Information criteria for efficient quantum state estimation

J. O. S. Yin and S. J. van Enk

Information criteria for efficient quantum state estimation

J. O. S. Yin and S. J. van Enk

Oregon Center for Optics, Department of Physicso
University of Oregon
Eugene, OR 97403

Recently several more efficient versions of quantum state tomography have been proposed, with the purpose of making tomography feasible even for many-qubit states. The number of state parameters to be estimated is reduced by tentatively introducing certain simplifying assumptions on the form of the quantum state, and subsequently using the data to rigorously verify these assumptions. The simplifying assumptions considered so far were (i) the state can be well approximated to be of low rank, or (ii) the state can be well approximated as a matrix product state, or (iii) only the permutationally invariant part of the density matrix is determined. We add one more method in that same spirit: we allow in principle any model for the state, using any (small) number of parameters (which can, e.g., be chosen to have a clear physical meaning), and the data are used to verify the model. The proof that this method is valid cannot be as strict as in above-mentioned cases, but is based on well-established statistical methods that go under the name of "information criteria." We exploit here, in particular, the Akaike Information Criterion (AIC). We illustrate the method by simulating experiments on (noisy) Dicke states.

I. INTRODUCTION

Quantum state estimation [1–3] remains one of the hot topics in the field of quantum information processing. The hope to recover each element in the density matrix, however, is impeded by the exponential growth of the number of matrix elements with the number of qubits, and the concomitant exponential growth in time and memory required to compute and store the density matrix. The task can become intimidating when 14 qubits are involved [4], and so efforts have been made to simplify quantum state tomography. One such effort focused on states that have high purity [5] so that the size of the state space shrinks significantly (from $\mathcal{O}(D^2)$ to $\mathcal{O}(D)$ for a system described by a D dimensional Hilbert space). Given that the measurement record is used to verify the assumptions made initially, this method avoids the trap of simplification through imposing *a priori* assumptions merely by *fiat*. Another recent effort [6] in the same spirit considered multi-qubit states that are well represented by matrix product states [7–9] (which require a number of parameters growing only polynomially with the number of qubits). Many states of interest, such as ground states of certain model Hamiltonians in condensed-matter physics, are of that form. Crucially, the particular form of the state can be verified by the data. Somewhat similarly, a recent paper [10] considered estimating the permutationally invariant part of a density matrix. This, too, requires surprisingly few measurements, and is relevant as many states of interest are permutationally invariant. In fact, the states we will consider in this paper possess that property.

Here we go one step further, and we will allow, tentatively, *any* parametrized form for the density matrix of the quantum system to be tested, possibly containing just a few parameters. In fact, we may have several different tentative ideas of how our quantum state is best parameterized. The questions are then, how the data reveal which of those descriptions work sufficiently well, and which description is the best. This idea corresponds to a well-developed field in statistics: model selection [11–13]. All mathematical descriptions of reality are in fact models (and a quantum state, pure or mixed, is an excellent example of a model), and they can be evaluated by judging their performance relative to that of the true model (assuming it exists). In order to quantify this relative performance, we will make use of the Kullback-Leibler divergence (aka mutual information, aka cross entropy, aka relative entropy) [14], which has the interpretation of the amount of information lost when a specific model is used rather than the true model. Based on the minimization of the Kullback-Leibler divergence over different models, the Akaike Information Criterion (AIC) [15] was developed as a ranking system so that models are evaluated with respect to each other, given measurement data. The only quantities appearing in the criterion are the maximum likelihood obtainable with a given model (i.e., the probability the observed data would occur according to the model, maximized over all model parameters), and the number of independent parameters of the model.

The minimization does not require any knowledge of the true model, only that the testing model is sufficiently close to the true model. The legitimate application of AIC should, therefore, in principle be limited to "good" models, ones that include the true model (in our case, the exact quantum state that generated the data), at least to a very good approximation. However this does not prevent one from resorting to the AIC for model evaluation when there is no such guarantee. In fact, Takeuchi studied the case where the true model does not belong to the model set and came up with a more general criterion, named the Takeuchi Information Criterion, TIC [16]. However the estimation of the term introduced by Takeuchi to counterbalance the bias of the maximum likelihood estimator used in the AIC, requires estimation of a $K \times K$ matrix (K being the number of independent

parameters used by a model) from the data, which, unfortunately, is prone to significant error. This reduces the overall charm and practical use of the TIC. Since in most cases the AIC is still a good approximation to the TIC [12], especially in the case of many data, we stick to the simpler and more robust criterion here.

Information criteria are designed to produce a relative (rather than absolute) ranking of models, so that fixing a reference model is convenient. Throughout this paper we choose the "full-parameter model" (FPM) as reference, that is, a model with just enough independent variables to fully parameterize the measurement on our quantum system. For tomographically complete measurements (discussed in detail in Sec. III B) the number of independent variables is given by the number of free parameters in the density matrix ($D^2 - 1$ for a D-dimensional Hilbert space). For tomographically incomplete measurements (see Sec. III D), the number of independent variables of FPM is smaller, and equals the number of independent observables. We will, in fact, not even need the explicit form of the FPM (which may be hard to construct for tomographically incomplete measurements), as its maximum possible likelihood can be easily upper-bounded.

We should note an important distinction between maximum likelihood estimation (MLE) [17], a technique often used in quantum tomography, and the method of information criteria and model selection. MLE produces the state that fits the data best. Now the data inevitably contains (statistical) noise, and the MLE state predicts, incorrectly, that same noise to appear in future data. Information criteria, on the other hand, have been designed to find the model that best predicts future data, and tries, in particular, to avoid overfitting to the data, by limiting the number of model parameters. This is how a model with a few parameters can turn out to be the best predictive model, even if, obviously, the MLE state will fit the (past) data better.

We also note that information criteria have been applied mostly in areas of research outside of physics. This is simply due to the happy circumstance that in physics we tend to know what the "true" model underlying our observations is (or should be), whereas this is much less the case in other fields. Within physics, information criteria have been applied to astrophysics [18], where one indeed may not know the "true" model (yet), but also to the problem of entanglement estimation [19]. In the latter case (and in quantum information theory in general) the problem is not that we do not know what the underlying model is, but that that model may contain far too many parameters. Hence the potential usefulness of information criteria. And as we recently discovered, the AIC has even been applied to quantum state estimation, not for the purpose of making it more efficient, but making it more accurate, by avoiding overfitting [20].

