Rank-based model selection for multiple ions quantum tomography

2012 New J. Phys. 14 105002

(http://iopscience.iop.org/1367-2630/14/10/105002)

New Journal of Physics
The open-access journal for physics

Rank-based model selection for multiple ions quantum tomography

Mădălin Guţă[1,3], **Theodore Kypraios**[1] **and Ian Dryden**[1,2]

[1] School of Mathematical Sciences, University of Nottingham, University Park, NG7 2RD Nottingham, UK

[2] Department of Statistics, University of South Carolina, Columbia, SC 29208, USA

E-mail: madalin.guta@nottingham.ac.uk

New Journal of Physics **14** (2012) 105002 (26pp)
Received 18 June 2012
Published 1 October 2012
Online at http://www.njp.org/
doi:10.1088/1367-2630/14/10/105002

Abstract. The statistical analysis of measurement data has become a key component of many quantum engineering experiments. As standard full state tomography becomes unfeasible for large dimensional quantum systems, one needs to exploit prior information and the 'sparsity' properties of the experimental state in order to reduce the dimensionality of the estimation problem. In this paper we propose model selection as a general principle for finding the simplest, or most parsimonious explanation of the data, by fitting different models and choosing the estimator with the best trade-off between likelihood fit and model complexity. We apply two well established model selection methods—the Akaike information criterion (AIC) and the Bayesian information criterion (BIC)—two models consisting of states of fixed rank and datasets such as are currently produced in multiple ions experiments. We test the performance of AIC and BIC on randomly chosen low rank states of four ions, and study the dependence of the selected rank with the number of measurement repetitions for one ion states. We then apply the methods to real data from a four ions experiment aimed at creating a Smolin state of rank 4. By applying the two methods together with the Pearson χ^2 test we conclude that the data can be

[3] Author to whom any correspondence should be addressed.

suitably described with a model whose rank is between 7 and 9. Additionally we find that the mean square error of the maximum likelihood estimator for pure states is close to that of the optimal over all possible measurements.

Contents

1. Introduction

Recent years have witnessed significant progress in the engineering and control of quantum systems [1–3]. From the preparation of exotic quantum states [4–7] to the implementation of accurate quantum protocols [8–11] experimentalists are confronted with the problem of reconstructing such mathematical objects *statistically*, from the outcomes of repeated measurements. The theoretical and experimental challenges have stimulated the development of a large array of new methods at the boundary between quantum theory and statistics: state estimation (or tomography) [12–17], tomography for incomplete data [18–20], permutationally invariant tomography [21, 22], design of experiments [23–25], quantum process and detector tomography [26, 27] construction of confidence regions (error bars) [28, 29], quantum tests [30–32], entanglement estimation [33], quantum homodyne tomography [34–36], asymptotic theory [37–39]; see also the monographs [40, 41] and the collections of papers [42, 43].

The importance and difficulty of quantum state tomography became evident in the landmark experiment [7] where entangled states of up to eight ions were created and fully characterized. More recently the same group succeeded in creating entangled states of 14 ions [44] but their statistical reconstruction is beyond current computational capabilities! Therefore, there is great interest in alternative methods aimed at reducing the dimensionality of the state estimation problem without making unwarranted or unrealistic assumptions. Among these we mention the development of quantum compressed sensing methods [45, 46]

which extend the 'classical' ℓ_1-minimization algorithms [47, 48] to the quantum set-up, and the estimation of many-body states based on lower dimensional families of matrix product states [49]. Both methods rely on the ansatz that the states produced in real experiments are not completely arbitrary, but have some *sparsity* structure that can be exploited for more efficient estimation, e.g. low rank in the first case and finite correlations in the second.

In this paper we propose and investigate a state tomography method which can also take advantage of the sparsity structure of the state, by adjusting the *rank* of the estimator (number of non-zero eigenvalues) according to the measurement data. However, although it shares with compressed sensing the goal of exploiting sparsity structures, our method is closer to the standard tomography set-up in the sense that it takes as input the dataset consisting of *measurement counts* rather than *estimates of observables expectations*, and it uses maximum likelihood (ML) for determining the estimator of a given rank. The philosophy of *rank-based model selection* is to choose an estimator which offers a good fit to the data, but at the same time contains a minimal number of parameters (Occam razor principle). For this, we construct a sequence of models consisting of states of fixed rank, and choose the model whose maximum likelihood estimator (MLE) achieves the best trade-off between fit (likelihood) and model complexity. To quantify the trade-off we use two model selection methods, the Akaike information criterion (AIC) [50] and the Bayesian information criterion (BIC) [51] which have been used extensively in model selection problems; see [52, 53] for an introduction to model selection methods, and [54–56] for applications in quantum statistics.

Although the method can be used for an arbitrary measurement set-up, we focus on the statistical model of multiple ions tomography (MIT) [7, 44], which constitutes a physically relevant testing ground for tomography of large dimensional systems. We emphasize that model selection does not assume any particular model, but rather lets the data select the model which gives the most suitable description.This offers the experimentalist an 'honest' but also minimal estimation framework. Since the states created in many experiments have a good degree of purity, one only needs to compute the ML over spaces of low rank, rather than full rank matrices. Furthermore, the principle of model selection can be applied to other families of models such as matrix product states, which may be more suitable in specific experimental conditions.

The paper is organized as follows. In section 2 we introduce the statistical model of MIT, and discuss some of the existing estimation methods. To gain more insight, in section 3 we investigate the problem of estimating *pure states* and in particular we find that the MIT measurement set-up is quasi-optimal in the sense that the mean norm-two distance squared is only slightly larger than that of the optimal measurement, asymptotically with the number of measured copies. Section 4 introduces the two rank-based model selection procedures based on the AIC and BIC, and discusses the implementation of the fixed rank models by using the Cholesky decomposition. The methods are applied to three randomly chosen states of ranks 1, 2 and 3 in section 5. We find that both criteria perform very well when the strictly positive eigenvalues are significant relative to the number of measurement samples, and explain this by analysing the asymptotics of the log-likelihood ratio statistic for different models. In section 6 we investigate the dependence of the selected rank on purity and number of measurement repetitions in a one ion study. In section 7 we apply BIC and AIC to experimental data provided by Rainer Blatt's group from the University of Innsbruck. We find that the maximum log-likelihood flattens from rank 10 (see figure 5), and the AIC and BIC predict ranks 9 and respectively 6, which capture the four significant eigenvalues and some of the eigenvalues of order 10^{-2} (see table 3). As an additional check, we use the Pearson χ^2 statistic to test the

hypothesis H_0 that state has rank of at most 10, and find that there is no evidence to reject H_0. Section 8 contains a summary of the paper and an outlook for future work.

The focus of our paper is to analyse the behaviour of the model selection methods and the dependence on the eigenvalues of the states, and the number of measurement repetitions. For these purposes it suffices to consider states of up to four ions. We leave it for the future work to explore how much the method can be pushed towards higher number of ions, of relevance to current experiments.

In several occasions we refer to 'asymptotics' as the body of statistical theory dealing with the behaviour of estimators and other statistics when the number of samples (measurement repetitions) tends to infinity [58, 66]. The advantage of asymptotic analysis is that it often offers a clearer view of the problem, revealing the generic, universal features such as the asymptotic normality of the MLE, the χ^2-distribution of certain test statistics, etc. The relevance of asymptotic results to high dimensional models with constrained parameter spaces such as encountered in quantum tomography, has been justly debated and needs to be carefully considered on a case by case basis. In this paper we contribute towards this goal with examples on both sides sides of the debate.

2. Background on multiple ions tomography (MIT)

In this section we review the statistical model describing the measurement data collected in MIT experiments [7, 11, 44], and comment briefly on the existing estimation methods, with an emphasis on maximum likelihood estimation. Throughout the paper 'hat' is used to denote estimators, e.g. $\hat{\rho}$ is a data dependent estimator of the state ρ.

The physical system consists of an array of trapped ions whose joint state can be manipulated by means of precisely tuned laser pulses. Since only two electronic energy levels are used for encoding the state, each ion can be describe mathematically as a two level system, so that the joint Hilbert space of k ions is $\mathbb{C}^{2^k}$. The state of the system is described by a density matrix ρ on this space, i.e. a $2^k \times 2^k$ complex selfadjoint matrix which is positive semidefinite and has trace one. Typically, the goal of the experiment is demonstrate the preparation of a certain target state to a sufficiently high degree of precision. To validate the result, a large number of preparation-measurement cycles are performed, and the collected measurement data are used to estimate the state produced in the preparation phase.

In a nutshell, the measurement procedure consists of performing simultaneous Pauli measurements on all ions, each combination of Pauli observables being repeatedly measured n times. More precisely, each measurement is defined by a setting $\mathbf{d}$ which specifies which of the three Pauli observables $\sigma_x, \sigma_y, \sigma_z$ is measured for each ion. For instance $\mathbf{d} := (x, y, z, z)$ is a four ions measurement setting, and in general for a k-ions state there are 3^k possible settings $\mathbf{d} \in \mathcal{D}_k := \{x, y, z\}^k$. For each fixed setting, the measurement produces random outcomes $\mathbf{s} \in \mathcal{O}_k := \{+1, -1\}^k$ with probability distribution

$$\mathbb{P}_\rho(\mathbf{s}|\mathbf{d}) := \mathrm{Tr}(\rho P_{\mathbf{s}}^{\mathbf{d}}) = \langle e_{\mathbf{s}}^{\mathbf{d}}|\rho|e_{\mathbf{s}}^{\mathbf{d}}\rangle, \tag{1}$$

where $P_{\mathbf{s}}^{\mathbf{d}}$ are one dimensional projections onto the vectors of the orthonormal basis

$$|e_{\mathbf{s}}^{\mathbf{d}}\rangle := |e_{s_1}^{d_1}\rangle \otimes \cdots \otimes |e_{s_k}^{d_k}\rangle, \quad \mathbf{s} \in \mathcal{O}_k := \{+1, -1\}^k, \tag{2}$$

formed by taking tensor products of eigenvectors of the Pauli matrices $\sigma_{d_1}, \ldots, \sigma_{d_k}$:

$$\sigma_d|e_s^d\rangle = s\,|e_s^d\rangle, \quad d \in \{x, y, z\}, \quad s \in \{+1, -1\}.$$

After repeating n times the measurement with setting d, the data can be summarized by counting the number of times that each possible outcome has occurred. The probability of a certain set of counts $\{N(\mathbf{s}|\mathbf{d}) : \mathbf{s} \in \mathcal{O}_k\}$ is given by the multinomial distribution with probabilities given by (1), so that

$$\mathbb{P}_\rho(\{N(\mathbf{s}|\mathbf{d}) : \mathbf{s} \in \mathcal{O}_k\}) = \frac{n!}{\prod_s N(\mathbf{s}|\mathbf{d})!} \prod_{\mathbf{s}} \mathbb{P}_\rho(\mathbf{s}|\mathbf{d})^{N(\mathbf{s}|\mathbf{d})}, \quad \mathbf{d} \in \mathcal{D}_k. \tag{3}$$

