

Research Article

# Quantum Probability for Statisticians: Some New Ideas

Inge Svein Helland<sup>1</sup>

1. Department of Mathematics, University of Oslo, Norway

It is argued from several points of view that quantum probabilities might play a role in statistical settings. New approaches toward quantum foundations have postulates that appear to be equally valid in macroscopic settings. One such approach is described here in detail, while two others are briefly sketched. In particular, arguments behind the Born rule, which gives the basis for quantum probabilities, are given. A list of ideas for possible statistical applications of quantum probabilities is provided and discussed. A particular area is machine learning, where there exists substantial literature on links to quantum probability. Here, an idea about model reduction is sketched and is motivated from a quantum probability model. Quantum models can play a role in model reduction, where the partial least squares regression model is a special case. It is shown that for certain experiments, a Bayesian prior given by a quantum probability can be motivated. Quantum decision theory is an emerging discipline that can be motivated by this author's theory of quantum foundations.

Corresponding author: Inge S. Helland, [ingeh@math.uio.no](mailto:ingeh@math.uio.no)

## 1. Introduction

The basis for nearly all articles in theoretical and applied statistics is Kolmogorovian probability<sup>[1]</sup>. Quantum probability is mostly looked upon by statisticians as an exotic tool with no relevance for statistical science and statistical inference. This implies that, in the statistical literature, there is very little discussion of possible links between statistical theory and quantum theory. (A good exception is the article by Barndorff-Nielsen et al.<sup>[2]</sup>, where quantum versions of exponential models, sufficiency, and Fisher information are discussed, among other related topics.)

In my opinion, this has led to a limitation of statistical science, both from a theoretical point of view and from an applied point of view. I relate this statement to the idea that traditional statistical theory only - at least in most cases - looks upon the parameter space as a set with no structure. Introducing more structure into the parameter space will give a richer theory. And I will show in the next Section that some such structures indeed lead to quantum probabilities. Later in this article, I will discuss a specific structure, symmetry introduced by letting a concrete group act on the parameter space, and the implications for model reduction.

It seems as if a discussion of links between quantum theory and statistical theory is more than due now. This can be underpinned by looking at the current revolution in artificial intelligence, in particular, machine learning, an area that engages more and more statisticians. It is probably not so well known among statisticians that discussions of connections between machine learning and quantum mechanics have appeared in the literature now; see, for instance, the review article by Dunjko & Briegel<sup>[3]</sup>. If we really mean that machine learning should be based upon statistics, connections between statistical theory and quantum theory should also be of high interest.

One can ask then: Is not quantum probabilities just of relevance to the microcosmos? My answer to this is no. To argue for that answer, one can look at recent derivations of quantum mechanics from various postulates. In two books<sup>[4][5]</sup> and in a series of articles<sup>[6][7][8][9][10][11][12][13][14][15]</sup>, this author has proposed a completely new foundation of quantum theory, and also, in this connection, discussed the interpretation of the theory. This foundation is summarized in Subsections 2.2 and 2.3 below, where the essence of quantum theory is deduced from 7 postulates. Looking at these postulates, most of them may be seen to be equally valid in a macroscopic setting.

Another basis for quantum theory is given in a series of articles by De Raedt and his collaborators. One of these is De Raedt et al.<sup>[16]</sup>. Here, essential elements of quantum mechanics are derived from what is called logical inference to experiments, an assumption that there are uncertainties about individual events and that the frequencies are robust with respect to small changes in the parameters. A basic tool in the derivations is that of Fisher information. Fisher information as a general basis for nearly all aspects of science has earlier been advocated by Frieden<sup>[17]</sup>.

Finally, it is of interest in this connection to look at the quantum-like models introduced by Khrennikov and collaborators; see Khrennikov<sup>[18][19]</sup> and Haven & Khrennikov<sup>[20][21]</sup>. These models are based upon quantum probabilities and are applied to several sciences, including cognitive

psychology, sociology, finance, and biology. An important element here is a theory of quantum decisions; see also Helland<sup>[10][14]</sup>.

While there are few examples of statistical papers related to quantum probability, there are several articles in the quantum theory literature that use basic statistical concepts and ideas. Very much discussion among theoretical physicists is concerned with the interpretation of quantum theory. Some still want to look upon it as an ontological theory, but more and more physicists now conclude with the view that quantum mechanics should be interpreted as an epistemic or epistemological theory: It is a theory of our knowledge of the world, in the same way that statistical theory is basically about knowledge from experiments and observations. A strong school in the quantum community is QBism, originally an abbreviation for Quantum Bayesianism, but now by its founders claimed to have a somewhat wider basis, see Caves et al.<sup>[22]</sup> and Fuchs<sup>[23][24]</sup>. A very recent article on Bayes' rule and related inference in quantum mechanics is Liu<sup>[25]</sup>.

The plan of this article is as follows: In Section 2, my own basis for quantum theory is discussed in some detail, together with other bases. One aim is to motivate the use of quantum probabilities in macroscopic settings. In Section 3, several potential ideas for the use of quantum probabilities in statistical settings are discussed. This includes model reductions, quantum probabilities as priors for certain experiments, decision theory, and machine learning. The final section gives some general discussion points.

## 2. Basic theories leading to quantum probabilities

### 2.1. *The traditional approach*

The traditional approach towards quantum theory, as found in numerous textbooks, is rather formal. To understand the main message of this article, it is not necessary to read this subsection in complete detail, but some of the elements discussed here will be used later.

As a start, quantum mechanics is a nearly 100-year-old theory, whose formal foundation is given by von Neumann<sup>[26]</sup>. That book summarizes the efforts of many prominent physicists at the beginning of the last century, physicists whose names need not be mentioned here. The book by von Neumann has, in turn, inspired a large literature, consisting of research articles, monographs, textbooks, and more popular books.

A thorough and more modern formal treatment is given by Hall<sup>[27]</sup>. The basic mathematical concept underlying all the formal literature on quantum mechanics is that of a Hilbert space, a complex linear space of vectors admitting a scalar product. The theory is much simpler when the dimension of the Hilbert space is finite, when it can be taken to be equal to  $\mathbb{C}^r$  for some  $r$ , where  $\mathbb{C}$  is the space of complex numbers. Physical states are represented by normalized vectors in the Hilbert space, and physical variables are represented by operators, in the finite-dimensional case, matrices.

Hall<sup>[27]</sup> gives axioms behind quantum mechanics that are mathematically precise, also for the case where the operators involved are unbounded, which implies for an operator  $A$  that it is only defined on a subset  $D = D_A$  of the Hilbert space. For this case, the adjoint  $A^\dagger$  of the operator  $A$  is defined: A vector  $|\phi\rangle$  belongs to the domain  $D^\dagger$  of  $A^\dagger$  if the linear functional  $\langle\phi, A\cdot\rangle$  defined on  $D$  is bounded, and for  $|\phi\rangle \in D^\dagger$ ,  $|\xi\rangle = A^\dagger|\phi\rangle$  is defined as the unique vector such that  $\langle\xi, \psi\rangle = \langle\phi, A\psi\rangle$  for all  $|\psi\rangle \in D$ , where  $\langle, \rangle$  is the scalar product of the Hilbert space. From this, the concept of a self-adjoint operator is defined by requiring that  $A^\dagger = A$ . A self-adjoint operator is characterized by having a real-valued spectrum, say, real-valued eigenvalues, so, by connecting such an operator to a physical variable, the eigenvalues can be interpreted as the possible values of this variable.

A simpler concept is that of a symmetric operator: An operator  $A$  is symmetric if  $\langle\phi, A\psi\rangle = \langle A\phi, \psi\rangle$  for all  $|\phi\rangle, |\psi\rangle \in D$ . Conditions under which a symmetric operator is self-adjoint are discussed in Chapter 9 of Hall<sup>[27]</sup>. It is noted that also symmetric operators have real-valued eigenvalues. All bounded operators that are symmetric are self-adjoint.

Of some interest to statisticians are also the Lecture Notes by Meyer<sup>[28]</sup> and Holevo<sup>[29]</sup>. What is common to those monographs is that they assume the superposition principle: If  $|\phi\rangle$  and  $|\psi\rangle$  are possible state vectors and  $a$  and  $b$  are complex numbers, then a normalized version of  $a|\phi\rangle + b|\psi\rangle$  is also a possible state vector. In the next subsection, I will limit the concept of state vector to vectors in the Hilbert space that are eigenvectors of some physically or statistically meaningful operator, and then the general superposition principle is not necessarily true. As shown in Helland<sup>[8][12]</sup>, this view can be used to resolve various quantum paradoxes.

In newer versions of this traditional approach to quantum probabilities, a basic notion is that of a POVM (positive operator-valued measurement): Let  $\mathcal{H}$  be a Hilbert space, and  $(\Omega, \mathcal{F})$  a measurable space with  $\mathcal{F}$  being the Borel  $\sigma$ -algebra on  $\Omega$ . A POVM is a function  $F$  defined on  $\mathcal{F}$  whose values are

positive bounded self-adjoint operators on  $\mathcal{H}$  such that for every  $|\psi\rangle \in \mathcal{H}$  the mapping from  $E$  to  $\langle F(E)\psi, \psi \rangle$  is a non-negative countably additive measure on  $\mathcal{F}$ , and  $F(\Omega) = I$  is the identity operator.

In the finite-dimensional case, a POVM can be constructed from a set of semi-definite self-adjoint matrices  $\{M_i\}$  such that  $\sum M_i = I$  by letting  $M_i = F(i)$  for  $i = 1, \dots, r$ . Then for a set of non-negative constants  $p_i$  adding to 1, and for a fixed orthonormal basis  $\{|\psi_i\rangle\}$  of the Hilbert space, define  $\langle M_i\psi_i, \psi_i \rangle = p_i$ . This gives for any subset  $E$  of  $\{1, \dots, r\}$ , and for any  $|\psi\rangle$  in the Hilbert space,  $\langle F(E)\psi, \psi \rangle = \sum_{i \in E} |\alpha_i|^2 p_i$  when  $|\psi\rangle = \sum_i \alpha_i |\psi_i\rangle$ .

A POVM is called a projection valued measure (PVM) if it is self-adjoint and  $F(E)$  is a projection operator for every  $E$ . Naimark's dilation theorem shows how any POVM  $F$  can be obtained from a PVM  $P$  acting on a larger Hilbert space  $\mathcal{K}$ : There is a unitary transformation  $U$  (a transformation such that  $UU^\dagger = I$ ) from  $\mathcal{K}$  to  $\mathcal{H}$  such that  $F(E) = UP(E)U^\dagger$  for every  $E$ .

The key property of a POVM is that  $\langle F(E)\psi, \psi \rangle$  can be interpreted as a probability of outcome  $E$  when measured in the state  $|\psi\rangle$  when this is a normalized vector in the Hilbert space. This is a version of Born's formula.

## 2.2. A new approach; the postulates and the first results

A completely new approach towards quantum foundations is proposed in Helland<sup>[12][13][14][15]</sup>, where the formal properties of quantum mechanics are derived from a rather simple set of postulates. These postulates will be repeated below.

As a possible general interpretation, the basis can be taken to be relative to an actor who is in some fixed (physical or statistical) context. In this context, there are *theoretical variables*, and some of these variables, say  $\theta, \lambda, \eta, \dots$  may be related to the actor  $C$ . Some of these variables are *accessible* to him, which means roughly that it is, in some future, given some estimation principle, in principle possible to obtain as accurate estimates as he wishes on the relevant variable. Other variables are *inaccessible*. In Helland<sup>[12][13][14][15]</sup>, physical examples are given. In the present article, applications to statistics will be the theme, and then in most cases, the theoretical variables will be parameters of some statistical model. However, I will also allow other interpretations of the variables: In special applications, they may be latent variables, future data, or combinations of parameters and data.