II. THE AKAIKE INFORMATION CRITERION - A SCHEMATIC DERIVATION

Suppose we are interested in measuring certain variables, summarized as a vector $\mathbf{x}$, and their probability of occurrence as outcome of our measurement. We denote $f(\mathbf{x})$ as the probabilistic model that truthfully reflects reality (assuming for convenience that such a model exists) and $g(\mathbf{x}|\vec{\theta})$ as our (approximate) model characterized by one or more parameters, summarized as a vector $\vec{\theta}$. The models satisfy the normalization condition $\int \mathrm{d}\mathbf{x} f(\mathbf{x}) = \int \mathrm{d}\mathbf{x} g(\mathbf{x}|\vec{\theta}) = 1$ for all $\vec{\theta}$. By definition, we say there is no information lost when $f(\mathbf{x})$ is used to describe reality. The amount of information lost when $g(\mathbf{x}|\vec{\theta})$ is used instead of the true model is defined to be the Kullback-Leibler divergence [14] between the model $g(\mathbf{x}|\vec{\theta})$ and the true model $f(\mathbf{x})$:

$$I(f, g_{\vec{\theta}}) = \int \mathrm{d}\mathbf{x} f(\mathbf{x}) \log(f(\mathbf{x}))$$
$$- \int \mathrm{d}\mathbf{x} f(\mathbf{x}) \log(g(\mathbf{x}|\vec{\theta})). \quad (1)$$

Eq. (1) can be conveniently rewritten as

$$I(f, g_{\vec{\theta}}) = E_{\mathbf{x}} \left[\log(f(\mathbf{x}))\right] - E_{\mathbf{x}} \left[\log(g(\mathbf{x}|\vec{\theta}))\right], \quad (2)$$

where $E_{\mathbf{x}}[\cdot]$ denotes an estimate with respect to the true distribution $f(\mathbf{x})$. We see that $\mathbf{x}$ is no longer a variable in the above estimator, as we integrated it out. The only variable that affects $I(f, g_{\vec{\theta}})$ is $\vec{\theta}$. Since the first term in Eq. (1) is irrelevant to the purpose of rank-ordering different models g (not to mention we cannot evaluate it when f is not known), we only have to consider the second term. Suppose there exists $\vec{\theta}_0$ such that $g(\mathbf{x}|\vec{\theta}_0) = f(\mathbf{x})$ for every $\mathbf{x}$, that is, the true model is included in the model set. Note that for this to hold, $\vec{\theta}$ does not necessarily contain the same number of parameters as the dimension of the system. To use a simpler notation without the integration over $\mathbf{x}$ we denote the second term in Eq. (1) (without the minus sign) as

$$S(\vec{\theta}_0 : \vec{\theta}) = \int \mathrm{d}\mathbf{x} g(\mathbf{x}|\vec{\theta}_0) \log(g(\mathbf{x}|\vec{\theta})), \quad (3)$$

where we have used $g(\mathbf{x}|\vec{\theta}_0)$ to represent the true model $f(\mathbf{x})$. The advantage of this estimator is that it can be approximated without knowing the true distribution $f(\mathbf{x})$. To do that we first consider the situation where $\vec{\theta}$ is close to $\vec{\theta}_0$. This assumption can be justified in the limit of large N, N being the number of measurement records, since the model $\vec{\theta}$ ought to approach $\vec{\theta}_0$ asymptotically (assuming, for simplicity, $\vec{\theta}_0$ is unique). We know that $S(\vec{\theta}_0 : \vec{\theta})$ must have a maximum when $\vec{\theta} = \vec{\theta}_0$, and we may then symbolically expand $S(\vec{\theta}_0 : \vec{\theta})$ in the vicinity of

$\vec{\theta}_0$ by

$$S(\vec{\theta}_0 : \vec{\theta}) = S(\vec{\theta}_0 : \vec{\theta}_0) - \frac{1}{2}||\vec{\theta} - \vec{\theta}_0||^2_{\vec{\theta}_0}$$
$$+ \mathcal{O}\left(||\vec{\theta} - \vec{\theta}_0||^{3/2}_{\vec{\theta}_0}\right), \qquad (4)$$

where

$$||\vec{\theta} - \vec{\theta}_0||^2_{\vec{\theta}_0} = \left(\vec{\theta} - \vec{\theta}_0\right)' \cdot \left.\frac{\partial^2 S(\vec{\theta}_0 : \vec{\theta})}{\partial\vec{\theta}^2}\right|_{\vec{\theta}=\vec{\theta}_0} \cdot \left(\vec{\theta} - \vec{\theta}_0\right)$$
$$(5)$$

denoting a *squared length* derived from a metric defined at $\vec{\theta}_0$. It can be proved that when N is sufficiently large $||\vec{\theta} - \vec{\theta}_0||^2_{\vec{\theta}_0}$ can be approximated by the χ^2_K distribution, with K equal to the number of independent parameters used by the model $\vec{\theta}$. From the properties of the χ^2_K distribution, we know the average value of $||\vec{\theta} - \vec{\theta}_0||^2_{\vec{\theta}_0}$ will approach K.

The next step is to evaluate the estimator $S(\vec{\theta}_0 : \vec{\theta}_0)$, where $\vec{\theta}_0$ is now considered a variable. Suppose we find the maximum likelihood estimate $\vec{\theta}_M$ from the measurement outcomes such that $S(\vec{\theta}_M : \vec{\theta}_M)$ is the maximum. Now $\vec{\theta}_M$ should also be close to the true model $\vec{\theta}_0$, when N is sufficiently large. Therefore we can similarly expand $S(\vec{\theta}_0 : \vec{\theta}_0)$ in the vicinity of $\vec{\theta}_M$ as

$$S(\vec{\theta}_0 : \vec{\theta}_0) = S(\vec{\theta}_M : \vec{\theta}_M) - \frac{1}{2}||\vec{\theta}_0 - \vec{\theta}_M||^2_{\vec{\theta}_M}$$
$$+ \mathcal{O}\left(||\vec{\theta} - \vec{\theta}_0||^{3/2}_{\vec{\theta}_M}\right), \quad (6)$$

$||.||_{\vec{\theta}_M}$ is a length similarly defined as in Eq. (4) and has the same statistical attributes as $||\vec{\theta} - \vec{\theta}_0||^2_{\vec{\theta}_0}$ since $\vec{\theta}_0$ is related to $\vec{\theta}_M$ the same way $\vec{\theta}$ is related to $\vec{\theta}_0$ and $\vec{\theta}_M$ is very close to $\vec{\theta}_0$. Its average value, therefore, approaches again K, according to the χ^2_K distribution. Thus we are able to rewrite Eq. (3) as

$$S(\vec{\theta}_0 : \vec{\theta}) \approx S(\vec{\theta}_M : \vec{\theta}_M) - K. \qquad (7)$$

We see that now our target estimator $S(\vec{\theta}_0 : \vec{\theta})$ is evaluated by the MLE solution $\vec{\theta}_M$ only (plus the number of parameters K of the model), with no knowledge of what the true model f is. The assumption that underlies this convenience is constituted by two parts: estimating $S(\vec{\theta}_0 : \vec{\theta})$ with its maximum $\vec{\theta}_0$ and estimating $S(\vec{\theta}_0 : \vec{\theta}_0)$ from the data by its optimum $\vec{\theta}_M$. The deviations from their respective maxima are equal and result simply in the appearance of the constant K.

