{X}_{(n)}) \geq \alpha$ (Schervish, 1995; Robert, 2007). Since there are many such regions, one must choose a set according to some heuristic or decision theoretic foundation. One popular choice are highest posterior density (HPD) regions. If the set is an interval it is called a credible interval, for instance for a unimodal density in $\mathbb{R}^1$ one may choose $(1 - \alpha)/2$ and $(1 + \alpha)/2$-quantiles. Decision theoretic set estimation is introduced in Schervish (1995) or Robert (2007), for instance.

Bayes procedures require the choice of a prior and, as we will later see (compare Le Cam and Lo Yang (2000) "*Bayes procedures behave miserably*") the choice of prior can influence consistency, and because of this, definition 4.1 considers consistency in a prior-a.s.-sense. Common methods to construct a prior are the use of information available before data collection, *conjugate* priors, which given a certain shape of the likelihood produce posteriors of the same family, or *Jeffreys* prior which is chosen to be proportional to the square root of the Fisher information. Further constructions and examples of priors can be found in Ferguson (1974); Delaigle and Hall (2010); Phadia (2016); Ghosal and van der Vaart (2017); Castillo (2024), etc.

We will briefly introduce the concept of exchangeability that has been mentioned by Castillo (2014).

Definition 1.54 (Exchangeability (Schervish, 1995))
A finite set $X_1, \ldots, X_n$ of random quantities is said to be exchangeable *if every permutation of $(X_1, \ldots, X_n)$ has the same joint distribution. An infinite collection is exchangeable if every finite subcollection is exchangeable.*

Theorem 1.55 (De Finetti's Representation Theorem (Schervish, 1995))
Let $(\mathscr{X}, \mathfrak{X}, \mu)$ be a probability space, and let $(\mathscr{Y}, \mathfrak{Y})$ be a Borel space. For each $n \in \mathbb{N}$, let $Y_n : \mathscr{X} \to \mathscr{Y}$ be $(\mathfrak{X}, \mathfrak{Y})$-mb.. The sequence $(Y_n)_{n=1}^{\infty}$ is exchangeable if and only if there is a random probability measure $\mathbb{P}$ on $(\mathscr{Y}, \mathfrak{Y})$ such that, conditional on $\mathbb{P} = P$, $(Y_n)_{n=1}^{\infty}$ are independent and identically distributed (iid) with distribution P. Furthermore, if the sequence is exchangeable, then the distribution of $\mathbb{P}$ is unique, and $\mathbb{P}_n\mathcal{Y}$ converges to $\mathbb{P}\mathcal{Y}$ almost surely for each $\mathcal{Y} \in \mathfrak{Y}$.

1.3 Asymptotic Statistics

Aad van der Vaart (1998) writes: "*Why asymptotic statistics? The use of asymptotic approximations is two-fold. First, they enable us to find approximate tests and confidence regions. Second, approximations can be used theoretically to study the quality (efficiency) of statistical procedures.*" This paragraph firstly establishes the concepts of different forms of convergence on probability spaces and introduces their notation, then the most relevant asymptotic results will be stated.

Definitions 1.56, 1.58, and 1.59 are based on the assumption that $X, X_1, X_2, \ldots$ are random variables defined on the common probability space $(\Omega, \mathfrak{A}, \mathbb{P})$.

Definition 1.56 (Almost Sure Convergence)
The sequence $(X_n)_{n\in\mathbb{N}}$ converges $\mathbb{P}$-a.s. to X iff

$$\mathbb{P}\left[\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right] = 1,$$

in other words, the convergence is pointwise $\mathbb{P}$-a.e.. In this case we write $X_n \overset{\mathbb{P}\ a.s.}{\to} X$.

Lemma 1.57 (Convergence of Products with Degenerated Limits)
Let $(\Omega, \mathfrak{A}, \mathbb{P})$ be a probability space and let $(X_n)_{n\in\mathbb{N}}, (Y_n)_{n\in\mathbb{N}}, X, Y$ be real-valued random variables on this space. The convergence of $X_n \overset{\mathbb{P}\ a.s.}{\to} X$ to a degenerated random variable X with $\mathbb{P}\circ X^{-1} = \delta_x$ will be denoted as $X_n \overset{\mathbb{P}\ a.s.}{\to} x$. Now let $X_n \overset{\mathbb{P}\ a.s.}{\to} x$ and $Y_n \overset{\mathbb{P}\ a.s.}{\to} y$. Then it follows that $X_nY_n \overset{\mathbb{P}\ a.s.}{\to} xy$. This type of convergence translates to any number of factors by induction.

Proof. By definition of almost sure convergence:

$$\begin{aligned} X_n \overset{\mathbb{P}\ \text{a.s.}}{\to} x &\Leftrightarrow \mathbb{P}[\{\omega \in \Omega : X_n(\omega) = X(\omega)\}] = 1 \\ &\Leftrightarrow \exists A_1 \in \mathfrak{A} : \mathbb{P}A_1 = 1 \wedge \forall \omega \in A_1 : \forall \varepsilon_1 > 0 : \exists n_1 \in \mathbb{N} : \forall n \geq n_1 : |X_n(\omega) - x| < \varepsilon_1. \end{aligned}$$

Similarly:

$$\exists A_2 \in \mathfrak{A} : \mathbb{P}A_2 = 1 \wedge \forall \omega \in A_2 : \forall \varepsilon_2 > 0 : \exists n_2 \in \mathbb{N} : \forall n \geq n_2 : |Y_n(\omega) - y| < \varepsilon_2.$$

Since $A_1 \subset A_1 \cup A_2$, the monotonicity of $\mathbb{P}$ implies $\mathbb{P}A_1 = 1 \leq \mathbb{P}(A_1 \cup A_2) \leq 1$, hence $\mathbb{P}(A_1 \cup A_2) = 1$. It follows that $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cup A_2) = 1$, i.e. inductively the intersection of any number of sets with probability one has probability one.
The following argument holds pointwise $\forall \omega \in A_1 \cap A_2$ and $\forall n \geq \max\{n_1, n_2\} =: m$.

$$\begin{aligned} \varepsilon_1\varepsilon_2 &> |X_n(\omega) - x| \cdot |Y_n(\omega) - y| \\ &= |X_n(\omega)Y_n(\omega) - xy - y(X_n(\omega) - x) - x(Y_n(\omega) - y)| \\ &\geq |X_n(\omega)Y_n(\omega) - xy| - |y|\,|X_n(\omega) - x| - |x|\,|Y_n(\omega) - y| \\ &> |X_n(\omega)Y_n(\omega) - xy| - |y|\varepsilon_1 - |x|\varepsilon_2 \\ \Rightarrow |X_n(\omega)Y_n(\omega) - xy| &< \varepsilon_1\varepsilon_2 + |y|\varepsilon_1 + |x|\varepsilon_2 =: \varepsilon. \end{aligned}$$

For the second inequality compare $|a| = |a - b + b| \leq |a - b| + |b| \Leftrightarrow |a - b| \geq |a| - |b|$, for any $a, b \in \mathbb{R}$. Since ε is deterministic and can take any value in $(0, \infty)$, $X_n Y_n \overset{\mathbb{P}\text{ a.s.}}{\to} xy$ can be concluded. ■

Definition 1.58 (Convergence in Probability / Stochastically)
The sequence $(X_n)_{n\in\mathbb{N}}$ converges in probability to X iff

$$\forall \varepsilon > 0 : \lim_{n\to\infty} \mathbb{P}\left[\{\omega \in \Omega : \|X_n(\omega) = X(\omega)\| > \varepsilon\}\right] = 0.$$

In this case we write $X_n \overset{\mathbb{P}}{\to} X$.

For a real-valued random variable X, let $F_X(t) := \mathfrak{L}(X)((-\infty, t])$ declare the cumulative distribution function (cdf). This function is monotonous and right-continuous with finite left limits (càdlàg) with left limit 0 and right limit 1.

Definition 1.59 (Convergence in Distribution)
The sequence $(X_n)_{n\in\mathbb{N}}$ converges in distribution to X iff $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for every x in the set of continuity points of F_X. In this case we write $X_n \rightsquigarrow X$.

The following theorem is often used to examine almost sure convergence of arithmetic means; the proof is omitted here but can be found in (Kallenberg, 2021).

Theorem 1.60 (Strong Law(s) of Large Numbers (SLLN) of Kolmogorov, Marcinkiewicz & Zygmund)
Let $X, X_1, X_2, \ldots$ be iid random variables on $(\Omega, \mathfrak{A}, \mathbb{P})$, put $S_n := \sum_{k\in\{1,\ldots,n\}} X_k$, and fix a $p \in (0, 2)$. Then $n^{-1/p} S_n$ converges a.s. iff these conditions hold, depending on the value of p:

- *for $p \in (0, 1]$: $X \in L^p$*
- *for $p \in (1, 2)$: $X \in L^p$ and $\mathbb{E}X = 0$.*

The limit equals $\mathbb{E}X$ when $p = 1$ is chosen and is 0 otherwise.

In particular, this version of the SLLN states that if random variables are iid and integrable, the arithmetic mean converges almost surely to the expectation. There are other versions of SLLNs, among them one needed for section4.1; the proof uses the Borel-Cantelli lemma and can be found in Schervish (1995).

Theorem 1.61 (SLLN for Bounded Conditionally independent and identically distributed (ciid) Random Variables (Schervish, 1995))
Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of bounded random variables and let θ be a random variable such that the X_n are iid conditional on $\sigma(\theta)$ with $\mathbb{E}^\theta X_n = m(\theta)$. Then $Y_n := n^{-1} \sum_{j=1}^n X_j$ converges a.s. to $m(\theta)$.

Finally, two theorems dealing with mappings of convergent random elements are presented.

Theorem 1.62 (Continuous Mapping Theorem (CMT) (van der Vaart, 1998))
Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be continuous at every point in $\mathcal{X}$ such that $\mathfrak{L}(X)(\mathcal{X}) = 1$. Then

(i) $X_n \rightsquigarrow X \Rightarrow g(X_n) \rightsquigarrow g(X)$,

(ii) $X_n \overset{\mathbb{P}}{\to} X \Rightarrow g(X_n) \overset{\mathbb{P}}{\to} g(X)$,

(iii) $X_n \xrightarrow{a.s.} X \Rightarrow g(X_n) \xrightarrow{a.s.} g(X)$.

Theorem 1.63 (Slutsky Theorem (van der Vaart, 1998))
Let $(X_n)_{n\in\mathbb{N}}, (Y_n)_{n\in\mathbb{N}}$ and $(Z_n)_{n\in\mathbb{N}}$ be sequences of real-valued random variables, and X, Y, Z random variables, all defined on the same probability space $(\Omega, \mathfrak{A}, \mathbb{P})$. Further let Y and Z be degenerated, in the sense that $\mathbb{P}_Y = \delta_y, \mathbb{P}_Z = \delta_z$ for some $y, z \in \mathbb{R}$. Moreover let

$$X_n \rightsquigarrow X, \quad Y_n \xrightarrow{\mathbb{P}} Y, \quad Z_n \xrightarrow{\mathbb{P}} Z$$

then it follows that

$$Y_n X_n + Z_n \rightsquigarrow YX + Z. \tag{1.9}$$

Asymptotic equality of sequences will be denoted as (Tsybakov, 2009)

$$a_n \asymp b_n \quad :\Leftrightarrow \quad 0 < \liminf_{n\to\infty} \frac{a_n}{b_n} \leq \limsup_{n\to\infty} \frac{a_n}{b_n} < \infty. \tag{1.10}$$

Definition 1.64 (Stochastic Order (van der Vaart, 1998))
The following symbols are often used for a more concise notation. Let $(X_n)_{n\in\mathbb{N}}$, $(Y_n)_{n\in\mathbb{N}}$, and $(R_n)_{n\in\mathbb{N}}$ be sequences of random variables on the common probability space $(\Omega, \mathfrak{A}.\mathbb{P})$.

(a) $X_n = o_{\mathbb{P}}(1) :\Leftrightarrow X_n \xrightarrow{\mathbb{P}} 0,\ n \to \infty$

(b) $X_n = o_{\mathbb{P}}(R_n) :\Leftrightarrow X_n = R_n Y_n,\ Y_n = o_{\mathbb{P}}(1)$

(c) $X_n = \mathcal{O}_{\mathbb{P}}(1) :\Leftrightarrow \forall \varepsilon > 0 : \exists M \in (0, \infty) : \exists n_0 \in \mathbb{N} : \forall n \geq n_0 : \mathbb{P}(|X_n| > M) < \varepsilon$

(d) $X_n = \mathcal{O}_{\mathbb{P}}(R_n) :\Leftrightarrow X_n = R_n Y_n,\ Y_n = \mathcal{O}_{\mathbb{P}}(1)$

Example 1.65. Let $X_1, \ldots$ be an iid sequence of real-valued random variables on $(\Omega, \mathfrak{A}, \mathbb{P})$ satisfying $X_1 \in \mathcal{L}^2$. Define the sequence of arithmetic means $\bar{X}_n := \frac{1}{n}\sum_{j=1}^n X_j$. Then, since $\mathbb{E}\bar{X}_n \overset{\text{LIN, ID}}{=} \mathbb{E}X_1 < \infty$ and $\mathbb{V}\bar{X}_n \overset{iid}{=} \frac{\mathbb{V}X_1}{n}$, Theorem 1.37 implies $\left|\bar{X}_n - \mathbb{E}X_1\right| = o_{\mathbb{P}}(1)$. This is called the weak law of large numbers (WLLN).

Theorem 1.66 (Lindeberg-Lévy Central Limit Theorem (CLT) (Henze, 2024))
Let $X_1, X_2, \ldots$ be iid random variables satisfying $\mathbb{E}X_1^2 < \infty$ and $\mathbb{V}X_1 > 0$. Take $(S_n)_{n\in\mathbb{N}}$ to be the sequence of partial sums $S_n := \sum_{j=1}^n X_j$. Then the standardised version of S_n converges in distribution to a standard normal random variable:

$$\frac{S_n - \mathbb{E}S_n}{\sqrt{\mathbb{V}S_n}} \rightsquigarrow \mathcal{N}(0, 1)\,.$$

Other CLTs are versions of de Moivre-Laplace, Lindeberg-Feller, or Lyapunov.

1.4 Asymptotic Bayesian Statistics

The focus of Bayesian statistical analysis is centred on the posterior distribution, which in the dominated setting is given by the Bayes formula. Asymptotic Bayesian Statistics studies the asymptotic behaviour of the posterior, such as lower and upper bounds of the contraction rates and limiting shapes (Castillo, 2014). For the asymptotic behaviour we will follow the approach of Ghosal and van der Vaart (2017) who interpret their approach: *The framework for such a study is frequentist. It assumes the data are generated according to some "true" distribution, and the question is whether and how the posterior distribution can recover this data-generating mechanism.*

Let us now introduce the concept of consistency. For this purpose $\mathbf{X}_{(n)}$ denotes an observation in the sample space $(\mathscr{X}^n, \mathfrak{X}^{(n)})$.

Definition 1.67 (Posterior Consistency (Ghosal and van der Vaart, 2017))
The Posterior distribution $\Pi_n(\cdot|\mathbf{X}_{(n)})$ *is said to be* weakly consistent *at* $\theta_0 \in \Theta$ *if* $\Pi_n(\mathcal{U}^c|\mathbf{X}_{(n)}) \to 0$ *in* $(\mathbb{P}^{\theta_0})^{(n)}$*-probability, as* $n \to \infty$*, for every neighbourhood* $\mathcal{U}$ *of* θ_0*. It is said to be* strongly consistent *if the convergence is in the almost sure sense.*

Definition 1.68 (Kullback-Leibler Property (Ghosal and van der Vaart, 2017))
A density p_0 *is said to possess the Kullback-Leibler property relative to a prior* Π *or belong to the Kullback-Leibler support of* Π *if* $\Pi(p : \mathcal{K}(p_0, p) < \varepsilon) > 0$ *for every* $\varepsilon > 0$*. We write* $p_0 \in KL(\Pi)$*.*

Definition 1.69 (Contraction Rate (Castillo, 2024))
Let (Θ, d) *be a metric parameter space. We say a sequence* ε_n *(often tending to zero as* n *diverges) is a contraction rate around* θ_0 *for the posterior* $\Pi * \cdot|X)$ *if*

$$\int \Pi(d(\theta, \theta_0) > \varepsilon_n|X) d\mathbb{P}^{\theta_0} = o(1).$$

Theorem 1.70 (Schwartz (Ghosal and van der Vaart, 2017))
If $p_0 \in KL(\Pi)$ *and for every neighbourhood* $\mathcal{U}$ *of* p_0 *there exist tests* ϕ_n *such that* $P_0^n \phi_n \to 0$ *and* $\sup_{p \in \mathcal{U}^c} P^n(1 - \phi_n) \to 0$*, then the posterior distribution* $\Pi_n(\cdot|X_1, \ldots, X_n)$ *in the model* $X_1, \ldots, X_n|p \overset{iid}{\sim} p$ *and* $p \sim \Pi$ *is consistent at* p_0*.*

Remark 1.71. In essence this means that the prior assigns positive probability to a relevant neighbourhood of $p_0 = p(\cdot|\theta_0)$ and we can find tests to separate the true parameter from those that are far enough from it with error probabilities that converge to zero.

A consistency result for parametric models is given by the following theorem.

Theorem 1.72 (Doob's Consistency Theorem (van der Vaart, 1998))
Suppose that the sample space $(\mathscr{X}, \mathfrak{X})$ *to be a subset of an Euclidean space equipped with the corresponding Borel* σ*-field. Suppose that the conditional distributions are identifiable in the sense* $\theta_1 \neq \theta_2 \Rightarrow \mathbb{P}^{\theta_1} \neq \mathbb{P}^{\theta_2}$*. Then for every prior measure* Π *on* $(\Theta, \mathfrak{T})$ *the sequence of posterior measures is consistent* Π*-a.s.*

A key concept of asymptotic Bayesian statistics are Laplace-Bernstein-von Mises (BvM)-type theorems (compare Le Cam (1986a); Castillo (2014)) and will be discussed in section2.2.

1.5 Results from Multivariate Analysis

Judith Rousseau (2016) states local asymptotic normality (LAN) as one of three sufficient conditions for the derivation of BvM-type theorems. As Aad van der Vaart (1998) describes, a sequence of statistical models indexed by an open parameter space in $\mathbb{R}^d$ is defined to be locally asymptotically normal, if the log-likelihood ratios allow a certain quadratic expansion. In the multivariate parametric setting of section 4.1 this expansion is realised using the following version of the Taylor formula (compare Cartan (1971)). For this purpose, take $\mathscr{O} \subset \mathbb{R}^d$ to be an open set and $f : \mathscr{O} \to \mathbb{R}^p$ be a mapping, then $\mathcal{C}^n(\mathscr{O}, \mathbb{R}^p)$, $n \in \mathbb{N}_0$ describes the set of n-times continuously differentiable mappings from $\mathscr{O}$ to $\mathbb{R}^p$. To understand the formula let us remark that Duistermaat and Kolk (2004) define $h^{(j)} := (0, \ldots, 0, h_j, 0, \ldots, 0) \in \times_{i=1}^k \mathbb{R}^d$ and $(h_1, \ldots, h_k) := \sum_{j=1}^k h^{(j)}$.

Theorem 1.73 (Taylor Formula (Duistermaat and Kolk, 2004))
Assume $\mathscr{O} \subset \mathbb{R}^d$ to be an open and convex set and let $f \in \mathcal{C}^{k+1}(\mathscr{O}, \mathbb{R}^p)$, $k \in \mathbb{N}_0$. Then we have, for all a and $a+h$ in $\mathscr{O}$, the following version of the Taylor formula with the integral formula for the k-th remainder $R_k(a,h)$:

$$f(a+h) = \sum_{j=0}^{k} \frac{1}{j!} D^j f(a)(h^j) + \frac{1}{k!} \int_0^1 (1-t)^k D^{k+1} f(a+th)(h^{k+1}) dt,$$

where $D^j f(a)(h^j)$ means $D^j f(a)(h, \dots, h)$, i.e. the product of the j-th derivative of f at point a times h^j. Moreover

$$\|R_k(a,h)\| = \mathcal{O}\left(\|h\|^{k+1}\right), \quad h \to 0.$$

Let it be remarked that it can be useful to think of h as the difference $x-a$. The first derivative of a vector-valued function will often be denoted as the gradient: $\nabla \cdot := (D\cdot)^\top$

Another concept used in section 4.1 is the square root of a $(n \times n)$-matrix. If $\mathscr{M}_n$ denotes the set of square $(n \times n)$-matrices, a matrix $B \in \mathscr{M}_n$ is called the square root of $A \in \mathscr{M}_n$ iff $B^2 = A$.

Theorem 1.74 (Existence of Matrix Square Roots (Horn and Johnson, 2012))
Let $A \in \mathscr{M}_n$ be Hermitian – i.e. self-adjoint such that $A = \bar{A}^\top$ – and positive semidefinite, let $r := rank A$, and let $k \in \{2, 3, \dots\}$.

(i) *There is a unique Hermitian positive semidefinite matrix B such that $B^k = A$.*

(ii) *There is a polynomial p with real coefficients such that $B = p(A)$. Consequently, B commutes with any matrix that commutes with A.*

(iii) *$range A = range B$, so $rank A = rank B$.*

(iv) *B is real if A is real.*

Example 1.75. Any diagonisable matrix $A = V \Lambda V^{-1}$ has the square root $V \Lambda^{1/2} V^{-1}$ where $\Lambda^{1/2}$ is the diagonal matrix of the square roots of the eigenvalues of A.

2 Motivation and Overview of Previous Findings

2.1 Group Testing

2.1.1 Logistic and Binary Regression

Generally, in regression models, the goal is to determine the influence of a (possibly multivariate) covariate (also called regressor, predictor or explanatory variable) to a response (regressand or target) variable. Logistic regression stems from the application of ideas of linear regression to classification problems. For instance, Hastie, Tibshirani, and Friedman (2009) present the classification problem where the random group indicator G can take values $k \in \{0, \dots, K-1\}$, and the probability that G was drawn from class k conditional on the realisation of the data $\{\mathbf{X} = \mathbb{X}\}$ is modelled as

$$\mathbb{P}(G = k|\mathbf{X} = \mathbb{X}) = H(\mathbb{X}\beta_k)$$

where $\mathbb{X}$ represents the data matrix, including a column of ones, and $\beta_k \in \mathbb{R}^{d+1}$ are the regression coefficients to be estimated, for instance by a maximum likelihood procedure, which often requires numeric optimisation. H is a cdf, for instance the *logit* or *probit* link.

Definition 2.1 (Logit and Probit Link Function)
(Hastie et al., 2009; Ghosal and van der Vaart, 2017)

(i) *The logit cdf is defined as* $H(t) := (1 + \exp(-t))^{-1}$,

(ii) *and the probit link as* $H(t) := \int \mathbb{1}_{(-\infty,t]}(y)(2\pi)^{-1/2} \exp(-y^2/2)dy$.

A special case of this problem arises when $K = 2$ and is called binary regression. For instance one might model conditionally on p

$$Y|\mathbf{X} \sim \text{Ber}(p(\mathbf{X}))$$

where p can be interpreted as some kind of risk function of the covariate $\mathbf{X}$ for the outcome Y, e.g. the risk of heart disease given the number of cigarettes smoked and their age.

Choudhuri, Ghosal, and Roy (2004) apply a nonparametric approach to model the 'response probability function'

$$p(\mathbf{x}) := \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) =: H \circ \xi(\mathbf{x}) \tag{2.1}$$

in a Bayesian setting using a Gaussian process prior. Regarding the link function they write: "*A completely nonparametric estimate may be obtained by keeping the link function fixed, but modelling* ξ *as an arbitrary function. This kind of flexible shape modelling can produce any useful shape, especially for spatial data. . . . In contrast, the approach of varying* H *and keeping the form of* ξ *fixed can produce only specifically shaped equal probability contours.*" They go on and present a Markov Chain Monte Carlo (MCMC) algorithm including a model with priors on the hyperparameters of the Gaussian process and discuss 'robustification' as well as a simulation study.