The above characterization of accessible and inaccessible variables will, in this article, mainly be related to a statistical implication of the theory. But the theory itself is purely mathematical and can be

made precise in different directions. In particular, the terms ‘accessible’ and ‘inaccessible’ can just be seen as primitive notions of the theory. In addition to the statistical implication, two other ways that the theory can be made precise are 1) ordinary quantum mechanics, where the theoretical variables are physical variables, in my interpretation connected to a fixed context and also to the mind of some actor; 2) quantum decision theory, where the variables are decision variables. From a mathematical point of view, it is only assumed that if  $\lambda$  is a theoretical variable and  $\theta = f(\lambda)$  for some function  $f$ , then  $\theta$  is a theoretical variable. And if  $\lambda$  is accessible, then  $\theta$  is accessible.

As said, in physical applications, the variables may also be connected to the mind of some actor. Note, however, that actors may communicate. The mathematical model developed in the articles mentioned above is equally valid relative to a group of people that can communicate about the various theoretical variables. This gives a new version of the theory, a version where all theoretical variables are defined jointly for such a group of actors. In physical applications, the actor or the communicating group of actors is important. In many statistical applications, we may take the group to be the set of all possible actors; see Section 3 below.

From a mathematical point of view, an accessible variable  $\theta$  is called maximal if there is no other accessible variable  $\lambda$  such that  $\theta = f(\lambda)$  for some non-invertible function  $f$ . In other words, the term ‘maximal’ will then be seen to be maximal with respect to the partial ordering of variables given by  $\alpha \leq \beta$  iff  $\alpha = f(\beta)$  for some function  $f$ .

To be precise, every accessible variable is assumed to vary over some topological space; in most cases, they are real-valued or vector-valued, and all functions discussed are assumed to be Borel-measurable.

A basic assumption in my theory is that there exists an inaccessible variable  $\phi$  such that all the accessible variables can be seen as functions of  $\phi$ . In simple physical applications, such a  $\phi$  may easily be defined explicitly. In statistical applications,  $\phi$  may be some total, inaccessible parameter, say, the set of all parameters that in some way may be included in a certain statistical model.

Two different accessible variables  $\theta$  and  $\eta$  are defined to be *related* if there is a transformation (group action)  $k$  in  $\phi$ -space and a function  $f$  such that  $\theta = f(\phi)$  and  $\eta = f(k\phi)$ .

As a summary of the above discussion, here are the first 3 postulates of the theory:

**Postulate 1.** *If  $\eta$  is a theoretical variable and  $\gamma = f(\eta)$  for some function  $f$ , then  $\gamma$  is also a theoretical variable.*

**Postulate 2.** *If  $\theta$  is accessible to  $C$  and  $\lambda = f(\theta)$  for some function  $f$ , then  $\lambda$  is also accessible to  $C$ .*

**Postulate 3.** *In the given context, there exists an inaccessible variable  $\phi$  such that all the accessible ones can be seen as functions of  $\phi$ . There is a transitive group  $K$  acting upon  $\phi$ .*

A definition is now needed for the fourth postulate:

**Definition 1.** *The accessible variable  $\theta$  is called maximal if  $\theta$  is maximal as an accessible variable under the partial ordering defined by  $\alpha \leq \beta$  iff  $\alpha = f(\beta)$  for some function  $f$ .*

Note that this partial ordering is consistent with accessibility: If  $\beta$  is accessible and  $\alpha = f(\beta)$ , then  $\alpha$  is accessible. Also,  $\phi$  from Postulate 3 is an upper bound under this partial ordering.

**Postulate 4.** *There exist maximal accessible variables relative to this partial ordering. For every accessible variable  $\theta$  there exists a maximal accessible variable  $\lambda$  such that  $\theta$  is a function of  $\lambda$ .*

Then, in my opinion, two different maximal accessible variables come very close to what Bohr called complementary variables; see Plotnitsky<sup>[30]</sup> for a thorough discussion. The term complementary originated in connection to the variables position and momentum, but has now reached a number of applications; see Steiner & Rendell<sup>[31]</sup> and Maccone<sup>[32]</sup> for example.

It is crucial what is meant by ‘different’ here. If  $\theta = f(\eta)$ , where  $f$  is a bijective function, (i.e., there is a one-to-one correspondence between  $\theta$  and  $\eta$ ), then  $\theta$  and  $\eta$  contain the same information, and they must be considered ‘equal’ in this sense.  $\theta$  and  $\eta$  are said to be ‘different’ if they are not ‘equal’ in this meaning. This is consistent with the partial ordering in Definition 1. The word ‘different’ is used in the same meaning in the Theorem below.

Postulate 4 can be motivated by using Zorn’s lemma – if this lemma, which is equivalent to the axiom of choice, is assumed to hold – and Postulate 3, but such a motivation is not necessary if Postulate 4 is accepted. Physical examples of maximal accessible variables are the position or the momentum of some particle, or the spin component in some direction. In a more general situation, the maximal accessible variable may be a vector, whose components are simultaneously measurable variables.

Assuming these postulates, the main result of Helland<sup>[6][12][15]</sup> is as follows:

**Theorem 1.** *Consider a context where there are two different maximal accessible variables  $\theta$  and  $\eta$ . Assume that both  $\theta$  and  $\eta$  are real-valued or real vectors, taking at least two values. Make the following additional assumptions:*

- i. On one of these two variables, say  $\theta$ , there can be defined a transitive group of actions  $G$  with a trivial isotropy group and with a left-invariant measure  $\rho$  on the space  $\Omega_\theta$ , the range space of the function  $f$  in  $\theta = f(\phi)$ .
- ii. There exists a group  $M$  on  $\Omega_\phi$ .

Then there exists a Hilbert space  $\mathcal{H}$  connected to the situation, and to every (real-valued or vector-valued) accessible variable there can be associated a symmetric operator on  $\mathcal{H}$ .

The main result is that each accessible variable  $\xi$  is associated with an operator  $A^\xi$ . The proof goes by first constructing  $A^\theta$  and  $A^\eta$ , then operators associated with other accessible variables are found by using the spectral theorem. For this, we need weak conditions<sup>[27]</sup> ensuring that the symmetric operators are self-adjoint.

In order to formulate in general the spectral theorem, first, the *spectrum*  $\sigma^A$  of an operator  $A$  is defined as the set of constants  $\lambda$  such that  $A - \lambda I$  does not have a bounded inverse. For self-adjoint operators, the spectrum is contained in the real line and contains all eigenvalues of  $A$ . For bounded operators, the spectrum is equal to the set of eigenvalues.

Then, in general, we have for any self-adjoint operator  $A$  that there exists a projection-valued measure  $E^A$  such that

$$A = \int_{\sigma^A} \lambda dE^A(\lambda). \tag{1}$$

From this spectral theorem (see Hall<sup>[27]</sup> for a proof), if  $A = A^\eta$  is the operator which according to Theorem 1 is associated with the maximal accessible variable  $\eta$ , we can define the operator associated with  $\xi = f(\eta)$  by

$$A^\xi = f(A) = \int_{\sigma^A} f(\lambda) dE^A(\lambda). \tag{2}$$

In particular, we have

$$\int_{\sigma^A} dE^A(\lambda) = I. \tag{3}$$

Note that here  $\eta$  may be any maximal accessible variable associated with some  $\theta$  which satisfies (i) and (ii) above. It is shown in Helland (2024d) that, as a consequence of the assumption (ii), the two variables  $\theta$  and  $\eta$  will be related.



By Axiom 4, for any accessible variable  $\xi$ , there exists a maximal variable  $\eta$  and a function  $f$  such that  $\xi = f(\eta)$ . In this way, operators associated with any accessible variable may be defined.

Groups as acting on a space are important in my approach. A group action  $G$  acting on a space  $\Omega$  is called transitive if for every  $\omega \in \Omega$ , the range of  $g\omega$  as  $g$  runs through  $G$  is the full space  $\Omega$ . If this holds for one  $\omega$ , it holds for every  $\omega \in \Omega$ . The isotropy group of  $G$  at  $\omega$  is the set of  $g \in G$  such that  $g\omega = \omega$ . In the transitive case, for different  $\omega$ , the isotropy groups are isomorphic. In particular, if the isotropy group is trivial for one  $\omega$ , it is trivial for every  $\omega \in \Omega$ .

When there is a transitive group  $G$  with a trivial isotropy group acting on  $\Omega$ , there will be a one-to-one correspondence between the points  $\omega \in \Omega$  and the group elements  $g \in G$ . This is important for the proof of Theorem 1.

It is shown in Helland<sup>[15]</sup> that under the assumptions of Theorem 1, there exists a variable  $\xi$  that is a bijective function of  $\eta$  and a transformation  $k$  on the  $\phi$ -space by which the variables  $\theta$  and  $\xi$  are related:  $\xi(\phi) = \theta(k\phi)$ .

An important special case of Theorem 1 is when the accessible variables take a finite number of values, say  $u_1, u_2, \dots, u_r$ . For this case, it is proved in Helland<sup>[12][15]</sup> that a group  $G$  and a transformation  $k$  with the above properties can always be constructed. The following Corollary then follows:

**Corollary 1.** *Assume that there exist two different maximal accessible variables  $\theta$  and  $\eta$ , each taking  $r$  values, and not in one-to-one correspondence. Then, there exists an  $r$ -dimensional Hilbert space  $\mathcal{H}$  describing the situation, and every accessible variable in this situation will have an associated self-adjoint operator in  $\mathcal{H}$ .*

In the finite case, the equations (1-3) take a simpler form. The operator  $A^\eta$  will have eigenvalues  $\{u_j\}$  and corresponding eigenvectors  $\{\mathbf{u}_j\}$ . The spectral theorem then reads

$$A^\eta = \sum_j u_j \mathbf{u}_j \mathbf{u}_j^\dagger, \quad (4)$$

where  $\mathbf{u}_j^\dagger$  is the complex conjugate row vector corresponding to  $\mathbf{u}_j$ . In the quantum mechanical literature, these vectors are often written as ket and bra vectors:  $\mathbf{u}_j = |j\rangle$  and  $\mathbf{u}_j^\dagger = \langle j|$ . The eigenvectors can be chosen as orthonormal:  $\mathbf{u}_i^\dagger \mathbf{u}_j = \langle i|j\rangle = \delta_{ij}$ . In the following, both notations will be used.

We then further have

$$A^{f(\eta)} = \sum_j f(u_j) \mathbf{u}_j \mathbf{u}_j^\dagger, \quad (5)$$

in particular

$$\sum_j \mathbf{u}_j \mathbf{u}_j^\dagger = I, \quad (6)$$

the corresponding resolution of the identity.

Theorem 1 and its Corollary constitute the first steps in a new proposed foundation of quantum theory. In statistical applications, the variables (parameters) involved will in most cases have a continuous variation. However, continuous parameters can be approximated by a sequence of parameters taking a finite number of values; see subsection 5.3 in Helland<sup>[4]</sup>. In this way, we may avoid both the symmetry assumptions of Theorem 1 and the technical issues relating symmetric and self-adjoint operators. Examples are given in Section 3 below.

The second step now is to prove the following: If  $k$  is the transformation connecting two related maximal accessible variables  $\theta$  and  $\eta$ , and  $A^\theta$  and  $A^\eta$  are the associated operators, then there is a unitary operator  $W(k)$  such that  $A^\eta = W(k)^{-1} A^\theta W(k)$ . This, and a more general related result, is proved as Theorem 5 in Helland<sup>[12]</sup>.

Given these results, a rich theory follows. The set of eigenvalues of the operator  $A^\theta$  is proved to be identical to the set of possible values of  $\theta$ . The variable  $\theta$  is shown to be maximal if and only if all eigenvalues of the corresponding operator are simple. In general, the eigenspaces of  $A^\theta$  are in one-to-one correspondence with questions ‘What is  $\theta$ ’/ ‘What will  $\theta$  be if we measure it?’ together with sharp answers  $\theta = u$  for some eigenvalue  $u$  of  $A^\theta$ . If  $\theta$  is a maximal accessible variable, then all eigenvectors of the operator  $A^\theta$  have a similar interpretation. In my theory, I wish to limit the concept of a state vector to vectors in the Hilbert space that are eigenvectors of some meaningful operator. Then this gives a simple interpretation of all possible state vectors.

