New Journal of Physics
The open access journal at the forefront of physics

Deutsche Physikalische Gesellschaft DPG IOP Institute of Physics

Optimal error regions for quantum state estimation

New Journal of Physics
The open access journal for physics

Optimal error regions for quantum state estimation

Jiangwei Shang[1,6], **Hui Khoon Ng**[1,2,4], **Arun Sehrawat**[1,5], **Xikun Li**[1] **and Berthold-Georg Englert**[1,3]

[1] Centre for Quantum Technologies, National University of Singapore,
3 Science Drive 2, Singapore 117543, Singapore
[2] Applied Physics Lab, DSO National Laboratories, 20 Science Park Drive,
Singapore 118230, Singapore
[3] Department of Physics, National University of Singapore, 2 Science Drive 3,
Singapore 117542, Singapore
E-mail: jiangwei.shang@quantumlah.org

New Journal of Physics **15** (2013) 123026 (26pp)
Received 20 July 2013
Published 13 December 2013
Online at http://www.njp.org/
doi:10.1088/1367-2630/15/12/123026

Abstract. An estimator is a state that represents one's best guess of the actual state of the quantum system for the given data. Such estimators are points in the state space. To be statistically meaningful, they have to be endowed with error regions, the generalization of error bars beyond one dimension. As opposed to standard ad hoc constructions of error regions, we introduce the maximum-likelihood region—the region of largest likelihood among all regions of the same size—as the natural counterpart of the popular maximum-likelihood estimator. Here, the size of a region is its prior probability. A related concept is the smallest credible region—the smallest region with pre-chosen posterior probability. In both cases, the optimal error region has constant likelihood on its boundary. This surprisingly simple characterization permits concise reporting of the error regions, even in high-dimensional problems. For illustration, we identify optimal error regions for single-qubit and two-qubit states from computer-generated data that simulate incomplete tomography with few measured copies.

[4] Current address: Yale-NUS College, 6 College Avenue East, Singapore 138614, Singapore.
[5] Current address: Departamento de Física, Universidad de Guadalajara, 44420 Guadalajara, Jalisco, Mexico.
[6] Author to whom any correspondence should be addressed.

Contents

1. Introduction

Quantum state estimation (see e.g. [1]) is central to many, if not all, tasks that process quantum information. The characterization of a source of quantum carriers, the verification of the properties of a quantum channel and the monitoring of a transmission line used for quantum key distribution—all three require reliable quantum state estimation, to name just the most familiar examples.

In the typical situation that we are considering, several independently and identically prepared quantum-information carriers are measured one-by-one by an apparatus that realizes a probability-operator measurement (POM), suitably designed to extract the wanted information. The POM has a number of outcomes, with detectors that register individual information carriers (photons in the majority of current experiments), and the data consist of the observed sequence of detection events (‘clicks’)[7].

The quantum state to be estimated is described by a statistical operator, the *state*, and the data can be used to determine an *estimator* for the state—another state that, so one hopes, approximates the actual state well. There are various strategies for finding such an estimator. Thanks to the efficient methods that Hradil, Řeháček, and their collaborators developed for calculating maximum-likelihood estimators (MLEs, reviewed in [2]; see also [3]), MLEs have

[7] It is advisable to verify that the observed sequence does not have systematic correlations that speak against the assumption of independently and identically prepared quantum-information carriers.

become the estimators of choice. For the given data, the MLE is the state for which the data are more likely than for any other state.

Whether one prefers the MLE or a point estimator found by another method, the data have statistical noise and, therefore, one needs to supplement the point estimator with error bars of some sort—*error regions*, more generally, for higher-dimensional problems. Many recipes, which are often ad hoc in nature, have been proposed for attaching a vicinity of states to an estimator. These usually rely on having a lot of data [4, 5], involve data resampling [6] or consider all data that one might have observed [7, 8]. By contrast, we systematically construct error regions from the data *actually* observed.

For this purpose, we propose *maximum-likelihood regions* (MLRs) and *smallest credible regions* (SCRs). These are regions in the space of quantum states (more precisely: in the reconstruction space; see section 2). The MLR is that region of pre-chosen size for which the given data are more likely than for any other region of the same size. The SCR is the smallest region with pre-chosen credibility—the credibility of a region being the probability of finding the actual state in the region, conditioned on the data (see e.g. [9]). Whether one chooses the MLR or the SCR as the optimal error region depends on the situation at hand.

Central to both concepts is the notion of the *size* of a region. In fact, some notion of size must underlie *any* useful definition of error regions, since one usually aims at reporting an error region that is not unnecessarily large—a judgement that can only be made with a suitable concept of size. We agree with Evans *et al* [10] that, in the context of state estimation, it is most natural to measure the size of a region by its prior—before any data are at hand—probability of finding the actual state in the region: regions with the same prior probability are considered as having the same size. Hence, the size of a region expresses the relative importance of that region of states.

The identification 'size $\equiv$ prior probability' is technically possible because both quantities simply add when disjointed regions are combined into a single region. While for some tasks one prefers not to assign a prior[8], since state estimation expresses our best attempt at guessing the state, any prior information we possess should be taken into account in the estimation process alongside the data. Much guidance on choosing priors can be found in the standard statistics literature; in appendix A, we provide a summary that focuses on points relevant in quantum contexts. Ultimately, the choice of prior is up to the user, but it should be *consistent*: the estimation results should be dominated by the data, not the prior, if many copies of the state are measured.

As we show below, the problems of finding the MLR and the SCR are duals of each other. In both cases, the optimal regions contain all states for which the likelihood of the data exceeds a threshold value. This provides a concise way of communicating one's uncertainty of the estimate. That the optimal error regions possess such a simple description is surprising, since our construction imposes no restriction on the shape of the regions to be considered. The shape of the optimal regions are uniquely determined by the likelihood function, in sharp contrast to the arbitrariness in the shape of a confidence region (see appendix B), a concept that is the subject of recent discussion [7, 8]. Yet the two are not unrelated: our SCRs provide natural starting points for the construction of the confidence regions considered in [7].

[8] For tasks like quantum key distribution, one may want to adopt a different attitude and assume the worst possible scenario rather than relying on one's information to assign a prior. Then, the confidence regions of [7, 8] are appropriate as error regions.

While the chosen MLR or SCR depends on the prior, the set of candidate regions is prior-independent: it depends only on the likelihood function for the given data. Also reassuring is the fact that every MLR or SCR is a small vicinity of the MLE, in the respective limits of small size or small credibility. This is reminiscent of standard ellipsoidal error regions constructed around the MLE, which are, however, applicable only in the limit of a large amount of data when the central limit theorem can be invoked and the uncertainty can be characterized by the Fisher information (see, for instance, [4]).

Here is a brief outline of the paper. We set the stage in section 2 where we introduce the reconstruction space, discuss the size of a region, and define the various joint and conditional probabilities. Equipped with these tools, we then formulate in section 3 the optimization problems that identify the MLRs and SCRs and find their solutions in section 4. We illustrate the matter by simulated single-qubit and two-qubit experiments in sections 5 and 6, and close with an outlook. Two appendices supply additional material: a guide for choosing priors, and a comparison with confidence regions.

2. Setting the stage

2.1. Reconstruction space

The outcomes Π_k, $k = 1, 2, \ldots, K$, of the POM, with which the data are acquired, are positive ($\Pi_k \geqslant 0$) Hilbert-space operators with $\sum_{k=1}^{K} \Pi_k = 1$. If the state ρ describes the system, the probability p_k that the kth detector clicks is

$$p_k = \mathrm{tr}\{\Pi_k \rho\} = \langle \Pi_k \rangle \quad \text{(Born rule)}. \tag{1}$$

The positivity of ρ and its normalization to unit trace ensure $p_k \geqslant 0$ for all k and $\sum_{k=1}^{K} p_k = 1$. Probabilities $p = (p_1, p_2, \ldots, p_K)$ for which there is a state ρ such that (1) holds are *permissible* probabilities. They make up the *probability space*. The probability space for a K-outcome POM is usually smaller than that of a tossed K-sided die because not all positive p_k with unit sum are permissible. The quantum nature of the state estimation problem enters *only* in these additional restrictions on p: quantum state estimation is standard statistical state estimation with constraints of quantum-mechanical origin. The rich methods of statistical inference immediately apply, modified where necessary to account for the restricted probability space.

Whereas p is uniquely determined by ρ through (1), the converse is true only if the POM is informationally complete. In any case, there is always a *reconstruction space* $\mathcal{R}_0$, a set of ρ that contains exactly one ρ for each permissible p, consistent with the Born rule. If there is more than one reconstruction space, it does not matter which one we choose. As an example, consider a harmonic oscillator with its infinite-dimensional state space. If the POM has two outcomes, with p_1 equal to the probability of finding the oscillator in its ground state and $p_2 = 1 - p_1$, the reconstruction space is the set of convex combinations of the projector to the ground state and another state with no ground-state component. In this situation, there are very many reconstruction spaces to choose from because *any* other state serves the purpose, and all one can infer from the data is an estimate of the ground-state probability.

Since the probability space is unique, while there can be many different reconstruction spaces, it is often more convenient to work in the probability space. In particular, the probability space has the desirable property that it is always convex; it is, however, not always possible to find a convex reconstruction space. The primary objective of state estimation is then to find an estimator, or a region of estimators, for the probabilities p. The conversion of p into a state ρ can be performed later, if at all. At this stage, if the POM is not informationally complete, one must invoke additional criteria—beyond what the data tells us—for a unique mapping $p \to \rho$. For example, one could follow Jaynes's guidance [11, 12] and maximize the entropy [13] (see also [14]). However, following the tradition in this topic, we will formally work in a reconstruction space $\mathcal{R}_0$, although all actual calculations are performed in the probability space. Estimators are states in $\mathcal{R}_0$, and regions are sets of states there.

2.2. *Size and prior content of a region*

The reconstruction space is an abstract construct that is often not endowed with a self-suggesting unique metric. Instead, a region's prior probability—the quantity that matters most in the present context of statistical inference—offers a natural notion of size. This relieves us of the need to invoke additional, possibly artificial, criteria for the assignment of size, for instance, one that has more to do with a simple parameterization of the state space than the relative importance of different regions in terms of our prior expectations.

We denote by $(\mathrm{d}\rho)$ the size ($\equiv$ prior probability) of the infinitesimal vicinity of state ρ in $\mathcal{R}_0$. The size $S_{\mathcal{R}}$ of a region $\mathcal{R}$ is then

$$S_{\mathcal{R}} = \int_{\mathcal{R}} (\mathrm{d}\rho) \quad \text{with} \int_{\mathcal{R}_0} (\mathrm{d}\rho) = 1. \tag{2}$$

A pertinent remark: we exclude pathological cases of improper priors where the prior density $(\mathrm{d}\rho)$ cannot be normalized; should an improper prior be useful in a particular context, it should come about as the limit of a well-defined sequence of proper priors.