In the context of binary regression, group testing can be seen as a statistical problem for which the pairs (X_j, Y_j) are pooled and not all Y_j can be observed. This specific design will now be introduced.

2.1.2 Group Testing

The group testing design was originally introduced by Dorfman (1943) in order to detect syphilis before army enlistment during World War II more cost efficiently. Dorfman argues that, for safety reasons, a *complete elimination of defective units is desired.* The process of testing involves

drawing blood from candidates and subjecting it to laboratory analysis to reveal the presence of antigen as an indicator of infection. Since laboratory analysis is costly, Dorfman suggests pooling the blood samples whilst retaining part of the original sample. Suppose $n \in \mathbb{N}$ samples are pooled. Since the response variable is binary, a negative test – assuming dilution, etc. not having a relevant effect – suggests that none of the samples contained antigen. However, if the test was positive, the retained samples are analysed again individually. If $p \in (0,1)$ denotes the prevalence, the number of tests needed is $T_n \sim (1-p)^n \delta_1 + (1-(1-p)^n)\delta_{n+1}$ and the expected number of tests necessary is

$$\mathbb{E}T_n = (n+1) - n(1-p)^n.$$

This design only makes sense, if there is a cost reduction on average, i.e. $\mathbb{E}T_n < n \Leftrightarrow p < 1 - n^{-1/n}$. This function takes its maximum of about 31% at $n = 3$, and therefore the design should only be applied if the prevalence is lower than this percentage, and if it is more practical to test pooled samples. Dorfman then calculates the *relative testing cost* as the ratio of the expected number of tests in the group design to that of the *individual technique*, i.e.

$$C_n := \frac{\mathbb{E}T_n}{n} = \frac{n+1}{n} - (1-p)^n$$

and the *percentage saving attainable* is one minus C_n. Table 2.1.2 gives an overview on optimal group sizes and attainable savings in percent given a certain prevalence in the population.

Further applications include SARS-CoV-2 / COVID-19 PCR-pool tests (Martin et al., 2021), detection of chlamydia, hepatitis (McMahan et al., 2012), water pollution (Matsushima et al., 2024; Bryan and Gershman, 1975), and many more, where the design explained above is reasonable.

Delaigle and Meister (2011) investigate a group testing regression problem for a model very similar to that of section 3, in particular the goal is to estimate $p(x)$ as in (2.1) for $x \in \mathbb{R}$. They assume their data (X_{ij}, Y_{ij}) to be iid for $j \in \{1, \dots, J\}, i \in \{1, \dots, n_j\}$ and presume that, after pooling, only the maximum value $Y_j^* := \max\{Y_{1j}, \dots, Y_{n_j j}\}$ is observed alongside the covariates. Derivation of the conditional expectation $\mathbb{E}^p[Y_j^* | X_{1j}, \dots, X_{n_j j}]$ is analogous to that in section 3. They then derive the relationship

$$p(x) = 1 - \frac{q}{\mu_Z^*} g(x),$$

and propose an unbiased estimator for μ_Z^*, a maximum likelihood estimator (MLE) based on the independent, but non-identically distributed variables $1 - Y_j^*$, as well as a local polynomial regression estimator (LPE) of degree l for g.

Table 1: Overview of efficiency increase by pooling relative to prevalence. Compare Dorfman (1943)

Prevalence Rate	Optimal Group Size	Relative Cost	Attainable Savings
0.0001	101	0.0200	0.9800
0.0010	32	0.0628	0.9372
0.0050	15	0.1391	0.8609
0.0100	11	0.1956	0.8044
0.0500	5	0.4262	0.5738
0.1000	4	0.5939	0.4061
0.1500	3	0.7192	0.2808
0.2000	3	0.8213	0.1787
0.2500	3	0.9115	0.0885
0.3000	3	0.9903	0.0097

LPEs can be found in Fan and Gijbels (2018) and utilise the height $\hat{\beta}_0$ of an estimator derived as

$$\hat{\beta} = \left(\mathbf{X}^\top \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{y} \tag{2.2}$$

for the local function value of $g(x)$ (in general $g(x) = \mathbb{E}(Y|X = x)$), where

$$\mathbf{X} := \begin{pmatrix} 1 & (X_1 - x) & \dots & (X_1 - x)^l \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (X_n - x) & \dots & (X_n - x)^l \end{pmatrix}, \ \mathbf{y} := \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \ \hat{\beta} := \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_l \end{pmatrix}, \ \mathbf{W} := \operatorname{diag}(K_h(X_j - x)),$$

for the covariate data matrix $\mathbf{X}$ and a kernel function K depending on the bandwidth h which is used to weight and smooth the observations around x.

Remark 2.2. Do not confuse the kernel function (Tsybakov, 2009) below with the probability kernel of Definition 1.47.

Definition 2.3 (Kernel Function)
A kernel function is a Lebesgue-integrable function $K : \mathbb{R}^d \to \mathbb{R}^1$ satisfying $\int K d\lambda^{(d)} = 1$.
The bandwidth-rescaled version is denoted as $K_h(\mathbf{x}) := h^{-d} K(h^{-1}\mathbf{x})$.
In the case $d = 1$, the kernel is of order l, iff $\forall j = 1, \dots, l : \int i d^j K d\lambda = 0$.

Note that such a LPE cannot be defined for the data directly, as the Y_{ij} are not observed – such a variable is called *latent* or *hidden*. Delaigle and Meister (2011) state that "*the success of the estimator depends crucially on the value of h, which must be chosen with much care.*" Their main result states that given $q \geq q_0 > 0$, some local boundedness and continuity conditions on the density of X_{11}, p, and the kernel, as well as the bandwidth rate condition $h \to 0$ and $Nh \to \infty$ as $N := \sum_{j=1}^{J} n_j \to \infty$, the estimator $\hat{p}(x)$ has some favourable asymptotic qualities, namely the (local) squared error $|\hat{p}(x) - p(x)|^2$ equals an asymptotic squared error plus a stochastically negligible term based on sample size N and bandwidth h. Furthermore they derive an optimal bandwidth of order $h \asymp N^{-\frac{1}{2l+3}}$. They go on to conduct a numerical study for the heteroscedastic data, proposing two automatic selection methods for the bandwidth, where the bandwidth is chosen as the argument minimising the asymptotic weighted mean integrated squared error (MISE).

Definition 2.4 (MISE (Tsybakov, 2009))
For iid real-valued random variables $X_1, \dots, X_n$ and an estimator $\hat{p}(x) = \hat{p}(x, X_1, \dots, X_n)$ for $p : \mathbb{R} \to [0, \infty)$, the MISE is defined as

$$\mathbb{E} \int |\hat{p}(x) - p(x)|^2 \, dx = \int \left(\mathbb{V}(x) + \mathbb{B}(x)^2 \right) dx,$$

where the second term stems from a bias-variance decomposition.

For the weighted version replace $dx := \lambda(dx)$ with $w(x)dx$. For a parametric estimation of the bias Delaigle and Meister (2011) propose a *rule of thumb* bandwidth, and a *plug-in* method if the bias is estimated nonparametrically. They also show the practical applicability of their method in a simulation study before applying it to data from the National Health and Nutrition Examination Survey (NHANES).

Works by McMahan, Tebbs, and Bilder (2012) and Delaigle and Hall (2015) extend this estimation problem taking the imperfection of the real world into account. Delaigle and Hall (2015) write that in practice the test "*is sensitive to the proportion of contaminated items in the group,*

rather than to the sheer existence of one or more contaminated items." As in the previous example, the data is $(X_{1j}, \dots, X_{n_j j}, Y_j^*)$, $j \in \{1, \dots, J\}$, but due to measurement errors Y_j^* may differ from the true disease or pollution status

$$\tilde{Y}_j^* := \max\{\tilde{Y}_{1j}, \dots, \tilde{Y}_{n_j j}\},$$

and the new goal is to estimate

$$m(x) := \mathbb{P}(\tilde{Y}_{ij} = 1 | X_{ij} = x). \tag{2.3}$$

They introduce a biomarker concentration level for the i-th individual that was put in the j-th group as the unobserved quantity B_{ij} and it is assumed that the pool j contains the average $\bar{B}_j := n_j^{-1} \sum_{i=1}^{n_j} B_{ij}$ which is measured as an optical density reading

$$W_j := \bar{B}_j + U_j, \quad \bar{B}_j \perp\!\!\!\perp U_j,\ U_j \sim f_U, U_j \perp\!\!\!\perp \tilde{Y}_j^*,$$

with a known error density f_U, and what is actually observed becomes

$$Y_j^* := \mathbb{1}_{(t_{0,j}, \infty)}(W_j), \tag{2.4}$$

where the cutoff point $t_{0,j}$ is assumed to be given. They add: "*In practice it can be chosen so in order to minimise the variance of an estimator of m.*" Beside $m(x)$, Delaigle and Hall (2015) estimate the probability, q, that an individual chosen at random is disease free, the *specificity* (true negatives)

$$\mathrm{Sp}_j = \mathbb{P}(W_j \le t_{0,j} | \tilde{Y}_j^* = 0),$$

and the *sensitivity* (true positives)

$$\mathrm{Se}_j = \mathbb{P}\left(W_j > t_{0,j} \middle| \sum_{i=1}^{n_j} \tilde{Y}_{ij} = k \right), \quad k \in \{1, \dots, n_j\}$$

of the test. As for the statistical estimation, Delaigle and Hall (2015) firstly propose an oracle local polynomial estimator of m, which would be consistent for the $(X_{ij}, \tilde{Y}_{ij})$ data that is not given, but the $\tilde{Y}_{ij}$ are replaced by Y_j^* in each group. Then they propose a nonparametric estimator for the specificity and sensitivity – modelling it as a *deconvolution problem* – utilising a kernel estimator whilst exploiting the inverse Fourier transform. After this they derive the likelihood $\mathscr{L}(q)$ as a function of the specificity, sensitivity and Y_j^* stating the MLE to be found after plugging in the estimators for the specificity and the sensitivity. Lastly, they propose a "*fully data-driven nonparametric estimator of m*".

Theoretical properties of the results are stated to depend on the rate of decay of the Fourier transform and the asymptotic properties are derived for two cases: one which encompasses Laplace distributions, and one including normal distributions. The main result states that despite the problems being ill-posed the "*estimators of q, Sp and Se are root-N consistent*" and the "*estimator of m(x) converges at the rate it would enjoy if q, Sp and Se were known*", which is specified in their last theorem.

Also included are simulations, where the bandwidth is chosen wrt the asymptotic MISE, similar to Delaigle and Meister (2011), proposals on the choice of the cutoffs and illustrations for data from nine Irish prisons regarding hepatitis B and C infections.

Another extension to the approach by Delaigle and Meister (2011) is the contribution by Delaigle and Hall (2012) where, instead of assigning the covariates X_i randomly to groups, the list is simply partitioned into pools of equal size n (*homogeneous group testing*). Their approach is to then assume that

$$\mathbb{E}(1 - Y_j^* | X_{1j}, \dots, X_{nj}) = \prod_{i=1}^{n} (1 - p(X_{ij}))$$

is described accurately enough by

$$\left(1 - p(\bar{X}_j)\right)^n, \quad \bar{X}_j := \frac{1}{n}\sum_{i=1}^{n} X_{ij}.$$

They then define $\mu(x) := (1-p(x))^n$, which is then estimated utilising the data $(\bar{X}_j, 1 - Y_j^*)$ by means of a *linear smoother* such as the LPE, thus

$$\hat{p}(x) := 1 - \hat{\mu}(x)^{1/n}.$$

To derive theoretical properties of their estimators, Delaigle and Hall (2012) express the prevalence as

$$p(x) = \delta(N)\pi(x),$$

where δ is a sequence of positive numbers which, in the extreme case, tends to zero as the sample size N diverges, and π is taken to be a fixed nonnegative function. Numerous conditions are developed, among them conditions on the smoother (compare Tsybakov (2009)), continuity of f_X, uniform boundedness of p away from 1 and Hölder continuity for the first two derivatives of π. Included are three theorems regarding the asymptotic behaviour of the bias when $n\delta$ diverges as N diverges, a comparison with the approach of Delaigle and Meister (2011), as well as generalisations to unequal groups and the multivariate case. They also apply their method to the NHANES data making it directly comparable in practice.

Their main result is a comparison between *moderate levels of pooling* to the case of "*over-pooling*". In the first case, more accurate nonparametric estimators are obtained by homogenous pooling and "*the same convergence rate as in the case of no pooling*" is achieved. In the latter case, they show a different rate of convergence, however, the disadvantages are "*no more than a logarithmic factor.*"

In "*New approaches to nonparametric and semiparametric regression for univariate and multivariate group testing data*", Delaigle, Hall, and Wishart (2014) summarise many methods and approaches and study new ideas.

Firstly, they review the previous two articles again. Then they consider a *partially linear model* (compare Härdle et al. (2000)) for the case where the covariates $\mathbf{X}_{ij} = (U_{ij}, \mathbf{V}_{ij}^\top)^\top$ consist of one continuous variable (e.g. weight) and a vector of discrete values (e.g. gender, number of cigarettes smoked). Note that number of cigarettes smoked could be considered approximately continuous while age could be discretised. The idea is to estimate a function g nonparametrically as above and to simultaneously estimate a parameter $\gamma \in \mathbb{R}^{d-1}$ for the model

$$p(\mathbf{X}_{ij}) = g(U_{ij}) + \gamma^\top \mathbf{V}_{ij}.$$

They develop an estimator for $1 - p(x)$ and also present a "*centralised bias*" version of it, and show asymptotic normality of their estimator of under a number of conditions.

Next, they extend the estimator of Delaigle and Meister (2011) to a multivariate setting, where they estimate $1 - p(x)$ via a local constant, i.e. Nadaraya-Watson estimator with a d-variate kernel function. As an alternative they present a "*single-index model*" where the dependence of p is modelled solely through the parameter β_0 such that

$$1 - Y_j^* = q_0^{n_j - 1} g\left(\beta_0^\top X_{ij}\right) + \epsilon_{ij}, \quad q_0 := 1 - p(X_{ij}), \quad \mathbb{E}(\epsilon_{ij}|X_{ij}) = 0.$$

A second theorem (partly in the supplementary material) including the asymptotic bias and variance term is given for this estimator.

Furthermore they offer computational advice for the choice of the bandwidth inspired by leave-one-out cross-validation.

Last but not least, they present examples in the form of simulations, and with data from NHANES.

A Bayesian approach to group testing regression has been presented by McMahan, Tebbs, Hanson, and Bilder (2017) where models are developed for different pooling designs and the estimators are evaluated via simulations: "*We use simulation to investigate the quality of our regression methods under a variety of group testing protocols and prior models.*" McMahan et al. (2017) name the three models considered

1. master pool testing, when only the grouped observations are available,
2. Dorfman testing, the hierarchical design introduced by Dorfman (1943),
3. and array testing, when the individual samples are arranged in an array, such that each sample is included in both a row group and a column group.

As in the contribution by Delaigle and Hall (2015) the $\tilde{Y}_i$ are assumed to be latent and the group assignments are defined by the index sets ("*protocols*")

$$P_j \subset \{1, \ldots, N\} : \cup_{j=1}^{J} P_j = \{1, \ldots, N\},$$

such that the unobserved group statuses can be represented as

$$\tilde{Y}_j^* = \mathbb{1}_{\{\sum_{i \in P_j} \tilde{Y}_i > 0\}}.$$

They assume that the distribution of $\tilde{Y}_i$ can be described by a link function H and a linear relationship via a regression parameter $\beta \in \mathbb{R}^{r+1}$ as

$$\mathbb{P}(\tilde{Y}_i = 1 | \mathbf{x}_i, \beta) = H^{-1}(\mathbf{x}_i^\top \beta).$$

Under the assumption that the $\tilde{Y}_i$ are conditionally independent (Definition 1.49) given the covariates, their conditional distribution of the vector $\mathbf{Y}^*$ is

$$\begin{aligned}
&\pi\left(\mathbf{Y}^* | \mathrm{Se}, \mathrm{Sp}, \mathbf{X}, \beta\right) \\
&= \sum_{\tilde{\mathbf{Y}} \in \{0,1\}} \prod_{j=1}^{J} \left[\mathrm{Se}_j^{Y_j^*} (1 - \mathrm{Se})^{1 - Y_j^*} \right]^{\tilde{Y}_j^*} \left[1 - \mathrm{Sp}_j^{Y_j^*} (\mathrm{Se})^{1 - Y_j^*} \right]^{1 - \tilde{Y}_j^*} \\
&\quad \cdot \prod_{i=1}^{N} H^{-1}(\mathbf{x}_i^\top \beta)^{\tilde{Y}_i} \left(1 - H^{-1}(\mathbf{x}_i^\top \beta)\right)^{1 - \tilde{Y}_i}.
\end{aligned}$$

They then derive the posterior distribution, given the prior $\beta \sim \mathcal{N}_{(r+1)}(\mathbf{a}, \mathbf{R})$, up to a proportional factor, which is then sampled via a Metropolis-Hastings (MH)-style algorithm (compare Chib and Greenberg (1995); Roberts and Smith (1994); Roberts and Rosenthal (2001); Kunick (2018)). They included an extension with unknown assay accuracies $\mathrm{Se}_j, \mathrm{Sp}_j$ which are "*pool-specific, reflecting that different pools could have different accuracies depending on how large the pools are and what type of assay is used.*" The model is then applied to simulation studies and Iowa chlamydia data.

Two more contributions to Bayesian group testing focussing on computational aspects are the works of Bai et al. (2019), where an adaptive algorithm with a Expectation Maximisation (EM)-style optimisation of computational complexity of $\mathcal{O}(N \log(N))$ is proposed, as well as the article by Tatsuoka et al. (2022), proposing a lattice-based model for group testing under dilution where the computation is optimised by a look-ahead backtracking algorithm. They state that their "*Bayesian halving algorithm, has attractive optimal convergence properties.*" Since their

approach was developed during and after the Coronavirus pandemic, Tatsuoka et al. (2022) add: "*This work has particular relevance given the pressing public health need to enhance testing capacity for coronavirus disease 2019 and future pandemics, and the need for wide-scale and repeated testing for surveillance under constantly varying conditions. The proposed Bayesian approach allows for dilution effects in group testing and for general test response distributions beyond just binary outcomes.*"

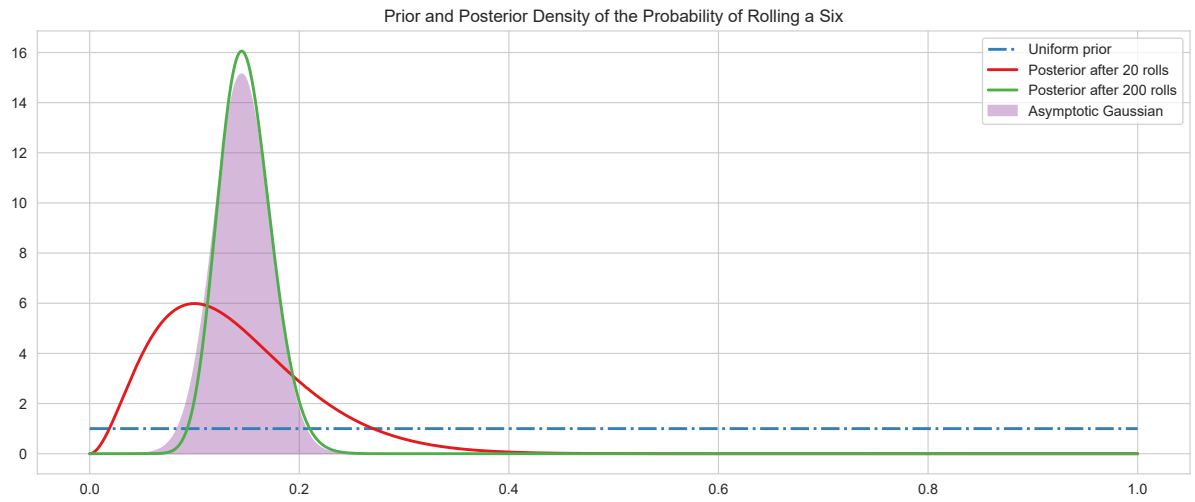

Figure 1: Illustration of Laplace's experiment for the example of n die rolls and the true probability of rolling a six 1/6 (fair dice).

2.2 Bernstein-von Mises (BvM) Theorems

"*The Laplace-Bernstein-von Mises theorem asserts that the posterior distribution of a parameter in a smooth finite-dimensional model is approximately a normal distribution if the number of observations tends to infinity.*" (Kleijn and van der Vaart, 2012)

Example 2.5. Laplace-Bernstein-von Mises theorems date back to the work of Pierre Simon Laplace published in 1774. Laplace considered an experiment of the form

$$\mathscr{E}_n := (\{0,\dots,n\}, \mathfrak{P}(\{0,\dots,n\}, \{\text{Bin}(n,\theta) : \theta \in (0,1)\}),$$

endowing θ with a uniform prior $\mathcal{U}_{(0,1)} = \text{Beta}(1,1)$, which is conjugate and produces the posterior $\Pi(\cdot|X) = \text{Beta}(X+1, n-X+1)$. "*Laplace observed and proved that this asymptotically looks like* $\mathcal{N}(X/n, \theta_0(1-\theta_0)/n)$" if the data had been generated by $\text{Bin}(n,\theta_0)$ (Castillo, 2024). This experiment is illustrated in Figure 1 for $\theta_0 = 1/6$ for a fair (generated by `numpy`) die roll.

Ghosal and van der Vaart (2017) write in *Fundamentals of Nonparametric Bayesian Inference*: "*The Bernstein-von Mises theorem for regular parametric models implies that the posterior distribution of the parameter centred at the MLE converges to the same normal distribution as that of the limit of the normalised MLE. Thus, asymptotically, Bayesian and sampling probabilities agree, so confidence regions of approximate frequentist validity may be generated from the posterior distribution. Cox, Freedman and others showed that such a result should not be expected for curve estimation problems. Some positive results have been obtained by Lo, Kim, Lee, Shen, Castillo, Nickl, Leahu, Bickel, Kleijn and others.*"

2.2.1 Intuitive explanation

In *Asymptotic Statistics*, van der Vaart (1998) presents the following intuition: Suppose $\mathbf{X}_{(n)} = (X_1,\dots,X_m)^\top$ was obtained as a random sample, i.e. independent and identically distributed, by repeating the experiment $\mathscr{E} = (\mathscr{X}, \mathfrak{B}(\mathscr{X}), \{\mathbb{P}^\theta : \theta \in \Theta\})$, where $\Theta \subset \mathbb{R}^k$ is open, $\mathscr{X} \subset \mathbb{R}^1$, and the experiment is assumed to be dominated as in the situation of section 1.2, where the dominating measures are $\mu = \lambda$ and $\nu = \lambda^{(k)}$. Then the posterior has a density of the form

$$\pi(\theta|\mathbf{X}_{(n)}) = \frac{\prod_{j=1}^n p(X_j|\theta)\pi(\theta)}{\int \prod_{j=1}^n p(X_j|t)\pi(t)dt}.$$

A rescaling of the parameter $h := \sqrt{n}(\theta - \theta_0) \Leftrightarrow \theta = \theta_0 + n^{-1/2}h$ leads to the consideration of

$$\pi(h|\mathbf{X}_{(n)}) = \frac{\prod_{j=1}^n p(X_j|\theta_0 + n^{-1/2}h)\pi(\theta_0 + n^{-1/2}h)}{\int \prod_{j=1}^n p(X_j|\theta_0 + n^{-1/2}t)\pi(\theta_0 + n^{-1/2}t)dt}.$$

Given some smoothness constraints, $\pi(\theta_0 + n^{-1/2}h)$ behaves asymptotically like $\pi(\theta_0)$ "*and* π *cancels from the expression for the posterior density.*" Moreover $\{\mathbb{P}^{\theta_0+n^{-1/2}h} : h \in \mathbb{R}^k\}$ is LAN and the "*likelihood ratio processes* $h \mapsto \prod_{j=1}^n p(X_j|\theta_0 + n^{-1/2}h)/p(X_j|\theta_0)$ *behave asymptotically like those of a normal experiment*", and we can expect the posterior density to behave like the Lebesgue density of $\mathcal{N}_k\left(h, \mathcal{I}(\theta_0)^{-1}\right)$.

