SCIENTIFIC REPORTS

OPEN

Received: 08 December 2015
Accepted: 07 November 2016
Published: 01 December 2016

Crystallizing highly-likely subspaces that contain an unknown quantum state of light

Yong Siah Teo[1,2], Dmitri Mogilevtsev[3], Alexander Mikhalychev[3], Jaroslav Řeháček[2] & Zdeněk Hradil[2]

In continuous-variable tomography, with finite data and limited computation resources, reconstruction of a quantum state of light is performed on a finite-dimensional subspace. In principle, the data themselves encode all information about the relevant subspace that physically contains the state. We provide a straightforward and numerically feasible procedure to uniquely determine the appropriate reconstruction subspace by extracting this information directly from the data for any given unknown quantum state of light and measurement scheme. This procedure makes use of the celebrated statistical principle of maximum likelihood, along with other validation tools, to grow an appropriate seed subspace into the optimal reconstruction subspace, much like the nucleation of a seed into a crystal. Apart from using the available measurement data, no other assumptions about the source or preconceived parametric model subspaces are invoked. This ensures that no spurious reconstruction artifacts are present in state reconstruction as a result of inappropriate choices of the reconstruction subspace. The procedure can be understood as the maximum-likelihood reconstruction for quantum subspaces, which is an analog to, and fully compatible with that for quantum states.

One of the scientifically established tenets in quantum mechanics is the ability to reconstruct any quantum state of an arbitrary quantum source[1,2]. Maturation of theoretical and experimental techniques in quantum tomography for continuous-variable (CV) measurements is of top priority for practical certifications in optical quantum cryptography[3–7], optomechanics[8,9], quantum metrology[10,11] and other quantum computation protocols[12–17].

Since measurement data and computation resources are always finite, the reconstruction of any quantum state of light, which in principle resides in an infinite-dimensional Hilbert space, is always performed on a finite-dimensional subspace. An unsolved problem in CV tomography is an objective systematic search for the appropriate reconstruction subspace. Ideally, an observer would hope for an analytical reasoning that leads to the optimal reconstruction subspace that minimizes some sort of tomographic accuracy measure. This thinking is, in some sense, naive as such an optimal subspace would always depend on the measurement scheme and the true quantum state of the source, an element that is certainly unknown to the observer. Furthermore, the positivity constraint on quantum states forbids any straightforward analysis on the problem.

Despite the aforementioned difficulties, there exist numerous studies on alternative solutions to this problem. These studies, nonetheless, involve making some assumptions about the source. If the observer knows, usually with low to moderate levels of confidence, that the source emits no more than $D_{\rm rec}$ photons, then in principle, she can prepare a set of CV measurement outcomes that is *informationally complete* on the $D_{\rm rec}$-dimensional Hilbert subspace[18]. The *maximum-likelihood* (*ML*) method[19,20], for instance, can be used to reconstruct the state on this subspace based on the measurement data. However, in refs 21 and 22 it was shown that such a simple approach often gives estimators that are far away from the true state $\rho_{\rm true}$, especially when there are features in high-dimensional sectors that are not obvious from simple deductions with mean photon numbers. In the same articles, the technique of *maximum-likelihood-maximum-entropy* (MLME) was used to reconstruct states on subspaces larger than the tomographic coverage of the measurement outcomes to reveal genuine quantum-state features and reduce reconstruction artifacts on average.

[1]BK21 Frontier Physics Research Division, Seoul National University, 1 Gwanak-ro, Gwanak-gu, 08826 Seoul, South Korea. [2]Department of Optics, Palacký University, 17. listopadu 12, 77146 Olomouc, Czech Republic. [3]Institute of Physics, Belarus National Academy of Sciences, F. Skarina Ave. 68, 220072 Minsk, Belarus. Correspondence and requests for materials should be addressed to Y.S.T. (email: yong.siah.teo@gmail.com)

Recently, methods employed in classical statistical-model selection were used to localize the signal (see, for example, refs 23–28). These methods involve the consideration of the popular Akaike criterion and the Bayesian information criterion to penalize the likelihood function for the problem and restrict models for up to a certain number of parameters. These strategies depend on the parametric models selected (which in our case corresponds to some pre-chosen reconstruction subspace) to optimize the signal locations, the accuracy of which depend heavily on the correctness of the chosen models. An average of these models may be carried out over the quantum state space to mitigate possible model inaccuracies[29]. Other methods of choosing reconstruction subspaces include the utilization of other prior knowledge about the source and assigning a partial dependence of the subspace dimension D_{rec} on the number of measurement settings or groups of outcomes[30].

In what follows, we shall present a systematic and practical procedure to locate subspaces that highly-likely contain a given unknown quantum state of light that is completely free of model (preconceived subspace) considerations and hard-to-justify assumptions about the source. This strategy converges to the appropriate "model" subspace based on information encoded in the collected measurement data alone. In a nutshell, the procedure makes use of the ML strategy to define an initial reconstruction subspace of low-dimension and gradually evolve the seed subspace to a reconstruction subspace of a stipulated dimension D_{rec}—much like a typical nucleation process in the formation of crystals. The termination of the ML nucleation process, and the subsequent determination of D_{rec}, is governed by the procedure of cross-validation, which is a prototypical statistical validation tool that ensures the reliability and predictive power of the resulting ML state estimator. This numerical nucleation process, which is naturally compatible with ML state estimation[1,2], makes use of *only* the acquired measurement data in an experiment. The underlying physical reason is that all encoded information in the data reflects the features of the unknown quantum state, albeit with some statistical fluctuation, and can thus be systematically extracted to obtain the optimal reconstruction subspace and state estimator.

Without loss of generality, we shall assume here that the data associated with the continuous-variable quantum measurement, although finite, are sufficiently large enough such that statistical fluctuation is minimized within typical experimental means. In this situation, the relevant reconstruction error of interest is primarily influenced by the choice of reconstruction subspace.

Results

The ML subspace nucleation process.

Suppose that the observer chooses to reconstruct the true quantum state ρ_{true} of the source using the ML method from a set of data. In CV tomography, the data are event occurrences $\sum_j n_j = N$ of N sampling events (say voltage detection) collected with a measurement described by a set of probability operator measurement (POM) $\sum_{j=1}^{M} \Pi_j = 1$ consisting of M outcomes $\Pi_j \geq 0$. In this scenario, it is natural to consider an assignment of the reconstruction subspace that is compatible with the ML principle. Clearly, if the desired ML estimator $\hat{\rho}_{\mathrm{ML}}$ is the one that maximizes the log-likelihood function of ρ for the data,

$$\log \mathcal{L}(\{n_j\};\, \rho) = \sum_{j=1}^{M} n_j \,\log p_j, \quad \text{where } p_j = \mathrm{tr}\{\rho \Pi_j\}, \tag{1}$$

the reconstruction subspace should then be a subspace of a certain dimension D_{rec} that optimizes this log-likelihood function. The problem now reduces to deciding the appropriate value of $D_{\mathrm{rec}} \geq 2$ and searching for the optimal subspace of this dimension.

