Covariance Matrix Preparation for Quantum Principal Component Analysis

Max Hunter Gordon [1,2,*] M. Cerezo,[3] Lukasz Cincio [2,4] and Patrick J. Coles[2]

[1]*Instituto de Física Teórica, UAM/CSIC, Universidad Autónoma de Madrid, Madrid, Spain*

[2]*Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*

[3]*Information Sciences, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*

[4]*Quantum Science Center, Oak Ridge, Tennessee 37931, USA*

Principal component analysis (PCA) is a dimensionality reduction method in data analysis that involves diagonalizing the covariance matrix of the dataset. Recently, quantum algorithms have been formulated for PCA based on diagonalizing a density matrix. These algorithms assume that the covariance matrix can be encoded in a density matrix, but a concrete protocol for this encoding has been lacking. Our work aims to address this gap. Assuming amplitude encoding of the data, with the data given by the ensemble $\{p_i, |\psi_i\rangle\}$, then one can easily prepare the ensemble average density matrix $\overline{\rho} = \sum_i p_i |\psi_i\rangle\langle\psi_i|$. We first show that $\overline{\rho}$ is precisely the covariance matrix whenever the dataset is centered. For quantum datasets, we exploit global phase symmetry to argue that there always exists a centered dataset consistent with $\overline{\rho}$, and hence $\overline{\rho}$ can always be interpreted as a covariance matrix. This provides a simple means for preparing the covariance matrix for arbitrary quantum datasets or centered classical datasets. For uncentered classical datasets, our method is so-called "PCA without centering," which we interpret as PCA on a symmetrized dataset. We argue that this closely corresponds to standard PCA, and we derive equations and inequalities that bound the deviation of the spectrum obtained with our method from that of standard PCA. We numerically illustrate our method for the Modified National Institute of Standards and Technology (MNIST) handwritten digit dataset. We also argue that PCA on quantum datasets is natural and meaningful, and we numerically implement our method for molecular ground-state datasets.

I. INTRODUCTION

Interpreting and analyzing large datasets is a technologically important task. Principal component analysis (PCA) can reduce the dimensionality of large datasets to improve their interpretability while minimizing information loss [1]. PCA was invented by Karl Pearson in 1901 as an analog of the principal axis theorem in mechanics [2]. It can be thought of as fitting an ellipsoid to the data, where the length of each axis of the ellipsoid quantifies the variance of the data along that axis. PCA is widely used in bioinformatics, facial recognition, quantitative finance, and many other applications.

PCA is typically performed by diagonalizing the covariance matrix of the training data and retaining only the largest-eigenvalue eigenvectors. The covariance matrix

quantifies correlations between different features in the data, with matrix elements given by

$$Q_{jk} = E[[X_j - E(X_j)][X_k - E(X_k)]] \tag{1}$$

for random variables X_j and X_k, with E denoting the expectation value. One can see that $Q = [Q_{jk}]$ is a positive semidefinite matrix: $Q \geqslant 0$. For large datasets, Q is often low rank, due to redundancy or correlation between features.

Quantum computers are naturally suited to solve linear algebra problems due to the underlying linear mathematics of quantum mechanics. Indeed, in 2014, Lloyd *et al.* [3] proposed a quantum algorithm for performing PCA called quantum principal component analysis (quantum PCA). Quantum PCA has the possibility of exponential speedup (over classical algorithms) when Q is low rank. The algorithm of Lloyd *et al.* uses multiple copies of a density matrix ρ in order to diagonalize ρ and read off its spectrum. A key assumption in this algorithm is that the covariance matrix Q can be encoded into a density matrix ρ. The plausibility of this assumption arises from

*mhuntergordon@gmail.com

the fact that both Q and ρ are positive semidefinite. Nevertheless, an explicit method for encoding Q into ρ has not been provided.