Since any given setting $\mathbf{d}$ gives information only about the diagonal of the density matrix ρ with respect to the basis (2), the above procedure is repeated for all possible settings to obtain the complete $2^k \cdot 3^k$ dataset consisting of counts $\{N(\mathbf{s}|\mathbf{d}) : (\mathbf{s}, \mathbf{d}) \in \mathcal{O}_k \times \mathcal{D}_k\}$ for all outcomes in each setting. As successive preparation-measurement cycles are independent of each other, the distribution over all possible datasets is the product of multinomials,

$$\mathbb{P}_\rho(\{N(\mathbf{s}|\mathbf{d}) : (\mathbf{s}, \mathbf{d}) \in \mathcal{O}_k \times \mathcal{D}_k\}) = \prod_{\mathbf{d}} \mathbb{P}_\rho(\{N(\mathbf{s}|\mathbf{d}) : \mathbf{s} \in \mathcal{O}_k\}). \tag{4}$$

Let us ponder for a moment on the structure of this statistical model. If no assumption is made on the state, the parameter space is the $(4^k - 1)$-dimensional convex set of density matrices $\mathcal{S}_k \subset M(\mathbb{C}^{2^k})$. We will verify that the above measurement scheme is *informationally complete*, or equivalently that the parameter ρ is *identifiable* in the sense that there is a one-to-one correspondence between ρ and the probability distribution $\mathbb{P}_\rho$ given in (4). Since $\{\sigma_x, \sigma_y, \sigma_z, \sigma_0 := \mathbf{1}\}$ form a basis in the space of 2×2 selfadjoint matrices, the tensor products

$$\tilde{\sigma}_{\mathbf{i}} := \frac{1}{2^{k/2}} \sigma_{i_1} \otimes \ldots \otimes \sigma_{i_k}, \quad \mathbf{i} := (i_1, \ldots, i_k) \in \{x, y, z, 0\}^k$$

form an orthonormal basis of the space of $2^k \times 2^k$ selfadjoint matrices with respect to the inner product $\langle A, B\rangle := \mathrm{Tr}(AB)$. Therefore, any state can be expanded as

$$\rho = \sum_i \rho_{\mathbf{i}} \tilde{\sigma}_{\mathbf{i}} := \sum_{\mathbf{i}} \langle \tilde{\sigma}_{\mathbf{i}}, \rho\rangle \tilde{\sigma}_{\mathbf{i}}, \tag{5}$$

and to estimate ρ it suffices to estimate the Fourier coefficients $\rho_{\mathbf{i}}$. A naive unbiased estimator can be easily constructed based on the counts of any particular measurement setting $\mathbf{d}$ for which $d_j = \mathbf{i}_j$ whenever $\mathbf{i}_j \neq 0$. For example when $k = 2$, to estimate $\rho_{(x,z)}$ we consider the counts from the setting $\mathbf{d} = (x, z)$, and define

$$\hat{\rho}_{(x,z)} := \frac{1}{\sqrt{2^2 n}} \left[N((+1, +1)|\mathbf{d}) + N((-1, -1)|\mathbf{d}) - N((+1, -1)|\mathbf{d}) - N((-1, +1)|\mathbf{d})\right].$$

While this proves that the state can be fully estimated, the naive estimator is generally not a bona-fide density matrix, and more importantly, has large estimation errors. The latter is due to the fact that $\hat{\rho}_{\mathbf{i}}$ is constructed from the counts of a *single* setting and does not harness the information contained in the counts of the others. Indeed, since the projectors $\{P_{\mathbf{s}}^{\mathbf{d}} : \mathbf{s} \in \mathcal{O}_k, \mathbf{d} \in \mathcal{D}_k\}$ form a (highly) overcomplete set of vectors in $M(\mathbb{C}^{2^k})$, any product of Pauli's $\sigma_{\mathbf{i}}$ can be expressed in (continuously) many ways as a linear combination of projectors, each producing a linear estimator which could in principle be combined to obtain a significantly reduced mean square error (MSE). However, finding the 'optimal linear estimator' is problematic due to the fact that the empirical frequencies $N_{\mathbf{s}}^{\mathbf{d}}/n$ are noisy estimates of the probabilities $\mathbb{P}(\mathbf{s}|\mathbf{d})$, and their covariance depends on the unknown state. An interesting proposal in this direction is the Kalman filter estimator developed in [14], but to our knowledge its performance in the case of

MIT has not been extensively investigated. Another proposal put forward in [57] is to combine the naive estimator with a second stage rank-penalized minimization of the norm-two square (Hilbert–Schmidt) distance to the final estimator.

ML is one of the most commonly used estimation methods across statistics. Its popularity is due to the intuitive interpretation, versatility, and strong theoretical underpinning. Under certain regularity conditions the MLE is asymptotically optimal (or efficient in statistical terminology) in the sense that its covariance achieves the Cramér–Rao bound in the limit of large samples [58], and has normal (Gaussian) limiting distribution, with covariance equal to the inverse of the Fisher information matrix. By discarding the constant factorial term in (3) and taking logarithm we can write the MLE for MIT as

$$\hat{\rho} := \arg\max_{\tau \in \mathcal{S}_k} \sum_{\mathbf{s},\mathbf{d}} N(\mathbf{s}|\mathbf{d}) \log \mathbb{P}_\tau(\mathbf{s}|\mathbf{d}), \tag{6}$$

where the maximum is computed over the set $\mathcal{S}_k$ of k-ions states τ. Note that the MLE is invariant under reparametrization, i.e. the ML estimator of a state functional $f := f(\rho)$ is $f(\hat{\rho})$. The MLE has been used extensively in quantum statistics [59], and an efficient iterative computational routine has been put forward in [60, 61]. Nevertheless, ML has been criticized for several perceived drawbacks [12, 13]. The first criticism is that the ML has the tendency to produce rank deficient estimators, i.e. which have some zero eigenvalues, when the true state has some small eigenvalues; this can be understood [12] by observing that the likelihood (seen as a function of the matrix elements) may attain its maximum at a point which lies outside the convex space of states $\mathcal{S}_k$, in which case the 'constrained' MLE $\hat{\rho}$ will fall on boundary of $\mathcal{S}_k$ by the concavity of the log-likelihood function. The second, and in our opinion more serious criticism is that the standard asymptotic theory does not apply as such to states which lie on the boundary, more precisely the estimated parameters are not normally distributed around the truth for large sample sizes. Note however that asymptotic normality *does* hold when restricting to pure states models as we will show in the next section, and also holds for the *unconstrained* MLE, for 'generic' states which satisfy $\mathbb{P}_\rho(\mathbf{s}|\mathbf{d}) > 0$ for all $\mathbf{s}, \mathbf{d}$. This may be used to prove the existence of the asymptotic distribution of the MLE (6), but the latter is likely to be complicated and impractical for establishing confidence regions (error bars). In this paper we focus on the performance of the proposed model selection estimation method, and refer to [28, 29] for two recent proposals for constructing confidence regions, and the forthcoming paper [62] for a comparative study of bootstrap and Fisher information methods.

3. Estimation of pure states in the MIT setting

Pure states are arguably the golden standard of most state preparation experiments [7, 44]. Therefore, we will start by considering the ideal situation in which the quantum state is assumed to be pure, and we would like to estimate it using the MIT dataset described in section 2. The goal is to get more insight into the statistical power of the measurement set-up and its asymptotic properties. This will prepare the ground for the next section where the purity assumption is lifted and the state is fitted to models of increasing rank, and model selection criteria are used for choosing a rank with a good trade-off between fit and model complexity. The findings are summarized at the end of the section, where we also clarify the relation to the compressed sensing set-up [45, 46] which uses as input the estimated values $\hat{\rho}_{\mathbf{i}}$ of the expectations of Pauli observables $\sigma_{\mathbf{i}}$.

The pure states MLE $\hat{\rho}$ can be computed as in (6), with the maximization restricted to the space of pure (rank one) states on $\mathbb{C}^{2^k}$. As figure of merit we consider the MSE

$$\mathrm{MSE}(\hat{\rho}) := \mathbb{E}(\|\rho - \hat{\rho}\|_2^2)$$

with the norm-two distance squared defined as

$$\|\rho - \hat{\rho}\|_2^2 := \sum_{i,j=1}^{2^k} |\rho_{i,j} - \hat{\rho}_{i,j}|^2 = \sum_{\mathbf{i}} |\rho_{\mathbf{i}} - \hat{\rho}_{\mathbf{i}}|^2, \tag{7}$$

where ρ_i are the Fourier coefficients with respect to the Pauli basis defined in (5). Note that for pure states the norm-two distance is related in a simple way to the (arguably more natural) norm-one distance $\|\rho - \hat{\rho}\|_1 := \mathrm{Tr}(|\rho - \hat{\rho}|)$ by the equality $\|\rho - \hat{\rho}\|_2 = \|\rho - \hat{\rho}\|_1/\sqrt{2}$.

We would like to address the following questions:

1. What is the MSE of the MLE?
2. Are we in an 'asymptotic regime'?
3. Is the multiple ions measurement 'optimal' in any sense?

The pure states form a compact manifold of dimension $2(2^k - 1)$ which can be identified with the complex projective space $\mathbb{C}P^{2^k+1}$. Therefore, when restricting the MIT statistical model to pure states, the standard asymptotic efficiency theory [58] is applicable. For simplicity, we assume that $|\psi\rangle$ has the expansion with respect to the standard basis $|\psi\rangle = \sum c_i |e_i\rangle$ such that $c_1 \neq 0$, in which case we can parametrize the state by the real and the imaginary parts of the remaining coefficients

$$\theta \to |\psi_\theta\rangle = \sqrt{1 - \|\theta\|^2}|e_1\rangle + \sum_{j=2}^{2^k} (\theta_j + \mathrm{i}\,\theta_{2^k-2+j})|e_j\rangle, \quad \theta \in \mathbb{R}^{2(2^k-1)}, \quad \|\theta\| < 1.$$

Note that due to the geometry of the projective space, any global parametrization must be singular unless some points are cut out as we did here. However, as we are interested in the asymptotic behaviour of the MLE, the global properties are unimportant and we can always choose an appropriate local parametrization for all practical purposes. The norm two-square distance (7) can be rewritten locally as a quadratic form [63]

$$\|\rho_\theta - \rho_{\hat{\theta}}\|_2^2 = (\hat{\theta} - \theta)G(\theta)(\hat{\theta} - \theta)^{\mathrm{t}} + o(\|\hat{\theta} - \theta\|^2), \tag{8}$$

where $G(\theta)$ is a positive definite matrix whose explicit form can be easily computed, and superscript 't' denotes the transpose. The MLE $\hat{\theta} = \hat{\theta}_n$ is efficient, i.e. as $n \to \infty$ its renormalized error converges in distribution (or law) to a normal