In real situations where detection losses are present, such that the efficiency of the overall quantum detection is less than unity, there exist many ways to cope with this additional detail depending on specific experimental situations. The *complete likelihood* is now one that accounts for both the detected and missing copies of quantum systems, with the sum of the measured probabilities $\sum_{j'} p_{j'} < 1$. If the source (usually photonic in this case) has a known Poissonian prior distribution for the total number of quantum systems (photons), then an effective likelihood function for the measured sampling events may be defined as an average of the complete likelihood over the missing quantum systems because of losses with this known prior distribution. Alternatively, we may carry on the maximum-likelihood philosophy and instead maximize the complete likelihood over the missing copies, and the result is an effective log-likelihood function of the form in Eq. (1) with p_j replaced by $p_j / \sum_{j'} p_{j'}$[1,21]. In any case, the effective log-likelihood function is now the proper log-likelihood function for subsequent optimizations.

Since all we have are the measurement data $\{n_j\}$, the most straightforward way to carry out the subspace search is *subspace nucleation*. Such a numerical nucleation process involves the surveillance of all possible d-dimensional discrete subspaces of some large Hilbert space of dimension D_{lim} that defines some limit for the state reconstruction. All the $L = \binom{D_{\mathrm{lim}}}{d}$ subspaces can be represented by a set of L projectors $\{S_{l,d}\}_{l=1}^{L}$ that are diagonal in the computational basis and consist of $D_{\mathrm{lim}} - d$ zeros and d ones. For any given operator A, its suboperator $A_{l,d}$ in the lth subspace is conveniently expressed as $A_{l,d} = S_{l,d} A S_{l,d}$. Analogous to nucleation in crystal formation, subspace nucleation begins with a seed subspace of a certain smallest pre-chosen dimension d. For the purpose of illustration, we take $d = 2$. The qubit subspace $S_2^{(\mathrm{ML})}$ appropriate for seeding the nucleation process is the one corresponding to the two- dimensional $\hat{\rho}_{\mathrm{ML}}$ of the largest maximal log-likelihood out of all possible projectors $S_{l,2}$. The subspace begins to grow along the trajectory of largest likelihood increment. The next optimal subspace to choose would be the subspace $S_4^{(\mathrm{ML})} = S_2^{(\mathrm{ML})} + S_2$, where S_2 is the optimal orthogonal subspace to $S_2^{(\mathrm{ML})}$, such that the corresponding four-dimensional $\hat{\rho}_{\mathrm{ML}}$ yields the largest maximal log-likelihood. Nucleation continues, this time establishing the next larger optimal subspace $S_6^{(\mathrm{ML})} = S_4^{(\mathrm{ML})} + S_2$ such that, again, $S_2 S_4^{(\mathrm{ML})} = 0$ and the corresponding six-dimensional $\hat{\rho}_{\mathrm{ML}}$ gives the largest maximal log-likelihood, and so on.

In this way, the reconstruction subspace matures in the direction of maximal sequential increase in the log-likelihood. Since the process evaluates the (log-)likelihood and maximizes it over subspaces, this process is entirely equivalent to a ML subspace estimation, which is fully analogous to a ML state estimation. The data alone contain all hidden signatures of the relevant subspace segments and the structures thereon, all of which are revealed by nothing else but the log-likelihood function. In the limit of large N of sampling events, which is an achievable commodity in homodyne tomography, for instance, these signatures accurately reflect those of ρ_{true}. This function thus serves as the only important objective function following which subspace crystallization takes place. No additional parametric model selection or spurious assumptions about the source are necessary. We have therefore established a fully objective numerical procedure for assigning reconstruction subspaces that is compliant with ML state estimation.

Computationally, the ML subspace nucleation process is a continuous iteration of the following simple numerical steps over k, starting with the smallest optimal d-dimensional seed subspace defined by $S_d^{(\mathrm{ML})}$ at $k = 1$ and proceeding till $k = \kappa$ that defines the final reconstruction-subspace dimension D_{rec}:

1. In the kth step, look for the full set of operators $\mathcal{S}_\perp = \{S_{l,d}\}$ that are orthogonal to $S_{kd}^{(\mathrm{ML})}$.
2. Set $S_{(k+1)d}^{(\mathrm{ML})} = S_{kd}^{(\mathrm{ML})} + S_d$ with $S_d \in \mathcal{S}_\perp$ that maximizes $\log\mathcal{L}(\{n_j\};\,\hat{\rho}_{\mathrm{ML}})$, where $\hat{\rho}_{\mathrm{ML}}$ resides in the subspace defined by $S_{(k+1)d}^{(\mathrm{ML})}$.

The **Methods** section provides more explicit details on the numerical procedure.

Criterion for nucleation termination. The final task is now to decide on the reasonable value of D_{rec}. Various statistical tools are available for this purpose, the choice of which depends on the application of the statistical operator $\hat{\rho}_{\mathrm{ML}}$. Typically, the estimator $\hat{\rho}_{\mathrm{ML}}$ is used for statistical prediction of probability distributions for future measurement schemes. Some measure of predictive power for the estimator is hence necessary to judge if the related subspace acquired from the nucleation process is sufficiently accurate in data prediction. Physically, the reconstruction subspace that best predicts data should be the largest possible subspace that tomographically covers all the possible datasets (infinite-dimensional in principle). In practice, however, all resources are finite and some sort of statistical certification is necessary to judge if the estimator of finite data that resides in a finite subspace is predictive enough.