By construction, $S_{\mathcal{R}}$ does not depend on the parameterization used for the numerical representation of $(\mathrm{d}\rho)$. The primary parameterization is in terms of the probabilities,

$$(\mathrm{d}\rho) = (\mathrm{d}p)\, w(p) \quad \text{with } (\mathrm{d}p) = \mathrm{d}p_1\, \mathrm{d}p_2 \cdots \mathrm{d}p_K, \tag{3}$$

where the prior density $w(p)$ is positive for all permissible probabilities and vanishes for non-permissible probabilities. To enforce positivity and normalization of p, $w(p)$ always contains

$$w_0(p) = \eta(p_1)\, \eta(p_2) \cdots \eta(p_K)\, \delta\left(\sum_k p_k - 1\right) \tag{4}$$

as a factor, where $\eta(\;)$ denotes Heaviside's unit step function and $\delta(\;)$ is Dirac's delta function. If there are no other constraints, we have the probability space of a K-sided die. For genuine quantum measurements, however, there are additional constraints, some accounted for by more delta-function factors, others by step functions. The delta-function constraints reduce the dimension of the reconstruction space from $K-1$ to the number of independent probabilities. Accordingly, there is a factor of constraint $w_{\mathrm{cstr}}(p)$ (containing $w_0(p)$) that specifies the probability space and appears in all possible priors. In particular, there are two specific priors we will employ as examples below: the *primitive prior*

$$(\mathrm{d}\rho) \propto (\mathrm{d}p)\, w_{\mathrm{cstr}}(p) \tag{5}$$

and the *Jeffreys prior* [15]

$$(\mathrm{d}\rho) \propto (\mathrm{d}p)\, w_{\mathrm{cstr}}(p)\, \frac{1}{\sqrt{p_1 p_2 \cdots p_K}}, \tag{6}$$

which is a popular choice of an unprejudiced prior [16].

For the harmonic-oscillator example in section 2.1, which has the same probability space as a tossed coin, the factor $w_0(p)$ selects the line segment with $0 \leqslant p_1 = 1 - p_2 \leqslant 1$ in the $p_1 p_2$ plane. If we choose the primitive prior $(\mathrm{d}\rho) = (\mathrm{d}p)\, w_0(p)$, the subsegment with $a \leqslant p_1 \leqslant b$ has size $b - a$. For the Jeffreys prior

$$(\mathrm{d}\rho) = (\mathrm{d}p)\, w_0(p) \frac{1}{\pi \sqrt{p_1 p_2}}, \tag{7}$$

the same subsegment has size $\frac{2}{\pi}[\sin^{-1}(\sqrt{b}) - \sin^{-1}(\sqrt{a})]$.

In this example, and also in those we use for illustration in section 5 below, it is easy to state quite explicitly the restrictions on the set of permissible probabilities that follow from the Born rule. In other situations, including the examples of section 6, this is more difficult. In yet more complicated situations it could be impossible. It is, however, possible to check numerically if a certain $\tilde{p} = (\tilde{p}_1, \ldots, \tilde{p}_K)$ is permissible. For example, one calculates a MLE (which can be done efficiently) for relative frequencies $n_k/N = \tilde{p}_k$ (see below), and if the resulting probabilities p are such that $p = \tilde{p}$, then $\tilde{p}$ is permissible; otherwise it is not. For practical reasons, it may be necessary to truncate the full state space—which can be, and often is, infinite-dimensional—to a test space of manageable size. With such a truncation one accepts that not all permissible probabilities are investigated. Therefore, a criterion for judging if the test space is large enough is to verify that the estimated probabilities do not change significantly when the space is enlarged. Examples for the artifacts that result from test spaces that are too small can be found in [17].

2.3. *Point likelihood, region likelihood and credibility*

The data D consist of a sequence of detector clicks, with n_k clicks in total of the kth detector after measuring $N = n_1 + n_2 + \cdots + n_K$ copies of the state[9]. The probability of obtaining D, *if* ρ is the state, is the *point likelihood*

$$L(D|\rho) = p_1^{n_1} p_2^{n_2} \cdots p_K^{n_K}, \tag{8}$$

which attains its maximum value when ρ is the MLE $\widehat{\rho}_{\mathrm{ML}} \in \mathcal{R}_0$; the MLE is fully determined by the relative frequencies n_k/N. The joint probability of finding the state ρ in the region $\mathcal{R}$ and obtaining the data D is

$$\mathrm{prob}(D \wedge \mathcal{R}) = \int_{\mathcal{R}} (\mathrm{d}\rho)\, L(D|\rho)\,. \tag{9}$$

For $\mathcal{R} = \mathcal{R}_0$, $\mathrm{prob}(D \wedge \mathcal{R}_0) = L(D)$ is the *prior likelihood*. Since one of the click sequences is surely observed, we have

$$\sum_D L(D|\rho) = 1, \quad \sum_D L(D) = \int_{\mathcal{R}_0} (\mathrm{d}\rho) = 1, \tag{10}$$

where the sum is taken over all possible data for N clicks.

[9] One can account for detector inefficiencies and dark counts, but such technical details, which are important for practical applications, are immaterial to the current discussion.

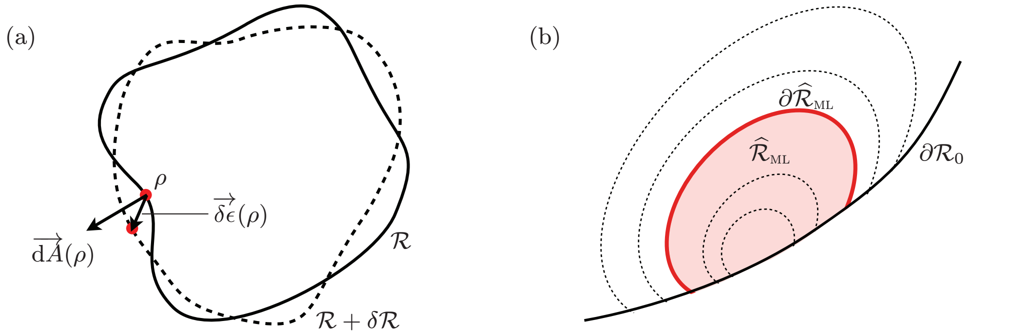
(a)
ρ
$\overrightarrow{\delta\epsilon}(\rho)$
$\overrightarrow{\mathrm{d}A}(\rho)$
$\mathcal{R}$
$\mathcal{R}+\delta\mathcal{R}$
(b)
$\partial\widehat{\mathcal{R}}_{\mathrm{ML}}$
$\widehat{\mathcal{R}}_{\mathrm{ML}}$
$\partial\mathcal{R}_0$

Figure 1. (a) Infinitesimal variation of region $\mathcal{R}$. The boundary $\partial\mathcal{R}$ of region $\mathcal{R}$ (solid line) is deformed to the boundary of region $\mathcal{R}+\delta\mathcal{R}$ (dashed line). $\overrightarrow{\mathrm{d}A}(\rho)$ is the vectorial surface element of $\partial\mathcal{R}$ at ρ, and $\overrightarrow{\delta\epsilon}(\rho)$ is the infinitesimal displacement of ρ. (b) Dotted lines indicate ILSs. The boundary $\partial\widehat{\mathcal{R}}_{\mathrm{ML}}$ of $\widehat{\mathcal{R}}_{\mathrm{ML}}$ can contain part of the surface $\partial\mathcal{R}_0$ of the reconstruction space.

We factor the joint probability in two ways

$$\mathrm{prob}(D\wedge\mathcal{R}) = L(D|\mathcal{R})S_{\mathcal{R}} = C_{\mathcal{R}}(D)L(D) \tag{11}$$

and so identify the *region likelihood* $L(D|\mathcal{R})$ and the *credibility* $C_{\mathcal{R}}(D)$. These are conditional probabilities: $L(D|\mathcal{R})$ is the probability of obtaining the data D if the actual state is in the region $\mathcal{R}$; $C_{\mathcal{R}}(D)$ is the probability that the actual state is in $\mathcal{R}$ if D were obtained—the posterior probability of $\mathcal{R}$.

3. Optimal error regions

For the given data D, we desire a region with the largest likelihood—the MLR. For this purpose, we maximize the region likelihood $L(D|\mathcal{R})$ under the constraint that only regions with a prechosen size s participate in the competition, with $0 < s < 1$,

$$\max_{\mathcal{R}\subseteq\mathcal{R}_0} L(D|\mathcal{R}) = L(D|\widehat{\mathcal{R}}_{\mathrm{ML}}) \quad \text{with } S_{\mathcal{R}} = s; \tag{12}$$

an unconstrained maximization of $L(D|\mathcal{R})$ is not meaningful as it gives the limiting region comprising nothing but the point $\widehat{\rho}_{\mathrm{ML}}$. The resulting MLR $\widehat{\mathcal{R}}_{\mathrm{ML}}$ depends on D and s. Since the size is fixed, we can equivalently maximize the joint probability $\mathrm{prob}(D\wedge\mathcal{R})$ under the size constraint.