2.2.2 Local Asymptotic Normality

We consider the experiment from section 2.2.1, where $\mathbb{P}^\theta$ is the image measure of X_1. We will then draw a ciid (conditional on $\sigma(\theta)$) sample such that

$$\mathscr{E}_{1,n} := (\mathscr{X}^n, \mathfrak{B}(\mathscr{X}^n), \{\otimes_{j=1}^n \mathbb{P}^\theta : \theta \in \Theta\}) \tag{2.5}$$

becomes our new experiment. We take $p(\cdot|\theta) := p_\theta(\cdot) := \frac{dP^\theta}{d\mu}$ to be our densities. LAN requires the model to be sufficiently smooth. The local nature stems from the reparameterisation $h := \sqrt{n}(\theta - \theta_0)$ such that experiments of the form

$$\mathscr{E}_{2,n} := (\mathscr{X}^n, \mathfrak{B}(\mathscr{X}^n), \{\otimes_{j=1}^n \mathbb{P}^{\theta_0+n^{-1/2}h} : h \in \mathbb{R}^k\}) \tag{2.6}$$

arise. van der Vaart (1998) warns that if the fixed point θ_0 is not an inner point one must assume this to be defined arbitrarily. Following are the main statements adapted from van der Vaart (1998). Let $\ell_\theta(x) := \ln p(x|\theta)$ be the log-likelihood and let a dot denote the derivative wrt θ. We call $\dot{\ell}_\theta$ the score function.

Definition 2.6 (Differentiability in Quadratic Mean (van der Vaart, 1998))
The model above is said to be differentiable in quadratic mean (DQM) at θ*, iff*

$$\int \left(\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2}h^\top \dot{\ell}_\theta \sqrt{p_\theta}\right)^2 d\mu = o(\|h\|^2), \quad h \to 0.$$

Theorem 2.7 (Order of Likelihood Ratios (van der Vaart, 1998))
Suppose that Θ *is an open subset of* $\mathbb{R}^k$ *and that the model above is DQM at* θ*. Then* $\mathbb{P}^\theta \dot{\ell}_\theta = \mathbb{E}_{\mathbb{P}^\theta}\dot{\ell}_\theta = 0$ *and the Fisher information matrix* $\mathcal{I}(\theta) = \mathbb{P}^\theta \dot{\ell}_\theta \dot{\ell}_\theta^\top$ *exists. Furthermore, for every converging sequence* $h_n \to h$ *as* $n \to \infty$ *one has*

$$\ln\left[\prod_{j=1}^n \frac{p_{\theta+n^{-1/2}h}}{p_\theta}(X_j)\right] = n^{-1/2}\sum_{j=1}^n h^\top \dot{\ell}_\theta(X_j) - \frac{1}{2}h^\top \mathcal{I}(\theta)h + o_{\mathbb{P}^\theta}(1).$$

The argument can be made with a Taylor expansion either of first or second degree. van der Vaart (1998) also provides a lemma giving the sufficient condition that if the square root of the likelihood is continuously differentiable in θ for every x, and t if $\mathcal{I}(\theta)$ exists and is continuous in θ, the corresponding model is differentiable in quadratic mean.

Example 2.8. van der Vaart (1998) states that exponential families

$$p(x|\theta) = d(\theta)h(x)\exp\left(\langle Q(\theta), t(x)\rangle\right)$$

generally fulfil the requirements for the asymptotic expansion above.

Example 2.9. A counterexample is given by the uniform distributions $\mathcal{U}_{[0,\theta]}$ as their support "*depends too much on the parameter*", and

$$\mathbb{P}^{\theta+h}(p_\theta = 0) = \int_{\mathbb{R}^1 \setminus [0,\theta]} \frac{1}{\theta+h} \mathbb{1}_{[0,\theta+h]} d\lambda = \frac{h}{\theta+h} = \mathcal{O}(h) \neq o(h^2), \quad h \to 0,$$

which van der Vaart (1998) states as a requirement.

The log-likelihood ratio of two Gaussian shift experiments looks similar to the asymptotic term in theorem 2.7:

$$\ln \frac{d\mathcal{N}_k\left(h, \mathcal{I}(\theta)^{-1}\right)}{d\mathcal{N}_k\left(0, \mathcal{I}(\theta)^{-1}\right)} = \langle \mathcal{I}(\theta)X, h\rangle - \frac{1}{2}\langle h, \mathcal{I}(\theta)h\rangle,$$

which, as the next theorem shows, is no coincidence.

Theorem 2.10 (LAN of DQM Experiments (van der Vaart, 1998))
Assume that the experiment of the form $\mathscr{E}_{1,n}$ is differentiable in quadratic mean at the point θ with nonsingular Fisher information matrix $\mathcal{I}(\theta)$. Let T_n be statistics in the experiments $\mathscr{E}_{2,n}$ such that the sequence T_n converges in distribution under every h.

Then there exists a randomised statistic T in the experiment $(\mathscr{X}, \mathfrak{B}(\mathscr{X}), \{\mathcal{N}_k\left(h, \mathcal{I}(\theta)^{-1}\right) : h \in \mathbb{R}^k\})$ such that $T_n \rightsquigarrow T$ for every h.

Related concepts discussed in Le Cam and Lo Yang (2000) are the concepts of experiments being locally asymptotically mixed normal (LAMN) or locally asymptotically quadratic (LAQ).

2.2.3 Selected Bernstein-von Mises Theorems

Theorem 2.11 (BvM with Testing Condition (van der Vaart, 1998; Castillo, 2024))
Let the experiment $\mathscr{E}_{1,n}$ be DQM at θ_0 (van der Vaart, 1998) / LAN at θ_0 (Castillo, 2024) and let the Fisher information matrix $\mathcal{I}(\theta_0)$ be nonsingular. Suppose that for every $\varepsilon > 0$ there exists a sequence of tests ϕ_n such that

$$\otimes_{j=1}^n \mathbb{P}^{\theta_0}\phi_n \to 0, \qquad \sup_{\theta:\|\theta-\theta_0\|\geq\varepsilon} \otimes_{j=1}^n \mathbb{P}^{\theta}(1-\phi_n) \to 0. \tag{2.7}$$

Furthermore let the prior measure be absolutely continuous wrt the Lebesgue measure in a neighbourhood of θ_0 with continuous and positive density at θ_0. Moreover let

$$\Delta_{n,\theta_0} := n^{-1/2} \sum_{j=1}^n \mathcal{I}(\theta_0)^{-1} \dot{\ell}_{\theta_0}(X_j). \tag{2.8}$$

Define the map $\tau : \theta \mapsto \sqrt{n}(\theta - \theta_0)$. Then the corresponding posterior distributions satisfy

$$d_{TV}\left(\Pi(\cdot|\mathbf{X}_{(n)}) \circ \tau^{-1}, \mathcal{N}_k\left(\Delta_{n,\theta_0}, \mathcal{I}(\theta_0)^{-1}\right)\right) = o_{\mathbb{P}^{\theta_0}}(1).$$

Remark 2.12. The posterior in the theorem is examined with respect to the local parameter $h = \tau(\theta)$.

Remark 2.13. van der Vaart (1998) writes that the random Gaussian measure in theorem 2.11 can instead be centred around the standardised asymptotically efficient estimators $\sqrt{n}(\hat{\theta}_n - \theta)$. This is also the version presented by Castillo (2024).

Remark 2.14. The testing condition (2.7) – for the hypotheses $H_0 : \theta = \theta_0,\ H_1 : \|\theta - \theta_0\| \geq \varepsilon$ – ensures that we can separate the true parameter from parameters that are far enough from it (complements of balls $\mathcal{B}(\theta_0, \varepsilon)$ with tests that have decreasing first and second type error probabilities, i.e. they are uniformly consistent (van der Vaart, 1998; Rousseau, 2016).

Remark 2.15. Ghosal and van der Vaart (2017) name this theorem BvM-Le Cam theorem (Theorem 12.1 in their book).

Example 2.16 (Continuation of the die roll example 2.5). For this example I have rolled a die 51 times and observed seven times the face value *six* among them. An exact $(1-\alpha)$-credibility or credible interval is constructed by taking a and b to be the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantile of the posterior $\Pi(\cdot|\mathbf{X}_{(n)})$. Using the BvM theorem one may construct approximate credible intervals using a Gaussian. This is illustrated in Figure 2. Castillo (2024) writes that the $(1-\alpha)$-credible interval can also be taken as an asymptotic confidence interval of level $(1-\alpha)$.

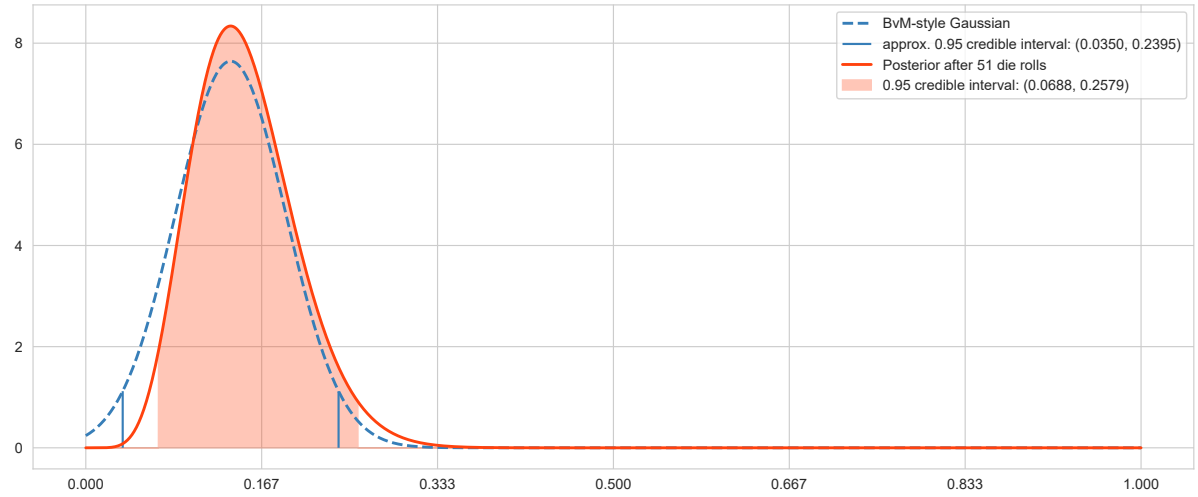

Figure 2: Illustration of the posterior distribution of the Bernoulli-parameter for rolling a six, an exact credible interval, and an approximate credible interval after 51 die rolls.

Lemma 2.17 (Sufficient Conditions for the BvM (Le Cam, 1986a,b; Le Cam and Lo Yang, 2000))

(i) $\theta_0 = 0$, Θ *is the intersection of a closed set with an open set.*

(ii) $\mathbb{P}^s = \mathbb{P}^t \Rightarrow s = t$.

(iii) *If* $t_n \to t$ *for the Eucledian topology of* Θ *then* $\int f d\mathbb{P}^{t_n} \to \int f d\mathbb{P}^t$ *for every bounded mb.* f.

(iv) *There is a compact* $K \subset \Theta$, $\varepsilon \in (0, \frac{1}{2})$ *and tests* ϕ_n *such that* $\int (1-\phi_n) d\mathbb{P}^{\theta_0} < \varepsilon$ *and* $\forall \theta \in K^c : \int \phi_n d\mathbb{P}^{\theta} < \varepsilon$. *(testing condition)*

(v) *A process* ξ *with covariance kernel* $\mathbb{E}\xi(s)\xi(t) = \int \sqrt{d\mathbb{P}^s d\mathbb{P}^t}$ *admits at* $\theta = 0$ *a derivative in quadratic mean that has a non-singular covariance matrix.*

(vi) *The* $\mathbb{P}^{\theta_0}$*-singular part of* $\mathbb{P}^t$ *satisfies* $|t|^{-2} \left\| \mathbb{P}^t_{\perp} \right\| \to 0, t \to 0$.

(vii) *The prior measure* Π *is independent of* n *and has a Lebesgue density* π.

(viii) *Let the ball around the true parameter with radius* ε *be denoted as* $\mathcal{B}(\theta_0, \varepsilon)$. *There is some number* $a > 0$ *such that*

$$\lim_{\varepsilon \to 0} \frac{1}{\lambda(\mathcal{B}(\theta_0, \varepsilon))} \int \mathbb{1}_{\mathcal{B}(\theta_0, \varepsilon)} |\pi(t) - a| \, \lambda(dt) = 0$$

and

$$\liminf_{\varepsilon \to 0} \frac{\lambda(\mathcal{B}(\theta_0, \varepsilon) \cap \Theta)}{\lambda(\mathcal{B}(\theta_0, \varepsilon))} > 0.$$

Remark 2.18. Condition (viii) says that for parameter values close to the *true parameter* the prior is close to a constant a. This condition is similar to the *Kullback-Leibler condition* stated by Rousseau (2016), which ensures that enough positive mass is given to neighbourhoods of the true parameter. Le Cam (1986a) phrases it as follows: "*the prior is not too thin around the true parameter.*"

Under these conditions, Le Cam (1986a) takes estimates $T_n \circ \mathbf{X}_{(n)} \in \mathbb{R}^k$ and $V_n \circ \mathbf{X}_{(n)} \in \mathbb{R}^{k \times k}$ assuming V_n to be positive definite. Then he defines measures

$$\begin{aligned} HB &:= \int \mathbb{1}_{B \cap \Theta} \exp\left(-\frac{n}{2} \langle t - T_n, V_n(t - T_n) \rangle\right) \lambda^{(k)}(dt), \\ G &:= \|H\|^{-1} H, \end{aligned}$$

and obtains the following result.

Theorem 2.19
Let the conditions of lemma 2.17 be satisfied. Let $\Pi(\cdot|\mathbf{X}_{(n)})$ be the posterior distribution of θ. Then there exist T_n, V_n such that

$$\int \|\Pi(\cdot|x) - G\| \, \mathbb{P}^{\theta_0}(dx)$$

tends to zero as n diverges.

Le Cam and Lo Yang (2000) also present three examples where "*Bayes procedures behave miserably.*", among them an inconsistency result from Freedman (1963) where the parameter set consists of all probability measures on $\mathbb{N}$ and the posteriors live on nowhere dense closed sets almost surely. Le Cam and Lo Yang (2000) interpret this as follows: "*except for those priors Π that belong to a meagre set, there will be only a meagre set of values of θ where the posterior measures do not wander about aimlessly and indefinitely. That will happen even if the support of Π is the entire space Θ. It is due to the fact that, for nearly all $\theta \in \Theta$, the prior measure gives little weight to small neighbourhoods of θ.*" They add, however, that this problem may be solved with careful prior construction, for instance using Dirichlet priors.

The next problematic example involves a p-measure symmetric around zero that gets shifted by some unknown quantity t. If $\mathfrak{F}$ is the set of the symmetric measures, one puts a prior $\Pi \otimes \nu$ on $\mathfrak{F} \times \mathbb{R}^1$. There are priors for which "*the posterior distribution of the one-dimensional parameter t will oscillate indefinitely and never concentrate around the true value t_0 of t. This occurs by virtue of a peculiar phenomenon. To ensure consistency for all $\mathbb{P} \in \mathfrak{F}$ one must scatter Π around. There will then be many $\mathbb{P}$'s that have neighbours that are bumpy, symmetric around zero, but with several modes of about equal heights away from zero. The posterior distribution of t will make it oscillate between those modes.*"

The next negative result is presented by Kleijn and van der Vaart (2012) where a misspecified model is investigated. In this situation the posterior is formed in the same way as presented above or in section 1.2, but the observations are sampled from $\mathbb{P}_0$ instead of $\mathbb{P}^{\theta_0}$. Kleijn and van der Vaart (2012) derive asymptotic normality of the posterior and show consistency in the sense of: "*In the misspecified situation the posterior distribution of a parameter shrinks to the point within the model at minimum Kullback-Leibler divergence to the true distribution, a consistency property that it shares with the maximum likelihood estimator.*" However, they prove that the Bayesian credible sets in the misspecified experiment do not yield confidence sets for the minimum Kullback-Leibler point.

There is another contribution of Kleijn and van der Vaart (2006) for misspecified models in a nonparametric, i.e. infinite-dimensional, setting. Kleijn and van der Vaart (2006) write: "*Given a prior distribution and a random sample from a distribution P_0, which may not be in*

the support of the prior, we show that the posterior concentrates its mass near the points in the support of the prior that minimise the Kullback-Leibler divergence with respect to P_0. An entropy condition and a prior-mass condition determine the rate of convergence. The method is applied to several examples, with special interest for infinite-dimensional models. These include Gaussian mixtures, nonparametric regression and parametric models."

Another slightly negative result is presented by Franssen and van der Vaart (2021), who write "*The Pitman-Yor process generates discrete probability distributions ... and can be used as a prior distribution in a nonparametric Bayesian analysis. ... It was previously shown that the resulting posterior distribution is consistent if and only if the true distribution of the data is discrete. For a general discrete distribution, the posterior distribution, although consistent, may contain a bias which does not converge to zero at the $\sqrt{n}$-rate and invalidates posterior inference. We propose a bias correction that solves this problem. We also consider the effect of estimating the type parameter from the data, both by empirical Bayes and full Bayes methods. In a small simulation study we illustrate that without bias correction the coverage of credible sets can be arbitrarily low even for some discrete distributions.*"

Castillo and Nickl (2014) prove BvM theorems for a variety of nonparametric Bayes procedures, including Gaussian nonparametric regression and an iid sampling model. They "*deduce several applications where posterior-based inference coincides with efficient frequentist procedures, including Donsker- and Kolmogorov-Smirnov theorems for the random posterior cumulative distribution functions, and show that multiscale posterior credible bands for the regression or density function are optimal frequentist confidence bands.*" Donsker's theorem can be seen as a functional CLT for empirical processes and versions can be found in van der Vaart and Wellner (1996); Dudley (2014); Giné and Nickl (2015); Henze (2024). Regarding Gaussian processes, recall lemma 1.34 (compare Bogachev (2015)).

Definition 2.20 (Gaussian Process (Giné and Nickl, 2015))
A stochastic process $X(t), t \in T$, is called a Gaussian process if for all $n \in \mathbb{N}, a_j \in \mathbb{R}$ and $t_j \in T$, the random variable $\sum_{j=1}^n a_j X(t_j)$ is normal, or equivalently, if all the finite-dimensional marginals of X are multivariate normal. X is a centred Gaussian process if all these random variables are normal with mean zero.

Example 2.21. A Gaussian process used in Castillo and Nickl (2013) and Castillo and Nickl (2014) is the *white noise process*. Given a separable Hilbert space $(\mathscr{H}, \langle\cdot,\cdot\rangle_{\mathscr{H}})$, the process has the characteristic that $\mathbb{E}X(h) = 0$ and the covariance is given by $\mathbb{E}X(h)X(g) = \langle h, g\rangle_{\mathscr{H}}$.

Theorem 2.22 (Donsker Theorem (Henze, 2024))
Let $Z_1, Z_2, \ldots$ be iid random variables satisfying $\mathbb{E}Z_1^2 < \infty$, $\mathbb{E}Z_1 = 0$, and $0 < \sigma^2 := \mathbb{V}Z_1 < \infty$. Define $S_0 := 0$, $S_n := \sum_{j=1}^n Z_j, n \in \mathbb{N}$, and further for $n \in \mathbb{N}$:

$$X_n(t) := \frac{1}{\sigma\sqrt{n}} S_{\lfloor nt \rfloor} + (nt - \lfloor nt \rfloor)\frac{Z_{\lfloor nt \rfloor}}{\sigma\sqrt{n}}, \quad t \in [0,1].$$

Then $X_n \rightsquigarrow W$, for W as in Definition 1.35.

For their framework, Castillo and Nickl (2014) define an *S-regular, $S \geq 0$, wavelet basis* (ψ_{lk}), Hölder-type spaces, a bounded Lipschitz-metric (compare Lévy-Prohorov metric) and multiscale

spaces. The latter are defined for monotone increasing weighting sequences w as

$$\mathcal{M}(w) := \left\{ x = (x_{lk}) : \|x\|_{\mathcal{M}(w)} := \sup_l \frac{\max_k |x_{lk}|}{w_l} < \infty \right\}$$
$$\mathcal{M}_0(w) := \left\{ x \in \mathcal{M}(w) : \|x\|_{\mathcal{M}(w)} := \lim_{l\to\infty} \max_k \frac{|x_{lk}|}{w_l} = 0 \right\}$$

and the metric they use is defined for p-measures on metric spaces $(\mathscr{S}, d)$ as

$$\beta_{\mathscr{S}}(\mu,\nu) := \sup_{F:\|F\|_{BL}\le 1} \left| \int_{\mathscr{S}} F(d\mu - d\nu) \right|$$
$$\|F\|_{BL} := \sup_{s\in\mathscr{S}} |F(s)| + \sup_{s,t\in\mathscr{S}:s\neq t} \frac{|F(s)-F(t)|}{d(s,t)}.$$

Definition 2.23 (S-regular Basis (Giné and Nickl, 2015))
*Let $S \in \mathbb{N}$. By a S-*regular basis $\{\psi_{lk} : l \in L \subset \mathbb{Z}, k \in Z_l \subset \mathbb{Z}\}$ *of L^2 and characteristic sequence a_l, we shall mean any of the following:*

(i) *$\psi_{lk} := e_l$ is S-times differentiable with all derivatives in L^2,* $\operatorname{card} Z_l = 1$, *$a_l =:= \max(2,|l|)$, and $\{e_l : l \in L\}$ is an orthonormal basis of L^2.*

(ii) *ψ_{lk} is S-times differentiable with all derivatives in L^2, $a_l := \operatorname{card} Z_l = 2^l$, and $\{\psi_{lk} : l \in L, k \in Z_l\}$ forms an orthonormal basis of L^2.*

The models considered are an iid sampling model where $X_1,\dots,X_n$ are iid from law P with density f on $[0,1]$. Here they estimate $\langle f, \psi_{lk}\rangle$ with $\langle P_n, \psi_{lk}\rangle = \frac{1}{n}\sum_{j=1}^n \psi_{lk}(X_j)$ and find that for fixed k and l

$$\sqrt{n}(P_n - P)(\psi_{lk}) \rightsquigarrow \mathbb{G}_P(\psi_{lk}) \sim \mathcal{N}(0, \mathbb{V}_P\psi_{lk}(X_1)).$$

Castillo and Nickl (2014) call this a P-white bridge process indexed by the Hilbert space $L^2(P) := \{f : [0,1] \to \mathbb{R}^1 : \int_0^1 f^2 dP < \infty\}$ with covariance $\mathbb{E}\mathbb{G}_P(g)\mathbb{G}_P(h) = \int_0^1 (g-Pg)(h-Ph)dP$. (Memo from Def. 1.5: $Pg = \int_0^1 g dP$). For this problem they define an *admissible* w such that it satisfies $l^{-1/2}w_l \uparrow \infty$.

