Quantum state tomography of a single qubit: comparison of methods

Roman Schmied

Taylor & Francis
Taylor & Francis Group

Quantum state tomography of a single qubit: comparison of methods

Roman Schmied

Department of Physics, University of Basel, Basel, Switzerland

ABSTRACT
The tomographic reconstruction of the state of a quantum-mechanical system is an essential component in the development of quantum technologies. We present an overview of different tomographic methods for determining the quantum-mechanical density matrix of a single qubit: (scaled) direct inversion, maximum likelihood estimation (MLE), minimum Fisher information distance and Bayesian mean estimation (BME). We discuss the different prior densities in the space of density matrices, on which both MLE and BME depend, as well as ways of including experimental errors and of estimating tomography errors. As a measure of the accuracy of these methods, we average the trace distance between a given density matrix and the tomographic density matrices it can give rise to through experimental measurements. We find that the BME provides the most accurate estimate of the density matrix, and suggest using either the pure-state prior, if the system is known to be in a rather pure state, or the Bures prior if any state is possible. The MLE is found to be slightly less accurate. We comment on the extrapolation of these results to larger systems.

1. Introduction

Quantum state tomography is the attempt to discover the quantum-mechanical state of a physical system, or more precisely, of a finite set of systems prepared by the same process [1]. The experimenter acquires a set of measurements of different non-commuting observables and tries to estimate what the density matrix of the systems must have been before the measurements were made, with the goal of being able to predict the statistics of future measurements generated by the same process. In this sense, quantum state tomography characterizes a state preparation process that is assumed to be stable over time [2].

In the context of the generation and characterization of non-classical states of Bose–Einstein condensates with internal degrees of freedom [3], the system under study is known to be in a totally symmetric state because of its Bose symmetry. These states are usually described in terms of total-spin observables, with the effective spin length equal to half the atom number. In this restricted framework, quantum-state reconstruction is much more feasible than for general many-particle systems; for this reason, the reconstruction of spin (or pseudo-spin) density matrices is an important real-world case for quantum state tomography. In practice, there are many different mathematical methods for determining a density matrix from a given experimental data-set, yielding sometimes very different results, and it is not obvious which of these

is objectively better, even when opinions and philosophical arguments are seemingly clear.

In order to see these methods more clearly and compare them, we apply them to the simplest possible quantum-mechanical problem of determining the density matrix of a two-level system (a qubit, or a spin of length 1/2), and compare the obtained results. We find that for qubits in general, Bayesian mean estimates (Section 2.4) are most accurate at determining a density matrix, in agreement with general statements of Refs. [4,5]. We generally consider mixed qubit states; for a review of *pure* qubit state estimation, see Ref. [6].

The quantum-mechanical state of any two-level system can be expressed as a 2×2 density matrix

$$\hat{\rho} = \frac{1}{2} \left(\mathbb{1} + x\hat{\sigma}_x + y\hat{\sigma}_y + z\hat{\sigma}_z \right) = \frac{1}{2} \left(\mathbb{1} + \boldsymbol{r} \cdot \hat{\boldsymbol{\sigma}} \right) \quad (1)$$

in terms of the Pauli matrices

$$\hat{\sigma}_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \hat{\sigma}_y = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}$$

$$\hat{\sigma}_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \qquad \mathbb{1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad (2)$$

and the vectors $\boldsymbol{r} = (x, y, z) \in \mathbb{R}^3$ and $\hat{\boldsymbol{\sigma}} = (\hat{\sigma}_x, \hat{\sigma}_y, \hat{\sigma}_z)$. Since the eigenvalues of $\hat{\rho}$ are $\lambda^{\pm} = \frac{1}{2}(1 \pm \sqrt{x^2 + y^2 + z^2})$ and must both be nonnegative, a *Bloch vector* $\boldsymbol{r}$ only represents a physical (positive semi-definite) state if

$\|\boldsymbol{r}\|^2 = x^2 + y^2 + z^2 \le 1$. The three-dimensional unit sphere of Bloch vectors, where every physically possible qubit density matrix can be represented as a point in space, is an appealing and convenient representation and will be used throughout this paper.

An alternative representation of a qubit density matrix is the spherical Wigner function [3,7]

$$W(\vartheta, \varphi) = \frac{1 + \sqrt{3}(\sin\vartheta\cos\varphi, \sin\vartheta\sin\varphi, \cos\vartheta) \cdot \boldsymbol{r}}{\sqrt{8\pi}} \tag{3}$$

defined as a pseudo-probability density on the *surface* of the unit sphere. It encodes the direction of the vector $\boldsymbol{r}$ in the angular distribution and the length of $\boldsymbol{r}$ in the amplitude of the pseudo-probability density. This representation is convenient for longer spins, where the Bloch vector representation is unavailable.

Many characteristics of a qubit state $\hat{\rho}$ can be expressed in terms of the length $r = \|\boldsymbol{r}\|$ of its Bloch vector alone, for example the *quantum Fisher information* [8] $F_Q(\hat{\rho}) = r^2$, the *purity* $\mathrm{Tr}(\hat{\rho}^2) = (1 + r^2)/2$, or the *von Neumann entropy*

$$S(\hat{\rho}) = -\frac{1+r}{2}\ln\left(\frac{1+r}{2}\right) - \frac{1-r}{2}\ln\left(\frac{1-r}{2}\right). \tag{4}$$

In what follows we consider only Stern–Gerlach type measurements on a single qubit: a projective measurement along an axis $\boldsymbol{n}$ (with $\|\boldsymbol{n}\| = 1$) is represented by the observable $\hat{\sigma}_{\boldsymbol{n}} = \boldsymbol{n} \cdot \hat{\boldsymbol{\sigma}}$, and has an expectation value

$$\langle\hat{\sigma}_{\boldsymbol{n}}\rangle = \mathrm{Tr}(\hat{\sigma}_{\boldsymbol{n}}\hat{\rho}) = \boldsymbol{n} \cdot \boldsymbol{r}. \tag{5}$$

The probabilities for detecting the qubit in the "up" state $|\boldsymbol{n}\uparrow\rangle$ satisfying $\hat{\sigma}_{\boldsymbol{n}}|\boldsymbol{n}\uparrow\rangle = +|\boldsymbol{n}\uparrow\rangle$, or in the "down" state $|\boldsymbol{n}\downarrow\rangle$ satisfying $\hat{\sigma}_{\boldsymbol{n}}|\boldsymbol{n}\downarrow\rangle = -|\boldsymbol{n}\downarrow\rangle$, are

$$p_\uparrow(\boldsymbol{n}) = \frac{1 + \boldsymbol{n}\cdot\boldsymbol{r}}{2}, \qquad p_\downarrow(\boldsymbol{n}) = \frac{1 - \boldsymbol{n}\cdot\boldsymbol{r}}{2}, \tag{6}$$

respectively.

If we identically prepare $N_{\boldsymbol{n}}$ qubits and measure the observable $\hat{\sigma}_{\boldsymbol{n}}$ on each one, we will find $N_{\boldsymbol{n}\uparrow}$ qubits in the $|\boldsymbol{n}\uparrow\rangle$ state and $N_{\boldsymbol{n}\downarrow}$ qubits in the $|\boldsymbol{n}\downarrow\rangle$ state, giving an estimate of the expectation value (sample mean)

$$\langle\!\langle\hat{\sigma}_{\boldsymbol{n}}\rangle\!\rangle = \frac{N_{\boldsymbol{n}\uparrow} - N_{\boldsymbol{n}\downarrow}}{N_{\boldsymbol{n}\uparrow} + N_{\boldsymbol{n}\downarrow}} = \frac{N_{\boldsymbol{n}\uparrow} - N_{\boldsymbol{n}\downarrow}}{N_{\boldsymbol{n}}}. \tag{7}$$

A statistical estimate of the error of this expectation value is given by the width of a binomial distribution with the same expectation value,

$$\Delta\langle\!\langle\hat{\sigma}_{\boldsymbol{n}}\rangle\!\rangle = \frac{2\sqrt{N_{\boldsymbol{n}\uparrow}N_{\boldsymbol{n}\downarrow}}}{N_{\boldsymbol{n}}^{3/2}}. \tag{8}$$

This error measure will be justified below through Equation (17).

In the absence of prior knowledge about the experimental system's state, the precision of the tomographic methods of Section 2 is highest if the measurement axes are arranged uniformly on the sphere. As detailed in Appendix 1, for a spin-1/2 system any angular distribution is considered uniform if its quadrupolar component vanishes. Examples of uniform sampling strategies according to this criterion are equal sampling along the three Cartesian axes, the four axes through the vertices of a tetrahedron [9] or a completely uniform distribution of measurement axes over the entire sphere. In what follows, we assume that the experimenter performs the same number of single-qubit measurements along each of the three Cartesian axes $\boldsymbol{n} = \boldsymbol{e}_x, \boldsymbol{e}_y, \boldsymbol{e}_z$, which is the simplest complete and uniform measurement strategy [10]. Using different axes or more than three axes generally makes all of the following tomography schemes more complicated, and if the measurements are not uniformly distributed (for example by making more measurements along $\boldsymbol{e}_z$ than along $\boldsymbol{e}_x$ or $\boldsymbol{e}_y$), the tomographic result will generally be less precise or even biased. However, if each measurement axis is adaptively chosen depending on the previous measurement results, efficiency can be improved over the Cartesian axes [11–13]. Also, multi-qubit joint measurements may yield information faster than sequential single-qubit measurements [14,15]. Such adaptations are not considered in the present work.

If our qubits are all in the state of Equation (1) and we perform N_x measurements along the x-axis, N_y along the y-axis and N_z along the z-axis, the probability of getting a certain set of results is

$$\begin{aligned}
&\mathcal{P}(N_{x\uparrow}, N_{x\downarrow}, N_{y\uparrow}, N_{y\downarrow}, N_{z\uparrow}, N_{z\downarrow}|\hat{\rho}) \\
&= \binom{N_x}{N_{x\uparrow}}\left(\frac{1+x}{2}\right)^{N_{x\uparrow}}\left(\frac{1-x}{2}\right)^{N_{x\downarrow}} \\
&\quad \times \binom{N_y}{N_{y\uparrow}}\left(\frac{1+y}{2}\right)^{N_{y\uparrow}}\left(\frac{1-y}{2}\right)^{N_{y\downarrow}} \\
&\quad \times \binom{N_z}{N_{z\uparrow}}\left(\frac{1+z}{2}\right)^{N_{z\uparrow}}\left(\frac{1-z}{2}\right)^{N_{z\downarrow}}, \tag{9}
\end{aligned}$$

where $N_x = N_{x\uparrow} + N_{x\downarrow}$ etc. and we will assume $N_x = N_y = N_z$ below. In such a set-up, the problem of quantum state tomography is to invert Equation (9): given a set of experimental results, what can we say about the qubits' density matrix that has given rise to these results? In what follows, we first present several tomographic methods and apply them to a single qubit (Section 2), make some comments about experimental and tomographic errors (Section 3) and then compare the accuracies of the different methods (Section 4).