We now denote $\mathrm{L}_M = S(\vec{\theta}_M : \vec{\theta}_M)$, which is the maximum likelihood obtainable by our model, with respect to a given set of measurement records. The indicator of the AIC is denoted by Ξ, which is defined by

$$\Xi = -2\mathrm{L}_M + 2K. \qquad (8)$$

Apart from the conventional factor 2, and a constant independent of the model $\vec{\theta}$, Ξ is an estimator of the quantity in Eq. (1) we originally considered, that is, the Kullback-Leibler divergence between a model that is used to describe the true model and the true model itself. Therefore a given model is considered better than another if it has a *lower* value of Ξ.

Finally, in the case that N is not so large yet that asymptotic relations hold to a very good approximation, one can include a correction factor to the AIC taking the deviation from asymptotic values into account. The corrected AIC gives rise to a slightly different criterion [21]:

$$\Xi_c = -2\mathrm{L}_M + 2K + \frac{2K(K+1)}{N-K-1}. \qquad (9)$$

III. RESULTS

A. Dicke states

We will apply the AIC to measurements on a popular family of entangled states, the Dicke states of four qubits [22–26]. We simulate two different experiments, one tomographically complete experiment, another measuring an entanglement witness. We include imperfections of a simple type, and we investigate how model selection, according to the AIC, would work. We consider cases where we happen to guess the correct model, as well as cases where our initial guess is, in fact, incorrect.

We consider the four-qubit Dicke states with one or two excitations $\left|D_4^{1,2}\right\rangle$ (with the state $|1\rangle$ representing an excitation):

$$\left|D_4^1\right\rangle = \left(|0001\rangle + |0010\rangle + |0100\rangle + |1000\rangle\right)/2, \quad (10a)$$
$$\left|D_4^2\right\rangle = \left(|0011\rangle + |0101\rangle + |0110\rangle + |1001\rangle\right.$$
$$\left. + |1010\rangle + |1100\rangle\right)\sqrt{6}. \qquad (10b)$$

For simplicity, let us suppose that white noise is the only random noise in the state generation, and that it corresponds to mixing of the ideal state with the maximally mixed state of the entire space (instead of the subspace with exactly one or two excitations, which could be a reasonable choice, too, depending on the actual implementation of the Dicke states). We thus write the states under discussion as

$$\rho^{1,2}(\alpha) = (1 - \alpha)\left|D_4^{1,2}\right\rangle\left\langle D_4^{1,2}\right| + \alpha\mathbb{1}/D, \qquad (11)$$

where $\mathbb{1}/D$ is the maximally mixed state for dimension $D = 2^4$, and $0 \le \alpha \le 1$. We will fix the actual states generating our data to be

$$\rho^{1,2}_{\mathrm{actual}} = \rho^{1,2}(\alpha = 0.2). \qquad (12)$$

This choice is such that the mixed state is entangled (as measured by our multi-qubit version of the negativity,

see below), even though the entanglement witness whose measurement we consider later in Sec. III D, just fails to detect it.

For our first model (to be tested by AIC) we wish to pick a one-parameter model (so, $K = 1$) that also includes a wrong guess. A straightforward model choice, denoted by $M_{1\phi}^{1,2}$, is

$$M_{1\phi}^{1,2} : \rho_\phi^{1,2}(q) = (1 - q) \left| \Psi_{\text{target}}^{1,2}(\phi) \right\rangle$$
$$\left\langle \Psi_{\text{target}}^{1,2}(\phi) \right| + q \mathbb{1}/D. \qquad (13)$$

We refer to the pure states appearing here as the *target* states $\left| \Psi_{\text{target}}^{1,2}(\phi) \right\rangle$, simulating the case where we (possibly incorrectly) think we would be creating a pure state of that form, if only the white noise were absent ($q = 0$). The phase ϕ is included not as a (variable) parameter of the model but as an inadvertently mis-specified property. In this case, it stands for us being wrong about a single relative phase in one of the qubits in state $|1\rangle$. Without loss of generality we assume the first qubit in our representation to carry the wrong phase, and we write

$$\left| \Psi_{\text{target}}^1(\phi) \right\rangle = \frac{1}{2} \left(|0001\rangle + |0010\rangle + |0100\rangle \right.$$
$$\left. + e^{i\phi} |1000\rangle \right), \qquad (14a)$$
$$\left| \Psi_{\text{target}}^2(\phi) \right\rangle = \frac{1}{\sqrt{6}} \left[|0011\rangle + |0101\rangle + |0110\rangle \right.$$
$$\left. + e^{i\phi} \left(|1001\rangle + |1010\rangle + |1100\rangle \right) \right]. \qquad (14b)$$

Alternatively, if we *do* consider this a two-parameter model (changing $K = 1$ to $K = 2$), then ϕ is variable, and we would optimize over ϕ. In our case, this optimum value should always be close to $\phi = 0$.

In order to avoid confusion, let us note that we consider model $M_{1\phi}^1$ when we simulate measurements on $\rho^1(\alpha = 0.2)$, and we consider model $M_{1\phi}^2$ when simulating measurements on $\rho^2(\alpha = 0.2)$.

B. Tomographically complete measurement

We first consider a tomographically complete measurement, in which a so-called SIC-POVM (symmetric informationally complete positive operator values measure [27]) with 4 outcomes is applied to each qubit individually. We first test our one-parameter model, and compare it to the FPM, which contains $255 (= 4^4 - 1)$ parameters, which is the number of parameters needed to fully describe a general state of 4 qubits. With definition Eq. (8) we have

$$\Xi(M_{1\phi}^{1,2}) = -2\mathrm{L}_M(M_{1\phi}^{1,2}) + 2, \qquad (15)$$

since $K = 1$ for $M_{1\phi}^{1,2}$. For the FPM we have

$$\Xi(\text{FPM}) = -2\mathrm{L}_M(\text{FPM}) + 2 \times 255, \qquad (16)$$
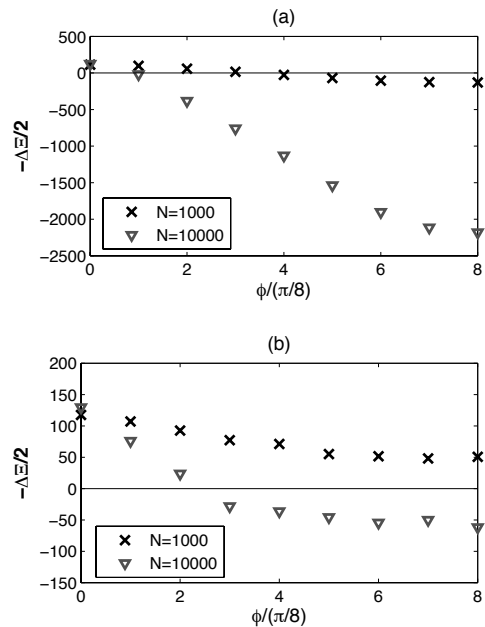