More recently, near-term approaches to quantum PCA have been developed [4–7] in the framework of variational quantum algorithms [8]. For example, variational quantum state diagonalization [4] unitarily rotates ρ towards a diagonal form, while estimating the distance to being diagonal using two copies of ρ. The variational quantum state eigensolver [9] uses a nondegenerate Hamiltonian to extract the principal components of ρ and only requires a single copy of ρ. Once again, while these variational methods are more near-term, they still do not address the issue of encoding Q in ρ.

We note that studies of quantum PCA are especially timely given that Huang *et al.* [10] recently established that quantum PCA can achieve exponential quantum advantage, at least for quantum data analysis. While classical data are still subject to dequantization arguments [11,12], this does not preclude the possibility of modest quantum speedups for classical data [13]. (We elaborate on these points in the Discussion section.)

In this work, we make a simple but technologically important observation. We consider a dataset that has been amplitude encoded [14–21], such that the dataset is described by an ensemble $\{p_i, |\psi_i\rangle\}$ of normalized quantum states $|\psi_i\rangle$. We note that it is straightforward to use classical randomness in order to prepare the ensemble average density matrix $\overline{\rho} = \sum_i p_i |\psi_i\rangle\langle\psi_i|$, as shown in Fig. 1. We ask the question: how is $\overline{\rho}$ related to the covariance matrix Q? For classical datasets (i.e., datasets stored on classical devices), one can show that we have a precise equality $\overline{\rho} = Q$ whenever the dataset is centered (i.e., the mean values for all features are zero).

For uncentered classical datasets $\overline{\rho} = Q + M$, where M is the outer product of the mean vector with itself. This implies that using $\overline{\rho}$ in place of the true covariance matrix corresponds to performing so-called "PCA without centering" [22], and we provide an interpretation of this as performing PCA on a symmetrized dataset. We derive relations that upper bound the deviation of the eigenvalues and eigenvectors of $\overline{\rho}$ and Q. For example, we show that the eigenvalues of $\overline{\rho}$ are interlaced with those of Q, and any eigenvector of $\overline{\rho}$ is shared with Q as long as it is either orthogonal or parallel to the mean vector.

However, we go a step further and argue that the equality $\overline{\rho} = Q$ holds more generally for all quantum datasets (i.e., datasets prepared on quantum devices from some physical process), due to the unphysical nature of global phase in quantum mechanics. The (somewhat subtle) argument is that any dataset that will be prepared on a quantum device will necessarily lose its global phase information, and consequently one can always assume that such datasets are inherently centered. Therefore, the equality $\overline{\rho} = Q$ holds for all datasets that are prepared on quantum devices.
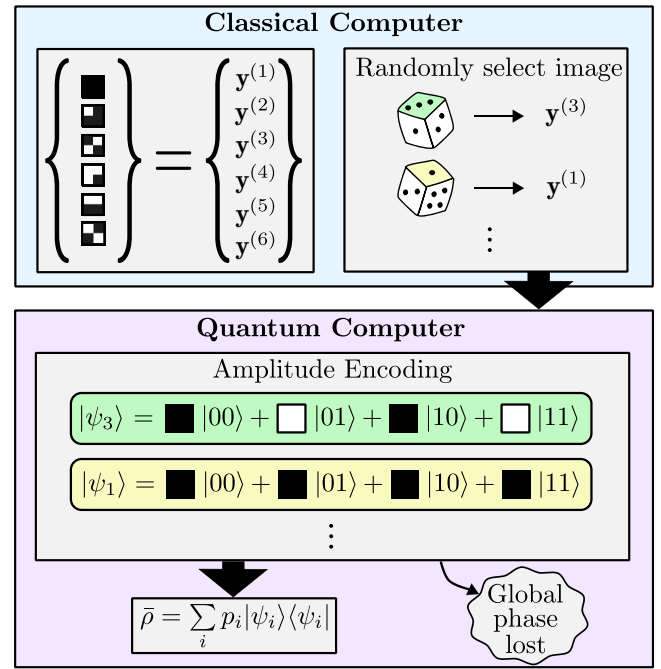