2. Tomographic methods

2.1. Direct inversion tomography

The simplest tomographic method, called a *direct inversion*, assumes that the sample mean $\langle\!\langle \hat{\sigma}_{\boldsymbol{n}} \rangle\!\rangle$ is a good and unbiased estimate of the population mean $\langle \hat{\sigma}_{\boldsymbol{n}} \rangle$ [16]. Combining Equations (5) and (7) along the three Cartesian axes fully defines an estimate of the qubits' Bloch vector,

$$\boldsymbol{r}_{\mathrm{d}} = \left(\frac{N_{x\uparrow} - N_{x\downarrow}}{N_{x\uparrow} + N_{x\downarrow}}, \frac{N_{y\uparrow} - N_{y\downarrow}}{N_{y\uparrow} + N_{y\downarrow}}, \frac{N_{z\uparrow} - N_{z\downarrow}}{N_{z\uparrow} + N_{z\downarrow}} \right). \quad (10)$$

We note that $\boldsymbol{r}_{\mathrm{d}}$ is the global maximum of Equation (9), which is a definition of $\boldsymbol{r}_{\mathrm{d}}$ that is readily extensible to different measurement schemes.

In Figures 1 and 2 this direct inversion Bloch vector is shown as a blue dot for two sets of experimental results, both found by performing 30 Stern–Gerlach measurements along each Cartesian axis. While in Figure 2 the Bloch vector is physically valid since $\|\boldsymbol{r}_{\mathrm{d}}\| \leq 1$, the Bloch vector in Figure 1 is invalid and points out a fundamental problem with the direct inversion method. Equation (10) can be seen as three individual parameter estimations for the three Cartesian components of the Bloch vector, and even though each parameter estimate is unconstrained on its own, the three estimates must satisfy the joint constraint $x_{\mathrm{d}}^2 + y_{\mathrm{d}}^2 + z_{\mathrm{d}}^2 \leq 1$. For any given state $\boldsymbol{r}$ of the qubit and for any number of measurements (N_x, N_y, N_z), there is a finite probability that direct inversion tomography will find a physically invalid Bloch vector that violates this joint constraint. For example, for the completely mixed state $\hat{\rho} = \frac{1}{2}\mathbb{1}$ with $\boldsymbol{r} = (0, 0, 0)$, measuring $N_x = N_y = N_z = 30$ times along each Cartesian direction, the probability of finding an unphysical $\boldsymbol{r}_{\mathrm{d}}$ is only 3×10^{-7}; but if we do the same measurements on the pure state $\hat{\rho} = |z\uparrow\rangle\langle z\uparrow|$ with $\boldsymbol{r} = (0, 0, 1)$, the chance of finding an unphysical $\boldsymbol{r}_{\mathrm{d}}$ is 98%. For higher dimensional quantum systems, this problem becomes even more severe (see Section 4.1). It has been argued recently [17] that the direct inversion method provides more accurate results because it is less biased than other methods (see Table 1 for an example of such biases); but we side with Ref. [18] in preferring physically valid density matrices despite their bias, and do not report direct inversion results in our comparison of methods. Many interesting quantities derived from the density matrix, particularly ones that go beyond linear operator expectation values and involve the entire density matrix, cannot be defined properly for density matrices that are not positive semi-definite.

Nevertheless, $\boldsymbol{r}_{\mathrm{d}}$ is an important starting point for many other tomographic techniques. In what follows, we broadly distinguish between tomographic methods

that minimize some distance between $\boldsymbol{r}_{\mathrm{d}}$ and the space of physically valid tomographic Bloch vectors (Sections 2.2 and 2.3), and methods not based on $\boldsymbol{r}_{\mathrm{d}}$ at all (Section 2.4).

2.2. Distance minimization to r_{d}

In order to find a valid tomographic density matrix even if $\|\boldsymbol{r}_{\mathrm{d}}\| > 1$, we search for a modified Bloch vector $\boldsymbol{r}_{\mathrm{tomo}}$ that (i) is physically valid, $\|\boldsymbol{r}_{\mathrm{tomo}}\| \leq 1$, and that (ii) lies *closest* to $\boldsymbol{r}_{\mathrm{d}}$ in terms of a distance to be defined. In Figure 1 the three dashed lines emanating from the blue dot indicate the locations of the points that minimize three types of distances to $\boldsymbol{r}_{\mathrm{d}}$ on concentric spherical shells around the origin ($\boldsymbol{r} = 0$); their intersections with the unit sphere surface (red circle), among others, provide useable tomographic Bloch vectors and are discussed in detail below.

2.2.1. Minimum p-distance of the Bloch vectors

The simplest family of distances between two *Bloch vectors* are the p-distances $\|\boldsymbol{r} - \boldsymbol{r}'\|_p = (|x - x'|^p + |y - y'|^p + |z - z'|^p)^{1/p}$ for $p \geq 1$. Even though the direct inversion Bloch vector $\boldsymbol{r}_{\mathrm{d}}$ can be located anywhere in the unit cube, the space of physically valid Bloch vectors has an intrinsic spherical symmetry around the fully mixed state $\boldsymbol{r} = 0$, which suggests that only the Euclidean distance $p = 2$ is to be used. In this case, the *scaled direct inversion* Bloch vector minimizing the Euclidean distance to $\boldsymbol{r}_{\mathrm{d}}$ over the space of physically valid Bloch vectors is

$$\boldsymbol{r}_{\mathrm{sd}} = \begin{cases} \boldsymbol{r}_{\mathrm{d}} & \text{if } \|\boldsymbol{r}_{\mathrm{d}}\| \leq 1, \\ \boldsymbol{r}_{\mathrm{d}}/\|\boldsymbol{r}_{\mathrm{d}}\| & \text{if } \|\boldsymbol{r}_{\mathrm{d}}\| > 1. \end{cases} \quad (11)$$

Radial scaling is shown in Figures 1 and 2 as a red line, with $\boldsymbol{r}_{\mathrm{sd}}$ indicated as a red dot.

2.2.2. Minimum Schatten p-distance of the density matrices

The simplest family of distances between two *density matrices* are the Schatten p-distances $\|\hat{\rho} - \hat{\rho}'\|_p$. They include the trace distance ($p = 1$) and the Frobenius or Hilbert–Schmidt distance ($p = 2$). The Schatten p-distance between two qubit density matrices $\hat{\rho} = \frac{1}{2}(\mathbb{1} + \boldsymbol{r} \cdot \hat{\boldsymbol{\sigma}})$ and $\hat{\rho}' = \frac{1}{2}(\mathbb{1} + \boldsymbol{r}' \cdot \hat{\boldsymbol{\sigma}})$ is $\|\hat{\rho} - \hat{\rho}'\|_p = 2^{1/p}\frac{1}{2}\|\boldsymbol{r} - \boldsymbol{r}'\|$, proportional to the Euclidean distance between their Bloch vectors. The minimum of any Schatten p-distance between the direct inversion tomography and the space of physically valid density matrices is therefore given by Equation (11).
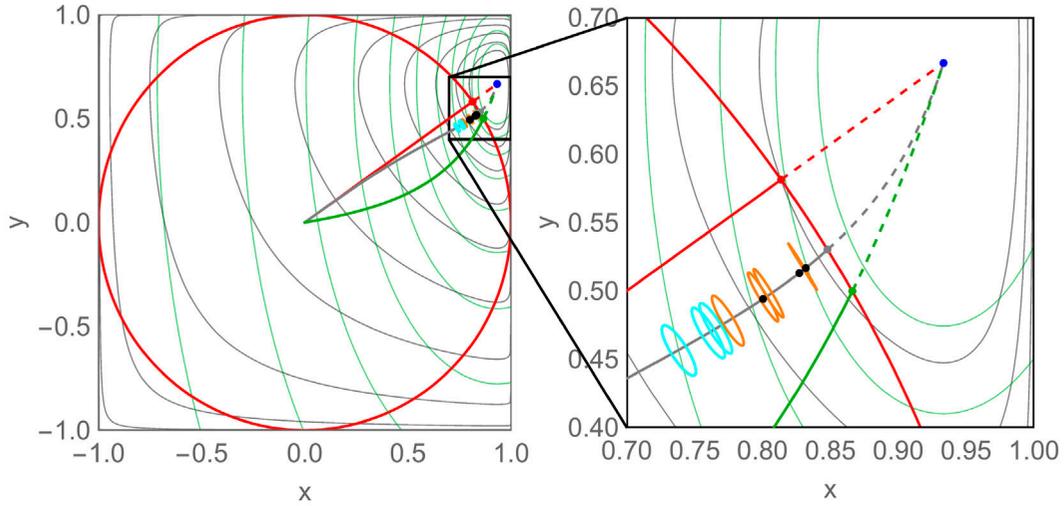
Figure 1. Example of a qubit tomography, assuming that 30 ideal measurements ($\eta = 1$, see Section 3.1) along each Cartesian quantization axes have resulted in $(N_{x\uparrow}, N_{x\downarrow}, N_{y\uparrow}, N_{y\downarrow}, N_{z\uparrow}, N_{z\downarrow}) = (29, 1, 25, 5, 15, 15)$. All quantities are restricted to the $z = 0$ plane. The blue dot shows the zero of the Kullback–Leibler divergence (log-likelihood) at $r_d = (\frac{14}{15}, \frac{2}{3}, 0)$ from Equation (10), which is outside of the physically allowed region $\|r\| \leq 1$ indicated by the red circle. The straight red line (Sections 2.2.1 and 2.2.2) connects r_d with the totally mixed state $r = 0$; the red dot shows the result of linear scaling $r_{sd} = (0.814, 0.581, 0)$ [Equation (11)]. The grey contours of the Kullback–Leibler divergence are at $D_{KL}(r_d|r) = 10^{n/4}$ for $n = 1 \ldots 10$ (outward from the blue dot). The grey line (Section 2.2.4) traces the constrained likelihood maximum, Equation (15), as a function of $\|r\|$; the likelihood maximum must be on this line if the prior depends only on $\|r\|$ (Haar measure), such as the grey dot showing the maximum of the likelihood at $r_{MLE}^{1<k\leq 2} = r_{MLE}^{Ch} = (0.848, 0.530, 0)$, or the black dots showing the maximum of the likelihood with entropy weight (24) for the Hilbert–Schmidt prior ($k = 2$) at $(0.800, 0.494, 0)$, the Bures prior ($k = \frac{3}{2}$) at $(0.827, 0.513, 0)$ and the Chernoff-information prior at $(0.832, 0.517, 0)$. The green contours of the Fisher information distance (Section 2.2.5) are at $D_{\mathcal{F}}(r - r_d) = 10^{n/4}$ for $n = 1 \ldots 11$; the green dot shows the point with minimum Fisher information distance at $r_{Fi} = (0.866, 0.500, 0)$, located on the green line of points tracing the minimum of the Fisher information distance as a function of $\|r\|$. The orange ellipses show the Bayesian means and variances of r weighted by the likelihood (Section 2.4): from left to right, they use radial priors with $k = 2$ (Hilbert–Schmidt measure), $k = \frac{3}{2}$ (Bures measure), the Chernoff-information measure (23) and $k = 1$ (pure states only); the cyan ellipses show the same with entropy weight (24): from left to right, they use radial priors with $k = 2$ (Hilbert–Schmidt measure), $k = \frac{3}{2}$ (Bures measure) and the Chernoff-information measure. (The colour version of this figure is included in the online version of the journal.)

