

# FOUNDATIONS OF STRUCTURAL STATISTICS: TOPOLOGICAL STATISTICAL THEORY

P. MICHL <PATRICK.MICHL@GMAIL.COM>

**ABSTRACT.** Topological Statistical Theory, provides the foundation for a new understanding of classical Statistics: Structural Statistics, which emphasizes intrinsically structured model spaces and structure preserving transformations as the central objects and morphisms of respective categories. The resulting language not only turns out to be highly compatible with classical statistical theory, but indeed outperforms it in simplicity and elegance for complicated model spaces. Maybe the most important present showcases for this new branch of statistics are therefore Machine-Learning and in particular Deep-Learning. There above it concerns topological-, geometric- as well as algebraic data analysis, which respectively derive statistical estimations, by the assumption of simplicial complexes, Riemannian manifolds and algebraic varieties.

## 1. INTRODUCTION

As a passionate advocate of observation based statistical inference in the first third of the 20<sup>th</sup> century RONALD A. FISHER introduced the concepts of *sufficient statistics*, as well as a measure of information, which is invariant w.r.t. them [2]. This *Fisher information* encouraged S. KULLBACK and R. A. LEIBLER in the mid-20<sup>th</sup> century to further investigations, regarding the correspondence between statistics and the uprising information theory of CLAUDE E. SHANNON [3]. In the course of their analysis a further invariant structure, today known as the *Kullback-Leibler divergence* [4] became apparent, which permitted to study the invariance property in an information theoretic context. As a consequence the rising theory of statistical *divergence functions*, increasingly provided an applicable framework to study the equivalence of statistical models in terms of reversible Markov processes [1, 5]. A generic theory of equivalence and invariance of statistical models, however, still is missing and first of all requires to substantiate the structures, that have to be conserved. The subsequent sections provide the foundation for a

topological statistical theory, by successively introducing the *Equivalence*, the *Topology* and the structural *Homology* of statistical models.

## 2. STATISTICAL EQUIVALENCE AND THE CATEGORY OF STATISTICAL MODELS

**Definition** (Statistical model). *Let  $(S, \Sigma)$  be a measurable space (the sample space),  $T: (\Omega, \mathcal{F}, P) \rightarrow (S, \Sigma)$  an i.i.d. random variable (a statistic) and  $\mathcal{M}$  a set of probability distributions over  $(S, \Sigma)$  (the model space). Then the 3-tuple  $(S, \Sigma, \mathcal{M})$  is termed a statistical model.*

The choice of a statistical model is generally subjected to the convenience of the experimenter and therefore far away from being unique. This ambiguity concerns the underlying sample space and statistic, by the applied physical measurement method, as well as the model space, by the interpretation of the underlying theory. As observation based statistical inference however is obligated to provide unique conclusions, that do not depend on these choices a canonical equivalence relationship between statistical models, referred as *observational equivalence*. Thereby observational equivalence describes the circumstance, that no possible observation of the underlying statistical population may be



In order to avoid an inconsistent notation the conditional probabilities are therefore represented by *transition operators* and afterwards generalized to a continuous notation.

**Lemma 1.** *Let  $(X, \mathcal{A}, \mu)$  be a discrete probability space with  $\mathcal{A} = \mathcal{P}(X)$  and  $(Y, \mathcal{B})$  a discrete measurable space with  $\mathcal{B} = \mathcal{P}(Y)$ . Let further be*

$$T: (X, \mathcal{A}) \rightarrow (Y, \mathcal{B})$$

*a measurable function, which induces the transition operator  $\tau$  by  $\tau(\mu) = T_*\mu$  and  $P(y | x)$  the conditional probabilities from  $X$  to  $Y$ , then:*

$$(2.1) \quad P(y | x) = \tau(\delta_x)(y), \forall (x, y) \in X \times Y$$

*Proof.* For  $x \in X$  let  $\delta_x$  denote the Dirac measure over  $(X, \mathcal{A})$ , then:

$$(2.2) \quad \tau(\delta_x)(T(x)) = \sum_{a \in X} \tau(\delta_x)(T(a))\delta_x(\{a\})$$

Furthermore by applying the law of total probability, it also holds that :

$$(2.3) \quad \tau(\delta_x)(T(x)) = \sum_{a \in X} P(T(x) | a)\delta_x(\{a\})$$

A subsequent substitution of equation 2.3 in 2.2 yields:

$$(2.4) \quad \sum_{a \in X} P(T(x) | a)\delta_x(\{a\}) = \sum_{a \in X} \tau(\delta_x)(T(a))\delta_x(\{a\})$$

And since  $\delta_x(\{a\}) = 1 \Leftrightarrow x = a$  equation 2.4 dissolves to a representation:

$$P(T(x) | x) = \tau(\delta_x)(T(x))$$

This proves Lemma 1 for  $y = T(x)$ . Since  $P(y | x) = 0$  for  $y \notin T(x)$  and

$$P(y_1 \cup y_2 | x) = P(y_1 | x)P(y_2 | x)$$

for  $y_1 \cap y_2 = \emptyset$  the generic case follows by complete induction over the discrete product measurable space  $(X \times Y, \mathcal{A} \otimes \mathcal{B})$ .  $\square$

**Definition** (Regular conditional probability). *Let  $(X, \mathcal{A}, \mu)$  be a probability space,  $(Y, \mathcal{B})$  a measurable space and  $T: (X, \mathcal{A}) \rightarrow (Y, \mathcal{B})$  a measurable function, which induces the transition operator  $\tau$  by  $\tau(\mu) = T_*\mu$ . Then the regular conditional probability from  $X$  to  $Y$  is given by:*

$$(2.5) \quad p(y | x) := \frac{d\tau(\delta_x)}{d\tau(\mu)}(y), \forall (x, y) \in X \times Y$$

Due to the definition regular conditional probabilities resolve the contradiction of infinitesimal measures by the properties of transition operators and therefore essentially by the Radon-Nikodym derivative. Although the definition only regards the case of a conditional probability  $p(y | x)$  from  $X$  to  $Y$  it implicates a dual conditional probability  $p(x | y)$  from  $Y$  to  $X$ , which existence and uniqueness is postulated by the following lemma.

**Lemma 2** (Dual conditional probability). *Let  $(X, \mathcal{A}, \mu)$  be a probability space,  $(Y, \mathcal{B})$  a measurable space and*

$$T: (X, \mathcal{A}) \rightarrow (Y, \mathcal{B})$$

*a measurable function, which induces the transition operator  $\tau$  by  $\tau(\mu) = T_*\mu$ . Then the regular conditional probability from  $Y$  to  $X$  is a.e. uniquely given by:*

$$(2.6) \quad p(x | y) \stackrel{a.e.}{=} \frac{\mu(x)}{\tau(\mu)(y)} \frac{d\tau(\delta_x)}{d\tau(\mu)}(y), \forall (x, y) \in X \times Y$$

*Proof.* Let  $\nu := T_*\mu$ , then Lemma 1 postulates the existence of a transition operator  $\tau$  that satisfies  $\tau(\mu) = \nu$ . By applying the definition of the regular conditional probability  $p(y | x)$ , a probability measure  $\pi$  over the product measurable space  $(X \times Y, \mathcal{A} \otimes \mathcal{B})$  is given by:

$$(2.7) \quad \pi(x, y) = \frac{d\tau(\delta_x)}{d\nu}(y)\mu(x)$$

Conversely Lemma 1 also postulates the existence of a further transition operator  $\tau^*$ , that satisfies  $\tau^*(\nu) = \mu$ . Then  $\tau^*$  induces a probability measure  $\pi^*$  over the dual product measurable space  $(Y \times$

$X, \mathcal{B} \otimes \mathcal{A}$ ) by:

$$(2.8) \quad \pi^*(y, x) = \frac{d\tau^*(\delta_y)}{d\mu}(x)\nu(y)$$

Let  $(\pi^*)^*$  be the dual product measure of  $\pi^*$ , then  $\pi = (\pi^*)^*$  for  $\pi$ -almost all  $(x, y) \in X \times Y$ . Let now be  $p(x | y)$  be the regular conditional probability, which is given by  $\tau^*$ , then the substitution of equations 2 and 2.7 yields a representation:

$$p(x | y) = \frac{d\tau^*(\delta_y)}{d\mu}(x) = \frac{\mu(x)}{\nu(y)} \frac{d\tau(\delta_x)}{d\nu}(y)$$

The property of  $\tau$  to be unique for  $\mu$ -almost all  $x \in X$  and of  $\tau^*$  to be unique for  $\nu$ -almost all  $y \in Y$  prove that  $p(x | y)$  is a.e. unique. Equation 2.6 therefore follows by the identity  $\nu = \tau(\mu)$ .  $\square$

**Theorem 3** (*Continuous Bayes' Theorem*). *Let  $(X, \mathcal{A}, \mu)$  and  $(Y, \mathcal{B}, \nu)$  be probability spaces with  $\mu \preceq \nu$  and  $T: (X, \mathcal{A}) \rightarrow (Y, \mathcal{B})$  a measurable function with  $\nu = T_*\mu$ . Then for  $x, y \in X \times Y$ :*

$$(2.9) \quad p(x | y) \stackrel{a.e.}{=} \frac{\mu(x)}{\nu(y)} p(y | x)$$

*Proof.* By applying Lemma 1 and Lemma 2 it immediate follows that for  $\nu$ -almost all  $(x, y) \in X \times Y$  it holds that:

$$p(x | y) \stackrel{2}{=} \frac{\mu(x)}{\tau(\mu)(y)} \frac{d}{d\tau(\mu)} \tau(\delta_x)(y) \stackrel{1}{=} \frac{\mu(x)}{\nu(y)} p(y | x)$$

$\square$

Let now  $(X, \mathcal{A}, \{\mu\})$  and  $(Y, \mathcal{B}, \{\nu\})$  be singleton statistical models of  $(\Omega, \mathcal{F}, P)$  and  $Z = \{T_1, T_2\}$  a Markov process which is given by the underlying statistics

$$T_X: (\Omega, \mathcal{F}, P) \rightarrow (X, \mathcal{A})$$

and

$$T_Y: (\Omega, \mathcal{F}, P) \rightarrow (Y, \mathcal{B})$$

Then the measurable function  $T = T_Y \circ T_X^{-1}$  allows an application of Bayes' theorem which in turn assures the existence of an equilibrium distribution  $\pi = \mu\nu$  and therefore  $Z$  to be reversible. However for non-singleton statistical models  $(X, \mathcal{A}, \mathcal{M})$  and

$(Y, \mathcal{B}, \mathcal{N})$  the transition operator, which is induced by  $T$  depends on the choice of  $P \in \mathcal{M}$ , by the condition:

$$\tau_P(P) = T_*P$$

Consequently also the conditional probability from  $X$  to  $Y$  which is induced by  $P$  depends on this choice, such that:

$$(2.10) \quad p_P(y | x) := \frac{d\tau_P(\delta_x)}{d\tau_P(P)}(y), \forall (x, y) \in X \times Y$$

As the Markov process, however is completely determined by the underlying conditional probabilities, the existence of a Markov process, that induces  $\mathcal{N}$  from  $\mathcal{M}$  requires  $p_P(y | x)$  and therefore  $\tau_P$  to be independent of the choice of  $P$ . Then there has to exist a transition operator  $\tau$  with  $\tau(\mathcal{M}) = \mathcal{N}$  that induces a fixed conditional probability  $p(y | x)$ . The dual conditional probability from  $Y$  to  $X$ , however still depends on the choice of  $P$ , by the continuous Bayes Theorem:

$$p_P(x | y) := \frac{P(x)}{\tau(P)(y)} \frac{d\tau(\delta_x)}{d\tau(P)}(y), \forall (x, y) \in X \times Y$$

Then the reversibility of the Markov process is precisely satisfied, if also the dual conditional probability does not depend on the choice of  $P$ . In this case also the equilibrium distribution

$$\pi_P(x, y) = P(x)\tau(P)(y)$$

does not depend on the choice of  $P$  and therefore allows to generalize the principle of detailed balance to model spaces.