The answer to this maximization problem is given in [10, corollary 4], which we translate into our present context as follows:

> The MLRs of various sizes s consist of all states ρ for which the point likelihood exceeds a threshold value, with higher thresholds for smaller sizes. (13)

This is justified by a proof of considerable mathematical sophistication in [10]. We offer an alternative argument that is more accessible to the working physicist. Consider infinitesimal variations of a region $\mathcal{R}$ by deforming its boundary. The maximum property of the MLR and its

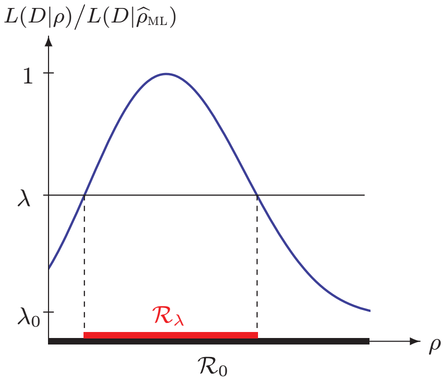

Figure 2. Illustration of a BLR: $\mathcal{R}_0$ is the reconstruction space; the region $\mathcal{R}_\lambda$ is a BLR, delineated by the threshold value $\lambda L(D|\widehat{\rho}_{\mathrm{ML}})$; λ_0 marks the minimum ratio $L(D|\rho)/L(D|\widehat{\rho}_{\mathrm{ML}})$ over $\mathcal{R}_0$.

fixed size require prob$(D \wedge \mathcal{R})$ and $S_\mathcal{R}$ to be stationary under all variations about $\mathcal{R} = \widehat{\mathcal{R}}_{\mathrm{ML}}$ (see figure 1(a)),

$$\begin{aligned} \delta S_\mathcal{R} &= \int_{\partial\mathcal{R}} \overrightarrow{\mathrm{d}A}(\rho) \cdot \overrightarrow{\delta\epsilon}(\rho) = 0, \\ \delta \mathrm{prob}(D \wedge \mathcal{R}) &= \int_{\partial\mathcal{R}} \overrightarrow{\mathrm{d}A}(\rho) \cdot \overrightarrow{\delta\epsilon}(\rho)\, L(D|\rho) = 0. \end{aligned} \tag{14}$$

Both hold only if $L(D|\rho)$ is constant on the boundary $\partial\widehat{\mathcal{R}}_{\mathrm{ML}}$ of $\widehat{\mathcal{R}}_{\mathrm{ML}}$, for an $\widehat{\mathcal{R}}_{\mathrm{ML}}$ entirely contained inside $\mathcal{R}_0$ (so that $\overrightarrow{\delta\epsilon}(\rho)$ can have any direction): $\partial\widehat{\mathcal{R}}_{\mathrm{ML}}$ is an *iso-likelihood surface* (ILS). Concavity of the log-likelihood further requires $\widehat{\mathcal{R}}_{\mathrm{ML}}$ to be the interior of this ILS. $\partial\widehat{\mathcal{R}}_{\mathrm{ML}}$ can also contain part of the boundary of $\mathcal{R}_0$ (see figure 1(b)), in which case only the part of $\partial\widehat{\mathcal{R}}_{\mathrm{ML}}$ inside $\mathcal{R}_0$ is an ILS. In either case, $\widehat{\mathcal{R}}_{\mathrm{ML}}$ is a *bounded-likelihood region* (BLR), comprising all states in $\mathcal{R}_0$ with point likelihood exceeding a certain threshold. BLRs have appeared previously in standard statistical analysis (see [18] and references therein).

We specify the threshold value as a fraction λ of the maximum value $L(D|\widehat{\rho}_{\mathrm{ML}})$ of the point likelihood; see figure 2. The BLR $\mathcal{R}_\lambda$ has the characteristic function

$$\chi_\lambda(\rho) = \eta\Big(L(D|\rho) - \lambda L(D|\widehat{\rho}_{\mathrm{ML}})\Big), \tag{15}$$

and

$$s_\lambda = \int_{\mathcal{R}_0} (\mathrm{d}\rho)\, \chi_\lambda(\rho) \tag{16}$$

is the size of $\mathcal{R}_\lambda$. We have $\mathcal{R}_\lambda = \mathcal{R}_0$ and $s_\lambda = s_0 = 1$ for $\lambda \leqslant \lambda_0$ with $\lambda_0 \geqslant 0$ given by $\min_\rho L(D|\rho) = \lambda_0\, L(D|\widehat{\rho}_{\mathrm{ML}})$. As λ increases from λ_0 to 1, s_λ decreases monotonically from 1 to 0. The size s specified in (12) is obtained for an intermediate λ value and the corresponding BLR is the looked-for MLR. Note that the MLE is contained in all MLRs: as $s \to 0$, the MLR becomes an infinitesimal vicinity of the MLE, and $L(D|\widehat{\mathcal{R}}_{\mathrm{ML}}) \to L(D|\widehat{\rho}_{\mathrm{ML}})$.

The MLR is the region for which the observed data are particularly likely. With a reversal of emphasis, we now look for a region that contains the actual state with high probability. Ultimately, this is the SCR $\widehat{\mathcal{R}}_{\mathrm{SC}}$—the smallest region with the pre-chosen credibility value c. For the given D, the optimization problem

$$\min_{\mathcal{R}\subseteq\mathcal{R}_0} S_\mathcal{R} = S_{\widehat{\mathcal{R}}_{\mathrm{SC}}} \quad \text{with } C_\mathcal{R}(D) = c \tag{17}$$

is dual to that of (12). Here, we minimize the size for given joint probability; there we maximize the joint probability for a given size. It follows that the BLRs are not only the MLRs but they are also the SCRs: each MLR is a SCR, each SCR is an MLR.

The BLR $\mathcal{R}_\lambda$ has credibility

$$c_\lambda = \frac{1}{L(D)} \int_{\mathcal{R}_0} (\mathrm{d}\rho)\, \chi_\lambda(\rho) L(D|\rho), \tag{18}$$

which, like s_λ, decreases monotonically from 1 to 0 as λ increases from λ_0 to 1. The credibility c specified in (17) is obtained for an intermediate value, and the corresponding BLR is the looked-for SCR.

That the general definitions of the MLR and the SCR, which allow for regions of arbitrary shapes, permit such a simple characterization in terms of BLRs is remarkable. BLRs are reminiscent of standard ellipsoidal error regions constructed by analyzing the neighborhood of the peak of the likelihood function—a procedure justified only for large enough N for the central limit theorem to apply (see, for instance, [4]); yet, our result employs no such assumption. Also surprising is that, while λ depends on the choice of prior, the set of regions that enter the competition is independent of that choice; the prior enters only in the size, region likelihood, and credibility of the MLR/SCR.

Once the data are obtained, there is *the* MLR and *the* SCR for these data, and other MLRs or SCRs associated with unobserved data play no role. This is in sharp contrast to confidence regions, whose construction requires consideration of all data that could have been obtained, since the confidence level is a property of the entire set of confidence regions, one for each possible data (see appendix B). Nevertheless, they are not unrelated. Christandl and Renner [7] showed that high-credibility regions offer starting points for constructing confidence regions—a set of SCRs with high credibility immediately suggests itself—and Blume–Kohout [8] argued that BLRs can be good confidence regions.

4. Reporting error regions

For a BLR, s_λ and c_λ are linked by

$$L(D)\frac{\partial}{\partial\lambda}c_\lambda = L(D|\widehat{\rho}_{\mathrm{ML}})\,\lambda\frac{\partial}{\partial\lambda}s_\lambda. \tag{19}$$

Therefore, once s_λ is known as a function of λ, we obtain c_λ from

$$c_\lambda = \frac{\lambda s_\lambda + \int_\lambda^1 \mathrm{d}\lambda'\, s_{\lambda'}}{\int_0^1 \mathrm{d}\lambda'\, s_{\lambda'}}. \tag{20}$$

This is, of course, consistent with the limiting values for $\lambda \leqslant \lambda_0$ and $\lambda = 1$, and also establishes that $c_\lambda > s_\lambda$ for $\lambda_0 < \lambda < 1$. Furthermore, (19) and (20) tell us that in the $\lambda \to 1$ limit, when both s_λ and c_λ vanish, their ratio is finite and exceeds unity,

$$\frac{c_\lambda}{s_\lambda} \to \frac{1}{\int_0^1 \mathrm{d}\lambda'\, s_{\lambda'}} \to \frac{L(D|\widehat{\rho}_{\mathrm{ML}})}{L(D)} > 1 \quad \text{for } \lambda \to 1. \tag{21}$$

We note that this provides the value of $L(D)$, since the maximal value $L(D|\widehat{\rho}_{\mathrm{ML}})$ of the point likelihood is computed earlier as it is needed when identifying the BLRs.

Inasmuch as the value of s_λ quantifies our prior belief that the actual state is in $\mathcal{R}_\lambda$, we are surprised when the data tell us that the probability for finding the state in $\mathcal{R}_\lambda$ is larger.

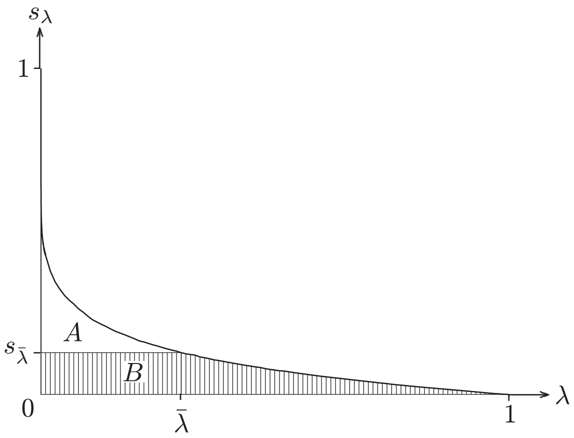

Figure 3. Geometrical meaning of the relation (20) between the size s_λ and the credibility c_λ. For the chosen value of λ, say $\bar\lambda$, the horizontal line from $(0, s_{\bar\lambda})$ to $(\bar\lambda, s_{\bar\lambda})$ divides the area under the graph of s_λ into the two pieces A and B indicated in the plot. The credibility is the fractional size of area B, that is: $c_{\bar\lambda} = B/(A+B)$.

Accordingly, the SCR is the region for which we are most surprised for the given prior belief. This matter and other aspects of Bayesian inference based on the concept of relative surprise are discussed in [10].

Relation (20) has a simple geometrical meaning in terms of areas under the graph of s_λ, as explained in figure 3. This relation is of considerable practical importance because we only need to evaluate the integrals of (16), but not those of (18). Since the latter integrals require well-tailored Monte Carlo methods to handle the typically sharply peaked point likelihood, the numerical effort is substantially reduced if we only need to evaluate (16). The error regions for the observed data are then concisely communicated by reporting s_λ and c_λ as functions of λ. With these, the end user interested in the MLR with the size s of his liking or the SCR of his wanted credibility c can determine the required value of λ. It is then easy to check if a state is inside the specified error region. The example of section 6 illustrates the matter for an eight-dimensional reconstruction space, for which the error regions will be impossible to visualize, but can still be easily specified through reporting the s_λ and c_λ values.

5. Example: incomplete single-qubit tomography

For a first illustration, we consider the simplest situation that exhibits the typical features: the quantum-information carriers have a qubit degree of freedom, which is measured by one of two standard POMs that are not informationally complete.

5.1. Probability-operator measurements and priors

For both POMs, the unit disk in the xy plane suggests itself for the reconstruction space $\mathcal{R}_0$. The first POM is the crosshair measurement that combines projective measurements of σ_x and

σ_y into a four-outcome POM ($K = 4$) with probabilities

$$\left.\begin{matrix} p_1 \\ p_2 \end{matrix}\right\} = \frac{1}{4}(1 \pm x), \quad \left.\begin{matrix} p_3 \\ p_4 \end{matrix}\right\} = \frac{1}{4}(1 \pm y) \quad \text{with } x = \langle\sigma_x\rangle, \;\; y = \langle\sigma_y\rangle\,. \tag{22}$$

The permissible probabilities are identified by

$$w_{\text{cstr}}(p) \doteq \eta(p)\,\delta(p_1 + p_2 - \tfrac{1}{2})\,\delta(p_3 + p_4 - \tfrac{1}{2})\,\eta(3 - 8p^2), \tag{23}$$

where

$$\eta(p) = \prod_{k=1}^{K} \eta(p_k) \quad \text{and} \quad p^2 = \sum_{k=1}^{K} p_k^2\,. \tag{24}$$

The dotted equals sign in (23) stands for ‘equal up to a multiplicative constant’, namely the factor that ensures the unit size of the reconstruction space.