2.2.3. Maximum fidelity

The fidelity $F(\hat{\rho}, \hat{\rho}') = \mathrm{Tr}\left(\sqrt{\sqrt{\hat{\rho}} \cdot \hat{\rho}' \cdot \sqrt{\hat{\rho}}}\right)$ is a frequently used measure of the overlap between two qubit density matrices [17]. Since it does not break the spherical symmetry of the space of Bloch vectors, maximizing the fidelity between two density matrices necessarily reduces to the purely radial scaling of Equation (11).

2.2.4. Kullback–Leibler divergence and the MLE

Bayes' theorem states that if we are given a set of experimental measurements $(N_{x\uparrow}, N_{x\downarrow}, N_{y\uparrow}, N_{y\downarrow}, N_{z\uparrow}, N_{z\downarrow})$, the likelihood that a certain density matrix $\hat{\rho} = \frac{1}{2}(\mathbb{1} + r \cdot \hat{\sigma})$ was at the source of these data is

$$\mathcal{L}(\hat{\rho}|N_{x\uparrow}, N_{x\downarrow}, N_{y\uparrow}, N_{y\downarrow}, N_{z\uparrow}, N_{z\downarrow})$$
$$\propto \mathcal{C}(\hat{\rho}) \times \mathcal{P}(N_{x\uparrow}, N_{x\downarrow}, N_{y\uparrow}, N_{y\downarrow}, N_{z\uparrow}, N_{z\downarrow}|\hat{\rho}), \quad (12)$$

where $\mathcal{C}(\hat{\rho})$ is a prior density in the space of density matrices, vanishing whenever $\|r\| > 1$. Choosing a prior density can be a matter of taste or actual prior knowledge; however, in almost all cases the prior density will depend

only on $\|r\|$ but not on the direction of r (i.e. it is a Haar measure with respect to the spherical symmetry group).

In this section, we only use the Hilbert–Schmidt measure

$$\mathcal{C}_{HS}(r) = \begin{cases} \text{const.} & \text{if } \|r\| \leq 1, \\ 0 & \text{if } \|r\| > 1 \end{cases} \quad (13)$$

as a prior density, which is uniform when viewed as the density of Bloch vectors within the unit sphere (but non-uniform when viewed in any other parametrization). While this is a simple and very common (often tacit) choice, it is not the most natural prior density; in Section 2.3 we discuss different prior densities and their application.

A popular tomography method is to search for the maximum of the likelihood (12) with $\mathcal{C}_{HS}(r)$ [19]. Since the global maximum of the probability $\mathcal{P}$, Equation (9), is at r_d, we see that whenever $\|r_d\| \leq 1$ the maximum-likelihood estimate (MLE) of the Bloch vector is simply $r_{MLE} = r_d$. If $\|r_d\| > 1$, on the other hand, we define the scaled log-likelihood, relative entropy or Kullback–

Leibler divergence [19,20]

$$
\begin{aligned}
D_{\mathrm{KL}}(\boldsymbol{r}_{\mathrm{d}}|\boldsymbol{r}) &= \ln\left[\frac{\mathcal{P}(\boldsymbol{r}_{\mathrm{d}})}{\mathcal{P}(\boldsymbol{r})}\right] \\
&= N_{x\uparrow}\ln\left(\frac{1+x_{\mathrm{d}}}{1+x}\right) + N_{x\downarrow}\ln\left(\frac{1-x_{\mathrm{d}}}{1-x}\right) \\
&\quad + N_{y\uparrow}\ln\left(\frac{1+y_{\mathrm{d}}}{1+y}\right) + N_{y\downarrow}\ln\left(\frac{1-y_{\mathrm{d}}}{1-y}\right) \\
&\quad + N_{z\uparrow}\ln\left(\frac{1+z_{\mathrm{d}}}{1+z}\right) + N_{z\downarrow}\ln\left(\frac{1-z_{\mathrm{d}}}{1-z}\right)
\end{aligned}
\tag{14}
$$

and minimize this distance over the space of physically valid density matrices $\|\boldsymbol{r}\| \le 1$[1]. Especially for large numbers of experimental data, the log-likelihood is easier to calculate in practice than the likelihood, as its dynamic range is much smaller; since the logarithm is monotonic, maximizing $\mathcal{P}$ is equivalent to minimizing D_{KL}. In Figures 1 and 2, the grey contours show the Kullback–Leibler divergence, and the gray dot in Figure 1 gives the likelihood maximum within the unit sphere. The grey line emanating from the blue dot is found by maximizing Equation (12), or minimizing Equation (14), for constant $\|\boldsymbol{r}\|$, assuming that the prior depends only on $\|\boldsymbol{r}\|$: we find that these extrema are located at

$$
\boldsymbol{r}_{\mathrm{MLE}}(\alpha) = \left[J_{\alpha/N_x}(x_{\mathrm{d}}), J_{\alpha/N_y}(y_{\mathrm{d}}), J_{\alpha/N_z}(z_{\mathrm{d}})\right]
\tag{15}
$$

with the analytic function $J_u(t)$ defined piecewise,

$$
J_u(t) = \begin{cases}
\operatorname{sign}(t) & \text{for } u \to -\infty \\
2\sqrt{\frac{u+1}{3u}}\operatorname{sign}(t)\cos\left[\frac{1}{3}\cos^{-1}\left(\frac{3}{2}|t|\sqrt{\frac{3u}{(u+1)^3}}\right)\right] & \text{if } u < -1 \text{ (branch cut at } t=0) \\
\operatorname{sign}(t)|t|^{1/3} & \text{if } u = -1 \\
2\sqrt{\frac{u+1}{-3u}}\sinh\left[\frac{1}{3}\sinh^{-1}\left(\frac{3}{2}t\sqrt{\frac{-3u}{(u+1)^3}}\right)\right] & \text{if } -1 < u < 0 \\
t & \text{if } u = 0 \\
2\sqrt{\frac{u+1}{3u}}\sin\left[\frac{1}{3}\sin^{-1}\left(\frac{3}{2}t\sqrt{\frac{3u}{(u+1)^3}}\right)\right] & \text{if } u > 0 \\
0 & \text{for } u \to +\infty.
\end{cases}
\tag{16}
$$

The grey line given by Equation (15) has the following properties as a function of the Lagrange multiplier α:

- $\boldsymbol{r}_{\mathrm{MLE}}(\alpha)$ maximizes $\mathcal{P}(N_{x\uparrow}, N_{x\downarrow}, N_{y\uparrow}, N_{y\downarrow}, N_{z\uparrow}, N_{z\downarrow}|\hat{\rho})$ and minimizes $D_{\mathrm{KL}}(\boldsymbol{r}_{\mathrm{d}}|\boldsymbol{r})$ under the constraint that $\|\boldsymbol{r}\| = \|\boldsymbol{r}_{\mathrm{MLE}}(\alpha)\|$,
- $\|\boldsymbol{r}_{\mathrm{MLE}}(\alpha)\|$ decreases monotonically with $\alpha \in \mathbb{R}$,
- $\lim_{\alpha \to -\infty} \boldsymbol{r}_{\mathrm{MLE}}(\alpha) = (\operatorname{sign}[x_{\mathrm{d}}], \operatorname{sign}[y_{\mathrm{d}}], \operatorname{sign}[z_{\mathrm{d}}])$,
- $\boldsymbol{r}_{\mathrm{MLE}}(0) = \boldsymbol{r}_{\mathrm{d}}$,
- $\lim_{\alpha \to \infty} \boldsymbol{r}_{\mathrm{MLE}}(\alpha) = 0$.

Since $J_u(t)$ has a branch cut discontinuity at $t = 0$ for $u < -1$, we must be careful when evaluating Equation (15) if any of the $(x_{\mathrm{d}}, y_{\mathrm{d}}, z_{\mathrm{d}})$ are zero (see below).

Thus the maximum likelihood method for Cartesian-axes qubit tomography is simpler than the $\hat{R} \cdot \hat{\rho} \cdot \hat{R}$ iteration used for larger systems [19], and consists of the following steps:

(1) Calculate $\boldsymbol{r}_{\mathrm{d}}$ from Equation (10).
(2) If $\|\boldsymbol{r}_{\mathrm{d}}\| \le 1$, set $\boldsymbol{r}_{\mathrm{MLE}} = \boldsymbol{r}_{\mathrm{d}}$.
(3) If $\|\boldsymbol{r}_{\mathrm{d}}\| > 1$, find $\alpha > 0$ such that $\|\boldsymbol{r}_{\mathrm{MLE}}(\alpha)\| = 1$.

2.2.5. Fisher information distance

When the direct inversion Bloch vector $\boldsymbol{r}_{\mathrm{d}}$ is only slightly outside the unit sphere of physically valid states, it may be sufficiently accurate to minimize the quadratic approximation of the Kullback–Leibler divergence (14),

$$
\begin{aligned}
D_{\mathrm{KL}}(\boldsymbol{r}_{\mathrm{d}}|\boldsymbol{r}) = \frac{1}{2}&\left[\left(\frac{x - x_{\mathrm{d}}}{\Delta\langle\!\langle\hat{\sigma}_x\rangle\!\rangle}\right)^2 + \left(\frac{y - y_{\mathrm{d}}}{\Delta\langle\!\langle\hat{\sigma}_y\rangle\!\rangle}\right)^2\right. \\
&\left. + \left(\frac{z - z_{\mathrm{d}}}{\Delta\langle\!\langle\hat{\sigma}_z\rangle\!\rangle}\right)^2\right] + \mathcal{O}[(\boldsymbol{r} - \boldsymbol{r}_{\mathrm{d}})^3],
\end{aligned}
\tag{17}
$$

given in terms of the error estimates of Equation (8). This approximation, called the *Fisher information distance* [21], is easier to use than the Kullback–Leibler divergence while mostly giving comparable results (see Table 2). In Figures 1 and 2 the green lines show the minima of the Fisher information distance on concentric shells around the origin $\boldsymbol{r} = 0$, and the green dot in Figure 1 minimizes this distance between $\boldsymbol{r}_{\mathrm{d}}$ and the space of physically valid states. In analogy to Equation (15), the green line is given by

$$
\boldsymbol{r}_{\mathrm{Fi}}(\alpha) = \\
\left(\frac{x_{\mathrm{d}}}{1 + \alpha[\Delta\langle\!\langle\hat{\sigma}_x\rangle\!\rangle]^2}, \frac{y_{\mathrm{d}}}{1 + \alpha[\Delta\langle\!\langle\hat{\sigma}_y\rangle\!\rangle]^2}, \frac{z_{\mathrm{d}}}{1 + \alpha[\Delta\langle\!\langle\hat{\sigma}_z\rangle\!\rangle]^2}\right)
\tag{18}
$$

and has similar properties, so that the three-step recipe of Section 2.2.4 can still be used. There are situations where no α exists that satisfies $\|\boldsymbol{r}_{\mathrm{Fi}}(\alpha)\| = 1$, but we have found that they are very unlikely to occur in an experiment (see Tables 1 and 2).