**Proposition 4** (*Generalized principle of detailed balance*). *Let  $Z$  be a Markov process from  $X$  to  $Y$ , which pushes  $\mathcal{M}$  to  $\mathcal{N}$ . Then  $Z$  is reversible, iff there exist conditional probabilities  $p(y | x)$  and  $p(x | y)$ , such that for all  $P \in \mathcal{M}$  and its respectively induced  $Q \in \mathcal{N}$ , it holds that:*

$$p(x | y)Q(y) = p(y | x)P(x), \forall (x, y) \in X \times Y$$

The generalized principle of detailed balance can be substantiated if the respective model spaces  $\mathcal{M}$  and  $\mathcal{N}$  are dominated by measures  $\mu$  and  $\nu$ . Then the conditional probabilities are induced by a transition operator

$$\tau : L^1(X, \mathcal{A}, \mu) \rightarrow L^1(X, \mathcal{A}, \nu)$$

that satisfies  $\tau(\mathcal{M}) = \mathcal{N}$ . This property facilitates the notation of *Statistical Morphisms*.

**Definition** (Statistical Morphism). *Let  $(X, \mathcal{A}, \mathcal{M})$  and  $(Y, \mathcal{B}, \mathcal{N})$  be statistical models and*

$$T : (X, \mathcal{A}) \rightarrow (Y, \mathcal{B})$$

*a measurable function. Then a mapping  $f : \mathcal{M} \rightarrow \mathcal{N}$  is termed a Statistical Morphism, iff there exists a transition operator*

$$\tau : L^1(X, \mathcal{A}, \mathcal{M}) \rightarrow L^1(Y, \mathcal{B}, \mathcal{N})$$

*such that the following diagram commutes:*

**Lemma 5** (Existence and Uniqueness). *Let  $(X, \mathcal{A}, \mathcal{M})$  and  $(Y, \mathcal{B}, \mathcal{N})$  be statistical models and*

$$T : (X, \mathcal{A}) \rightarrow (Y, \mathcal{B})$$

*a measurable function with  $\mathcal{N} = T_*\mathcal{M}$ . Then  $T$  induces a Statistical Morphism  $f : \mathcal{M} \rightarrow \mathcal{N}$  which is a.e. unique up to the quotient  $\mu/f(\mu)$ .*

*Proof.* Lemma 2 postulates the existence of a transition operator  $\tau$  with  $\tau(\mu) = T_*\mu$ , which is unique for  $\nu$ -almost all  $y \in Y$ . Furthermore from  $\mathcal{M} \preceq \mu$  it follows, that

$$\mathcal{M} \subseteq L^1(X, \mathcal{A}, \mu)$$

and from  $\mathcal{N} \preceq \nu$  that  $T_*\mu \preceq \nu$ . Since however  $\mathcal{N} = T_*\mathcal{M}$  it also follows, that:

$$\mathcal{N} \subseteq L^1(Y, \mathcal{B}, T_*\mu)$$

Then for any  $P \in \mathcal{M}$  it holds, that  $\tau(P) \in \mathcal{N}$ , and therefore that  $\text{img}\tau|_{\mathcal{M}} \subseteq \mathcal{N}$ . Conversely for any  $Q \in \mathcal{N}$  there exists an  $P \in \mathcal{M}$  with  $Q = \tau(P)$  such that also  $\mathcal{N} \subseteq \text{img}\tau|_{\mathcal{M}}$ . This proves that  $\mathcal{N} = \text{img}\tau|_{\mathcal{M}}$  and thereupon that  $f := \tau|_{\mathcal{M}}$  yields a Statistical Morphism which is unique for  $\nu$ -almost all  $y \in Y$ . As  $f$  in turn depends on the choice of  $\mu$  it is unique up to the quotient  $\mu/f(\mu)$  and therefore essentially a.e. unique.  $\square$

Since transition operators are bounded linear operators with regard to the underlying  $L^1$ -spaces, they induce a proximity structure within the model spaces, which is determined by its image and its kernel. Thereby the image of the transition operator is uniquely given by:

$$\text{img}(\tau) = L^1(Y, \mathcal{B}, T_*\mu)$$

and its kernel by  $\ker(\tau) = \tau^{-1}(0_{\mathcal{B}})$ , where:

$$0_{\mathcal{B}} := \{\nu \in \text{img}(\tau) \mid \nu(B) = 0, \forall B \in \mathcal{B}\}$$

With regard to a Statistical Morphism  $f$  however, the situation is slightly more sophisticated, as the model spaces do not provide a canonical vector space structure. Due to the requirement of commutativity with regard to a measurable function  $T$  however the image is given by  $\text{img}(f) = T_*\mathcal{M}$ . Furthermore the kernel of  $f$  corresponds to a partition of  $\mathcal{M}$ , which is induced by the subspace topology from  $L^1$ . Thereby  $P, Q \in \mathcal{M}$  are  $L^1$ -identical over  $\mathcal{A}$  and denoted by  $P \stackrel{A}{=} Q$  or by the equivalence class  $P \in \text{id}_{\mathcal{A}}(Q)$ , iff:

$$P(A) = Q(A), \forall A \in \mathcal{A}$$

Furthermore for any  $Q \in \text{img}(f)$  the kernel of  $f$  at  $Q$  is given by the preimage

$$\ker_Q(f) = f^{-1}(\text{id}_{\mathcal{B}}(Q))$$

**Definition** (Statistical Epi-/Mono-/Isomorphism). Let  $f: \mathcal{M} \rightarrow \mathcal{N}$  be a Statistical Morphism which is induced by:

$$T: (X, \mathcal{A}) \rightarrow (Y, \mathcal{B})$$

Then a  $f$  is termed a Statistical Monomorphism, iff it satisfies:

$$\ker_{f(P)}(f) \subseteq \text{id}_{\mathcal{A}}(P), P \in \mathcal{M}$$

In this case it follows, that:

$$f(P) \stackrel{\mathcal{B}}{=} f(Q) \Rightarrow P \stackrel{\mathcal{A}}{=} Q, \forall P \in \mathcal{M}, \forall Q \in \mathcal{M}$$

Furthermore  $f$  is termed a Statistical Epimorphism iff for all  $Q \in \mathcal{N}$ , there exists an  $P \in \mathcal{M}$ , such that  $f(P) \stackrel{\mathcal{B}}{=} Q$  and  $\text{img}(f) \stackrel{\mathcal{B}}{=} \mathcal{N}$ . Finally  $f$  is termed a Statistical Isomorphism, iff it is a Statistical Monomorphism and a Statistical Epimorphism.

**Lemma 6.** Let  $(X, \mathcal{A}, \mathcal{M})$  and  $(Y, \mathcal{B}, \mathcal{N})$  be statistical models and  $f: \mathcal{M} \rightarrow \mathcal{N}$  be a Statistical Morphism, then  $f$  is a Statistical Isomorphism, iff there exists a further Statistical Morphism  $f^*: \mathcal{N} \rightarrow \mathcal{M}$ , such that (i)

$$(f^* \circ f)(P) \stackrel{\mathcal{A}}{=} P, \forall P \in \mathcal{M}$$

and (ii)

$$(f \circ f^*)(Q) \stackrel{\mathcal{B}}{=} Q, \forall Q \in \mathcal{N}$$

*Proof.* “ $\Rightarrow$ ” Let  $f$  be a statistical isomorphism, then a further Statistical Morphism  $f^*: \mathcal{N} \rightarrow \mathcal{M}$  is given by  $f^*(Q) := \ker_Q(f)$ , which provides a complete system of representatives of the kernel, such that

$$(f^* \circ f)(P) \stackrel{\mathcal{A}}{=} P, \forall P \in \mathcal{M}$$

and

$$(f \circ f^*)(Q) \stackrel{\mathcal{B}}{=} Q, \forall Q \in \mathcal{N}$$

“ $\Leftarrow$ ” Conversely if there exists any Statistical Morphism  $f^*: \mathcal{N} \rightarrow \mathcal{M}$  such that  $f$  and  $f^*$  satisfy (i) and (ii), then for any  $P, Q \in \mathcal{M}$  with  $P \notin \text{id}_{\mathcal{A}}(Q)$  it follows that:

$$f(P) \notin \text{id}_{\mathcal{B}}(f(Q))$$

Otherwise let  $P \notin \text{id}_{\mathcal{A}}(Q)$  with  $f(P) \in \text{id}_{\mathcal{B}}(f(Q))$ , then it would follow, that also

$$(f^* \circ f)(P) \stackrel{\mathcal{A}}{=} (f^* \circ f)(Q)$$

However since  $(f^* \circ f)(P) \stackrel{\mathcal{A}}{=} P$  and  $(f^* \circ f)(Q) \stackrel{\mathcal{A}}{=} Q$  it would also follow, that  $P \notin \text{id}_{\mathcal{A}}(Q)$  which contradicts to the initial assumption  $P \notin \text{id}_{\mathcal{A}}(Q)$ . This proves, that  $f$  is a statistical monomorphism. Furthermore let  $Q \in \mathcal{N}$  and  $P = f^*(Q)$ . Then  $f(P) = (f \circ f^*)(Q)$  and therefore  $f(P) \stackrel{\mathcal{B}}{=} Q$ . This proves, that  $f$  is a Statistical Epimorphism and therefore a statistical isomorphism.  $\square$

**Proposition 7.** Let  $(X, \mathcal{A}, \mathcal{M})$  and  $(Y, \mathcal{B}, \mathcal{N})$  be statistical models with  $\mathcal{M} \preceq \mu$  and  $\mathcal{N} \preceq \nu$ , then  $(X, \mathcal{A}, \mathcal{M})$  and  $(Y, \mathcal{B}, \mathcal{N})$  are statistical equivalent, iff there exists a statistical isomorphism  $f: \mathcal{M} \rightarrow \mathcal{N}$ .