The second POM is the three-outcome trine measurement ($K = 3$), whose outcomes are subnormalized projectors on the eigenstates of σ_x and $(-\sigma_x \pm \sqrt{3}\,\sigma_y)/2$ with eigenvalue +1. It has the probabilities

$$p_1 = \frac{1}{3}(1 + x), \quad \left.\begin{matrix} p_2 \\ p_3 \end{matrix}\right\} = \frac{1}{6}(2 - x \pm \sqrt{3}\,y), \tag{25}$$

for which

$$w_{\text{cstr}}(p) \doteq \eta(p)\,\delta(p_1 + p_2 + p_3 - 1)\,\eta(1 - 2p^2) \tag{26}$$

summarizes the constraints that the permissible values of p_1, p_2, p_3 obey.

Both POMs have the same primitive prior

$$(\mathrm{d}\rho) = \mathrm{d}x\,\mathrm{d}y\,\frac{1}{\pi}\,\eta(1 - x^2 - y^2) = \mathrm{d}r^2\,\frac{\mathrm{d}\varphi}{2\pi}, \tag{27}$$

where $0 \leqslant r \leqslant 1$ and φ covers any convenient range of 2π. This prior is uniform in x and y, and in r^2 and φ. The polar-coordinate version, that is $x + \mathrm{i}y = r\,\mathrm{e}^{\mathrm{i}\varphi}$, is the more natural parameterization of the unit disk. The Jeffreys prior for the four-outcome POM is

$$(\mathrm{d}\rho) = \frac{2}{\pi^2}\,\frac{\mathrm{d}r\,r\,\mathrm{d}\varphi}{\sqrt{1 - r^2 + \frac{1}{4}r^4\sin(2\varphi)^2}}\,. \tag{28}$$

For the three-outcome POM, we have the Jeffreys prior

$$(\mathrm{d}\rho) = \frac{1}{4\pi - 24\sin^{-1}\left(\frac{1}{3}\right)}\,\frac{\mathrm{d}r\,r\,\mathrm{d}\varphi}{\sqrt{1 - \frac{3}{4}r^2 + \frac{1}{4}r^3\cos(3\varphi)}}\,. \tag{29}$$

5.2. Computer-generated data

Figures 4(a) and (b) show SCRs obtained for simulated experiments in which $N = 24$ copies of a qubit state are measured. The actual state used for the simulation has $(x, y) = (0.6, 0.2)$. Its position in the reconstruction space is indicated by the red star (★).

In figure 4(a), we see the SCRs for the four-outcome POM. Two experiments were simulated, with $(n_1, n_2, n_3, n_4) = (8, 5, 10, 1)$ and $(6, 3, 10, 5)$ clicks of the detectors, respectively, and the triangles (△) show the positions of the corresponding MLEs. For each data, the plot reports the SCRs with credibility $c = 0.5$ and 0.9, both for the primitive prior of (27)

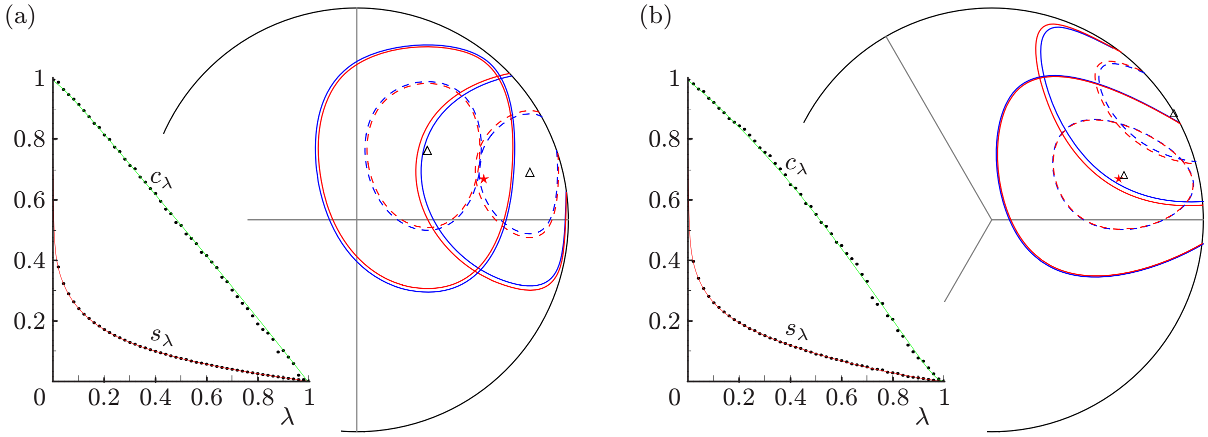

Figure 4. Smallest credible regions for simulated experiments. Twenty-four copies are measured by the POMs of section 5.1, which have the unit disk of figure A.1 as the reconstruction space. Plot (a) is for the four-outcome POM with the crosshair indicating the orientations of the two projective measurements. Plot (b) is for the three-outcome measurement with the orientation of the trine indicated. The red star (⋆) at $(x, y) = (0.6, 0.2)$ marks the actual state that was used for the simulation. For each POM, there are SCRs for the data of two simulated experiments, with black triangles (△) indicating the respective MLEs. The boundaries of the SCRs with credibility $c = 0.9$ are traced by the continuous lines; all of these SCRs contain the actual state. The dashed lines are the boundaries of the SCRs with credibility $c = 0.5$; the actual state is inside half of these SCRs. The red lines are for the primitive prior of (27), the blue lines are for the Jeffreys priors of (28) and (29), respectively.—the insets in the lower left corners show the size s_λ and the credibility c_λ for the BLRs of two different sets of data. Inset (a) is for (6, 3, 10, 5) counts for the four-outcome POM and the Jeffreys prior. Inset (b) is for (13, 7, 4) counts for the three-outcome POM and the primitive prior. The dots show the values computed with a Monte Carlo algorithm. There is much more scatter in the c_λ values than the s_λ values. The red lines are fits to the s_λ values, with the fits using twice as many values as there are dots in the insets. The green lines that approximate the c_λ values are obtained from the red lines with the aid of (20).

and for the Jeffreys prior of (28). The actual state is inside two of the four SCRs with credibility $c = 0.5$ and is contained in all four SCRs with credibility $c = 0.9$.

Not unexpectedly, we get quite different regions for the two rather different sets of detector click counts. Yet, we observe that the choice of prior has little effect on the SCRs, although the total number of measured copies is too small for relying on the consistency of the priors. The same remarks apply to the SCRs for the three-outcome POM in figure 4(b); here we counted $(n_1, n_2, n_3) = (15, 8, 1)$ and $(13, 7, 4)$ detector clicks in the simulated experiments.

In section 4 we remarked that the estimator regions are properly communicated by reporting s_λ and c_λ as functions of λ. This is accomplished by the insets in figure 4 for two of the four simulated experiments. The dots give the values obtained by numerical integration that uses a Monte Carlo algorithm. The scatter of these numerical values confirms the expected: the computation of s_λ only requires sampling the probability space in accordance with the prior and determining the fraction of the sample that is in $\mathcal{R}_\lambda$; for the computation of c_λ we need to add the values of $L(D|\rho)$ for the sample points inside $\mathcal{R}_\lambda$; and since $L(D|\rho)$ is a sharply peaked function of the probabilities, the s_λ values are more trustworthy than the c_λ values for the same computational effort. The line fitted to the s_λ values is a Padé approximant (see e.g. [19, section 5.12]) that takes the analytic forms near $\lambda = \lambda_0 = 0$ and $\lambda = 1$ into account. The line approximating the c_λ values is then computed in accordance with (20).

6. Example: incomplete two-qubit tomography

For a second illustration, we consider the situations that arise in the quantum-key-distribution schemes by Bennett and Brassard (BB84 [20]) and the trine–antitrine (TAT) scheme of reference [21]. Both schemes can be implemented by having a source of entangled qubit pairs distribute one qubit each to the two communicating parties. Prior to any key generation, the two-qubit state emitted by the source needs to be characterized. It is desirable to achieve quantum state estimation with reliable error regions without sacrificing many data that are then not available for the key generation.

6.1. Probability-operator measurements and computer-generated data

In the BB84 scheme, each qubit is measured by the crosshair POM of (22); the resulting two-qubit POM has sixteen outcomes that obey eight constraints that give delta-function factors in $w_{\rm cstr}(p)$. In the TAT scheme, one qubit is measured by the trine POM of (25) and the other qubit by the antitrine POM that has the signs of x and y reversed in (25); the resulting two-qubit POM has nine outcomes subject to the single delta-function constraint of unit sum. Accordingly, the probability space is eight-dimensional for both schemes[10], and we cannot report the SCRs by showing the optimal error regions in the reconstruction space, as was possible for the two-dimensional probability space in figure 4. Therefore, we employ the strategy of section 4 and report the size s_λ and the credibility c_λ of the respective BLRs as functions of λ.

For the generation of the simulated data, we first add noise to the singlet state by putting it through a random Pauli channel[11]. The resulting true state has the probabilities for the two-qubit POMs given in the top row of table 1. For example, the '12' entry in the 4×4 table for the double-crosshair POM is the probability for outcome $\Pi_1 \otimes \Pi_2 = \frac{1}{4}(1+\sigma_x) \otimes \frac{1}{4}(1-\sigma_x)$. The '11' entry of the 3×3 table for the TAT POM is 16/9 times that number. Note that all marginal probabilities (sums of rows and sums of columns) are equal; this is so because the

[10] In actual experiments, the probability space is nine-dimensional because one must account for the no-click probability of the qubit pairs that do not give rise to coincidence clicks. Furthermore, the state estimation could also exploit the data collected for single-qubit detection without the coincidental detection of the partner qubit. Consistent with the footnote in section 2.3, we are here content with the idealized situation of perfect detection devices because our objective is to give an example for a higher-dimensional space rather than evaluating real experimental data.