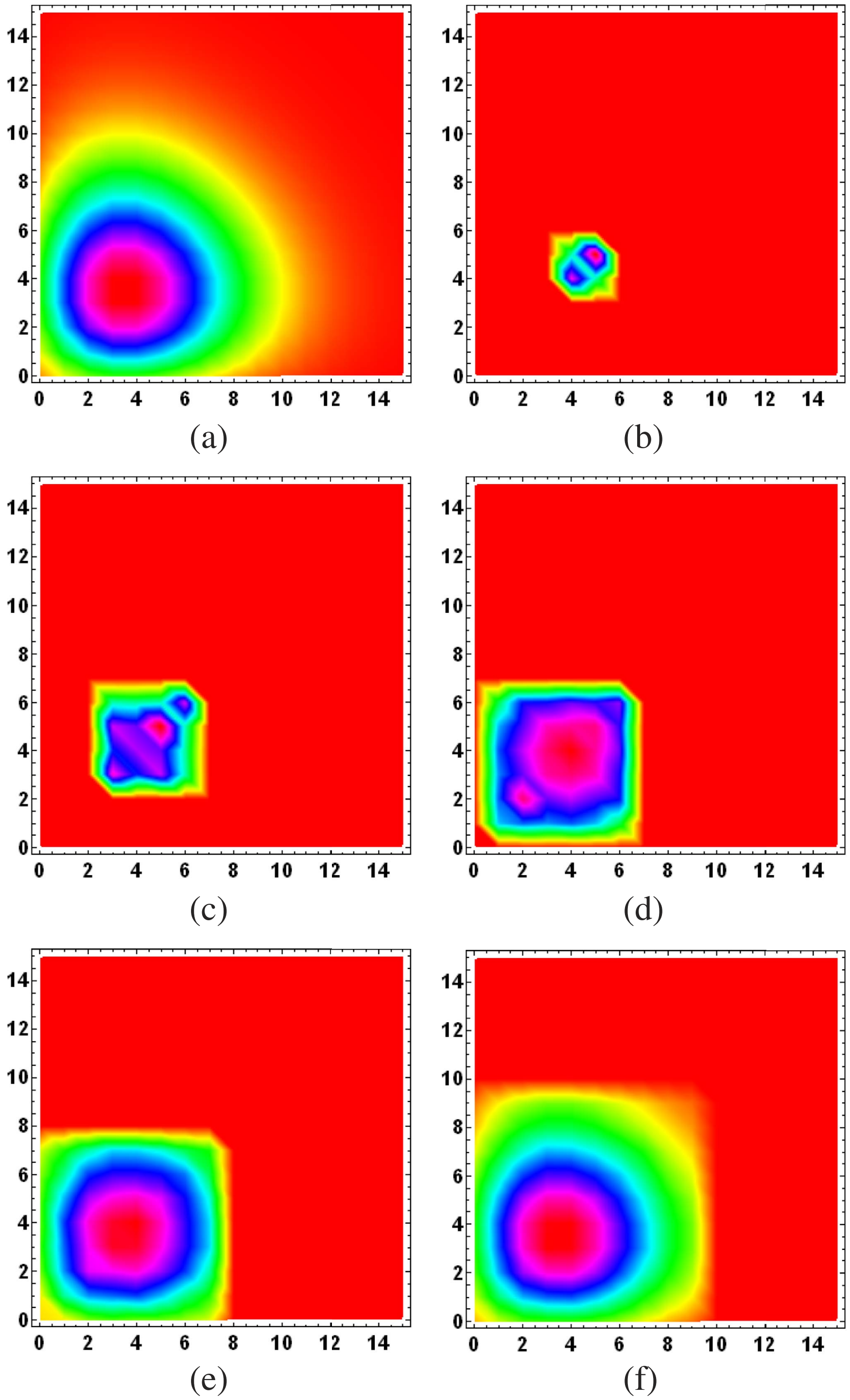
Cross-validation[31–33] (and a simplified variant discussed in ref. 34) is a decent certification tool of choice to judge if $\hat{\rho}_{\mathrm{ML}}$ resides in a large enough reconstruction subspace of reasonable coverage relative to the data of a given POM. Typically, when the estimator $\hat{\rho}_{\mathrm{ML}}$ fits the data from a POM according to the log-likelihood function, the estimator may not necessarily predict other data from the same POM, or any other POM for that matter, especially in a situation where the reconstruction subspace does not tomographically include the measurement data sufficiently. If, on average, $\hat{\rho}_{\mathrm{ML}}$ predicts different sets of data of the same POM, then the subspace yields a predictive $\hat{\rho}_{\mathrm{ML}}$. For demonstrating the principles of cross-validation, we consider a *two-fold* cross-validation strategy and split the M measurement data into two datasets of equal size. Borrowing the language of machine learning, one dataset, the *training set*, is used to obtain $\hat{\rho}_{\mathrm{ML}}$, and the other *testing set* is used to test the predictive power of $\hat{\rho}_{\mathrm{ML}}$. The roles of both datasets are then switched, and training and testing are performed again. The average chi-square metric between the test data and the ML probabilities,

$$\mathrm{PrErr} = \frac{1}{M}\sum_{k=1}^{2}\sum_{j=1}^{M/2}\frac{(n_j/N - \hat{p}_j^{(\mathrm{ML})})^2}{\hat{p}_j^{(\mathrm{ML})}}\Bigg|_{k\text{th testing set}}, \tag{2}$$

describes the predictive power for $\hat{\rho}_{\mathrm{ML}}$ in terms of the prediction error for the given measurement scheme.

Like all numerical algorithms, there are many ways to terminate the nucleation procedure. The observer may choose to set a pre-chosen tolerance level for PrErr beyond which the procedure stops; or compare the change in the current PrErr value relative to the preceding value and accept the reconstruction if the change falls below certain threshold; or simply repeat the procedure a pre-chosen number of times. The numerical stabilization of PrErr can serve as an indication that continuing the procedure will not give appreciable improvement in the resulting ML estimator and ML subspace. Since the value of PrErr fluctuates for every experimental run, its value for each reconstruction-subspace dimension D_{rec} should be accompanied by a statistical quantifier for its reliability. As a typical choice, we shall assign confidence intervals to reflect the level of confidence (or signifcance) for these values. These confidence intervals are calculated using a known method of *bootstrapping* on the PrErr values (see **Methods**).

Numerical Experiments. To put the ML subspace nucleation procedure to the test, for a given true state ρ_{true}, we simulate an experimental run for a CV POM involving $M = 1000$ random rank-one POM outcomes distributed uniformly according to the Haar measure[35]. In this run, a total of $N = 10^7$ sampling events is measured with the POM and the resulting data are accumulated through Monte Carlo methods. To demonstrate the proposed numerical method, we investigate three examples, namely a coherent state, even coherent state and squeezed coherent state. Figures 1, 2 and 3 provide a visualization of the respective nucleation processes for these three states. Figure 4 presents the results of the nucleation process with statistical descriptions for the PrErr values.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 1. Subspace nucleation process from one set of data for (**a**) the coherent state defined by $|\alpha\rangle$ of mean-photon number equal to $|\alpha|^2 = 4$, projected onto the 16-dimensional Hilbert space for visualization. The seed subspace is of dimension $d = 2$. Subspaces of (**b**) $D_{\mathrm{rec}} = 2$, (**c**) 4, (**d**) 6, (**e**) 8 and (**f**) 10 that respectively maximize the log-likelihood are shown here for $M = 1000$ measurement outcomes. The interpolated hue for each integer coordinate (position of the matrix element) in the plots visually indicates the relative magnitudes of neighboring matrix elements of the real parts of all quantum states in the computational basis. Here the ten-dimensional optimal ML subspace already captures most of the important features of the state.

The results obtained with simulated experiments verify the decreasing behavior for the values of the prediction error PrErr with increasing reconstruction-subspace dimension D_{rec}. This behavior confirms that, logically, a larger subspace would more adequately accomodate the measurement outcomes and more accurately predict any data derived from these outcomes.