*Proof.* “ $\Rightarrow$ ” Let  $(X, \mathcal{A}, \mathcal{M})$  and  $(Y, \mathcal{B}, \mathcal{N})$  be statistical equivalent, then there exists a Markov Process  $Z$  from  $(X, \mathcal{A})$  to  $(Y, \mathcal{B})$ , that induces  $\mathcal{N}$  from  $\mathcal{M}$  and therefore a measurable function

$$T: (X, \mathcal{A}) \rightarrow (Y, \mathcal{B})$$

with  $\mathcal{N} = T_*\mathcal{M}$ . Then due to Lemma 5  $T$  induces a Statistical Morphism  $f: \mathcal{M} \rightarrow \mathcal{N}$  and therefore a transition operator

$$\tau: L^1(X, \mathcal{A}, \mu) \rightarrow L^1(Y, \mathcal{B}, \nu)$$

with  $\tau(\mathcal{M}) = \mathcal{N}$  such that  $f := \tau|_{\mathcal{M}}$ . Since  $Z$  is reversible also  $\nu$  induces a Statistical Morphism  $f^*: \mathcal{N} \rightarrow \mathcal{M}$  and therefore a dual transition operator

$$\tau^*: L^1(Y, \mathcal{B}, \nu) \rightarrow L^1(X, \mathcal{A}, \mu)$$

with  $\tau^*(\mathcal{N}) = \mathcal{M}$ , such that  $f^* := \tau^*|_{\mathcal{N}}$ . Thereby  $f$  and  $f^*$  are unique up to the quotients  $\mu/f(\mu)$  and  $\nu/f^*(\nu)$ . Furthermore  $f$  and  $f^*$  have to satisfy the generalized principle of detailed balance. Therefore independent of  $\mu/f(\mu)$  and  $\nu/f^*(\nu)$  it follows, that:

$$(f^* \circ f)(P) \stackrel{\mathcal{A}}{=} P, \forall P \in \mathcal{M}$$

and conversely that:

$$(f \circ f^*)(Q) \stackrel{\mathcal{B}}{=} Q, \forall Q \in \mathcal{N}$$

Due to Lemma 6 it follows, that  $f$  is a statistical isomorphism. “ $\Leftarrow$ ” Let  $f$  be a statistical isomorphism. Then there exists a transition operator  $\tau$  with  $\tau(\mathcal{M}) = \mathcal{N}$ , such that the conditional probabilities given by equation 2.10 do not depend on the choice of  $P \in \mathcal{M}$ . Then  $\tau$  defines a Markov process  $Z$  from  $(X, \mathcal{A})$  to  $(Y, \mathcal{B})$ , that induces  $\mathcal{N}$  from  $\mathcal{M}$ . Since  $f$  is also an isomorphism, the argument mutatis mutandis also applies to a dual transition operator  $\tau^*$  such that  $Z$  is reversible.  $\square$

Proposition 7 shows, that the abstract concept of statistical equivalence has a very precise interpretation given by Statistical Isomorphisms. This motivates to reconsider statistical models as the objects of a category.

**Definition** (Category of statistical models). *The category of statistical models, denoted by **Stat**, consists of:*

- (1) *a class of objects  $\text{ob}(\mathbf{Stat})$ , that comprises all statistical models*
- (2) *a class of morphisms  $\text{hom}(\mathbf{Stat})$ , that comprises all statistical morphisms*

### 3. SUFFICIENCY AND COMPLETENESS IN THE CATEGORY OF STATISTICAL MODELS

Of course any new mathematical framework at first has to proof it's usability w.r.t. established concepts. In the following it is therefore shown, that the classical concepts of sufficiency and completeness have a very narrow meaning in the category of statistical models, respectively given by statistical monomorphisms and statistical epimorphisms. For the beginning, it is shown, that that statistical morphisms are closely related to the properties of its underlying measurable function  $T$ .

**Proposition 8.** *Let*

$$(X, \mathcal{A}, \mathcal{M}) \in \text{ob}(\mathbf{Stat})$$

$$(Y, \mathcal{B}, \mathcal{N}) \in \text{ob}(\mathbf{Stat})$$

*with  $\mathcal{M} \preceq \mu$  and  $\mathcal{N} \preceq \nu$  and let*

$$T: (X, \mathcal{A}) \rightarrow (Y, \mathcal{B})$$

*be a bimeasurable function with:*

$$\mathcal{N} = T_*\mathcal{M}$$

*then  $(X, \mathcal{A}, \mathcal{M})$  and  $(Y, \mathcal{B}, \mathcal{N})$  are statistical equivalent.*

*Proof.* Due to Lemma 5  $T$  induces a Statistical Morphism  $f: \mathcal{M} \rightarrow \mathcal{N}$  with

$$f(P) = T_*P, \forall P \in \mathcal{M}$$

which is a.e. unique up to the quotient  $\mu/f(\mu)$ . Since  $T$  is bimeasurable, there also exists an inverse measurable function  $T^{-1}$  with  $T^{-1} \circ T = \text{id}_{\mathcal{A}}$ . Then also  $T^{-1}$  induces a Statistical Morphism  $f^*: \mathcal{N} \rightarrow \mathcal{M}$  with

$$f^*(Q) = T_*^{-1}Q, \forall Q \in \mathcal{N}$$

which is a.e. unique up to the quotient  $\nu/f^*(\nu)$ . Then it follows, that:

$$(f^* \circ f)(P) = (T^{-1} \circ T)_*(P) \stackrel{\mathcal{A}}{=} P, \forall P \in \mathcal{M}$$

and that:

$$(f \circ f^*)(Q) = (T \circ T^{-1})_*(Q) \stackrel{\mathcal{B}}{=} Q, \forall Q \in \mathcal{N}$$

By Lemma 6 it follows, that  $f$  is a statistical isomorphism and therefore by 7 that  $(X, \mathcal{A}, \mathcal{M})$  and  $(Y, \mathcal{B}, \mathcal{N})$  are statistically equivalent.  $\square$

The criterion for statistical equivalence, given by 8 requires the statistic  $T$  to be a bimeasurable function, which is a very strong assumptions. In the purpose to characterise statistical equivalence by the underlying measurable function, this requirement therefore has to be weakened in the sense, to only regard events, that are crucial to the respectively model spaces. For example if only very few

distinct events are necessary for the determination of the  $L^1$ -identity, then the measurable function  $T$  in particular only has to preserve the distinction of those events and not of all events, as in the case of a bimeasurable function. In order to formalize this approach of “coarse graining”, the injectivity and surjectivity of a bimeasurable function are respectively weakened by *sufficiency* and *completeness*.

**Definition** (Sufficiency). *Let*

$$(X, \mathcal{A}, \mathcal{M}) \in \text{ob}(\mathbf{Stat})$$

*let*  $(Y, \mathcal{B})$  *be a measurable space and*

$$T: (X, \mathcal{A}) \rightarrow (Y, \mathcal{B})$$

*a measurable function. Let further be*  $\mathcal{M}_S \subseteq \mathcal{M}$  *with*  $\mathcal{M}_S \preccurlyeq \mu$ , *then*  $T$  *is sufficient for*  $\mathcal{M}_S$ , *iff:*

$$p_P(x | y) = p_\mu(x | y), \forall P \in \mathcal{M}_S$$

Intuitively a measurable function  $T: (X, \mathcal{A}) \rightarrow (Y, \mathcal{B})$ , that is sufficient for a set of probability distributions  $\mathcal{M}_S \subseteq \mathcal{M}$ , is “fine enough”, to preserve the  $L^1$ -identity of  $\mathcal{M}_S$  within its image  $T_*\mathcal{M}_S$ , such that for arbitrary  $P, Q \in \mathcal{M}_S$  it holds that:

$$T_*P \stackrel{\mathcal{B}}{=} T_*Q \Rightarrow P \stackrel{\mathcal{A}}{=} Q$$

If  $T$  is even sufficient for the whole  $L^1$ -space, such that

$$\mathcal{M} = L^1(X, \mathcal{A}, \mu)$$

with  $\mathcal{M} \preccurlyeq \mu$ , then  $T$  is a bimeasurable function. This shows, that sufficiency indeed may be regarded as a generalization of injectivity and furthermore is closely related to statistical monomorphisms.

**Lemma 9.** *Let*

$$(X, \mathcal{A}, \mathcal{M}) \in \text{ob}(\mathbf{Stat})$$

$$(Y, \mathcal{B}, \mathcal{N}) \in \text{ob}(\mathbf{Stat})$$

*with*  $\mathcal{M} \preccurlyeq \mu$  *and*  $\mathcal{N} \preccurlyeq \nu$  *and let*

$$T: (X, \mathcal{A}) \rightarrow (Y, \mathcal{B})$$

*be a measurable function with*  $T_*\mathcal{M} \stackrel{\mathcal{B}}{=} \mathcal{N}$ . *Then*  $T$  *is sufficient for*  $\mathcal{M}$ , *iff*  $T$  *induces a statistical monomorphism*  $f: \mathcal{M} \rightarrow \mathcal{N}$ .