FIG. 1: *How AIC ranks the one- and two-parameter models vs. the full-parameter model (FPM):*
Plot of the difference between Ξ values of our models, the one-parameter model $M_{1\phi}^1$ (a) defined in (13) and the two-parameter model $M_{2\phi}^1$ (b) defined in (18), and the FPM, *i.e.*, $-\Delta\Xi = \Xi(\text{FPM}) - \Xi(M_{1\phi}^1)$ or $-\Delta\Xi = \Xi(\text{FPM}) - \Xi(M_{2\phi}^1)$, for various numbers of SIC-POVM measurements, N, with $\left| \Psi_{\text{target}}^1 \right\rangle$ as the target state, as functions of the angle ϕ (in units of $\pi/8$). The horizontal line demarcates $\Delta\Xi = 0$: points above (below) that line correspond to cases where the model with fewer (more) parameters is preferred. The figures with $\left| \Psi_{\text{target}}^2 \right\rangle$ as the target state look very similar (see FIG. 2 for an example of this similarity).

where $\mathrm{L}_M(\text{FPM})$ is the log of the maximum likelihood obtainable by the FPM. The latter can be bounded from above by noting that the best possible FPM would generate probabilities that exactly match the actual observed frequencies of all measurement outcomes. In the following we will always use that upper bound, rather than the actual maximum likelihood. Even though it is possible to find the maximum likelihood state in principle (and even in practice for small enough Hilbert spaces), we are only concerned with the FPM's ranking according to the AIC, which does not require its density matrix representation. For $M_{1\phi}^{1,2}$ to beat the FPM we require

$$-\Delta\Xi := \Xi(\text{FPM}) - \Xi(M_{1\phi}^{1,2}) > 0. \qquad (17)$$

This is a sufficient but not necessary requirement, as we use the above-mentioned upper bound to the FPM likelihood.

We plot the difference $\Delta\Xi$ between the two rankings in FIG. 1(a) for various values of the number of measurements, and for various values of the phase ϕ. We observe the following: The simple model is, correctly, judged bet-
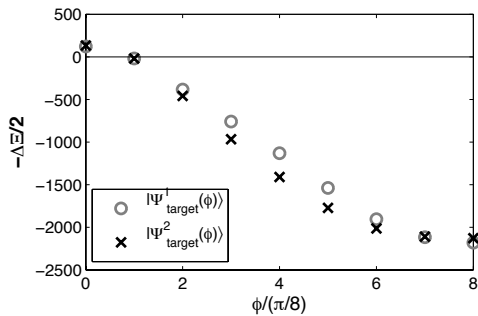
FIG. 2: *Comparing single- and double-excitation Dicke states*: The difference between Ξ's of $M_{1\phi}^{1,2}$ and the FPM as functions of ϕ (in units of $\pi/8$) for N SIC-POVM measurements. There are two different simulations here, one on a noisy Dicke state with 1 excitation using model $M_{1\phi}^{1}$ and another on a noisy Dicke state with two excitations using model $M_{1\phi}^{2}$. Both simulations contain $N = 10000$ measurements. The horizontal line demarcates $\Delta\Xi = 0$.

ter than the FPM when the phase ϕ is sufficiently small. The more measurements one performs, the smaller ϕ has to be for the AIC to still declare the model superior to the FPM (i.e., for the points to stay above the solid line, at $\Delta\Xi = 0$).

Although the correction to the AIC mentioned in Eq. (9) is not very small for the FPM for $N = 1000$, applying that correction still does not shift the second and third point below zero: that is, $N = 1000$ measurements is still not sufficiently large for the Ξ_c to recognize that $\phi = \pi/4$ and $\phi = \pi/2$ are incorrect guesses. One can argue about what the cause of this is: it could be that N is just too small for the derivation of the Ξ (or even the Ξ_c) to be correct. Or it could be that the AIC ranking is unreliable because the assumption that the true model is included in the model, is violated. Or it could be that, even with a perfectly valid criterion (perhaps the TIC), the statistical noise present in the data would still be too large.

If we consider the phase ϕ as a second (variable) parameter (thus creating a two-parameter model), then we can give FIG. 1 a different interpretation: we would pick $\phi = 0$ as the best choice, and we would increase K by 1. The latter correction is small on the scale of the plots, and so we find the two-parameter model to be superior to the one-parameter model for any nonzero plotted value of ϕ, and to the FPM. This is a good illustration of the following rather obvious fact: even if one has the impression that a particular property of one's quantum source is (or ought to be) known, it still might pay off to represent that property explicitly as a variable parameter (at the small cost of increasing K by 1), and let the data determine its best value.

C. Cross modeling

Suppose one picked a one-parameter model with a wrong (nonzero) value of ϕ, and the AIC has declared the model to be worse than the FPM. How can one improve the model in a systematic way when one lacks a good idea of which parameters to add to the model (we assume we already incorporated all parameters deemed important *a priori*). Apart from taking more and different measurements, one could use a hint from the existing data. One method making use of the data is to apply "cross modeling," where half the data is used to construct a modification to the model, and the remaining half is used for model validation, again by evaluating Ξ on just that part of the data. So suppose N measurements generate a data sequence $\mathcal{F} = \{f_1, f_2, ..., f_N\}$. One takes, *e.g.*, the first $N/2$ data points, $\{f_1, ..., f_{N/2}\}$, as the training set, and acquires the MLE state ρ_{MLE}, or a numerically feasible approximation thereof, with respect to the training set. We then create a model with two parameters like so:

$$M_{2\phi}^{1,2} : \rho_\phi(\epsilon, q) = (1 - \epsilon)\left[(1 - q)\rho_{\mathrm{MLE}}\right.$$
$$\left. + q\left|\Psi_{\mathrm{target}}^{1,2}(\phi)\right\rangle \left\langle \Psi_{\mathrm{target}}^{1,2}(\phi)\right|\right] + \epsilon\mathbb{1}/D. \quad (18)$$

For practical reasons ρ_{MLE} does not need to be strictly the MLE state, in particular when the dimension of the full parameter space is large. One would only require it to explain $\{f_1, ..., f_{N/2}\}$ well enough to make sure that part of the data is properly incorporated in the model. Thus, one could, for example, use one of the numerical shortcuts described in [28]. The rest of the data $\{f_{N/2+1}, ..., f_N\}$ is used to evaluate $M_{2\phi}^{1,2}$ against the FPM.

We note the resemblance of this procedure with the method of "cross-validation" [29]. In cross-validation one tries to find out how well a *given* predictive model performs by partitioning the data set into training set and validation set (exactly the same idea as given above). One uses multiple different partitions, and the results are averaged and optimized over those partitions. It can be shown [30] that under certain conditions cross-validation and the AIC are asymptotically equivalent in model selection. This virtually exempts one from having to check multiple partitions of the data set, by applying the AIC to the whole data set.

It is worth emphasizing that what we do here is different in two ways. First, our model is not fixed but modified, based on information obtained from one half of the data. Second, we partition the data set only once, and the reason is, that it would be cheating to calculate the (approximate) MLE state of the full set of data (or, similarly, check many partitions and average), and then consider the resulting MLE state a parameter-free model.