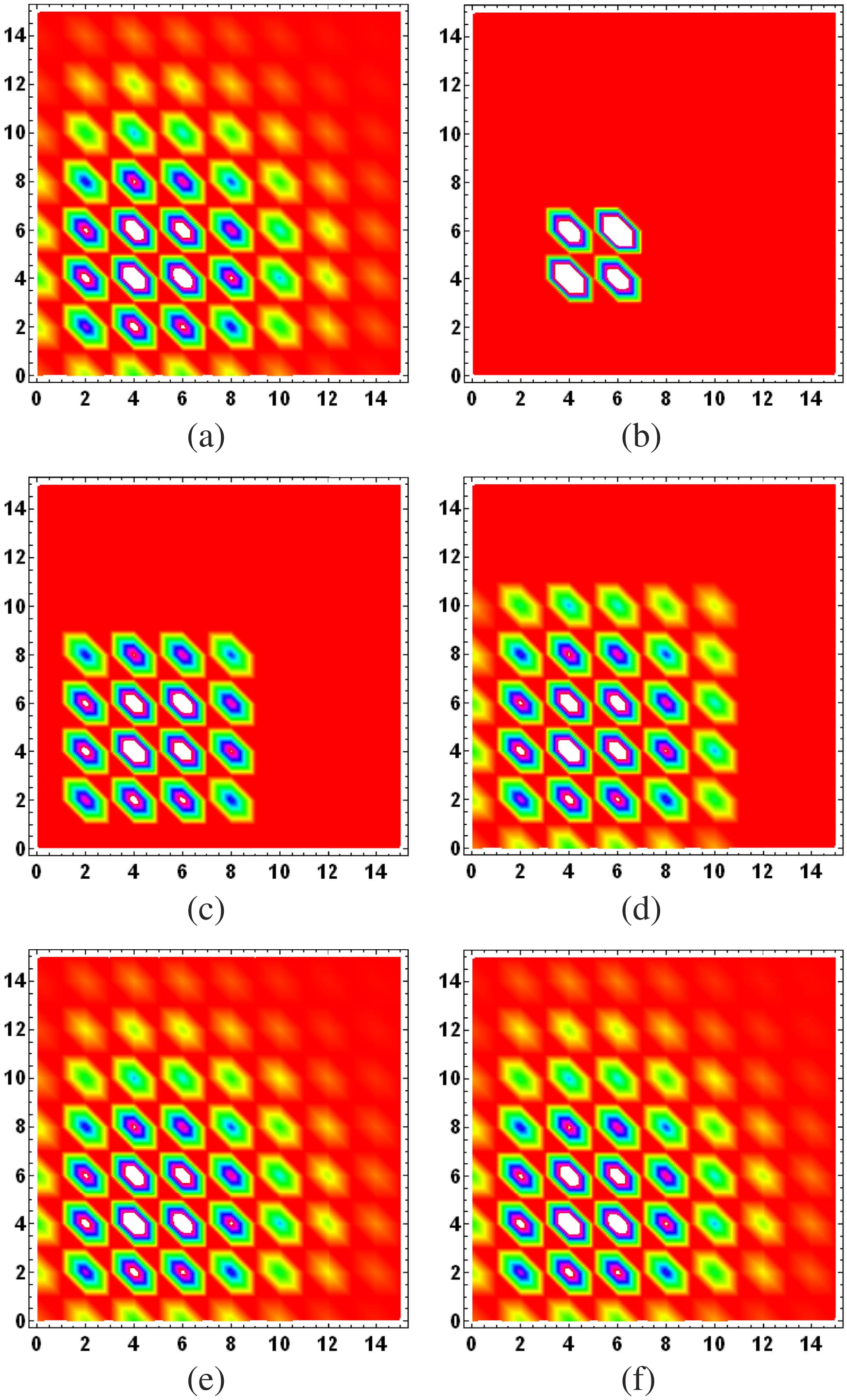


Figure 2. Subspace nucleation process from one set of data for (**a**) the even coherent state defined by $\mathcal{N}(|\alpha\rangle + |-\alpha\rangle)$ with $\alpha = \sqrt{5}$ and a proper normalization. All other figure specifications are as described in Fig. 1. For this state, the eight-dimensional optimal ML subspace is sufficient for a rather accurate ML reconstruction.

Practical Aspects Of The Nucleation Methodology. *Subspace coverage.* In the usual situation, the observer has already an intended target quantum state ρ_{targ} for the source in mind before setting up the experiment for a particular quantum protocol. Owing to experimental imperfections, the target state she intends to prepare is never the same as the true state ρ_{true} to which she asymptotically measures. Nevertheless, if the control of the source is done well, the observer may have reasons to believe that ρ_{true} should be close to ρ_{targ}. For a given basis, the reconstruction dimension D_{rec}, and hence the limit dimension D_{lim}, should at least be large enough to encompass all the significant matrix elements of ρ_{true}. Usually, the choice of D_{lim} is decided from ρ_{targ} by trusting that it is close enough to the unknown ρ_{true}.