*Proof.* Let  $\mathcal{N}_C = T_*\mathcal{M} \subseteq \mathcal{N}$  then  $\mathcal{N}_C \preccurlyeq \nu$ . Due to Lemma 5  $T$  induces a Statistical Morphism  $f: \mathcal{M} \rightarrow \mathcal{N}_C$  and therefore a transition operator

$$\tau: L^1(X, \mathcal{A}, \mathcal{M}) \rightarrow L^1(Y, \mathcal{B}, \mathcal{N}_C)$$

with  $\tau(\mathcal{M}) = \mathcal{N}_C$ , which is a.e. unique up to  $\mu/\tau(\mu)$ . In particular the  $p_P(y | x)$  do not depend on the choice of  $P \in \mathcal{M}$ . “ $\Rightarrow$ ” Let now be  $T$  sufficient for  $\mathcal{M}$ , then by definition also the induced dual conditional probabilities  $p_P(x | y)$  do not depend on the choice of  $P \in \mathcal{M}$ . Due to the generalized principle of detailed balance there exists a reversible Markov process from  $X$  to  $Y$ , that pushes  $\mathcal{M}$  to  $\mathcal{N}_C$ , such that  $(X, \mathcal{A}, \mathcal{M})$  and  $(Y, \mathcal{B}, \mathcal{N}_C)$  are statistical equivalent. Then by 7 it follows, that  $f|_{\mathcal{N}_C}$  is a statistical isomorphism and  $f$  is a statistical monomorphism. “ $\Leftarrow$ ” Let  $f$  be a statistical monomorphism, then  $f|_{\mathcal{N}_C}$  is a statistical isomorphism and 7 postulates that  $(X, \mathcal{A}, \mathcal{M})$  and  $(Y, \mathcal{B}, \mathcal{N}_C)$  are statistical equivalent such that there exists a reversible Markov process from  $X$  to  $Y$ , that pushes  $\mathcal{M}$  to  $\mathcal{N}_C$ . Then the generalized principle of detailed balance postulates, that the  $p_P(x | y)$  do not depend on the choice of  $P \in \mathcal{M}$ , such that  $T$  is sufficient for  $\mathcal{M}$ .  $\square$

Conversely also the question arises, if a measurable function  $T$  is “fine enough”, to generate the  $L^1$ -identity over its codomain. Let therefore  $(X, \mathcal{A})$  be a measurable space,  $(Y, \mathcal{B}, \mathcal{N})$  a statistical model with  $\mathcal{N} \preccurlyeq \nu$  and

$$T: (X, \mathcal{A}) \rightarrow (Y, \mathcal{B})$$

a measurable function. Then for any  $\rho \in L^1(Y, \mathcal{B}, \nu)$  an approximation of  $\rho$  is given by the conditional



expectation  $E_\nu(\rho \mid T(\mathcal{A}))$  and  $T$  generates the  $L^1$ -identity over  $(Y, \mathcal{B}, \mathcal{N})$ , if for any

$$\rho \in L^1(Y, \mathcal{B}, \nu)$$

with

$$E_\nu(\rho \mid T(\mathcal{A})) \stackrel{\mathcal{B}}{=} 0_{\mathcal{B}}$$

it follows, that  $\rho \stackrel{\mathcal{B}}{=} 0_{\mathcal{B}}$ . This provides the notation of completeness.

**Definition** (Completeness). *Let  $(X, \mathcal{A})$  be a measurable space,  $(Y, \mathcal{B}, \mathcal{N}) \in \text{ob}(\mathbf{Stat})$  and*

$$T: (X, \mathcal{A}) \rightarrow (Y, \mathcal{B})$$

*a measurable function. Let further be  $\mathcal{N}_C \subseteq \mathcal{N}$  with  $\mathcal{N}_C \preccurlyeq \nu$ , then  $T$  is termed complete for  $\mathcal{N}_C$ , iff for all*

$$\rho \in L^1(Y, \mathcal{B}, \nu)$$

*with*

$$E_\nu(\rho \mid T(\mathcal{A})) \stackrel{\mathcal{B}}{=} 0_{\mathcal{B}}$$

*it follows, that  $\rho \stackrel{\mathcal{B}}{=} 0_{\mathcal{B}}$ .*

In particular if a measurable function  $T$  is complete for  $\mathcal{N}_C$ , then for arbitrary  $P, Q \in \mathcal{N}_C$  with  $P \notin \text{id}_{\mathcal{B}}(Q)$  there exists an  $A \in \mathcal{A}$ , such that:

$$P(T(A)) \neq Q(T(A))$$

By assuming an underlying statistical model  $(X, \mathcal{A}, \mathcal{M})$  and  $\mathcal{N}$  to be induced by  $T$ , such that  $\mathcal{N} = T_*\mathcal{M}$ , then the condition may also be pulled back to  $\mathcal{M}$ . Then  $T$  is complete for  $\mathcal{N}_C$ , if for arbitrary for  $P, Q \in \mathcal{N}_C$  with  $P \notin \text{id}_{\mathcal{B}}(Q)$  there exists an  $A \in \mathcal{A}$ , such that

$$(T^*P)(A) \neq (T^*Q)(A)$$

This however is equivalent to the claim, that  $T$  generates the  $L^1$ -identity over  $\mathcal{N}_C$ , such that for all  $Q \in \mathcal{N}_C$ , there exists an  $P \in \mathcal{M}$ , such that  $T_*P \stackrel{\mathcal{B}}{=} Q$ . This shows, that completeness generalizes surjectivity and furthermore is closely related to Statistical Epimorphisms.

**Lemma 10.** *Let*

$$(X, \mathcal{A}, \mathcal{M}) \in \text{ob}(\mathbf{Stat})$$

$$(Y, \mathcal{B}, \mathcal{N}) \in \text{ob}(\mathbf{Stat})$$

*with  $\mathcal{M} \preccurlyeq \mu$  and  $\mathcal{N} \preccurlyeq \nu$  and*

$$T: (X, \mathcal{A}) \rightarrow (Y, \mathcal{B})$$

*a measurable function with  $T_*\mathcal{M} \stackrel{\mathcal{B}}{=} \mathcal{N}$ . Then  $T$  is complete for  $\mathcal{N}$ , iff  $T$  induces a Statistical Epimorphism  $f: \mathcal{M} \rightarrow \mathcal{N}$ .*

*Proof.* “ $\implies$ ” Since  $T_*\mathcal{M} \stackrel{\mathcal{B}}{=} \mathcal{N}$  it follows, that  $T_*\mathcal{M} \subseteq \mathcal{N}$ . Since  $T$  is complete for  $\mathcal{N}$ , for all  $Q \in \mathcal{N}$  there exists an  $P \in \mathcal{M}$ , such that  $T_*P \stackrel{\mathcal{B}}{=} Q$  and therefore  $\mathcal{N} \subseteq T_*\mathcal{M}$ , such that  $T_*\mathcal{M} = \mathcal{N}$ . Then due to Lemma 5  $T$  induces a Statistical Morphism  $f: \mathcal{M} \rightarrow \mathcal{N}$  with  $f(\mathcal{M}) = T_*\mathcal{M}$  and since

$$\text{img}(f) = T_*\mathcal{M} \stackrel{\mathcal{B}}{=} \mathcal{N}$$

it follows, that  $f$  is a Statistical Epimorphism. “ $\impliedby$ ” Let  $f$  be a Statistical Epimorphism, then  $\text{img}(f) \stackrel{\mathcal{B}}{=} \mathcal{N}$  where

$$f(\mathcal{M}) = T_*\mathcal{M} \stackrel{\mathcal{B}}{=} \mathcal{N}$$

Then for any  $Q \in \mathcal{N}$  there exists an  $P \in \mathcal{M}$ , such that  $T_*P \stackrel{\mathcal{B}}{=} Q$  and  $T$  is complete for  $\mathcal{N}$ .  $\square$

The properties of sufficiency for  $\mathcal{M}$  and completeness for  $\mathcal{N}$  are still satisfied if they are only required for the subsets  $\mathcal{M}/\text{id}_{\mathcal{A}} \subseteq \mathcal{M}$  and  $\mathcal{N}/\text{id}_{\mathcal{B}} \subseteq \mathcal{N}$ , that generate the respective  $L^1$ -identity. Intuitively this represents the property that for the statistical equivalence of statistical models it simply doesn't matter if the model spaces are generated by distribution assumptions, that are observational indistinguishable, as long as the distinguishable probability distributions are distinguishable in both statistical models. This allows to characterise statistical equivalence by the underlying measurable function.

**Theorem 11.** *Let*

$$(X, \mathcal{A}, \mathcal{M}) \in \text{ob}(\mathbf{Stat})$$

$$(Y, \mathcal{B}, \mathcal{N}) \in \text{ob}(\mathbf{Stat})$$

*with  $\mathcal{M} \preceq \mu$  and  $\mathcal{N} \preceq \nu$ . Then the following statements are equivalent:*

(i) *There exists a measurable function*

$$T: (X, \mathcal{A}) \rightarrow (Y, \mathcal{B})$$

*with  $T_*\mathcal{M} \stackrel{\mathcal{B}}{=} \mathcal{N}$ , such that  $T$  is sufficient for  $\mathcal{M}/\text{id}_{\mathcal{A}}$  and  $T$  is complete for  $\mathcal{N}/\text{id}_{\mathcal{B}}$*

(ii)  *$(X, \mathcal{A}, \mathcal{M})$  and  $(Y, \mathcal{B}, \mathcal{N})$  are statistical equivalent*

(iii) *There exists a statistical isomorphism  $f: \mathcal{M} \rightarrow \mathcal{N}$*

*Proof.* “(i)  $\Rightarrow$  (ii)” Since  $T$  is sufficient for  $\mathcal{M}/\text{id}_{\mathcal{A}}$ , it is also sufficient for  $\mathcal{M}$  and therefore by Lemma 9 it follows, that  $T$  induces a statistical monomorphism  $f: \mathcal{M} \rightarrow \mathcal{N}$ . Furthermore since  $T$  is complete for  $\mathcal{N}/\text{id}_{\mathcal{B}}$ , it is also complete for  $\mathcal{N}$  and therefore by Lemma 10 and the a.e. uniqueness of  $f$  it follows, that  $f$  is also a Statistical Epimorphism and therefore a statistical isomorphism. Then due to 7  $(X, \mathcal{A}, \mathcal{M})$  and  $(Y, \mathcal{B}, \mathcal{N})$  are statistical equivalent. “(ii)  $\Rightarrow$  (i)” Let  $(X, \mathcal{A}, \mathcal{M})$  and  $(Y, \mathcal{B}, \mathcal{N})$  be statistical equivalent, then 7 postulates the existence of a statistical isomorphism  $f: \mathcal{M} \rightarrow \mathcal{N}$ . “(ii)  $\Leftrightarrow$  (iii)” Has already been proven by 7.  $\square$

#### 4. TOPOLOGY OF STATISTICAL MODELS

Apart of an underlying reversible Markov process, it would be pleasant to characterise statistical equivalence directly by a proximity relationship of the model spaces. In 2 this observable structure has intuitively been introduced by a pairwise comparison of probability distributions, with regard to the collection of events in the sample space. What is common to them, is the capability to induce a unique topology to the model space, that characterises the statistical inference w.r.t. the evaluations.

The intuition behind the concept of statistical models is to provide a framework for *statistical inference*, which eventually allows to derive conclusions about the “true model” by coincidences between observable and structural beliefs. Consequentially the fundamental ability to derive any conclusions depends on the distinguishability of distribution assumptions w.r.t. realizations of finite random samples. More precisely for a hypothesis Space  $\mathcal{H}$  any distribution assumptions  $H_P, H_Q \in \mathcal{H}$  are *observable distinguishable*, iff their induced probability distributions  $P, Q \in \mathcal{M}$  are distinguishable by an event  $A \in \Sigma$ , such that  $P(A) \neq Q(A)$ . Beyond the observable distinguishability however, the capability to decide which distribution assumption is “closer” to an observation, requires the existence of a proximity structure within the model space. The *observable structure* of a statistical model therefore provides the foundation of observation based statistical inference. Although for arbitrary statistical models, there is no natural selection of an observable structure, this deficiency may be resolved, by requiring the induced probability distributions to admit a density function. This restriction yields the notation of *continuous statistical models*.