What is lacking in the above theory is a foundation of Born’s formula, necessary for the computation of quantum probabilities. Several versions of Born’s formula are proved from two new postulates in Helland<sup>[4][15]</sup>. The first postulate is as follows:

**Postulate 5.** *The likelihood principle holds.*

As is well known, the likelihood principle is a principle that many statisticians base their inference on. In its strict form, it is controversial; see, for instance, the discussion in Schweder and Hjort<sup>[33]</sup>. Elsewhere<sup>[4]</sup>, I have advocated the view that the principle should be restricted to a specific context,

and then it is less controversial. For a basic historical discussion of the principle, see Berger and Wolpert<sup>[34]</sup>.

Recall that the likelihood principle runs as follows: *Relative to any experiment, the experimental evidence is always a function of the likelihood.* Here, the term ‘experimental evidence’ is left undefined and can be made precise in several directions. But as everybody would agree, an experiment is always done in a context, and such a context should include a well-defined experimental question.

In a quantum mechanical setting, a potential or actual experiment is seen in relation to an actor  $C$  or to a communicating group of actors. Concentrate here on the first scenario. In the simplest case, assuming a discrete-valued variable, we assume that  $C$  knows the state  $|a; i\rangle$  of a physical system and that this state can be interpreted as the knowledge that  $\theta^a = u_i$  for some maximal accessible variable  $\theta^a$ . Then assume that  $C$  has focused upon a new maximal accessible variable  $\theta^b$ , and we are interested in the probability distribution of this variable.

The last postulate is connected to the scientific ideals of  $C$ , ideals that either are given by certain conscious or unconscious principles, or are connected to some concrete persons. These ideals are then modeled by some ‘higher being’  $D$  that  $C$  considers to be perfectly rational with respect to any aspect of the relevant theoretical variables.

**Postulate 6.** *Consider in the context  $\tau$  an experiment where the likelihood principle is assumed to be satisfied, and the whole situation is observed by an experimentalist  $C$  whose decisions can be shaped or influenced by a ‘superior being’  $D$ . Assume  $D$ ’s probability for some given outcome  $E$  is  $q$ , that  $D$  is seen by  $C$  to be perfectly rational in agreement with the Dutch Book Principle, and that  $q$  is assumed to be the real probability for  $E$ .*

The Dutch Book Principle says as follows: No choice of payoffs in a series of bets shall lead to a sure loss for the bettor.

A situation where Postulate 6 holds will be called a rational epistemic setting. It will be seen in the next subsection to imply essential aspects of quantum probability. As shown in Helland<sup>[13]</sup>, it also gives a foundation for probabilities in quantum decision theory.

In Helland<sup>[4][14]</sup>, a generalized likelihood principle is proved from the ordinary likelihood principle: Given some experiment, assumed here to have a discrete, maximal accessible parameter  $\theta^a$ , and assume a context  $\tau$  connected to the experiment, any experimental evidence will under the above assumptions be a function of the so-called likelihood effect  $F = F^a$ , defined by

$$F^a(\mathbf{u}; z, \tau) = \sum_i p(z|\tau, \theta^a = u_i) |a; i\rangle \langle a; i| \quad (7)$$

In particular, the probability  $q$  of Postulate 6 must be a function of  $F$ :  $q(F|\tau)$ .

In many textbooks, quantum mechanics is restricted to discrete-valued variables as above. For a continuous variable  $\theta^a$ , the likelihood effect can be defined by appealing to the spectral theorem for the operator  $A = A^{\theta^a}$  and using the probability density  $p(z|\tau, \theta^a = u)$  for the data:

$$F^a(\mathbf{u}; z, \tau) = \int_{\sigma^A} p(z|\tau, \theta^a = u) dE^A(u). \quad (8)$$

### 2.3. The Born rule from the approach above

Using these postulates and a version of Gleason's Theorem due to Busch<sup>[35]</sup>, the following variant of Born's formula is proved in Helland<sup>[4],[14]</sup>:

**Theorem 2** [Born's formula, simple version] *Assume a rational epistemic setting and assume two different discrete maximal accessible variables  $\theta^a$  and  $\theta^b$ . In the above situation, we have:*

$$P(\theta^b = v_j | \theta^a = u_i) = |\langle a; i | b; j \rangle|^2. \quad (9)$$

Here,  $|a; i\rangle$  is the state given by  $\theta^a = u_i$  and  $|b; j\rangle$  is the state given by  $\theta^b = v_j$ . In this version of the Born formula, I have assumed *perfect measurements*: there is no experimental noise, so that the experiment gives a direct value of the relevant theoretical variable. Another assumption is that the events  $\theta^a = u_i$  and  $\theta^b = v_j$  are contained in the experimental questions related to the respective experiments.

A last postulate is needed to compute probabilities of independent events. A version of such a postulate is

**Postulate 7.** *If the probability of an event  $E_1$  is computed by a probability amplitude  $z_1$  from the Born rule in the Hilbert space  $\mathcal{H}_1$ , the probability of an event  $E_2$  is computed by a probability amplitude  $z_2$  from the Born rule in the Hilbert space  $\mathcal{H}_2$ , and these two events are independent, then the probability of the event  $E_1 \cap E_2$  can be computed from the probability amplitude  $z_1 z_2$ , associated with the tensor product of the Hilbert spaces  $\mathcal{H}_1$  and  $\mathcal{H}_2$ .*

This postulate can be motivated by its relation to classical probability theory: If  $P(E_1) = |z_1|^2$  and  $P(E_2) = |z_2|^2$ , then  $P(E_1 \cap E_2) = P(E_1)P(E_2) = |z_1|^2 |z_2|^2 = |z_1 z_2|^2$

The simple Born formula can now be generalized to the case where the variables are continuous, and where the accessible variables  $\theta^a$  and  $\theta^b$  are not necessarily maximal. There is also a variant for a

mixed state involving  $\theta^a$ .

Define first the mixed state operator associated with any accessible variable  $\theta$  by using the spectral theorem for  $A = A^\theta$ :

$$\rho^\theta = \int_{\sigma^A} p^\theta(u) dE^A(u). \quad (10)$$

In the continuous case,  $p^\theta(u)$  is the probability density for  $\theta$ . In the discrete case,  $p^\theta(u)$  is the point probability of  $\theta$ , and (10) reads

$$\rho^\theta = \sum_j p^\theta(u_j) \mathbf{u}_j \mathbf{u}_j^\dagger, \quad (11)$$

where  $\mathbf{u}_j$  is the eigenvector of  $A^\theta$  corresponding to the eigenvalue  $u_j$ .

The probability distribution for  $\theta$  assumed in (10) can be of many kinds. For a Bayesian, it can be a prior or posterior distribution. For a frequentist, it can be a confidence distribution of the kind discussed by Schweder and Hjort<sup>[33]</sup>. Also, statisticians that follow some fiducial school operate with probability distributions of parameters.

Assume now first that  $\theta^b$  in Theorem 2 is discrete, but not necessarily maximal. Then  $\theta^b$  is a function  $f$  of a maximal accessible variable  $\eta$ , and it follows by summation over  $j$ , assuming that  $\eta = v_j$  belongs to some set  $B_1$  defined by  $f(v_j) \in B$  for some given set  $B$ , that

$$P(\theta^b \in B | \theta^a = i) = \langle a; i | \Pi_B | a : i \rangle, \quad (12)$$

where  $\Pi_B$  is the projection upon the space spanned by the eigenvectors  $|b; j\rangle$  of  $A^\eta$  for which the eigenvalues  $v_j$  are in the set of indices  $j$  such that  $f(v_j) \in B$ .

Now, by approximating a continuous  $\theta^b$  by discrete variables  $\theta^{bn}$  such that  $\theta^{bn} \rightarrow \theta^b$  as  $n \rightarrow \infty$ , it is easy to show that (12) holds in general, where now  $\Pi_B$  is interpreted as the projection upon the eigenspace of the indicator for the set  $\theta^b \in B$  corresponding to the value 1 for this indicator. More precisely, we should use  $\Pi_B = \int_{B \cap \sigma^b} dE^b(u)$ , where  $\sigma^b$  is the spectrum of the operator  $A^{\theta^b}$ , and  $\{E^b\}$  is found from the spectral theorem of the same operator.

In the same way, a continuous  $\theta^a$  may be approximated by discrete  $\theta^{ar}$ , assumed to be functions  $f_r$  of some maximal accessible variables  $\xi_r$ , replacing the variable  $\theta^a$  in Theorem 2. Then, using the definition (10) (which can be generalized to the case of non-maximal accessible variables), we can prove

**Theorem 3** [Born's formula, general version]. Assume Postulate 5 and Postulate 6, and assume that we have two different accessible variables  $\theta^a$  and  $\theta^b$  in the relevant context. Assume that the knowledge of  $\theta^a$  is given by the density matrix  $\rho^a$ . Then for any Borel set  $B$  in  $\Omega_{\theta^b}$  we have

$$P(\theta^b \in B | \rho^a) = \text{trace}(\rho^a \Pi_B). \quad (13)$$

This result is not necessarily associated with a microscopic situation, a fact that I will come back to in examples in Section 3.

As a corollary, we have

$$E(\theta^b | \rho^a) = \text{trace}(\rho^a A^{\theta^b}). \quad (14)$$

Here  $A^{\theta^b}$  is the operator corresponding to  $\theta^b$ .

Finally, one can generalize to the case where the final measurement is not necessarily perfect. Let us assume future data  $z^b$  instead of a perfect theoretical variable  $\theta^b$ , which is now taken to be the parameter of the experiment. Note that we only need the likelihood principle (together with postulate 6) for perfect experiments in order to prove that (13) is valid. Then we can define an operator corresponding to  $z^b$  by

$$A^{z^b} = \int_{\sigma^A} z^b p(z^b | \theta^b = u) dE^b(u), \quad (15)$$

where  $\{E^b\}$  is found from the spectral theorem used on the operator  $A = A^{\theta^b}$ . Then from the version (13) of the Born formula, we obtain

$$E(z^b | \rho^a) = \text{trace}(\rho^a A^{z^b}). \quad (16)$$

and

$$P(z^b \in F) = \text{trace}(\rho^a \int_{z \in (F \cap \sigma^{A^1})} dE^{z^b}(z)), \quad (17)$$

where  $\{E^{z^b}\}$  is found from the spectral theorem used on the operator  $A^1 = A^{z^b}$ .

Again, elementary quantum mechanics uses discrete data and discrete parameters, a setting unfamiliar to statisticians, but useful as an approximation. Then  $p(z^b | \theta^b = u_i)$  is the point probability of the data, we define

$$A^{z^b} = \sum_i p(z^b | \theta^b = u_i) \mathbf{u}_i \mathbf{u}_i^\dagger, \quad (18)$$

and Born's formula gives  $E(z^b | \rho^a) = \text{trace}(\rho^a A^{z^b})$  and

$$P(z^b \in F) = \text{trace}(\rho^a \sum_{z_j \in F} \mathbf{v}_j \mathbf{v}_j^\dagger), \quad (19)$$

where  $\mathbf{v}_j$  is the eigenvector of  $A^{z^b}$  corresponding to the eigenvalue  $z_j$ .

All this not only points to a new foundation of quantum theory, but it also suggests a general epistemic interpretation of the theory: Quantum Theory is not directly a theory about the world, but a theory about an actor's knowledge of the world. In particular, the probabilities in the Born formula can be interpreted as probabilities attached to a single observer, or to a communicating group of observers. It is crucial that the probabilities at the outset according to Postulate 6 are seen as probabilities as evaluated by the 'superior actor' D.

#### 2.4. Quantum theory as the most robust description of reproducible experiments

In De Raedt et al.<sup>[16][36]</sup> another approach to the foundation of quantum theory is discussed. This approach will be described very briefly here.