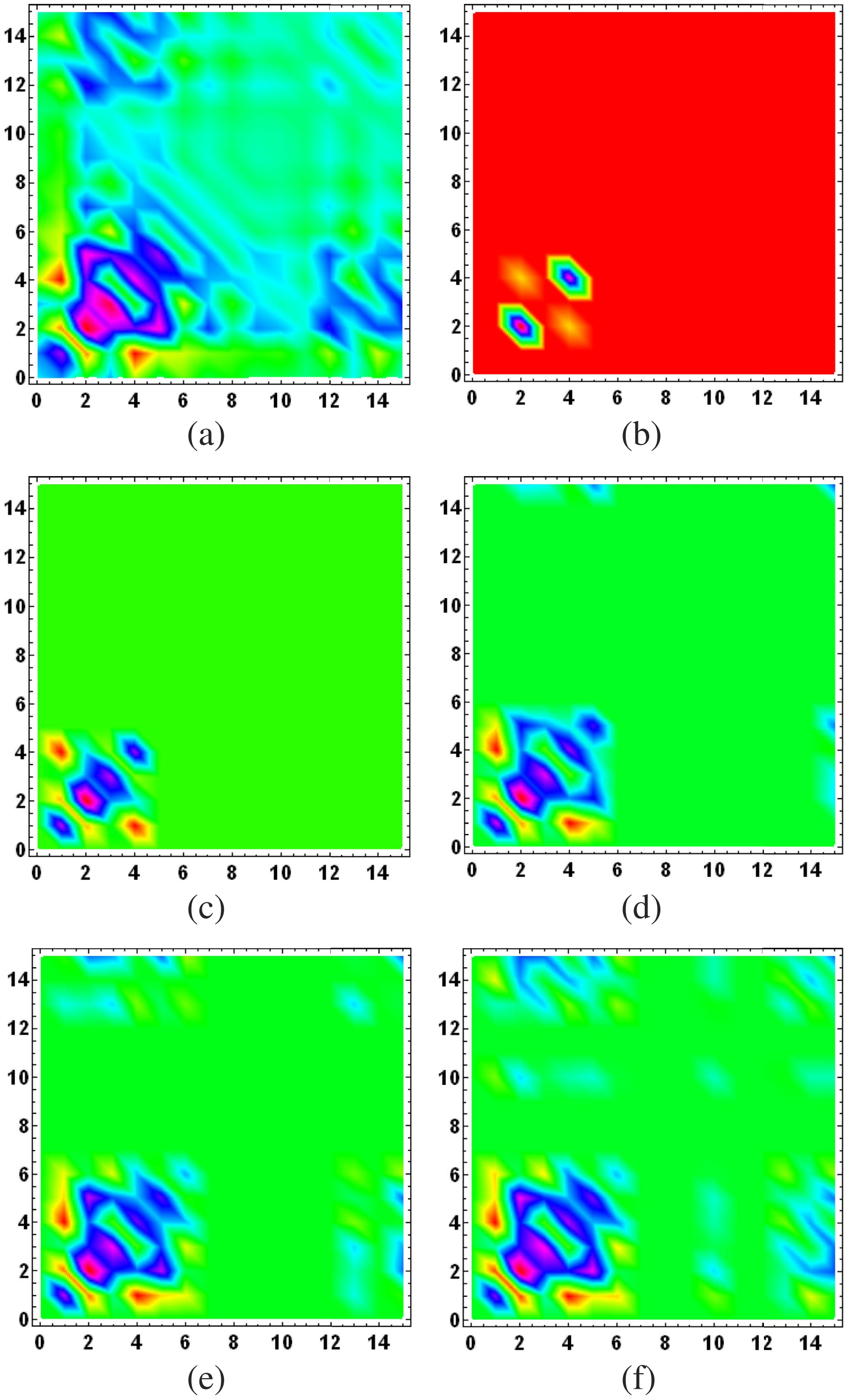


Figure 3. Subspace nucleation process from one set of data for (**a**) a squeezed coherent state of squeeze parameter $z = 2\mathrm{e}^{\mathrm{i}\frac{\pi}{4}}$ and $\alpha = \sqrt{5}$. All other figure specifications are as described in Fig. 1.

However, such a gut feeling is not a necessary ingredient to pick the value of D_{lim}, for the data themselves already contain all encoded information about the quantum-state features. If D_{lim} is too small to cover the significant features of ρ_{true}, the data should be able to tell us just that, which they do indeed. One way of capturing the tell-tale signs from the data is to inspect the PrErr values. If the PrErr saturates at a value that is large, then this is an indication that the subspace does not cover the state features very well, and the corresponding estimator does not explain the data obtained and will have limited predictive power. In this case, one would need to increase the value of D_{lim}. The behavior of PrErr with D_{rec} would eventually stabilize for sufficiently large D_{lim}.

As an example, Fig. 5 shows the behavior of PrErr for different values of D_{lim}, obtained from a fixed set of simulated data of a coherent state with mean photon number 30. As D_{lim} increases, the saturation of PrErr lowers and vanishes for sufficiently large D_{lim}. In this way, the choice of D_{lim} is optimized without the need for a prior belief.

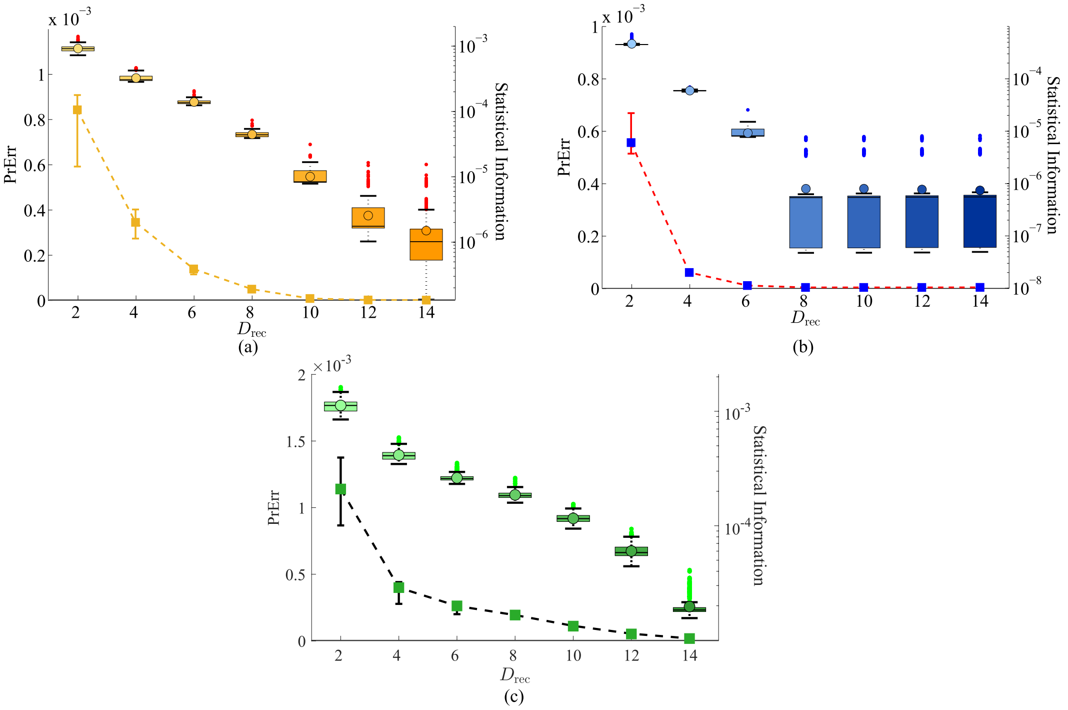


Figure 4. Superimposed plots of the PeErr values with error bars (squares and dashed lines plotted on a linear scale) and corresponding statistical information about each value (plotted on a log scale) against D_{rec}, respectively for (**a**) the coherent state, (**b**) the even coherent state and (**c**) the squeezed coherent state. Each error bar is computed from the relevant bootstrap distribution. Small error bars are not visible in the figure. The statistical information for each value of D_{rec} is based on a bootstrap distribution of 500 Monte-Carlo-generated PrErr values. This information includes the first and third quartiles of these points (respectively the bottom and top edges of the rectangle), the median or second quartile (solid line in the rectangle), the mean (circle), the lowest datum still within 1.5 interquartile range (IQR) of the first quartile and the highest datum still within 1.5 IQR of the third quartile (respectively the bottom and top solid lines of the whisker). Outliers, which are outside the whisker, are plotted as vertically-aligned dots.

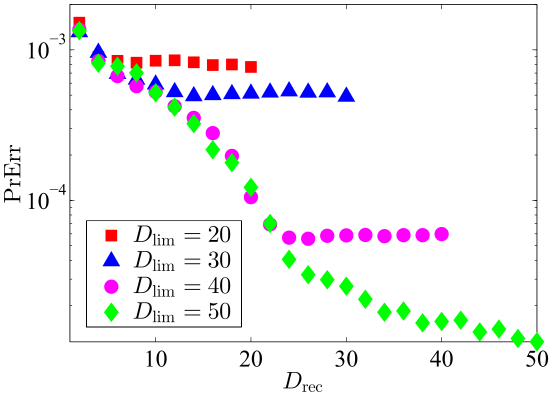


Figure 5. A plot of PrErr (log-scale) against D_{rec} for a fixed set of data and various limit dimensions D_{lim}. The true state is a coherent state of mean photon number 30.