**Definition** (Continuous statistical model). *Let  $(S, \Sigma, \mathcal{M})$  be a statistical model. Then  $(S, \Sigma, \mathcal{M})$  is termed continuous iff (i) the sample space  $(S, \Sigma)$  is a Borel-space and (ii) all  $P \in \mathcal{M}$  are absolutely continuous over  $(S, \Sigma)$ . **Remark:** In the following all statistical models are assumed to be continuous unless stated otherwise.*

Due to its definition, the model space of a continuous statistical model  $(S, \Sigma, \mathcal{M})$  is embedded within the space of Lebesgue measurable functions over  $(S, \Sigma)$ , which naturally induces the  $L_1$ -topology to  $\mathcal{M}$ . The underlying statistic  $T: (\Omega, \mathcal{F}, P) \rightarrow (S, \Sigma)$  then allows to pull back the topology to  $\mathcal{H}$  and therefore endows the distribution assumptions

with a proximity structure. In particular distribution assumptions are observable distinguishable iff they are distinguishable within the induced  $L_1$ -topology. However in order to quantify the proximity between probability assumptions and observations also the respective “ $L_1$ -neighbourhoods” of empirical distributions have to be pulled back. A technical difficulty however arises from the fact that empirical distributions are not Lebesgue measurable. Nevertheless since the empirical distributions are at least Bochner measurable, it suffices to (i) extend  $L^1(S, \Sigma)$  by the empirical distributions of a finite random sample  $X$  and (ii) derive the observable structure from the extended topology. In this purpose let  $X$  be an  $(S, \Sigma)$ -valued finite random sample of length  $N$  and  $\mathcal{E}(X)$  the set of all empirical distributions  $P_n(X)$  of  $X$  with  $n \leq N$ . Let further be  $d_1$  the  $L_1$ -distance over  $L^1(S, \Sigma)$ , such that  $d_1(\mu, \nu) := \|\mu - \nu\|_1$  for all  $\mu, \nu \in L^1(S, \Sigma)$  and  $L_X^1(S, \Sigma)$  the smallest Bochner space over  $(S, \Sigma)$ , that covers  $L^1(S, \Sigma)$  as well as  $\mathcal{E}(X)$ . Then due to the finite number of jump discontinuities within the empirical distributions, the  $L_1$ -distance  $d_1$  may naturally be extended to  $L_X^1(S, \Sigma)$  by a formal integration by parts. Consequentially the extended  $L_1$ -topology, which is generated by  $d_1$  in particular makes  $L_X^1(S, \Sigma)$  a topological vector space, which covers  $\mathcal{M}$  and  $\mathcal{E}(X)$  and therefore provides a natural choice for an observable structure of continuous statistical models. The next step towards an effective approximation of the sample distribution concerns the convergence rate of the empirical distributions. Although the  $L_1$ -convergence of empirical distributions is assured in their asymptotic limit, the  $L_1$ -topology may be too strong, to capture the remaining uncertainty of finite random samples. This is due to the fact that the  $L_1$ -topology considers every single aspect of the sample distribution by the complete evaluation of individual events. Therefore if some aspects are considered to be “more

important” than others, then a more commensurate topology may be obtained by a restriction of the evaluation to these aspects. This restriction is performed by *estimands*.

**Definition** (Continues estimand). *Let  $(S, \Sigma, \mathcal{M})$  be a continuous statistical model and  $X$  a finite random sample in  $S$  of length  $n$ , which generates the  $\sigma$ -Algebra  $\mathcal{A}^n$ . Then a mapping  $\epsilon: \mathcal{M} \times \mathcal{A}^n \rightarrow \mathbb{R}$  is a continuous estimand, iff (i)  $\epsilon$  is  $L^1$ -continuous in its first argument and (ii)  $\epsilon$  is  $\mathcal{A}^n$ -symmetric in its second argument.*

Estimands provide the opportunity to restrict an evaluation of probability distributions to assorted aspects. As this evaluation is performed w.r.t. given realizations  $A \in \mathcal{A}^n$  they conditionally dependent on the respective realizations and therefore are given by the notation  $\epsilon(P \mid X \in A) := \epsilon(P, A)$ . This notation implicates the abbreviations “ $\epsilon(P \mid A)$ ”, to emphasize a value in  $\mathbb{R}$  and “ $\epsilon(P \mid X)$ ”, to emphasize a function in  $L_X^1(S, \Sigma)$ . Thereby the requirement of the function  $\epsilon(P \mid X): A \mapsto \epsilon(P \mid A)$  to be symmetric w.r.t.  $A = \{A_i\} \in \mathcal{A}^n$  assures permutation invariance and therefore preserves the independence of the individual observations  $A_i$ . Furthermore the requirement of the operator  $\epsilon: P \mapsto \epsilon(P \mid X)$  to be  $L_1$ -continuous induces a topology within  $L_X^1(S, \Sigma)$ , by the quotient topology w.r.t. its kernel. In particular this induced  $\epsilon$ -topology is identical to the  $L_1$ -topology iff  $\epsilon$  is an  $L_1$ -homeomorphism. Consequently the question for the existence of continuous estimands arises.

**Lemma 12.** *Let*

$$(S, \Sigma, \mathcal{M}) \in \text{ob}(\mathbf{Stat})$$

*$X$  a finite sample over  $(S, \Sigma)$ ,  $\epsilon$  an estimand of  $\mathcal{M}$  over  $X$ . Then for any  $P \in \mathcal{M}$  there exists a unique coarsest topology  $\mathcal{T}_P$  over  $\mathcal{M}$ , that makes  $\epsilon$  continuous w.r.t.  $P$ .*

*Proof.* Let  $\mathcal{A}$  be the induced  $\sigma$ -algebra of the sample  $X$ . Then the definition of  $\epsilon$  assures the existence of a  $\sigma$ -finite measure  $\mu$  over  $(X, \mathcal{A})$  with  $\mathcal{M} \subseteq L^1(S, \Sigma, \mu)$ , such that for any  $Q \in \mathcal{M}$  the function  $\epsilon_Q: A \mapsto \epsilon(Q | A)$  is absolutely continuous w.r.t.  $\mu$  and therefore  $\epsilon_Q \in L^1(S, \Sigma, \mu)$ . This allows the definition of a distance by:

$$d(P, Q) := \|\epsilon_P - \epsilon_Q\|_1$$

Let  $S_P = \{d(P, Q) | Q \in \mathcal{M}\}$ , then

$$\epsilon_P, \epsilon_Q \in L^1(S, \Sigma, \mu)$$

assures the existence of  $a, b \in \mathbb{R}$  with  $S_P \subseteq [a, b]$ . Let  $\mathcal{T}_S$  be the subspace topology of  $S_P$  in  $\mathbb{R}$ , then for any  $V \in \mathcal{T}_S$  let

$$U_V = \{Q \in \mathcal{P} | d(P, Q) \in V\}$$

Then  $\mathcal{T}_P = \{U_V | V \in \mathcal{T}_S\}$  defines a topology over  $\mathcal{M}$ , which is second countable, since  $\mathcal{T}_S$  has a countable base. Furthermore any  $Q \in \mathcal{P}$  is topologically distinguishable from  $P$  iff there exists an  $A \in \Sigma$  with

$$\epsilon(P | A) \neq \epsilon(Q | A)$$

□

**Theorem 13** (*Initial Theorem*). *Let  $(S, \Sigma, \mathcal{M})$  be a statistical model,  $X$  a finite sample over  $(S, \Sigma)$  and  $\epsilon$  an estimand of  $\mathcal{M}$  over  $X$ . Then there exists a unique coarsest topology  $\mathcal{T}$  over  $\mathcal{M}$ , that makes  $\epsilon$  continuous.*

*Proof.* Due to Lemma 12 for any  $P \in \mathcal{M}$  the evaluation  $\epsilon$  induces a topology  $\mathcal{T}_P$  in  $\mathcal{M}$ , which inherits the property of a countable base. Let  $\mathcal{U}$  be the unification of all topologies  $\mathcal{T}_P$  and  $\mathcal{T}$  the coarsest topology over  $\mathcal{M}$ , that covers  $\mathcal{U}$ . Then  $\mathcal{T}$  has a countable base and is the coarsest topology over  $\mathcal{M}$ , that preserves the continuity of  $\epsilon$  within  $P$ . In particular  $\mathcal{T}$  preserves the distinction of any  $P, Q \in \mathcal{M}$  w.r.t.  $\epsilon$ , since  $P$  and  $Q$  are distinguishable in  $\mathcal{T}$  iff there exists an  $A \in \Sigma$  with  $\epsilon(P | A) \neq \epsilon(Q | A)$ . □

The Initial Theorem of Topological Statistical Theory for any given estimand assures the existence of a unique coarsest topology, that makes it continuous. Without a given estimand, however, one might of course also like to be able to derive a proximity structure. A canonical choice for an  $L_1$ -homeomorphic estimand can be obtained by the “evaluation of all individual events”. Therefore let  $X$  be a finite random sample of length  $n$ , that generates the sigma algebra  $\mathcal{A}^n$ . Then any  $P \in \mathcal{M}$  induces a canonical probability distribution over  $\mathcal{A}^n$  by its product measure, which is known as the *likelihood function*.