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} N(0, I(\theta)^{-1}), \tag{9}$$

where $N(0, I(\theta)^{-1}))$ is the centred normal distribution with covariance matrix $I(\theta)^{-1}$ which is the inverse of the (classical) Fisher information matrix $I(\theta)$. In particular, from (8) and (9) we get

$$\lim_{n\to\infty} \mathbb{E}(\|\rho_\theta - \rho_{\hat{\theta}}\|_2^2) = \mathrm{Tr}(G(\theta)I^{-1}(\theta)). \tag{10}$$

To verify these results we simulated 100 datasets from a fixed but randomly chosen pure state of $k = 4$ ions, each dataset consisting of counts for $n = 100$ measurements per setting. Figure 1 shows the histogram of the square error $\|\rho_\theta - \rho_{\hat{\theta}}\|_2^2$ whose empirically estimated MSE

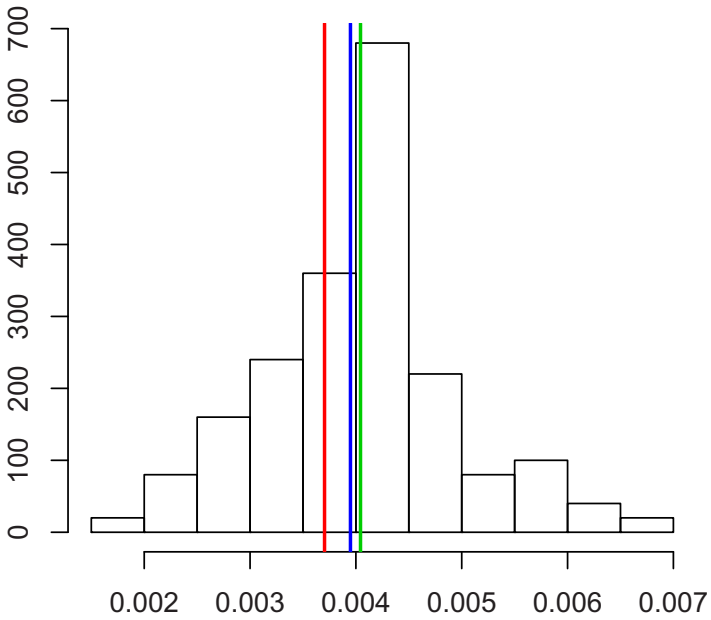

Figure 1. Histogram of the norm-two error $\|\rho - \hat{\rho}\|_2^2$ of the MLE $\hat{\rho}$ for 100 samples from a fixed pure state ρ. The mean square error (green line) is very close to the classical Cramér–Rao bound (blue line) as predicted by asymptotic theory, and the latter is only slightly larger than the 'quantum optimal bound' (red line), showing that for pure states the ions measurement is almost almost optimal among *all* measurements.

(green line) is very close to the asymptotic prediction (blue line) computed from (10) which is equal to 3.9×10^{-3}. More interestingly, we find that the MSE is also very close to the 'quantum optimal bound' (red line) which describes the best MSE achievable with *any* measurement! The latter can be obtained by using the machinery of quantum Cramér–Rao theory and is given by the simple formula (see [64] and references therein)

$$\mathrm{QMSE} = \frac{\sharp\text{parameters}}{\sharp\text{samples}} = \frac{2(2^k - 1)}{3^k n} = \frac{2(2^4 - 1)}{3^4 \times 100} = 3.7 \times 10^{-3}. \tag{11}$$

This example shows that the MSE of the MLE agrees with the asymptotic theory when $n = 100$ and $k = 4$; we expect that the same holds for fixed n and larger k due to the favourable scaling of the number of samples $n \cdot 3^k$ with respect to the number of parameters $2(2^k - 1)$. Moreover, the multiple ions measurement set-up appears to be quasi-optimal. This implies that adaptive strategies (where the measurement settings are chosen in a way that depends on previous measurement outcomes) cannot offer a significant improvement, but does not exclude the possibility that a similar performance can be achieved with a fraction of the settings. To further emphasize the point that the MIT dataset is *very informative*, and that ML can optimally extract this information from the data, we compare the above results with the MSE of the naive estimator discussed in section 2, and with the asymptotic MSE of a dataset consisting of *estimates of the Pauli products* obtained by lumping together the counts of each measurement setting into a single statistic.

3.1. Mean square error (MSE) of the naive estimator

With the square error defined as in (7), we note that the MSE of each coefficient $\hat{\rho}_{\mathbf{i}}$ for which $i_1, \ldots, i_k \neq 0$, is of the order $1/(n \cdot 2^k)$ since we are essentially dealing with the problem of estimating the mean of a random variable with values $\{+2^{-k/2}, -2^{-k/2}\}$. Therefore these coefficients alone (not counting those for which some i_j are zero) bring a contribution of the order $3^k/(n \cdot 2^k)$ which is larger than QMSE (11) by a factor $(9/4)^k/2$. For the particular example of $k = 4$ and $n = 100$ this gives an MSE of 5×10^{-2} which is an order of magnitude larger than that of the MLE.

3.2. MSE of the coarse grained data

At this point, it is natural to ask the following question. Suppose that we are given the 3^k empirical averages of the Pauli products $\sigma_{\mathbf{i}}$

$$\hat{\rho}_{\mathbf{i}} \approx \mathrm{Tr}(\rho\tilde{\sigma}_{\mathbf{i}}) = \langle\psi|\tilde{\sigma}_{\mathbf{i}}|\psi\rangle, \quad i_1, \ldots, i_k \neq 0 \tag{12}$$

which are obtained by computing one empirical average for each column of the original dataset. Is there a more efficient method to estimate the pure state $|\psi\rangle$, from the data (12) and what is its MSE? Two important candidates are the *compressed sensing* and *lasso* algorithms [46] (with the slight difference that they would use a smaller number of settings, but proportionally more measurements per setting). Both methods aims at estimating the state by trying to match the empirical expectations $\hat{\rho}_{\mathbf{i}}$ with those of a selfadjoint matrix, while at the same time penalizing the trace norm of the matrix. Testing these methods is beyond the scope of this paper, but the asymptotic efficiency theory offers a shortcut to the answer of the above question. Applying the same methodology as before, but to the coarse grained data (12) we can predict that (asymptotically) the MSE of any estimator is bounded from below by that of the MLE $\hat{\rho}_{\mathrm{cg}}$ which in turn satisfies

$$\lim_{n\to\infty} n\mathbb{E}\left(\|\hat{\rho}_{\mathrm{cg}} - \rho\|_2^2\right) = \mathrm{Tr}(G(\theta)I_{\mathrm{cg}}(\theta)^{-1}), \tag{13}$$

where the only difference with (11) is the Fisher information matrix which satisfies the inequality $I_{\mathrm{cg}}(\theta) \leqslant I(\theta)$. Figure 2 shows histograms of the asymptotic MSE (11) for the full MIT data (left panel) versus the MSE (13) of the coarse grained data (right panel). The histograms were produced with 250 randomly chosen pure states, $k = 4$ and $n = 100$. Note that the MSE of the coarse grained data is smaller than the (partial) estimated contribution of the naive estimator. However, the MSE is still an order of magnitude higher that that of the full dataset, due to the fact that a significant amount of information has been discarded in the process of retaining the Pauli products expectations.

To summarize, we conclude that MIT works because the different settings 'overlap' with each other in the sense that the one dimensional projections $|e_{\mathbf{d}}^{\mathbf{s}}\rangle\langle e_{\mathbf{d}}^{\mathbf{s}}|$ form an overcomplete set of size $2^k \times 3^k$ which is significantly larger than the dimension of the space of matrices 4^k even for small k. Therefore, the measurement data is structured so that the counts for each setting provide a relatively small amount of information, but the dataset as a whole is very informative about the state. Reducing this dataset to a small number of expectations may be advantageous for the purpose of devising fast estimation algorithms, but underperforms from the viewpoint of statistical errors, for a given number of state re-preparations. This statement may seem to contradict the simulation results illustrated in figure 1 of [46], where compressed

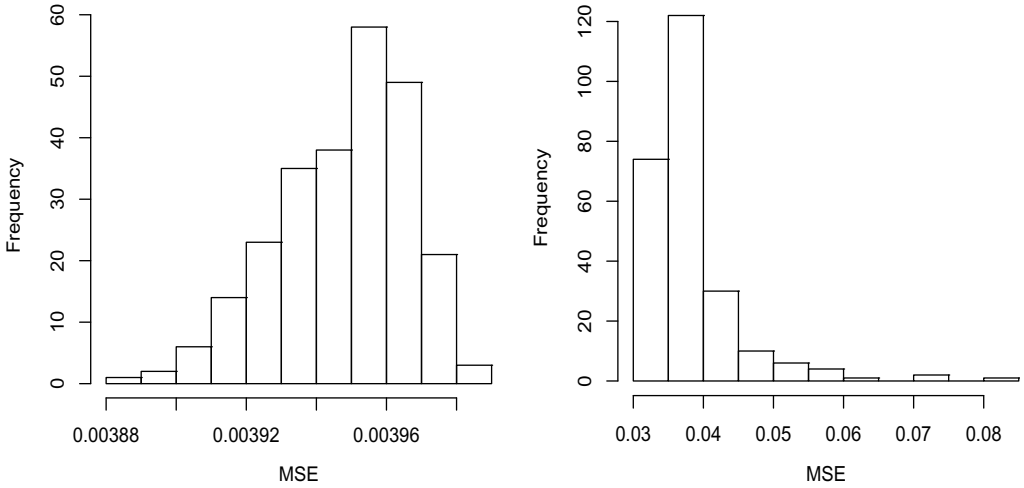

Figure 2. Histograms of asymptotic MSE's for 250 randomly chosen pure states, with $k = 4$ and $n = 100$. Left panel: full counts dataset. Right panel: coarse grained dataset. Keeping only the empirical means of the Pauli products leads to a 10-fold increase in the MSE.

sensing and lasso are found to perform *better* than ML on datasets of the type (12). This apparent contradiction is lifted by the following observations:

(1) Our comparison is between the MSE of efficient estimators for two *different types of data*. Based on this we conclude that *any* estimator using the coarse grained data will asymptotically underperform ML based on the full counts experimental dataset.

(2) The comparison in [46] is different; it regards the performance of ML versus compressed sensing and lasso for the coarse grained data. Since a completely unknown state is *not identifiable* for the coarse grained model, the MLE is *not consistent*, and arguably should not be used in this case.

4. Model selection for quantum tomography

In the previous sections we discussed the extreme scenarios of 'full' quantum tomography and estimation of pure states. In reality, the states produced in experiments tend to have one or few significant eigenvalues and a large number of small eigenvalues of different orders of magnitude, which account for the imperfections in the preparation procedure. Therefore, neither of the two settings seem to be suitable: the former underfits the real state while the latter overfits by trying to estimate eigenvalues that may not be statistically significant.