FIG. 1. Preparation of the ensemble average density matrix $\overline{\rho}$. For illustration, we show the well-known bars-and-stripes dataset. To prepare $\overline{\rho}$, images from this dataset are randomly selected according to a probability distribution $\{p_i\}$ and then amplitude encoded on a quantum device. The latter involves encoding each pixel's color into the amplitudes of a quantum state $|\psi_i\rangle$, expanded in the standard basis. Once the data are physically represented on a quantum device, the global phase information is lost and they can be represented as density matrices, of the form $|\psi_i\rangle\langle\psi_i|$. The average state, averaged over many samplings from $\{p_i\}$, is what we call the ensemble average density matrix $\overline{\rho} = \sum_{i=1}^{N} p_i |\psi_i\rangle\langle\psi_i|$. Preparing $\overline{\rho}$ simply adds a linear overhead (linear in the number of datapoints N) to the complexity of state preparation for each datapoint. See Sec. V for a detailed analysis of the complexity of our method when used as a subroutine of quantum PCA algorithms.

As a consequence, for quantum datasets, we provide a simple means to prepare the covariance matrix as a density matrix, filling in the missing ingredient for quantum PCA algorithms.

To illustrate our method, we numerically simulate both PCA and quantum PCA for the MNIST dataset of handwritten digits. In this case, loss of global phase information (due to amplitude encoding) amounts to losing the information about whether the image colors are white or black. Nevertheless, this loss of information is fairly trivial, and quantum PCA gives principal components that are very similar to those given by PCA. Moreover, quantum PCA performs as good as (or better than) PCA at compressing the MNIST dataset into a small number of features.

We also argue that it is natural to apply quantum PCA to quantum datasets, i.e., data that are inherently quantum mechanical and hence that do not require an amplitude

encoding step. The covariance of complex random variables (such as quantum amplitudes) has a clear meaning, and therefore so does PCA on quantum states. Moreover, the loss of global phase information (mentioned above) does not apply to quantum datasets since there is no amplitude encoding step, so quantum PCA for quantum data is even more natural than for classical data. We illustrate quantum PCA for quantum datasets by applying it to a set of molecular ground states for different interatomic distances. Our quantum PCA simulation allows us to accurately compress these molecular ground states into a subspace of small dimension.

II. BACKGROUND

Let us first give some background on covariance and PCA. We also discuss the covariance for complex random variables. This is particularly relevant to quantum systems, since quantum amplitudes are complex, in general.

A. Covariance

For two real random variables X and Y, their covariance is given by

$$\begin{aligned} \mathrm{Cov}(X,Y) &= E[[X - E(X)][Y - E(Y)]] \\ &= E(XY) - E(X)E(Y). \end{aligned} \tag{2}$$

The sign of the covariance indicates whether the variables correlate (positive sign) or anticorrelate (negative sign), and the magnitude of the covariance quantifies the degree of correlation.

Now suppose that X and Y are complex random variables. In this case, it is helpful to think of X and Y as vectors in the complex plane, with some associated randomness. Their covariance is given by

$$\begin{aligned} \mathrm{Cov}(X,Y) &= E[[X - E(X)]\overline{[Y - E(Y)]}] \\ &= E(X\overline{Y}) - E(X)E(\overline{Y}), \end{aligned} \tag{3}$$

where the overline indicates the complex conjugate. There is a geometric interpretation of this complex covariance, as follows. The covariance is a vector in the complex plane, with a direction and magnitude. The magnitude of this vector quantifies how correlated the two variables are, while the direction of this vector captures how out of phase the two variables are. When the covariance is positive, X and Y are in phase. When the covariance is negative, X and Y are completely out of phase. When the covariance is complex, X and Y are partially out of phase. For example, if the covariance is purely imaginary then the two variables are $90°$ out of phase. Figure 2 gives an illustration of this geometric interpretation for the covariance of complex random variables.