Theorem 2.24
Let w be admissible and let P have a density f in $\mathscr{C}^\gamma[0,1], \gamma \ge 0$. Take j_n such that

$$\sqrt{n}2^{-j_n(\gamma+\frac{1}{2})}w_{j_n}^{-1} = o(1), \quad \frac{2^{j_n}j_n}{n} = \mathcal{O}(1).$$

Then, as $n\to\infty$, in $\mathcal{M}_0(w)$:

$$\sqrt{n}(P(j_n) - P) \rightsquigarrow \mathbb{G}_P.$$

Their second model is a Gaussian white noise model that is also investigated in Castillo and Nickl (2013):

$$d\mathbf{X}_{(n)}(t) = f(t)dt + n^{-1/2}dW(t) \Leftrightarrow \mathbb{X}_{(n)} = f + n^{-1/2}\mathbb{W}. \tag{2.9}$$

They write that for admissible w this is in $\mathcal{M}_0(w)$. Denote $P_n^f := \mathfrak{L}(\mathbb{X}_{(n)})$. Considering $\mathbb{X}_{(n)}$ was generated by the true $P_n^{f_0}$ they define a weak BvM phenomenon in $\mathcal{M}_0(w)$ such that the posterior under the mapping $\tau : f \mapsto \sqrt{n}(f - T_n)$ converges in the bounded Lipschitz metric to a Gaussian

$$\beta_{\mathcal{M}_0(w)}(\Pi(\cdot|\mathbb{X}_{(n)}) \circ \tau^{-1}, \mathbb{W}) \xrightarrow{P^{f_0}} 0.$$

They present various BvM theorems and also construct credible regions, which they prove to hold asymptotically for P^{f_0} with the same level.

Castillo and Rousseau (2013) develop BvM theorems for semiparametric models, i.e. in the setting of $\mathscr{E}_{1,n}$ in (2.5) the interest lies on a functional $\psi(\theta)$, $\psi : \Theta \to \mathbb{R}^1$ of the parameter. The bounded Lipschitz distance between the image of the posterior under the mapping

$$\tau : x \mapsto \sqrt{n}(x - T_n), \quad T_n := \theta_0 + \Delta_{n,\theta_0}, \tag{2.10}$$

where Δ_{n,θ_0} is defined in (2.8), and a centred Gaussian is then shown to converge to zero in $\mathbb{P}^{\theta_0}$-probability. Versions of the theorem for the Gaussian white noise model, and a density model are then derived from the main result. A generic LAN framework and semiparametric BvM are also presented in Castillo (2024) as follows.

The framework involves a *LAN expansion* of the log-likelihood for a parameter in the Hilbert space $(\mathscr{H}, \langle\cdot,\cdot\rangle_{\mathscr{H}})$:

$$\ell_n(\theta) - \ell_n(\theta_0) = -\frac{n}{2}\|\theta - \theta_0\|^2_{\mathscr{H}} + \sqrt{n}\, W_n(\theta - \theta_0) + R_n(\theta, \theta_0),$$

where $W_n : h \mapsto W_n(h)$ is linear $\mathbb{P}^{\theta_0}$-a.s. and $W_n \rightsquigarrow \mathcal{N}\left(0, \|h\|^2_{\mathscr{H}}\right)$ as n diverges.

Secondly a *functional expansion*: Suppose the functional ψ can be expanded around $\vartheta \in \mathscr{H}$ as

$$\psi(\theta) - \psi(\theta_0) = \langle \vartheta, \theta - \theta_0 \rangle_{\mathscr{H}} + r(\theta, \theta_0),$$

and define for a small, fixed $t \in \mathbb{R}^1$ a path through η, such that $\theta_t \in \Theta$ as

$$\theta_t := \theta - \frac{t\vartheta}{\sqrt{n}}.$$

Control the remainder terms: Assume the existence of measurable sets A_n satsifying $\Pi(A_n|X) = 1 + o_{\mathbb{P}^{\theta_0}}(1)$, such that $\theta - \theta_0 \in \mathscr{H}$ for all $\theta \in A_n$, n sufficiently large, and for fixed $t \in \mathbb{R}^1$:

$$\sup_{\theta \in A_n} \left| t\sqrt{n}\, r(\theta, \theta_0) + R_n(\theta, \theta_0) \right| = o_{\mathbb{P}^{\theta_0}}(1).$$

For ϑ and W_n as above, further define

$$\hat{\psi} := \psi(\theta_0) + n^{-1/2} W_n(\theta_0), \quad V_0 := \|\vartheta\|^2_{\mathscr{H}}.$$

Theorem 2.25 (Semiparametric BvM (Castillo, 2024))
Let Π be a prior distribution on θ and assume the LAN framework above. If for any $t \in \mathbb{R}^1$

$$\frac{\int_{A_n} \exp\left(\ell_n(\theta_t)\right) \Pi(d\theta)}{\int \exp\left(\ell_n(\theta_t)\right) \Pi(d\theta)} = 1 + o_{\mathbb{P}^{\theta_0}}(1), \tag{2.11}$$

then the posterior distribution of $\sqrt{n}(\psi(\theta) - \hat{\psi})$ converges weakly in $\mathbb{P}^{\theta_0}$-probability to a centred Gaussian distribution with variance V_0.

Rousseau (2016) offers three sufficient conditions for the existence of a semiparametric BvM, which are

1. concentration of the posterior,
2. local asymptotic normality of the likelihood, and
3. smoothness of the functional.

Giné and Nickl (2015) present an approach of deriving BvM theorems that we will adapt in section 4.2, namely to project the posterior onto a finite-dimensional subspace (at first). In chapter 7.3 Giné and Nickl (2015) put a prior Borel probability measure on L^2 for their Gaussian white noise regression model as in Castillo and Nickl (2014) from (2.9). They take V to be a finite-dimensional projection subspace of L^2 spanned by the ψ_{lk} of Definition 2.23, and the projection of $f = (f_{lk})$ onto V is denoted as π_V. The same scale-and-shift transformation that we have seen before is then used to study the posterior push-forward through this transformation $\Pi(\cdot|\mathbf{X}_{(n)}) \circ T_{z,V}^{-1}$:

$$T_z := T_{z,V} : f \mapsto \sqrt{n}\, \pi_V(f - z),$$

and they write that this "*carries a natural Lebesgue product measure on it.*"

Lemma 2.26 (Condition 7.3.17 (Giné and Nickl, 2015))
Suppose that Π *is a product measure on the span of the* $\{psi_{lk}\}$ *and that* $\Pi \circ \pi_V^{-1}$ *has a Lebesgue density* $d\Pi_V$ *in a neighbourhood of* $\pi_V(f_0)$*. Suppose also that for every* $\varepsilon > 0$ *there exists a fixed* L^2*-norm ball* $\mathcal{B} = \mathcal{B}(\tilde{f}, \varepsilon)$ *in* V *such that, for* n *large enough,*

$$\mathbb{E}_{\mathbb{P}^{f_0}} \left[\Pi(\cdot|\mathbf{X}_{(n)}) \circ T_{f_0}^{-1} \right] (\mathcal{B}^c) < \delta.$$

Theorem 2.27 (Giné and Nickl (2015))
Consider $\mathbf{X}_{(n)} \sim \mathbb{P}^{f_0}$ *in the context of (2.9) under a fixed* $f_0 \in L^2$*. Suppose Condition 7.3.17 holds. Then*

$$d_{TV}\left(\Pi(\cdot|\mathbf{X}_{(n)}) \circ T_{\mathbb{X}}^{-1}, \mathcal{N}(0, \mathbf{I})\right) = o_{\mathbb{P}^{f_0}}(1), \quad n \to \infty.$$

They then build on this theorem to derive BvM results in negative-order Sobolev spaces and multiscale spaces, vide supra (v.s.). For nonparametric settings, Giné and Nickl (2015) warn: "*Likelihood-based procedures ... This typically relies on the assumption that the true parameter* θ_0 *is interior to* Θ *so that by consistency* $\hat{\theta}_n$ *will eventually also be. In the infinite-dimensional setting, even if we can define an appropriate notion of derivative, this approach is usually not viable because* $\hat{p}_n$ *is, as we shall see, never an interior point in the parameter space, even when* p_0 *is.*"

BvM-type results for regression models are obtained by Bontemps (2011) or Castillo, Schmidt-Hieber, and van der Vaart (2015), and Ghosal and van der Vaart (2017), for instance. Ning, Jeong, and Ghosal (2018) write: "*We quantify the uncertainty for the regression coefficients with frequentist validity through a BvM type theorem. The result leads to selection consistency for the Bayesian method.*", and later: "*To establish selection consistency, Castillo et al. (2015) devised a key technique through a distributional approximation for the posterior distribution. As in a BvM theorem, the posterior distribution of the regression parameter is approximated by a relatively simpler distribution, but unlike in a traditional BvM theorem for increasing dimensional parameters (Bontemps, 2011), ..., the approximating distribution is a mixture of multivariate normal instead of a single one.*"

The book *Fundamentals of nonparametric Bayesian inference* by Ghosal and van der Vaart (2017) features construction of priors and consistency of NONPARAMETRIC BINARY REGRESSION. Contraction rates are derived for a Dirichlet Process prior, and adaptation and model selection is discussed for the logit link function as "*it allows a direct relationship between the Hellinger distance, Kullback-Leibler divergence and* L^2*-norm.*" Ghosal and van der Vaart (2017) also give computational advice in the form of a MCMC algorithm for Gaussian Process priors.

To the best of our knowledge, there are no Laplace-Bernstein-von Mises-type theorems for group testing regression problems yet.

3 Derivation of Likelihood and Posterior Distribution for Group Testing Regression

We are interested in the distribution of a binary random variable Y given the covariate X (*random design*), and we assume the pairs $((X_{ij}, Y_{ij}))_{1\leq j\leq J,\, 1\leq i\leq n_j}$ to be ciid given the σ-field generated by the function p. *Fixed design*, arises when the regressor $X = x$ is known or chosen beforehand. The goal is to estimate the conditional probability

$$\mathbb{P}^p[Y = 1|X] = p(X) = H \circ f(X), \tag{3.1}$$

where $p : \mathbb{R} \supset \mathscr{X} \to [0, 1]$ defines a functional dependence of the Bernoulli-parameter given the regressor, $f : \mathbb{R} \supset \mathscr{X} \to \mathbb{R}$ is called the *regression function*, and $H : \mathbb{R} \to [0, 1]$ is a fixed cdf, called the link function (compare section 2.1 and Definition 2.1). Note that the dependence on the parameter p is often indicated by a subscript, i.e. $\mathbb{P}_p$, but in the Bayesian context, p is assumed to be a random element itself, thus it is denoted by a superscript in accordance with Theorem 1.44. And the probability measure in (3.1) will be interpreted as a regular version of a conditional distribution as introduced in Definition 1.47 such that we can identify a probability kernel $\kappa_p(X, \cdot) := \mathbb{P}^p_Y[\cdot|X] \ll \zeta$ if ζ is taken to be the counting measure on $\mathfrak{P}(\{0, 1\})$, i.e. $\delta_0 + \delta_1$. The parameter function p is also defined by the relationship

$$\mathbb{P}^p[Y = y|X] = p(X)^y(1 - p(X))^{1-y}, \quad y \in \{0, 1\}, \tag{3.2}$$

For economic reasons explained in section 2.1, the N individuals will be split into J groups consisting of n_j individuals in each group, such that $N = \sum_{j=1}^{J} n_j$. After grouping one may however only observe

$$Y_j^* := \max_{i\in\{1,\dots,n_j\}} Y_{ij} \tag{3.3}$$

alongside the covariates. Moreover, it will be assumed that the $X_{ij} \sim G$ are ciid (compare Lemma 1.42) on the induced probability space $(\mathscr{X}, \mathfrak{X}, G)$ with the marginal distribution $G \ll \mu$, which has a μ-density $g := \frac{dG}{d\mu}$ for some σ-finite measure μ. To simplify notation, define:

- $\mathbf{Y}^* := (Y_1^*, \dots, Y_J^*)^\top$,
- $\mathbf{X} := (X_{11}, \dots, X_{n_1 1}, \dots, X_{1J}, \dots, X_{n_J J})^\top$,
- $\mathbf{X}_j := (X_{1j}, \dots, X_{n_j j})^\top$, and let
- $g_N : \mathscr{X}^N \to [0, \infty),\ \mathbf{x} \mapsto \prod_{j=1}^{J} \prod_{i=1}^{n_j} g(x_{ij})$,
- $G_N : \mathfrak{X}^{(N)} \to [0, 1],\ A \mapsto \int \mathbb{1}_A g_N d\mu^{(N)}$, as well as
- $g_{n_j} : \mathscr{X}^{n_j} \to [0, \infty),\ \mathbf{x}_j \mapsto \prod_{i=1}^{n_j} g(x_{ij})$,
- $G_{n_j} : \mathfrak{X}^{(n_j)} \to [0, 1],\ A \mapsto \int \mathbb{1}_A g_{n_j} d\mu^{(n_j)}$.

To derive the likelihood, similarly to the approach presented by Delaigle and Meister (2011), it is useful to exploit the properties of the Bernoulli distribution and the fact that if the maximum of binary values equals zero, so do all of them, thus linking the conditional distribution of the Y_j^* to that of the Y_{ij}. Since the (X_{ij}, Y_{ij}) are ciid, it follows that the Y_{ij} are ciid and thus the $Y_{ij}|\sigma(X_{1j}, \dots, X_{n_j j})$ are stochastically independent for each $j \in 1, \dots, J$ $(*)$; moreover the pairwise independences $\forall j \in \{1, \dots, J\} : \forall i \in \{1, \dots, n_j\} : \forall k \in \{1, \dots, n_j\} \setminus \{i\} : Y_{ij} \perp\!\!\!\perp X_{kj}$

follow directly. Thus

$$\begin{aligned}
\mathbb{E}^p[Y_j^*|X_{1j},\dots,X_{n_jj}] &= \mathbb{P}^p[Y_j^*=1|X_{1j},\dots,X_{n_jj}] \\
&= 1-\mathbb{P}^p[Y_j^*=0|X_{1j},\dots,X_{n_jj}] \\
&= 1-\mathbb{P}^p\left[\cap_{i=1}^{n_j}\{Y_{ij}=0\}|X_{1j},\dots,X_{n_jj}\right] \\
&= 1-\prod_{i=1}^{n_j}\mathbb{P}^p[Y_{ij}=0|X_{1j},\dots,X_{n_jj}] && (*) \\
&= 1-\prod_{i=1}^{n_j}\mathbb{P}^p[Y_{ij}=0|X_{ij}] && Y_{ij}\perp\!\!\!\perp X_{kj}, k\neq i \\
&= 1-\prod_{i=1}^{n_j}(1-p(X_{ij})) \\
&=: p_j^* = p_j^*(X_{1j},\dots,X_{n_jj}).
\end{aligned}$$

Hence $Y_j^*|\sigma(X_{1j},\dots,X_{n_jj},p)\sim \mathrm{Ber}(p_j^*)$, and

$$\mathbb{P}^p\left[\cap_{j=1}^{J}\{Y_j^*=y_j^*\}|\sigma(X_{11},\dots,X_{n_11},\dots,X_{n_JJ})\right] \tag{3.4}$$

$$=\prod_{j=1}^{J}\left(1-\prod_{i=1}^{n_j}(1-p(X_{ij}))\right)^{y_j^*}\left(\prod_{i=1}^{n_j}(1-p(X_{ij}))\right)^{1-y_j^*} \tag{3.5}$$

$$=\exp\left(\sum_{j=1}^{J}y_j^*\ln\left[1-\prod_{i=1}^{n_j}(1-p(X_{ij}))\right]+(1-y_j^*)\sum_{i=1}^{n_j}\ln\left[1-p(X_{ij})\right]\right). \tag{3.6}$$

From this the likelihood of an observation given p, in this case the conditional $\zeta^{(J)}\otimes\mu^{(N)}$-density of $(\mathbf{Y}^*,\mathbf{X})$, is obtained by multiplying with the marginal density of $\mathbf{X}$:

$$\mathscr{L}(\mathbf{Y}^*,\mathbf{X}|p):=\exp\left(\sum_{j=1}^{J}y_j^*\ln\left[1-\prod_{i=1}^{n_j}(1-p(X_{ij}))\right]+(1-y_j^*)\sum_{i=1}^{n_j}\ln\left[1-p(X_{ij})\right]\right)g_N(\mathbf{X}) \tag{3.7}$$

Let Π denote the prior distribution of p, i.e. a probability distribution on the measurable space $(\mathscr{T},\mathfrak{T})$. Since the model is dominated by the measure $\zeta^{(J)}\otimes\mu^{(N)}$ the posterior probability can be calculated by means of the Bayes formula, as presented in Ghosal and van der Vaart (2017), as:

$$\Pi_N\left[\mathcal{T}|\mathbf{Y}^*,\mathbf{X}\right]:=\mathbb{P}[p\in\mathcal{T}|\sigma(\mathbf{Y}^*,\mathbf{X})] \tag{3.8}$$

$$=\frac{\int\mathbb{1}_{\mathcal{T}}(p)\left(\prod_{j=1}^{J}\left(1-\prod_{i=1}^{n_j}\left[1-p(x_{ij})\right]\right)^{y_j^*}\left(\prod_{i=1}^{n_j}\left[1-p(x_{ij})\right]\right)^{1-y_j^*}\right)\prod_{j=1}^{J}\prod_{i=1}^{n_j}g(x_{ij})\Pi(dp)}{\int\left(\prod_{j=1}^{J}\left(1-\prod_{i=1}^{n_j}\left[1-p(x_{ij})\right]\right)^{y_j^*}\left(\prod_{i=1}^{n_j}\left[1-p(x_{ij})\right]\right)^{1-y_j^*}\right)\prod_{j=1}^{J}\prod_{i=1}^{n_j}g(x_{ij})\Pi(dp)} \tag{3.9}$$

for a set $\mathcal{T}\in\mathfrak{T}$, if p has the prior measure Π on $(\mathscr{T},\mathfrak{T})$.

Remark 3.1. In later sections we will confine ourselves to the case of subgroups of equal size, i.e. $\forall j\in\{1,\dots,J\}: n_j=n$, hence $N=nJ$, resulting in the $\mathbf{X}_j$ to be ciid G_n, and the pairs $(Y_j^*,\mathbf{X}_j)_{j=1,\dots,J}$ to be ciid, as well, a requirement for theorem 4.14. In practical applications μ will often be the Lebesgue measure λ which is assumed from this point on. Moreover, assume $\mathscr{X}$ to be a compact space. As another restriction, p will be parameterised by some multivariate ϑ in a compact subset of $\mathbb{R}^K$.

4 Local Asymptotic Normality and Bernstein-von Mises Theorems

4.1 Bernstein-von Mises Theorem for Finite-Dimensional Parameter Spaces in the General Case

Assume $(\Omega, \mathfrak{A}, \mathbb{P})$ to be a probability space. The focus will lie on the induced probability space of $X_1, \dots, X_n$, which are assumed to be conditionally independent and identically distributed random variables: $X_1 : \Omega \to \mathscr{X}$ given the σ-field generated by ϑ. Thus the experiments of interest are of the form

$$\mathscr{E}_n := \left(\mathscr{X}^n, \mathfrak{X}^{(n)} := \otimes_{j=1}^n \mathfrak{X}, \left\{ \mathbb{P}^\vartheta_{\mathbf{X}_{(n)}} := \otimes_{j=1}^n \mathbb{P}^\vartheta_{X_j} = \left(\mathbb{P}^\vartheta_{X_1}\right)^{(n)} : \vartheta \in \Theta \right\} \right), \tag{4.1}$$

and for each ϑ in the parameter space Θ, which is assumed to be a Borel subset of a Polish space, $X_1|\vartheta$ has the conditional image measure $\mathbb{P}^\vartheta_{X_1}$, which is assumed to be dominated by some σ-finite measure μ. Because we study an iid model, the vector $\mathbf{X}_{(n)} = \mathbf{X}_{(n)}(\omega) := (X_1(\omega), \dots, X_n(\omega))^\top$ follows the corresponding product measure $\mathbb{P}^\vartheta_{\mathbf{X}_{(n)}} \ll \mu^{(n)}$. Subsequently, μ is assumed to be the Lebesgue-Borel measure λ, and $\mathscr{X}$ is a subset of $\mathbb{R}$, and $\mathfrak{X} = \mathfrak{B}(\mathscr{X})$.

In the Bayesian context, the parameter ϑ is itself assumed to be a random variable, more specifically, the measurable space $(\Theta, \mathfrak{T})$, will be equipped with the prior measure Q dominated by some σ-finite measure ν. This prior can be seen as the image measure of ϑ, and as Hastie et al. (2009) put it, *reflects our knowledge about the parameter before we see the data.* In this setting we can identify a probability kernel $\kappa : \Theta \times \mathfrak{X} \to [0,1]$ such that $\mathbb{P}(X_1 \in \cdot|\vartheta) = \kappa(\vartheta, \cdot)$. We will from now on assume $\Theta \subset \mathbb{R}^k$ and $\mathfrak{T} = \mathfrak{B}(\Theta)$.

As phrased by Giné and Nickl (2015), we take the *frequentist Bayes approach*, considering ϑ_0 to be the "*true*" but unknown underlying parameter of interest (compare section 5). Let the Radon-Nikodym derivative of the likelihood be denoted by

$$f_n(\cdot|\vartheta) := \frac{d\mathbb{P}^\vartheta_{\mathbf{X}_{(n)}}}{d\mu^{(n)}}$$

– remember that $f_n(\mathbf{X}_{(n)}|\vartheta) = \prod_{j=1}^n f(X_j|\vartheta)$, where $f(\cdot|\vartheta) := \frac{d\mathbb{P}^\vartheta_{X_j}}{d\mu}$. Then, the posterior density with respect to to the likelihood f_n and the prior density $q := \frac{dQ}{d\nu}$ can be calculated applying the Bayes formula (1.5) and (1.6):

$$\begin{aligned} g(\vartheta|\mathbf{X}_{(n)}) &:= \frac{q(\vartheta)\prod_{j=1}^n f(X_j|\vartheta)}{\int q(\vartheta')\prod_{j=1}^n f(X_j|\vartheta')\nu(d\vartheta')} \\ &= \frac{q(\vartheta)\prod_{j=1}^n \frac{f(X_j|\vartheta)}{f(X_j|\vartheta_n(\mathbf{X}_{(n)}))}}{\int q(\vartheta')\prod_{j=1}^n \frac{f(X_j|\vartheta')}{f(X_j|\vartheta_n(\mathbf{X}_{(n)}))}\nu(d\vartheta')} \\ &=: \frac{q(\vartheta)\prod_{j=1}^n \frac{f(X_j|\vartheta)}{f(X_j|\vartheta_n)}}{C(\mathbf{X}_{(n)})}. \end{aligned}$$

For a more concise notation, we denote $G = G_n := Q(\cdot|\mathbf{X}_{(n)}) = \mathfrak{L}(\vartheta|\mathbf{X}_{(n)})$ and define it via its ν-probability density g. For this approach, ϑ_n is taken to be the MLE

$$\vartheta_n(\mathbf{X}_{(n)}) := \arg\sup_{\vartheta\in\Theta} f_n(\mathbf{X}_{(n)}|\vartheta)$$

and the posterior is expanded in a way such that the likelihood ratios are regarded. We will also require the MLE to be strongly consistent in the following sense.

Definition 4.1 (Strong Consistency)
We call ϑ_n strongly consistent iff for Q-a.a. $\theta \in \Theta$ one has

$$\vartheta_n \overset{\mathbb{P}^\theta \text{ a.s.}}{\to} \theta \;\Leftrightarrow\; \mathbb{P}\left[\left\{\omega \in \Omega : \lim_{n\to\infty} \vartheta_n \circ \mathbf{X}_{(n)}(\omega) = \theta\right\} | \theta\right] = 1. \tag{4.2}$$

In particular, ϑ_n is strongly consistent for ϑ_0 if the convergence holds for $\theta = \vartheta_0$.

Remark 4.2. For technical simplification we will now assume (4.2) to hold for each $\vartheta \in \Theta$.

Moreover, take the sphere

$$\Theta_0 := \mathcal{B}(\vartheta_0, \rho) := \{\vartheta \in \mathbb{R}^k : \|\vartheta - \vartheta_0\| \leq \rho\} \tag{4.3}$$

to be a subset of the support of q, i.e. $\operatorname{supp}(q) := \{\vartheta \in \Theta : q(\vartheta) > 0\} =: \Theta_1 \supset \Theta_0$, and henceforth that of g. The inclusion of ϑ_0 in this set is important for consistency in accordance with theorem 1.70.

Remark 4.3. This also means that ϑ_0 is an inner point of Θ (compare van der Vaart (1998), Chapter 7).