This is a well known phenomenon in statistics, that occurs in high (or infinite) dimensional problems such as estimating the probability density of a real valued random variable, or in nonlinear regression where an unknown function is 'learned' from estimates of its values at certain points. In such cases the MLE overfits the data and is very 'noisy'. A possible solution is to use a MLE, which maximizes the difference between the log-likelihood and a penalty measuring the complexity of the estimator, e.g. the number of non-zero coefficients with respect to an appropriate basis. More generally, one can design a class of statistical models with various degrees of complexity, and decide which model and which estimator from that model is most

suitable for describing the data. Our aim is to apply the model selection methodology to state tomography, the models being the families of states of a given rank. The same methods can be used in tasks such as quantum homodyne tomography [34] where the state to be estimated is that of a light pulse (one mode continuous variables system), and a model could be the set of states with a given maximum number of photons [65].

To select the rank of the state we will use two well established methods: the AIC [50] and the BIC [51]. Both methods amount to penalizing the log-likelihood function by a factor proportional to the dimension of the model, and choosing the MLE with the smallest value of the information criterion. In the next section we give a brief general description of AIC and BIC, after which we discuss the parametrization of the fixed rank quantum models, and the implementation of the model selection procedure.

4.1. Akaike information criterion (AIC) versus Bayesian information criterion (BIC) model selection

Occam's razor is an old scientific principle which states that when trying to explain a phenomenon, one should choose *the simplest* model that adequately fits the data. A very complex model will be able to fit the given data almost perfectly but it will not be able to generalize very well. On the other hand, very simple models will not be able to describe the essential features of the data. Therefore, we must make a compromise and choose a model which is as simple as possible, but no simpler. It is not surprising that many approaches have been proposed over the years for dealing with this key aspect in statistical modelling.

The general framework of model selection is the following. We are given n samples $\mathbf{X} = \{X_1, \ldots, X_n\}$ from some unknown distribution $\mathbb{P}$ which we try to fit with a distribution from one of several possible statistical models

$$\mathcal{M}_r := \{\mathbb{P}_{\theta_r} : \theta_r \in \Theta_r \subset \mathbb{R}^{p(r)}\}, \quad r = 1, \ldots, D,$$

where $\mathcal{M}_r$ has a parameter θ_r of dimension $p(r)$. For simplicity we assume that $X_i \in \{1, \ldots, a\}$ are discrete random variables, as in the case of MIT measurements. We also assume that at least one of the D models contains the true distribution, or at least gives a reasonable approximation to it. We denote by $\hat{\theta}_r$ the MLE for the model $\mathcal{M}_r$, and by $\ell_{\theta_r} = \ell_{\theta_r}(\mathbf{X}) := \sum_{i=1}^{n} \log \mathbb{P}_{\theta_r}(X_i)$ log-likelihood function at θ_r.

4.1.1. Akaike information criterion.

The AIC for model $\mathcal{M}_r$ is [50]

$$\mathrm{AIC}(r) = -2\ell_{\hat{\theta}_r} + 2p(r),$$

and the chosen model is the one with the minimum AIC. Since $p(r)$ is larger for more complex models, the AIC formally biases against overly complicated models. Although the derivation of AIC is outside the scope of this paper, we briefly explain the idea behind the choice of penalty. Having computed the MLEs for different models we would like to select the 'best' one in the sense that the corresponding distribution $\mathbb{P}_{\hat{\theta}_r}$ is the closest to the 'truth' $\mathbb{P}$ with respect to the Kullback–Leibler distance (or relative entropy)

$$K(\mathbb{P}|\mathbb{P}_{\hat{\theta}_r}) := \sum_{i=1}^{a} \mathbb{P}(i) \log(\mathbb{P}(i)) - \sum_{i=1}^{a} \mathbb{P}(i) \log(\mathbb{P}_{\hat{\theta}_r}(i)).$$

However, this quantity cannot be computed since $\mathbb{P}$ is unknown. Since the first terms on the right side is the same for all models, it can be neglected, and one can focus on estimating the second

term, which nevertheless still depends on $\mathbb{P}$. If instead of $\hat{\theta}_r$ we had a fixed parameter θ_r, this term would be the expected value of the log-likelihood at θ_r and could be estimated by $\ell_{\theta_r}(X)/n$, by the law of large numbers. However $\ell_{\hat{\theta}_r}(\mathbf{X})/n$ is a biased estimator of the second term, due to the fact that the data has been already used in computing $\hat{\theta}_r$. Akaike showed that under the regularity conditions required by the asymptotic normality theory, the bias is approximately $p(r)/n$, so that ML which is the closest to the truth is approximately given by the minimizer of the AIC.

4.1.2. Bayesian information criterion. The BIC for model $\mathcal{M}_r$ is defined as [51]

$$\mathrm{BIC}(r) = -2\ell_{\hat{\theta}_r} + p(r)\log(n),$$

where n is the sample size. Note that the BIC differs from the AIC only in the second term which increases with n, so that BIC favours simpler models (that is models with a smaller number of parameters) compared to AIC. But despite the superficial similarity between the AIC and BIC the latter is derived in a very different way, within a Bayesian framework.

For simplicity, suppose that there are two competing models, $\mathcal{M}_1$ and $\mathcal{M}_2$ with parameters θ_1 and θ_2 respectively. One begins by assigning prior probabilities q_1 and $q_2 = 1 - q_1$ to the event that the observed data have been generated from either model. One also assigns prior distributions $\pi_1(\theta_1)$ and $\pi_2(\theta_2)$ to the model parameters in each model. Then one can compute the marginal likelihoods which can be interpreted as the probability of observing the data if model $\mathcal{M}_i$ is correct, having integrated out our ignorance about the parameters θ_1 and θ_2 in each model. Hence, one can apply Bayes theorem to evaluate the probability of model $\mathcal{M}_i$ being the true model given the observed data. A measure of the extent to which the data support model $\mathcal{M}_2$ over $\mathcal{M}_1$ is given by the *posterior odds*

$$\frac{\mathbb{P}(\mathcal{M}_2|\mathbf{X})}{\mathbb{P}(\mathcal{M}_1|\mathbf{X})} = \frac{\mathbb{P}(\mathbf{X}|\mathcal{M}_2)}{\mathbb{P}(\mathbf{X}|\mathcal{M}_1)}\frac{q_2}{q_1}.$$

The first fraction on the right-hand side is called the *Bayes factor* and the second is known as the prior odds. The Bayes factor is a fundamental quantity in Bayesian theory and can be interpreted as a measure of the extent to which the data support model $\mathcal{M}_2$ over $\mathcal{M}_1$ when the prior odds are equal to one. The difference $\mathrm{BIC}(1) - \mathrm{BIC}(2)$ can be shown to be a large sample approximation to the logarithm of the Bayes factor, so that the second model is chosen if the difference is positive.

4.2. Parametrizing models with fixed rank

Here we describe the fixed rank models which will be used in model selection. Let $\mathcal{D}(d, r)$ be the set of rank r states of a d-dimensional quantum system, i.e. those states which have exactly r non-zero eigenvalues, and let

$$\mathcal{R}(d, r) := \bigcup_{i=1}^{r} \mathcal{D}(d, r)$$

be the set of states of rank at most r. Every state ρ has a unique spectral decomposition

$$\rho = \sum_{i=1}^{r} \lambda_i P_i,$$

where $\lambda_i > 0$ are its *distinct* eigenvalues, and P_i is an eigenprojector whose dimension is equal to the multiplicity m_i of λ_i. The spectral information $(\lambda_1, P_1, \ldots, \lambda_r, P_r)$ can be used to construct a parametrization of $\mathcal{D}(d, r)$ and $\mathcal{R}(d, r)$, which has the advantage of a direct physical interpretation. However, the practical implementation of such a parametrization for computing the MLE is less straightforward due to the orthogonality constraints for the eigenvectors, and the singularities appearing on lower dimensional manifolds consisting of states with non-trivial sets of multiplicities. A variation on this would be to parametrize the state by the set of eigenvalues and an eigenbasis, in which case the singularity problem is replaced by the *non-identifiability* of the different basis vectors corresponding to the same eigenvalue.

We will describe an alternative parametrization which is related to the Cholesky factorization of the state. Recall that any *positive definite* matrix $A \in M(\mathbb{C}^d)$ has a unique decomposition

$$A = T^* T, \tag{14}$$

where T is an upper triangular matrix with strictly positive diagonal elements. Therefore there exists a one-to-one correspondence between full-rank states ρ and matrices T as described above, with the additional constraint

$$\mathrm{Tr}(T^* T) = \sum_{ij} |T_{ij}|^2 = \mathrm{Tr}(\rho) = 1. \tag{15}$$

We parametrize such a matrix T by the vector of real numbers $\theta := (R, I, D) \in \mathbb{R}^{d^2-1}$ with

$$\begin{cases} R := (\mathrm{Re}(T_{12}), \ldots, \mathrm{Re}(T_{d-1,d})), \\ I := (\mathrm{Im}(T_{12}), \ldots, \mathrm{Im}(T_{d-1,d})), \\ D := (T_{22}, \ldots, T_{dd}), \end{cases} \tag{16}$$

such that R, I are the real and imaginary parts of the off-diagonal elements ordered from the first to the $d-1$ row, and from left to right along each row. By (14) and (15), θ must satisfy the constraints $D > 0$ and $\|R\|^2 + \|I\|^2 + \|D\|^2 < 1$, and the left-top element of T is equal to

$$T_{11} = T_{11}(\theta) = (1 - \|R\|^2 + \|I\|^2 + \|D\|^2)^{1/2} > 0.$$

The Cholesky parametrization of the full rank matrices can be extended, albeit with some caveats, to the spaces of rank-deficient matrices $\mathcal{D}(d, r)$ and $\mathcal{R}(d, r)$. The idea is to consider a decomposition as in (14), but with T belonging to the set $\mathcal{T}^+(d, r)$ of $d \times d$ upper triangular matrices with the bottom $d - r$ rows equal to zero, and satisfying $T_{11}, \ldots, T_{rr} > 0$; equivalently, one can consider $r \times d$ trapezoidal matrices obtained by removing the zero lines of the triangular matrices. Since every $T \in \mathcal{T}^+(d, r)$ is of rank r, this guarantees that the corresponding state ρ has the same property. However, not all states of rank r can be decomposed in this way! Indeed it is easy to verify that if $\rho = T^* T$ then the $r \times r$ top-left principal minor of ρ must be of rank k, and therefore such a parametrization excludes states in $\mathcal{D}(d, r)$ which do not satisfy this property. Nevertheless, 'generic' matrices of rank r *do* have principal minors of rank r, in the sense that those with smaller rank principal minors form a lower dimensional subset of $\mathcal{D}(d, r)$. If we restrict our attention to the subset $\mathcal{D}(d, r)^+ \subset \mathcal{D}(d, r)$ which excludes the 'deficient' states, we find that the Cholesky decomposition exists and is unique, so that

$$\mathcal{D}(d, r)^+ := \{\rho = T^* T : T \in \mathcal{T}^+_{d,r}\} \subset \mathcal{D}(d, r).$$

What can we say about the complement $\mathcal{D}(d, r) \setminus \mathcal{D}(d, r)^+$? In order to have a Cholesky decomposition we need to relax the condition $T_{11}, \ldots, T_{rr} > 0$ and consider the set $\mathcal{T}(d, r)$ of

r-lines upper triangular matrices, with non-negative elements on the diagonal. In this case, the root T not only exists but is in general not unique.