2.3. MLE with radial prior

The Bayesian prior density $\mathcal{C}(\hat{\rho})$ used in Equation (12) contains two components that are sometimes difficult to distinguish. On the one hand, it contains a measure on the space of density matrices, which is a way of saying how "finely grained" this space is in its different regions, or from what distribution a purely random density matrix should be drawn in the absence of concrete knowledge about the system [22]. This first part is likely invariant under unitary transformations (i.e. a Haar measure). On the other hand, $\mathcal{C}(\hat{\rho})$ can contain prior knowledge about the particular situation in which we are determining density matrices, gained for example from previous experiments. This second part need not be invariant under unitary transformations. Expressed in a given parametrization, which in our case is the Bloch vector $\boldsymbol{r}$ and Equation (1), the prior density $\mathcal{C}(\boldsymbol{r})$ is the product of the measure expressed in terms of $\boldsymbol{r}$ and the density gained from prior knowledge. It is important to note that concrete prior knowledge in the absence of a measure on the space of density matrices is useless.

In the previous section, we have used the Hilbert–Schmidt measure on the space of Bloch vectors (13) because of its simplicity, ubiquity and geometric appeal. However, this prejudice is misleading, and the Hilbert–Schmidt measure is neither the only nor the most natural density of quantum states of a qubit. In this section, we discuss different density-matrix measures, and then use these to generalize the maximum-likelihood method to non-trivial priors.

2.3.1. Radial prior densities of quantum states

There is much freedom in defining a measure on the space of Bloch vectors. In order to focus on more natural measures, we use a physical argument for defining such a measure: a constructive procedure related to quantum state purification [4,22,23].

We start from the observation that the density of *pure* states of a d-dimensional quantum system is uniquely defined as a Haar measure over the unitary group $U(d)$; that is, since every pure state is related to every other pure state by a unitary transformation, and since all unitary transformations can be parametrized as points on the surface of a $(d^2 - 1)$-dimensional hypersphere, we can use the geometric measure on this hypersphere's surface as the natural measure in the space of pure states.

Next, we consider the joint tensor-product quantum state of our two-dimensional qubit ($D = 2$) and a k-dimensional ancillary system, for a total dimension $d = D + k$. For every pure state of this $(2 + k)$-dimensional system, we can trace out the ancillary dimensions to find a reduced qubit density matrix (1). The reverse is also true, called quantum state purification: for every qubit density matrix (1) we can find a pure state of a system of $d \geq 2D = 4$ dimensions, of which our state is the partial trace. This partial trace operation therefore constructs a unique measure of qubit density matrices, depending only on the ancillary dimension k. Expressed as a density in the space of qubit Bloch vectors (the unit sphere), the resulting density (measure) is

$$\mathcal{C}_k(\boldsymbol{r}) = \begin{cases} \frac{\Gamma(k+\frac{1}{2})}{\pi^{3/2}\Gamma(k-1)}(1 - \|\boldsymbol{r}\|^2)^{k-2} & \text{if } \|\boldsymbol{r}\| < 1 \\ 0 & \text{if } \|\boldsymbol{r}\| > 1 \end{cases} \quad (19)$$

for $k > 1$, where $\Gamma(z)$ is the Euler gamma function. As expected, this measure only depends on the length of the Bloch vector but not on its direction. The mean squared Bloch vector of this measure is $\langle\|\boldsymbol{r}\|^2\rangle = 3/(2k + 1)$: for larger values of k, mixed states carry more weight than pure states.

How can we choose a value for the ancillary dimension k? While the derivation of Equation (19) assumes that k is an integer, we can use the resulting prior density for any value of k. Not all values of k are equally natural; we deem the following choices meaningful:

- $k = 1$ *pure states*   In the limit $k \to 1^+$ the measure (19) becomes fully concentrated on the surface of the unit sphere ($\|\boldsymbol{r}\| = 1$), meaning that only pure states have a nonzero likelihood in Equation (12). While this is not a natural choice, as (strictly speaking) pure states do not exist in nature, it can be of interest for theoretical considerations or in cases where the state purity is known to be very high.
- $k = \frac{3}{2}$ *Bures measure*   In general, the Bures measure [22,24,25] is considered the most natural density of mixed states [22], as it is the Jeffreys prior [26]. For qubits, its distribution is formally that of tracing over $k = \frac{3}{2}$ ancillary dimensions, and has the radial density [27][2]

$$\mathcal{C}_{\mathrm{B}}(\boldsymbol{r}) = \mathcal{C}_{\frac{3}{2}}(\boldsymbol{r}) = \begin{cases} \frac{1}{\pi^2\sqrt{1-\|\boldsymbol{r}\|^2}} & \text{if } \|\boldsymbol{r}\| < 1, \\ 0 & \text{if } \|\boldsymbol{r}\| > 1, \end{cases} \quad (20)$$

shown as a solid blue line in Figure 3. If nothing at all is known about the expected tomographic density matrix, then this Jeffreys prior density should be used.
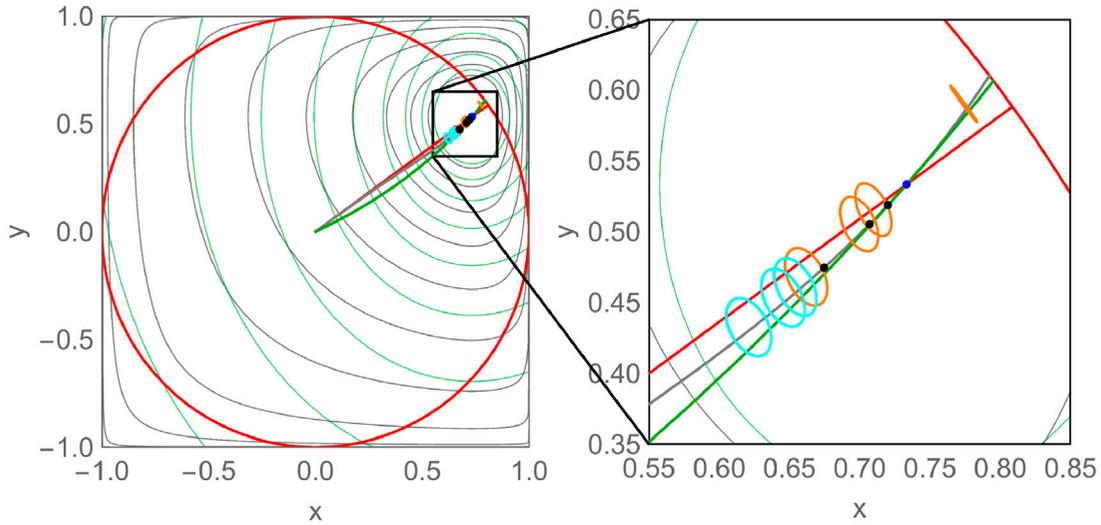
Figure 2. Same as Figure 1 but for $(N_{x\uparrow}, N_{x\downarrow}, N_{y\uparrow}, N_{y\downarrow}, N_{z\uparrow}, N_{z\downarrow}) = (26, 4, 23, 7, 15, 15)$. The global likelihood maximum $r_d = (0.733, 0.533, 0)$ (blue dot) now lies inside the unit sphere and coincides with the red, grey and green dots of Figure 1. (The colour version of this figure is included in the online version of the journal.)

Note that for systems with Hilbert space dimension $D > 2$, the Bures measure cannot be constructed by choosing a particular value of k.

In the sphere of Bloch vectors r, the Bures measure assigns a higher density of states to purer states (large $r = \|r\|$) than to more mixed states (small r). We can introduce a transformed radial coordinate $s = \left[\frac{2}{\pi}\left(\sin^{-1}(r) - r\sqrt{1-r^2}\right)\right]^{1/3}$, in terms of which the Bures measure is homogeneous:

$$\mathcal{C}_B(s) = \begin{cases} \frac{3}{4\pi} & \text{if } \|s\| < 1, \\ 0 & \text{if } \|s\| > 1. \end{cases} \quad (21)$$

This shows that the flatness of the measure depends on the chosen parametrization, and cannot be used as a criterion to prefer one measure over another.

- $k = 2$ *Hilbert–Schmidt measure* The previously used Hilbert–Schmidt measure of Equation (13) is found by setting the ancillary dimension equal to the system dimension, $k = D = 2$. It is equal to the Euclidean measure in the unit sphere of Bloch vectors, meaning that it gives every Bloch vector equal *a priori* weight in the simplest geometric sense (solid red line in Figure 3). This prior is used very frequently in practice, mainly due to its mathematical simplicity; but it must be noted that it does not represent the natural density of qubit states [22].

- $k \gg 2$ *highly mixed states* For large ancillary dimensions the density matrix measure becomes Gaussian and peaked around the fully mixed state,

$$\mathcal{C}_k(r) \approx \left(\frac{k+\frac{1}{2}}{\pi}\right)^{\frac{3}{2}} e^{-(k+\frac{1}{2})\|r\|^2} \text{ for } k \gg 1 \quad (22)$$
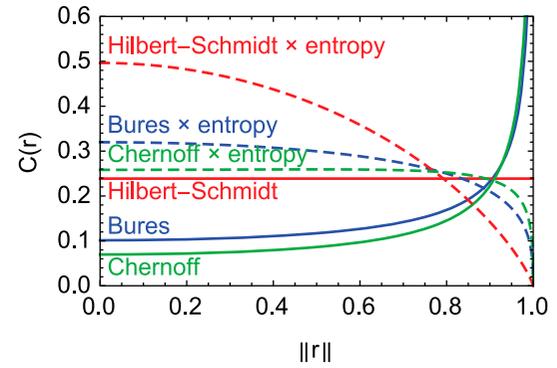
Figure 3. A few spherically symmetric prior densities of Bloch vectors. Solid red line: the Hilbert–Schmidt measure $\mathcal{C}_{HS}(r) = \mathcal{C}_2(r)$, Equation (13). Solid blue line: the Bures measure $\mathcal{C}_B(r) = \mathcal{C}_{\frac{3}{2}}(r)$, Equation (20). Solid green line: the Chernoff-information measure $\mathcal{C}_{Ch}(r)$, Equation (23). The dashed lines are the same measures weighted by the entropy as in Equation (24), and normalized. (The colour version of this figure is included in the online version of the journal.)

This measure can be used for tomographies where the state is known to be highly mixed.