It will further be assumed that

$$\Theta_1 := \mathcal{B}(\vartheta_0, \rho_1),\ \rho_1 > \rho. \tag{4.4}$$

Now, take H to be a Gaussian measure with $h := \frac{dH}{d\nu}$ being the corresponding ν-density, more specifically it will be assumed that $H = H_n(\mathbf{X}_{(n)}, \vartheta_0) = \mathcal{N}_k(\vartheta_n, n^{-1}\mathcal{I}(\vartheta_0)^{-1}))$, and we can identify a probability kernel $\pi : \mathscr{X}^n \times \mathfrak{T} \to [0,1]$ with it, such that $\pi(\mathbf{X}_{(n)}, \cdot) = H_n(\mathbf{X}_{(n)}, \vartheta_0)(\cdot)$. It shall be remarked that the simplified notation $h = h_{\vartheta_n(\mathbf{X}_{(n)}),\vartheta_0}$ will be used when convenient to increase readability. In the definition of $H_n(\mathbf{X}_{(n)}, \vartheta_0)$, $\mathcal{I}(\vartheta_0)$ is the Fisher information (matrix) of ϑ with respect to the likelihood $f_n(\mathbf{X}_{(n)}|\vartheta)$ at ϑ_0 and it is clear that this matrix must be regular and symmetric from lemma 1.34.

Definition 4.4 (Modified Posterior Density)
As the support of g and h may not necessarily coincide, a technical modification is used to compare these densities asymptotically:

$$\tilde{H} = \tilde{H}_n(\mathbf{X}_{(n)}, \vartheta_0) := H(\Theta_1^c) = \int_{\Theta_1^c} h_{\vartheta_n(\mathbf{X}_{(n)}),\vartheta_0}(\vartheta)\nu(d\vartheta) \tag{4.5}$$

$$\tilde{g} = \tilde{g}(\vartheta|\mathbf{X}_{(n)}) := (1-\tilde{H})g\mathbb{1}_{\Theta_1} + h_{\vartheta_n(\mathbf{X}_{(n)}),\vartheta_0}\mathbb{1}_{\Theta_1^c} \tag{4.6}$$

$$= (1 - \tilde{H}_n(\mathbf{X}_{(n)}, \vartheta_0))g(\vartheta|\mathbf{X}_{(n)})\mathbb{1}_{\Theta_1}(\vartheta) + h_{\vartheta_n(\mathbf{X}_{(n)}),\vartheta_0}(\vartheta)\mathbb{1}_{\Theta_1^c}(\vartheta) \tag{4.7}$$

Here $\Theta_1^c = \{\vartheta \in \mathbb{R}^k : \|\vartheta - \vartheta_0\| > \rho_1\}$ for some $\rho_1 > \rho$ since $\Theta_1^c \subset \Theta_0^c$.

Lemma 4.5
The function $\tilde{g}$ is a ν-(probability) density. $\tilde{H}$, as defined above, converges to zero as $n \to \infty$ $\mathbb{P}^{\vartheta_0}$-almost surely.

Proof. The second part of lemma 4.5 will be proven first. Consider the subsitution

$$\vartheta \mapsto n^{1/2}(\vartheta - \vartheta_n(\mathbf{X}_{(n)})) = \theta \tag{4.8}$$

and henceforth the transformed region

$$\tilde{\Theta}_1^c = \tilde{\Theta}_1^c(\mathbf{X}_{(n)}(\omega)) := \left\{\theta \in \mathbb{R}^k : \left\|n^{-1/2}\theta + \vartheta_n \circ \mathbf{X}_{(n)}(\omega) - \vartheta_0\right\| > \rho_1\right\}, \tag{4.9}$$

which will be empty asymptotically as $n \to \infty$, $\mathbb{P}^{\vartheta_0}$-almost surely, i.e.

$$\exists A \in \mathfrak{A} : \mathbb{P}^{\vartheta_0}(A) = 1 \ \wedge \ \forall \omega \in A : \lim_{n\to\infty} \mathbb{1}_{\tilde{\Theta}_1^c(\mathbf{X}_{(n)}(\omega))} = 0.$$

The following argument holds for all $\omega \in A$, i.e. almost surely in the above sense.

$$\begin{aligned}
&\int \mathbb{1}_{\Theta_1^c} h_{\vartheta_n \circ \mathbf{X}_{(n)}(\omega),\vartheta_0} \, d\nu \\
&= \int \mathbb{1}_{\Theta_1^c}(\vartheta) \ (2\pi)^{-k/2} n^{k/2} \sqrt{\det(\mathcal{I}(\vartheta_0))} \exp\left(-\frac{n}{2}(\vartheta - \vartheta_n)^\top \mathcal{I}(\vartheta_0)(\vartheta - \vartheta_n)\right) \nu(d\vartheta) \\
&= \int \mathbb{1}_{\tilde{\Theta}_1^c(\mathbf{X}_{(n)}(\omega))}(\theta) \ (2\pi)^{-k/2} \sqrt{\det(\mathcal{I}(\vartheta_0))} \exp\left(-\frac{1}{2}\theta^\top \mathcal{I}(\vartheta_0)\theta\right) \nu(d\theta)
\end{aligned}$$

A dominating term is given by the ν-density of the multivariate normal distribution

$$\mathcal{N}_k\left(0, \mathcal{I}(\vartheta_0)^{-1}\right),$$

which integrates to one. Since the integrand converges to zero, it follows from dominated convergence that $\lim_{n\to\infty} \tilde{H} = 0$ on A.
For the first part, consider that $\tilde{H} \in [0, 1)$, because $\tilde{\Theta}_1^c \subsetneq \mathbb{R}^k$ and h is a probability density, then exploit the linearity of measure integrals, remembering that $\operatorname{supp}(g) = \Theta_1$:

$$\begin{aligned}
\int \tilde{g} d\nu &= (1 - H(\Theta_1^c)) \int g \mathbb{1}_{\Theta_1} d\nu + \int h_{\vartheta_n(\mathbf{X}_{(n)}),\vartheta_0} \mathbb{1}_{\Theta_1^c} d\nu \\
&= (1 - H(\Theta_1^c)) + H(\Theta_1^c) = 1.
\end{aligned}$$

■

The main interest of this section is the asymptotic connection between the posterior distribution G as defined ($\nu - a.e.$) by its density g ($\tilde{G}$ defined by $\tilde{g}$, respectively) and the distribution H. For this purpose, their total variation distance will be examined. Applying the triangle inequality one can see that:

$$d_{\mathrm{TV}}(G, H) \leq d_{\mathrm{TV}}\left(G, \tilde{G}\right) + d_{\mathrm{TV}}\left(\tilde{G}, H\right).$$

Using the Scheffé lemma 1.18 and lemma 4.5:

$$\begin{aligned}
2d_{\mathrm{TV}}\left(G, \tilde{G}\right) &= \int |\tilde{g} - g| d\nu = \int_{\Theta_1} \left|g - \tilde{H}g - g\right| d\nu + \int_{\Theta_1^c} |h - 0| \, d\nu \\
&= \tilde{H} \int_{\Theta_1} g d\nu + \int_{\Theta_1^c} h d\nu = 2\tilde{H} \xrightarrow[n\to\infty]{} 0 \quad \mathbb{P}^{\vartheta_0} - a.s.
\end{aligned}$$

and therefore the distance between the modified and unmodified posterior vanishes asymptotically a.s., as intended, and we can focus on the asymptotic comparison between $\tilde{G}$ and H. As an application of the Pinsker bound (theorem 1.20) and corollary 1.21, the total variation distance of two measures converges to zero if the respective Kullback-Leibler divergence converges to zero. Thus, the remainder of this section studies the asymptotic behaviour of the following term, which will be split into four summands:

$$\begin{aligned}
\mathcal{K}\left(H, \tilde{G}\right) &= \int \ln\left(\frac{h}{\tilde{g}}\right) h d\nu \\
&= \int \ln(h) \, h d\nu - \int \ln(\tilde{g}) \, h d\nu \\
&= \int \ln(h) \, h d\nu - \int \mathbb{1}_{\Theta_1} \ln(\tilde{g}) \, h d\nu - \int \mathbb{1}_{\Theta_1^c} \ln(\tilde{g}) \, h d\nu \\
&= \int \ln(h) \, h d\nu - \int_{\Theta_1} \ln\left((1 - \tilde{H})g\right) h d\nu - \int_{\Theta_1^c} \ln(h) \, h d\nu \\
&= \underbrace{\int \ln(h) h d\nu}_{=:(\mathtt{iv})} - \ln\left(1 - \tilde{H}\right) \int_{\Theta_1} h d\nu - \underbrace{\int_{\Theta_1} \ln(g) h d\nu}_{=:(\mathtt{ii})} - \int_{\Theta_1^c} \ln(h) \, h d\nu.
\end{aligned}$$

Since h is a ν-density, the integral $\int_{\Theta_1} h d\nu$ is bounded by one, thus the second summand converges to zero as n approaches infinity almost surely according to lemma 4.5.

To calculate term (iv), consider the substitution

$$\vartheta \mapsto n^{1/2}\mathcal{I}(\vartheta_0)^{1/2}(\vartheta - \vartheta_n) = \theta, \tag{4.10}$$

which is well defined thanks to theorem 1.74. It follows that $\theta \sim \mathcal{N}_k(0, \mathbf{I}_k)$ and therefore its components follow a standard normal distribution and are independent. As a first step define c_2 to be the normalising constant of h to enhance readability:

$$c_2 := (2\pi)^{-k/2}(\det(n^{-1}\mathcal{I}(\vartheta_0)^{-1}))^{-1/2} = (2\pi)^{-k/2}n^{k/2}\sqrt{\det(\mathcal{I}(\vartheta_0))}.$$

Hence the integral evaluates to

$$\begin{aligned}
(\mathtt{iv}) &= \int \ln(h) h d\nu \\
&= \int \left(\ln(c_2) - \frac{n}{2}(\vartheta - \vartheta_n)^\top \mathcal{I}(\vartheta_0)(\vartheta - \vartheta_n)\right) c_2 \exp\left(-\frac{n}{2}(\vartheta - \vartheta_n)^\top \mathcal{I}(\vartheta_0)(\vartheta - \vartheta_n)\right) \nu(d\vartheta) \\
&= \ln(c_2) - \frac{1}{2}\int \|\theta\|^2 (2\pi)^{-k/2} \exp\left(-\frac{1}{2}\|\theta\|^2\right) \nu(d\theta) \\
&= \ln(c_2) - \frac{1}{2}\mathbb{E}_{\mathcal{N}_k(0,\mathbf{I}_k)} \|\theta\|^2 \\
&= \ln(c_2) - \frac{1}{2}\mathbb{E}_{\mathcal{N}_k(0,\mathbf{I}_k)} \sum_{j=1}^{k} \theta_j^2 \\
&= \ln(c_2) - \frac{1}{2}\sum_{j=1}^{k} \mathbb{E}_{\mathcal{N}(0,1)} \theta_j^2 \\
&= \ln(c_2) - \frac{k}{2} \\
&= -\frac{k}{2}\ln(2\pi) + \frac{k}{2}\ln(n) + \frac{1}{2}\ln(\det \mathcal{I}(\vartheta_0)) - \frac{k}{2}.
\end{aligned}$$

At first glance, the integral $\int \mathbb{1}_{\Theta_1^c} \ln(h) h d\nu$ looks similar to (iv), however it is slightly more tricky to evaluate. As a first step consider substitution (4.10) again. The transformed region will be defined as

$$\tilde{\Theta}_1^c = \tilde{\Theta}_1^c(\mathbf{X}_{(n)}(\omega)) := \left\{\theta \in \mathbb{R}^k : \left\| n^{-1/2}\mathcal{I}(\vartheta_0)^{-1/2}\theta + \vartheta_n \circ \mathbf{X}_{(n)}(\omega) - \vartheta_0 \right\| > \rho_1 \right\}, \tag{4.11}$$

and for each $\vartheta \in \Theta$ the indicator $\mathbb{1}_{\tilde{\Theta}_1^c}(\vartheta)$ converges to zero $\mathbb{P}^{\vartheta_0}$-almost surely as n diverges. Since $\theta \sim \mathcal{N}_k(0, \mathbf{I}_k)$ it follows that $\forall j \in \{1, \ldots, k\} : \theta_j \sim \mathcal{N}(0, 1)$ and $\tilde{\theta} := \|\theta\|^2 \sim \chi^2(k)$. Now split the integral into two parts:

$$\int \mathbb{1}_{\Theta_1^c} \ln(h) h d\nu = \underbrace{\ln(c_2) H(\Theta_1^c)}_{=:\mathtt{A}} - \underbrace{\int_{\Theta_1^c} \frac{n}{2}(\vartheta - \vartheta_n)^\top \mathcal{I}(\vartheta_0)(\vartheta - \vartheta_n) h(\vartheta) \nu(d\vartheta)}_{=:\mathtt{B}}.$$

As for part A the question is, whether the indicator $\mathbb{1}_{\tilde{\Theta}_1^c}$ converges to (the) zero (function) fast enough compared to $\ln(c_2) = \mathcal{O}(\frac{k}{2}\ln(n))$.

$$\begin{aligned}
\mathtt{A} &= \left(-\frac{k}{2}\ln(2\pi) + \frac{k}{2}\ln(n) + \frac{1}{2}\ln(\det \mathcal{I}(\vartheta_0))\right) \\
&\quad \cdot \int \mathbb{1}_{\tilde{\Theta}_1^c} (2\pi)^{-k/2} n^{k/2} \sqrt{\det(\mathcal{I}(\vartheta_0))} \exp\left(-\frac{n}{2}(\vartheta - \vartheta_n)^\top \mathcal{I}(\vartheta_0)(\vartheta - \vartheta_n)\right) \nu(d\vartheta) \\
&= \left(-\frac{k}{2}\ln(2\pi) + \frac{k}{2}\ln(n) + \frac{1}{2}\ln(\det \mathcal{I}(\vartheta_0))\right) \mathbb{E}_{\mathcal{N}_k(0,\mathbf{I}_k)} \mathbb{1}_{\tilde{\Theta}_1^c}
\end{aligned}$$

It turns out that this question does not need to be answered now, because term `A` can be used to compensate for other terms asymptotically, which will be done when combining all terms in section 4.1.1.

For part $\texttt{B} = \mathbb{E}_{\mathcal{N}_k(0,\mathbf{I}_k)} \mathbb{1}_{\tilde{\Theta}_1^c}(\theta) \|\theta\|^2$ consider the dominating integral

$$\mathbb{E}_{\mathcal{N}_k(0,\mathbf{I}_k)} \|\theta\|^2 = \mathbb{E}_{\chi^s(k)} \tilde{\theta} = k < \infty, \tag{4.12}$$

and since the integrand $\mathbb{1}_{\tilde{\Theta}_1^c}(\theta) \|\theta\|^2$ converges to zero $\mathbb{P}^{\vartheta_0}$-a.s. – applying the same reasoning as in the proof of lemma 4.5 – `B` converges to zero $\mathbb{P}^{\vartheta_0}$-a.s. thanks to dominated convergence (theorem 1.8).

Term (`ii`) will firstly be split into three summands. Then, pointwise on $\mathscr{X}^n$, the function $f(X_j|\vartheta)$ will be expanded around $\vartheta_n(\mathbf{X}_{(n)})$ applying the Taylor formula in theorem 1.73.

$$\begin{aligned}
(\texttt{ii}) &= \int_{\Theta_1} \left(\sum_{j=1}^n [\ln f(X_j|\vartheta) - \ln f(X_j|\vartheta_n)] + \ln q(\vartheta) \right) h(\vartheta)\nu(d\vartheta) - \int_{\Theta_1} \ln(C(\mathbf{X}_{(n)})) h(\vartheta)\nu(d\vartheta) \\
&= \int_{\Theta_1} \left(\sum_{j=1}^n [\ln f(X_j|\vartheta) - \ln f(X_j|\vartheta_n)] + \ln q(\vartheta) \right) h(\vartheta)\nu(d\vartheta) - \ln(C(\mathbf{X}_{(n)})) H(\Theta_1) \\
&= \sum_{j=1}^n \int_{\Theta_1} [\ln f(X_j|\vartheta) - \ln f(X_j|\vartheta_n)] \, h(\vartheta)\nu(d\vartheta) + \int_{\Theta_1} \ln(q) \, hd\nu - \ln(C(\mathbf{X}_{(n)})) H(\Theta_1) \\
&= \sum_{j=1}^n \int_{\Theta_1} \left\langle \frac{\nabla_\vartheta f(X_j|\vartheta_n)}{f(X_j|\vartheta_n)}, \vartheta - \vartheta_n \right\rangle h(\vartheta)\nu(d\vartheta) \qquad (4.13) \\
&\quad + \sum_{j=1}^n \int_{\Theta_1} \int_0^1 \left\langle \vartheta - \vartheta_n, D_\vartheta^2 \ln f(X_j|\vartheta_n + t(\vartheta - \vartheta_n))(\vartheta - \vartheta_n) \right\rangle h(\vartheta)(1-t)dt \, \nu(d\vartheta) \qquad (4.14) \\
&\quad + \underbrace{\int_{\Theta_1} \ln(q) hd\nu}_{=:(\texttt{vi})} - \underbrace{\ln(C(\mathbf{X}_{(n)}))}_{=:(\texttt{iii})} H(\Theta_1)
\end{aligned}$$

To evaluate term (`ii`) one needs regularity conditions on the likelihood similar to those in Henze (2024) or Ibragimov and Has'minskii (1981).

Lemma 4.6 (Regularity Conditions on the Likelihood)
The following conditions allow the exchange of integration and differentiation.

(i) *Let $f(\omega|\vartheta)$ be differentiable wrt ϑ for each $\omega \in \Omega$, and let all partial derivatives be bounded as follows for every ϑ in a neighbourhood $\mathcal{U}(\theta) := \mathcal{B}(\theta, \rho_\theta) \subset \Theta$ of a $\theta \in \Theta$:*

$$\forall l \in \{1, \dots, k\} : \left| \frac{\partial}{\partial \vartheta_l} f(\omega|\vartheta) \right| \leq k_1(\omega) \text{for } \mu\text{-a.a. } \omega \in \Omega.$$

Furthermore let k_1 be integrable: $\int k_1 d\mu < \infty$. Under these conditions it follows that $\int \nabla_\vartheta f(\cdot|\vartheta) d\mu = 0$.

(ii) *If furthermore for every $\vartheta \in \mathcal{U}(\theta)$ and $\omega \in \Omega$ the second order partial derivatives exist and are bounded as follows:*

$$\forall (l, m) \in \{1, \dots, k\}^2 : \left| \frac{\partial^2}{\partial \vartheta_l \partial_m} f(\omega|\vartheta) \right| \leq k_2(\omega) \text{for } \mu\text{-a.a. } \omega \in \Omega.$$

If k_2 is integrable it follows that $\int D_\vartheta^2 f(\cdot|\vartheta) d\mu = 0$.

(iii) *Continuity of the second order partial derivatives for each* $\vartheta \in \mathcal{U}(\theta)$.

Proof. (i) Since the integral is evaluated component-wise:

$$\int \nabla_\vartheta f(\cdot|\vartheta) d\mu := \begin{pmatrix} \int \frac{\partial}{\partial \vartheta_1} f(\cdot|\vartheta) d\mu \\ \vdots \\ \int \frac{\partial}{\partial \vartheta_k} f(\cdot|\vartheta) d\mu \end{pmatrix},$$

it suffices to show that each coordinate converges to zero.
For ease of readability define $\eta(\vartheta_l) := f(\cdot|\vartheta_1, \ldots, \vartheta_l, \ldots, \vartheta_k)$ and note that

$$\lim_{m\to\infty} m\left(\eta\left(\vartheta_l + \frac{1}{m}\right) - \eta(\vartheta_l)\right) = \frac{\partial}{\partial \vartheta_l} f(\cdot|\vartheta) =: \xi(\vartheta_l).$$

Applying the Mean Value Theorem (compare (Bartle, 2011)) the following inequality holds for each $m \in \mathbb{N}$, if $\vartheta_{l_0} \in \left(\vartheta_l, \vartheta_l + \frac{1}{m}\right)$:

$$\int \left| m\left(\eta\left(\vartheta_l + \frac{1}{m}\right) - \eta(\vartheta_l)\right)\right| d\mu = \int \left|\frac{\partial}{\partial \vartheta_l}\eta(\vartheta_{l_0})\right| d\mu \leq \int k_1 d\mu < \infty.$$

Thanks to the linearity of measure integrals this allows the DCT to be applied to

$$\xi_m(\vartheta_l) := m\left(\eta\left(\vartheta_l + \frac{1}{m}\right) - \eta(\vartheta_l)\right),$$

noting that the derivative is zero since f is a μ-probability density:

$$\begin{aligned} \int \xi(\vartheta_l) d\mu &= \int \frac{\partial}{\partial \vartheta_l} f(\cdot|\vartheta) d\mu \\ &\overset{\text{DCT}}{=} \lim_{m\to\infty} \int \xi_m(\vartheta_l) d\mu \\ &\overset{\text{LIN}}{=} \lim_{m\to\infty} m\left[\int \eta\left(\vartheta_l + \frac{1}{m}\right) d\mu - \int \eta(\vartheta_l) d\mu\right] \\ &\overset{\text{DEF}}{=} \frac{\partial}{\partial \vartheta_l} \int f(\cdot|\vartheta) d\mu = 0. \end{aligned}$$

(ii) Apply the reasoning in the proof of part (i) to every component of the $(k \times k)$-matrix $D^2_\vartheta f(\cdot|\vartheta)$, taking the partial derivatives of $\frac{\partial}{\partial \vartheta_l} f(\omega|\vartheta)$, $l \in \{1, \ldots, k\}$, instead of $f(\omega|\vartheta)$. ■

Corollary 4.7
Given the conditions of lemma 4.6 (i) *for* $\vartheta \in \mathcal{U}(\theta)$ *it follows that*

$$\mathbb{E}_{\mathbb{P}^\theta} \nabla_\vartheta \ln f(\cdot|\theta) = 0;$$

in particular, this holds for $\theta = \vartheta_0$.
If the conditions of lemma 4.6 (ii) *and* (iii) *are also fulfilled, the Fisher information matrix can be calculated as*

$$\mathcal{I}(\theta) = -\mathbb{E}_{\mathbb{P}^\theta} D^2_\theta \ln f(\cdot|\theta).$$

Proof. The proof will be given for $\theta = \vartheta_0$ without loss of generality (WLOG):

$$\mathbb{E}_{\mathbb{P}^{\vartheta_0}} \nabla_\vartheta \ln f(\cdot|\vartheta_0) = \int \frac{\nabla_\vartheta f(\cdot|\vartheta_0)}{f(\cdot|\vartheta_0)} f(\cdot|\vartheta_0) d\mu = 0.$$