[11] A random Pauli channel is used as a simple model for noise in a communication protocol. The channel acts on an input state as $\rho \to \sum_{jk} r_{jk}\, \sigma_j \otimes \sigma_k\, \rho\, \sigma_j \otimes \sigma_k$, where $j, k = 0, x, y, z$ (σ_0 denotes the single-qubit identity operator), and the r_{jk} are sixteen randomly chosen probabilities. The 60 copies of the true state come from passing 60 copies of the singlet state through one instance of the random Pauli channel, i.e., the r_{jk} are randomly picked once, with r_{00} given a higher weight of 0.7 to simulate weak noise.

Table 1. Computer-generated data for the estimation of a two-qubit state from measuring 60 identically prepared copies. The first row gives the joint probabilities of the true state. The broken second row shows the number of detector-click pairs obtained in the simulated experiment (and their expected values) together with the single-qubit marginals. The third row reports the joint probabilities of the MLEs for the data in the second row. In each row, we have a 4×4 table on the left for the double-crosshair POM of the BB84 scenario and a 3×3 table on the right for the nine-outcome POM of the TAT scheme. The rows of a 4×4 table for the double-crosshair POM refer to the four Π_j of the first qubit in the pair and the columns refer to the Π_k of the second qubit; entry 'jk' is the probability for outcome $\Pi_j \otimes \Pi_k$. Analogously, entry 'jk' in a 3×3 table for the TAT scheme is the expectation value of $\Pi_j \otimes \Pi_k$ with trine outcome Π_j and antitrine outcome Π_k.

		Double-crosshair POM				
True-state probabilities		1	2	3	4	
	1	0.0206	0.1044	0.0625	0.0625	
	2	0.1044	0.0206	0.0625	0.0625	
	3	0.0625	0.0625	0.0212	0.1038	
	4	0.0625	0.0625	0.1038	0.0212	
Computer-generated data (expected number of clicks)		1	2	3	4	
	1	0 (1.24)	4 (6.26)	6 (3.75)	4 (3.75)	14 (15)
	2	6 (6.26)	3 (1.24)	8 (3.75)	4 (3.75)	23 (15)
	3	3 (3.75)	1 (3.75)	0 (1.27)	8 (6.23)	12 (15)
	4	1 (3.75)	7 (3.75)	4 (6.23)	1 (1.27)	13 (15)
		10 (15)	15 (15)	18 (15)	17 (15)	
MLE probabilities		1	2	3	4	
	1	0.0056	0.1012	0.0497	0.0571	
	2	0.0939	0.0493	0.0821	0.0611	
	3	0.0630	0.0344	0.0025	0.0949	
	4	0.0365	0.1160	0.1293	0.0232	

		TAT POM			
True-state probabilities		1	2	3	
	1	0.1856	0.0739	0.0739	
	2	0.0739	0.1848	0.0747	
	3	0.0739	0.0747	0.1848	
Computer-generated data (expected number of clicks)		1	2	3	
	1	11 (11.14)	4 (4.43)	5 (4.43)	20 (20)
	2	2 (4.43)	10 (11.09)	5 (4.48)	17 (20)
	3	4 (4.43)	6 (4.48)	13 (11.09)	23 (20)
		17 (20)	20 (20)	23 (20)	
MLE probabilities		1	2	3	
	1	0.1833	0.0667	0.0833	
	2	0.0333	0.1667	0.0833	
	3	0.0667	0.1000	0.2167	

reduced single-qubit states of the true state are completely mixed. For the same reason, both tables are symmetric and the lower-left and upper-right 2×2 subtables of the 4×4 table have entries of $1/16 = 0.0625$. More generally, there is a one-to-one correspondence between the 16 permissible probabilities in the 4×4 table and the nine permissible probabilities in the 3×3 table, because all table entries are determined by the expectation values of $A \otimes B$ with $A, B = 1, \sigma_x, \sigma_y$.

Simulated measurements of 60 qubit pairs in the true state for each POM produced the counts of detector-click pairs in the second row of table 1; expected values are given in parentheses. Owing to the statistical fluctuations, the tables of counts are not symmetric[12] and the marginal counts are not equal.

The third row of table 1 shows the corresponding MLE probabilities. These probabilities are equal to the relative frequencies of the counts for the 9-outcome POM but are different from the relative frequencies for the 16-outcome POM. This tells us that the computer-generated data are not typical for the double-crosshair POM, whereas we have typical data for the TAT POM.

6.2. Size and credibility of the bounded-likelihood regions

As noted in section 4, the primary task of the data evaluation is the computation of the multi-dimensional integrals that give the size s_λ of the BLRs for the whole range of $0 < \lambda < 1$. For the data in table 1, these are integrals over eight-dimensional regions. We used a random-sampling technique for this purpose.

As a preparation, we generated a random sample of 648 785 permissible sets of probabilities, uniformly distributed in accordance with the primitive prior (see section 6.3 below). In view of the one-to-one correspondence between the permissible probabilities of the 16-outcome POM and the 9-outcome POM, the same random sample can be, and was, used for both POMs.

The actual data processing consists of two steps. In the first step, we determine the size s_λ for the 160 values of λ with $-\log_{10}\lambda = 0.1(0.1)16.0$. This requires a simple counting of how many samples are inside the BLR $\mathcal{R}_\lambda$ if the primitive prior is used. In the case of the Jeffreys prior, one adds the weights $(p_1 p_2 \cdots)^{-1/2}$ of the samples inside the BLR. The correct normalization follows from $s_{\lambda=0} = 1$.

In the second step, the integrals needed in (20) are evaluated, for which a simple linear interpolation between adjacent (λ, s_λ) pairs is sufficiently accurate. Then, c_λ is known as a function of λ and the λ values for which we have 99% or 95% credibility are determined.

We show s_λ and c_λ as functions of λ in figure 5. Table 2 reports the λ values of the 99% and 95% credibility thresholds. We observe that for the 16-outcome POM, the true state is inside the SCRs with 99% credibility for both the primitive prior and the Jeffreys prior, whereas it is inside the 95% SCR only for the primitive prior but not for the Jeffreys prior. This is more evidence that these data are untypical. By contrast, for the 9-outcome POM, the true state is inside all SCRs for both priors and both values of the credibility.

Typicality, or lack thereof, can also be noticed in figure 5. Since the Jeffreys prior gives more weight to the regions near the boundary of the probability space than the primitive prior, and less weight to regions deep inside, one expects that the values of s_λ for the primitive prior are larger than those for the Jeffreys prior if the data are typical and, accordingly, the MLE is not close to the boundary. This is indeed the case for the TAT data, but not for the double-crosshair data.

[12] One usually restores the symmetry by so-called ‘twirling’ before the key generation protocol is executed. The characterization of the source, however, should be done without the twirling.

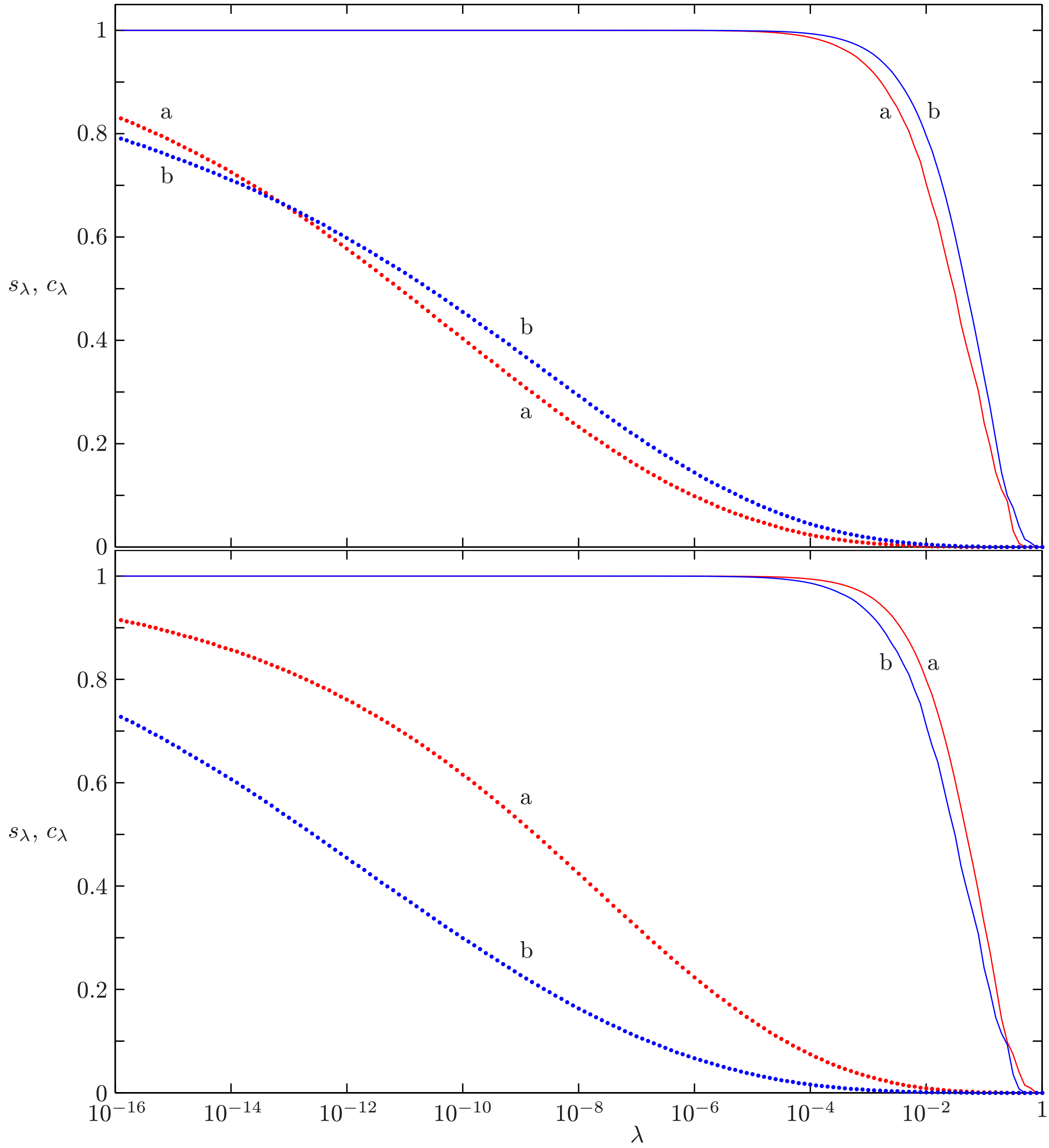

Figure 5. The size s_λ (dotted lines) and the credibility c_λ (solid lines) as functions of λ for the data of table 1. The top plot is for the double-crosshair POM, the bottom plot is for the TAT POM. Curves 'a' are for the primitive prior, curves 'b' are for the Jeffreys prior. The abscissa is linear in $\log\lambda$. For $\lambda \lesssim 1$, the BLRs are so small that only very few sample points are inside and the sizes s_λ have comparatively large fluctuation errors. This statistical noise is visible in the bottom-right corners of the plots. It has no bearing, however, on the accuracy of the credibility c_λ in the important range of smaller λ values, as one notes upon recalling figure 3.