Let $\Theta(d,r)^+$ be the set of real parameters $\theta := (R, I, D)$ of a matrix $T = T_\theta \in \mathcal{T}(d,r)^+$ which are defined similarly to equation (16), and let $\Theta(d,r)$ be the set of parameters associated to matrices in $\mathcal{D}(k,r)$. We define two *sequences* of quantum statistical models:

$$\mathcal{Q}^+(d,r) := \{\rho_\theta = T_\theta^* T_\theta : \theta \in \Theta(d,r)^+\}, \quad r = 1,\ldots,d, \tag{17}$$

$$\mathcal{Q}(d,r) := \{\rho_\theta = T_\theta^* T_\theta : \theta \in \Theta(d,r)\}, \quad r = 1,\ldots,d, \tag{18}$$

the first one consisting of rank r matrices with rank r principal minor, the second one describing (albeit not always uniquely) all matrices of rank up to r. The reason why we mention the two models is that each has some appealing features and some disadvantages. For $\mathcal{Q}(d,r)$ the advantage is that we deal with a *nested* set of models

$$\mathcal{Q}(d,1) \subset \mathcal{Q}(d,2) \subset \cdots \subset \mathcal{Q}(d,d).$$

The disadvantage is that the Cholesky parametrization is not one-to-one in this case. On the other hand, $\mathcal{Q}(r,d)^+$ offers a one-to-one parametrization of rank r matrices in $\mathcal{D}(d,r)^+$, with the disadvantage that the models are not nested, but instead $\mathcal{Q}(d,r)^+$ lies on the boundary of $\mathcal{Q}(d,r+1)^+$. While these facts are relevant to a theoretical analysis, for practical purposes the distinction between the two models is less important, and in all our numerical experiments we used the models $\mathcal{Q}^+(d,r)$.

4.3. The implementation of AIC and BIC model selection for rank-based models

We return now to the state estimation problem, and describe how AIC and BIC model selection is applied to the family of rank-based models described above for a system consisting of k ions, i.e. $d = 2^k$. Let

$$\ell_\theta = \ell_\theta\left(\{N(\mathbf{s}|\mathbf{d}) : \mathbf{s} \in \mathcal{O}_k, \mathbf{d} \in \mathcal{D}_k\}\right) := \sum_{\mathbf{s},\mathbf{d}} N(\mathbf{s}|\mathbf{d}) \log \mathbb{P}_{\rho_\theta}(\mathbf{s}|\mathbf{d})$$

be the log-likelihood of the measurement data, ignoring the constant factorial terms. The MLEs $\hat{\theta}_r$ and $\hat{\rho}_r$ for the model $\mathcal{Q}(2^k,r)^+$ are

$$\hat{\theta}_r := \arg \max_{\theta \in \Theta(2^k,r)^+} \ell_\theta, \quad \hat{\rho}_r := \rho_{\hat{\theta}_r}.$$

In order to choose between the different models we compute the AIC and the BIC for each rank and select the model with the smallest value. In our case the two criteria are given by

$$\begin{cases} \mathrm{AIC}(r) := -2\ell_{\hat{\theta}_r} + 2p(2^k,r), \\ \mathrm{BIC}(r) := -2\ell_{\hat{\theta}_r} + p(2^k,r)\log(n \cdot 3^k), \end{cases} \tag{19}$$

with

$$p(d,r) = 2dr - r^2 - 1, \tag{20}$$

the dimension of the space of rank r matrices, and $n \cdot 3^k$ is the total number of measurements. In practice each criterion decreases with the rank until it reaches the minimum value after which it increases, so one only needs to compute the MLE up to the rank where the criterion begins to increase. For low rank states, this offers the advantage of having to compute the MLE on models of dimension approximately rd rather than $d^2 - 1$ as standard ML. The disadvantage

Table 1. AIC and BIC performance for 100 datasets generated by three randomly chosen states of ranks 1, 2 and 3. The tables shows the number of times AIC and BIC choose rank 1, 2 or 3 for each state.

	AIC rank			
True rank	1	2	3	4
1	82	9	9	0
2	0	74	26	0
3	0	1	80	19

	BIC rank			
True rank	1	2	3	4
1	99	0	1	0
2	7	90	3	0
3	0	5	95	0

is that the likelihood function is not concave as in the full rank model, and may have several local maxima.

To implement the ML estimation numerically, we used a standard maximization routine of the statistics package R. Additionally, we developed an array of statistical analysis tools such as Fisher information, square errors, bootstrap, Pearson χ^2 statistic which will be made available online. Although the computation of the log-likelihood was optimized for faster speed, the maximization can probably be improved significantly by using more sophisticated routines.

In the next sections we will discuss the results of several investigations on the performance of BIC and AIC model selection, using simulated and real data.

5. Study 1: randomly chosen low rank states

In a first simulation study we chose three 'random' states of ranks 1, 2, and 3 of $k = 4$ ions, and generated 100 datasets from each state, each dataset with $n = 100$ measurement repetitions. We then computed the MLEs for the ranks between 1 and 4 and used AIC and BIC to select the optimal rank. The exact procedure used to generate 'random' states is not very important, but it will be relevant that all non-zero eigenvalues of the states are significant. As illustrated in table 1, BIC selected the correct rank for each state in roughly 90% of the cases while for AIC the rate is about 80%. Due to the different penalties, the AIC tends to over-estimate the rank of the state, while BIC has a slight tendency to under-estimate it. While at first sight this may appear to be a surprisingly good performance, we will show that it agrees very well with the predictions of asymptotic theory. For illustration, we consider the state of rank $r = 2$ denoted ρ, and show that the distributions of BIC(3) − BIC(2) and BIC(1) − BIC(2) concentrate on the positive axis, so that BIC chooses the correct rank. Since their behaviours are determined by different mechanisms, we will study each BIC difference separately. A similar analysis can be performed for AIC.

In the first case,

$$\begin{aligned} \mathrm{BIC}(3) - \mathrm{BIC}(2) &= -2(\ell_{\hat{\theta}_3} - \ell_{\hat{\theta}_2}) + \log(n \cdot 3^4)(p(4,3) - p(4,2)) \\ &= -2(\ell_{\hat{\theta}_3} - \ell_{\hat{\theta}_2}) + 242.98, \end{aligned} \tag{21}$$

so the problem is to show that the *log-likelihood ratio statistic*

$$\Lambda := 2(\ell_{\hat{\theta}_3} - \ell_{\hat{\theta}_2})$$

is typically smaller than the penalty 242.98. For 'regular' nested models, the asymptotic distribution of Λ is described by Wilks' theorem as discussed in the appendix. However, this

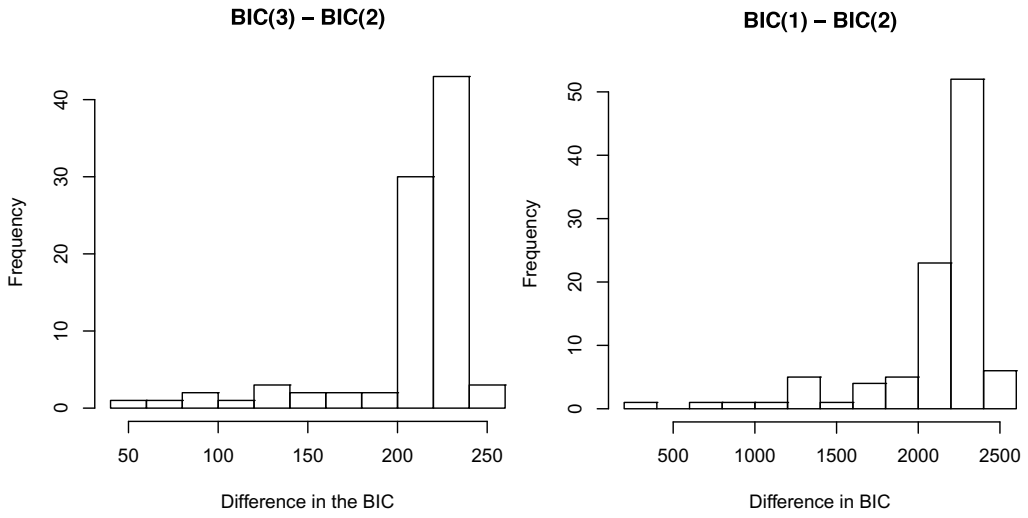

Figure 3. Histogram of BIC differences for the rank 2 state. Left panel BIC(3) − BIC(2); right panel BIC(1) − BIC(2). The values are in good agreement with the asymptotic predictions.

is not directly applicable here since the rank 2 model lies on the boundary of the rank 3 one, due to the positivity constraints. Nevertheless, Wilks' theorem can be extended to more general situations where the two hypotheses can be 'linearized' locally (see [66, chapter 16]), in which case the limiting distribution depends on the local geometry of the two models and the Fisher information at each point. We will not pursue this analysis here but limit ourselves to giving a stochastic upper bound to the limiting distribution which will be sufficient for our purposes. The idea is to note that

$$\Lambda := 2(\ell_{\hat{\theta}_3} - \ell_{\hat{\theta}_2}) \leqslant 2(\ell_{\tilde{\theta}_3} - \ell_{\hat{\theta}_2}), \tag{22}$$

where $\ell_{\tilde{\theta}_r}$ is the 'unconstrained' MLE obtained by maximizing over the space $\tilde{\mathcal{D}}(d, r) \supset \mathcal{D}(d, r)$ consisting of matrices ρ of rank r which are not necessarily positive but must respect the property that $\mathbb{P}_\rho(\cdot|\mathbf{d})$ is a probability distribution for each $\mathbf{d}$. The unconstrained MLE is easier to analyse theoretically and can be used to explain why MLE often produces rank deficient states when the true state has high purity [12]. Now, assuming that we are in the generic situation where all probabilities for the true rank-two state ρ are non-zero, this means that locally around ρ the rank two model is a regular submodel of the extended rank 3 model, and we can apply Wilks' theorem to conclude that

$$2(\ell_{\tilde{\theta}_3} - \ell_{\hat{\theta}_2}) \xrightarrow{\mathcal{L}} \chi^2(p(4,3) - p(4,2)).$$

From (22) we get that Λ is stochastically bounded from above by $\chi^2(27)$ and similarly BIC(3) − BIC(2) is bounded from below by $242.98 - \chi^2(27)$ which agrees with the simulations results illustrated in the left panel of figure 3. Note that as n increases, the probability of BIC choosing the rank 3 model converges to zero due to the presence of the log$n$ factor in the penalty, while AIC is *not rank consistent* in the sense that it choses the higher rank with a probability which does not vanish with n, in agreement with the results illustrated in table 1. Let us consider now the second difference BIC(1) − BIC(2), and note that its values are much larger, as illustrated in the right panel of figure 3. It turns out that in this case the behaviour is

Table 2. Performance of BIC and AIC model selection for three states: pure (state 1), almost pure (state 2), and mixed (state 3). For each choice of number of repetitions, we record the number of times the BIC and AIC select the *correct* rank out of a total of 1000 simulations.