First, the authors define inference-probability as any conditional probability satisfying the three rules

1.  $P(A|Z) + P(\bar{A}|Z) = 1$ , where  $\bar{A}$  denotes not- $A$ .
2.  $P(AB|Z) = P(A|BZ)P(B|Z) = P(B|AZ)P(A|Z)$ , where  $AB$  denotes essentially what statisticians call  $A \cap B$ .
3.  $P(A\bar{A}|Z) = 0$  and  $P(A + \bar{A}|Z) = 1$ , where  $A + B$  denotes essentially what statisticians call  $A \cup B$ .

These rules are the same as the rules for the concept of plausibility, derived from reasonable assumptions and discussed in detail by Jaynes<sup>[37]</sup>. In op. cit.,  $A$ ,  $B$  and  $Z$  are propositions, and to be precise,  $AB$  denotes that both propositions  $A$  and  $B$  are true, and  $A + B$  denotes that at least one of the propositions is true.

Next, De Raedt et al.<sup>[16][36]</sup> assume the following conditions, which are made precise in these articles:

**Conditions 1.** *There may be uncertainty about each event. The condition under which the experiment is carried out may be uncertain. The frequencies with which events are observed are reproducible and robust against small changes in the conditions. Individual events are independent.*

Using these assumptions, they first derive quantum probabilities for the Einstein-Podolsky-Rosen-Bohm thought experiment. This experiment consists of a source  $S$ , a router  $R_1$  to the left of the source

oriented by a chosen unit vector  $\mathbf{a}_1$ , a router  $R_2$  to the right of the source oriented by a chosen unit vector  $\mathbf{a}_2$ , and two detectors connected to each router.

The experiment is the same as what is called the Bell experiment, an experiment that has now been shown experimentally to give outcomes as predicted by quantum theory, but not by 'common sense' use of classical arguments. The Bell experiment has been discussed under various assumptions by many authors, including Helland<sup>[7][11]</sup>

De Raedt aims at making a minimal set of assumptions about the experiment; for details, see De Raedt et al.<sup>[16][36]</sup> and references therein:

- a. Each time the source  $S$  is activated, it sends a signal to the right router and a signal to the left router.
- b. After this, each router sends a signal to one of its two detectors, depending upon the orientations  $a_i$ .
- c. The detectors register the signal and operate with 100 % efficiency.

The whole experiment is then run a large number  $N$  of times, giving frequencies for the various outcomes. Probabilities are obtained by letting  $N \rightarrow \infty$ . It turns out that these probabilities are in agreement with quantum theory, which deviates from the common sense use of classical arguments.

This is a special experiment, but it is an important experiment, distinguishing between quantum predictions and classical conditions. A simpler experiment is the Stern-Gerlach experiment. Here the source  $S$ , activated at times  $n = 1, 2, \dots, N$ , sends a particle carrying a magnetic moment  $\mathbf{s}$  to a magnet  $M$  with its magnetization  $\mathbf{a}$ . Depending on  $\mathbf{a}$  and  $\mathbf{s}$ , the particle is detected with 100 % by one of two assumed detectors. It is crucial that this depends only on the scalar product  $\mathbf{a} \cdot \mathbf{s}$ . Again, by a long argument, the predictions of quantum theory are derived under very weak assumptions.

In addition to this, there is a discussion of the Schrödinger equation in De Raedt et al.<sup>[16]</sup>, a theme that I will not discuss in detail in the present article.

It is crucial that fundamental quantum predictions are derived in op. cit., using plausible assumptions together with the basic Conditions 1. There is nothing microscopic connected to these assumptions, supporting my main thesis that quantum theory also may be relevant in a macroscopic context.

Finally, these derivations are also consistent with my epistemic interpretation of quantum theory, as shown by the following citation: '... current scientific knowledge derives, through cognitive processes



in the human brain, from the discrete events which are observed in laboratory experiments and from the relationships between those experiments that we, humans, discover.’

As said in the Introduction above, it is also highly relevant that the concept of Fisher information is used in their detailed arguments.

In a recent, long, and detailed article<sup>[38]</sup>, the Bell experiment and the violation of the so-called CHSH inequality is discussed from the general point of view of mathematical models and discrete data. It is concluded that discrete data recorded by experiments and mathematical models used to describe relevant features belong to different, separate universes and should be treated accordingly. This is a conclusion that seemingly has large consequences for ordinary statistical modelling.

In my opinion, this conclusion indeed seems to be supported in experiments like the Bell experiment. It is connected to the fact that any human being, including a theoretician that makes mathematical models, meets limitations that apply to these models: It may in certain cases be impossible to include more than two maximal accessible parameters in the models. More precisely: If the model contains two, really different, related maximal accessible parameters  $\theta$  and  $\eta$ , it cannot at the same time contain a parameter  $\zeta$  which is related to  $\theta$ , but not to  $\eta$ . The notion of being related can be given a precise definition connected to the mind of a person or to the joint minds of a group of communicating persons (see Subsection 2.2). All this follows from the discussion in Helland<sup>[7][11]</sup>, a discussion which is also briefly given in Helland (2024c).

## 2.5. Ubiquitous quantum structure

In a number of books and articles, see Khrennikov<sup>[18][19]</sup> and Haven & Khrennikov<sup>[20][21]</sup> for some of them, Andrei Khrennikov has advocated what he calls quantum-like models. I will start with a particular argument given in Khrennikov<sup>[18]</sup>.

He there takes the point of departure that the law of total probability

$$P(B|C) = \sum_a P(A_a|C)P(B|A_a \cap C), \quad (20)$$

where  $\{A_a\}$  is a disjoint partition of the probability space, does not hold under all circumstances. He assumes in particular that a term of the form

$$2\lambda(B|a, C)\sqrt{\prod_a (P(A_a|C)P(B|A_a \cap C))} \quad (21)$$

where  $\prod_a$  denotes the product, may have to be added to the right-hand side of (20).

He then shows that quantum probabilities may be derived under the specific condition that  $|\lambda(B|a, C)| \leq 1$  for all  $B, a, C$ . One then can write  $\lambda = \cos(\phi)$  for an angle  $\phi$ . Specialize to the case where the partition has two elements  $A$  and  $\bar{A}$ , and use the elementary formula  $x + y + 2\sqrt{xy}\cos(\phi) = |\sqrt{x} + e^{i\phi}\sqrt{y}|^2$  to arrive at

$$P(B|C) = |\psi(B)|^2, \quad (22)$$

where

$$\psi(B) = \sqrt{P(A|C)P(B|A \cap C)} + e^{i\phi}\sqrt{P(\bar{A}|C)P(B|\bar{A} \cap C)}. \quad (23)$$

which can be seen as a special case of Born's formula.

This is a minor technical argument, but it again shows that quantum probabilities may be derived under conditions that are not necessarily microscopic. In fact, the many examples and the various other arguments in Khrennikov<sup>[18][19]</sup> underline this point. Let me end this subsection with a citation from the preface of Khrennikov<sup>[19]</sup>.

'Quantum-like modeling is built on the methodology and the mathematical apparatus of quantum theory and it is directed to applications outside of physics, namely to biology, cognition, psychology, decision making, economics, finances, social and political sciences, and artificial intelligence.'

Note that statistics as a science aims at a similar list of applications.

## 2.6. Some of the properties of quantum probabilities

I have already mentioned that the law of total probability does not hold for quantum probabilities. This implies also that Savage's sure thing principle<sup>[39]</sup> does not hold:

If an event  $B$  is true (has probability 1) under condition  $A$ , and it also is true under condition  $\bar{A}$ , then it is always true.

This may seem counterintuitive, but it can be understood by the fact that quantum mechanics allows states that can be seen as superpositions of the states specified by the condition  $A$  and by the condition  $\bar{A}$ , states where other conditions are focused upon. (I do not need the general superposition principle here; the 'other conditions' can be related to a complementary theoretical variable.)

Another property of quantum probabilities is that the probability of successive events depends on the order of the events. An example mentioned in the quantum decision literature is an opinion poll, where American citizens were asked about their opinions on Al Gore and their opinion on Bill Clinton.

Empirically, it was shown that answers were dependent upon the order in which the questions were asked.

To prove this property, we also have to assume the well-known collapse rule of quantum mechanics: Start with a pure state  $|\psi\rangle$ , so that, according to the Born rule, the probability of an event  $B$  is  $\|\Pi^B|\psi\rangle\|^2$ , where  $\Pi^B$  is the projection on the indicator of  $B$ . Then the collapse rule says that after the measurement, the state changes to

$$|\psi^B\rangle = \frac{\Pi^B|\psi\rangle}{\|\Pi^B|\psi\rangle\|}. \quad (24)$$

For a derivation of this rule from a knowledge-based perspective, see Shrapnel et al.<sup>[40]</sup>

This gives for successive events  $B$  and  $C$ :

$$P(B \text{ and then } C) = P(B|\psi)P(C|B) = \|\Pi^B|\psi\rangle\|^2 \|\Pi^C \frac{\Pi^B|\psi\rangle}{\|\Pi^B|\psi\rangle}\|^2 = \|\Pi^C \Pi^B|\psi\rangle\|^2, \quad (25)$$

which in general is different from  $P(C \text{ and then } B)$ .

Finally, it follows from (25) and a simple geometric argument that one in certain cases may have

$$P(B \text{ and then } C) > P(C),$$

which to some may seem counterintuitive, but can be illustrated by the so-called Linda paradox, discussed by several authors, for instance Busemeyer and Bruza<sup>[41]</sup>.

### 3. Statistical applications

In the statistical applications below, I will only in some special cases go into details concerning the related quantum-mechanical calculations, which may be complicated. The main purpose of this Section is to point at some ideas under which such calculations may possibly enlighten or complement a statistical analysis.

#### 3.1. An experiment that can be analyzed by quantum probabilities

In a medical experiment, let  $\mu_a, \mu_b, \mu_c$  and  $\mu_d$  be continuous inaccessible parameters, the hypothetical effects of treatment  $a, b, c$  and  $d$ , respectively. Assume that the focus of the experiment is to compare treatment  $b$  with the mean effect of the other treatments, which is supposed to give the parameter

$\frac{1}{3}(\mu_a + \mu_c + \mu_d)$ . One wants to do a pairwise experiment, but it turns out that the maximal parameter which can be estimated is

$$\theta^b = \text{sign}(\mu^b - \frac{1}{3}(\mu_a + \mu_c + \mu_d)).$$

(Imagine, for example, that one has four different ointments against rash. A patient is treated with ointment  $b$  on one side of his back; a mixture of the other ointments on the other side of his back. It is only possible to observe which side improves best, but this observation is assumed to be very accurate. One can in principle do the experiment on several patients and select out the patients where the difference is clear.)

Described in this way, it may be natural, after the data are collected, to do a Bayesian analysis with a prior given by  $P(\theta^b = -1) = P(\theta^b = +1) = 1/2$ . But assume now that we have the following modification of the experiment:

The experiment is done on a selected set of experimental units, on whom it is known from earlier accurate experiments that the corresponding parameter

$$\theta^a = \text{sign}(\mu^a - \frac{1}{3}(\mu_b + \mu_c + \mu_d))$$

takes the value  $+1$ . In other words, for a Bayesian analysis, one is interested in the priors

$$\pi = P(\theta^b = +1 | \theta^a = +1).$$

$$1 - \pi = P(\theta^b = -1 | \theta^a = +1).$$

Consider first a full Bayesian approach, also toward these priors. Natural priors for  $\mu_a, \dots, \mu_d$  are independent  $N(\nu, \sigma^2)$  with the same  $\nu$  and  $\sigma$ . By location and scale invariance, there is no loss in generality by assuming  $\nu = 0$  and  $\sigma = 1$ . Then the joint prior of  $\zeta^a = \mu_a - \frac{1}{3}(\mu_b + \mu_c + \mu_d)$  and  $\zeta^b = \mu_b - \frac{1}{3}(\mu_a + \mu_c + \mu_d)$  is multinormal with mean  $\mathbf{0}$  and covariance matrix

$$\begin{pmatrix} \frac{4}{3} & -\frac{4}{9} \\ -\frac{4}{9} & \frac{4}{3} \end{pmatrix}.$$

A numerical calculation from this gives

$$\pi = P(\zeta^b > 0 | \zeta^a > 0) \approx 0.43.$$

This result can also be assumed to be valid when  $\sigma \rightarrow \infty$ , a case which can be considered as independent objective priors for  $\mu_a, \dots, \mu_d$ , more precisely, a joint non-informative prior for the

parameters under the translation group; see Helland<sup>[42]</sup>.