FIG. 1(b) shows results for $M_{2\phi}^{1}$ and SIC-POVM measurements. When the number of measurements is $N = 1000$, all $M_{2\phi}^{1}$ models are considered better than the FPM, regardless of the phase error ϕ assumed for the
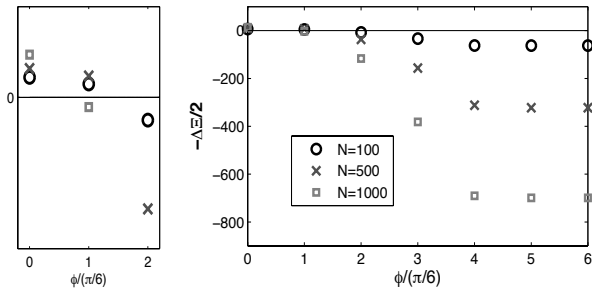
FIG. 3: *How the AIC ranks our one-parameter model vs. the FPM for an entanglement witness measurement*: The difference between Ξ's of $M_{1\phi}^2$ (with $\left|\Psi_{\text{target}}^2(\phi)\right\rangle$) and the FPM, i.e., $-\Delta\Xi = \Xi(\text{FPM}) - \Xi(M_{1\phi}^2)$, for different numbers of witness measurements, as functions of ϕ. The horizontal line demarcates $\Delta\Xi = 0$. The left figure zooms in on the leftmost three sets of data points in the right figure.

FIG. 4: *Does the witness $W_{J_{xy}}$ detect entanglement if there is a phase error?*: Witness performance $\langle W_{J_{xy}}\rangle$ for different states (defined as in Eq. (20)) as a function of ϕ. A negative expectation value detects entanglement.

target state. The reason is that around $\phi = \pi/2$ the approximate MLE state obtained from the first half of the data is able to "predict" the measurement outcomes (including their large amount of noise!) on the second half better than the 1-parameter model with the wrong phase.

On the other hand, when $N = 10000$ the AIC recognizes only the simple models with small phase errors ($\phi = 0, \pi/8, \pi/4$) as better than the FPM. So, neither the approximate MLE state, nor the 1-parameter model with wrong phase are performing well. This indicates how many measurements are needed to predict a single phase to a given precision.

D. Witness measurement

For states that are close to symmetric Dicke states $\left|D_N^{N/2}\right\rangle$, their entanglement can be verified by using measurements that require only two different local settings, *e.g.*, spins (or polarizations) either *all* in the x-direction or *all* in the y-direction. In particular, when $N = 4$, an efficient witness is $W_{Jxy} = 7/2 + \sqrt{3} - J_x^2 - J_y^2$ [31], where $J_{x,y} = \sum_j \sigma_{x,y}^{(j)}/2$, with $\sigma_{x,y}^{(j)}$ the Pauli matrices for the j-th subsystem. This witness detects (by having a negative expectation value) Dicke states with a white noise background, *i.e.*, $\rho(\alpha) = (1-\alpha)\left|D_4^2\right\rangle\left\langle D_4^2\right| + \alpha\mathbb{1}/D$ whenever $0 \le \alpha < 0.1920$.

So we suppose we perform $N/2$ measurements on all of the four spins in the x-direction simultaneously, and another $N/2$ similar measurements in the y-direction. Instead of calculating the witness W_{Jxy} and ending up with one single value determining entanglement, we make use of the full record of all individual outcomes in order to evaluate (and then maximize) likelihoods. For example, for the measurement of all four spins in the x-direction simultaneously, we can count the number
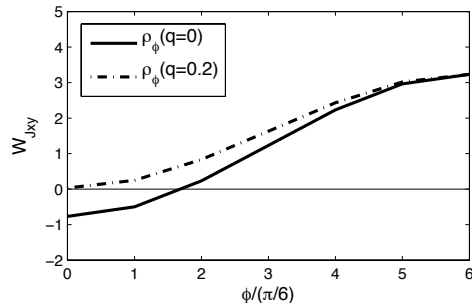
of times they are projected onto the $|x+x+x+x+\rangle$ state, the $|x+x+x+x-\rangle$ state, etc. In both x- or y-directions, the number of independent observables (i.e., the number of independent joint expectation values) is 15, which can be seen as follows: Any density matrix of M qubits can be expressed in terms of the expectation values of 4^M tensor products of the 3 Pauli operators and the identity $\mathbb{1}$, but the expectation value of the product of M identities equals 1 for any density matrix, thus leading to $4^M - 1$ independent parameters encoded in a general density matrix. From having measured just σ_x on all M qubits, we can evaluate all expectation values of all operators that are tensor products of σ_x and the identity. There are 2^M such products, and subtracting the trivial expectation value for $\mathbb{1}^{\otimes M}$ leaves $2^M - 1$ independent expectation values.

This means it only takes $2 \times 15 = 30$ independent parameters to form the FPM, and we have $K = 30$. Similar to the tomographically complete case, we do not need the concrete form of the whole 255-30 dimensional manifold of MLE states, nor do we need to explicitly parameterize the 30-parameter FPM states, as we can simply upper bound the maximum likelihood for this model, $\mathbb{L}_M(\text{FPM})$, by noting the best one could possibly do is reproduce exactly the observed frequencies of all possible measurement outcomes.

E. Estimating entanglement

Our state $\rho_{\text{actual}} = \rho(\alpha = 0.2)$ is just not detected by the witness W_{Jxy}, but still contains a considerable amount of entanglement. We choose to quantify this entanglement by means of three entanglement monotones (of which only two are independent), simply constructed from all bipartite negativities. If the four parties are denoted A, B, C and D, the generalized negativities [32–34]
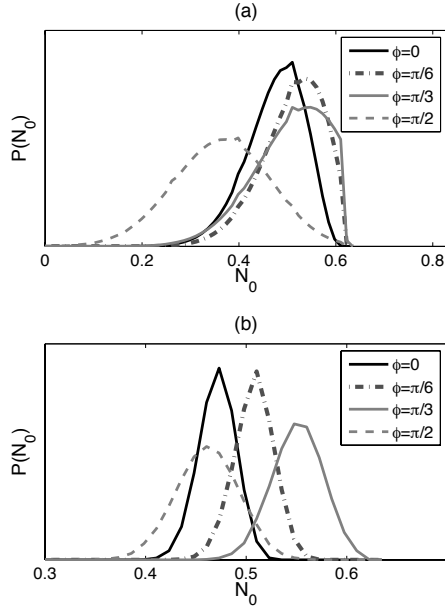
FIG. 5: *How one quantifies entanglement from a witness measurement*: The posterior probability distributions of $\mathcal{N}_0$ for different number of witness measurements ($N = 100$ in (a), $N = 1000$ in (b)), from model $M_{1\phi}^2$ (with target state $\left|\Psi_{\text{target}}^2\right\rangle$), where $\phi = 0, \pi/6, \pi/3$. $\mathcal{N}_0(\rho_{\text{actual}}) = 0.4770$. The posterior distributions are obtained by binning the results into 50 equal-sized bins from the minimum to the maximum value of $\mathcal{N}_0$ for each distribution separately. The prior distribution is assumed to be uniform on [0,1] for both ϵ and q. The distributions of $\mathcal{N}_1$ and $\mathcal{N}_2$ are similar (up to a simple shift).