		Measurement repetitions				
		10	50	100	250	500
State 1	BIC	987	990	994	992	996
	AIC	945	944	919	927	930
State 2	BIC	25	83	183	394	706
	AIC	77	312	502	802	942
State 3	BIC	384	973	998	997	988
	AIC	594	992	998	997	988

not dominated by the complexity penalty but by the *bias* of the lower rank model with respect to the 'correct' one, and in particular the distribution of the difference is state dependent. The key is to observe that while the rank 2 MLE $\hat{\rho}_2$ converges to the true state ρ, the rank one MLE $\hat{\rho}_1$ converges to the state ρ_1^* whose corresponding distribution $\mathbb{P}_{\rho_1^*}$ is the closest to the true distribution $\mathbb{P}_\rho$ with respect to the relative entropy (or Kullback–Leibler divergence)

$$\rho_1^* := \operatorname*{arg\,min}_{\tau \in \mathcal{D}(2^4,1)} K(\mathbb{P}_\rho | \, \mathbb{P}_\tau).$$

In conjunction with the law of large numbers we then obtain the almost sure convergence

$$\frac{\Lambda}{2n} = \frac{1}{n}\left(\ell_{\hat{\theta}_2} - \ell_{\hat{\theta}_1}\right) \longrightarrow K\left(\mathbb{P}_\rho \,\middle|\, \mathbb{P}_{\rho_1^*}\right) \quad \text{as } n \to \infty.$$

For our particular example we used one of the rank one MLEs to compute an approximate value $K(\mathbb{P}_\tau | \, \mathbb{P}_\rho) \approx 11.33$ which gives an estimate

$$\begin{aligned} \mathrm{BIC}(1) - \mathrm{BIC}(2) = & -2(\ell_{\hat{\theta}_1} - \ell_{\hat{\theta}_2}) + \log(n \cdot 3^4)(p(4,1) - p(4,2)) \\ & \approx 2 \times 11.33 \times 100 + \log(100 \times 3^4)(p(4,1) - p(4,2)) \\ & \approx 2266 - 261 = 2005, \end{aligned}$$

in agreement with the histogram illustrated in the right panel of figure 3.

In conclusion, for low rank states with eigenvalues which are not very close to zero, the BIC and to lesser extent AIC, identify the correct rank with high probability, the latter having a tendency to overfit the true model. On the other hand, as we will see in the next section, the BIC may underfit the true model when one or more eigenvalues are small.

6. Study 2: one ion simulations

We have seen that the performance of the model selection criteria depends on the spectrum of eigenvalues of the true state, and on the number of measurement repetitions. To investigate this dependence we performed a statistical experiment with three one-ion states ($k = 1$) of different degrees of purity: a pure state, one with eigenvalues $(0.95, 0.05)$, and the other with eigenvalues $(0.72, 0.28)$. For each state we simulated datasets with a varying number of repetitions $n = 10, 50, 100, 250, 500$. Table 2 shows the number of times (out of 1000

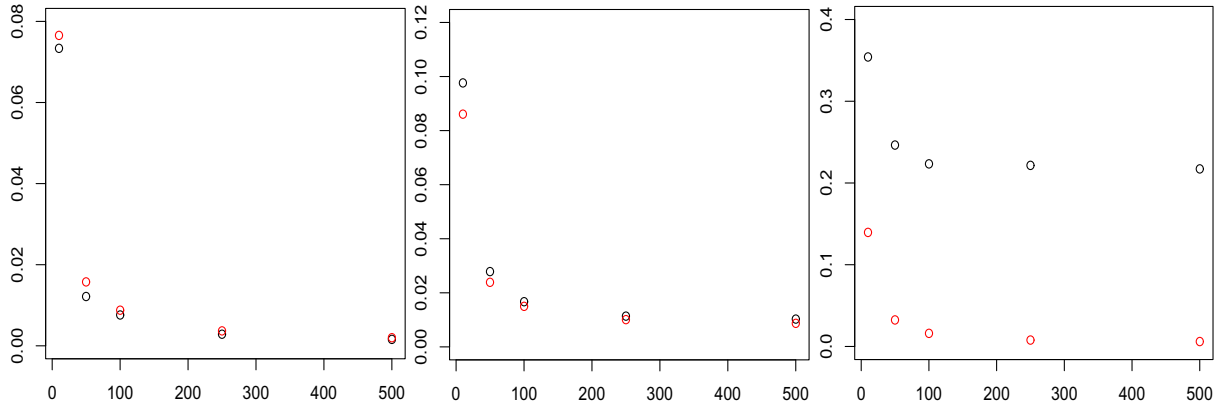

Figure 4. Mean square error for rank 1 (black circle) and rank 2 (red circle) estimators, as function of the number of measurement repetitions $n = 10, 50, 100, 250, 500$. Left: state 1 (pure); middle: state 2 (almost pure); right: state 3 (mixed).

samples) BIC and AIC choose the correct rank of the state, for all possible choices of states and measurement repetitions. As expected, in the case of the the pure and the mixed states both criteria require a small number of repetitions (of the order of 50) to give the correct answer. In the case of the almost pure state, we see a clear dependence with n: for small n the difference between the log-likelihoods does not off-set the complexity penalty and both criteria choose rank one; at $n = 500$ the balance tips in favour of the rank 2 model, with AIC switching faster than BIC, on average.

Figure 4 shows the mean square errors (MSE) of the two MLEs

$$\mathrm{MSE}(r) := \mathbb{E}(\|\hat{\rho}_r - \rho\|_2^2), \quad r = 1, 2$$

as a function of n for each of the three states, with the pure state (rank one) estimator in black and the mixed state (rank two) estimator in red. For the pure state (left panel), the rank two estimator has a larger MSE due to the variance contribution from the third parameter, but the relative difference between the two MSE's is small for all n. In this case the rank one estimator proposed by both criteria is optimal both from the point of view of parsimony, as well as estimation error. For the mixed state (right panel), the rank one estimator has a large bias which dominates the MSE, while the rank two MSE decreases at rate $1/n$, as expected. At $n = 50$ the *relative* difference in risk is significant and both criteria choose the optimal rank-two estimator. The most interesting case is that of the almost pure state (middle panel). Here we see that the relative difference in MSE is not significant for small and medium number of repetitions ($n = 10, 50, 100$), but for larger n the error of the pure state estimator is dominated by its bias while the variance of the full state estimator becomes very small. This behaviour is picked up by the model selection criteria, which on average switch to the more complex model when n is in the interval between 200 and 500.

In conclusion, the study shows that both methods become more sensitive to the true rank of the state as the number of repetitions increases, and the switching point increases (on average) with the purity of the state. As for the estimation error, the switch to the higher rank model appears to happen in the region where the MSE's of the two estimators starts to diverge from each other, which shows that even if the result is suboptimal for small n, the relative difference

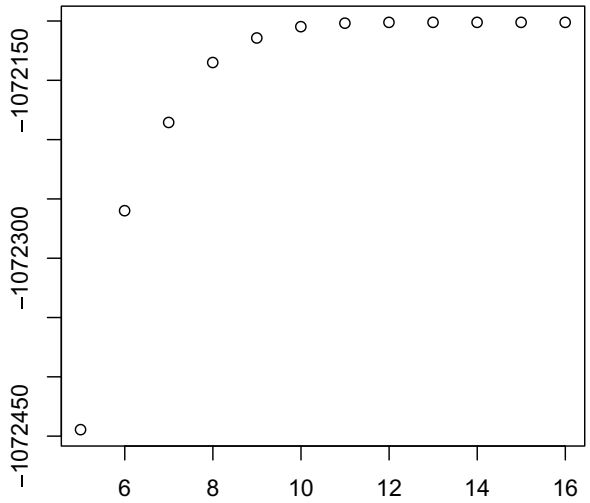

Figure 5. Log-likelihood values for the MLE as a function of rank.

in errors remains small. Finally, the BIC is more aggressive in selecting the lower complexity model, due to the additional log-factor in the penalty.

7. Study 3: model selection for four ions real data

In the third study, we applied the model selection methods to experimental data provided by Rainer Blatt's group from the University of Innsbruck. The aim of the experiment [11] was to create a particular four ions bound entangled state of rank 4 called Smolin state [67], and the measurement dataset consisted of counts for the 3^4 measurement settings, with a number $n = 4800$ of repetitions for each setting. We computed the MLEs $\hat{\rho}_r$ for all ranks r between 1 and 16, and found that the corresponding log-likelihoods reach a plateau at rank 10 (see figure 5) which indicates that the rank 10 model is already sufficiently rich to describe the measurement data. Reinforcing this conclusion, we found that the value of the ML for rank 10 (and the subsequent ones) was slightly *larger* than that of that of the ML over all states computed with Hradil's iterative method [59], probably due to the fact that the latter had not reached the true maximum after 1000 iterations.

The values of the AIC and BIC for all ranks are shown on the left side of table 3. The two criteria reach minima at ranks $r = 6$ and respectively $r = 9$, as a result of the trade-off between the increasing penalty and the log-likelihood. As expected, the BIC chooses a smaller rank due to its larger complexity penalty, but both methods capture the top four eigenvalues of order 10^{-1} and a few of the following ones of order 10^{-2} which account for experimental imprecision in creating the state. On the right side of table 3 we listed for comparison the eigenvalues of the rank 10 and full rank estimators, showing perfect agreement in the first two decimal places. We emphasize that results should be taken as an indication that the experimental data is consistent with models whose rank could be chosen somewhere between 6 and 10, rather than answering the ill posed question 'what is the rank of the state'. To make a more informed decision on the final choice of model, one can additionally use different *model testing* procedures such as the Pearson chi-square test discussed below. As we will see, the various arguments converge towards the conclusion that the rank 6 estimator may be too conservative, while the rank 10 model already fits the data very well.

Table 3. Left: values of AIC and BIC for the MLEs of ranks 1–16. The minimum values of the two criteria are attained at ranks $r = 9$ and respectively $r = 6$. Right: eigenvalues of the MLEs of rank 10 and 16 in decreasing order.