For the second part consider the following calculations:

$$\begin{aligned}
\mathbb{V}_{\mathbb{P}^{\vartheta_0}} \nabla_\vartheta \ln f(\cdot|\vartheta_0) &\overset{\text{(i)}}{=} \mathbb{E}_{\mathbb{P}^{\vartheta_0}} \left(\nabla_\vartheta \ln f(\cdot|\vartheta_0)\right) \left(\nabla_\vartheta \ln f(\cdot|\vartheta_0)\right)^\top \\
&\overset{\text{(ii)}}{=} -\int D^2_\vartheta f(\cdot|\vartheta_0) d\mu + \mathbb{E}_{\mathbb{P}^{\vartheta_0}} \left(\nabla_\vartheta \ln f(\cdot|\vartheta_0)\right) \left(\nabla_\vartheta \ln f(\cdot|\vartheta_0)\right)^\top \\
&= -\mathbb{E}_{\mathbb{P}^{\vartheta_0}} \frac{D^2_\vartheta f(\cdot|\vartheta_0)}{f(\cdot|\vartheta_0)} + \mathbb{E}_{\mathbb{P}^{\vartheta_0}} \left(\nabla_\vartheta \ln f(\cdot|\vartheta_0)\right) \left(\nabla_\vartheta \ln f(\cdot|\vartheta_0)\right)^\top \\
&\overset{\bullet}{=} \left(-\int \frac{f(\cdot|\vartheta_0) \frac{\partial^2}{\partial\vartheta_i \partial\vartheta_j} f(\cdot|\vartheta_0) - \frac{\partial}{\partial\vartheta_j} f(\cdot|\vartheta_0) \frac{\partial}{\partial\vartheta_i} f(\cdot|\vartheta_0)}{(f(\cdot|\vartheta_0))^2} f(\cdot|\vartheta_0) d\mu \right)_{i,j=1\ldots k} \\
&= -\mathbb{E}_{\mathbb{P}^{\vartheta_0}} D_\vartheta \frac{\nabla_\vartheta f(\cdot|\vartheta_0)}{f(\cdot|\vartheta_0)} = -\mathbb{E}_{\mathbb{P}^{\vartheta_0}} D^2_\vartheta \ln f(\cdot|\vartheta_0),
\end{aligned}$$

where $\bullet$ holds thanks to the Schwarz-Clairaut theorem (equality of mixed partial derivatives (Duistermaat and Kolk, 2004)) and lemma 4.6 (iii). ■

Due to the linearity of measure integrals the term in line (4.13) disappears because ϑ_n is assumed to be the maximum likelihood estimator, i.e.

$$0 = \nabla_\vartheta \ln f_n(\mathbf{X}_{(n)}|\vartheta_n) = \nabla_\vartheta \ln \left(\prod_{j=1}^n f(X_j|\vartheta_n) \right) = \sum_{j=1}^n \frac{\nabla_\vartheta f(X_j|\vartheta_n)}{f(X_j|\vartheta_n)}.$$

In order to examine the remainder $R_{\mathbf{X}_{(n)}(\omega)}(\vartheta)$ of the Taylor expansion in (4.14), consider substitution (4.8) from lemma 4.5 again and remember that for the indicator of the transformed region (4.9) we have $\lim_{n\to\infty} \mathbb{1}_{\tilde{\Theta}_1}(\theta) = 1$ $\mathbb{P}^{\vartheta_0}$-a.s.

$$\begin{aligned}
&\sum_{j=1}^n \int_{\Theta_1} \int_0^1 \left\langle \vartheta - \vartheta_n, D^2_\vartheta \ln f(X_j|\vartheta_n + t(\vartheta - \vartheta_n))(\vartheta - \vartheta_n) \right\rangle h(\vartheta)(1-t) dt\, \nu(d\vartheta) \\
&= \int_{\Theta_1} \int_0^1 \left\langle \vartheta - \vartheta_n, D^2_\vartheta \sum_{j=1}^n \ln f(X_j|\vartheta_n + t(\vartheta - \vartheta_n))(\vartheta - \vartheta_n) \right\rangle h(\vartheta)(1-t) dt\, \nu(d\vartheta) \\
&= \int_{\tilde{\Theta}_1} \int_0^1 \left\langle \theta, \frac{1}{n} \sum_{j=1}^n D^2_\vartheta \ln f(X_j|\vartheta_n + tn^{-1/2}\theta)\theta \right\rangle n^{-k/2} h(n^{-1/2}\theta + \vartheta_n)(1-t) dt\, \nu(d\theta) \\
&= \int_0^1 \int_{\tilde{\Theta}_1} \left\langle \theta, \frac{1}{n} \sum_{j=1}^n D^2_\vartheta \ln f(X_j|\vartheta_n + tn^{-1/2}\theta)\theta \right\rangle \\
&\qquad (2\pi)^{-k/2} \sqrt{\det \mathcal{I}(\vartheta_0)} \exp\left(-\frac{1}{2} \theta^\top \mathcal{I}(\vartheta_0) \theta \right) \nu(d\theta)(1-t) dt
\end{aligned}$$

Where the last equality holds thanks to the theorems of Fubini 1.9 and Tonelli 1.10 if $D^2_\vartheta \ln f$ is bounded, because, in that case, the integrand is bounded by a moment of a Gaussian distribution. According to the SLLN, the term $\frac{1}{n} \sum_{j=1}^n D^2_\vartheta \ln f(X_j|\vartheta_0)$ converges $\mathbb{P}^{\vartheta_0}$-almost surely to

$$\mathbb{E}_{\mathbb{P}^{\vartheta_0}_{X_1}} D^2_\vartheta \ln f(\cdot|\vartheta_0) = -\mathcal{I}(\vartheta_0)$$

as $n \to \infty$, assuming the regularity conditions in lemma 4.6 hold.

If it holds true that also $D^2_\vartheta \ln f(X_j|\vartheta_n + tn^{-1/2}\theta) \overset{\mathbb{P}^{\vartheta_0}\text{a.s.}}{\to} -\mathcal{I}(\vartheta_0)$, the limit of the remainder could then be evaluated using a similar argument as for the substitution in (4.10):

$$\begin{aligned}
&-\int_0^1 \int \langle\theta, \mathcal{I}(\vartheta_0)\theta\rangle (2\pi)^{-k/2}\sqrt{\det(\mathcal{I}(\vartheta_0))}\exp\left(-\frac{1}{2}\langle\theta, \mathcal{I}(\vartheta_0)\theta\rangle\right)\nu(d\theta)(1-t)dt\\
&= -\frac{1}{2}\mathbb{E}_{\mathcal{N}_k(0,\mathcal{I}(\vartheta_0)^{-1})}\langle\theta, \mathcal{I}(\vartheta_0)\theta\rangle\\
&= -\frac{1}{2}\mathbb{E}_{\mathcal{N}_k(0,\mathbf{I}_k)}\|\theta'\|^2 = -\frac{1}{2}\mathbb{E}_{\chi^2(k)}\tilde{\theta} = -\frac{k}{2}.
\end{aligned}$$

The following lemma proves that only continuity is needed for the assumption above.

Lemma 4.8

(a) Assuming $D^2_\vartheta \ln f(X_j|\vartheta)$ to be continuous in $\vartheta_0 \in \Theta$ and $\vartheta_n \overset{\mathbb{P}^{\vartheta_0}\,a.s.}{\to} \vartheta_0$ the following statement holds:

$$\mathbb{P}^{\vartheta_0}\left[\lim_{n\to\infty}\left\|\frac{1}{n}\sum_{j=1}^n D^2_\vartheta \ln f(X_j|\vartheta_n + tn^{-1/2}\theta) - \frac{1}{n}\sum_{j=1}^n D^2_\vartheta \ln f(X_j|\vartheta_0)\right\| = 0\right] = 1.$$

(b) If the likelihood fulfills the conditions in (a), as well as the regularities listed in lemma 4.6 $\frac{1}{n}\sum_{j=1}^n D^2_\vartheta \ln f(X_j|\vartheta_n + tn^{-1/2}\theta) \overset{\mathbb{P}^{\vartheta_0}\,a.s.}{\to} -\mathcal{I}(\vartheta_0)$.

Proof. (a) Almost sure convergence is equivalent to pointwise convergence on a set with probability one, more specifically

$$\vartheta_n(\mathbf{X}_{(n)}(\omega)) \overset{\mathbb{P}^{\vartheta_0}\text{a.s.}}{\to} \vartheta_0 \Leftrightarrow \exists A \in \mathfrak{A} : \mathbb{P}^{\vartheta_0}(A) = 1 \wedge \forall\omega \in A : \lim_{n\to\infty}\vartheta_n(\mathbf{X}_{(n)}(\omega)) = \vartheta_0.$$

Thus continuity of $D^2_\vartheta \ln f(X_j(\omega)|\vartheta)$ in ϑ_0 ensures, that

$$\lim_{n\to\infty} D^2_\vartheta \ln f(X_j(\omega)|\vartheta_n \circ \mathbf{X}_{(n)}(\omega) + tn^{-1/2}\theta) = D^2_\vartheta \ln f(X_j(\omega)|\vartheta_0)$$

pointwise on A, q.e.d.

(b) According to the SLLN $\frac{1}{n}\sum_{j=1}^n D^2_\vartheta \ln f(X_j|\vartheta_0)$ converges almost surely to $\mathbb{E}D^2_\vartheta \ln f(\cdot|\vartheta_0)$, and the regularity conditions in lemma 4.6 ensure that Dominated Convergence can be applied and therefore the expected value equals $-\mathcal{I}(\vartheta_0)$. Now

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^n D^2_\vartheta \ln f(X_j|\vartheta_n + tn^{-1/2}\theta)\\
&= \frac{1}{n}\sum_{j=1}^n D^2_\vartheta \ln f(X_j|\vartheta_n + tn^{-1/2}\theta) - D^2_\vartheta \ln f(X_j|\vartheta_0) + D^2_\vartheta \ln f(X_j|\vartheta_0) \overset{\mathbb{P}^{\vartheta_0}\text{a.s.}}{\to} 0 - \mathcal{I}(\vartheta_0).
\end{aligned}$$

■

To evaluate term (`vi`) apply substitution (4.8) again. Then, the reasoning below holds for $\mathbb{P}^{\vartheta_0}$-a.a. ω (REMINDER: $\vartheta_n = \vartheta_n \circ \mathbf{X}_{(n)}(\omega)$):

$$\begin{aligned}
(\texttt{vi}) &= \int_{\Theta_1} \ln(q) h d\nu = \mathbb{E}_H \ln(q)\mathbb{1}_{\Theta_1}\\
&= \int_{\Theta_1} \ln(q(\vartheta))(2\pi)^{-k/2}n^{k/2}\sqrt{\det(\mathcal{I}(\vartheta_0))}\exp\left(-\frac{n}{2}(\vartheta - \vartheta_n(\mathbf{X}_{(n)}))^\top\mathcal{I}(\vartheta_0)(\vartheta - \vartheta_n(\mathbf{X}_{(n)}))\right)\nu(d\vartheta)\\
&= \int_{\tilde{\Theta}_1(\mathbf{X}_{(n)})} \ln(q(\vartheta_n(\mathbf{X}_{(n)}) + n^{-1/2}\theta))\underbrace{(2\pi)^{-k/2}\sqrt{\det(\mathcal{I}(\vartheta_0))}\exp\left(-\frac{1}{2}\theta^\top\mathcal{I}(\vartheta_0)\theta\right)}_{=:f(\theta)}\nu(d\theta).
\end{aligned}$$

Firstly we can state that $\forall\theta\in\Theta:\lim_{n\to\infty}\vartheta_n(\mathbf{X}_{(n)})+n^{-1/2}\theta=\vartheta_0\ \mathbb{P}^{\vartheta_0}$-a.s. Now assume q to be bounded for each θ in a neighbourhood $\mathcal{U}(\vartheta_0)$ of ϑ_0. Consequently

$$\sup_{\theta\in\mathcal{U}(\vartheta_0)}|\ln(q(\theta)|\leq c_q<\infty, \tag{4.15}$$

and therefore

$$\int\mathbb{1}_{\tilde{\Theta}_1}(\theta)\left|\ln(q(\vartheta_n+n^{-1/2}\theta))f(\theta)\right|\nu(d\theta)\leq\int c_q f(\theta)\ \nu(d\theta)=c_q.$$

In essence, the DCT is applicable and implies (`iv`) $\mathbb{P}^{\vartheta_0}$-a.s. converging to

$$\int\ln(q(\vartheta_0))f(\theta)\nu(d\theta)=\ln(q(\vartheta_0)).$$

The next term in line is (`iii`) $H(\Theta_1)$. Remembering lemma 4.5 we observe

$$\lim_{n\to\infty}H_n(\mathbf{X}_{(n)},\vartheta_0)(\Theta_1)=1\ \mathbb{P}^{\vartheta_0}-a.s.$$

Subsequently expand (`iii`) applying theorem 1.73 and notice the first term disappearing by construction:

$$\begin{aligned}
(\texttt{iii})&=\ln\left[\int\prod_{j=1}^{n}\frac{f(X_j|\vartheta')}{f(X_j|\vartheta_n(\mathbf{X}_{(n)}))}q(\vartheta')\nu(d\vartheta')\right]\\
&=\ln\left[\int\exp\left(\sum_{j=1}^{n}\ln f(X_j|\vartheta')-\ln f(X_j|\vartheta_n)\right)q(\vartheta')\nu(d\vartheta')\right]\\
&=\ln\left[\int\exp\left(\sum_{j=1}^{n}\left\langle\frac{\nabla_\vartheta f(X_j|\vartheta_n)}{f(X_j|\vartheta_n)},\vartheta'-\vartheta_n\right\rangle\right.\right.\\
&\quad\left.\left.+\int_0^1(1-t)\left\langle\vartheta-\vartheta_n,\sum_{j=1}^{n}D^2_\vartheta\ln f(X_j|\vartheta_n+t(\vartheta'-\vartheta_n))(\vartheta-\vartheta_n)\right\rangle dt\right)q(\vartheta')\nu(d\vartheta')\right].
\end{aligned}$$

Since ϑ_n was taken to be the MLE the term $\frac{\nabla_\vartheta f(X_j|\vartheta_n)}{f(X_j|\vartheta_n)}$ disappears, under regularity conditions that follow from lemma 4.6. What is left is the function of the remainder term. Before its evaluation consider the subsequent lemma.

Lemma 4.9
The function $y\mapsto\int_0^1\xi\left(y+\frac{t}{n}\right)dt$ is continuous if ξ is continuous and uniformly bounded.

Proof. Continuity ensures that $\lim_{m\to\infty}\xi\left(y+\frac{1}{m}+\frac{t}{n}\right)-\xi\left(y+\frac{t}{n}\right)=0$. If ξ is also uniformly bounded the consideration of $\left|\xi\left(y+\frac{1}{m}+\frac{t}{n}\right)-\xi\left(y+\frac{t}{n}\right)\right|\leq2\left\|\xi\right\|_\infty$ ensures the applicability of the DCT and therefore

$$\begin{aligned}
&\lim_{m\to\infty}\int_0^1\xi\left(y+\frac{1}{m}+\frac{t}{n}\right)dt-\int_0^1\xi\left(y+\frac{t}{n}\right)dt=\lim_{m\to\infty}\int_0^1\left[\xi\left(y+\frac{1}{m}+\frac{t}{n}\right)-\xi\left(y+\frac{t}{n}\right)\right]dt\\
&=\int_0^1 0dt=0.
\end{aligned}$$

■

$$
\begin{aligned}
(\texttt{iii}) &= \ln\left[\int \exp\left(\int_0^1 (1-t)\left\langle \vartheta-\vartheta_n, \sum_{j=1}^n D^2_\vartheta \ln f(X_j|\vartheta_n + t(\vartheta-\vartheta_n))(\vartheta'-\vartheta_n)\right\rangle\right) q(\vartheta')\nu(d\vartheta')dt\right] \\
&= \ln\left[\int \exp\left(\int_0^1 (1-t)\left\langle \theta, \frac{1}{n}\sum_{j=1}^n D^2_\vartheta \ln f(X_j|\vartheta_n + tn^{-1/2}\theta)\theta\right\rangle dt\right) q(\vartheta_n + n^{-1/2}\theta) n^{-k/2}\nu(d\theta)\right] \\
&= -\frac{k}{2}\ln(n) + \\
&\qquad \ln\left[\int \exp\left(\int_0^1 (1-t)\left\langle \theta, \frac{1}{n}\sum_{j=1}^n D^2_\vartheta \ln f(X_j|\vartheta_n + tn^{-1/2}\theta)\theta\right\rangle dt\right) q(\vartheta_n + n^{-1/2}\theta)\nu(d\theta)\right]
\end{aligned}
$$

If q is continuous in ϑ_0 the CMT implies $\vartheta_n + n^{-1/2}\theta \to \vartheta_0$ $\mathbb{P}^{\vartheta_0}$-a.s., thus $q(\vartheta_n + n^{-1/2}\theta) \to q(\vartheta_0)$ $\mathbb{P}^{\vartheta_0}$-a.s. If, furthermore, $D^2_\vartheta \ln f(X_j|\vartheta)$ is continuous in ϑ_0 and bounded (as already assumed above), lemma 4.8 becomes applicable to the term $(\texttt{iii}) + \frac{k}{2}\ln(n)$:

$$
\begin{aligned}
\lim_{n\to\infty} (\texttt{iii}) + \frac{k}{2}\ln(n) &= \ln\left[\int \exp\left(\int_0^1 (1-t)\left\langle \theta, -\mathcal{I}(\vartheta_0)\theta\right\rangle dt\right) q(\vartheta_0)\nu(d\theta)\right] \ \mathbb{P}^{\vartheta_0} - a.s. \\
&= \ln\left[\int f(\theta)\frac{(2\pi)^{k/2}}{\sqrt{\det \mathcal{I}(\vartheta_0))}} q(\vartheta_0)\nu(d\theta)\right] \\
&= \ln(q(\vartheta_0)) - \frac{1}{2}\ln(\det \mathcal{I}(\vartheta_0)) + \frac{k}{2}\ln(2\pi). \qquad (4.16)
\end{aligned}
$$

By applying lemmata 1.57 and 4.5 to $H(\Theta_1)\left((\texttt{iii}) + \frac{k}{2}\ln(n)\right)$ one can see that the product almost surely converges to (4.16), as well.

4.1.1 Combining the Terms

After expanding the Kullback-Leibler divergence $\mathcal{K}\left(H, \tilde{G}\right)$ we have developed conditions under which some of terms converge $\mathbb{P}^{\vartheta_0}$-almost surely to degenerate limits, if k is fixed, while other terms remain. It will now be shown that combining these terms cancels them out and one can follow the desired result of $d_{\mathrm{TV}}(G, H)$ converging to zero $\mathbb{P}^{\vartheta_0}$-a.s.

$$
\begin{aligned}
&\mathcal{K}\left(H, \tilde{G}\right) = (\texttt{iv}) - (\texttt{ii}) - \ln(1-\tilde{H}) \int_{\Theta_1} h d\nu - \int_{\Theta_1^c} \ln(h)\, h d\nu \\
&= (\texttt{iv}) - (4.13) - (4.14) - (\texttt{vi}) + H(\Theta_1)\left[(\texttt{iii}) + \frac{k}{2}\ln(n)\right] - H(\Theta_1)\frac{k}{2}\ln(n) \\
&\quad - \ln(1-\tilde{H}) \int_{\Theta_1} h d\nu - \texttt{A} + \texttt{B} \\
&= -\frac{k}{2}\ln(2\pi) + \frac{k}{2}\ln(n) + \frac{1}{2}\ln(\det \mathcal{I}(\vartheta_0)) - \frac{k}{2} \\
&\quad - \sum_{j=1}^{n} \int_{\Theta_1} \int_0^1 \left\langle \vartheta - \vartheta_n, D_\vartheta^2 \ln f(X_j | \vartheta_n + t(\vartheta - \vartheta_n))(\vartheta - \vartheta_n) \right\rangle h(\vartheta)(1-t) dt\, \nu(d\vartheta) \\
&\quad - \int_{\Theta_1} \ln(q) h d\nu \\
&\quad + \ln\left[\int \exp\left(\int_0^1 (1-t) \left\langle \theta, \frac{1}{n} \sum_{j=1}^{n} D_\vartheta^2 \ln f(X_j | \vartheta_n + t n^{-1/2}\theta)\theta \right\rangle dt \right) q(\vartheta_n + n^{-1/2}\theta) \nu(d\theta) \right] \\
&\quad \cdot H(\Theta_1) - \frac{k}{2}\ln(n) H(\Theta_1) \\
&\quad - \int_{\Theta_1} \ln(1-\tilde{H}) h d\nu \\
&\quad - \texttt{A} + \texttt{B} \\
&= -\frac{k}{2}\ln(2\pi) + \frac{1}{2}\ln(\det \mathcal{I}(\vartheta_0)) - \frac{k}{2} \\
&\quad - \sum_{j=1}^{n} \int_{\Theta_1} \int_0^1 \left\langle \vartheta - \vartheta_n, D_\vartheta^2 \ln f(X_j | \vartheta_n + t(\vartheta - \vartheta_n))(\vartheta - \vartheta_n) \right\rangle h(\vartheta)(1-t) dt\, \nu(d\vartheta) \\
&\quad - \int_{\Theta_1} \ln(q) h d\nu \\
&\quad + \ln\left[\int \exp\left(\int_0^1 (1-t) \left\langle \theta, \frac{1}{n} \sum_{j=1}^{n} D_\vartheta^2 \ln f(X_j | \vartheta_n + t n^{-1/2}\theta)\theta \right\rangle dt \right) q(\vartheta_n + n^{-1/2}\theta) \nu(d\theta) \right] \\
&\quad \cdot H(\Theta_1) \\
&\quad - \int_{\Theta_1} \ln(1-\tilde{H}) h d\nu + \texttt{B} \\
&\quad - H(\Theta_1^c)\left(-\frac{k}{2}\ln(2\pi) + \frac{k}{2}\ln(n) + \frac{1}{2}\ln(\det \mathcal{I}(\vartheta_0)\right) + \frac{k}{2}\ln(n) - \frac{k}{2}\ln(n) H(\Theta_1) \\
&\overset{\mathbb{P}^{\vartheta_0}\text{ a.s.}}{\to} 0 \quad (n \to \infty)
\end{aligned}
$$

Remark 4.10. Term A was combined with parts of terms (iii) and (iv) to form

$$
\begin{aligned}
&H(\Theta_1^c)\left(-\frac{k}{2}\ln(2\pi) + \frac{k}{2}\ln(n) + \frac{1}{2}\ln(\det \mathcal{I}(\vartheta_0)\right) + \frac{k}{2}\ln(n) + \frac{k}{2}\ln(n) H(\Theta_1) \\
&= H(\Theta_1^c)\left(-\frac{k}{2}\ln(2\pi) + \frac{1}{2}\ln(\det \mathcal{I}(\vartheta_0)\right), \qquad (4.17)
\end{aligned}
$$

which converges to zero $\mathbb{P}^{\vartheta_0}$-a.s. according to lemma 4.5 for any finite $k \in \mathbb{N}$.

Remark 4.11. Our BvM theorem implicitly assumes consistency of the prior, which requires a condition similar to (viii) in lemma 2.17 from Le Cam (1986a,b); Le Cam and Lo Yang (2000). For this we also (implicitly) suppose the testing condition (2.7) or lemma 2.17 (iv) to hold. Since Θ is a Borel subset of a Polish space and $\Theta_0 \subset \text{supp}(q)$, a modified version of theorem 1.70 using the Euclidean metric should let us construct such tests without too much difficulty. Thus this assumption shall be considered justified. More on consistency in parametric models is presented in van der Vaart (1998), chapter 10.

Lemma 4.12 (Consistency of the Prior, Prior Mass Condition)
The prior in the experiment above is consistent given

$$\int \mathbb{1}_{\Theta_0} |q(t) - q(\vartheta_0)| \, \lambda^{(k)}(dt) = o(\rho_1^k), \quad \rho_1 \to 0.$$

Proof. Le Cam's second part of the condition is easily proved by considering

$$\liminf_{\rho_1 \to 0} \frac{\lambda^{(k)}(\Theta_0 \cap \Theta)}{\lambda^{(k)}(\Theta_0)} \equiv 1 > 0.$$

The first part involves the convergence of q to $a = q(\vartheta_0)$ fast enough. Now take $\lambda^{(k)}(\Theta_0) = \frac{\pi^{k/2}}{\Gamma\left(\frac{k}{2}+1\right)} \rho_1^k = o(\rho_1^{k-1})$ (Smith and Vamanamurthy, 1989), then the first term is $o(1)$ as $\rho_1 \to 0$ given the condition in the lemma. This holds for any fixed $k \in \mathbb{N}$. ■

Remark 4.13. Another reason to assume consistency is the applicability of Doob's theorem 1.72 as long as the conditional image measure of the observations is identifiable in the sense of the theorem. Also compare lemma 2.17 (ii).