Now consider quantum probabilities for the same priors. Since again scale is irrelevant, a natural group on  $\mu_a, \dots, \mu_d$  is a 4-dimensional rotation group around a point  $(\nu, \dots, \nu)$  together with a translation of  $\nu$ . Furthermore,  $\zeta^a$  and  $\zeta^b$  are contrasts, that is, linear combinations with coefficients adding to 0. The space of such contrasts is a 3-dimensional subspace of the original 4-dimensional space, and by a single orthogonal transformation, the relevant subset of the 4-dimensional rotations can be transformed into the group  $G$  of 3-dimensional rotations on this latter space, and the translation in  $\nu$  is irrelevant. One such orthogonal transformation is given by

$$\begin{aligned}\psi_0 &= \frac{1}{2}(\mu_a + \mu_b + \mu_c + \mu_d), \\ \psi_1 &= \frac{1}{2}(-\mu_a - \mu_b + \mu_c + \mu_d), \\ \psi_2 &= \frac{1}{2}(-\mu_a + \mu_b - \mu_c + \mu_d), \\ \psi_3 &= \frac{1}{2}(-\mu_a + \mu_b + \mu_c - \mu_d).\end{aligned}$$

Let  $G$  be the group of rotations orthogonal to  $\psi_0$ . We find

$$\begin{aligned}\zeta^a &= -\frac{2}{3}(\psi_1 + \psi_2 + \psi_3), \\ \zeta^b &= -\frac{2}{3}(\psi_1 - \psi_2 - \psi_3).\end{aligned}$$

The rotation group element transforming  $\zeta^a$  into  $\zeta^b$  under  $G$  is strongly related to the group element  $g_{ab}$  transforming  $a = -\frac{1}{\sqrt{3}}(1, 1, 1)$  into  $b = -\frac{1}{\sqrt{3}}(1, -1, -1)$  under a group of rotations of unit vectors.

Furthermore, let  $G^a$  be the maximal subgroup of  $G$  under which  $\zeta^a$  is permissible. The following definition was given in Helland<sup>[42][43]</sup> and is further discussed in these two articles:

**Definition 2.** Let  $G$  be a group acting upon a parameter  $\eta$ , and let  $\zeta$  be a function of  $\eta$ . We say that  $\zeta$  is permissible if  $\zeta(\eta_1) = \zeta(\eta_2)$  implies  $\zeta(g\eta_1) = \zeta(g\eta_2)$  for all  $g \in G$ .

In general, if  $\zeta$  is a permissible parameter in this way, one can define group actions  $h \in H$  on  $\zeta$  by  $h(\zeta(\eta)) = \zeta(g\eta)$  for  $g \in G$ .

The subgroup  $G^a$  is here isomorphic with the unit vector transformation group of rotations around  $a$  together with a reflection in the plane perpendicular to  $a$ . The action by the group  $H^a$  induced on  $\zeta^a$  by  $G^a$  is just a reflection together with the unit element.

Again, all these groups have their analogues in relation to the rotation group of unit vectors.

In conclusion, the whole situation is completely equivalent to the spin-example discussed in many books, for instance, Helland<sup>[4]</sup>, and may be assumed to satisfy the postulates of Subsection 2.2 above. This implies, by an application of the Born rule (see Proposition 7 in Helland<sup>[4]</sup>):

$$\pi = P(\text{sign}(\zeta^b) = +1 | \text{sign}(\zeta^a) = +1) = \frac{1}{2}(1 + a \cdot b) = \frac{1}{3}.$$

So the two analyses give different results for the desired prior. Which solution should one recommend? Here is my opinion: Both solutions are based upon symmetries implied by group actions. The full Bayesian solution is based upon a prior distribution on the inaccessible parameters  $\mu_a, \mu_b, \mu_c$  and  $\mu_d$ , which could be related to group actions upon these parameters. In the quantum solution, one ends up with a group acting upon the accessible parameters  $\theta^a = \text{sign}(\zeta^a)$  and  $\theta^b = \text{sign}(\zeta^b)$ . In general, in applied statistics, it is crucial that the parameter space is not too large. From an applied point of view, symmetries based upon accessible parameters should be preferred when compared to rather abstract symmetries based upon inaccessible parameters. Therefore, even though the arguments are more complicated, I will here prefer the quantum probability solution. Related arguments are discussed more generally in the next subsections.

### 3.2. Model reduction under symmetry

In applied statistics, it is crucial that the parameter space is not too large. For instance, in regression problems, when the number of regression variables  $p$  is larger than the number of units  $n$ , ordinary least squares regression runs into problems. Let, in general,  $\phi$  be the set of all thinkable parameters that a statistician  $C$  wants to include in his model, and, for the sake of the argument, let us assume that there exists a group  $M$  acting on the space  $\Omega_\phi$ .

The total parameter  $\phi$  may, in many cases, be so large that it cannot be estimated from the available data, using some estimation principle like unbiased estimation or equivariant estimation. Here, I will concentrate on the last estimation principle.

In general, let a group  $G$  act upon the space  $\Omega_\theta$  over which a parameter  $\theta$  varies. In many cases, this group may be induced by a group  $\hat{G}$  acting upon the sample space  $\Omega$ , based upon some statistical model  $P^\theta(x)$ :

$$P^{g\theta}(\hat{g}x) = P^\theta(x) \text{ for all } x \in \Omega. \quad (26)$$

This introduces a *homomorphism* from  $\hat{G}$  to  $G$ : If  $\hat{g}_1$  is mapped to  $g_1$  and  $\hat{g}_2$  is mapped to  $g_2$ , then  $\hat{g}_1 \hat{g}_2$  is mapped to  $g_1 g_2$ .

An estimator  $\hat{\theta}(x)$  of the parameter  $\theta$  is said to be *equivariant* if  $\hat{\theta}(\hat{g}x)$  is the estimator of  $g\theta$  whenever  $\hat{g}$  is mapped to  $g$ . There can be many arguments given to concentrate upon equivariant estimators.

Under very weak conditions, there exists a right-invariant measure  $\mu$  on  $\Omega_\theta$  under the group  $G$ : First, a right-invariant Haar measure  $\nu$  is defined on the group  $G$  itself by  $\nu(D \cdot g) = \nu(D)$  for all  $D \subseteq G$  and  $g \in G$ . Then  $\mu$  is said to be right-invariant if  $\mu(A) = \nu(G_A)$  with  $G_A = \{g : g\theta_0 \in A\}$  for some  $\theta_0$ . Left-invariant measures have a similar definition. In many cases, the left-invariant measure is equal to the right-invariant measure,

This introduces an invariant measure on every *orbit* of the group  $G$ : An orbit is the set  $\{g\theta_0\}$  for some  $\theta_0 \in \Omega_\theta$ . The space  $\Omega_\theta$  is always divided into a disjoint set of orbits. If  $\theta_1$  and  $\theta_2$  belong to the same orbit, this orbit can equivalently be characterized by either  $\{g\theta_1\}$  or  $\{g\theta_2\}$ . If there is only one orbit of  $G$  in  $\Omega_\theta$ , the group is said to be *transitive*.

An objective Bayes estimator with respect to  $G$  is an estimator that uses the right-invariant measure as a prior. In Helland<sup>[42]</sup>, 12 different reasons for using such a prior are given; among other things, it can be proved that credibility sets with some credibility probability are equal to frequentist confidence sets with the same confidence probability.

Turning to quantum theory, it is important for the foundation that there exists a transitive group on the variable space (see point (i) in Theorem 1 of Subsection 2.2). If  $G$  should not be transitive, we can introduce the following model reduction principle:

**Principle 1.** *Reduce  $\Omega_\theta$  to an orbit of the group  $G$ . Choose the orbit such that a subparameter  $\zeta$  of interest is permissible.*

In Helland<sup>[44]</sup> and Helland<sup>[4]</sup>, this principle is used on the electron spin, a qubit. It is shown that a classical model of spin can be reduced to a quantum model using this principle. In Helland et al.<sup>[45]</sup>, the same principle was used to motivate the model reduction in multiple regression, leading to the partial least squares regression model.

Go back to the example in Subsection 3.1. By a change of notation, let  $G$  be the group given by reflections of three-dimensional unit vectors  $\mathbf{a}$  together with rotations around  $\mathbf{a}$ . This group is intransitive, and its orbits are found by fixing some  $\mathbf{a}$ . This corresponds to what was called  $G^a$  there, and a group reduction gives the quantum-mechanical interpretation of the example.

More statistical theory related to transitive and intransitive groups defined on the parameter space and the sample space is given in Helland<sup>[42]</sup>. It is of independent interest that the statistical model

corresponding to partial least squares regression can be motivated by model reduction to orbits of a certain group defined on the parameter space<sup>[4.6][4.7]</sup>.

### 3.3. Partial least squares regression and model reduction

Partial least squares regression is an algorithmic method for estimating the regression parameter  $\beta$ , intended for the case of collinearity. It was connected to a statistical model in Helland<sup>[4.8]</sup>. Briefly, this model can be formulated as follows: Let  $\Sigma_x$  be the covariance matrix of the  $p$  explanatory variables  $x_i; i = 1, \dots, p$ , assumed to be random, and let  $\{\mathbf{d}_i\}$  be the eigenvectors of  $\Sigma_x$ . Decompose the regression vector  $\beta$  of the predictor variable  $y$  upon  $\mathbf{x} = (x_1, \dots, x_p)$  as

$$\beta = \sum_{i=1}^p \gamma_i \mathbf{d}_i, \quad (27)$$

and then introduce the following hypothesis for some  $m < p$ :

$$H_m : \text{There are exactly } m \text{ nonzero terms in (27).}$$

There are two mechanisms by which the number of terms can be reduced: 1) Some terms are really zero; 2) There are coinciding eigenvalues of  $\Sigma_x$ , and then the eigenvectors may be rotated in such a way that there is only one in the relevant eigenspace that is along  $\beta$ .

Then the following is proved in Helland<sup>[4.8]</sup>: The parametric version of the partial least squares regression algorithm stops after  $m$  steps under the hypothesis  $H_m$ . Using the resulting partial least squares regression for prediction seems to give a good solution to the collinearity problem.

In Helland<sup>[4.7]</sup>, this is studied further. Among other results, one can prove the following: (Theorem 5 in op. cit.) Using a least squares criterion, the partial least squares model under certain technical conditions gives the best possible model reduction for linear prediction.

In discussing this and related results, it turns out to be of some relevance to use results from the foundation of quantum theory, in particular Theorem 1 from Subsection 2.2. Specifically, let  $\phi = (\beta, \Sigma_x, \sigma^2)$  be the full parameter of the model, where  $\sigma^2 = \text{Var}(y|x_i; i = 1, \dots, p)$ , let  $\theta = \theta(\phi)$  be the model reduced  $\beta$  under the hypothesis  $H_m$ , and let  $\eta = \eta(\phi)$  be any other  $m$ -parametric model reduction of  $\beta$ . Then it is shown in Helland<sup>[4.7]</sup> that the assumptions of Theorem 1 are satisfied.

Using this, it is shown: The technical condition ensuring that the PLS regression model ( $\theta$ ) is better than the arbitrary reduction  $\eta$  holds if a statistician  $B$  has a non-informative prior on  $\eta$ .