Subspace truncation and rate of convergence. If the observer insists, she can certainly make use of an educated prior belief for the true state to enhance the ML subspace nucleation procedure in an objective way. To understand how, consider the simple case where ρ_{true} is the single-photon Fock state $|n=1\rangle\langle n=1|$. If one carries out the nucleation procedure in the computational basis with (hypothetical) noiseless data, then the procedure will terminate after just one step. This is because already after the very first step, the optimal qubit subspace is, of course, any of the subspaces that covers the $n=1$ sector, and the resulting ML estimator $\hat{\rho}_{\text{ML}}$ is precisely ρ_{true} since all other matrix elements are zero in the computational basis.

More generally, for (hypothetical) noiseless data and a rank-one ρ_{true}, if the basis in which subspace truncation of the Hilbert space is performed happens to be the basis with one of the basis kets being the ket of ρ_{true}, then the

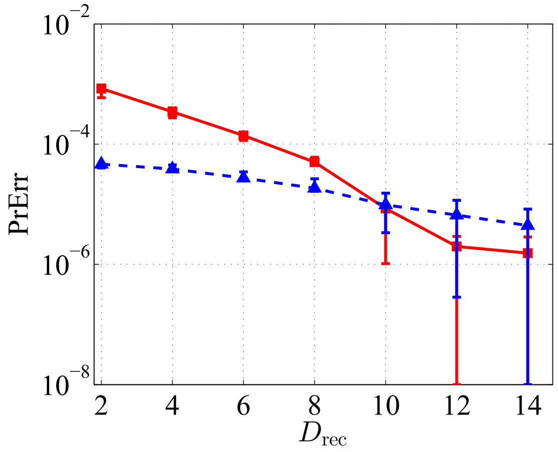


Figure 6. A plot of PrErr (log-scale) against D_{rec} for the coherent state discussed in Fig. 1 and a fixed set of simulated data with slight statistical fluctuation. Comparing with the rate of the subspace nucleation process in the standard computational (Fock) basis (red solid lines and square markers), a basis transformation based on the target coherent state of mean photon number five gives a relatively faster convergence (blue dashed line and triangular markers). The intersection of the data-point sequences at $D_{\mathrm{rec}} = 10$ and larger arises from the finite precision of the ML estimation.

nucleation procedure yields ρ_{true} after the very first step, because in this basis all matrix elements are zero except for the one corresponding to the ket of ρ_{true}. This argument is easily extended to any ρ_{true}, in which case truncation in a basis containing all eigenstates of ρ_{true} will give ρ_{true} as the estimator in no more than rank $\{\rho_{\mathrm{true}}\}$ steps. The exact number of steps would depend on the dimension d of the seed subspaces. Even for a realistic situation when the observer has access to only the target state ρ_{targ}, not ρ_{true}, and real data with statistical fluctuation, finding the right basis with a reasonable ρ_{targ} to perform subspace truncation for the nucleation procedure can greatly speed up the nucleation procedure if ρ_{true} is reasonably close to the target state, especially in the limit of large number of sampling events N.

It is now clear how the prior belief enters the nucleation procedure—it is simply used to set up the appropriate basis for subspace truncation in order to carry out subspace nucleation with significantly fewer steps. *In no way* is the final state estimator $\hat{\rho}_{\mathrm{ML}}$ dependent on the prior belief, only the rate of convergence to $\hat{\rho}_{\mathrm{ML}}$, for the entire nucleation process is still controlled by data inspection alone once the basis is set up. Mathematically, if U is the unitary operator that converts the Fock basis to the appropriate basis, then a basis transformation $\rho \rightarrow U\rho U^{\dagger}$ on quantum states in the optimization routine is entirely equivalent to an inverse basis transformation $\Pi_j \rightarrow U^{\dagger}\Pi_j U$ on measurement outcomes due to the symmetry in the Born rule. If the observer believes that the target state $\rho_{\mathrm{targ}} = |\ \rangle\langle\ |$ is the likely candidate for describing the source, she may take this and generate a D_{lim}-dimensional eigenbasis of ρ_{targ} and construct U out of this eigenbasis. The results in Fig. 6 further confirms the possibility of a significant improvement in nucleation convergence to the final optimal subspace and state estimator for a given set of data after a basis transformation.

Discussions

We have shown, from these findings, that the maximum-likelihood subspace nucleation procedure is a numerically feasible procedure for obtaining the valid optimal reconstruction subspace that contains the unknown quantum state and, at the same time, maximizes the (log-)likelihood with respect to the measured data. Throughout the procedure, no other assumptions about the source are required. The complete elimination of this requirement turns our proposed procedure into an extremely robust method for real experiments, where such assumptions are sometimes difficult to justify precisely. The reporting of all results on experimental state reconstructions and diagnostics using continuous-variable measurement schemes can now be done more reliably once this restriction is lifted, since the concern of reconstruction artifacts that typically arise from an unsuitable or a suboptimal choice of reconstruction subspace is now out of the picture.

The methods of cross-validation and bootstrapping are used to justify the appropriate size of the optimal reconstruction subspace by investigating its predictive power of future data from the same measurement scheme. Other statistical tools can also be invoked depending on the way the observer uses the resulting quantum-state estimator. In general, all these statistical tools would have to be improved in order to address statistical problems related to the quantum-state space, as the positivity constraint plays an important role in altering the probability distribution of any set of data generated from a quantum state, which would in general be different from its classical counterpart. The study of the implications of the positivity constraint on these statistical methods is beyond the scope of this article.

It should be emphasized that the nucleation methodology is completely general and applicable to quantum-state estimation strategies that are not necessarily invoking the maximum-likelihood principle. Very similar nucleation procedures may be implemented for strategies such as linear-inversion or weighted linear-inversion, for instance. The only difference is that the objective function is no longer the likelihood function, but some other function compatible with the chosen estimation strategy, and the quantum positive constraint can additionally be imposed on all such strategies. The subspace nucleation procedures for these strategies proceed as usual otherwise. The bottom line—the set of data obtained with any CV measurement scheme is the only essential element for an accurate subspace and state reconstruction.

Methods

Detailed numerical procedure for the ML subspace nucleation process.

In this section, we shall also assume that the largest possible subspace for an efficient reconstruction has dimension defined by some large integer $D_{\mathrm{lim}} \gg d$—the limit for the state reconstruction. The "lth (reconstruction) subspace of dimension d" can therefore be synonymously understood as the D_{lim}-dimensional projection operator $S_{l,d}$.