Theorem 4.14 (Bernstein-von Mises Theorem for a General Posterior)
Consider the dominated experiment from (4.1)

$$\mathscr{E}_n := \left(\mathscr{X}^n, \mathfrak{X}^{(n)}, \left\{ \mathbb{P}^{\vartheta}_{\mathbf{X}_{(n)}} : \vartheta \in \Theta \right\} \right),$$

and take the parameter space $\Theta \subset \mathbb{R}^k$, where $(\Theta, \mathfrak{T}, Q)$ is a probability space and $Q = \mathfrak{L}(\vartheta) \ll \nu = \lambda^{(k)}$ is the image measure of ϑ. The prior density $q := \frac{dQ}{d\nu}$ must be defined on supp$(q) := \Theta_1 \supset \Theta_0$ as defined in (4.3) and (4.4), and is required to be continuous in ϑ_0 and to be bounded as in (4.15):

$$\sup_{\theta \in \mathcal{U}(\vartheta_0)} |\ln(q(\theta)| \leq c_q < \infty.$$

Assume the existence of a strongly consistent MLE in the sense of (4.2) for each $\vartheta \in \Theta$.
If the likelihood meets the regularity conditions of lemmata 4.6 and 4.8, and if $D^2_\vartheta \ln f$ is also bounded, and the Fisher information matrix regular and positive definite,
then *the posterior $G = Q(\cdot|\mathbf{X}_{(n)})$ defined by its density*

$$g(\cdot|\mathbf{X}_{(n)}) := \frac{dQ(\cdot|\mathbf{X}_{(n)})}{d\nu}, \quad g(\vartheta|\mathbf{X}_{(n)}) = \frac{q(\vartheta) \prod_{j=1}^n f(X_j|\vartheta)}{\int q(\vartheta') \prod_{j=1}^n f(X_j|\vartheta') \nu(d\vartheta')}$$

approaches the Gaussian

$$H_n(\mathbf{X}_{(n)}, \vartheta_0) = \mathcal{N}_k(\vartheta_n \circ \mathbf{X}_{(n)}, n^{-1} \mathcal{I}(\vartheta_0)^{-1}))$$

in the sense that

$$\lim_{n \to \infty} d_{TV}\left(H_n(\mathbf{X}_{(n)}, \vartheta_0), Q(\cdot|\mathbf{X}_{(n)}) \right) = 0, \quad \mathbb{P}^{\vartheta_0} - a.s.$$

4.2 Bernstein-von Mises Theorem for a Finite-Dimensional Subspace in Group Testing

In the setting of section 3 with equally sized subgroups the parameter p will be projected onto a finite-dimensional subspace such that, in essence, a parametric setting is achieved, similar to the initial approach by Giné and Nickl (2015). More specifically:

$$p_\vartheta : \mathscr{X} \to [0,1], \quad x \mapsto \sum_{k=1}^{K} \vartheta_k \varphi_k(x), \tag{4.18}$$

where $\vartheta := (\vartheta_1, \dots, \vartheta_K)^\top$ will be interpreted as a random variable on the measurable space $(\mathcal{U}(\theta), \mathfrak{B}(\mathcal{U}(\theta)))$ equipped with the Borel-σ-algebra, where ϑ is assumed to be defined on a compact space $\mathcal{U}(\theta) \subset \mathbb{R}^K$ which includes the *true* parameter θ in accordance with lemma 4.6. Additionally it is assumed that the prior is dominated by the K-dimensional Lebesgue measure: $\Pi \ll \lambda^{(K)}$.
Furthermore μ is assumed to be the Lebesgue-measure and $\mathscr{X}$ to be compact set in $\mathbb{R}$, and as mentioned in section 3.

The task now is to check the conditions of theorem 4.14 for the likelihood at hand. As a first step it will be shown that the necessary regularity conditions are met, and to do so the partial derivatives of the likelihood and the second order derivative of the log-likelihood $\ell(Y_j^*, \mathbf{X}_j|\vartheta) := \ln\left(\mathscr{L}(Y_j^*, \mathbf{X}_j|\vartheta)\right)$ are calculated. The regularity conditions can then be stated as:

(i) Dominated convergence of the first order derivative:

$$\sup_{\vartheta \in \mathcal{U}(\theta)} \left| \frac{\partial}{\partial \vartheta_l} \mathscr{L}(Y_j^*, \mathbf{X}_j|\vartheta) \right| < k_1(Y_j^*, \mathbf{X}_j)$$

where $\int k_1 d(\zeta \otimes \mu^{(n)}) < \infty$.

(ii) Dominated convergence of the second order derivative:

$$\sup_{\vartheta \in \mathcal{U}(\theta)} \left| \frac{\partial^2}{\partial \vartheta_l \partial \vartheta_m} \mathscr{L}(Y_j^*, \mathbf{X}_j|\vartheta) \right| < k_2(Y_j^*, \mathbf{X}_j)$$

where $\int k_2 d(\zeta \otimes \mu^{(n)}) < \infty$.

(iii) And continuity of the second order derivative of the log-likelihood in $\vartheta_0 \in \mathbb{R}^K$, i.e.

$$\forall \varepsilon > 0 : \exists \delta > 0 : \|\vartheta - \vartheta_0\| < \delta \Rightarrow \left\| D_\vartheta^2 \ell(Y_j^*, \mathbf{X}_j|\vartheta) - D_\vartheta^2 \ell(Y_j^*, \mathbf{X}_j|\vartheta_0) \right\| < \varepsilon.$$

(iv) Boundedness of the second derivative of the log-likelihood.

(v) Symmetry and positive definiteness of the Fisher information matrix.

The likelihood for this problem is

$$\begin{aligned}
&\mathscr{L}(Y_j^*, \mathbf{X}_j|\vartheta) \\
&= \exp\left(Y_j^* \ln\left[1 - \prod_{i=1}^{n}(1 - \sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij}))\right] + (1-Y_j^*)\sum_{i=1}^{n}\ln\left[1 - \sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij})\right]\right) g_n(\mathbf{X}_j),
\end{aligned}$$

and the first and second order partial derivatives are given below:

$$\begin{aligned}
&\frac{\partial}{\partial\vartheta_l}\mathscr{L}(Y_j^*, \mathbf{X}_j|\vartheta) \\
&= \exp\left(Y_j^* \ln\left[1 - \prod_{i=1}^{n}(1 - \sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij}))\right] + (1-Y_j^*)\sum_{i=1}^{n}\ln\left[1 - \sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij})\right]\right) g_n(\mathbf{X}_j) \\
&\quad \cdot \left[Y_j^* \frac{\sum_{i=1}^{n}\varphi_l(X_{ij})\prod_{r\in\{1,\dots,n\}\setminus\{i\}}\left[1 - \sum_{k=1}^{K}\vartheta_k\varphi_k(X_{rj})\right]}{1 - \prod_{i=1}^{n}(1 - \sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij}))} - (1-Y_j^*)\sum_{i=1}^{n}\frac{\varphi_l(X_{ij})}{1 - \sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij})}\right],
\end{aligned}$$

$$\begin{aligned}
&\frac{\partial^2}{\partial\vartheta_l\partial\vartheta_m}\mathscr{L}(Y_j^*, \mathbf{X}_j|\vartheta) \\
&= \exp\left(Y_j^* \ln\left[1 - \prod_{i=1}^{n}(1 - \sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij}))\right] + (1-Y_j^*)\sum_{i=1}^{n}\ln\left[1 - \sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij})\right]\right) g_n(\mathbf{X}_j) \\
&\quad \cdot \left[Y_j^* \frac{\sum_{i=1}^{n}\varphi_m(X_{ij})\prod_{r\in\{1,\dots,n\}\setminus\{i\}}\left[1 - \sum_{k=1}^{K}\vartheta_k\varphi_k(X_{rj})\right]}{1 - \prod_{i=1}^{n}(1 - \sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij}))} - (1-Y_j^*)\sum_{i=1}^{n}\frac{\varphi_m(X_{ij})}{1 - \sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij})}\right] \\
&\quad \cdot \left[Y_j^* \frac{\sum_{i=1}^{n}\varphi_l(X_{ij})\prod_{r\in\{1,\dots,n\}\setminus\{i\}}\left[1 - \sum_{k=1}^{K}\vartheta_k\varphi_k(X_{rj})\right]}{1 - \prod_{i=1}^{n}(1 - \sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij}))} - (1-Y_j^*)\sum_{i=1}^{n}\frac{\varphi_l(X_{ij})}{1 - \sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij})}\right] \\
&\quad + \exp\left(Y_j^* \ln\left[1 - \prod_{i=1}^{n}(1 - \sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij}))\right] + (1-Y_j^*)\sum_{i=1}^{n}\ln\left[1 - \sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij})\right]\right) g_n(\mathbf{X}_j) \\
&\quad \cdot \Bigg\{ Y_j^* \Bigg\{ \left[\prod_{i=1}^{n}\left[1 - \sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij})\right] - 1\right] \\
&\quad \cdot \left[\sum_{i=1}^{n}\varphi_l(X_{ij}) \sum_{r\in\{1,\dots,n\}\setminus\{i\}}\varphi_m(X_{rj}) \prod_{s\in\{1,\dots,n\}\setminus\{i,r\}}\left[1 - \sum_{k=1}^{K}\vartheta_k\varphi_k(X_{sj})\right]\right] \\
&\quad - \left[\sum_{i=1}^{n}\varphi_m(X_{ij}) \prod_{r\in\{1,\dots,n\}\setminus\{i\}}\left[1 - \sum_{k=1}^{K}\vartheta_k\varphi_k(X_{rj})\right]\right] \\
&\quad \cdot \left[\sum_{i=1}^{n}\varphi_l(X_{ij}) \prod_{r\in\{1,\dots,n\}\setminus\{i\}}\left[1 - \sum_{k=1}^{K}\vartheta_k\varphi_k(X_{rj})\right]\right] \Bigg\} \cdot \left[1 - \prod_{i=1}^{n}(1 - \sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij}))\right]^{-2} \\
&\quad - (1-Y_j^*)\sum_{i=1}^{n}\frac{\varphi_l(X_{ij})\varphi_m(X_{ij})}{\left(1 - \sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij})\right)^2} \Bigg\}.
\end{aligned}$$

The Hessian matrix of the log-likelihood $D^2_\vartheta \ln\left(\mathscr{L}(Y_j^*, \mathbf{X}_j|\vartheta)\right) = \left(D^2_\vartheta \ell(Y_j^*, \mathbf{X}_j|\vartheta)\right)$

$= \left(\frac{\partial^2}{\partial\vartheta_l\partial\vartheta_m}\ell(Y_j^*, \mathbf{X}_j|\vartheta)\right)_{(l,m)\in\{1,\dots,K\}^2}$ is defined by its entries

$$
\begin{aligned}
&\frac{\partial^2}{\partial\vartheta_l\partial\vartheta_m}\ell(Y_j^*, \mathbf{X}_j|\vartheta) \\
&\quad = Y_j^*\Bigg\{ \left[\prod_{i=1}^{n}\left[1-\sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij})\right]-1\right] \\
&\quad \cdot \left[\sum_{i=1}^{n}\varphi_l(X_{ij})\sum_{r\in\{1,\dots,n\}\setminus\{i\}}\varphi_m(X_{rj})\prod_{s\in\{1,\dots,n\}\setminus\{i,r\}}\left[1-\sum_{k=1}^{K}\vartheta_k\varphi_k(X_{sj})\right]\right] \\
&\quad - \left[\sum_{i=1}^{n}\varphi_m(X_{ij})\prod_{r\in\{1,\dots,n\}\setminus\{i\}}\left[1-\sum_{k=1}^{K}\vartheta_k\varphi_k(X_{rj})\right]\right] \\
&\quad \cdot \left[\sum_{i=1}^{n}\varphi_l(X_{ij})\prod_{r\in\{1,\dots,n\}\setminus\{i\}}\left[1-\sum_{k=1}^{K}\vartheta_k\varphi_k(X_{rj})\right]\right]\Bigg\}\cdot\left[1-\prod_{i=1}^{n}(1-\sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij}))\right]^{-2} \\
&\quad - (1-Y_j^*)\sum_{i=1}^{n}\frac{\varphi_l(X_{ij})\varphi_m(X_{ij})}{\left(1-\sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij})\right)^2}.
\end{aligned}
$$

As a first step it can be shown that condition (i) holds true if the following three additional conditions are met:

1. All basis functions are bounded, i.e. $\forall k \in \{1,\dots,K\} : \forall x \in \mathscr{X} : \varphi_k(x) \leq |\varphi_k(x)| \leq \|\varphi_k\|_\infty \leq \phi < \infty$,
2. p_ϑ must be bounded from below, more precisely: $\exists l > 0 : \forall\vartheta \in \mathcal{U}(\theta) : \forall x \in \mathscr{X} : p_\vartheta(x) > l$,
3. and from above: $\exists u < 1 : \forall\vartheta \in \mathcal{U}(\theta) : \forall x \in \mathscr{X} : p_\vartheta(x) < u$.

Now, note that

$$\exp\left(Y_j^*\ln\left[1-\prod_{i=1}^{n}(1-\sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij}))\right]+(1-Y_j^*)\sum_{i=1}^{n}\ln\left[1-\sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij})\right]\right)g_n(\mathbf{X}_j)$$

is a $(\zeta\otimes\mu^{(n)})$-(probability) density, which, using the conditions above and the triangle inequality, can be bounded from above by

$$\exp\left(1\cdot\ln\left(1-(1-u)^n\right)+1\cdot n\ln(1-l)\right)g_n(\mathbf{X}_j) = (1-(1-u)^n)(1-l)^n g_n(\mathbf{X}_j),$$

since the exponential and logarithm functions are strictly monotonous and both exp as well as g_n are nonnegative. Moreover, one can bound the additional factor

$$
\begin{aligned}
&\left|Y_j^*\frac{\sum_{i=1}^{n}\varphi_l(X_{ij})\prod_{r\in\{1,\dots,n\}\setminus\{i\}}\left[1-\sum_{k=1}^{K}\vartheta_k\varphi_k(X_{rj})\right]}{1-\prod_{i=1}^{n}(1-\sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij}))}-(1-Y_j^*)\sum_{i=1}^{n}\frac{\varphi_l(X_{ij})}{1-\sum_{k=1}^{K}\vartheta_k\varphi_k(X_{ij})}\right| \\
&< \left|1\cdot\frac{(1-l)^{n-1}}{1-(1-l)^n}\sum_{i=1}^{n}\varphi_l(X_{ij})\right|+\left|1\cdot\frac{1}{1-u}\sum_{i=1}^{n}\varphi_l(X_{ij})\right| \\
&\leq \frac{(1-l)^{n-1}}{1-(1-l)^n}n\phi+\frac{1}{1-u}n\phi \\
&= n\phi\left(\frac{(1-l)^{n-1}}{1-(1-l)^n}+\frac{1}{1-u}\right)\in(0,\infty).
\end{aligned}
$$

Hence

$$\begin{aligned}
&\int k_1 d(\zeta \otimes \mu^{(n)}) \\
&= \int (1-(1-u)^n)(1-l)^n g_n(\mathbf{x}) n\phi \left(\frac{(1-l)^{n-1}}{1-(1-l)^n} + \frac{1}{1-u} \right) (\zeta \otimes \mu^{(n)})(dy, d\mathbf{x}) \\
&= (1-(1-u)^n)(1-l)^n n\phi \left(\frac{(1-l)^{n-1}}{1-(1-l)^n} + \frac{1}{1-u} \right) \int_{\mathscr{X}^n} g_n(\mathbf{x}) \int_{\{0,1\}} \zeta(dy) \mu^{(n)}(d\mathbf{x}) \\
&= 2(1-(1-u)^n)(1-l)^n n\phi \left(\frac{(1-l)^{n-1}}{1-(1-l)^n} + \frac{1}{1-u} \right) < \infty.
\end{aligned}$$

In a similar fashion condition (ii) is shown below, considering

$$l < p_\vartheta \Rightarrow \left[1 - \prod_{i=1}^{n} (1 - \sum_{k=1}^{K} \vartheta_k \varphi_k(X_{ij})) \right]^{-2} < (1-(1-l)^n)^{-2},$$

and

$$p_\vartheta < u \Rightarrow \left(1 - \sum_{k=1}^{K} \vartheta_k \varphi_k(X_{ij}) \right)^{-2} < (1-u)^{-2}.$$

Hence $\forall \vartheta \in \mathcal{U}(\theta)$:

$$\begin{aligned}
\left| \frac{\partial^2}{\partial\vartheta_l \partial\vartheta_m} \mathscr{L}(Y_j^*, \mathbf{X}_j|\vartheta) \right| &< (1-(1-u)^n)(1-l)^n g_n(\mathbf{X}_j) n^2\phi^2 \left(\frac{(1-l)^{n-1}}{1-(1-l)^n} + \frac{1}{1-u} \right)^2 \\
&\quad + (1-(1-u)^n)(1-l)^n g_n(\mathbf{X}_j) \\
&\quad \cdot \left\{ \frac{((1-l)^n - 1)(1-l)^{n-2} n(n-1)\phi^2 + n^2\phi^2(1-l)^{2n-2}}{(1-(1-l)^n)^2} + \frac{n\phi^2}{(1-u)^2} \right\} \\
&= g_n(\mathbf{X}_j) n\phi^2 (1-(1-u)^n)(1-l)^n \left[n \left(\frac{(1-l)^{n-1}}{1-(1-l)^n} + \frac{1}{1-u} \right)^2 \right. \\
&\quad \left. + \frac{((1-l)^n - 1)(1-l)^{n-2}(n-1) + n(1-l)^{2n-2}}{(1-(1-l)^n)^2} + \frac{1}{(1-u)^2} \right] \\
&=: g_n(\mathbf{X}_j)\ c_{\mathscr{L},2}(n, \phi, l, u),
\end{aligned}$$

and thus $\int k_2 d(\zeta \otimes \mu^{(n)}) = 2c_{\mathscr{L},2}(n, \phi, l, u) < \infty$.

The condition (iii) holds since the log-likelihood can be seen as a composition of continuous functions regarding the components of ϑ.
To show condition (iv) it suffices to state that the upper bound

$$\left\{ \frac{((1-l)^n - 1)(1-l)^{n-2} n(n-1)\phi^2 + n^2\phi^2(1-l)^{2n-2}}{(1-(1-l)^n)^2} + \frac{n\phi^2}{(1-u)^2} \right\}$$

holds for the absolute value of each entry since this implies boundedness of the Hessian matrix in finite dimensions.

In order to show condition (v), it is useful to simplify the entries of the log-likelihood using the notation $\psi_\vartheta(\mathbf{X}_j) := \prod_{i=1}^{n} (1 - p_\vartheta(X_{ij}))$ and to calculate the Fisher information exploiting the tower property of conditional expectations and the fact that $\mathbf{X}_j$ is $\sigma(\mathbf{X}_j)$-measurable; moreover the regularity conditions shown above allow for the application of corollary 4.7 as follows:

$$\mathcal{I}(\vartheta)_{l,m} = -\mathbb{E} \frac{\partial^2}{\partial\vartheta_l \partial\vartheta_m} \ell(Y_j^*, \mathbf{X}_j|\vartheta) = -\mathbb{E}\mathbb{E} \left(\frac{\partial^2}{\partial\vartheta_l \partial\vartheta_m} \ell(Y_j^*, \mathbf{X}_j|\vartheta) \Big| \mathbf{X}_j \right) =: \mathbb{E} h_{\vartheta,l,m}(\mathbf{X}_j). \quad (4.19)$$

Here, the second equality holds because of theorem 1.48.

$$
\begin{aligned}
h_{\vartheta,l,m}(\mathbf{X}_j) &= -\sum_{y\in\{0,1\}} \frac{\partial^2}{\partial\vartheta_l\partial\vartheta_m}\ell(y,\mathbf{X}_j|\vartheta)\,\mathrm{Bin}(p_\vartheta^*(\mathbf{X}_j))(\{y\}) \\
&= \psi_\vartheta(\mathbf{X}_j)\sum_{i=1}^n \frac{\varphi_l(X_{ij})}{1-p_\vartheta(X_{ij})}\frac{\varphi_m(X_{ij})}{1-p_\vartheta(X_{ij})} \\
&\quad + (1-\psi_\vartheta(\mathbf{X}_j))^2\left[\sum_{\{(i,r)\in\{1,\dots,n\}^2:i\neq r\}} \varphi_l(X_{ij})\varphi_m(X_{rj}) \prod_{s\in\{1,\dots,n\}\setminus\{i,r\}}(1-p_\vartheta(X_{sj}))\right] \\
&\quad \cdot (1-\psi_\vartheta(\mathbf{X}_j))^{-2} \\
&\quad + (1-\psi_\vartheta(\mathbf{X}_j))\left[\sum_{i=1}^n \varphi_m(X_{ij}) \prod_{r\in\{1,\dots,n\}\setminus\{i\}}[1-p_\vartheta(X_{rj})]\right] \\
&\quad \cdot \left[\sum_{i=1}^n \varphi_l(X_{ij}) \prod_{r\in\{1,\dots,n\}\setminus\{i\}}[1-p_\vartheta(X_{rj})]\right](1-\psi_\vartheta(\mathbf{X}_j))^{-2} \\
&= \psi_\vartheta(\mathbf{X}_j)\sum_{i=1}^n \frac{\varphi_l(X_{ij})}{1-p_\vartheta(X_{ij})}\frac{\varphi_m(X_{ij})}{1-p_\vartheta(X_{ij})} + \psi_\vartheta(\mathbf{X}_j)\sum_{\{(i,r)\in\{1,\dots,n\}^2:i\neq r\}} \frac{\varphi_l(X_{ij})}{1-p_\vartheta(X_{ij})}\frac{\varphi_m(X_{rj})}{1-p_\vartheta(X_{rj})} \\
&\quad + \frac{\psi_\vartheta(\mathbf{X}_j)^2}{1-\psi_\vartheta(\mathbf{X}_j)}\left[\sum_{i=1}^n \frac{\varphi_m(X_{ij})}{1-p_\vartheta(X_{ij})}\right]\left[\sum_{i=1}^n \frac{\varphi_l(X_{ij})}{1-p_\vartheta(X_{ij})}\right] \\
&= \psi_\vartheta(\mathbf{X}_j)\sum_{i=1}^n\sum_{r=1}^n \frac{\varphi_l(X_{ij})}{1-p_\vartheta(X_{ij})}\frac{\varphi_m(X_{rj})}{1-p_\vartheta(X_{rj})} + \frac{\psi_\vartheta(\mathbf{X}_j)^2}{1-\psi_\vartheta(\mathbf{X}_j)}\left[\sum_{i=1}^n \frac{\varphi_m(X_{ij})}{1-p_\vartheta(X_{ij})}\right]\left[\sum_{i=1}^n \frac{\varphi_l(X_{ij})}{1-p_\vartheta(X_{ij})}\right]
\end{aligned}
$$

It can thus be seen that l and m play symmetric roles in each entry of the Fisher information, i.e. it is symmetric and therefore self-adjoint. To prove that it is positive definite it remains to be shown that for any $v\in\mathbb{R}^K\setminus\{0\}$ the number $\langle\mathcal{I}(\vartheta)v,v\rangle$ is positive (compare Duistermaat and Kolk (2004)). In order to do so, the linearity of the expectation can be exploited and the sums will be reordered.