are defined as

$$\mathcal{N}_1 = (\mathcal{N}_{AB-CD}\mathcal{N}_{AC-BD}\mathcal{N}_{AD-BC})^{1/3}, \tag{19a}$$

$$\mathcal{N}_2 = (\mathcal{N}_{A-BCD}\mathcal{N}_{B-CDA}\mathcal{N}_{C-DAB}\mathcal{N}_{D-ABC})^{1/4}, \tag{19b}$$

$$\mathcal{N}_0 = \left(\mathcal{N}_1^3\mathcal{N}_2^4\right)^{1/7}, \tag{19c}$$

where $\mathcal{N}_{AB-CD}$ denotes the negativity with respect to partition AB against CD, etc. The main advantage of the generalized negativities is that they are all efficiently computable directly from the density matrix. We have for our state $\mathcal{N}_1 = 0.6293$, $\mathcal{N}_2 = 0.3875$, and $\mathcal{N}_0 = 0.4770$.

Similarly to the tomographically complete case, we first consider the following one-parameter model:

$$M_{1\phi}^2: \ \rho_\phi(q) = (1-q)\left|\Psi_{\text{target}}^2(\phi)\right\rangle$$
$$\left\langle\Psi_{\text{target}}^2(\phi)\right| + q\mathbb{1}/D, \tag{20}$$

where $\left|\Psi_{\text{target}}^2(\phi)\right\rangle$ is defined in Eg. (14b). Ξ's for $M_{1\phi}^2$ and FPM are

$$\Xi(M_{1\phi}^2) = -2\text{L}_M(M_{1\phi}^2) + 2, \tag{21}$$

$$\Xi(\text{FPM}) = -2\text{L}_M(\text{FPM}) + 2 \times 30. \tag{22}$$

FIG. 3 shows that, as before, the marks above the horizontal solid line correspond to models deemed better than FPM. Compared to the case of full tomography (FIG. 1(a)), here the value of $\Xi(M_{1\phi}^2)$ is larger than $\Xi(\text{FPM})$ by a much smaller amount, even when the phase term is correct ($\phi = 0$). The absolute value of the difference is not relevant, though, and what counts is its sign. The obvious reason for the smaller difference is that the number of independent parameters for the FPM has dropped from 255 to 30. In addition, the FPM in this case does not refer to a specific 30-parameter model. On the contrary, since the number of degrees of freedom of the quantum system is still 255, there is a whole subspace of states, spanning a number of degrees of freedom equal to 225 (=255-30), all satisfying the maximum likelihood condition.

The witness measurement is very sensitive to the phase error, even when the number of measurements is still small. When $N = 1000$, the estimation of ϕ is within an error of $\pi/6$, as the second point plotted is already below the line $\Delta\Xi = 0$. Compared to FIG. 1(a), this precision is only reached when $N = 10000$.

An interesting comparison can be made between AIC and the entanglement-detecting nature of witness $W_{J_{xy}}$. FIG. 4 shows the performance of $\langle W_{J_{xy}}\rangle$ for the pure state $\left|\Psi_{\text{target}}^2(\phi)\right\rangle$ ($\rho_\phi(q = 0)$, solid curve) and the mixed with 20% of identity mixed in ($\rho_\phi(q = 0.2)$, dot-dashed curve). Even when the state is pure, $\langle W_{J_{xy}}\rangle$ will not be able to witness any entanglement if the phase error is larger than $\pi/3$, just about when AIC declares such a model deficient. Entanglement in the mixed $\rho_\phi(q = 0.2)$ of course is never witnessed. This means $\langle W_{J_{xy}}\rangle$ is only an effective witness in the vicinity of $\left|D_4^2\right\rangle$, with limited tolerance of either white noise or phase noise in even just one of the four qubits. (Of course, one would detect the entanglement in the pure state by appropriately rotating the axes in the spin measurement on the first qubit over an angle ϕ.)

To test whether a few-parameter model correctly quantifies entanglement if that model is preferred over the FPM by AIC, we estimate a (posterior, Bayesian) probability distribution over the generalized negativities (defined above). We see that the first three curves in FIG. 5(a) and the first two curves in FIG. 5(b), which correspond to the data points above the horizontal line in FIG. 3, all give consistent estimates of $\mathcal{N}_0$, compared to the actual value of $\mathcal{N}_0$ for the true state (and the same holds for $\mathcal{N}_{1,2}$ (not shown)). Conversely, the estimate cannot be trusted when AIC deems the simple model inferior to the FPM (of course, it may still happen to be a correct estimate, but one could not be sure). This gives additional evidence for the success of AIC.

F. Cross modeling for a witness measurement

We now construct a two-parameter model $M_{2\phi}^2$ similar in spirit to that discussed for tomographically com-
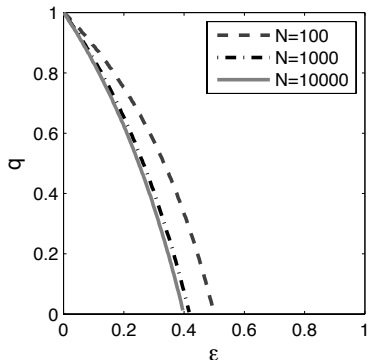
FIG. 6: *What fraction of the model Eq. (23) describes physical states?*: The lower left part separated by the curves is where $\rho(\epsilon, q)$ of Eq. (23) is unphysical (and so is not actually included in the model), for different number of measurements N.

FIG. 7: *How the AIC ranks our two-parameter model vs. the FPM, for a witness measurement*: The difference between Ξ's of $M_{2\phi}^2$ and the FPM, *i.e.*, $-\Delta\Xi = \Xi(\text{FPM}) - \Xi(M_{2\phi}^2)$, for different numbers of witness measurements as a function ϕ. The target state is $\left|\Psi_{\text{target}}^2\right\rangle$. The horizontal line demarcates $\Delta\Xi = 0$. The left figure zooms in on the leftmost three sets of data points in the right figure.

plete measurements: half the data [on which half the time $(\sigma_x)^{\otimes 4}$ is measured, and half the time $(\sigma_y)^{\otimes 4}$] are used to generate a better model, which is then tested on the other half of the data (also containing both types of measurements equally). We write

$$\rho(\epsilon, q) = (1 - \epsilon) \left[(1 - q)\rho_{\text{observation}} + q \left|\Psi_{\text{target}}^2(\phi)\right\rangle \left\langle\Psi_{\text{target}}^2(\phi)\right|\right] + \epsilon \mathbb{1}/D. \quad (23)$$