There are other ways of defining a measure on the space of qubit density matrices. As an example, the Chernoff-information measure [27]

$$\mathcal{C}_{Ch}(r) = \begin{cases} \frac{(1-\|r\|^2)^{-\frac{1}{2}} - 1}{2\pi(\pi-2)\|r\|^2} & \text{if } \|r\| < 1 \\ 0 & \text{if } \|r\| > 1 \end{cases} \quad (23)$$

follows from the experimental distinguishability of density matrices, and is shown as a green line in Figure 3.

Once a measure has been chosen for the space of density matrices, the prior density in Equation (12) can

be taken directly from Equations (19) and (23), or other, or it can be further multiplied by a weight of our choice, for example representing concrete prior knowledge. As an example, we may use the entropy (4) as a radial weight, in combination with an underlying state measure:

$$\mathcal{C}(\boldsymbol{r}) \propto \mathcal{C}_k(\boldsymbol{r}) \times S(\boldsymbol{r}), \tag{24}$$

biasing the likelihood (12) towards less pure states. Figure 3 shows a few examples of prior densities, including entropy weights.

2.3.2. MLEs with different priors

The MLE of Section 2.2.4 is easily adapted to any spherically symmetric prior density $\mathcal{C}(\hat{\rho}) = \mathcal{C}(\|\boldsymbol{r}\|)$. Since Equation (15) maximizes the likelihood on each concentric shell $\|\boldsymbol{r}\| = \|\boldsymbol{r}_{\mathrm{MLE}}(\alpha)\|$, maximizing the likelihood globally thus means finding the value of $\alpha \in \mathbb{R}$ that maximizes the likelihood $\mathcal{L}[\boldsymbol{r}_{\mathrm{MLE}}(\alpha)]$, Equation (12). For this maximization we can distinguish different classes of priors:

- *pure or pure-peaked*   If the prior density is singular at $\|\boldsymbol{r}\| = 1$, for example Equation (19) with $1 \leq k < 2$ or Equation (23), it is sufficient to look for the value of α for which $\|\boldsymbol{r}_{\mathrm{MLE}}(\alpha)\| = 1$. Two special cases are important: if two or three components of $\boldsymbol{r}_{\mathrm{d}}$ are zero, then the likelihood maximum is not unique and the MLE should not return a value; the same is true if only one component of $\boldsymbol{r}_{\mathrm{d}}$ is zero and the determined value of $(-\alpha)$ is larger than the number of measurements along this axis.[3]
- *monotonic pure-biased*   We distinguish three cases:
  - $\|\boldsymbol{r}_{\mathrm{d}}\| > 1$ find the value of $\alpha > 0$ for which $\|\boldsymbol{r}_{\mathrm{MLE}}(\alpha)\| = 1$.
  - $\|\boldsymbol{r}_{\mathrm{d}}\| = 1$ the likelihood maximum is at $\boldsymbol{r}_{\mathrm{d}}$.
  - $\|\boldsymbol{r}_{\mathrm{d}}\| < 1$ find the value of $\alpha < 0$, with $\|\boldsymbol{r}_{\mathrm{MLE}}(\alpha)\| \leq 1$, that maximizes the likelihood $\mathcal{L}[\boldsymbol{r}_{\mathrm{MLE}}(\alpha)]$.
- *uniform (Hilbert–Schmidt)*   see Section 2.2.4 for an extended discussion.
- *monotonic mixed-biased*   find the value $\alpha > 0$ for which $\boldsymbol{r}_{\mathrm{MLE}}(\alpha)$ maximizes Equation (12).
- *non-monotonic*   find the value α, with $\|\boldsymbol{r}_{\mathrm{MLE}}(\alpha)\| \leq 1$, for which $\boldsymbol{r}_{\mathrm{MLE}}(\alpha)$ globally maximizes Equation (12).

In Figures 1 and 2 the likelihood maxima are shown for several different prior densities. We can see that for many priors and experimental results, the MLE is rank-deficient ($\|\boldsymbol{r}_{\mathrm{MLE}}\| = 1$), which is a serious drawback of this method [4].

2.4. Bayesian mean estimate

Instead of reporting only the maximum of the likelihood (12), we can interpret the likelihood as a density in the state space and use it to calculate a weighted mean state. This Bayesian mean estimate [4] is

$$\overline{\hat{\rho}}_{\mathrm{BME}} = \frac{\int \hat{\rho}\, \mathcal{L}(\hat{\rho})\, \mathcal{D}\hat{\rho}}{\int \mathcal{L}(\hat{\rho})\, \mathcal{D}\hat{\rho}}, \tag{25}$$

where $\mathcal{D}\hat{\rho}$ represents the chosen measure on the space of density matrices, and $\mathcal{L}(\hat{\rho})$ contains the experimental knowledge including prior knowledge (see the discussion of Section 2.3 on the two components of the prior density). In practice this integral is done by averaging the components of the Bloch vector,

$$\overline{\boldsymbol{r}}_{\mathrm{BME}} = \frac{\int_{\|\boldsymbol{r}\| \leq 1} \boldsymbol{r}\mathcal{L}(\boldsymbol{r})\mathrm{d}^3\boldsymbol{r}}{\int_{\|\boldsymbol{r}\| \leq 1} \mathcal{L}(\boldsymbol{r})\mathrm{d}^3\boldsymbol{r}}, \tag{26}$$

where the measure on the space of density matrices is now *included* in the definition of the likelihood (12) and expressed in terms of the geometric Bloch vector measure $\mathrm{d}^3\boldsymbol{r}$, as in Equation (19). The Bayesian mean is generally more plausible than the likelihood maximum [4] because it is never rank-deficient.

This method can be naturally extended to higher moments of the density matrix, from which we can calculate a covariance matrix: for example, with

$$\overline{x^2}_{\mathrm{BME}} = \frac{\int_{\|\boldsymbol{r}\| \leq 1} x^2 \mathcal{L}(\boldsymbol{r})\mathrm{d}^3\boldsymbol{r}}{\int_{\|\boldsymbol{r}\| \leq 1} \mathcal{L}(\boldsymbol{r})\mathrm{d}^3\boldsymbol{r}} \tag{27}$$

we can define the variance $(\Delta x_{\mathrm{BME}})^2 = \overline{x^2}_{\mathrm{BME}} - \overline{x}^2_{\mathrm{BME}}$, and similarly the entire covariance matrix for the components of $\overline{\boldsymbol{r}}_{\mathrm{BME}}$. In Figures 1 and 2 we show these covariances as orange and cyan ellipses around the Bayesian mean estimates for different choices of the prior density $\mathcal{C}(\boldsymbol{r})$.

At this point we need to distinguish between two kinds of uncertainty: firstly, there is the quantum-mechanical uncertainty within a single density matrix (Bloch vector), which is typically a statistical mixture of pure states, and leads to the well-known stochastic outcomes of observables through Born's rule; and secondly, there is the uncertainty in the density matrix (Bloch vector) parametrization coming from tomographic uncertainties, given above by the covariance matrix of the components of the Bloch vector.

When we calculate the linear expectation value and variance of an operator $\hat{A}$, these two types of uncertainty cannot be distinguished, and the expectation value and variance are estimated with

$$\langle \hat{A} \rangle = \frac{\int_{\|\boldsymbol{r}\| \leq 1} \mathrm{Tr}[\hat{\rho}(\boldsymbol{r})\hat{A}]\mathcal{L}(\boldsymbol{r})\mathrm{d}^3\boldsymbol{r}}{\int_{\|\boldsymbol{r}\| \leq 1} \mathcal{L}(\boldsymbol{r})\mathrm{d}^3\boldsymbol{r}} = \mathrm{Tr}[\overline{\hat{\rho}}_{\mathrm{BME}}\hat{A}],$$

(28a)

$$(\Delta A)^2 = \langle \hat{A}^2 \rangle - \langle \hat{A} \rangle^2$$
$$= \mathrm{Tr}[\overline{\hat{\rho}}_{\mathrm{BME}}\hat{A}^2] - \mathrm{Tr}[\overline{\hat{\rho}}_{\mathrm{BME}}\hat{A}]^2,$$

(28b)

where the average density matrix is given in Equations (25) and (26). In this sense, the mean density matrix $\overline{\hat{\rho}}_{\mathrm{BME}} = \hat{\rho}(\overline{\boldsymbol{r}}_{\mathrm{BME}})$ represents the statistical mixture containing both the quantum uncertainty in each Bloch vector and the uncertainty of the parametrization of the Bloch vector itself.

The situation is different when we estimate the Bayesian mean value of a non-linear quantity such as the mean purity $\langle \mathrm{Tr}(\hat{\rho}^2) \rangle$ or the mean entropy $\langle S(\hat{\rho}) \rangle$: in these cases, the covariance of the Bloch vector, Equation (27), becomes important. For example,

$$\langle S(\hat{\rho}) \rangle = \frac{\int_{\|\boldsymbol{r}\| \leq 1} S[\hat{\rho}(\boldsymbol{r})]\mathcal{L}(\boldsymbol{r})\mathrm{d}^3\boldsymbol{r}}{\int_{\|\boldsymbol{r}\| \leq 1} \mathcal{L}(\boldsymbol{r})\mathrm{d}^3\boldsymbol{r}} \neq S(\overline{\hat{\rho}}_{\mathrm{BME}}),$$

(29a)

$$[\Delta S(\hat{\rho})]^2 = \frac{\int_{\|\boldsymbol{r}\| \leq 1} S^2[\hat{\rho}(\boldsymbol{r})]\mathcal{L}(\boldsymbol{r})\mathrm{d}^3\boldsymbol{r}}{\int_{\|\boldsymbol{r}\| \leq 1} \mathcal{L}(\boldsymbol{r})\mathrm{d}^3\boldsymbol{r}} - \langle S(\hat{\rho}) \rangle^2 \quad (29b)$$

must be calculated by taking the uncertainty in the Bloch vector parametrization, given by $\mathcal{L}(\boldsymbol{r})$, into account.

3. Error considerations

3.1. Including experimental errors

In a real experiment, the outcomes of Stern–Gerlach measurements are never perfect. For simplicity, we assume that independently of the measurement direction, every measurement has a probability η of giving the correct result and a probability $1 - \eta$ of giving a random result (i.e. passing through a depolarizing channel [28]), or equivalently, a probability of $(1 + \eta)/2$ of giving the correct result and $(1 - \eta)/2$ of giving the wrong result.

Many sources of experimental errors can be expressed in this form, apart from simple detection errors. For example, if the experimental Stern–Gerlach axes fluctuate around their respective mean directions with a variance $4\langle \sin^2(\chi/2) \rangle$ (with χ the angle between the desired axis and the true experimental axis), the experimental error can be described by $\eta = \langle \cos(\chi) \rangle = 1 - 2\langle \sin^2(\chi/2) \rangle$. If several independent sources of errors $\eta_1, \eta_2, \dots$ are present, the total error is described by their product $\eta = \eta_1 \eta_2 \dots$.