Furthermore, assume Postulate 5 and Postulate 6. Then the general version of the Born formula holds. In Postulate 6, we may specify the superior actor  $D$  to represent general scientific ideals connected to any statistician making the statistical analysis. Note that the probabilities here must be interpreted as probabilities as calculated by  $D$ , that is, probabilities assuming general scientific ideals.

### 3.4. Continuous parameters and complementarity

The discussion in this subsection must be considered tentative. There are mathematical issues that should be resolved.

In the traditional approach to quantum mechanics, the Hilbert space is directly determined by the variable considered. If this is the position of a particle, say, the Hilbert space is  $L^2(\mathbb{R}, d\mu)$ , where  $\mu$  is Lebesgue measure, and the operator corresponding to position  $x$  is just a multiplication by  $x$ .

Similarly, for a continuous statistical parameter  $\theta$  that varies over the whole space, and where the relevant group is the translation group, we can take the Hilbert space to be  $L^2(\mathbb{R}, d\mu)$ , and take the operator  $A^\theta$  corresponding to  $\theta$  to be multiplication of  $f \in L^2(\mathbb{R}, d\mu)$  by  $\theta$ .

This also determines the operator for any  $\xi = \xi(\theta)$ : By the spectral theorem, we have  $A^\theta = \int_{\sigma A^\theta} \lambda dE^{A^\theta}$ , which gives  $A^\xi = \int_{\sigma A^\theta} \xi(\lambda) dE^{A^\theta}$ . This reduces to multiplication with  $\xi(\theta)$  in this case.

Complementarity is a notion due to Niels Bohr, who called the position and momentum of a particle complementary variables. In my theory, in a statistical context, two parameters are called complementary if they are really different and both are maximal accessible variables. By Theorem 1, the existence of two such complementary parameters is the essential basis for the development of quantum phenomena in a statistical setting.

Subsection 3.1 gave a setting where two discrete complementary parameters implied quantum probability as a possible prior. For continuous parameters, the theory is more complicated. The point is that the Hilbert space  $L^2(\mathbb{R}, d\mu)$  is not separable and does not have a countable basis. On many occasions, also in this article, it may nevertheless be useful to think in terms of a finite set of basis vectors. This corresponds to parameters  $\theta$  and  $\eta$  taking a finite number of values. Continuous parameters may be approximated by such finite-valued parameters. For mathematicians, a direct strictly precise theory is given in Hall<sup>[27]</sup>.

#### Example

Consider a modified version of the example from Subsection 3.1. Assume that in the rash–medicine illustration, we really are able to measure the difference in improvement on the sides of each patient’s back, and thus in one set of measurements get an estimate of  $\zeta^a = \mu_a - \frac{1}{3}(\mu_b + \mu_c + \mu_d)$  and in another set of measurements get an estimate of  $\zeta^b = \mu_b - \frac{1}{3}(\mu_a + \mu_c + \mu_d)$ . These are contrasts, but not orthogonal contrasts.

We are interested in the contrast  $\zeta^b$ , but relative to a particular population for which we have information about the contrast  $\zeta^a$ . This information is obtained from a previous experiment that has been analyzed by either frequentist or Bayesian methods. In the first case, we have obtained a confidence distribution of the contrast  $\zeta^a$ , and in the last case, a posterior probability distribution. In either case, we possess now a probability density  $p^a(\zeta)$  for  $\zeta^a$ . We will use this to find a prior for the experiment on  $\zeta^b$ , relevant for the resulting population.

Now we use Theorem 1. The assumptions may be shown to be satisfied, see below, with  $\theta = \zeta^b$ ,  $\eta = \zeta^a$  and  $G$  equal to the translation group on  $\theta$ , and  $M$  equal to the translation group on  $\phi = (\mu_a, \mu_b, \mu_c, \mu_d)$ . The result is two symmetric operators,  $A^a$  corresponding to  $\zeta^a$  and  $A^b$  corresponding to  $\zeta^b$ . We can take  $A^a$  to be of the simple multiplication form. The operators will be self-adjoint. Using the spectral theorem for  $A^a$  ( $A^a = \int \zeta dE^a(\zeta)$ ) and the probability density  $p^a(\zeta)$ , we find a density operator  $\rho^a = \int p(\zeta) dE^a(\zeta)$ , and the spectral theorem for  $A^b$  ( $A^b = \int \eta dE^b(\eta)$ ) gives a projection operator  $\Pi_B = \int_B dE^b(\eta)$  for each Borel set  $B$ . Then, from Born’s formula, a prior for  $\zeta^b$  is given by

$$P(\zeta^b \in B | \rho^a) = \text{trace}(\rho^a \Pi_B). \quad (28)$$

It is crucial for this example that both the parameters  $\zeta^a$  and  $\zeta^b$  can be seen to be maximal as accessible parameters; see Definition 1 of Subsection 2.2. They belong to different experiments and must be maximal in these experiments: For instance, any parameter  $\lambda$  for the first experiment such that  $\zeta^a$  is a function of  $\lambda$  which is not bijective must be inaccessible, not possible to estimate from the available data. This is a rather strong requirement.

One such potential  $\lambda$  will be the vector  $(\mu_a, \frac{1}{3}(\mu_b + \mu_c + \mu_d))$ , and another will be the vector  $(\mu_a, \mu_b, \mu_c, \mu_d)$ . These must be inaccessible; it is only possible to measure contrasts/ differences between sides of the back. (This can be seen as a somewhat strange requirement, but it is here crucial for my arguments.) Furthermore, one must argue that priors should only depend upon accessible parameters. Then, a quantum prior determined by (28) can be motivated for the second experiment.

Note that in this example, if we take the basic inaccessible variable as  $\phi = (\mu_a, \mu_b, \mu_c, \mu_d)$ , and  $M$  as the translation group acting on  $\phi$ , then the contrast function  $\theta = \zeta^b = \mu_b - \frac{1}{3}(\mu_a + \mu_c + \mu_d)$  is not permissible with respect to  $M$ . (See Definition 2 of Subsection 3.1.) The group  $G$  acting on  $\theta$  is not induced by the full group  $M$ .

In the discussion above, I had assumed that we really had knowledge about  $\zeta^a$  for each unit, so that this was available for the selection of units. For the more realistic case, see below.

The essence of this example can be generalized. Assume a statistical model  $P(z|\phi)$  with density  $p(z|\phi)$  with some large parameter space  $\Omega_\phi$ . Let  $M$  be a group acting upon  $\Omega_\phi$ , and let  $\theta = \theta(\phi)$  be some focus parameter. Assume that  $\theta$  is maximally accessible in the given situation, that is, 1) It can be estimated using some estimation principle; 2) If  $\theta = f(\xi)$  for a non-invertible function  $f$ , then the parameter  $\xi$  cannot be estimated. Assume that there is a transitive group  $G$  with a trivial isotropy group acting on  $\Omega_\theta$ . Consider some experimental units, and make an experiment in accordance with the model on these units. This gives a Bayesian posterior or a confidence distribution (more generally, a fiducial distribution<sup>[49]</sup>) with density  $p(\theta)$ . In the Bayesian case, it is natural, if possible, to use as a prior the invariant measure associated with the group  $G$ .

Then do a new experiment on a selected set of units, selected according to the probability distribution  $p(\theta)$ . Let  $\eta = \eta(\phi)$  be another maximal accessible parameter, a focus parameter on the second experiment, and essentially different from  $\theta$ . Then, according to Theorem 1, there exists a Hilbert space  $\mathcal{H}$ , and two symmetric operators  $A^\theta$  and  $A^\eta$  in  $\mathcal{H}$ , one associated with  $\theta$  and one with  $\eta$ . Let  $A^\theta = \int \theta dE^\theta(\theta)$  be the spectral decomposition of  $A^\theta$ , and define  $\rho^\theta = \int p(\theta) dE^\theta(\theta)$ . Then it can be argued that a prior  $\pi$  for the second experiment should be chosen such that  $\pi(B) = \text{trace}(\rho^\theta \Pi_B)$ , where  $\Pi_B$  is the projection operator defined by  $\Pi_B = \int_B dE^\eta(\eta)$ , with  $\{E^\eta\}$  chosen such that  $A^\eta = \int \eta dE^\eta(\eta)$ .

All this assumes that units can be chosen by values of  $\theta$  that are really known. If not, we define  $u = \hat{\theta}$ , where  $\hat{\theta}$  is the chosen estimator of  $\theta$ , and we must replace  $\rho^\theta$  by  $\rho^u = \int r(u) dE^u(u)$ , where  $r(u) = \int q(u|\theta) p(\theta) d\theta$  with  $q$  being the density in the distribution of the estimator, assumed to only depend upon  $\theta$ , and where  $\{E^u\}$  is found from the spectral distribution of  $A^u = \int u q(u|\theta) dE^\theta(\theta)$ . Units are then chosen from data  $z$  of the first experiment according to the density  $r(u(z))$ .

Similar discussions can, in principle, be made in very complicated statistical models. For many such models, the groups  $G$  and  $M$  of Theorem 1 can be defined. An example from the design of experiments

where such groups can be defined is the model for randomized experiments discussed by Bailey<sup>[50]</sup>. But to discuss examples where links to quantum theory can be found, one has to have a basic inaccessible parameter  $\phi$ , and two really different maximal accessible (complementary) parameters.

An even more general class of statistical models is discussed by McCullagh<sup>[51]</sup> using category theory. Group theory can be seen as a special case of category theory, and using this, examples with groups  $G$  and  $M$  can be found. But again, it is a challenge to find applied examples with two different maximal accessible parameters.

It is obvious that more research in this area is required.

### 3.5. Decisions

For the purpose of this subsection, it is crucial that my foundation of quantum theory is also valid when relevant accessible variables are macroscopic. In the language of Khennikov<sup>[18][19]</sup>, I also want to include quantum-like models, which are applicable in biology, economics, psychology, and many other disciplines. Quantum structure may be ubiquitous. A special case is quantum cognition theory<sup>[41]</sup>, which includes quantum decision theory.

The traditional tool for decision-making in statistics is to minimize the expected value of a loss function. However, there are many decisions that are made in a statistical analysis, which cannot be seen in this way: The choice of model, the choice of method in the analysis, the choice of variables or set of variables to include in a multiple regression setting, or the choice to report or not report a  $p$ -value. These are examples of decisions made by a statistician or by a communicating group of statisticians, decisions that sometimes can be modeled by quantum decision theory<sup>[52]</sup>.

Decisions can be made on the basis of knowledge, on the basis of beliefs, or both. They are always made in a concrete context. Single persons can make decisions, and joint decisions can be made by a group of communicating persons.

Consider a person  $C$  or a group of persons in some decision situation. Say that he or she or they has/have the choice between a finite set of actions  $a_1, \dots, a_r$ . Relative to this situation, we can define a finite-valued decision variable  $\theta$ , taking the different values  $1, \dots, r$ , such that  $\theta = i$  corresponds to the action  $a_i$  ( $i = 1, \dots, r$ ). If  $C$  (or the group) really is (are) able to make a decision here and carry out the actions, we say that  $\theta$  is an accessible variable. As discussed for the general situation in Subsection

2.2, the variable  $\theta$  is in relation to a person  $C$  or to the group of persons; in fact, here  $\theta$  belongs to the mind of  $C$  or to the joint minds of the group.

This must be made precise. A decision problem is said to be maximal if  $C$  (or the group) is (are) just able to make his (their) mind(s) with respect to this decision; if the problem is made slightly more complicated, he (they) is (are) not able to make a decision. Let two (completely) different maximal decision variables be  $\theta$  and  $\eta$ , where  $\theta = i$  corresponds to the action  $a_i$  ( $i = 1, \dots, r$ ), and  $\eta = j$  corresponds to the action  $b_j$  ( $j = 1, \dots, r$ ). Then, by the theory of Subsection 2.2, we can model the situation by using quantum theory.