**Definition** (Likelihood). *Let  $(S, \Sigma, \mathcal{M})$  be a continuous statistical model and  $X$  a finite random sample in  $S$  of length  $n$ , which generates the  $\sigma$ -Algebra  $\mathcal{A}^n$ . Then the likelihood of  $P \in \mathcal{M}$  w.r.t.  $A \in \mathcal{A}^n$  is given by:*

$$L(P | X \in A) := \prod_{i=1}^n P(X_i \in A_i)$$

Due to its definition the likelihood function is symmetric with regard to realizations  $A \in \mathcal{A}^n$ . Furthermore the operator  $L: P \mapsto L(P | X \in A)$  is  $L_1$ -continuous and bijective w.r.t. its image within  $L_1(S^n, \mathcal{A}^n)$  and since conversely also the projection  $\pi_1: L(P | X \in A) \mapsto P(X_1 \in A_1)$  is  $L_1$ -continuous and bijective it follows that  $L$  is an  $L_1$ -homeomorphism. This shows, that the likelihood function is a continuous estimand which induces the  $L_1$ -topology. In the following it is shown, that the

**Corollary 14.** *Let  $(S, \Sigma, \mathcal{M})$  be a statistical model. Then the likelihood function induces the unique coarsest topology  $\mathcal{T}$  over  $\mathcal{M}$ , that makes  $\mathcal{M}$  continuous w.r.t. any estimands over  $(S, \Sigma)$ .*

*Proof.* By choosing the likelihood function  $L$  as an estimand Theorem 13 postulates the existence of

a unique coarsest topology  $\mathcal{T}$ , that makes  $L$  continuous. Then any  $P, Q \in \mathcal{M}$  are topological distinguishable in  $\mathcal{T}$ , iff there exists an  $A \in \Sigma$ , such that

$$L(P | A) \neq L(Q | A)$$

By applying the definition of the likelihood function this is equivalent to the condition, that  $P(A) \neq Q(A)$ , which proves that  $\mathcal{T}$  preserves the observational distinguishability of probability distributions in  $\mathcal{M}$  over  $\Sigma$ .  $\square$

**Definition** (Canonical topology). *Let  $(S, \Sigma, \mathcal{M})$  be a statistical model and  $\mathcal{T}$  the unique coarsest topology, that makes  $\mathcal{M}$  continuous w.r.t.  $(S, \Sigma)$ . Then  $\mathcal{T}$  is termed the canonical topology of  $\mathcal{M}$  over  $(S, \Sigma)$ .*

The canonical topology of a statistical model, describes a topology of the model space, that preserves the “proximity” of probability distributions with regard to their evaluation over the sample space. 14 thereby postulates, that the canonical topology is unique and therefore an intrinsic structure of a statistical model. This property gives rise to study the statistical equivalence of statistical models in terms of topological spaces. This description provides the notation of *topological statistical models*.

**Definition** (Topological statistical model). *Let*

$$(S, \Sigma, \mathcal{M}) \in \text{ob}(\mathbf{Stat})$$

*and let  $\mathcal{T}$  be a topology of  $\mathcal{M}$  over  $S$ . Then the 4-tuple*

$$(S, \Sigma, \mathcal{M}, \mathcal{T})$$

*is termed a topological statistical model and canonical, iff  $\mathcal{T}$  is the canonical topology of  $\mathcal{M}$  over  $(S, \Sigma)$ .*

In the purpose to describe statistical equivalence in the context of topological statistical models it

comes naturally to mind to utilize homeomorphisms. Thereby these homeomorphism are only required to imply observational distinguishable probability distributions. In the context of Statistical Morphisms, in Theorem 11 this circumstance has *mutatis mutandis* been satisfied by a formulation, that uses the quotient spaces  $\mathcal{M}/\text{id}_{\mathcal{A}}$  and  $\mathcal{N}/\text{id}_{\mathcal{B}}$ . These quotients naturally extend to quotient topologies of their respective observable topologies by  $\mathcal{T}_{\mathcal{A}}/\text{id}_{\mathcal{A}}$  and  $\mathcal{T}_{\mathcal{B}}/\text{id}_{\mathcal{B}}$ . Since the probability distributions within a model space however are observational identical iff they are topological indistinguishable, the corresponding quotient spaces of  $(\mathcal{M}, \mathcal{T}_{\mathcal{A}})$  and  $(\mathcal{N}, \mathcal{T}_{\mathcal{B}})$  may also directly be defined by their topological indistinguishability. This provides the notation of *Kolmogorov quotients*.

**Definition** (Kolmogorov quotient). *Let  $(\mathcal{M}, \mathcal{T})$  be a topological space. Then the Kolmogorov quotient  $\text{KQ}(\mathcal{M}, \mathcal{T})$  denotes the quotient space of  $(\mathcal{M}, \mathcal{T})$  w.r.t. the equivalence of topological indistinguishability.*

**Lemma 15.** *Let  $(X, \mathcal{A}, \mathcal{P}, \mathcal{T}_{\mathcal{A}})$  and  $(Y, \mathcal{B}, \mathcal{Q}, \mathcal{T}_{\mathcal{B}})$  be topological statistical models and  $T: (X, \mathcal{A}) \rightarrow (Y, \mathcal{B})$  a measurable function. Then the following statements are equivalent:*

(i) *There exists a Statistical Morphism*

$$f: \mathcal{P} \rightarrow \mathcal{Q}$$

(ii) *There exists a linear operator*

$$f: L^1(X, \mathcal{A}, \mathcal{P}) \rightarrow L^1(Y, \mathcal{B}, \mathcal{Q})$$

*such that*

$$f: \text{KQ}(\mathcal{P}, \mathcal{T}_{\mathcal{A}}) \rightarrow \text{KQ}(\mathcal{Q}, \mathcal{T}_{\mathcal{B}})$$

*is continuous*

*Proof.* “(i)  $\implies$  (ii)”: Let  $f$  be a Statistical Morphism. Then the definition of  $f$  implicates the existence of a transition operator

$$\tau: L^1(X, \mathcal{A}, \mathcal{P}) \rightarrow L^1(Y, \mathcal{B}, \mathcal{Q})$$

such that  $f = \tau|_{\mathcal{P}}$ . Since  $\tau$  is a continuous linear operator it follows, that for  $g := \tau$  also

$$g: \text{KQ}(\mathcal{P}, \mathcal{T}_{\mathcal{A}}) \rightarrow \text{KQ}(\mathcal{Q}, \mathcal{T}_{\mathcal{B}})$$

is continuous. “(ii)  $\implies$  (i)”: Let

$$f: \text{KQ}(\mathcal{P}, \mathcal{T}_{\mathcal{A}}) \rightarrow \text{KQ}(\mathcal{Q}, \mathcal{T}_{\mathcal{B}})$$

be continuous. By assuming, that  $f$  would be not be continuously linearised with regard to  $T$ , then  $f$  would not be continuous, which contradicts to the preliminary condition. Let therefore

$$g: \text{KQ}(\mathcal{Q}, \mathcal{T}_{\mathcal{B}}) \rightarrow \text{KQ}(\mathcal{Q}, \mathcal{T}_{\mathcal{B}})$$

be a homeomorphism, such that  $g \circ f$  commutes with  $T$ ,  $\pi_{\mathcal{A}}: \mathcal{P} \rightarrow \mathcal{P} / \text{id}_{\mathcal{A}}$  a natural projection and  $h := g \circ f \circ \pi_{\mathcal{A}}$ . Then  $h$  commutes with  $T$ ,  $\text{dom}(h) = \mathcal{P}$  and  $\text{img}(h) = \mathcal{Q}$ , such that  $h: \mathcal{P} \rightarrow \mathcal{Q}$  is a Statistical Morphism.  $\square$

**Definition** (Kolmogorov equivalence). *Let  $(\mathcal{P}, \mathcal{T}_{\mathcal{A}})$  and  $(\mathcal{Q}, \mathcal{T}_{\mathcal{B}})$  be topological spaces, then  $(\mathcal{P}, \tau_{\mathcal{A}})$  and  $(\mathcal{Q}, \tau_{\mathcal{B}})$  are termed Kolmogorov equivalent, iff there exists a homeomorphism*

$$f: \text{KQ}(\mathcal{P}, \mathcal{T}_{\mathcal{A}}) \rightarrow \text{KQ}(\mathcal{Q}, \mathcal{T}_{\mathcal{B}})$$

**Theorem 16.** *For the statistical population  $(\Omega, \mathcal{F}, \text{P})$  let  $(X, \mathcal{A}, \mathcal{P})$  and  $(Y, \mathcal{B}, \mathcal{Q})$  be statistical models and  $\mathcal{T}_{\mathcal{A}}$  and  $\mathcal{T}_{\mathcal{B}}$  their corresponding canonical topologies. Then the following statements are equal:*

- (i)  $(X, \mathcal{A}, \mathcal{P})$  and  $(Y, \mathcal{B}, \mathcal{Q})$  are statistical equivalent
- (ii)  $\text{KQ}(\mathcal{P}, \mathcal{T}_{\mathcal{A}})$  and  $\text{KQ}(\mathcal{Q}, \mathcal{T}_{\mathcal{B}})$  are linear isomorphic and homeomorphic

*Proof.* “(i)  $\implies$  (ii)”: Let  $(X, \mathcal{A}, \mathcal{P})$  and  $(Y, \mathcal{B}, \mathcal{Q})$  be statistical equivalent. Then there exists a Statistical Morphism  $\mathfrak{T}: \mathcal{P} \rightarrow \mathcal{Q}$ , as well as a dual Statistical Morphism  $\mathfrak{T}^*: \mathcal{P} \rightarrow \mathcal{Q}$ , such that:

$$(\mathfrak{T}^* \circ \mathfrak{T})(P) \in \text{id}_{\mathcal{A}}(P), \forall P \in \mathcal{P}$$

and

$$(\mathfrak{T}^* \circ \mathfrak{T})(Q) \in \text{id}_{\mathcal{B}}(Q), \forall Q \in \mathcal{Q}$$

For

$$f: \text{KQ}(\mathcal{P}, \mathcal{T}_{\mathcal{A}}) \rightarrow \text{KQ}(\mathcal{Q}, \mathcal{T}_{\mathcal{B}})$$

$$f(P) := \mathfrak{T}(P), \forall P \in \mathcal{P} / \text{id}_{\mathcal{A}}$$

and

$$f^*: \text{KQ}(\mathcal{Q}, \mathcal{T}_{\mathcal{B}}) \rightarrow \text{KQ}(\mathcal{P}, \tau_{\mathcal{A}})$$

$$f^*(Q) := \mathfrak{T}^*(Q), \forall Q \in \mathcal{Q} / \text{id}_{\mathcal{B}}$$

it then follows, that:

$$(f^* \circ f)(P) = P, \forall P \in \mathcal{P} / \text{id}_{\mathcal{A}}$$

$$(f \circ f^*)(Q) = Q, \forall Q \in \mathcal{Q} / \text{id}_{\mathcal{B}}$$

such that  $f$  is invertible. Since due to Lemma 15  $f$ , as well as its inverse  $f^*$  are continuous,  $f$  is a homeomorphism. This proves, that there exists a homeomorphism

$$f: \text{KQ}(\mathcal{P}, \mathcal{T}_{\mathcal{A}}) \rightarrow \text{KQ}(\mathcal{Q}, \mathcal{T}_{\mathcal{B}})$$

and therefore that  $(\mathcal{P}, \mathcal{T}_{\mathcal{A}})$  and  $(\mathcal{Q}, \mathcal{T}_{\mathcal{B}})$  are Kolmogorov equivalent .