In the presence of such experimental errors, the probability of measuring a certain data-set is modified from Equation (9) to

$$\mathcal{P}_\eta(N_{x\uparrow}, N_{x\downarrow}, N_{y\uparrow}, N_{y\downarrow}, N_{z\uparrow}, N_{z\downarrow}|\hat{\rho})$$
$$= \binom{N_x}{N_{x\uparrow}}\left(\frac{1 + \eta x}{2}\right)^{N_{x\uparrow}}\left(\frac{1 - \eta x}{2}\right)^{N_{x\downarrow}}$$
$$\times \binom{N_y}{N_{y\uparrow}}\left(\frac{1 + \eta y}{2}\right)^{N_{y\uparrow}}\left(\frac{1 - \eta y}{2}\right)^{N_{y\downarrow}}$$
$$\times \binom{N_z}{N_{z\uparrow}}\left(\frac{1 + \eta z}{2}\right)^{N_{z\uparrow}}\left(\frac{1 - \eta z}{2}\right)^{N_{z\downarrow}}$$

(30)

with $\eta \in [0, 1]$. For $\eta = 1$ the measurements are perfect and we recover Equation (9); for $\eta = 0$ the measurements contain no information about $\hat{\rho}$.

The form of Equation (30) is strictly that of Equation (9) where the Bloch vector $\boldsymbol{r}$ is replaced by $\eta\boldsymbol{r}$. All the tomographic methods of Section 2 can therefore be used to determine the vector $\eta\boldsymbol{r}$, with the caveat that the prior density depends on $\boldsymbol{r}$ and not on $\eta\boldsymbol{r}$. The direct inversion Bloch vector (10), which does not depend on the prior density, is now $\boldsymbol{r}_{\mathrm{d}}(\eta) = \boldsymbol{r}_{\mathrm{d}}/\eta$: it contains *more* structure than $\boldsymbol{r}_{\mathrm{d}}$, since $\|\boldsymbol{r}_{\mathrm{d}}(\eta)\| \geq \|\boldsymbol{r}_{\mathrm{d}}\|$, in order to compensate for the loss of information during the measurement. This observation remains true for the more complicated tomographic methods discussed above, and invites the following distinction:

- For $\eta < 1$ we can interpret any tomographic $\boldsymbol{r}(\eta)$ as a *platonic state* representing the ideal of the system, which is poorly measured in our experiment using Equation (30). If we could perform a more accurate measurement, we would find a state more closely resembling this $\boldsymbol{r}(\eta)$.
- We can define a *positivist state* $\tilde{\boldsymbol{r}}(\eta) = \eta\boldsymbol{r}(\eta)$ that already includes the effects of imprecise measurements; experimental outcomes can be predicted in terms of perfect measurements of this positivist state, using Equation (9).

The author believes that platonic ideals such as $\boldsymbol{r}(\eta)$ should be discouraged in quantum mechanics, as they do not represent what can currently be measured, but instead hypothesize knowledge that may forever remain out of experimental reach. Instead, we suggest using the state $\tilde{\boldsymbol{r}}(\eta)$ as a fair representation of the experimenter's current and actual knowledge about the system.

3.2. Estimating the tomographic uncertainty

For every tomographic reconstruction, it is important to be able to give an estimate of the uncertainty of the resulting density matrix [29]. While the Bayesian mean of Section 2.4 gives such an estimate via Equation (27), shown as orange and cyan ellipses in Figures 1 and 2, the
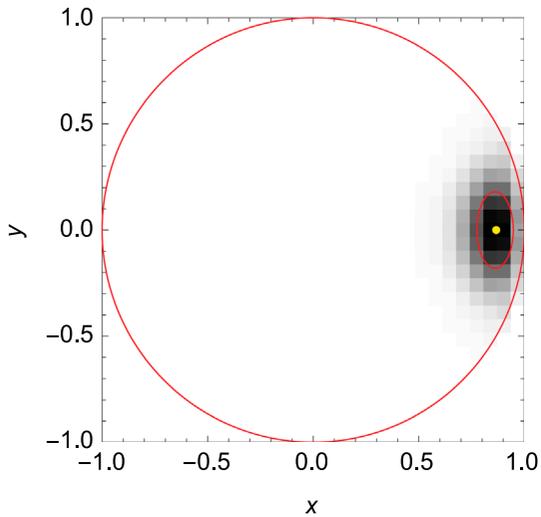
Figure 4. Bootstrapping (Section 3.2) the state $r = (13/15, 0, 0)$ (yellow dot) via scaled direct inversion tomography (Section 2.2.1). To generate this graphic, all 29,791 possible experimental outcomes $(N_{x\uparrow}, N_{x\downarrow}, N_{y\uparrow}, N_{y\downarrow}, N_{z\uparrow}, N_{z\downarrow})$ of performing $N_x = N_y = N_z = 30$ Stern–Gerlach measurements on the state $\hat{\rho}(r)$ were subjected to a tomographic state reconstruction, and the resulting Bloch vectors (projected into the xy plane) were plotted in a 2D histogram using weights from Equation (9), in 31×31 bins. The mean reconstructed state vector is at $(0.862 \pm 0.086, 0 \pm 0.180, 0 \pm 0.180)$, as listed in Table 1. The rms trace distance to r, Equation (31), is 0.135. (The colour version of this figure is included in the online version of the journal.)

other methods presented here do not give natural error estimates.

A widely used method for nonetheless finding such an error bar, called *bootstrapping* [30] or *case resampling*, goes as follows: once we have tomographically determined a density matrix from experimental data, we can use this density matrix to generate new "fake" data-sets using the same measurement operators and the probabilities of Equation (9) or (30); the argumentation is that in principle, each one of these fake data-sets could have been measured, instead of the set we have measured in reality. For each such fake data-set we can then do a tomography, and finally average any observables (or just the density matrix) over these tomographies. As is shown in Figure 4 and Table 1, this procedure can be used to calculate the covariance matrix of the components of the tomographic Bloch vector. While these covariances correctly estimate the uncertainty we are looking for [30], the bootstrap method contains systematic biases for the different tomographic methods, as shown in Table 1. Ideally, the weighted mean of all fake-data tomographies would be equal to the input state, such that we can use this technique to extract a covariance matrix without introducing a bias; however, this is not the case [17]. Nevertheless, the covariance matrix found in this way can still be used in order to get an idea of the tomographic uncertainty.

We find that the bootstrapped covariances of the Bloch vector components are slightly smaller than the covariances estimated with the Bayesian mean (Section 2.4) for a single experimental data-set. The similar magnitudes of these two sets of error estimates lead us to the conclusion that the bootstrap method can be a valid tool for estimating tomographic uncertainties. We believe that the cautionary footnote of Ref. [4] concerning the absurd results of bootstrapping for, e.g. $(N_x, N_y, N_z) = (0, 0, 1)$ do not apply when several non-commuting observables are measured, as we do in this text with $N_x = N_y = N_z \geq 1$.

Concerning the choice of input state r for the generation of fake data-sets, the *non-parametric* bootstrap method (direct re-sampling of measured data) requires us to use the direct-inversion Bloch vector r_d, Equation (10), even if $\|r_d\| > 1$ is unphysical. While this r_d-based non-parametric bootstrap is closest to the experimental data and may therefore be expected to be least biased, it is unrealistic in the case $\|r_d\| > 1$ because it neglects the physical condition that any density matrix used for predicting experimental outcomes, including generating fake data-sets, must be positive semi-definite. If we use a different Bloch vector, for example r_{MLE}, the method is called a *parametric bootstrap* and is physically better justified, albeit biased.

4. Comparison of tomography methods

In this section, we quantify the performance of the different tomographic methods discussed above. We use the following procedure, similar to suggestions in Refs. [4,5,19], to calculate an accuracy measure for each method:

(1) For a given input state r and a desired number of measurements along the Cartesian axes, in our case $N_x = N_y = N_z = 30$, we enumerate all $31^3 = 29,791$ possible experimental outcomes $\mathcal{N} = (N_{x\uparrow}, N_{x\downarrow}, N_{y\uparrow}, N_{y\downarrow}, N_{z\uparrow}, N_{z\downarrow})$ of Stern–Gerlach measurements, together with their probabilities $\mathcal{P}(\mathcal{N}|r)$ from Equation (9).

(2) For each possible experimental outcome $\mathcal{N}$ we reconstruct the tomographic Bloch vector $r_{\mathrm{tomo}}(\mathcal{N})$. This step is done differently for the various methods discussed in Section 2. As an example, Figure 4 shows a 2D histogram of the Bloch vectors reconstructed with scaled direct inversion, binned with their weight given in Equation (9). Table 1 compares the performance of the different tomographic methods when applied to the example of Figure 4.

(3) For each possible experimental outcome $\mathcal{N}$ we quantify the tomographic error through the trace

Table 1. The bootstrapping covariance ellipses of Figure 4 for the different tomography methods, using the exemplary input state $r = (13/15, 0, 0)$ and $N_x = N_y = N_z = 30$ measurements. For each method, the corresponding color in Figure 1 is indicated. The ellipses are centered at $\langle r_{\text{tomo}} \rangle = \{\langle x_{\text{tomo}} \rangle, 0, 0\}$ and have the given radii $\{\Delta x, \Delta y, \Delta z\}$ in the Cartesian directions. Δ_{tomo} gives the mean accuracy in terms of the rms trace distance to r, Equation (31). The most accurate methods (and up to 5% higher) are highlighted in green, and poorly performing methods are highlighted in red. The last column gives the failure rate for methods that do not always give a well-defined result.