Note that both  $\theta$  and  $\eta$  may be vector variables. Say that  $\theta = (\theta_1, \dots, \theta_n)$ , where each  $\theta_j$  is a simpler decision variable. Then this corresponds to a situation where the actor(s), in addition to the difficult decision given by  $\eta$ , is (are) faced with  $n$  more simple decisions. Such a situation is not uncommon. In each situation where we shall make a difficult decision, we will be in a context where also a number of trivial decisions may have to be made just in order to survive and to function well in the given context. For many people, these trivial decisions occupy a large portion of their mind, such a large part that the vector decision variable  $\theta$  also must be considered to be maximal. The assumption that both  $\theta$  and  $\eta$  take the same number of values  $r$ , can be satisfied by artificially adding some actions to one of the decision problems.

It is crucial that both  $\theta$  and  $\eta$  can be seen as functions of some large inaccessible variable  $\phi$ . The solution here depends on which philosophy one has. One psychological theory might be that our more automatic decisions depend upon our culture and upbringing, which, modeled in some way, can be seen as a part of  $\phi$ . In addition,  $\phi$  must contain something that may be called our free will.

The simple model above does not cover all situations. Sometimes we have a choice between an infinite number of possibilities, and sometimes the outer context changes during the decision process. Nevertheless, the simple model is a good starting point.

It is well known that our minds may be limited, for instance, when faced with difficult decisions. I will first mention a side result in this direction from the present development.

In Helland<sup>[7]</sup>, Theorem 2 says essentially: Imagine a person  $C$  which, in some context, has two related maximal accessible related variables  $\theta$  and  $\eta$  in his mind. Impose a specific symmetry assumption. Then  $C$  cannot simultaneously have in mind any other maximal accessible variable which is related to  $\theta$ , but not related to  $\eta$ . It was claimed in Helland<sup>[7]</sup> that the violation of a famous inequality by

practical Bell experiments can be understood on the basis of this theorem. See also Helland (2023e), where a corresponding theorem is formulated without any symmetry assumption.

Note that this result has the qualification ‘at the same time’ and indicates a specific restriction to two maximal variables. But the human mind is very flexible. Taking time into account, we can think of very many variables, even ones that are not related.

For the present article, however, the direct results from Subsection 2.2 are equally important. Consider again a decision situation, and assume the simple model of the present subsection. In particular, let  $C$  at the same time be confronted with at least two different maximal related decision processes. Then the following hold:

- Each decision variable  $\eta$  is associated with a self-adjoint operator  $A^\eta$ , whose eigenvalues are the possible values of  $\eta$ .
- The decision process is maximal if and only if each eigenvalue of the corresponding operator is single.
- In the maximal case, the eigenvectors of the operator can be given interpretations: They are coupled to one particular decision process and a specific choice in this decision process: In concrete terms, the eigenvectors  $\mathbf{u}$  are in one-to-one correspondence with 1) some maximal accessible decision variable  $\eta$ , and 2) a specific value  $u$  of  $\eta$ . In other words, the possible eigenvectors are in one-to-one correspondence with 1) the question ‘Which decision process?’ and 2) ‘Which action did this decision process lead to?’.
- In the general case, the eigenspaces of the operator have a similar interpretation.

This can be taken as a starting point of quantum decision theory, but to develop this theory further, we need to be able to calculate probabilities for the various decisions. For this, I refer to the discussion of the Born rule in Subsection 2.3. In particular, note that the probabilities are assumed to be calculated in a way that can be associated with some (abstract or concrete) superior being, assumed to be perfectly rational. The interpretation of this point also depends on our philosophy. My own view is discussed in Helland<sup>[53]</sup>.

### 3.6. Machine Learning

The literature on Artificial Intelligence, in particular Machine Learning, has exploded in recent years. For the purpose of this article, I will focus on a simple neural network with one hidden layer and a

single output, as described from a statistical point of view in Efron and Hastie<sup>[54]</sup>. Here, assume  $p$  predictors (features)  $\mathbf{x} = (x_1, x_2, \dots, x_p)$  which for simplicity are centered on zero expectation,  $k$  hidden units  $a_l = g(\sum_{j=1}^p w_{jl}x_j)$  ( $l = 1, \dots, k$ ) and output  $z = h(\sum_{l=1}^k v_l a_l)$ , where  $g$  and  $h$  are non-linear, monotonically increasing functions satisfying  $g(0) = 0$  and  $h(0) = 0$ , and  $\{w_{jl}\}$  and  $\{v_l\}$  are the parameters of the model, the weights. The  $\{x_j\}$  and  $y$  are observed on  $n$  units, and our task - the learning of the network - is to estimate the weights. In many applications,  $n$  is very large, and procedures such as backpropagation are used. I will here also consider the case where  $n$  is moderate, perhaps smaller than  $p$ . Then a model reduction may be called for.

Before discussing this, I take a brief look at some of the recent literature concerning links between machine learning and quantum theory.

In their abstract, Dunjko and Briegel<sup>[3]</sup> mention 3 points: 1) Quantum computing is finding vital applications in providing speed-ups for machine learning problems. 2) Machine learning may become instrumental in advanced quantum technologies. 3) One can consider quantum generalizations of learning and artificial intelligence concepts.

Op. cit. is a review article with many references to recent papers. Quantum models of relevance to machine learning are discussed in detail. Historically, the first such model was the quantum Turing machine<sup>[55]</sup>, but there are many more modern models. In general, machine learning can be divided into supervised and unsupervised learning. In supervised learning, we start with a training set  $(y_i, \mathbf{x}_i), i = 1, \dots, n$ , where  $\mathbf{x}_i$  is a vector, and the task is to predict  $y$  from  $\mathbf{x}$  on a new unit, so multiple regression can be seen as a special case of machine learning. Quantum information is a more general concept, where data are replaced by quantum states. This is a large area with independent literature. Again, specializing to the multiple regression case, one can mention works by Wiebe et al.<sup>[56]</sup>, Wang<sup>[57]</sup>, and Schuld et al.<sup>[58]</sup>.

In two recent articles, Wu et al.<sup>[59]</sup> and Zhu et al.<sup>[60]</sup> develop network models that can simultaneously predict multiple quantum properties and the behavior of an unknown quantum process.

Quantum foundations seek to understand and develop the mathematical and conceptual basis for quantum theory. Bharti et al.<sup>[61]</sup> survey representative works at the interface of machine learning and quantum foundations. Special topics considered are entanglement, Bell-type inequalities, and contextuality. It is proposed that neural networks can be seen as 'hidden' variable models for quantum systems.

Go back to the simple model introduced in the first paragraph of this Section. We are interested in model reduction. The whole neural net is usually learned by gradient descent<sup>[54][62]</sup>. For the purpose of model reduction, I will concentrate on a single perceptron

$$a = g\left(\sum_{j=1}^p w_j x_j\right). \quad (29)$$

I now, as a model, assume that the features  $\mathbf{x} = (x_1, \dots, x_p)$  have a random distribution with expectation  $\mathbf{0}$  and covariance matrix  $\Sigma$ , assumed to be positive definite. Expand  $\mathbf{w} = (w_1, \dots, w_p)$  in terms of eigenvectors  $\mathbf{d}_i$  of  $\Sigma$

$$\mathbf{w} = \sum_{i=1}^p \gamma_i \mathbf{d}_i. \quad (30)$$

Then, completely in analogy with partial least squares regression (Subsection 3.3), introduce the model reduction  $(\theta_m)$

$H_m$  : There are exactly  $m$  nonzero terms in (30).

The theory of Helland<sup>[47]</sup> carries over. We can represent  $H_m$  with the parameter  $\theta_m = (\gamma_1 \mathbf{d}_1, \dots, \gamma_m \mathbf{d}_m)$ , which is a function of  $\phi = (\mathbf{w}, \Sigma)$ . Assume any other model reduction  $\eta_m$  to  $m$  terms. Assume a non-informative probability distribution of  $\eta_m$ . Then, it can be proved that by a least squares criterion,  $\theta_m$  gives a better model reduction than  $\eta_m$ .

In proving this in op. cit., essential use was made of a joint quantum model for  $\theta_m$  and  $\eta_m$ . There, and here, the assumptions of Theorem 1 (Subsection 2.2) can be shown to be satisfied. The group  $G$  acting on  $\theta$  is given by orthogonal transformations of the  $\mathbf{d}_j$ 's and  $\gamma_j \mapsto \alpha_j \gamma_j$  ( $j = 1, 2, \dots, m$ ). It is convenient to let the group  $M$  on  $\phi$  be defined by orthogonal transformations of all the  $\mathbf{d}_i$ 's ( $i = 1, \dots, p$ ) and by  $\gamma_i \mapsto g(\alpha_i \gamma_i)$ . Then the orbits of the group  $M$  are given by  $m$  and the hypothesis  $H_m$ . (Theorem 2 in Helland<sup>[47]</sup>). This is the reason why I have chosen the constraint  $g(0) = 0$ .

To carry out the model reduction for the perceptron in theory, and also in practice, the whole literature on partial least squares can be taken over if we base ourselves on  $\mathbf{x}$  and  $y = g^{-1}(a) [= \sum_{j=1}^p w_j x_j]$ . The theoretical population algorithm is given in Appendix 1 of Helland<sup>[47]</sup>. In practice, we have data on  $n$  copies of  $(\mathbf{x}, y)$ , and in the algorithm, theoretical variances and covariances must be replaced by estimated variances and covariances. The size  $m$  can be determined by cross-validation.



Then all this should be incorporated into the algorithm for the whole neural network. Look again at the simple model defined at the beginning of this Subsection. We are given data  $(x, z)$  on  $n$  units, where all these variables are centered by their means. We use a feedforward procedure, which means that we start the estimation by first looking at the transition from the  $x$ -data to the variables  $a_i$ , then from  $a_i$  to  $z$ . The now well-known procedures here can be found in the machine learning literature. Each of these steps must now be replaced by a series of steps of the type described in the previous paragraph. I omit the details here, since the purpose of the present article is to introduce ideas.

A model reduction of this kind can be expected to give an advantage when there is not too much data ( $n$ ) compared to the number  $p$  of variables, or perhaps more accurately, as shown in the partial least case by recent asymptotics by Cook and Forzani<sup>[63]</sup>, in the abundant case where many predictors  $x$  contribute information about the response  $z$ , often correlated information.

## 4. Concluding remarks

The social scientist Ralph D. Stacey once said: 'Culture is a set of attitudes, opinions and convictions that a group of humans share, about how one shall behave against each other, how things shall be evaluated and done, which questions that are important and which answers that are accepted. The most important elements of culture are unconscious, and cannot be forced upon us from the outside.'

From this perspective, statistical theory and quantum theory, as they have functioned up to now, may be seen as connected to separate cultures. It is hoped that this article may help to bridge the gap between these two cultures.

The investigations here started with Helland<sup>[64]</sup>, where mathematical models in various sciences were discussed from several points of view. With the present article, this discussion can be said to lead to concrete results.

Of course, there are differences between models in quantum theory on the one side and statistically related models on the other side. It is important that quantum models are always seen in a context, and that they often are related to an observer or a group of communicating observers. Contextual quantum measurements have been discussed from several points of view by Khrennikov<sup>[65]</sup>. By contrast, statistical models are more universal. But this does not contradict the fact that quantum-like models are ubiquitous, cf. Khrennikov<sup>[18][19]</sup>.

## Acknowledgements

I really want to thank Andrei Khrennikov for his yearly, very enlightening conferences on quantum foundations and related topics. I have learned a lot by attending a few of these conferences. I also want to thank Christopher Fuchs and Richard Gill for their patience in trying to understand my basic message. Finally, I am grateful to Wolfgang Tiefenbrunner for making me aware of the works by Hans De Raedt et al. I am grateful to Solve Sæbø and Trygve Almøy for discussions, and I am grateful to Gudmund Hermansen for doing numerical calculations in connection to the experiment in Subsection 3.1.