“(ii)  $\implies$  (i)”: Let now  $(\mathcal{P}, \mathcal{T}_{\mathcal{A}})$  and  $(\mathcal{Q}, \mathcal{T}_{\mathcal{B}})$  be Kolmogorov equivalent, then there exists a homeomorphism

$$f: \text{KQ}(\mathcal{P}, \mathcal{T}_{\mathcal{A}}) \rightarrow \text{KQ}(\mathcal{Q}, \mathcal{T}_{\mathcal{B}})$$

Let

$$T_X: (\Omega, \mathcal{F}, \text{P}) \rightarrow (X, \mathcal{A})$$

$$T_Y: (\Omega, \mathcal{F}, \text{P}) \rightarrow (Y, \mathcal{B})$$

be the statistics, that generate the sample spaces. Then  $T := T_Y \circ T_X^{-1}$  is a measurable function with:

$$T: (X, \mathcal{A}) \rightarrow (Y, \mathcal{B})$$

Then due to the proof of Lemma 15 there exists a homeomorphism

$$g: \text{KQ}(\mathcal{Q}, \mathcal{T}_{\mathcal{B}}) \rightarrow \text{KQ}(\mathcal{Q}, \mathcal{T}_{\mathcal{B}})$$

such that  $g \circ f$  defines a Statistical Morphism

$$g \circ f: \mathcal{P} / \text{id}_{\mathcal{A}} \rightarrow \mathcal{Q} / \text{id}_{\mathcal{B}}$$

which commutes with  $T$ . Since  $f$ , as well as  $g$  are homeomorphisms also  $g \circ f$  is a homeomorphism

and its inverse  $(g \circ f)^{-1}$  is continuous. Let now

$$\pi_A: \mathcal{P} \rightarrow \mathcal{P} / \text{id}_A$$

$$\pi_B: \mathcal{Q} \rightarrow \mathcal{Q} / \text{id}_B$$

be natural projections that satisfy:

$$\mathfrak{T} := g \circ f \circ \pi_A$$

$$\mathfrak{T}^* := (g \circ f)^{-1} \circ \pi_B$$

Then  $\mathfrak{T}$  and  $\mathfrak{T}^*$  are Statistical Morphisms with:

$$(\mathfrak{T}^* \circ \mathfrak{T})(P) \in \text{id}_A(P), \forall P \in \mathcal{P}$$

$$(\mathfrak{T}^* \circ \mathfrak{T})(Q) \in \text{id}_B(Q), \forall Q \in \mathcal{Q}$$

This proves, that  $\mathfrak{T}: \mathcal{P} \rightarrow \mathcal{Q}$  is a statistical isomorphism.  $\square$

## 5. HOMOLOGY

The natural correspondence between a statistical model and its corresponding canonical topological statistical model, provides a characterisation of statistical equivalence with regard to the underlying topology. This property allows an intuitive description of many classical concepts of statistics, like sufficiency and completeness by homeomorphic embeddings and surjective continuous functions. Apart from its characterising qualities however a generic structural representation provides the opportunity to incorporate a priori knowledge in terms of *structural assumptions*. As this knowledge usually regards the structure of the statistical population, it has to be transferred to the statistical model. In particular for statistical models, which are based on natural observations, the statistic  $T$  thereby does not preserve all statistical parameters, such that also the observable structure only partially represents the structure of the underlying statistical population. In order to incorporate structural assumptions within the statistical model it is therefore useful to embed the statistical model within an “extended statistical model”, that is able to represent the

structure of the statistical population. Then structural assumptions regarding the statistical population correspond to constraints to the structural embedding and therefore implicitly effect the observable structure. Although the structural embedding theoretically does not demand the interpretation as a statistical model, it facilitates the transfer of concepts. In this purpose, an underlying sample space is constructed, to extend the sample space, which is given by the statistic  $T$ , by the set of unobserved statistical parameters. This provides the notation of a partially observable measurable space.

**Definition** (Partially observable measurable space).

Let  $T: (\Omega, \mathcal{F}, P) \rightarrow (S, \Sigma)$  be a statistic. Then a measurable space  $(E, \mathcal{E})$  is termed *partially observable by  $T$* , iff there exists a measurable space  $(H, \mathcal{H})$ , such that  $(E, \mathcal{E}) = (S \times H, \Sigma \otimes \mathcal{H})$ .

Due to this definition any  $E$ -valued random variable over  $(\Omega, \mathcal{F}, P)$  splits into a  $S$ -valued *observable* random variable  $v$  and a  $H$ -valued *latent* random variable  $h$ . Furthermore the probability distributions over  $(E, \mathcal{E})$  are defined by marginal probability distributions  $\mathcal{P}$  over  $(S, \Sigma)$  and  $\mathcal{Q}$  over  $(H, \mathcal{H})$ , as well as by regular conditional probabilities  $p(v | h, P, Q)$  and  $p(h | v, P, Q)$ , where  $P \in \mathcal{P}$  and  $Q \in \mathcal{Q}$ . This allows the definition of an extended statistical model, which underlying sample space also incorporates unobserved statistical parameters by latent random variables.

**Definition** (Latent variable model). Let  $(E, \mathcal{E})$  be a partially observable measurable space and  $\mathcal{M}$  a set of probability distributions over  $(E, \mathcal{E})$ . Then the tuple  $(E, \mathcal{E}, \mathcal{M})$  is termed a *latent variable model*.

The embedding of a given statistical model into a latent variable statistical model that represents the statistical population provides an intuitive description of the structural properties of the statistical model. In particular however structural assumptions do not only effect “*known unknowns*”,

as in the case of individual distributions assumptions, but indeed “*unknown unknowns*”. This is of particular importance for the statistical modeling of complex dynamical systems and especially for living cells, where it has to be assumed, that the underlying biomolecular mechanisms have only partially been identified. In order to provide statistical inference in living cells hence structural assumptions are required, that are able (i) to generate tractable model spaces and (ii) to facilitate statistical inference. Apart of the theoretical elegance of a structural embedding this approach at first sight however seems rather impractical, since the canonical topology is usually not tangible to an observer. Indeed without any structural knowledge, the determination of the canonical topology would require a complete observation of the underlying statistical model. The situation is quite different, if the extended statistical model may be embedded within a topological vector space  $V$ . Then (i) the subspace topology of the embedding of its observable random variables identifies the canonical topology and (ii) the algebraic structure of the vector space facilitates the traversal of the extended model space and therefore statistical inference. This embedding however has to be invariant with regard to statistical equivalent statistical models. To this end in a preceding step, a surjective Statistical Morphism  $\pi: \mathcal{M}/\text{id}_\Sigma \twoheadrightarrow \mathcal{M}_\theta$ , termed a *statistical projection* maps the model space  $\mathcal{M}$  onto a *parametric family*  $\mathcal{M}_\theta \subseteq \mathcal{M}$  and a subsequent structural embedding  $e: \mathcal{M}_\theta \hookrightarrow V$  into a vector space  $V$ . This defines a *parametrisation* of  $\mathcal{M}$ .

**Definition** (Parametrisation). *Let  $(S, \Sigma, \mathcal{M})$  be a statistical model,  $V$  a  $\mathbb{K}$ -vector space,*

$$\pi: \mathcal{M}/\text{id}_\Sigma \twoheadrightarrow \mathcal{M}_\theta \subseteq \mathcal{M}$$

*a statistical projection and*

$$e: \mathcal{M}_\theta \hookrightarrow \Theta \subseteq V$$

*an embedding. Then a mapping*

$$\theta: V \rightarrow \mathcal{M}/\text{id}_\Sigma$$

*is termed a parametrisation of  $\mathcal{M}$  over  $V$ , iff the following diagram commutes:*

$$(5.1) \quad \begin{array}{ccc} & \theta & \\ V & \xrightarrow{\quad} & \mathcal{M}/\text{id}_\Sigma \\ & \searrow e & \swarrow \\ & \mathcal{M}_\theta & \end{array}$$

Generally a parametrisation  $\theta$  is not required to be a function in terms of unique image elements, but only w.r.t. the preimage of  $e \circ \pi$ , such that  $\text{img}\theta = \mathcal{M}/\text{id}_\Sigma$ . Therefore an arbitrary parametrisation does not provide the ability to identify individual probability distributions by different parameters. As the probability distributions in the parametric family  $\pi(\mathcal{M}) = \mathcal{M}_\theta \subseteq \mathcal{M}$  however are trivially observational distinguishable, the additional claim that  $\pi$  is not only a Statistical Morphism, but indeed a statistical isomorphism provides that  $\mathcal{M}/\text{id}_\Sigma \simeq \mathcal{M}_\theta/\text{id}_\Sigma$  and since  $\mathcal{M}_\theta/\text{id}_\Sigma = \mathcal{M}_\theta$  it follows, that  $\mathcal{M}/\text{id}_\Sigma \simeq \mathcal{M}_\theta$ . Since furthermore embeddings  $e: \mathcal{M}_\theta \hookrightarrow V$  are injective it also holds that  $\mathcal{M}/\text{id}_\Sigma \simeq \text{img}(e \circ \pi)$ , and therefore with regard to the *parameter space*  $\Theta := \text{dom}\theta = \text{img}(e \circ \pi)$ , that  $\mathcal{M}/\text{id}_\Sigma \simeq \Theta$ . Then for any probability distributions  $P, Q \in \mathcal{M}$ , that are parametrised by parameter vectors  $\theta_P, \theta_Q \in \Theta$  with  $\theta_P \neq \theta_Q$  it follows that  $P \notin \text{id}_\Sigma Q$  and therefore, that  $P$  and  $Q$  are observational distinguishable. Consequentially the requirement of the projection  $\pi$  to be a statistical isomorphism implicates that different parameter vectors provide different probability distributions. Then  $\theta$  is termed *identifiable*.

**Definition** (Identifiable parametrisation). *Let*

$$(S, \Sigma, \mathcal{M}) \in$$

*be a statistical model,  $V$  a  $\mathbb{K}$ -vector space and  $\theta$  a parametrisation of  $\mathcal{M}$  over  $V$ . Then  $\theta$  is termed*



identifiable, iff its underlying statistical projection  $\pi$  is a statistical isomorphism.