Hence, the covariance of complex random variables has a clear conceptual interpretation, and it generalizes the
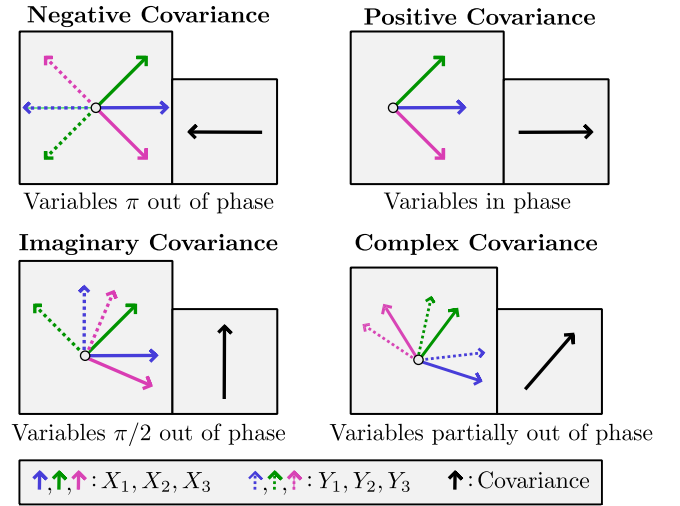

FIG. 2. Covariance for complex random variables. For two complex random variables X and Y, one can sample from their joint probability distribution $P(X, Y)$, and here we show three different samplings (with the purple, pink, and green colors). Each sampling leads to a pair of vectors in the complex plane (solid vector represents X, dashed vector represents Y). In the case of negative covariance, the vectors X and Y are $180°$ out of phase. For positive covariance, the vectors are in phase. Imaginary covariance corresponds to vectors that are $90°$ out of phase. In general, complex covariance arises from variables that are partially out of phase. In all cases, the direction of the covariance vector is indicated with a black arrow.

covariance for real random variables in perhaps the most natural way possible. This implies that applying the covariance to quantum amplitudes (which are complex numbers, in general) is conceptually meaningful. In turn, this implies that PCA on quantum states (i.e., vectors of quantum amplitudes) is conceptually meaningful.

If one has a set of random variables then one can look at all of the pairwise covariances, and this is essentially the covariance matrix. Suppose that $\mathbf{X} = \{X_1, \ldots, X_d\}$ is a d-dimensional vector of complex random variables. Then the (j, k)th entry of the covariance matrix Q is given by

$$\begin{aligned} Q_{jk} &= E[[X_j - E(X_k)]\overline{[X_k - E(X_k)]}] \\ &= E(X_j\overline{X_k}) - E(X_k)E(\overline{X_k}). \end{aligned} \tag{4}$$

B. Principal component analysis

In the context of PCA, one starts with a dataset $\{\mathbf{y}^{(i)}\}_{i=1}^N$ of N datapoints. Each datapoint $\mathbf{y}^{(i)}$ is vector over a d-dimensional feature space:

$$\mathbf{y}^{(i)} = \{y_1^{(i)}, \ldots, y_d^{(i)}\}^T. \tag{5}$$

In classical data analysis, the data typically consist of real numbers such that $y_j^{(i)} \in \mathbb{R}$. For quantum data, the $\mathbf{y}^{(i)}$ vectors would be quantum states and the $y_j^{(i)}$ would be

complex amplitudes. Hence, for generality, we can allow for $y_j^{(i)} \in \mathbb{C}$.