Table 2. Threshold λ values for 99% and 95% credibility for the data of table 1 and figure 5, and the sizes of the respective BLRs. The true state is inside the $\mathcal{R}_\lambda$s with $\lambda < 3.368 \times 10^{-3}$ for the 16-outcome POM (with its untypical data), and inside the BLRs with $\lambda < 0.2486$ for the 9-outcome POM.

	16-outcome POM			9-outcome POM		
	λ	s_λ	c_λ	λ	s_λ	c_λ
Primitive prior	6.70×10^{-5}	0.0279	0.99	1.92×10^{-4}	0.0601	0.99
	6.03×10^{-4}	0.0106	0.95	1.44×10^{-3}	0.0268	0.95
	λ	s_λ	c_λ	λ	s_λ	c_λ
Jeffreys prior	1.73×10^{-4}	0.0374	0.99	6.74×10^{-5}	0.0186	0.99
	1.35×10^{-3}	0.0161	0.95	6.20×10^{-4}	0.0070	0.95

6.3. Numerical effort

The two steps of data evaluation, the computation of the size s_λ and then the credibility c_λ, take a few seconds of central processing unit (CPU) time. The preparation of the random sample of permissible probabilities, which could be done ahead of the data taking, lasts much longer. Our sample of 648 785 probabilities took almost 100 h of CPU time on a standard desktop (Intel i7-870 CPU, using one of the four cores and 8 GB RAM). The procedure we employed was simple and reliable but not optimized for speed. There is clearly much room for improvement (see section 7).

For each potential sample of probabilities we first generate nine random numbers $x_1, x_2, \ldots, x_9$ uniformly and independently between 0 and 1. Then, the nine probabilities $p_k = \log(x_k)/\log(x_1x_2\cdots x_9)$ constitute a sample in the eight-dimensional simplex of the classical nine-sided die, and the samples are distributed in accordance with the primitive prior. The sample $p = (p_1, p_2, \ldots, p_9)$ is accepted if it is a permissible set of probabilities for the TAT POM with its nine outcomes. Only 9.27% of the 7×10^6 candidate probabilities generated were accepted.

Whereas the generation of another sample p is fast, the test of permissibility is the part that consumes most of the CPU time. After identifying the candidate p with the relative frequencies of a measurement with the 9-outcome POM, we calculate a MLE for these frequencies. If the probabilities of the MLE are equal to p, this sample probability is accepted, otherwise it is rejected.

7. Outlook

For the given data and chosen size or credibility, the MLR or the SCR is a neighborhood of the MLE. In this sense, one can regard them as systematically constructed error regions for the MLE. Although there are efficient methods for computing the MLE [2, 3], equally efficient algorithms for finding the MLR and the SCR need to be developed. In particular, there remains the challenge of evaluating the multi-dimensional integrals that give s_λ.

Random-sampling techniques are our methods of choice. In section 6, a simple strategy required 100 h of CPU time for creating a sample of permissible probabilities. However, this can be substantially reduced by optimizing the sampling process. Parallelizing the sampling over many different computers and, later, combining into a single dataset was suggested. The chance that a candidate probability is permissible (only 9.27% in the sampling of the example) can be much increased by employing cleverer Monte Carlo methods, where one makes use of information at the current physical point to stay within the physical state space. We have just begun to explore this and will report on our progress in due course. It is also worth noting that this computational time is an overhead that is incurred only once and the sampling can be done ahead of any actual data-taking in the laboratory.

Often, only a few parameters computed from the state are of interest. It is, therefore, possible to reduce the dimensionality of the problem by discarding nuisance parameters. A variant of the methodology described here can be used to determine small regions of high credibility in the few-parameter space of interest, without first determining SCRs in the reconstruction space. We will return to this matter on another occasion.

Acknowledgments

We have benefited greatly from our discussions with David Nott and thank him in particular for bringing [10, 16] to our attention. The Centre for Quantum Technologies is a Research Centre of Excellence funded by the Ministry of Education and the National Research Foundation of Singapore.

Note added in proof. Since the completion of this paper, we have improved the sampling algorithm and thereby reduced the CPU time consumed to less than 10% of what it was before. This speed-up results from a faster test of permissibility that is still easy to implement. Further improvement is likely.

Appendix A. Choosing the prior

The assignment of prior probabilities to regions in the reconstruction space should be done in an unprejudiced manner while taking into account all prior information that might be available. We cannot do justice to the rich literature on this subject and are content with noting that [16] reviews various approaches to constructing unprejudiced priors. Here, we discuss some criteria that are useful when choosing a prior, illustrating with examples familiar in quantum contexts.

A general remark: the chosen prior should give some weight to (almost) all states. It should not give extremely high weight to states in some part of the state space and extremely low weight to other states. This is to say, the prior should be *consistent* in the sense that the credibility of a region—its posterior content—is dominated by the data rather than by the prior if a reasonably large number N of copies is measured. In the examples of figure 4, $N = 24$ is close to being 'reasonably large', while $N = 2$ in figure B.1 is clearly not. Also, $N = 60$ in section 6 is not large enough to ensure data dominance because the λ values in table 2 for the primitive prior and for the Jeffreys prior are quite different and, hence, correspond to quite different BLRs.

Below, we describe a few criteria for choosing priors. We begin in section A.1 with the common choice of a uniform prior; section A.2 discusses priors motivated by the utility of the estimated state; section A.3 invokes symmetry arguments to restrict considerations to priors that possess some symmetry properties; section A.4 presents form-invariant prior constructions; section A.5 deals with the situation where one has a target state in mind; and section A.6 is about priors induced by marginalization of full-state-space priors according to what the data can tell us.

A.1. Uniformity

The time-honored strategy of choosing a uniform prior on $\mathcal{R}_0$ in which all states are treated equally gets us into a circular argument. Our identification of the size of a region with its prior content amounts to assigning equal probabilities to regions of equal sizes, prior to acquiring any data. But that just means that we now have to declare how we measure the size of a region without prejudice, and we are again faced with the original question about a uniform prior.

In fact, there is no unique meaning of the uniformity of a prior. In the sense that each prior tells us how to quantify the size of a region, each prior is uniform with respect to its induced size measure. To illustrate, reconsider the harmonic-oscillator example of section 2.1. For the primitive prior of (5), the parameterization

$$\begin{aligned} &p_1 = \tfrac{1}{2}(v+u), \quad p_2 = \tfrac{1}{2}(v-u), \\ &\mathrm{d}p_1\,\mathrm{d}p_2 = \mathrm{d}u\,\mathrm{d}v\,\tfrac{1}{2} \end{aligned} \tag{A.1}$$

gives

$$\begin{aligned} (\mathrm{d}\rho) = {} & \mathrm{d}u\,\mathrm{d}v\,\tfrac{1}{2}\eta(v+u)\eta(v-u)\delta(v-1) \\ & \to \mathrm{d}u\,\tfrac{1}{2} \quad \text{with} \quad -1 \leqslant u \leqslant 1, \end{aligned} \tag{A.2}$$

where we integrate over v in the last step and so observe that the primitive prior is uniform in u, that is, the size of the region $u_1 < u < u_2$ is proportional to $u_2 - u_1$. Likewise, the parameterization

$$\begin{aligned} &p_1 = v(\sin\alpha)^2, \quad p_2 = v(\cos\alpha)^2, \\ &\mathrm{d}p_1\,\mathrm{d}p_2 = \mathrm{d}\alpha\,\mathrm{d}v\,v\sin(2\alpha) \end{aligned} \tag{A.3}$$

gives

$$(\mathrm{d}\rho) \to \mathrm{d}\alpha\,\frac{2}{\pi} \quad \text{with} \quad 0 \leqslant \alpha \leqslant \frac{\pi}{2} \tag{A.4}$$

for the Jeffreys prior of (6), which is uniform in α. Other priors can be treated analogously, each of them yielding a uniform prior in an appropriate single parameter.

Visualization of the uniformity for qubit priors can be found in figure A.1. Plot (b) shows uniform tiling of the unit disk by tiles of equal size. Here size is measured by the primitive prior of (27), which is uniform in x and y, and in r^2 and φ (the latter is used for the plot). Plots (c1) and (c2) show uniform tilings of the unit disk for the Jeffreys prior for the four-outcome POM of (28), while plots (d1) and (d2) show those for the three-outcome POM of (29). The crosshair symmetry of the four-outcome POM and the trine symmetry of the three-outcome POM are manifest in their respective uniform tilings.

The parameterizations in (A.1) and (A.3), and the tilings of figure A.1 exhibit in which explicit sense the primitive prior and the Jeffreys prior are uniform. However, the priors are what they are, irrespective of how they are parameterized. They are explicitly uniform in a particular parameterization and implicitly uniform in all others. Uniformity, it follows, cannot serve as a principle that distinguishes one prior from another.

This ubiquity of uniform priors for a continuous set of infinitesimal probabilities is in marked contrast to situations in which prior probabilities are assigned to a finite number of discrete possibilities, such as the 38 pockets of a double-zero roulette wheel. Uniform

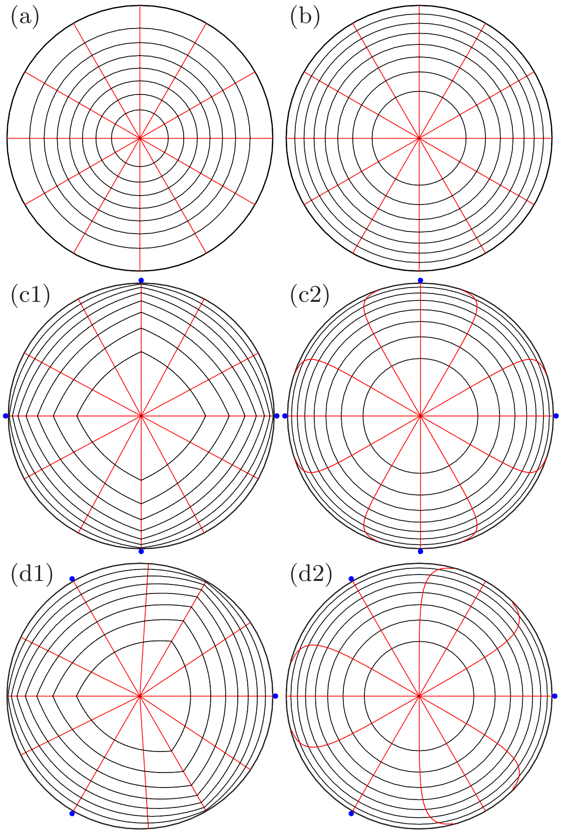

Figure A.1. Uniform tilings of the unit disk for four different priors. The disk is in the xy plane, with the x-axis horizontal, the y-axis vertical, and the disk center at $x = y = 0$. Tiling (a) depicts the marginal prior of (A.14); tiling (b) is for the primitive prior of (27); tilings (c1) and (c2) illustrate the Jeffreys prior of (28) with the blue dots (●) just outside the unit circle indicating the four directions onto which the POM outcomes project; and, tilings (d1) and (d2) are for the Jeffreys prior of (29), the blue dots marking the three directions of the trine projectors. In each tiling, we identify 96 regions of equal size by dividing the disk into eight 'tree rings' of equal size and twelve 'pie slices' of equal size. In the tilings (a), (b), (c1) and (d1), the boundaries of the pie slices are (red) rays and an arc of the unit circle. In the tilings (a), (b), (c2) and (d2), the tree rings have concentric circles as their boundaries.

probabilities of $1/38$ suggest themselves, are meaningful, and clearly distinguished from other priors, all of which have a bias.