Rank	AIC	BIC	Eigenvalues MLE rank 10	Eigenvalues MLE rank 16
1	2 397 395	2 397 722	2.337×10^{-1}	2.332×10^{-1}
2	2 217 096	2 217 738	2.290×10^{-1}	2.277×10^{-1}
3	2 170 638	2 171 573	2.258×10^{-1}	2.253×10^{-1}
4	2 146 295	2 147 502	1.725×10^{-1}	1.721×10^{-1}
5	2 145 157	2 146 614	4.599×10^{-2}	4.487×10^{-2}
6	2 144 830	**2 146 515**	2.656×10^{-2}	2.445×10^{-2}
7	2 144 719	2 146 611	2.385×10^{-2}	2.229×10^{-2}
8	2 144 652	2 146 728	1.948×10^{-2}	1.884×10^{-2}
9	**2 144 641**	2 146 880	1.226×10^{-2}	1.155×10^{-2}
10	2 144 648	2 147 028	1.067×10^{-2}	1.001×10^{-2}
11	2 144 664	2 147 164	0	6.057×10^{-3}
12	2 144 680	2 147 279	0	2.751×10^{-3}
13	2 144 694	2 147 369	0	6.779×10^{-4}
14	2 144 704	2 147 433	0	5.278×10^{-6}
15	2 144 710	2 147 472	0	2.153×10^{-6}
16	2 144 712	2 147 484	0	1.702×10^{-16}

7.1. Pearson χ^2-test

As an additional tool for probing the conclusions of the model selection procedures, we recast the problem as that of testing between the hypotheses:

$$\begin{cases} H_0 = \text{'the dataset is generated by a state of rank at most } r\text{'}, \\ H_1 = \text{'the dataset is generated by a state of rank higher than } r\text{'}. \end{cases}$$

A standard approach to such a problem is based on using the Pearson χ^2-statistic. Following the general procedure described in the appendix, we consider the rank r MLE $\hat{\rho}_r$ with expected number of counts $E(\mathbf{s}|\mathbf{d}) := n\mathbb{P}_{\hat{\rho}_r}(\mathbf{s}|\mathbf{d})$, and define the Pearson χ^2-statistic

$$T(r) = \sum_{\mathbf{s},\mathbf{d}} \frac{(N(\mathbf{s}|\mathbf{d}) - E(\mathbf{s}|\mathbf{d}))^2}{E(\mathbf{s}|\mathbf{d})}, \tag{23}$$

where $N(\mathbf{s}|\mathbf{d})$ are the counts from the real data. Under the hypothesis H_0, the Pearson statistic has an asymptotic χ^2 distribution with number of degrees of freedom equal to the number of free parameters of the dataset minus the number of parameters of the model

$$\mathrm{df}(r) := 3^4 \times (2^4 - 1) - p(r, 4).$$

Therefore one can define the (asymptotically) level α test

$$\begin{cases} \text{if } T \leqslant t_\alpha: \quad \text{accept } H_0, \\ \text{if } T > t_\alpha: \quad \text{accept } H_1, \end{cases}$$

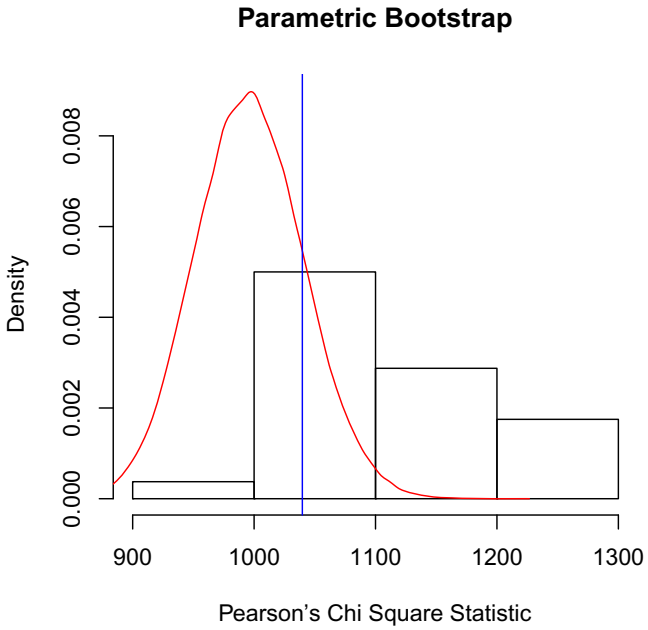

Figure 6. Pearson χ^2 statistic T (blue line), the limit χ^2 distribution (red curve) and the parametric bootstrap distribution for 100 bootstrap samples. The boostrap distribution is shifted with respect to the χ^2 due to the fact that the state is close to the boundary of the states space and the asymptotic theory does not hold. Based on the value of T we conclude that the hypothesis H_0 is not rejected for any reasonable significance level.

where the threshold t_α is chosen such that $\mathbb{P}(Y > t_\alpha) = \alpha$ for a $\chi^2(\mathrm{d}f(r))$-distributed random variable Y. In practice the χ^2 approximation works well for pure states ($r = 1$), and small rank states which have only a few small eigenvalues. However, if the state has a significant number of small eigenvalues, the distribution of $T(r)$ may differ significantly from the asymptotic χ^2 distribution. We will not pursue a theoretical analysis here, but instead use *bootstrap* techniques [58] to estimate the distribution of $T(r)$ and then perform the test with respect to the bootstrap distribution. The idea of bootstrap is to use the measurement data itself to construct a distribution which (under the hypothesis H_0) approximates that of $T(r)$, and therefore can be used to define the threshold t_α instead of the χ^2 distribution.

The bootstrap distributions are constructed as follows:

(1) compute the MLE $\hat{\rho}_r$ and its probability distribution $\mathbb{P}_{\hat{\rho}_r}(\mathbf{s}|\mathbf{d})$;
(2) generate a large number N of independent datasets from the distribution $\mathbb{P}_{\hat{\rho}_r}(\mathbf{s}|\mathbf{d})$ of the maximum likelihood estimation;
(3) compute the MLEs $\hat{\rho}_1^{\mathrm{boot}}, \ldots, \hat{\rho}_N^{\mathrm{boot}}$ for the bootstrap datasets;
(4) compute the Pearson χ^2 statistic for each bootstrap sample and its MLE as in (23);
(5) apply the χ^2 test using the empirical distribution of the boostrap χ^2 statistics.

The results of applying the χ^2 test based on the bootstrap distribution to the rank 10 model are illustrated in figure 6. The value of the test is $T = 1039$ is which means that the hypothesis H_0 is accepted. A similar test for the rank 9 leads to the same conclusion but in the case of the

rank 6 model the hypothesis H_0 is rejected. Therefore, the Pearson test together with the model selection criteria indicate that a good choice for the rank of the estimator is between 7 and 9.

8. Conclusions and outlook

Statistical inference has become a key tool in interpreting the measurement data in quantum engineering experiments, which require precise, efficient and informative estimation methods. However, standard full state tomography becomes unfeasible for large dimensional quantum systems [44]. In this paper we proposed model selection as a general principle for approaching state estimation problems. As in [45, 46, 49] the aim is to reduce the dimensionality of the problem by taking advantage of the 'sparsity' properties of quantum states in realistic experimental settings. The route to this goal is however different. The philosophy in model selection is to try to find the simplest, or most *parsimonious* explanation of the data, by fitting different models (often of increasing complexity) and choosing the estimator with the best trade-off between likelihood and complexity. Concretely, we looked at the problem of selecting the rank of the estimator, by using two well known methods: AIC and the BIC. In both cases the fit-complexity trade-off is realized by penalizing the log-likelihood of the data with a measure of complexity proportional to the number of parameters of the fixed rank model. We have tested AIC and BIC in several real data and simulation studies which we summarize here.

Pure states. We studied the performance of (rank one) ML for pure states and found a very good agreement with the asymptotic predictions based on Fisher information and the efficiency of the MLE. More interestingly, we found that the MSE is only slightly larger than the MSE of the *best* possible measurement predicted by quantum version of the asymptotic theory. In particular this rules out the possibility of significantly improving the MSE by means of adaptive measurement design techniques. The (asymptotic) MSE of the full counts dataset was compared to that of the 'coarse grained' data obtained by estimating the means of the Pauli products corresponding to each measurement setting, as used in compressed sensing algorithms [45, 46]. For four ions, the latter is an order of magnitude larger than the former due to the loss of information when discarding the full counts statistics.

Study 1. For four ions states of ranks between 1 and 3 we found that both AIC and BIC identify the correct rank in 80–90% of the cases, when the smallest non-zero eigenvalue is not too close to zero. The results are explained by using the ML asymptotic theory.

Study 2. We analysed the performance of AIC and BIC as a function of the number of measurement repetitions and the purity of the state, for a toy example consisting of one ion state. With only a small number of repetitions, both methods identify the correct rank for pure and 'pretty mixed' states. For an 'almost pure' state, the model choice switches to rank 2 as the number of repetitions increases. The switching happens roughly at the point where the MSE of the 'wrong' rank 1 estimator becomes significantly larger that that of the correct model, indicating that model selection is only slightly suboptimal in terms of the MSE.

Study 3. We applied model selection to the four ions experimental data provided by Rainer Blatt's group from the University of Innsbruck. The target state of the experiment was an equal mixture of four orthogonal pure states, and BIC and AIC selected rank 6 and respectively 9, with both estimators capturing the principle eigenvalues and (some of) the noisy components due to imperfections in the preparation procedure, of the order 10^{-2}. While the BIC prediction seems too conservative, we find that a rank 10 estimator gives a very good explanation of the data from several perspectives: log-likelihood values, eigenvalues of the estimators, hypothesis testing.

Overall, the numerical results indicate that model selection gives sensible answers, and can be used as an alternative to full tomography and compressed sensing. The results presented in this paper and other ongoing studies which were not included lead to the following conclusions regarding the accuracy of the rank-selected estimators in comparison to standard (full rank) MLE. The former is more accurate (with respect to the MSE) for pure states (cf sections 3 and 6), and states with eigenvalues which are well separated from zero, cf section 5. In the real data example we found that the rank 10 estimator had slightly larger likelihood that the estimator obtained by applying Hradil's iterative method [59], which indicates that their errors are probably very similar. On the other hand, model selection can have higher MSE than standard MLE for small number of measurement repetitions when the state has several very small eigenvalues (cf section 6), due to the bias introduced by the projection onto the lower rank space. However the in-depth one ion study shows that the relative difference in MSE is not significant, with AIC being closer to MSE optimality, in broad agreement with the theoretical properties. A more general study which goes beyond the scope of this paper, should compare these and other rank selection methods, e.g. [57] based on the spectral properties of the state and the number of measurement repetitions.

In principle the rank selection method works for any state, but is designed to take advantage of the lower complexity of small rank states. The drawback is the computational complexity of finding the MLE over states of fixed rank. Therefore it would be interesting to see whether ideas from the different methods can be combined in a fast, scalable and statistically efficient estimator. A possible future direction is apply model selection to state estimation for other types of models such as classes of matrix product states, and to system identification problems. Another topic of interest is the computation of confidence intervals (error-bars). Last but not least, there is a need for a deeper theoretical understanding of the quantum tomography statistical model. We mention two important questions: how does the state's proximity to the boundary affect the standard asymptotic theory, and how does the model behave for a large number of ions? This would hopefully lead to improved estimation algorithms and information criteria for model selection tailored to quantum tomography.

Acknowledgments

We thank Rainer Blatt's group for providing us experimental data, and in particular Thomas Monz and Philipp Schindler for many fruitful discussions and hospitality during our visits to Innsbruck. MG's research is funded by the EPSRC Fellowship EP/E052290/1. MG and TK acknowledge financial support from the University of Nottingham Additional Sponsorship grant EP/J501499/1.