We can compute the empirical mean of the data as the vector

$$\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_d\}^T, \tag{6}$$

where $\mu_j = (1/N) \sum_{i=1}^{N} y_j^{(i)}$ is the empirical mean value for the jth feature when the distribution over datapoints is uniform. More generally, it is possible that each datapoint $\mathbf{y}^{(i)}$ could come with an associated probability $p^{(i)}$, and the probability distribution $\{p^{(i)}\}$ might not be uniform. In that case we have

$$\mu_j = \sum_{i=1}^{N} p^{(i)} y_j^{(i)}. \tag{7}$$

For notational convenience, we define the centered dataset $\{\mathbf{b}^{(i)}\}_{i=1}^{N}$ by subtracting off the mean value to each datapoint:

$$\mathbf{b}^{(i)} = \mathbf{y}^{(i)} - \boldsymbol{\mu}. \tag{8}$$

The covariance between the jth and kth features is then given by

$$
\begin{aligned}
Q_{jk} &= \sum_{i=1}^{N} p^{(i)} b_j^{(i)} \overline{b_k^{(i)}} \\
&= \sum_{i=1}^{N} p^{(i)} (y_j^{(i)} - \mu_j) \overline{(y_k^{(i)} - \mu_k)} \\
&= \left(\sum_{i=1}^{N} p^{(i)} y_j^{(i)} \overline{y_k^{(i)}} \right) - \mu_j \overline{\mu_k}.
\end{aligned}
\tag{9}
$$

In the special case of the uniform distribution, this is

$$Q_{jk} = \frac{1}{N} \sum_{i=1}^{N} b_j^{(i)} \overline{b_k^{(i)}}. \tag{10}$$

(In some literature, one uses the factor $1/(N-1)$ instead of $1/N$ due to Bessel's correction. This normalization factor does not play a significant role in our analysis, so we ignore this issue.)

The PCA method involves diagonalizing Q and keeping the largest eigenvalues. This diagonalization can be done with a unitary matrix, i.e., we can write the diagonal form of the covariance matrix as

$$\Lambda_Q = U_Q Q U_Q^\dagger \tag{11}$$

for some unitary matrix U_Q whose columns correspond to the principal components.

C. PCA without centering

Let us elaborate here on what is often called "PCA without centering." Instead of considering the covariance matrix, one can consider the matrix of (uncentered) second moments, denoted T, whose matrix elements are given by

$$T_{jk} = \sum_{i=1}^{N} p^{(i)} y_j^{(i)} \overline{y_k^{(i)}}. \tag{12}$$

Analogous to standard PCA, one can then diagonalize the T matrix with a unitary matrix U_T, to find the diagonal form

$$\Lambda_T = U_T T U_T^\dagger. \tag{13}$$

So-called "PCA without centering" then involves using the eigenvalues and eigenvectors of T in place of those of Q, keeping only the principal components of T.

As we will see below, our results indicate that "PCA without centering" is relevant to our proposed approach to quantum PCA. Hence, it is important for us to review what is known about this subroutine.

First of all, the name "PCA without centering" is questionable, and we believe an alternative name could be more appropriate. Diagonalizing an uncentered matrix of second moments is not necessarily PCA, and hence the term "PCA without centering" is not entirely accurate. Our results below give a different perspective. We argue that "PCA without centering" actually corresponds to performing PCA on a symmetrized version of the dataset. So it could be called PCA with data symmetrization. See Fig. 3 for an illustration of this, and also see Eq. (51) below for an example of a symmetrized dataset.

Regardless of the terminology, there remains the intriguing question of how PCA with data symmetrization is
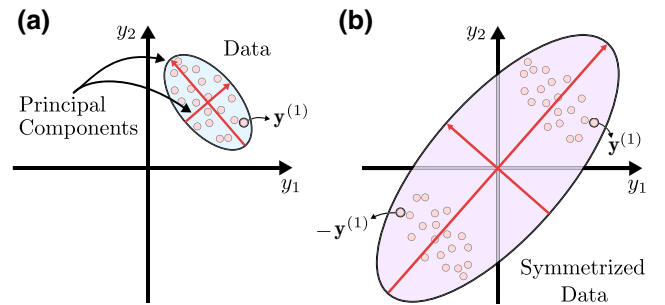
FIG. 3. Comparison of standard PCA and "PCA without centering." (a) Standard PCA involves diagonalizing Q, with the interpretation of fitting an ellipsoid to the data. (b) So-called "PCA without centering" can be interpreted as PCA on a symmetrized version of the dataset, e.g., where the dataset is duplicated with the duplicate copy having a minus sign applied. "PCA without centering" then corresponds to fitting an ellipsoid to the symmetrized dataset. [See Eq. (51) in the text below for an example of a symmetrized dataset.]

related to standard PCA. This question has been studied in detail by Cadima and Jolliffe [22]. Their comprehensive work on this topic is extremely useful for our purposes. Therefore, let us review some of their findings here.