## References

- <sup>1</sup> <sup>^</sup>Kolmogorov AN (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin. English translation: Kolmogorov AN (1956). *Foundations of the Theory of Probability*. 2. edition. Chelsea, New York.
- <sup>2</sup> <sup>^</sup>Barndorff-Nielsen OE, Gill RD, Jupp PE (2003). "On quantum statistical inference". *J R Statist Soc B*. 65 (4): 1-31.
- <sup>3</sup> <sup>a</sup>, <sup>b</sup>Dunjko V, Briegel HJ (2019). "Machine learning & artificial intelligence in the quantum domain: a review of recent progress". *Rep Prog Phys*. 81: 074001.
- <sup>4</sup> <sup>a</sup>, <sup>b</sup>, <sup>c</sup>, <sup>d</sup>, <sup>e</sup>, <sup>f</sup>, <sup>g</sup>, <sup>h</sup>, <sup>i</sup>Helland IS (2021). *Epistemic Processes. A Basis for Statistics and Quantum Theory*. 2. edition. Springer, Berlin.
- <sup>5</sup> <sup>^</sup>Helland IS, Parthasarathy H (2024). *Theoretical Variables, Quantum Theory, Relativistic Quantum Field Theory, and Quantum Gravity*. Manakin Press, New Dehli.
- <sup>6</sup> <sup>a</sup>, <sup>b</sup>Helland IS (2022a). "On reconstructing parts of quantum theory from two related maximal conceptual variables". *International Journal of Theoretical Physics*. 61: 69.
- <sup>7</sup> <sup>a</sup>, <sup>b</sup>, <sup>c</sup>, <sup>d</sup>, <sup>e</sup>Helland IS (2022b). "The Bell experiment and the limitation of actors". *Foundations of Physics*. 52: 55.
- <sup>8</sup> <sup>a</sup>, <sup>b</sup>Helland IS (2023). "A new foundation of quantum decision theory." *arXiv:2310.12762 [quant-ph]*.
- <sup>9</sup> <sup>^</sup>Helland IS (2023). "Possible connections between relativity theory and a version of quantum theory based upon theoretical variables." *arXiv:2305.15435 [physics.hist-ph]*.
- <sup>10</sup> <sup>a</sup>, <sup>b</sup>Helland IS (2023). "A simple quantum model linked to decisions." *Foundations of Physics* 53, 12.
- <sup>11</sup> <sup>a</sup>, <sup>b</sup>, <sup>c</sup>Helland IS (2023). "On the Bell experiment and quantum foundation." *arXiv:2305.05299 [quant-ph]*. *J. Mod. Appl. Phys.* 6 (2), 1-5.

12. <sup>a, b, c, d, e, f, g</sup>Helland IS (2024). "An alternative foundation of quantum theory." *Foundations of Physics* 54, 3. arXiv:2305.06727 [quant-ph].
13. <sup>a, b, c, d</sup>Helland IS (2024). "On probabilities in quantum mechanics." arXiv:2401.17717 [quant-ph]. *APL Quantum* 1, 036116. doi:10.1063/5.0218982.
14. <sup>a, b, c, d, e, f</sup>Helland IS (2024). "A new approach toward the quantum foundation and some consequences." *Academia Quantum* 1. doi:AcadQuant7282.
15. <sup>a, b, c, d, e, f, g</sup>Helland IS (2024). "Some mathematical issues regarding a new approach towards quantum foundation." arXiv:2411.13113 [quant-ph].
16. <sup>a, b, c, d, e</sup>De Raedt H, Katsnelson MJ, Michelson K (2014). "Quantum theory as the most robust description of reproducible experiments". *Annals of Physics*. 347: 45–73.
17. <sup>^</sup>Frieden BR (2004). *Science from Fisher Information: A Unification*. Cambridge University Press, Cambridge.
18. <sup>a, b, c, d, e, f</sup>Khrennikov A (2010). *Ubiquitous Quantum Structure. From Psychology to Finance*. Springer, Berlin.
19. <sup>a, b, c, d, e, f</sup>Khrennikov A (2023). *Open Quantum Systems in Biology, Cognitive and Social Sciences*. Springer Nature Switzerland.
20. <sup>a, b</sup>Haven E, Khrennikov A (2013). *Quantum Social Science*. Cambridge University Press, Cambridge.
21. <sup>a, b</sup>Haven E, Khrennikov A (2016). "Statistical and subjective interpretations of probability in quantum-like models in cognition and decision-making". *J Mathematical Psychology*. 74: 82–91.
22. <sup>^</sup>Caves CM, Fuchs CA, Schack R (2002). "Quantum probabilities as Bayesian probabilities". *Physical Review A* 65: 022305.
23. <sup>^</sup>Fuchs CA (2003). "Quantum mechanics as quantum information, mostly". *J Mod Opt*. 50: 987–1023.
24. <sup>^</sup>Fuchs CA (2010). "QBism, the perimeter of quantum Bayesianism". arXiv: 1003.5209 [quant-ph].
25. <sup>^</sup>Liu H (2024). "A quantum Bayes' rule and related inference." *Quantum Information Processing* 23:271, 1–26.
26. <sup>^</sup>von Neumann J (1932). *Mathematische Grundlagen der Quantenmechanik*. Springer, Berlin. Translated: *Mathematical Foundation of Quantum Mechanics*. Princeton University Press, Princeton.
27. <sup>a, b, c, d, e, f</sup>Hall BC (2013). *Quantum Theory for Mathematicians*. Graduate Texts in Mathematics, 267, Springer, Berlin.

28. <sup>^</sup>Meyer P-A (1995). *Quantum Probability for Probabilists*. 2. edition. *Lecture Notes in Mathematics* 1538, Springer-Verlag, Berlin.
29. <sup>^</sup>Holovo AS (2001). *Statistical Structure of Quantum Theory*. *Lecture Notes in Physics*. Springer-Verlag, Berlin.
30. <sup>^</sup>Plotnitsky A (2013). *Niels Bohr and Complementarity. An Introduction*. Springer, New York.
31. <sup>^</sup>Steiner M, Rendell R (2024). "Complementary relationships between entanglement and measurement". *Academia Quantum*. 1.
32. <sup>^</sup>Maccone L (2024). "Schrödinger cats and quantum complementarity." *Foundations of Physics* 54, 17.
33. <sup>a, b</sup> Schweder T, Hjort NL (2016). *Confidence, Likelihood, Probability. Statistical Inference with Confidence Distributions*. Cambridge University Press, Cambridge.
34. <sup>^</sup>Berger JO, Wolpert RL (1988). *The likelihood principle*. Institute of Mathematical Statistics, Hayward, C A.
35. <sup>^</sup>Busch P (2003). "Quantum states and generalized observables: A simple proof of Gleason's Theorem". *Physical Review Letters*. 91 (12): 120403.
36. <sup>a, b, c</sup> De Raedt H, Katsnelson MJ, Michelson K (2016). "Quantum theory as plausible reasoning applied to data obtained by robust experiments". *Phil Trans Roy Soc A*. 374: 20150233.
37. <sup>^</sup>Jaynes ET (2003). *Probability Theory: The Logic of Science*. Ed.: G.L. Bretthorst. Cambridge University Press, Cambridge.
38. <sup>^</sup>De Raedt H, Katsnelson MJ, Jattana MS, Mehta V, Willsch M, Willsch D, Michelson K, Jin F (2023). "Einstein-Podolsky-Rosen-Bohm experiments: A discrete data driven approach". *Annals of Physics*. 453: 169-314.
39. <sup>^</sup>Savage IJ (1954) *The Foundations of Statistics*. Wiley, New York.
40. <sup>^</sup>Shrapnel S, Costa F, Milburn G (2017). "Updating the Born rule." *arXiv:1702.01845 [quant-ph]*.
41. <sup>a, b</sup> Bussemeyer JR, Bruza PD (2012). *Quantum Models of Cognition and Decision*. Cambridge University Press, Cambridge.
42. <sup>a, b, c, d</sup> Helland IS (2004). "Statistical inference under symmetry". *Int Stat Rev*. 72 (3): 409-422.
43. <sup>^</sup>Helland IS (2006). "Extended statistical modeling under symmetry: The link toward quantum mechanics". *Ann Statist*. 34 (1): 42-77.
44. <sup>^</sup>Helland IS (2008). "Quantum mechanics from focusing and symmetry". *Found Phys*. 38: 818-842.
45. <sup>^</sup>Helland IS (2000). "Model reduction for prediction in regression models". *Scand J Statist*. 27 (1): 1-20.

46. <sup>^</sup>Helland IS, Sæbø S, Tjelmeland H (2012). "Near optimal prediction from relevant components." *Scand. J. Statist.* 39, 695-713.
47. <sup>a, b, c, d, e, f</sup>Helland IS (2024). On optimal linear prediction. *arXiv:2444412.19186 [math.ST]*.
48. <sup>a, b</sup>Helland IS (1990). "Partial least squares regression and statistical models". *Scand J Statist.* 17: 97-114.
49. <sup>^</sup>Murph AC, Hannig J, Williams JP (2024). "Introduction to generalized fiducial inference." In: Berger J, Meng X-L, Reid N, Xie M M-g (2024). *Handbook of Bayesian, Fiducial, and Frequentist Inference*. Chapman and Hall.
50. <sup>^</sup>Bailey RA (1991). "Strata for randomized experiments". *J Roy Statist Soc B.* 51 (1): 27-68.
51. <sup>^</sup>McCullagh P (2002). "What is a statistical model?" *Ann. Statist.* 30 (5), 1225-131.
52. <sup>^</sup>Helland IS (2024). "On the foundation of quantum decision theory." *arXiv:2310.12762 [quant-ph]*.
53. <sup>^</sup>Helland IS (2023). Quantum mechanics as a theory that is consistent with the existence of God. *Dialogo Conf J.* 10(1):127-34.
54. <sup>a, b</sup>Efron B, Hastie T (2021). *Computer Age Statistical Inference 2. Edition*. Cambridge University Press, Cambridge.
55. <sup>^</sup>Deutch D (1985). "Quantum theory, the Church-Turing principle and the universal quantum computer". *Proc R Soc A.* 400: 97-117.
56. <sup>^</sup>Wiebe N, Braun D, Lloyd S (2012). "Quantum algorithm for data fitting". *Phys. Rev. Lett.* 109: 050505.
57. <sup>^</sup>Wang G (2017). "New quantum algorithm for linear regression". *Phys. Rev. A.* 96: 012335.
58. <sup>^</sup>Schuld M, Sinaiskiy I, Petruccione F (2016). "Prediction by linear regression on a quantum computer." *Phys. Rev. A* 94, 022342.
59. <sup>^</sup>Wu Y-D, Zhu Y, Wang Y, Chirabella G (2023). "Learning and discovering quantum properties with multi-task neural networks". *arXiv: 2310.11807 [quant-ph]*.
60. <sup>^</sup>Zhu Y, Wu Y-D, Liu Q, Wang Y, Chirabella G (2023). "Predictive modelling of quantum process with neural networks". *arXiv: 2308.08815 [quant-ph]*.
61. <sup>^</sup>Bharti K, Haug T, Vedral V, Kwek L-C (2020). "Machine learning meets quantum foundations: A brief survey". *AVS Quantum Sci.* 2: 034101.
62. <sup>^</sup>Charniak E (2018). *Introduction to Deep Learning*. The MIT Press, Cambridge, Massachusetts.
63. <sup>^</sup>Cook RD, Forzani L (2019). "Partial least squares prediction in high-dimensional regression". *Ann Stat.* 47 (2): 884-908.

64. <sup>^</sup>Helland IS (2022). "On the diversity and similarity of mathematical models in science." *American Review of Mathematics and Statistics* 10 (1), 1-10.

65. <sup>^</sup>Khrennikov A (2024). "Contextual measurement model and quantum theory." *Royal Society Open Science* 11, 231953.

## **Declarations**

**Funding:** No specific funding was received for this work.

**Potential competing interests:** No potential competing interests to declare.