To find a $\rho_{\text{observation}}$—there are many equivalent ones for predicting the outcomes of the witness measurements—we recall that a generic four-qubit state can be expressed as

$$\rho = \sum_{jklm} c_{jklm}\sigma_j \otimes \sigma_k \otimes \sigma_l \otimes \sigma_m, \quad (24)$$

where $j, k, l, m = 1, 2, 3, 4$ where $\sigma_{1,2,3}$ denote the Pauli matrices $\sigma_{x,y,z}$ and $\sigma_4 = \mathbb{1}$. The witness measures the coefficients c_{jklm} where j, k, l, m can be combinations of only 1 and 4 or combinations of only 2 and 4 (*e.g.*, c_{1441} or c_{4222}). We label the c_{jklm}'s that can be recovered from witness measurement as c_{jklm}^w (w as in *witness*). We do not include in c_{jklm}^w the coefficient c_{4444}, which always equals 1/16, so that it does not depend on measurement outcomes. We define

$$\rho_{\text{observation}} = \sum_{jklm} c_{jklm}^w \sigma_j \otimes \sigma_k \otimes \sigma_l \otimes \sigma_m + \mathbb{1}/16. \quad (25)$$

Note that $\rho_{\text{observation}}$ can be considered as a trace-one *pseudostate*, since it is not necessarily positive semi-definite. But the most attractive property of $\rho_{\text{observation}}$ is that it preserves the measurement outcomes. It is in fact the unique *pseudostate* that reproduces the exact frequencies of all measurement outcomes *and* that has vanishing expectation values for all other *un*performed collective Pauli measurements. As a component of $\rho(\epsilon, q)$,
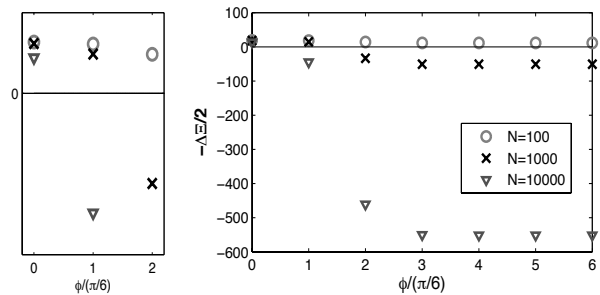
we allow $\rho_{\text{observation}}$ to be unphysical, but we only keep those $\rho(\epsilon, q)$ that are positive semi-definite. We checked numerically for what values of ϵ and q the states end up being physical, and how this depends on the number of measurements performed. Physical states are located in the upper right part of the square in FIG. 6. That is, only if ϵ and/or q are sufficiently large, so that a sufficiently large amount of $\left|\Psi_{\text{target}}^2\right\rangle$ and/or $\mathbb{1}/16$ has been mixed in, does $\rho(\epsilon, q)$ become physical. Depending on the number of measurements, the area of the upper right part is about 69%-77% of the whole square. The physical/unphysical boundary shifts closer to the origin as the number of measurements increases.

We test the two-parameter model (the physical part of it), and show the results in FIG. 7. We find that for $N = 100$ the AIC ranks $\rho_{2\phi}(\epsilon, q)$ better than the FPM, even when the guess about ϕ is very imprecise: 100 witness measurements are, unsurprisingly, not enough for a correct reconstruction of the state. When $N = 1000$, AIC only prefers the models with a value for ϕ within $\pi/6$ of the correct value. And when $N = 10000$, the accepted values of ϕ are even closer to the true value.

The corresponding posterior distributions of negativities $\mathcal{N}_2$ are plotted in FIG. 8 for the three better guesses, $\phi = 0, \pi/6, \pi/3$. When $N = 100$ all three give decent predictions of $\mathcal{N}_2$ (and indeed, AIC ranks those models highly). For $N = 1000$ and $N = 10000$, we would only trust the estimates arising from the lower two values of ϕ, or just the correct value of ϕ, respectively. This trust is rewarded in FIG. 7(b) and FIG. 7(c), as those estimates are indeed correct, within the error bars. In addition, the untrusted estimate for $\phi = \pi/6$ for $N = 10000$ still happens to be correct, too.
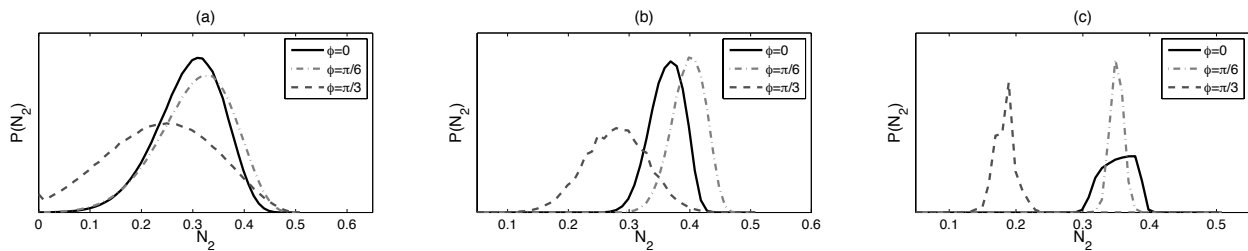
FIG. 8: *Quantifying entanglement from a witness measurement*: The posterior distributions for $\mathcal{N}_2$, for different numbers of measurements ($N = 100$ in (a), $N = 1000$ in (b), $N = 10000$ in (c)), using model $M_{2\phi}^2$ with $\rho_{\mathrm{observation}}$ and target state $\left|\Psi_{\mathrm{target}}^2\right\rangle$, where $\phi = 0, \pi/6, \pi/3$. $\mathcal{N}_2(\rho_{\mathrm{actual}}) = 0.3875$. The posterior distributions are obtained by binning the results into 50 equal-sized bins from the minimum to the maximum value of $\mathcal{N}_2$ for each distribution separately. The same prior is used as in FIG. 5. Whenever the AIC declares a model superior to the FPM, the estimated entanglement agrees, within error bars, with the actual value, but may be wrong otherwise.
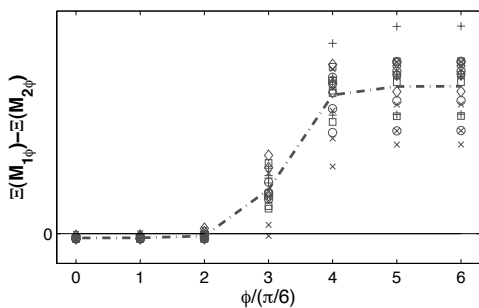

FIG. 9: *Comparing one- and two-parameter models directly*: The difference between Ξ's of $M_{2\phi}^2$ and $M_{1\phi}^2$ for 20 different sets of witness measurements ($N = 50$) as functions of ϕ. The target state is $\left|\Psi_{\mathrm{target}}^2\right\rangle$. The horizontal line demarcates $\Delta\Xi = 0$. The dotted-dashed line is the average of all 20 points at each different ϕ.