Method and prior density			$\langle x_{\text{tomo}} \rangle$	Δx	$\Delta y, z$	Δ_{tomo}	P_{fail}
Scaled direct inversion (Section 2.2.1)		●	0.862	0.086	0.180	0.135	
Fisher information distance (Section 2.2.5)		●	0.866	0.091	0.168	0.127	5×10^{-10}
MLE:	$1 \le k < 2$ and Chernoff measure	●	0.924	0.045	0.269	0.193	3%
(Section 2.3)	$k = 2$ (Hilbert–Schmidt)	●	0.864	0.088	0.174	0.131	
	Chernoff with entropy weight	●	0.853	0.084	0.165	0.124	
	$k = \frac{3}{2}$ with entropy weight	●	0.844	0.085	0.160	0.122	
	$k = 2$ with entropy weight	●	0.816	0.083	0.149	0.116	
BME:	$k = 1$ (pure states)	○	0.907	0.044	0.224	0.161	
(Section 2.4)	Chernoff measure	○	0.842	0.101	0.167	0.129	
	$k = \frac{3}{2}$ (Bures measure)	○	0.830	0.077	0.162	0.122	
	$k = 2$ (Hilbert–Schmidt measure)	○	0.797	0.077	0.148	0.117	
	Chernoff with entropy weight	○	0.790	0.084	0.146	0.118	
	$k = \frac{3}{2}$ with entropy weight	○	0.781	0.076	0.142	0.116	
	$k = 2$ with entropy weight	○	0.756	0.075	0.136	0.117	

Table 2. Comparison of the accuracies $\Delta_{\text{tomo}}(r)$ of various tomography methods, Equation (31). For each method, the corresponding color in Figure 1 is indicated. For a given input state (Bloch vector r), $N_x = N_y = N_z = 30$ measurements are simulated along each Cartesian quantization axis, and all 29,791 possible experimental outcomes $\mathcal{N} = (N_{x\uparrow}, N_{x\downarrow}, N_{y\uparrow}, N_{y\downarrow}, N_{z\uparrow}, N_{z\downarrow})$ are fed into each tomography method (see Figure 4 for an example), in the same way as bootstrapping (Section 3.2 and Table 1); finally, the root-mean-square (rms) of the trace distances $\frac{1}{2}\|r - r_{\text{tomo}}\|$ of the results (see Section 2.2.1) are computed with weights from Equation (9). Smaller values indicate better accuracy of the tomography method; the most accurate methods for each input state (and up to 5% higher) are highlighted in green, and poor methods are highlighted in red. We show exemplary results for six input states r, as in Table 1, and the last column gives the rms of these accuracies averaged over all possible input states r with the given method's prior density using Equation (32). (*) The probability for these methods to give an ill-defined result was at most 0.2%, except where noted; the given mean values only include well-defined results. The strongly mixed states with large failure rates only contribute minimally to $\bar{\Delta}_{\text{tomo}}$ due to the given prior density weighting. (§) For the six pure states along the Cartesian axes, this method always gives exactly the correct result. (†) Averages were done with the Hilbert–Schmidt measure.

Method and prior density			$r = (0,0,0)$	$(0,0,\frac{1}{2})$	$(0,0,0.9)$	$(0,0,1)$	$(1,1,0)/\sqrt{2}$	$(1,1,1)/\sqrt{3}$	$\bar{\Delta}_{\text{tomo}}$
Scaled direct inversion (Section 2.2.1)		●	0.158	0.151	0.132	0.123	0.116	0.114	**0.137**[†]
Fisher information distance (Section 2.2.5)*		●	0.158	0.151	0.119	0[§]	0.126	0.123	**0.139**[†]
MLE:	$k = 1$ (pure states)*	●	37% failure	19% failure	3% failure	2% failure	0.113	0.107	**0.111**
(Section 2.3)	Chernoff measure* (23)	●	37% failure	19% failure	3% failure	2% failure	0.113	0.107	**0.167**
	$k = \frac{3}{2}$ (Bures measure)* (20)	●	37% failure	19% failure	3% failure	2% failure	0.113	0.107	**0.179**
	$k = 2$ (Hilbert–Schmidt) (13)	●	0.158	0.151	0.125	0.087	0.117	0.118	**0.137**
	Chernoff with entropy weight	●	0.158	0.150	0.118	0.087	0.118	0.121	**0.135**
	$k = \frac{3}{2}$ with entropy weight	●	0.156	0.148	0.116	0.087	0.120	0.124	**0.135**
	$k = 2$ with entropy weight	●	0.150	0.142	0.112	0.090	0.127	0.133	**0.135**
BME:	$k = 1$ (pure states)	○	0.443	0.306	0.145	0.086	0.111	0.109	**0.110**
(Section 2.4)	Chernoff measure	○	0.316	0.634	0.118	0.089	0.124	0.133	**0.274**
	$k = \frac{3}{2}$ (Bures measure)	○	0.154	0.149	0.116	0.090	0.121	0.125	**0.126**
	$k = 2$ (Hilbert–Schmidt)	○	0.148	0.141	0.112	0.095	0.131	0.136	**0.131**
	Chernoff with entropy weight	○	1.65	1.53	0.112	0.097	0.139	0.161	**1.08**
	$k = \frac{3}{2}$ with entropy weight	○	0.146	0.139	0.112	0.099	0.136	0.142	**0.132**
	$k = 2$ with entropy weight	○	0.141	0.134	0.115	0.106	0.144	0.151	**0.133**

distance $\frac{1}{2}\|r - r_{\text{tomo}}\|$ (see Section 2.2.2). We choose the trace distance here because it quantifies the experimental distinguishability of the two involved density matrices; but since all Schatten p-distances are equivalent for qubits, this is the same as the Hilbert–Schmidt distance quantifier of Ref. [19].

(4) We calculate the weighted root-mean-square (rms) trace distance of all possible experimental outcomes $\mathcal{N}$ with

$$\Delta_{\text{tomo}}(r) = \frac{1}{2}\left[\sum_{\mathcal{N}} \mathcal{P}(\mathcal{N}|r) \times \|r_{\text{tomo}}(\mathcal{N}) - r\|^2\right]^{1/2}.$$

(31)

In Table 2 these accuracies are shown for several tomographic methods and for several input states $\boldsymbol{r}$.

(5) These $\boldsymbol{r}$-dependent accuracies are further averaged using the prior density appropriate for each method,

$$\bar{\Delta}_{\text{tomo}} = \left[\int_{\|\boldsymbol{r}\| \leq 1} \mathcal{C}(\boldsymbol{r}) \Delta^2_{\text{tomo}}(\boldsymbol{r}) \mathrm{d}^3 \boldsymbol{r} \right]^{1/2}. \quad (32)$$

This is a single number characterizing a given combination of a tomographic method and a prior density, without any further parameters. In the last column of Table 2 we show these averaged accuracies for the methods considered here, where for methods with no inherent prior we use the Hilbert–Schmidt prior $\mathcal{C}_{\text{HS}}(\boldsymbol{r})$, Equation (13).

These tomographic accuracy quantifiers contain variance contributions from both the tomographic method in question and the randomness of the measurement process (quantum projection noise). However, as the latter is independent of the tomographic method, we can nonetheless use Equations (31) and (32) to compare the accuracies of different tomography methods with each other. Even though the differences in the accuracies are sometimes small, we therefore compare them carefully and hope to extrapolate the findings to higher-dimensional quantum state tomographies.

We make the following observations for the different tomographic methods:

- *Scaled direct inversion*   While this method never gives the most accurate results, it is very simple, never fails, and provides a baseline against which we can compare the more complex methods. The overall performance of this method is comparable to that of the MLE with Hilbert–Schmidt prior, since their results are very often the same.

- *Fisher information distance*   This method gives good results for mixed states, but for pure states its results are worse than those of the scaled direct inversion, except on the Cartesian axes where the results are perfect. This inconsistent behaviour leads us to discourage the use of this method.

- *MLE*   The maximum likelihood method fails for priors that are singular for pure states ($k = 1$, the Bures prior, and the Chernoff information prior) by not giving unique results. The frequently used Hilbert–Schmidt prior, on the other hand, performs well, comparable to the scaled direct inversion. When mixed states are known to predominate, adding an entropy weight gives even slightly more accurate results.

It is often argued that the maximum likelihood method with Hilbert–Schmidt prior (Section 2.2.4) is the "best" method since we cannot gain by giving an answer that is less likely, such as we do when giving a Bayesian mean estimate [19]. This argument is misleading, however. Firstly, the question of which prior density $\mathcal{C}(\hat{\rho})$ to use in the definition of the likelihood (12) is not answered to our satisfaction by tacitly using the Hilbert–Schmidt prior (13), especially in situations where the experimenter knows *a priori* that the generated states are nearly pure. The argument that the flatness of the Hilbert–Schmidt prior makes it most natural is contingent on the chosen parametrization, see Equation 21 and Ref. [31]. While the more natural Bures prior (20) fails to give satisfactory results, see Table 2, other priors are possible, and this degree of freedom casts at least some uncertainty on the optimality of the MLE method with HS prior. Secondly, as discussed below, other methods give *on average* more accurate tomographic results according to our quantifiers (31) and (32).

- *BME*   The BME is not only more plausible than the MLE because it is of full rank [4], but according to our overall quantifier (32) the BME is the most accurate method studied here (except when using the Chernoff information measure, see below). As we can see in Figures 1 and 2, the BME is strongly influenced by the choice of the prior density $\mathcal{C}(\hat{\rho})$: in general, the MLE is not even contained within the corresponding BME uncertainty ellipsoid. In Table 2 we see that in experiments where pure states are expected, using a pure-state ($k = 1$) prior gives the best results of our study; if the purity of the state is not known *a priori*, using a Bures or Hilbert–Schmidt prior is the optimal choice.

It may be surprising that we find the Bayesian mean estimate to be more accurate *on average* than the likelihood maximum, even though the former suffers from rather strong prior-dependent biases and the latter has been shown to be the most efficient estimation strategy [19]. This discrepancy comes from the observation, seen in Table 2, that while the BME is generally more accurate for mixed states, the MLE method is more accurate for pure states; however, in our averaging procedure, Equation (32), mixed states carry much more weight than the pure states near the surface of the sphere of Bloch vectors. In real experiments, the experimenter often tries to generate rather pure quantum states, and for these the MLE method indeed does give more accurate results if the Hilbert–Schmidt prior is assumed. However, we argue that in this case the Hilbert–Schmidt prior is not the correct one to use, but either the pure-state prior $\mathcal{C}_1(\boldsymbol{r})$ or the

Bures prior $\mathcal{C}_B(\boldsymbol{r})$, which correctly prioritize pure states; and in these cases, the BME *does* out-perform the MLE method [4], quantitatively for $\mathcal{C}_1(\boldsymbol{r})$ and qualitatively for $\mathcal{C}_B(\boldsymbol{r})$ (since in this case the MLE method always returns a pure state, which is not justified *a priori*).

We make the following observations for the different prior densities:

- $k = 1$ *pure states*   If we can be sure *a priori* that the experimental data have been generated by measurements on a pure state, then the BME method with prior $\mathcal{C}_1(\hat{\rho})$ is slightly more accurate on average than the MLE; also, the MLE method has a small probability of not giving a unique result at all. Therefore, the BME is preferred in this case, keeping in mind that the BME is never a pure state.
- $k = \frac{3}{2}$ *Bures measure*   On average, the BME method gives much more accurate results than the MLE method, and it never fails. For pure states, however, the MLE is more accurate than the BME; but in this case the pure-state prior is more appropriate.
- $k = 2$ *Hilbert–Schmidt measure*   On average, the BME method gives slightly more accurate results than the MLE method. Again, for pure states the MLE is more accurate than the BME; but in this case the Hilbert–Schmidt measure is an inappropriate choice.
- *Chernoff information measure*   While the Chernoff information measure gives good results for the MLE, comparable to those of the Bures measure, it gives very poor results for the BME of mixed states. The reason for this is that the Chernoff information measure is strongly peaked at pure states; but even for the BME of pure states its results are less accurate than those obtained with the Bures measure. For this reason, we discourage the use of the Chernoff information measure.
- *Entropy weights*   In general, both the MLE and the BME give very good results with entropy-weighted prior densities. An exception is the entropy-weighted Chernoff information measure, which gives very poor results for mixed states.