The overall conclusion of Cadima and Jolliffe was that standard PCA and PCA with data symmetrization can be rigorously proven to be closely related. In particular, the eigenvalues and eigenvectors obtained with one method are often similar to those of the other method. For example, they showed that the eigenvalues are interlaced:

$$t_d \geqslant q_d \geqslant t_{d-1} \geqslant q_{d-1} \geqslant \cdots \geqslant t_1 \geqslant q_1 \qquad (14)$$

with $\{t_j\}_{j=1}^d$ and $\{q_j\}_{j=1}^d$ the eigenvalues of T and Q, respectively, listed in nondecreasing order. They also provided sufficient conditions for the sets of eigenvectors of T and Q to perfectly match. For example, one such sufficient condition is when either T or Q has an eigenvector that is identical to the normalized mean vector: $\boldsymbol{\mu}/\|\boldsymbol{\mu}\|$. It is often the case that the first eigenvector of T is almost colinear with $\boldsymbol{\mu}$, and consequently it is often observed that the eigenvectors of T and Q are quite similar. We refer the reader to Ref. [22] for additional theoretical results that establish a close connection between standard PCA and "PCA without centering."

Two minor differences between the setting considered by Cadima and Jolliffe and the setting we consider is that they restrict to real random variables ($y_j^{(i)} \in \mathbb{R}$) and uniform probability distributions over datapoints ($p^{(i)} = 1/N$ for all i). In our results (see Sec. IV), we allow for complex random variables ($y_j^{(i)} \in \mathbb{C}$) and nonuniform probability distributions ($p^{(i)}$ arbitrary). Hence we consider a slightly more general setting than Cadima and Jolliffe.

III. DEFINITIONS

Before stating our results let us first define some notation.

A. Datasets

Let us consider a dataset in the form of a set of normalized quantum states $\{|\psi^{(i)}\rangle\}$. For a classical dataset, we can imagine that the $y^{(i)}$ vectors (defined above) are encoded in the $|\psi^{(i)}\rangle$ states through an amplitude encoding procedure. There is a large body of literature on amplitude encoding [14–18], including near-term approaches [19–21], and hence we refer the reader to this literature. Hence, in what follows, we can restrict to datasets composed of quantum states.

We denote a dataset of statevectors as

$$D_{|\psi\rangle} = \{|\psi^{(i)}\rangle\}_{i=1}^N, \qquad (15)$$

and for this dataset, we denote the corresponding dataset of density matrices as

$$D_\rho = \{\rho^{(i)}\}_{i=1}^N = \{|\psi^{(i)}\rangle\langle\psi^{(i)}|\}_{i=1}^N. \qquad (16)$$

We note that, more generally, one could have a dataset of mixed states, which we can also denote as $D_\rho = \{\rho^{(i)}\}_{i=1}^N$. We give a full treatment of mixed-state datasets in Appendix C. There, we discuss how our results apply to an effective dataset that is constructed from the pure-state decompositions of each mixed state $\rho^{(i)}$. We refer the reader to that appendix for further discussion.

1. Classical and quantum datasets

Some of our theoretical results apply to classical datasets, while others apply to quantum datasets. Hence, let us define these terms here.

We use the term *classical dataset* to refer to a dataset that is stored on a classical computer or classical device. For classical datasets, the information about global phases for each datapoint is preserved. In other words, multiplicative factors applied to each datapoint have a nontrivial effect on the dataset. This is an important point that distinguishes classical datasets from quantum datasets.