A statistical model  $(S, \Sigma, \mathcal{M})$  is termed *identifiable* if there exists an identifiable parametrisation  $\theta$  over a vector space  $V$ . In this case it follows, that  $\mathcal{M}/\text{id}_\Sigma \simeq \mathcal{M}_\theta/\text{id}_\Sigma = \mathcal{M}_\theta$  and since the diagram 5.1 is required to commute, that  $\mathcal{M}_\theta = \mathcal{M}$ . Consequently the notation  $(S, \Sigma, \mathcal{M}_\theta)$  naturally implicates an identifiable parametrisation  $\theta$  and therefore an identifiable statistical model. In this case it follows that  $\Theta = \text{dom}\theta \simeq \text{img}\theta \simeq \mathcal{M}_\theta$ , such that any probability distribution  $P_\theta \in \mathcal{M}_\theta$  uniquely identifies a parameter vector  $\theta_P \in \Theta$  by  $\theta_P := \theta^{-1}(P)$  and vice versa by  $P_\theta := \theta(\theta_P)$ . Then the parameter space  $\Theta$  is a *parametric representation* of the statistical model and identifiable statistical models may be compared by their coincidence in a common parametric representation. Thereby statistical models  $(X, \mathcal{A}, \mathcal{M})$  and  $(Y, \mathcal{B}, \mathcal{N})$  are *equivalent in representation*, iff there exist identifiable parametrisations  $\theta_{\mathcal{M}}: \Theta_{\mathcal{M}} \rightarrow \mathcal{M}$  and  $\theta_{\mathcal{N}}: \Theta_{\mathcal{N}} \rightarrow \mathcal{N}$  such that  $\Theta_{\mathcal{M}} = \Theta_{\mathcal{N}}$ . This statistically invariant property however is completely described by the cardinality of their identifiable parametrisations.

**Example** (Cardinality of a parametrisation). *Let  $(S, \Sigma, \mathcal{M})$  be a statistical model,  $V$  a vector space, and  $\theta: \Theta \rightarrow \mathcal{M}$  a parametrisation of  $\mathcal{M}$  over  $V$ . Then the cardinality of  $\theta$  is given by the cardinality of the parameter space  $\Theta$ .*

Since the cardinality of an identifiable parametrisation quantifies the number of observational distinguishable probability distributions within the underlying model space, statistical models are equivalent in representation, iff they have identifiable parametrisations with equal cardinality. Then there exist parametrisations  $\theta_{\mathcal{M}}$  and  $\theta_{\mathcal{N}}$ , and such that  $\mathcal{M} = (\theta_{\mathcal{M}} \circ \theta_{\mathcal{N}}^{-1})(\mathcal{N})$  and  $\mathcal{N} = (\theta_{\mathcal{N}} \circ \theta_{\mathcal{M}}^{-1})(\mathcal{M})$ . In this case the functions  $\theta_{\mathcal{N}\mathcal{M}} = \theta_{\mathcal{M}} \circ \theta_{\mathcal{N}}^{-1}$  and  $\theta_{\mathcal{M}\mathcal{N}} = \theta_{\mathcal{N}} \circ \theta_{\mathcal{M}}^{-1}$  are termed *re-parametrisations*. Thereby it

is important to notice, that without further requirements re-parametrisations only preserve the primordial distinguishability of distribution assumptions but not their proximity structure, such that distribution assumptions, which exclude each other in one statistical model may be nearly equivalent in another. More for topological statistical model  $(\mathcal{M}, \tau)$  and a topological vector space  $(V, \tau_V)$  an identifiable parametrisation  $\theta: V \subseteq \Theta \rightarrow \mathcal{M}$ , does not provide statistical equivalence with regard to the subspace topology  $\tau_\Theta$  of  $\Theta$  in  $V$ , but only to the induced topology  $\theta^{-1}(\tau)$ . Therefore re-parametrisations have to be claimed to preserve the topology, in order to obtain statistical equivalence. As  $\tau$  however is generally not tractable by an observer it is more reasonable to require the individual parametrisations to be continuous. This provides the notation of a *continuous parametrisation*.

**Definition** (Continuous parametrisation). *Let*

$$(X, \mathcal{A}, \mathcal{M}, \mathcal{T})$$

*be a topological statistical model,  $(V, \mathcal{T}_V)$  a topological vector space and  $\theta$  a parametrisation of  $\mathcal{M}$  over  $V$ . Then  $\theta$  is continuous, iff*

$$\theta^{-1}: \text{KQ}(\mathcal{M}, \mathcal{T}) \rightarrow (V, \mathcal{T}_V)$$

*is a continuous embedding.*

By assuming a continuous identifiable parametrisation  $\theta: \Theta \rightarrow \mathcal{M}$ , then the parameter space  $\Theta$  naturally extends to a *continuous representation*  $(\Theta, \mathcal{T}_\Theta)$  of  $(\mathcal{M}, \mathcal{T})$ , where  $\mathcal{T}_\Theta$  denotes the subspace topology of  $\Theta$  in  $V$ . Then also

$$\theta: (\Theta, \mathcal{T}_\Theta) \rightarrow \text{KQ}(\mathcal{M}, \mathcal{T})$$

is a homeomorphism and since

$$\text{KQ}(\Theta, \mathcal{T}_\Theta) = (\Theta, \mathcal{T}_\Theta)$$

Theorem 16 proposes the existence of a statistical isomorphism between  $\Theta$  and  $\mathcal{M}$ , such that statistical inference may also be derived within  $(\Theta, \mathcal{T}_\Theta)$  as

a subspace of  $V$ . Moreover any further statistical model, that has a continuous representation, which is homeomorphic to  $(\Theta, \mathcal{T}_\Theta)$  is statistical equivalent to  $(\mathcal{M}, \mathcal{T})$ , such that  $(\Theta, \mathcal{T}_\Theta)$  is a continuous representation of both models. In particular identifiable statistical models are statistical equivalent, iff they are equivalent in a common continuous representation. Therefore structure preserving re-parametrisations naturally generalize statistical equivalence. Thereby the re-parametrisations represent isomorphisms of a *structural category*  $C$ , which provides the notation of a generic *structural equivalence*.

**Definition** (Structural equivalence). *Let  $(X, \mathcal{A}, \mathcal{M})$  and  $(Y, \mathcal{B}, \mathcal{N})$  be statistical models,  $C$  a category and  $(\mathcal{M}, \varphi)$  and  $(\mathcal{N}, \vartheta)$  objects in  $C$ . Then  $(X, \mathcal{A}, \mathcal{P})$  and  $(Y, \mathcal{B}, \mathcal{Q})$  are structural equivalent w.r.t  $C$ , iff  $(\mathcal{M}, \varphi) \stackrel{C}{\simeq} (\mathcal{Q}, \vartheta)$ .*

As the structural equivalence of statistical models may be defined for arbitrary underlying structural categories, it does not necessarily correspond with statistical equivalence. In the purpose of statistical inference however this correspondence has to be obtained by structural assumptions. For example if the statistical models are assumed to be identifiable, then due to Theorem 16 statistical equivalence corresponds to Kolmogorov equivalence of identifiable topological statistical models. In this case the subset of statistical isomorphisms that satisfy the requirement of identifiability are precisely the homeomorphic re-parametrisations of “identifiable topological statistical models”. Then arbitrary additional assumptions at least have to preserve the topology of continuous representations, such that statistical inference may completely be derived by the evaluation of estimands along paths within continuous representations. The great advantage of “structural equivalence” over “statistical equivalence” however is, that it is not restricted

to observable structures. This means, that structural assumptions regarding the statistical population may also implicitly affect the observable structure. Then statistical inference may be obtained by a projection of an structural estimator to the “closest” observable distribution assumption. However the structural assumptions thereby have to be “compatible” with the observable structure and therefore have to be statistical invariants, as well as structural invariants of the structural category. An example of such a *structural property* of a statistical model has already been given by the cardinality of an identifiable parametrisation, which for the category of “identifiable topological statistical models” provides the number of observational distinguishable probability distributions. The cardinality however is unsuitable to distinguish statistical models that comprise infinite observational distinguishable distribution assumptions. Nevertheless in the very same manner as the cardinality of an identifiable parametrisation determines the number of distinguishable probability distributions, the minimal length of structure preserving identifiable parametrisations determines the number of distinguishable dimensions.

**Example** (Length of a parametrisation). *Let  $(\mathcal{M}, \mathcal{T})$  be a topological statistical model,  $V$  a vector space, and  $\theta: \Theta \rightarrow \mathcal{M}$  a parametrisation of  $\mathcal{M}$  over  $V$ . Then the length of  $\theta$  is given by the dimension of the parameter space  $\Theta$  in  $V$ .*

The length of a given identifiable parametrisation is obviously not a structural property of a statistical model, since it depends on the chosen parametrisation. The minimal length of a structure preserving identifiable parametrisation however, only depends on the structural category, that has to be preserved. Therefore it provides a reasonable assumption, under the precondition of an underlying category. If this structural category is then induced by a structural assumption, then the

minimal length of a structure preserving identifiable parametrisation is indeed a statistical property. This relationship therefore provides a justification of *minimal parametrisations* as canonical representations of parametric families.

**Definition** (Minimal parametrisation). *Let  $(\mathcal{M}, \mathcal{T})$  be a topological statistical model,  $C$  a category and  $\theta: \Theta \rightarrow \mathcal{M}$  a parametrisation of  $\mathcal{M}$  over  $V$ . Then  $\theta$  is termed minimal w.r.t  $C$  if (i)  $\theta$  is identifiable, (ii)  $\theta$  is an isomorphism within  $C$  and (iii)  $\theta$  has a minimal length under the previous constraints.*

In the purpose to derive a minimal parametrisation, the choice of the structural category is crucial. For example by regarding model spaces as sets, then in the structural category **Set** the reparametrisations are bijections and a minimal parametrisation always has length 1, as long as the cardinality of the parametrisation is smaller than that of the vector space  $V$ , otherwise there exists no identifiable parametrisation over  $V$ . For many other structural categories however, the derivation of a minimal parametrisation is only hardly tractable. Then for a given statistical model and a given vector space  $V$  the question arises, if at least an identifiable parametrisation of  $\mathcal{M}$  over  $V$  exists with finite length. For example if the structural category extends the canonical topology by a global Euclidean structure, then the structural isomorphisms are linear mappings and the length of a minimal parametrisation equals the minimal number of  $V$ -valued, linear independent random variables, that are needed to describe any probability distribution in  $\mathcal{M}$ . If  $\mathcal{M}$  then contains at least one probability distribution, which is not finite dimensional over  $V$ , then there exists no identifiable parametrisation with finite length. In particular with a growing degree of structure it is getting more difficult, or even impossible to obtain structure preserving embeddings within vector spaces

and therefore to find appropriate identifiable parametrisations. In the purpose to derive statistical inference by structural inference it is therefore crucial to study the intrinsic structural properties of statistical models.

## REFERENCES

- [1] I. Csisz r. Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoffschen Ketten. *Magyar. Tud. Akad. Mat. Kutato Int. Kozl.*, (8):85–108, 1963.
- [2] Ronald Aylmer Fisher. The Logic of Inductive Inference. *Journal of the Royal Statistical Society*, 98(1):39–82, 1935.
- [3] Solomon Kullback. *Information Theory and Statistics*, volume 1. Wiley, 1959.
- [4] Solomon Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- [5] T. Morimoto. Markov processes and the H-theorem. *J. Phys. Soc. Jpn*, 18(3):328–331, 1963.