Appendix. Pearson χ^2-statistic and Wilks' theorem

For reader's convenience we collect here two important results used in the paper. We refer to [58, 66] for more details.

Theorem A.1 (Pearson's χ^2 statistic). *Let $X_1, \dots X_n$ be i.i.d. samples from the discrete distribution $\mathbb{P}_\theta$ over $\{1, \dots, p\}$, where $\mathcal{P} := \{\mathbb{P}_\theta : \theta \in \Theta \subset \mathbb{R}^m\}$ is a sufficiently regular model with Θ an open set. Let $N(i)$ be the number of counts of the outcome i in the sample, and let*

$E(i) = n\mathbb{P}_\theta(i)$ be the expected counts. Then, the Pearson χ^2 statistic

$$T := \sum_i \frac{(N(i) - E(i))^2}{E(i)}$$

converges in law as $n \to \infty$ to the χ^2 distribution with m degrees of freedom.

Theorem A.2 (Wilks' Theorem). *Let $\mathcal{P} := \{\mathbb{P}_\theta : \theta \in \Theta = \mathbb{R}^m\}$ be a sufficiently regular model and let $\mathcal{P}_0$ be the submodel with parameter space $\Theta_0 := \{\theta \in \Theta : \theta_1 = \ldots = \theta_k = 0\}$ for some $k \leqslant m$. Let $\mathbf{X} := \{X_1, \ldots, X_n\}$ be i.i.d. samples from $\mathbb{P}_\theta$ and let Λ be the log-likelihood ratio statistic*

$$\Lambda := 2\left[\sup_{\theta' \in \Theta} \ell_{\theta'}(\mathbf{X}) - \sup_{\theta'_0 \in \Theta_0} \ell_{\theta'_0}(\mathbf{X})\right].$$

If $\theta \in \Theta_0$, then Λ converges in law as $n \to \infty$ to the χ^2 distribution with k degrees of freedom.

In both cases, it is essential that the parameter does not lie on the boundary, in order to be able to apply the asymptotic normality theory of the MLE. This condition is violated for states whose rank is strictly smaller than that of the fixed rank model in which they are considered. Therefore care must be taken before applying these results directly, and indeed our results show that the χ^2 asymptotics fail in some cases. A more refined asymptotic analysis taking into account the boundary effects will be pursued elsewhere.

References

[1] Southwell K (ed) 2008 Quantum Coherence *Nature Insight Suppl.* **453** 1003
[2] Haroche S and Raimond J M 2006 *Exploring the Quantum* (Oxford: Oxford University Press)
[3] Dowling J P and Milburn G J 2003 *Phil. Trans. R. Soc.* A **361** 1655
[4] Smithey D T, Beck M, Raymer M G and Faridani A 1993 *Phys. Rev. Lett.* **70** 1244–7
[5] Resch K J, Walther P and Zeilinger A 2005 *Phys. Rev. Lett.* **94** 070402
[6] Zavatta A, Viciani S and Bellini M 2004 *Science* **306** 660–2
[7] Häffner H *et al* 2005 *Nature* **438** 643–6
[8] Altepeter J B, Branning D, Jeffrey E, Wei T C, Kwiat P G, Thew R T, O'Brien J L, Nielsen M A and White A G 2003 *Phys. Rev. Lett.* **90** 193601
[9] O'Brien J L, Pryde G J, White A G, Ralph T C and Branning D 2003 *Nature* **264** 264
[10] Riebe M, Kim K, Schindler P, Monz T, Schmidt P O, Körber T K, Hänsel W, Häffner H, Roos C F and Blatt R 2006 *Phys. Rev. Lett.* **97** 220407
[11] Barreiro J T, Schindler P, Gühne O, Monz T, Chwalla M, Roos C F, Hennrich M and Blatt R 2010 *Nature Phys.* **6** 943–6
[12] Blume-Kohout R 2010 *New J. Phys.* **12** 043034
[13] Blume-Kohout R 2010 *Phys. Rev. Lett.* **105** 200504
[14] Audenaert K M R and Scheel S 2009 *New J. Phys.* **11** 023028
[15] Khoon Ng H and Englert B G 2012 A simple minimax estimator for quantum states arXiv:1202.5136
[16] Smolin J A, Gambetta J M and Smith G 2012 *Phys. Rev. Lett.* **108** 070502
[17] Heinosaari T, Mazzarella L and Wolf M M 2011 Quantum tomography under prior information arXiv:1109.5478v1
[18] Bužek V 2004 *Lect. Notes Phys.* **649** 189
[19] Teo Y S, Zhu H, Englert B G, Rehacek J and Hradil Z 2011 *Phys. Rev. Lett.* **107** 020404
[20] Teo Y S, Englert B G, Rehacek J and Hradil Z 2011 *Phys. Rev.* A **84** 062125

[21] Tóth G, Wieczorek W, Gross D, Krischek R, Schwemmer C and Weinfurter H 2010 *Phys. Rev. Lett.* **105** 250403
[22] Moroder T, Hyllus P, Tóth G, Schwemmer C, Niggebaum A, Gaile S, Gühne O and Weinfurter H 2012 Permutationally invariant state reconstruction arXiv:1205.4941
[23] Smith G A, Silberfarb A, Deutsch I H and Jessen P S 2006 *Phys. Rev. Lett.* **97** 180403
[24] Merkel S T, Riofrío C A, Flammia S T and Deutsch I H 2010 *Phys. Rev.* A **81** 032126
[25] Nunn J, Smith B J, Puentes G, Walmsley I A and Lundeen J S 2010 *Phys. Rev.* A **81** 042109
[26] Rahimi-Keshari S, Scherer A, Mann A, Rezakhani A T, Lvovsky A and Sanders B C 2011 *New J. Phys.* **13** 013006
[27] Lundeen J, Feito A, Coldenstrodt-Ronge H, Pregnell K L, Silberhorn C, Ralph T C, Eisert J, Plenio M B and Walmsley I A 2008 *Phys. Rev.* A **5** 27
[28] Blume-Kohout R 2012 Robust error bars for quantum tomography arXiv:1202.5270
[29] Christandl M and Renner R 2011 Reliable quantum state tomography arXiv:1108.5329v1
[30] Jupp P E, Kim P T, Koo J Y and Pasieka A 2012 Testing for state purity *J. R. Stat. Soc. C: Appl. Stat.* C at press (doi: 10.1111/j.1467-9876.2012.01040.x)
[31] Temme K and Verstraete F 2011 Quantum chi-squared and goodness of fit testing arXiv:1112.6343v1
[32] Moroder T, Kleinmann M, Schindler P, Monz T, Gühne O and Blatt R 2012 Detection of systematic errors in quantum experiments arXiv:1204.3644v1
[33] Landon-Cardinal O and Poulin D 2012 Practical learning method for multi-scale entangled states arXiv:1204.0792
[34] Vogel K and Risken H 1989 *Phys. Rev.* A **40** 2847–9
[35] Leonhardt U, Paul H and D'Ariano G M 1995 *Phys. Rev.* A **52** 4899–907
[36] Lvovsky A I and Raymer M G 2009 *Rev. Mod. Phys.* **81** 299–332
[37] Kahn J and Guţă M 2009 *Commun. Math. Phys.* **289** 597–652
[38] Hayashi M and Matsumoto K 2008 *J. Math. Phys.* **49** 102101
[39] Audenaert K M R, Nussbaum M, Szkola A and Verstraete F 2008 *Commun. Math. Phys.* **279** 251–83
[40] Holevo A S 1982 *Probabilistic and Statistical Aspects of Quantum Theory* (Amsterdam: North-Holland)
[41] Helstrom C W 1976 *Quantum Detection and Estimation Theory* (New York: Academic)
[42] Paris M G A and Řeháček J (ed) 2004 *Quantum State Estimation* (Berlin: Springer)
[43] Hayashi M (ed) 2005 *Asymptotic Theory of Quantum Statistical Inference: Selected Papers* (Singapore: World Scientific)
[44] Monz T, Schindler P, Barreiro J T, Chwalla M, Nigg D, Coish W A, Harlander M, Haensel W, Hennrich M and Blatt R 2011 *Phys. Rev. Lett.* **106** 130506
[45] Gross D, Liu Y K, Flammia S, Becker S and Eisert J 2010 *Phys. Rev. Lett.* **105** 150401
[46] Flammia S T, Gross D, Liu Y K and Eisert J 2012 Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators arXiv:1205.2300
[47] Candés E J and Tao T 2006 *IEEE Trans. Inform. Theory* **52** 5406–25
[48] Tibshirani R 1996 *J. R. Stat. Soc.* B **58** 267–88
[49] Cramer M, Plenio M B, Flammia S, Gross D, Bartlett S D, Somma R, Landon-Cardinal O, Liu Y K and Poulin D 2010 *Nature Commun.* **1** 149
[50] Akaike A 1974 *IEEE Trans. Autom. Control* **19** 716–23
[51] Schwarz G E 1978 *Ann. Stat.* **6** 461–4
[52] Kadane J B and Lazar N A 2004 *J. Am. Stat. Assoc.* **99** 279–90
[53] Zucchini W 2000 *J. Math. Psychol.* **44** 41–6
[54] Kahn J 2009 *ESAIM: Probab. Stat.* **13** 363–99
[55] Usami K, Nambu Y, Tsuda Y, Matsumoto K and Nakamura K 2003 *Phys. Rev.* A **68** 022314
[56] Yin J O S and van Enk S J 2011 *Phys. Rev.* A **83** 062110
[57] Alquier P, Butucea B, Hebiri M and Meziani K 2012 Rank penalized estimation of a quantum system arXiv:1206.1711v1

[58] Young G A and Smith R L 2005 *Essentials of Statistical Inference* (Cambridge: Cambridge University Press)
[59] Hradil Z, Řeháček J, Fiurášek J and Ježek M 2004 *Quantum State Estimation* vol 649 ed M G A Paris and J Řeháček (Berlin: Springer) pp 59–112
[60] Hradil Z 1997 *Phys. Rev.* A **55** R1561
[61] Řeháček J, Hradil Z, Knill E and Lvovsky A I 2007 *Phys. Rev.* A **75** 042108
[62] Monz T, Schindler P, Kypraios T and Guţă M Error estimation and validation in quantum tomography in preparation
[63] Gill R D and Massar S 2000 *Phys. Rev.* A **61** 042312
[64] Gill R D and Guţă M 2011 On asymptotic quantum statistical inference *From Probability to Statistics and Back: High-Dimensional Models and Processes. A Festschrift in Honor of Jon Wellner* (*Lecture Notes Monograph Series*) ed M Banerjee and F Bunea (Hayward, CA: Institute of Mathematical Statistics) at press (arXiv:1112.2078v2)
[65] Artiles L, Gill R D and Guţă M 2005 *J. R. Stat. Soc.* B **67** 109–34
[66] van der Vaart A 1998 *Asymptotic Statistics* (Cambridge: Cambridge University Press)
[67] Smolin J A 2001 *Phys. Rev.* A **63** 032306