Uniformity in a particularly natural parameterization of the probability space might also be meaningful. This, however, invokes a notion of 'natural' that others may not share.

A.2. Utility

In many applications, estimating the state is not a purpose in itself, but only an intermediate step on the way to determining some particular property of the physical system. The objective is to find the value of a parameter that quantifies the *utility* of the state.

For example, one could be interested in the fidelity of the actual state with a target state, or in an entanglement measure of a two-partite state, or in another quantity that tells us how useful are the quantum-information carriers for their intended task. In a situation of this kind, one should, if possible, use a prior that is uniform in the utility parameter of interest. In contrast to the situation of the previous section, where requiring uniformity in $\mathcal{R}_0$ may be ill-advised because uniformity is a parameterization-dependent notion, here we specify uniformity for the parameter we are interested in.

To illustrate, consider a single qubit. Suppose the utility parameter is the purity $\xi(\rho) = \mathrm{tr}\{\rho^2\}$ of the state ρ. With the Bloch-ball representation of a qubit state, $\rho = \frac{1}{2}(1 + \boldsymbol{\varrho}\cdot\boldsymbol{\sigma})$, where $\boldsymbol{\varrho} = \mathrm{tr}\{\boldsymbol{\sigma}\rho\} = \langle\boldsymbol{\sigma}\rangle$ is the Bloch vector and $\boldsymbol{\sigma}$ is the vector of Pauli matrices, the purity is

$$\xi(\rho) = \tfrac{1}{2}(1 + \varrho^2) \quad \text{with } \varrho = |\boldsymbol{\varrho}|\,. \tag{A.5}$$

A prior uniform in purity induces a prior on the state space according to

$$(\mathrm{d}\rho) \propto \mathrm{d}\xi\,\mathrm{d}\Omega \propto \varrho\,\mathrm{d}\varrho\,\mathrm{d}\Omega\,, \tag{A.6}$$

where we parameterize the Bloch ball by spherical coordinates (ϱ, θ, ϕ). Here, $\mathrm{d}\Omega$ is the prior for the angular coordinates; the prior for the radial coordinate ϱ is fixed by our choice of uniformity in ξ. Irrespective of what we choose for $\mathrm{d}\Omega$, the marginal prior for ϱ is uniform in ξ.

If one can quantify the utility of an estimator by a cost function, an optimal prior can be selected by a minimax strategy. For each prior in the competition one determines the maximum of the cost function over the states in the reconstruction space, and then chooses the prior for which the maximum cost is minimal. In classical statistics, such minimax strategies are common (see, for instance, [22, chapter 5]); for an example in the context of quantum state estimation, see [23].

A.3. Symmetry

Symmetry considerations are often helpful in narrowing the search for the appropriate prior. For a particularly instructive example, see section 12.4.4 in Jaynes's posthumous book [24].

Returning to the uniform-in-purity prior of (A.6), one can invoke rotational symmetry in favor of the usual solid-angle element, $\mathrm{d}\Omega = \sin\theta\mathrm{d}\theta\,\mathrm{d}\phi$, as the choice of angular prior. The reasoning is as follows: the purity of a qubit state does not change under unitary transformations; unitarily equivalent states have the same purity. Now, regions that are turned into each other by a unitary transformation have identical radial content whereas the angular dependences are related by a rotation. Invariance under rotations, in turn, requires that the prior is proportional to the solid angle, hence the identification of $\mathrm{d}\Omega$ with the differential of the solid angle. Note that the resulting prior element $(\mathrm{d}\rho)$ is different from the usual Euclidean volume element, $\varrho^2\mathrm{d}\varrho\,\sin\theta\mathrm{d}\theta\,\mathrm{d}\phi$, which would be natural if the Bloch ball were an object in the physical three-dimensional space, but it is not.

Symmetry arguments should be used carefully and not blindly. For a fairly tossed coin, the prior should not be affected if the probabilities for heads and tails are interchanged, $w(p_1, p_2) = w(p_2, p_1)$. However, for the harmonic-oscillator example of section 2.1, which has the same reconstruction space as the coin, there is poor justification for requiring this symmetry because the two probabilities—of finding the oscillator in its ground state, or not—are not on equal footing.

A.4. Invariance

When one speaks of an *invariant prior*, one does not mean the invariance under a change of parameterization—all priors are invariant in this respect—but rather a *form-invariant* construction in terms of a quantity that, preferably, has an invariant significance. We consider two particular constructions that make use of the metric induced by the response of the selected function to infinitesimal changes of its variables.

The first construction begins with a quantity $F(p)$ that is a function of all probabilities $p=(p_1,\ldots,p_K)$. We include the square root of the determinant of the dyadic second derivative in the prior density as a factor,

$$(\mathrm{d}\rho)=(\mathrm{d}p)\left|\det\left\{\left(\frac{\partial^2 F}{\partial p_j\,\partial p_k}\right)_{jk}\right\}\right|^{1/2} w_{\mathrm{cstr}}(p), \tag{A.7}$$

where $w_{\mathrm{cstr}}(p)$ contains all the delta-function and step-function factors of constraint as well as the normalization factor that ensures the unit size of the reconstruction space. The prior defined by (A.7) is invariant in the sense that a change of parameterization, from p to α, say, does not affect its structure

$$(\mathrm{d}\rho)=(\mathrm{d}\alpha)\left|\det\left\{\left(\frac{\partial^2 F}{\partial \alpha_j\,\partial \alpha_k}\right)_{jk}\right\}\right|^{1/2} w_{\mathrm{cstr}}(p(\alpha)), \tag{A.8}$$

because the various Jacobian determinants take care of each other. Since $w_{\mathrm{cstr}}(p)$ enforces all constraints, the p_k are independent variables when $F(p)$ and $G(p,\nu)$ are differentiated in (A.7) and (A.9), respectively.

For the second construction, we use a data-dependent function $G(p,\nu)$ of the probabilities p and the frequencies $\nu=(\nu_1,\nu_2,\ldots,\nu_K)$ with $\nu_j=n_j/N$. Here, the square root of the determinant of the expected value of the dyadic square of the p-gradient of G is a factor in the prior density

$$(\mathrm{d}\rho)=(\mathrm{d}p)\left|\det\left\{\overline{\left(\frac{\partial G}{\partial p_j}\frac{\partial G}{\partial p_k}\right)_{jk}}\right\}\right|^{1/2} w_{\mathrm{cstr}}(p), \tag{A.9}$$

where $\overline{f(\nu)}$ denotes the expected value of $f(\nu)$,

$$\overline{f(\nu)}=\sum_D L(D|\rho)f(\nu)\,. \tag{A.10}$$

We have, in particular, the generating function

$$\overline{\exp\left(\sum_{k=1}^{K} a_k\nu_k\right)}=\left(\sum_{k=1}^{K}\mathrm{e}^{a_k/N}p_k\right)^{N} \tag{A.11}$$

for the expected values of products of the ν_k. The prior defined by (A.9) is form-invariant in the same sense, and for the same reason, as the prior of (A.7).

Table A.1. Form-invariant priors constructed by one of the two methods described in the text. The '$\sqrt{\det}$' column gives the p-dependent factors only and omits all p-independent constants. The first method of (A.7) proceeds from functions of the probabilities that have extremal values when all probabilities are equal or all vanish save one. The second method of (A.9) uses functions that quantify how similar are the probabilities and the frequencies. The 'hedged prior' is named in analogy to the 'hedged likelihood' [25].

Method	Primary function	$\sqrt{\det}$
1st	$-\sum_k p_k \log p_k$ (Shannon entropy)	$\dfrac{1}{\sqrt{p_1 p_2 \cdots p_K}}$ (Jeffreys prior)
1st	$\sum_k p_k^2$ (purity)	1 (primitive prior)
2nd	$\sum_k \nu_k p_k$ (inner product)	$\sqrt{p_1 p_2 \cdots p_K}$ (hedged prior)
2nd	$\sum_k \nu_k \log(\nu_k / p_k)$ (relative entropy)	$\dfrac{1}{\sqrt{p_1 p_2 \cdots p_K}}$ (Jeffreys prior)

Table A.1 reports a few examples of '$\sqrt{\det}$' factors constructed by one of these two methods. It is worth noting that the Jeffreys prior can be obtained from the entropy of the probabilities by the first method as well as from the relative entropy between the probabilities and the frequencies by the second method. The latter is a variant of Jeffreys's original derivation [15] in terms of the Fisher information.

A.5. Conjugation

Sometimes there are reasons to expect that the actual state is close to a certain target state with probabilities $t = (t_1, t_2, \ldots, t_K)$. This is the situation, for example, when a source is designed to emit the quantum-information carriers in a particular state. A *conjugate prior*

$$(\mathrm{d}\rho) = (\mathrm{d}p)\left(p_1^{t_1} p_2^{t_2} \cdots p_K^{t_K}\right)^{\beta} w_{\mathrm{cstr}}(p) \quad \text{with } \beta > 0 \tag{A.12}$$

could then be a natural choice. Such priors are called 'conjugate' in standard statistics literature because the $(\cdots)^{\beta}$ factor has the same structure as the point likelihood: a product of powers of the detection probabilities. The $(\cdots)^{\beta}$ factor is maximal for $p = t$ and the peak is narrower when β is larger.

The conjugate prior can be understood as the 'mock posterior' for the primitive prior that results from pretending that β copies have been measured in the past and data obtained that are most typical for the target state. Therefore, a conjugate prior is quite a natural way of expressing the expectation that the apparatus is functioning well. The posterior content of a region will be data-dominated only if N is much larger than β.