G. Comparing one- and two-parameter models directly

Finally, the AIC can compare the one- and two-parameter models $M_{1\phi}$ and $M_{2\phi}$ directly. For that purpose one needs to use the *same* validation set of data, which implies that the two-parameter model needs *additional* data to generate $\rho_{\mathrm{observation}}$. Here we display results for just 50 witness measurements, and an additional set of 50 measurements for $M_{2\phi}^2$. FIG. 9 shows that even such a small number of additional data is useful if the angle ϕ is wrong, and, similarly, it shows that the same small number suffices to detect a wrong single-qubit phase when it is larger than $\pi/3$.

IV. CONCLUSIONS

We applied information criteria, and the Akaike Information Criterion (AIC) developed in Ref. [15] in partic-

ular, to quantum state estimation. We showed it to be a powerful method, provided one has a reasonably good idea of what state one's quantum source actually generates.

For each given model, which may include several parameters describing error and noise, as well as some parameters—call them the ideal-state parameters— describing the state one would like to generate in the ideal (noiseless and error-free) case, the AIC determines a ranking from the observed data. One can construct multiple models, for instance, models where some ideal-state parameters and some noise parameters are fixed (possibly determined by previous experiments in the same setup), with others still considered variable. Crucially, the AIC also easily ranks the full-parameter model (FPM), which uses in principle all exponentially many parameters in the full density matrix, and which is, therefore, the model one would use in full-blown quantum state tomography. This ranking of the FPM can be accomplished without actually having to find the maximum-likelihood state (or its likelihood)—which quickly would run into insurmountable problems for many-qubit systems—by using a straightforward upper bound.

This way, observed data is used to justify *a posteriori* the use of the few-parameter models—namely, if the AIC ranks that model above the FPM—and thus our method is in the same spirit as several other recent proposals [5, 6] to simplify quantum tomography, by tentatively introducing certain assumptions on the quantum state generated, after which data is used to certify those assumptions (and if the certification fails, one at least knows the initial assumptions were incorrect).

We illustrated the method on (noisy and mis-specified) four-qubit members of the family of Dicke states, and demonstrated its effectiveness and efficiency. For instance, we showed that one can detect mis-specified ideal-state parameters and determine noise and error parameters. We also showed by example the successful application of the method to a specific and useful subtask, that of quantifying multi-qubit entanglement.

Acknowledgement

This work was supported by NSF grant PHY-1004219.

[1] M. Paris and J. Řeháček, *Quantum state estimation (Lecture notes in physics)* (Springer, 2004).
[2] Z. Hradil, Phys. Rev. A **55**, R1561 (1997).
[3] R. Blume-Kohout, New J. Phys. **12**, 043034 (2010).
[4] T. Monz, P. Schindler, J. T. Barreiro, M. Chwalla, D. Nigg, W. A. Coish, M. Harlander, W. Häsel, M. Heinrich, and R. Blatt, pp. arXiv:1009.6126v1 [quant–ph] (2010).
[5] D. Gross, Y.-K. Liu, S. Flammia, S. Becker, and J. Eisert, Phys. Rev. Lett. **105**, 150401 (2010).
[6] M. Cramer, M. B. Plenio, S. T. Flammia, R. Somma, D. Gross, S. D. Bartlett, O. Landon-Cardinal, D. Poulin, and Y.-K. Liu, Nature Communications **1**, 149 (2010).
[7] A. Klumper, A. Schadschneider, and J. Zittartz, J. Phys. A **24**, L955 (1991).
[8] M. Fannes, B. Nachtergaele, and R. F. Werner, Comm. Math. Phys. **144**, 443 (1992).
[9] D. Perez-Garcia, F. Verstraete, M. M. Wolf, and J. I. Cirac, Quantum Inf. Comput. **7**, 401 (2007).
[10] G. Tóth, W. Wieczorek, D. Gross, R. Krischek, C. Schwemmer, and H. Weinfurter, Phys. Rev. Lett. **105**, 250403 (2010).
[11] H. Jeffreys, *Theory of probability* (Oxford University Press, 3rd Ed., 1961).
[12] K. P. Burnham and D. R. Anderson, *Model selection and multimodel inference: a practical information-theoretic approach* (Springer-Verlag, 2002).
[13] W. Zucchini, J. Math. Psych. **44**, 41 (2000).
[14] S. Kullback and R. A. Leibler, Ann. Math. Stat. **22**, 79 (1951).
[15] H. Akaike, IEEE Transactions on Automatic Control **19**, 716 (1974).
[16] K. Takeuchi, Suri-Kagaku (Mathematic Sciences) **153**, 12 (1976).
[17] K. Banazek, G. M. D'Ariano, M. G. A. Paris, and M. F. Sacchi, Phys. Rev. A **61**, 010304R (1999).
[18] A. R. Liddle, Mon. Not. Roy. Astron. Soc. Lett. **377**, L74 (2007).
[19] P. Lougovski and S. J. van Enk, Phys. Rev. A **80**, 052324 (2009).
[20] K. Usami, Y. Nambu, Y. Tsuda, K. Matsumoto, and K. Nakamura, Phys. Rev. A **68**, 022314 (2003).
[21] N. Sugiura, Commun. Stat. Theo. Math. A **7**, 13 (1978).
[22] R. H. Dicke, Phys. Rev. **93**, 99 (1954).
[23] H. Häffner, W. Hänsel, C. F. Roos, J. Benhelm, D. C. al kar, M. Chwalla, T. Körber, U. D. Rapol, M. Reibe, P. O. Schmidt, et al., Nature **438**, 643 (2005).
[24] N. Kiesel, C. Schmidt, G. Tòth, E. Solano, and H. Weinfurter, Phys. Rev. Lett. **98**, 063604 (2007).
[25] S. B. Papp, K. S. Choi, H. Deng, P. Lougovski, S. J. van Enk, and H. J. Kimble, Science **324**, 764 (2009).
[26] K. S. Choi, A. Goban, S. B. Papp, S. J. van Enk, and H. J. Kimble, Nature **468**, 412 (2010).
[27] J. M. Renes, R. Blume-Kohout, A. J. Scott, and C. M. Caves, Journal of Mathematical Physics **45**, 2171 (2004).
[28] M. S. Kaznady and D. F. V. James, Phys. Rev. A **79**, 022109 (2009).
[29] P. Stoica, P. Eykhoff, P. Janssen, and T. Söderström, International Journal of Control **43**, 1841 (1986).
[30] M. Stone, J. Roc. Stat. Soc. B **39**, 44 (1977).
[31] G. Tòth, J. Opt. Soc. Am. B **24**, 275 (2007).
[32] G. Vidal and R. F. Werner, Phys. Rev. A **65**, 032314 (2002).
[33] S. Sabin and G. Garcia-Alcaine, Eur. Phys. J. D **48**, 435 (2008).
[34] J. O. S. Yin and S. J. van Enk, Phys. Rev. A **83**, 022326 (2011).