We conclude that the Bayesian mean estimate is the preferred method for single-qubit quantum state tomography. The instances where the maximum-likelihood method performs better, namely when pure states are reconstructed with the Hilbert–Schmidt or Bures prior, are not well justified since these priors are ill adapted to the experimental situation concerning pure states.

4.1. Extrapolation to larger systems

For systems with larger Hilbert spaces (dimension $D \gg 2$), the direct inversion result is very likely to be non-physical. The reason for this is that experimental quantum states are mostly of very low rank, and the tomographic estimates of the zero eigenvalues of the density matrix are statistically scattered around zero [32], with the probability of all of them being positive becoming exponentially small with the system dimension.

For this reason, the MLE becomes independent of the choice of prior, since any measure $\mathcal{C}_k$ for $k \leq D$, as well as the Bures measure, will yield the same rank-deficient result (the equivalent of a Bloch vector on the surface of the unit sphere for the single-qubit case, the grey dot in Figure 1). In this sense, the usual choice of the Hilbert–Schmidt measure $k = D$ [19] is valid.

For $D \gg 2$, the BME becomes computationally difficult to evaluate and must be calculated with a Monte Carlo algorithm. While in principle the BME is still the preferred method [4], such practical difficulties may discourage its use in large systems.

In our experimental practice with two-component Bose–Einstein condensates [3,33,34], we apply these insights and use the MLE with Hilbert–Schmidt prior for quantum state reconstruction, as is done in other groups [35]. We have found that this is the only computationally feasible and physically valid method for systems comprising hundreds or even thousands of particles, and consider it fortuitous that the present comparative study deems it appropriate. The problem of its general rank-deficiency is considered acceptable.

Notes

1. Even though the Kullback–Leibler divergence is not a distance because it is not symmetric in its arguments, we can still minimize it with respect to one of its arguments since it is a premetric.
2. The Bures measure for qubits (20) is the spherical equivalent of the Jeffreys prior of the Bernoulli trial, $\mathcal{C}(p) = 1/[\pi\sqrt{p(1-p)}]$.
3. $J_u(t)$ has a branch cut discontinuity at $t = 0$ for $u < -1$, and any results based on these values are ill-defined.

Acknowledgements

Disclosure statement

References

[1] Paris, M.; Řeháček, J. *Quantum State Estimation*; Lecture Notes in Physics 649; Springer: Berlin, Heidelberg, 2004.

[2] van Enk, S.J.; Blume-Kohout, R. *New J. Phys.* **2013,** *15,* 025024.

[3] Schmied, R.; Treutlein, P. *New J. Phys.* **2011,** *13,* 065019.

[4] Blume-Kohout, R. *New J. Phys.* **2010,** *12,* 043034.

[5] Ng, H.K.; Englert, B.-G. *Int. J. Quantum Inform.* **2012,** *10,* 1250038.

[6] Bagan, E.; Monras, A.; Muñoz-Tapia, R. *Phys. Rev. A* **2005,** *71,* 062318.

[7] Dowling, J.P.; Agarwal, G.S.; Schleich, W.P. *Phys. Rev. A* **1994,** *49,* 4101–4109.

[8] Pezzè, L.; Smerzi, A. *Phys. Rev. Lett.* **2009,** *102,* 100401.

[9] Řeháček, J.; Englert, B.-G.; Kaszlikowski, D. *Phys. Rev. A* **2004,** *70,* 052321.

[10] Scott, A.J. *J. Phys. A: Math. Gen.* **2006,** *39,* 13507–13530.

[11] Embacher, F.; Narnhofer, H. *Ann. Phys.* **2004,** *311,* 220–244.

[12] Mahler, D.H.; Rozema, L.A.; Darabi, A.; Ferrie, C.; Blume-Kohout, R.; Steinberg, A.M.; *Phys. Rev. Lett.* **2013,** *111,* 183601.

[13] Kravtsov, K.S.; Straupe, S.S.; Radchenko, I.V.; Houlsby, N.M.T.; Huszár, F.; Kulik, S.P. *Phys. Rev. A* **2013,** *87,* 062122.

[14] Massar, S.; Popescu, S. *Phys. Rev. Lett.* **1995,** *74,* 1259–1263.

[15] Keyl, M.; Werner, R.F. *Phys. Rev. A* **2001,** *64,* 052311.

[16] Newton, R.G.; Young, B.-L. *Ann. Phys.* **1968,** *49,* 393–402.

[17] Schwemmer, C.; Knips, L.; Richart, D.; Weinfurter, H.; Moroder, T.; Kleinmann, M.; Gühne, O. *Phys. Rev. Lett.* **2015,** *114,* 080403.

[18] Shang, J.; Ng, H.K.; Englert, G. *Quantum State Tomography: Mean Squared Error Matters, Bias Does Not,* 2014. arXiv:1405.5350 [quant-ph]

[19] Hradil, Z.; Řeháček, J.; Fiurášek, J.; et al. *Quantum State Estimation*; Lecture Notes in Physics 649, chap. 3; Springer: Berlin Heidelberg, 2004, pp. 59–112.

[20] Kullback, S.; Leibler, R.A. *Ann. Math. Stat.* **1951,** *22,* 79–86.

[21] Gourieroux, C.; Monfort, A. *Statistique et modèles économétriques*, Economica, Paris: 1989; Vol. 1.

[22] Sommers, H.-J.; Życzkowski, K. *J. Phys. A: Math. Gen.* **2004,** *37,* 8457–8466.

[23] Życzkowski, K.; Sommers, H.-J. *J. Phys. A: Math. Gen.* **2001,** *34,* 7111–7125.

[24] Hall, M.J.W. *Phys. Lett. A* **1998,** *242,* 123–129.

[25] Osipov, V.A.; Sommers, H.-J.; Życzkowski, K. *J. Phys. A: Math. Theor.* **2010,** *43,* 055302.

[26] Jeffreys, H. *Proc. R. Soc. London A* **1946,** *186,* 453–461.

[27] Sentís, G.; Bagan, E.; Calsamiglia, J.; Muñoz-Tapia, R. *Phys. Rev. A* **2010,** *82,* 042312.

[28] Nielsen, M.A.; Chuang, I.L. *Quantum Computation and Quantum Information*; Cambridge University Press, Cambridge, 2000.

[29] Christandl, M.; Renner, R. *Phys. Rev. Lett.* **2012,** *109,* 120403.

[30] Efron, B.; Tibshirani, R.J. *Stat. Sci.* **1986,** *1,* 54–77.

[31] Shang, J.; Ng, H.K.; Sehrawat, A.; Li, X.; Englert, B.-G. *New J. Phys.* **2013,** *15,* 123026.

[32] Wigner, E.P. *Ann. Math.* **1958,** *67,* 325–327.

[33] Riedel, M.F.; Böhi, P.; Li, Y.; Hänsch, T.W.; Sinatra, A.;Treutlein, P. *Nature* **2010,** *464,* 1170–1173.

[34] Ockeloen, C.F.; Schmied, R.; Riedel, M.F.; Treutlein, P. *Phys. Rev. Lett.* **2013,** *111,* 143001.

[35] Strobel, H.; Muessel, W.; Linnemann, D.; Zibold, T.; Hume, D.B.; Pezzè, L.; Smerzi, A.; Oberthaler, M.K.; *Science* **2014,** *345,* 424–427.

Appendix 1.   Choice of measurement axes

In this section, we motivate the set of Cartesian measurement axes used for spin-1/2 tomography in this work (see Section 1). We find the conditions under which a set of measurement axes can be used for efficient and accurate quantum state tomography.

A.1.   Direct inversion (filtered backprojection)

The direct inversion (filtered backprojection) technique of Ref. [3] is the simplest general tomographic method for quantum-mechanical spins of arbitrary length j. Given a true density matrix with spherical tensor coefficients ρ_{kq} and a set of M measurement axes (ϑ_n, φ_n) with measurement weights c_n, the average tomographically reconstructed spherical coefficients of the density matrix (averaged over all possible measurement results) are found by inserting Equations (6) into (4) of Ref. [3],

$$\langle \rho_{kq}^{(\mathrm{fbp})} \rangle = (2k+1) \sum_{n=1}^{M} c_n D_{q0}^k(\varphi_n, \vartheta_n, 0) \sum_{m=-j}^{j} p_m(\vartheta_n, \varphi_n) t_{k0}^{jmm}$$

$$= \sum_{q'=-k}^{k} \rho_{kq'} \times 4\pi \sum_{n=1}^{M} c_n \left[Y_k^q(\vartheta_n, \varphi_n) \right]^* Y_k^{q'}(\vartheta_n, \varphi_n). \quad \text{(A1)}$$

This is the correct result $\langle \rho_{kq}^{(\mathrm{fbp})} \rangle = \rho_{kq}$ if the measurement axis orientations (ϑ_n, φ_n) and their weights c_n satisfy

$$4\pi \sum_{n=1}^{M} c_n \left[Y_k^q(\vartheta_n, \varphi_n) \right]^* Y_k^{q'}(\vartheta_n, \varphi_n) = \delta_{qq'} \quad \text{(A2)}$$

for all $k = 0, 1, \ldots, 2j$ and $q, q' = -k, -k+1, \ldots, +k$. If we decompose the angular density of measurement axes into a sum of spherical harmonicswith coefficients

$$s_{kq} = \sum_{n=1}^{M} c_n \left[Y_k^q(\vartheta_n, \varphi_n) \right]^*, \quad \text{(A3)}$$

then Equation (A2) is satisfied whenever $s_{kq} = 0$ for all $k = 2, 4, 6, \ldots, 4j$.

For $j = 1/2$, as used in this work, this implies that direct inversion tomography [3] is correct on average if the distribution of measurement axes satisfies $s_{2,-2} = s_{2,-1} = s_{2,0} = s_{2,1} = s_{2,2} = 0$, i.e. if there is no quadrupolar anisotropy in the distribution of measurement axes. The smallest set of measurement axes that satisfies these conditions is the Cartesian-axes tomography set with equal numbers of measurements along each axis, as used in this text.

A.2. Maximum likelihood estimation

In contrast to direct inversion (Section A.1), maximum likelihood estimation is intrinsically biased due to the physicality constraint on the density matrix (see Table 1). We can nonetheless ask: what conditions must a set of measurement axes fulfil so that the mean tomographic estimate is closest to the true density matrix?

For the tomographic reconstruction of a completely mixed spin-1/2 state (Bloch vector $r = 0$), we can quantify this question by calculating the variance $\langle \| r_{\mathrm{MLE}} \|^2 \rangle$ of the resulting Bloch vector, averaged over all possible sets of experimental results. We find numerically that $\langle \| r_{\mathrm{MLE}} \|^2 \rangle$ is smallest if the distribution of measurement axes has $s_{2,-2} = s_{2,-1} = s_{2,0} = s_{2,1} = s_{2,2} = 0$, which is the same vanishing-quadrupole condition as found in Section A.1.