We use the term *quantum dataset* to refer to a dataset that is stored on a quantum computer or quantum device. For quantum datasets, the information about global phases for each datapoint is lost or erased. In other words, multiplicative factors applied to each datapoint have a trivial (or nonphysical) effect on the dataset.

B. Ensembles

We also introduce the notion of ensembles. Ensembles of quantum states are commonly used in quantum information theory [23]. Ensembles include datapoints and their associated probabilities $P = \{p^{(i)}\}_{i=1}^N$. A statevector ensemble is denoted as

$$E_{|\psi\rangle} = \{p^{(i)}, |\psi^{(i)}\rangle\}_{i=1}^N, \qquad (17)$$

and the associated ensemble of density matrices is

$$E_\rho = \{p^{(i)}, \rho^{(i)}\}_{i=1}^N = \{p^{(i)}, |\psi^{(i)}\rangle\langle\psi^{(i)}|\}_{i=1}^N. \qquad (18)$$

C. Mapping statevectors to density matrices

For convenience, we denote the mapping that takes a statevector to a density matrix as

$$\mathcal{P}(|\psi\rangle) = |\psi\rangle\langle\psi|, \qquad (19)$$

and we call this the outer product mapping, since $|\psi\rangle\langle\psi|$ is the outer product.

With a slight abuse of notation, we can act with this map on a statevector dataset to get the corresponding density matrix dataset. Let $D_{|\psi\rangle} = \{|\psi^{(i)}\rangle\}_{i=1}^N$ be a statevector dataset. Then the corresponding density matrix dataset is

$$\mathcal{P}(D_{|\psi\rangle}) = \{\mathcal{P}(|\psi^{(i)}\rangle)\}_{i=1}^N = \{|\psi^{(i)}\rangle\langle\psi^{(i)}|\}_{i=1}^N. \quad (20)$$

Similarly, we can act with this map on a statevector ensemble to get the corresponding density matrix ensemble. Let $E_{|\psi\rangle} = \{p^{(i)}, |\psi^{(i)}\rangle\}_{i=1}^N$ be a statevector ensemble. Then the corresponding density matrix ensemble is

$$\mathcal{P}(E_{|\psi\rangle}) = \{p^{(i)}, \mathcal{P}(|\psi^{(i)}\rangle)\}_{i=1}^N = \{p^{(i)}, |\psi^{(i)}\rangle\langle\psi^{(i)}|\}_{i=1}^N. \quad (21)$$

We remark that $\mathcal{P}$ is not an invertible map, since global phase information is lost via the outer product.

Nevertheless, one can think about the set of statevectors that are consistent with a given density matrix. We borrow terminology from the open-quantum-system literature and refer to a particular choice of statevector (for a given density matrix) as an *unraveling* [24]. For example, we say that $e^{i\phi}|\psi\rangle$ is a particular unraveling of the density matrix $|\psi\rangle\langle\psi|$. We also use this same language for ensembles. That is, we say that $E_{|\psi\rangle}$ is an unraveling of E_ρ if it holds that $\mathcal{P}(E_{|\psi\rangle}) = E_\rho$.

D. Ensemble average density matrix

Imagine a simple protocol whereby one samples from the probability distribution $\{p^{(i)}\}$, and if outcome i occurs then one prepares the state $|\psi^{(i)}\rangle$ on a quantum device. This is the protocol previously depicted in Fig. 1. The result of this protocol is to effectively prepare the state

$$\overline{\rho} = \sum_{i=1}^N p^{(i)} |\psi^{(i)}\rangle\langle\psi^{(i)}|, \quad (22)$$

which we call the ensemble average density matrix for the ensemble $E_{|\psi\rangle} = \{p^{(i)}, |\psi^{(i)}\rangle\}_{i=1}^N$.

IV. THEORETICAL RESULTS

We now proceed to state our theoretical results. We emphasize that, while these results are not mathematically deep, they are conceptually nontrivial and technologically important.