$$
\begin{aligned}
\langle\mathcal{I}(\vartheta)v,v\rangle &= \sum_{l=1}^K\sum_{m=1}^K v_l v_m \mathbb{E}_{G_n} h_{\vartheta,l,m}(\mathbf{X}_j) \\
&= \mathbb{E}_{G_n}\psi_\vartheta(\mathbf{X}_j)\sum_{l=1}^K\sum_{m=1}^K v_l v_m \sum_{i=1}^n\sum_{r=1}^n \frac{\varphi_l(X_{ij})}{1-p_\vartheta(X_{ij})}\frac{\varphi_m(X_{rj})}{1-p_\vartheta(X_{rj})} \\
&\quad + \mathbb{E}_{G_n}\frac{\psi_\vartheta(\mathbf{X}_j)^2}{1-\psi_\vartheta(\mathbf{X}_j)}\sum_{l=1}^K\sum_{m=1}^K v_l v_m \left[\sum_{i=1}^n \frac{\varphi_m(X_{ij})}{1-p_\vartheta(X_{ij})}\right]\left[\sum_{i=1}^n \frac{\varphi_l(X_{ij})}{1-p_\vartheta(X_{ij})}\right] \\
&= \mathbb{E}_{G_n}\psi_\vartheta(\mathbf{X}_j)\left[\sum_{l=1}^K v_l \sum_{i=1}^n \frac{\varphi_l(X_{ij})}{1-p_\vartheta(X_{ij})}\right]\sum_{m=1}^K v_m \sum_{r=1}^n \frac{\varphi_m(X_{rj})}{1-p_\vartheta(X_{rj})} \\
&\quad + \mathbb{E}_{G_n}\frac{\psi_\vartheta(\mathbf{X}_j)^2}{1-\psi_\vartheta(\mathbf{X}_j)}\sum_{l=1}^K v_l \left[\sum_{i=1}^n \frac{\varphi_l(X_{ij})}{1-p_\vartheta(X_{ij})}\right]\sum_{m=1}^K v_m \left[\sum_{i=1}^n \frac{\varphi_m(X_{ij})}{1-p_\vartheta(X_{ij})}\right] \\
&= \mathbb{E}_{G_n}\frac{\psi_\vartheta(\mathbf{X}_j)}{1-\psi_\vartheta(\mathbf{X}_j)}\left[\sum_{l=1}^K v_l \sum_{i=1}^n \frac{\varphi_l(X_{ij})}{1-p_\vartheta(X_{ij})}\right]^2
\end{aligned}
$$

Note that $\forall\vartheta\in\mathcal{U}(\theta):\frac{\psi_\vartheta}{1-\psi_\vartheta} > \frac{(1-u)^n}{1-(1-u)^n} > 0$ since $u\in(0,1)$ and therefore, the monotonicity of the integral implies $\langle\mathcal{I}(\vartheta)v,v\rangle\geq 0$, i.e. $\mathcal{I}(\vartheta)$ is positive semidefinite. To ensure strict positivity

remember that $l \in (0,1)$ and thus $\frac{1}{1-p_\vartheta} > \frac{1}{1-l}$. Now bound $\langle \mathcal{I}(\vartheta)v, v\rangle$ and apply corollary 1.38:

$$\begin{aligned}\langle \mathcal{I}(\vartheta)v, v\rangle &\geq \frac{(1-u)^n}{1-(1-u)^n} \cdot \frac{1}{(1-l)^2}\, \mathbb{E}_{G_n}\left[\sum_{l=1}^K v_l \sum_{i=1}^n \varphi_l(X_{ij})\right]^2 \\ &= \frac{(1-u)^n}{1-(1-u)^n} \cdot \frac{1}{(1-l)^2}\, \mathbb{E}_{G_n} \langle v, \Phi(\mathbf{X}_j)\rangle^2,\end{aligned}$$

if we define

$$\Phi : \mathscr{X}^n \to \mathbb{R}^K, \quad \mathbf{x} \mapsto \begin{pmatrix} \sum_{i=1}^n \varphi_1(x_i) \\ \vdots \\ \sum_{i=1}^n \varphi_K(x_i) \end{pmatrix}. \tag{4.20}$$

Assuming this number to be zero implies $\langle v, \Phi\rangle = 0$, G_n-a.s. If we impose the condition that $\langle v, \Phi\rangle$ cannot be zero G_n-a.s. on the basis functions, we can conclude by contradiction that $\mathbb{E}_{G_n}\langle v, \Phi(\mathbf{X}_j)\rangle^2 \neq 0$ and thus $\mathbb{E}_{G_n}\langle v, \Phi(\mathbf{X}_j)\rangle^2 > 0$, in other words this condition ensures $\mathcal{I}(\vartheta)$ to be positive definite. This condition can be further simplified applying lemma 1.42; it should be noted that for fixed j, l the random variables $v_l\varphi_l(X_{1j}), \ldots, v_l\varphi_l(X_{nj})$ are even iid.

$$\begin{aligned}&\mathbb{E}_{G_n}\left[\sum_{l=1}^K v_l \sum_{i=1}^n \varphi_l(X_{ij})\right]^2 = \mathbb{V}_{G_n}\sum_{l=1}^K v_l \sum_{i=1}^n \varphi_l(X_{ij}) + \left[\mathbb{E}_{G_n}\sum_{l=1}^K v_l \sum_{i=1}^n \varphi_l(X_{ij})\right]^2 \\ &= \mathbb{V}_{G_n}\sum_{i=1}^n\sum_{l=1}^K v_l\varphi_l(X_{ij}) + \left[\sum_{i=1}^n \mathbb{E}_G \sum_{l=1}^K v_l\varphi_l(X_{ij})\right]^2 = n\mathbb{V}_G\sum_{l=1}^K v_l\varphi_l(X_{1j}) + n^2\left[\mathbb{E}_G\sum_{l=1}^K v_l\varphi_l(X_{1j})\right]^2 \\ &\geq n\mathbb{V}_G\sum_{l=1}^K v_l\varphi_l(X_{1j}) + n\left[\mathbb{E}_G\sum_{l=1}^K v_l\varphi_l(X_{1j})\right]^2 = n\mathbb{E}_G\left[\sum_{l=1}^K v_l\varphi_l(X_{1j})\right]^2 \\ &= n\mathbb{E}_G\sum_{l=1}^K\sum_{m=1}^K v_l v_m\varphi_l(X_{1j})\varphi_m(X_{1j}) = n\sum_{l=1}^K\sum_{m=1}^K v_l v_m \mathbb{E}_G\varphi_l(X_{1j})\varphi_m(X_{1j}) \\ &\overset{\bullet}{=} n\sum_{l=1}^K\sum_{m=1}^K v_l v_m \mathbb{E}_G\varphi_l(X_{1j})\varphi_m(X_{1j})\end{aligned}$$

Now let $\operatorname{supp}(g) =: \mathscr{G}$ and assume that g is bounded from below on its support, i.e. $\forall x \in \mathscr{G} : g(x) \geq c_g > 0$. To ensure the term in equality $\bullet$ is strictly bounded from below, a sufficient condition is that $\mathbb{E}_G\varphi_l(X_{1j})\varphi_m(X_{1j}) = \delta_{l,m}$, i.e. the $\varphi_1, \ldots, \varphi_K$ form an ONS G-a.s. Alternatively, we now assume the $\varphi_1, \ldots, \varphi_K$ to be supported on $\mathscr{G}$ or $\mathscr{G} = \mathscr{X}$ and to form an ONS. This assumption is reasonable since the regression function cannot be estimated outside $\mathscr{G}$. With the latter assumption the inequality can be further simplified:

$$\begin{aligned}\mathbb{E}_{G_n}\left[\sum_{l=1}^K v_l \sum_{i=1}^n \varphi_l(X_{ij})\right]^2 &\geq n\sum_{l=1}^K\sum_{m=1}^K v_l v_m \mathbb{E}_G\varphi_l(X_{1j})\varphi_m(X_{1j}) \\ &= nc_g\sum_{l=1}^K\sum_{m=1}^K v_l v_m \int_{\mathscr{G}} \varphi_l(x)\varphi_m(x)\mu(dx) \\ &= nc_g\sum_{l=1}^K\sum_{m=1}^K v_l v_m\delta_{l,m} = nc_g\sum_{l=1}^K v_l^2 \\ &= nc_g\|v\|^2 > 0.\end{aligned}$$

Theorem 4.15 (BvM for a Finite-Dimensional Subspace in Group Testing)
Take p_ϑ as defined in (4.18) such that it maps to $[0,1]$ for any $\vartheta \in \mathcal{U}(\theta)$, and let it be bounded such that $\forall x \in \mathscr{X} : 0 < l < p_\vartheta(x) < u < 1$.
Assume the density g of the covariates to be supported on $\mathscr{G} = \mathscr{X}$, and to be bounded from below, i.e. $\forall x \in \mathscr{G} : g(x) \geq c_g > 0$.
If, furthermore, the basis functions $\varphi_1, \ldots, \varphi_K$ are uniformly bounded by $\phi < \infty$, and if they form an ONS on $L^2(\mathscr{X}, \mathfrak{X}, \mu)$, all prerequisites of theorem 4.14 are met.

Remark 4.16. Again, we assume the natural prior mass condition that is required for consistency.

5 Discussion, Conclusions and Possible Extensions

In section 2.1 we have given an overview of group testing regression models and various approaches to derive estimators and their consistency. Section 2.2 provided us with tools and conditions to study the asymptotic distributional behaviour of different Bayesian models as they were introduced in section 1.2.

In section 3 we have combined methodology from sections 2.1 and 1.2 to derive a posterior distribution for a group testing regression problem when only pooled data is considered. As a second step one may additionally model the individual testing stage, which we have omitted. Since our study of the posterior in section 4.1 relies on conditionally independent and identically distributed observations, the last paragraph of section 3 is dedicated to restrictions on the posterior model. Indeed, pooling samples of equally sized groups does not seem far-fetched in practice and is implicitly assumed by Dorfman (1943) in his paper. One might, however argue, that taking the marginal distribution of the covariate to be dominated by the Lebesgue measure may be somewhat of a simplification, and in fact Delaigle et al. (2014) propose a model with discrete and continuous covariates. If this really poses a problem, might depend on the application and the particular model at hand. Age, for instance, is a continuous value that may very well be made discrete by clustering it (e.g age in years), while other values, such as cigarettes smoked, while being discrete, might be considered approximately continuous.

Section 4.1 builds on section 2.2 and develops sufficient conditions for a strong Laplace-Bernstein-von Mises theorem, which is then combined with the insights of section 3 to propose a strong BvM theorem for a finite-dimensional projection of the functional Bernoulli parameter, which, to the best of our knowledge is the first of its kind for this particular kind of problem.

Desirable extensions of theorem 4.15 could be the consideration of a Bayesian nonparametric estimation of p instead of the finite-dimensional projection p_ϑ. For such an experiment, a few problems would arise in the argumentation we have given. For instance, taking k (section 4.1)/K (section 4.15) to diverge, one could no longer reason with dominated convergence in (4.12) for the "*outside-of-*Θ_1*-part*" of $\int \ln(h) h d\nu$. Additionally, the almost sure convergence of the term (4.17) is no longer guaranteed without restrictions: the divergence of k must be bounded in order to be compensated by the convergence of the random variable $\tilde{H}$, i.e. the rate of convergence must be determined or another solution has to be found.

Moreover, one might want to modify theorem 4.14, such that it becomes *fully Bayesian*, by allowing the convergence to be $\mathbb{P}^\vartheta \otimes Q$-a.s., or, similarly, in $\mathbb{P}^\vartheta \otimes Q$-probability, which means to relax the consistency of ϑ_n to only hold for Q-almost all $\theta \in \Theta$ as proposed in definition 4.1. In this sense, the prior consistency condition is not required as an extra, but as Bayes theorem 1.51 tells us, the posterior is dominated by the prior, and therefore prior consistency implies posterior consistency in a fully Bayesian approach.

Another restriction that might be lifted is the shape of Θ_1. While our definition still allows arbitrary shapes of priors, one might modify the support of the prior to contain Θ_1 rather than to be Θ_1. While this shouldn't be a big problem, the restriction was made for technical reasons, similar to restrictions in Le Cam and Lo Yang (2000).

Last but not least, for practical purposes it could be useful to model $p_\vartheta = F \circ \xi_\vartheta$ for a distribution function $F : \mathbb{R}^1 \to (0,1)$ as proposed by Ghosal and van der Vaart (2017), for example. If the link function F is continuous (with respect to the Lebesgue measure), which is usually the case, conditions (i), (ii), (iii), and (iv) in section 4.15 follow immediately from the proof already given. Regarding condition (v) the calculation will be slightly more complex and one would derive a different set of conditions on the φ_k.

Furthermore, one could relax the nature of consistency of the MLE to be weaker, i.e. convergence in probability.

Other interesting aspects are the contraction rates, where upper, lower bounds, as well as ex-

act rates would be of interest, in particular to justify approximations in practical applications. Regarding practical applications computational aspects and concrete algorithms could be developed.

References

Yechao Bai, Qingsi Wang, Chun Lo, Mingyan Liu, Jerome P. Lynch, and Xinggan Zhang. Adaptive Bayesian group testing: Algorithms and performance. *Signal Processing*, 156:191–207, March 2019. ISSN 0165-1684. doi: 10.1016/j.sigpro.2018.11.006.

Robert Gardner Bartle. *Introduction to Real Analysis*. Wiley, 2011. ISBN 9780471433316.

Vladimir I. Bogachev. *Measure Theory*. Springer Berlin Heidelberg, 2007. ISBN 9783540345145. doi: 10.1007/978-3-540-34514-5.

Vladimir I. Bogachev. *Gaussian Measures*. American Mathematical Society, 2015. ISBN 9781470418694.

Vladimir I. Bogachev and Oleg G. Smolyanov. *Real and Functional Analysis*. Springer International Publishing, 2020. doi: 10.1007/978-3-030-38219-3. URL `https://doi.org/10.1007/978-3-030-38219-3`.

Georg Bol. *Induktive Statistik. Lehr- und Arbeitsbuch.* Oldenbourg, 2002. ISBN 9783486272765.

Dominique Bontemps. Bernstein von Mises theorems for Gaussian regression with increasing number of regressors. *Annals of Statistics 39, 5 (2011) 2557-2584*, September 2011. doi: 10.1214/11-AOS912.

J. Bretagnolle and C. Huber. *Estimation des densités: Risque minimax*, pages 342–363. Springer Berlin Heidelberg, 1978. ISBN 9783540358565. doi: 10.1007/bfb0064610.

T.A. Bryan and M. Gershman. A semi-quantitative test for water supply. *Poultry Science*, 54 (6):2136–2137, November 1975. ISSN 0032-5791. doi: 10.3382/ps.0542136.

Clément L. Canonne. A short note on an inequality between kl and tv. February 2022. URL `http://arxiv.org/abs/2202.07198v2`.

Henri Paul Cartan. *Differential Calculus*. Houghton Mifflin Co, 1971. ISBN 9780395120330.

Ismaël Castillo. *Bayesian nonparametrics, convergence and limiting shape of posterior distributions*. Habilitation à diriger des recherches, Université Paris Diderot Paris 7, November 2014. URL `https://theses.hal.science/tel-01096755`.

Ismaël Castillo. *Bayesian nonparametric statistics, St-Flour lecture notes*. February 2024. URL `https://arxiv.org/pdf/2402.16422v1`.

Ismaël Castillo and Richard Nickl. Nonparametric Bernstein-von Mises theorems in Gaussian white noise. *Annals of Statistics 2013, Vol. 41, No. 4, 1999-2028*, 41(4), August 2013. doi: 10.1214/13-AOS1133.

Ismaël Castillo and Richard Nickl. On the Bernstein-von Mises phenomenon for nonparametric Bayes procedures. *Annals of Statistics 2014, Vol. 42, No. 5, 1941-1969*, 42(5), October 2014. doi: 10.1214/14-AOS1246.

Ismaël Castillo and Judith Rousseau. A Bernstein-von Mises theorem for smooth functionals in semiparametric models. *Annals of Statistics 2015, Vol. 43, No. 6, 2353-2383*, May 2013. doi: 10.1214/15-AOS1336.

Ismaël Castillo, Johannes Schmidt-Hieber, and Aad van der Vaart. Bayesian linear regression with sparse priors. *Annals of Statistics 2015, Vol. 43, No. 5, 1986-2018*, March 2015. doi: 10.1214/15-AOS1334.

Siddhartha Chib and Edward Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, nov 1995. doi: 10.1080/00031305.1995.10476177.

Nidhan Choudhuri, Subhashis Ghosal, and Anindya Roy. Bayesian estimation of the spectral density of a time series. *Journal of the American Statistical Association*, 99(468):1050–1059, December 2004. ISSN 1537-274X. doi: 10.1198/016214504000000557.

K. L. Chung and R. J. Williams. *Introduction to Stochastic Integration*. Birkhäuser Boston, 1990. ISBN 9781461244806. doi: 10.1007/978-1-4612-4480-6.

A. Delaigle and P. Hall. Nonparametric methods for group testing data, taking dilution into account. *Biometrika*, 102(4):871–887, nov 2015. doi: 10.1093/biomet/asv049.

A. Delaigle, P. Hall, and J. R. Wishart. New approaches to nonparametric and semiparametric regression for univariate and multivariate group testing data. *Biometrika*, 101(3):567–585, aug 2014. doi: 10.1093/biomet/asu025.

Aurore Delaigle and Peter Hall. Defining probability density for a distribution of random functions. *The Annals of Statistics*, 38(2):1171–1193, apr 2010. doi: 10.1214/09-aos741.

Aurore Delaigle and Peter Hall. Nonparametric regression with homogeneous group testing data. *The Annals of Statistics*, 40(1):131–158, feb 2012. doi: 10.1214/11-aos952.

Aurore Delaigle and Alexander Meister. Nonparametric regression analysis for group testing data. *Journal of the American Statistical Association*, 106(494):640–650, Jun 2011. doi: 10.1198/jasa.2011.tm10520.

Robert Dorfman. The detection of defective members of large populations. *The Annals of Mathematical Statistics*, 14(4):436–440, December 1943. ISSN 0003-4851. doi: 10.1214/aoms/1177731363.

R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, October 2002. ISBN 9780511755347. doi: 10.1017/cbo9780511755347.

R. M. Dudley. *Uniform Central Limit Theorems Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 2014. ISBN 9780521738415.

J. J. Duistermaat and J. A. C. Kolk. *Multidimensional Real Analysis I*. Cambridge University Press, May 2004. doi: 10.1017/cbo9780511616716. URL `https://doi.org/10.1017/CBO9780511616716`.

Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, April 2019. ISBN 9781108473682. doi: 10.1017/9781108591034.

J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications*. Routledge, May 2018. ISBN 9780203748725. doi: 10.1201/9780203748725.

Thomas S. Ferguson. Prior distributions on spaces of probability measures. *The Annals of Statistics*, 2(4), July 1974. ISSN 0090-5364. doi: 10.1214/aos/1176342752.

S. E. M. P. Franssen and A. W. van der Vaart. The Bernstein-von Mises theorem for the Pitman-Yor process of nonnegative type. February 2021. URL `https://arxiv.org/pdf/2102.06059v2`.

David A. Freedman. On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, 34(4):1386–1403, December 1963. ISSN 0003-4851. doi: 10.1214/aoms/1177703871.

Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, jun 2017. doi: 10.1017/9781139029834.

Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, nov 2015. doi: 10.1017/cbo9781107337862.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer New York, 2009. ISBN 9780387848587. doi: 10.1007/978-0-387-84858-7.

Norbert Henze. *Asymptotische Stochastik: Eine Einführung mit Blick auf die Statistik*. Springer Berlin Heidelberg, 2024. ISBN 9783662684467. doi: 10.1007/978-3-662-68446-7.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, October 2012. ISBN 9780521548236. doi: 10.1017/cbo9781139020411.

Tailen Hsing and Randall Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley, May 2015. ISBN 9781118762547. doi: 10.1002/9781118762547.

Wolfgang Härdle, Hua Liang, and Jiti Gao. *Partially Linear Models*. Physica-Verlag HD, 2000. ISBN 9783642577000. doi: 10.1007/978-3-642-57700-0.

I. A. Ibragimov and R. Z. Has'minskii. *Statistical Estimation*. Springer New York, 1981. ISBN 9781489900272. doi: 10.1007/978-1-4899-0027-2.

Olav Kallenberg. *Foundations of Modern Probability*. Springer International Publishing, 2021. doi: 10.1007/978-3-030-61871-1.

B. J. K. Kleijn and A. W. van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics*, 34(2), April 2006. ISSN 0090-5364. doi: 10.1214/009053606000000029.

B.J.K. Kleijn and A.W. van der Vaart. The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6(none), January 2012. ISSN 1935-7524. doi: 10.1214/12-ejs675.

Achim Klenke. *Probability Theory: A Comprehensive Course*. Springer International Publishing, 2020. ISBN 9783030564025. doi: 10.1007/978-3-030-56402-5.

Jonathan Kunick. Metropolis, Metropolis-Hastings, Gibbs, or yet another sampler? Sampling with application to Bayesian estimation of several parameters in survival analysis. Master's thesis, Universität Rostock, 2018.

Lucien Le Cam. On the Bernstein-von Mises theorem. Technical report, Department of Statistics, University of California, Berkley, 1986a.

Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer New York, 1986b. ISBN 9781461249467. doi: 10.1007/978-1-4612-4946-7.

Lucien Le Cam and Grace Lo Yang. *Asymptotics in Statistics*. Springer New York, 2000. ISBN 9781461211662. doi: 10.1007/978-1-4612-1166-2.

Alexandra Martin, Alexandre Storto, Quentin Le Hingrat, Gilles Collin, Barbara André, Allison Mallory, Rémi Dangla, Diane Descamps, Benoit Visseaux, and Olivier Gossner. High-sensitivity sars-cov-2 group testing by digital pcr among symptomatic patients in hospital settings. *Journal of Clinical Virology*, 141:104895, August 2021. ISSN 1386-6532. doi: 10.1016/j.jcv.2021.104895.

Hiroyasu Matsushima, Yusuke Tajima, Xiao-Nan Lu, and Masakazu Jimbo. Efficient pooling designs and screening performance in group testing for two type defectives. May 2024.

C. S. McMahan, J. M. Tebbs, and C. R. Bilder. Regression models for group testing data with pool dilution effects. *Biostatistics*, 14(2):284–298, November 2012. ISSN 1468-4357. doi: 10.1093/biostatistics/kxs045.

Christopher S. McMahan, Joshua M. Tebbs, Timothy E. Hanson, and Christopher R. Bilder. Bayesian regression for group testing data. *Biometrics*, 73(4):1443–1452, April 2017. ISSN 1541-0420. doi: 10.1111/biom.12704.

Bo Ning, Seonghyun Jeong, and Subhashis Ghosal. Bayesian linear regression for multivariate responses under group sparsity. July 2018.

Eswar G. Phadia. *Prior Processes and Their Applications*. Springer International Publishing, 2016. ISBN 9783319327891. doi: 10.1007/978-3-319-32789-1.

Christian P. Robert. *The Bayesian Choice*. Springer, 2007. ISBN 9780387715988.

Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367, nov 2001. doi: 10.1214/ss/1015346320.

G.O. Roberts and A.F.M. Smith. Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and their Applications*, 49(2):207–216, feb 1994. doi: 10.1016/0304-4149(94)90134-1.

Judith Rousseau. On the frequentist properties of Bayesian nonparametric methods. *Annual Review of Statistics and Its Application*, 3(1):211–231, June 2016. ISSN 2326-831X. doi: 10.1146/annurev-statistics-041715-033523.

Mark J. Schervish. *Theory of Statistics*. Springer New York, 1995. doi: 10.1007/978-1-4612-4250-5. URL `https://doi.org/10.1007/978-1-4612-4250-5`.

Jun Shao. *Mathematical statistics*. Springer, 2003. ISBN 0387953825.

David J. Smith and Mavina K. Vamanamurthy. How small is a unit ball? *Mathematics Magazine*, 62(2):101–107, April 1989. ISSN 1930-0980. doi: 10.1080/0025570x.1989.11977419.

Gunnar Taraldsen. Optimal learning from the Doob-Dynkin lemma. January 2018. URL `https://arxiv.org/pdf/1801.00974v1`.

Curtis Tatsuoka, Weicong Chen, and Xiaoyi Lu. Bayesian group testing with dilution effects. *Biostatistics*, 24(4):885–900, April 2022. ISSN 1468-4357. doi: 10.1093/biostatistics/kxac004.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer New York, 2009. ISBN 9780387790527. doi: 10.1007/b13794.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Oct 1998. doi: 10.1017/cbo9780511802256.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer New York, 1996. ISBN 9781475725452. doi: 10.1007/978-1-4757-2545-2.