The nucleation process for a particular CV measurement scheme makes use of a list of L seed subspaces $\{S_{l,d}\}_{l=1}^{L}$ of a pre-chosen dimension d. As an example, we shall take $d=2$ and $D_{\mathrm{lim}}=16$, which are the settings for the simulations. Each ($D_{\mathrm{lim}}=16$)-dimensional projector $S_{l,2}$ is used to compute the maximum log-likelihood with respect to the data. The operators involved in this computation are the state $\rho_{l,d=2}$ and all the POM outcomes $\Pi_j^{(l,d=2)}$ on this particular qubit subspace. More explicitly, for a given l, the two-dimensional $\rho_{l,2}$ is simply represented as a two-dimensional positive, unit-trace matrix defined as

$$\rho_{l,2} = \frac{A_{l,2}^{\dagger} A_{l,2}}{\mathrm{tr}\{A_{l,2}^{\dagger} A_{l,2}\}} \tag{3}$$

using an auxiliary complex operator $A_{l,2}$. The jth outcome $\Pi_j^{(l,2)}$ residing on this subspace is represented by a positive 2×2 matrix extracted out of the original 16×16 positive matrix $\Pi_j \mathrel{\hat{=}} M^{(j)}$ describing this outcome. For instance, suppose that $S_{l,2}$ is the 16×16 diagonal matrix having only two "ones" respectively for the second and fifth diagonal entries. Then the 2×2 positive matrix is simply

$$\Pi_j^{(l,2)} \mathrel{\hat{=}} \begin{pmatrix} M_{2,2}^{(j)} & M_{2,5}^{(j)} \\ M_{5,2}^{(j)} & M_{5,5}^{(j)} \end{pmatrix}. \tag{4}$$

Positivity in $\Pi_j^{(l,2)}$ is trivially preserved for every j since this matrix is just the matrix representing $S_{l,d}\Pi_j S_{l,d}$ with only matrix elements on the relevant subspace retained. The sum of all $\Pi_j^{(l,2)}$ s is typically not the identity. The ML method regarding such cases are discussed in, for instance, refs 1 and 2. Once the two-dimensional ML estimator $\hat{\rho}_{\mathrm{ML}}$ for every value of l is computed, the maximal log-likelihood values are then sorted in descending value and the subspace that yields the largest maximal log-likelihood value is then chosen to seed the nucleation process. The set of $\binom{D_{\mathrm{lim}} = 16}{d = 2} = 120$ projectors is then reduced to the set of $\binom{14}{2} = 91$ projectors which now corresponds to a set of subspaces that are orthogonal to the optimal subspace.

The next larger ($d=4$)-dimensional ML subspace is built from this seed by accomodating the optimal qubit seed subspace that is both orthogonal to the current subspace and maximizes the log-likelihood. The subsequent computation is very similar to that described for the previous case, only that $\rho_{l,4}$ and $\Pi_j^{(l,4)}$ are now four-dimensional operators. After this computation, the set of $\binom{14}{2} = 91$ projectors is then reduced to the set of $\binom{12}{2} = 66$ projectors that are orthogonal to all the selected projectors. The computation rate for this numerical scheme increases with each step as the set of seed subspaces on which ML estimation is performed decreases in size. The procedure continues in this manner until $\hat{\rho}_{\mathrm{ML}}$ fulfils some fixed criterion that would eventually terminate the nucleation process.

Cross-validation and bootstrapping.

Cross-validation. If the observer wants to use $\hat{\rho}_{\mathrm{ML}}$ to predict future measurement data, then the technique of cross-validation is a suitable approach to verify if this ML estimator is predictive. Here, cross-validation is used to verify its predictive power on at least the same measurement scheme. A common technique known as K-fold cross-validation involves the splitting of a set of M data into K datasets of equal size. A total of $K-1$ datasets are chosen as training sets to obtain an ML estimator $\hat{\rho}_{\mathrm{ML}}$. The remaining dataset, the testing set, is then used to test whether $\hat{\rho}_{\mathrm{ML}}$ gives ML probabilities that are close to these data on average. Other variants of cross-validation exists, some of which possess high computational complexities[32]. So far, no systematic studies of cross-validation has been performed for quantum tomography, as such the implications of the positivity constraint, if any, on the quantum-state space are not known. For the simulations, K is set to two to ensure that both the training set and testing set are equally large enough. For the specifications of typical homodyne experiments, the binned data are suitable for numerical computation for this value of K.

The predictive power of $\hat{\rho}_{\mathrm{ML}}$ is summarized by the prediction error

$$\mathrm{PrErr} = \frac{1}{M}\sum_{k=1}^{K}\sum_{j=1}^{M/K} \frac{(n_j/N - \hat{p}_j^{(\mathrm{ML})})^2}{\hat{p}_j^{(\mathrm{ML})}}\Bigg|_{k\text{th testing set}}, \tag{5}$$

where, without loss of generality, we have assumed that M is divisible by K. For a sufficiently large reconstruction subspace, PrErr would in principle approach zero if not for the slight statistical fluctuations of the measurement data. We mention in passing that in the case where a source drift is present, the true state of the source is no longer stable and describing the source with a single ML estimator using all the measurement data would result in an average bias for PrErr.

Bootstrapping. Since PrErr is statistical, it is in principle necessary to assign some statistical quantifier to it. Any statistical quantifier that describes the reliability of PrErr would generally require a sample of PrErr values for each $D_{\rm rec}$. With only one set of data, a viable option is to perform bootstrapping on this set of data to generate new sets of pseudodata for the construction of the quantifier. Without the assumption of a model for bootstrapping, the non-parametric bootstrap method is suitable and has been proven to give sample points that follow a distribution close to the population distribution. However, this convergence comes with strings attached, such as the adherence to a list of other assumptions, and these assumptions are not always satisfied for some cases, especially in the presence of the quantum positivity constraint.

A workaround is to suggest that since the PrErr decreases with increasing $D_{\rm rec}$ (if N is large enough that is), we may take the $\hat{\rho}_{\rm ML}$ estimator with the smallest PrErr for bootstrapping—the parametric bootstrapping strategy. This choice of model for the bootstrap data asymptotically guarantees that the resulting bootstrap distribution of random PrErr values converges to the actual population distribution from the true state as long as $N \gg 1$ (typical situation in CV experiments) and PrErr $\ll 1$. The procedure for generating a PrErr value from a set of pseudodata obtained from a run of parametric bootstrapping is exactly the same as in the case of real data. Parametric bootstrapping is then repeated to accumulate a sample of bootstrap PrErr values for each $D_{\rm rec}$.

The quantifier chosen as an example is the confidence interval that representatively quantifies the confidence level for each PrErr value. For a given significance level $0 < \alpha < 1$ that is small, we first compute the $1 - \alpha/2$ and $\alpha/2$ percentiles from each bootstrap sample. Upon denoting the percentiles respectively by $\mathrm{PrErr}_{1-\alpha/2}$ and $\mathrm{PrErr}_{\alpha/2}$, the confidence interval is defined as the percentile interval $[2\,\mathrm{PrErr} - \mathrm{PrErr}_{1-\alpha/2}, 2\,\mathrm{PrErr} - \mathrm{PrErr}_{\alpha/2}]$. The advantage of this interval is that it is computationally efficient and captures approximately some essence of the sample dstributions. A more accurate interval can be acquired by performing a second-level bootstrapping for the standard deviation of each sample, which is often computationally intractable. As an estimate for the spread of PrErr, the percentile confidence interval provides sufficiently reliable information for general purposes. Besides, other statistical information is usually needed to supplement this interval for a more thorough data analysis.