In this context, it may be worth noting that the Bayesian mean state,

$$\widehat{\rho}_{\mathrm{BM}} = \int_{\mathcal{R}_0} (\mathrm{d}\rho)\, \rho, \tag{A.13}$$

computed with the conjugate prior above, is usually not the target state unless β is large. One could construct priors for which $\widehat{\rho}_{\mathrm{BM}}$ is the target state, but the presence of the $w_{\mathrm{cstr}}(p)$ factor requires a case-by-case construction.

A.6. Marginalization

All priors used as examples—the ones in (A.2), (A.4) and (A.12), and in table A.1—have in common that they are defined in terms of the probabilities and, therefore, they refer to the particular POM with which the data are collected. While this takes the significance of the data duly into account, it does not seem to square with the point of view that prior probabilities are solely a property of the physical processes that put the quantum-information carriers into the state that is then diagnosed by the POM.

When adopting this viewpoint, one begins with a prior density defined on the entire state space. In addition to the parameters that specify the reconstruction space (essentially the probabilities p), this full-space prior will depend on parameters whose values are not determined by the data. There could be very many nuisance parameters of this kind. In the harmonic-oscillator example of section 2.1, the data tell us only about the ground-state population but nothing about the population in any specific excited state. For a prior assigned on the formally infinite-dimensional state space, all but the ground-state population are nuisance parameters. Upon integrating the full-space prior over the nuisance parameters, one obtains a *marginal prior* on the reconstruction space. As a function on the reconstruction space, the marginal prior is naturally parameterized in terms of the probabilities and so fits into the formalism we are using throughout.

We note that the invoking of 'additional criteria' for a unique mapping from p to ρ, as mentioned at the end of section 2.1, is exactly what would be required if one wishes to report estimated values of the nuisance parameters. That, however, goes beyond making statements that are solidly supported by the data and is, therefore, outside the scope of our present discussion.

The symmetric uniform-in-purity prior of sections A.2 and A.3 provides an example for marginalization if the POM only gives information about $x = \langle\sigma_x\rangle$ and $y = \langle\sigma_y\rangle$ but not about $z = \langle\sigma_z\rangle$. We express the full-space prior in Cartesian coordinates, integrate over z, and arrive at

$$\begin{aligned}(\mathrm{d}\rho) &= \mathrm{d}x\,\mathrm{d}y\,\frac{1}{2\pi}\int_{-\infty}^{\infty}\mathrm{d}z\,\frac{\eta(1-x^2-y^2-z^2)}{\sqrt{x^2+y^2+z^2}}\\ &= \mathrm{d}x\,\mathrm{d}y\,\frac{1}{\pi}\,\eta(1-x^2-y^2)\cosh^{-1}\frac{1}{\sqrt{x^2+y^2}}\,.\end{aligned} \tag{A.14}$$

This marginal prior is a function on the unit disk in the xy plane, which is the natural choice of reconstruction space here. When one expresses $(\mathrm{d}\rho)$ in polar coordinates, one sees that $(\mathrm{d}\rho)$ is uniform in φ and in $r^2\cosh^{-1}(1/r) - \sqrt{1-r^2}$, which increases monotonically from -1 to 0 on the way from the center of the disk at $r = 0$ to the unit circle where $r = 1$. Plot (a) in figure A.1 illustrates the matter.

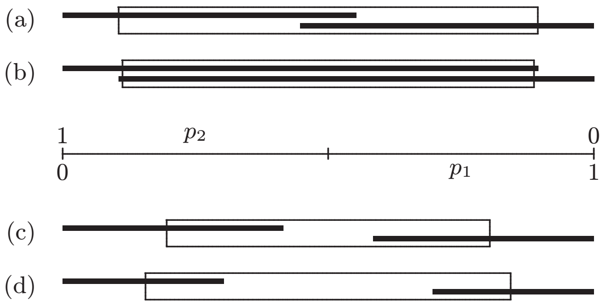

Figure B.1. Confidence regions and smallest credible regions. The bars indicate intervals of $p_1 = 1 - p_2$ for the harmonic-oscillator example of section 2.1, which has the reconstruction space of a tossed coin. Two copies are measured. The left solid bars indicate the regions for $(n_1, n_2) = (0, 2)$ counts; the right solid bars are for $(n_1, n_2) = (2, 0)$; and, the central open bars are for $(n_1, n_2) = (1, 1)$. Cases (a) and (b) show two sets of confidence regions for confidence level $\gamma = 0.8$. Regions (c) and (d) are the SCRs for the primitive prior and the Jeffreys prior, respectively, both for credibility $c = 0.8$.

Appendix B. Confidence regions

The confidence regions that were recently studied by Christandl and Renner [7], and independently by Blume-Kohout [8], are markedly different from the MLRs and the SCRs. The MLR and the SCR represent inferences drawn about the unknown state ρ from the data D that have actually been observed. By contrast, confidence regions are a set of regions, one region for each data, whether observed or not, from the measurement of N copies. The confidence regions would contain *any* state in, at least, a certain fraction of many N-copy measurements, if many measurements were performed. This fraction is the confidence level.

When denoting by $\mathcal{C}_D$ the confidence region for data D, the confidence level γ of the set $\mathbf{C}$ of $\mathcal{C}_D$s for all conceivable data (for fixed N) is

$$\gamma(\mathbf{C}) = \min_{\rho} \sum_{D} L(D|\rho)\, \eta_{\mathcal{C}_D}(\rho), \tag{B.1}$$

where the minimum is reached in the 'worst case'. For example, in the security analysis of a protocol for quantum key distribution, one wishes a large value of γ to protect against an adversary who controls the source and prepares the quantum-information carriers in the state that is best for her.

Any set $\mathbf{C}$, for which γ has the desired value, serves the purpose. A smaller set $\mathbf{C}'$, in the sense that $\mathcal{C}'_D$ is contained in $\mathcal{C}_D$ for all D, is preferable, but usually there is no smallest set of confidence regions. Here, 'smaller' is solely in this inclusion sense, with no reference to a quantification of the size of a region and, therefore, there is no necessity for specifying the prior probability of any region. Since the transition from set $\mathbf{C}$ to the smaller set $\mathbf{C}'$ requires the shrinking of some of the $\mathcal{C}_D$s without enlarging even a single one, it is easily possible to have two sets of confidence regions with the same confidence level and neither set smaller than the other.

For illustration, we consider the harmonic-oscillator example of section 2.1 yet another time. Figure B.1 shows two sets of confidence regions ($\gamma = 0.8$) and the corresponding three SCRs ($c = 0.8$) for the primitive prior and the Jeffreys prior. Both sets of confidence regions are optimal in the sense that one cannot shrink even one of the regions without decreasing the confidence level, but neither set is smaller than the other. In the absence of additional criteria that specify a preference, both work equally well as sets of confidence regions. This generic non-uniqueness of confidence regions, and the arbitrariness associated with it, are in marked contrast to the SCRs, which are always unique.

We also observe in this example that confidence regions tend to overlap a lot, which is indeed unavoidable if a large confidence level is desired; whereas, the SCRs for different data usually do not overlap unless the data are quite similar. In figure B.1, there is no overlap of the SCRs for $(n_1, n_2) = (0, 2)$ and $(2, 0)$.

Another important difference of considerable concern in all practical applications is the following. Once the data are obtained, there is *the* MLR and *the* SCR for these data, and it plays no role what other MLRs or SCRs are associated with different data that have not been observed. To find a confidence region for the actual data, however, one must first specify the whole set **C** of confidence regions because the confidence level of (B.1) is a property of the whole set.

References

[1] Paris M and Řeháček J (ed) 2004 *Quantum State Estimation* (*Lecture Notes in Physics* vol 649) (Heidelberg: Springer)

[2] Hradil Z, Řeháček J, Fiurášek J and Ježek M 2004 Maximum-likelihood methods in quantum mechanics *Quantum State Estimation* (*Lecture Notes in Physics* vol 649) ed M Paris and J Řeháček (Heidelberg: Springer) chapter 3

[3] Teo Y S 2012 Numerical estimation schemes for quantum tomography *PhD Thesis* Singapore (arXiv:1302:3399 [quant-ph])

[4] Řeháček J, Mogilevtsev D and Hradil Z 2008 *New J. Phys.* **10** 043022

[5] Audenaert K M R and Scheel S 2009 *New J. Phys.* **11** 023028

[6] Efron B and Tibshirani R J 1993 *An Introduction to the Bootstrap* (New York: Chapman and Hall)

[7] Christandl M and Renner R 2012 *Phys. Rev. Lett.* **109** 120403

[8] Blume-Kohout R 2012 Robust error bars for quantum tomography arXiv:1202.5270 [quant-ph]

[9] Berger J O 1985 *Statistical Decision Theory and Bayesian Analysis* 2nd edn (New York: Springer) chapter 4

[10] Evans M J, Guttman I and Swartz T 2006 *Can. J. Stat.* **34** 113

[11] Jaynes E T 1957 *Phys. Rev.* **106** 620

[12] Jaynes E T 1957 *Phys. Rev.* **108** 171

[13] Teo Y S, Zhu H, Englert B-G, Řeháček J and Hradil Z 2011 *Phys. Rev. Lett.* **107** 020404

[14] Bužek V 2004 Quantum tomography from incomplete data via MaxEnt principle *Quantum State Estimation* (*Lecture Notes in Physics* vol 649) ed M Paris and J Řeháček (Heidelberg: Springer) chapter 6

[15] Jeffreys H 1946 *Proc. R. Soc. Lond.* A **186** 453

[16] Kass R E and Wasserman L 1996 *J. Am. Stat. Assoc.* **91** 1343

[17] Teo Y S, Stoklasa B, Englert B-G, Řeháček J and Hradil Z 2012 *Phys. Rev.* A **85** 042317

[18] Wasserman L A 1989 *Ann. Stat.* **17** 1387

[19] Press W H, Teukolsky S A, Vetterling W T and Flannery B P 2007 *Numerical Recipes: The Art of Scientific Computing* 3rd edn (Cambridge: Cambridge University Press)

[20] Bennett C and Brassard G 1984 *IEEE Conf. Computers, Systems and Signal Processing, Bangalore, India* (New York: IEEE) p 175

[21] Tabia G and Englert B-G 2011 *Phys. Lett.* A **375** 817

[22] Lehmann E L and Casella G 1998 *Theory of Point Estimation* 2nd edn (Berlin: Springer)

[23] Ng H K, Phuah K T B and Englert B-G 2012 *New J. Phys.* **14** 085007

[24] Jaynes E T 2003 *Probability Theory—The Logic of Science* (Cambridge: Cambridge University Press)

[25] Blume-Kohout R 2010 *Phys. Rev. Lett.* **105** 200504