References

1. Teo, Y. S. *Introduction to Quantum-State Estimation* Ch. 1, 1–5 (World Scientific Publishing Co., 2015).
2. Paris, M. & Řeháček, J. *Quantum State Estimation* (eds Paris, M. & Řeháček, J.) Ch. 1, 1–4 (Springer, 2004).
3. Ralph, T. C. Continuous variable quantum cryptography. *Phys. Rev. A* **61,** 010303(R) (1999).
4. Hillery, M. Quantum cryptography with squeezed states. *Phys. Rev. A* **61,** 022309 (2000).
5. Gottesman, D. & Preskill, J. Secure quantum key distribution using squeezed states. *Phys. Rev. A* **63,** 022309 (2001).
6. Silberhorn, C., Ralph, T. C., Lütkenhaus, N. & Leuchs, G. Continuous Variable Quantum Cryptography: Beating the 3 dB Loss Limit. *Phys. Rev. Lett.* **89,** 167901 (2002).
7. Navascués, M., Grosshans, F. & Acín, A. Optimality of Gaussian Attacks in Continuous-Variable Quantum Cryptography. *Phys. Rev. Lett.* **97,** 190502 (2006).
8. Aspelmeyer, M., Kippenberg, T. J. & Marquardt, F. Cavity optomechanics. *Rev. Mod. Phys.* **86,** 1391 (2014).
9. Woolley, M. J. & Clerk, A. A. Two-mode squeezed states in cavity optomechanics via engineering of a single reservoir. *Phys. Rev. A* **89,** 063805 (2014).
10. Eberle, T. *et al.* Quantum Enhancement of the Zero-Area Sagnac Interferometer Topology for Gravitational Wave Detection. *Phys. Rev. Lett.* **104,** 251102 (2010).
11. Demkowicz-Dobrzański, R., Banaszek, K. & Schnabel, R. Fundamental quantum interferometry bound for the squeezed-light-enhanced gravitational wave detector GEO 600. *Phys. Rev. A* **88,** 041802(R) (2013).
12. Vaidman, L. Teleportation of quantum states. *Phys. Rev. A* **49,** 1473 (1994).
13. Braunstein, S. L. & Kimble, H. J. Teleportation of Continuous Quantum Variables. *Phys. Rev. Lett.* **80,** 869 (1998).
14. Ban, M. Quantum dense coding via a two-mode squeezed-vacuum state. *J. Opt. B: Quantum Semiclass. Opt.* **1,** L9 (1999).
15. Braunstein, S. L. & Kimble, H. J. Dense coding for continuous variables. *Phys. Rev. A* **61,** 042302 (2000).
16. Cerf, N. J. & Iblisdir, S. Optimal *N*-to-*M* cloning of conjugate quantum variables. *Phys. Rev. A* **62,** 040301(R) (2000).
17. Braunstein, S. L., Cerf, N. J., Iblisdir, S., van Loock, P. & Massar, S. Optimal Cloning of Coherent States with a Linear Amplifier and Beam Splitters. *Phys. Rev. Lett.* **86,** 4938 (2000).
18. Sych, D., Řeháček, J., Hradil, Z., Leuchs, G. & Sánchez-Soto, L. L. Informational completeness of continuous-variable measurements. *Phys. Rev. A* **86,** 052123 (2012).
19. Řeháček, J., Hradil, Z., Knill, E. & Lvovsky, A. I. Diluted maximum-likelihood algorithm for quantum tomography. *Phys. Rev. A* **75,** 042108 (2007).
20. Banaszek, K., D'Ariano, G. M., Paris, M. G. A. & Sacchi, M. F. Maximum-likelihood estimation of the density matrix. *Phys. Rev. A* **61,** 010304(R) (2000).
21. Teo, Y. S., Zhu, H., Englert, B.-G., Řeháček, J. & Hradil, Z. Quantum-State Reconstruction by Maximizing Likelihood and Entropy. *Phys. Rev. Lett.* **107,** 020404 (2011).
22. Teo, Y. S., Englert, B.-G., Řeháček, J. & Hradil, Z. Adaptive schemes for incomplete quantum process tomography. *Phys. Rev. A* **84,** 062125 (2011).
23. Anraku, K. An information criterion for parameters under a simple order restriction. *Biometrika* **86,** 141 (1999).
24. Hughes, A. W. & King, M. L. Model selection using AIC in the presence of one-sided information. *J. Stat. Plan. Inference* **115,** 397 (2003).
25. Usami, K., Nambu, Y., Tsuda, Y., Matsumoto, K. & Nakamura, K. Accuracy of quantum-state estimation utilizing Akaike's information criterion. *Phys. Rev. A* **68,** 022314 (2003).
26. Guţă, M., Kypraios, T. & Dryden, I. Rank-based model selection for multiple ions quantum tomography. *New J. Phys.* **14,** 105002 (2012).
27. Yin, J. O. S. & van Enk, S. J. Information criteria for efficient quantum state estimation. *Phys. Rev. A* **83,** 062110 (2011).
28. Granade, C., Combes, J. & Cory, D. G. Practical Bayesian tomography. *New J. Phys.* **18,** 033024 (2016).
29. Ferrie, C. Quantum model averaging. *New J. Phys.* **16,** 093035 (2014).
30. Artiles, L. M., Gill, R. & Guţă, M. An invitation to quantum tomography. *J. Roy. Statist. Soc. B* **67,** 109 (2005).

31. Bajorski, P. *Statistics for Imaging, Optics, and Photonics* (Wiley and SPIE Press, 2012).
32. Nadeau, C. & Bengio, Y. Inference for the generalization error. *Mach. Learn.* **52,** 239 (2003).
33. Grandvalet, Y. & Bengio, Y. Hypothesis Testing for Cross-Validation. *Montreal Universite de Montreal, Operationnelle DdIeR* **1285** (2006).
34. Mogilevtsev, D., Hradil, Z., Řeháček, J. & Shchesnovich, V. S. Cross-Validated Tomography. *Phys. Rev. Lett.* **111,** 120403 (2013).
35. Bengtsson, I. & Życzkowski, K. *Geometry of Quantum States—An Introduction to Quantum Entanglement* Ch. 14, 352–358 (Cambridge University Press, 2006).

Acknowledgements

Y.S.T. acknowledges support of the BK21 Plus Program (21A20131111123) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF). A.M. and D.M. acknowledge the support of the National Academy of Sciences of Belarus through the program "Convergence", the project 686731 SUPERTWIN of Horizon 2020, and FAPESP grant 2014/21188-0 (D.M.). Finally, Y.S.T., J.Ř., and Z.H. acknowledge the support of the Grant Agency of the Czech Republic (Grant No. 15-03194S), and the IGA Project of the Palacký University (Grant No. PRF 2015-002).

Author Contributions

D.M. and Y.S.T. contributed to the development of the theory. The manuscript was written by Y.S.T., with input and discussions from all other authors. A.M., J.Ř., and Z.H. supported and enhanced the research work.

Additional Information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Teo, Y. S. *et al.* Crystallizing highly-likely subspaces that contain an unknown quantum state of light. *Sci. Rep.* **6**, 38123; doi: 10.1038/srep38123 (2016).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.