We also note that the extension of our results to mixed-state datasets is given in Appendix C. In that appendix, we argue that our theoretical results apply to such datasets provided that we consider the pure states that decompose each mixed state as datapoints.

A. Results for classical datasets

Let us first consider the case of classical datasets defined in Sec. III.

1. Centered classical datasets

We begin by considering centered classical datasets. Such datasets have $\boldsymbol{\mu} = \mathbf{0}$, i.e., $\mu_j = 0$ for all j. The following proposition gives a simple equality in this case.

Proposition 1: *Consider a classical dataset of pure states* $D_{|\psi\rangle} = \{|\psi^{(i)}\rangle\}_{i=1}^N$. *With the ensemble denoted as* $E_{|\psi\rangle} = \{p^{(i)}, |\psi^{(i)}\rangle\}_{i=1}^N$, *the corresponding ensemble average density matrix in Eq. (22) is given by*

$$\overline{\rho} = Q \quad (23)$$

if the dataset is centered, i.e., if $\boldsymbol{\mu} = \mathbf{0}$.

Proof. The proof follows as a special case of Proposition 2 below (i.e., by setting $\boldsymbol{\mu} = \mathbf{0}$ in Proposition 2). ∎

Proposition 1 is already useful, as it identifies a method for preparing the covariance matrix for centered datasets, i.e., the method shown in Fig. 1.

2. Uncentered classical datasets

The following proposition gives the general relationship between $\overline{\rho}$ and Q, regardless of whether the data are centered or uncentered. We note that Eq. (24) below is mathematically related to (although conceptually different from) a result obtained in Ref. [22].

Proposition 2: *Consider a classical dataset of pure states* $D_{|\psi\rangle} = \{|\psi^{(i)}\rangle\}_{i=1}^N$. *With the ensemble denoted as* $E_{|\psi\rangle} = \{p^{(i)}, |\psi^{(i)}\rangle\}_{i=1}^N$, *the corresponding ensemble average density matrix in Eq. (22) is given by*

$$\overline{\rho} = Q + M, \quad (24)$$

where Q is the covariance matrix for the ensemble $E_{|\psi\rangle}$. Here, $M = \boldsymbol{\mu}\boldsymbol{\mu}^\dagger$ has matrix elements $M_{jk} = \mu_j \overline{\mu_k}$ with $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_d\}^T$ being the vector of mean values for the d features of the data.

Proof. Let us expand each state in terms of its amplitudes in the standard basis:

$$|\psi^{(i)}\rangle = \sum_{j=1}^d y_j^{(i)} |j\rangle. \quad (25)$$

Here we denote the amplitudes as $y_j^{(i)}$ because we assume that one has performed amplitude encoding of the data, and we recall that our notation above used $y_j^{(i)}$ to denote the jth feature value of the ith datapoint.

Then we have

$$|\psi^{(i)}\rangle\langle\psi^{(i)}| = \sum_{j,k=1}^{d} y_j^{(i)}\overline{y_k^{(i)}}|j\rangle\langle k|. \qquad (26)$$

Hence we can write

$$\overline{\rho} = \sum_{i=1}^{N}\sum_{j,k=1}^{d} p^{(i)}y_j^{(i)}\overline{y_k^{(i)}}|j\rangle\langle k|$$

$$= \sum_{j,k=1}^{d} \overline{\rho}_{jk}|j\rangle\langle k|. \qquad (27)$$

Here the matrix elements of $\overline{\rho}$ are given by

$$\overline{\rho}_{jk} = \sum_{i=1}^{N} p^{(i)}y_j^{(i)}\overline{y_k^{(i)}} = T_{jk}, \qquad (28)$$

where T_{jk} was defined in Eq. (12). Using Eq. (9), we can rewrite Eq. (28) in terms of the covariance matrix elements as

$$\overline{\rho}_{jk} = Q_{jk} + \mu_j\overline{\mu_k}, \qquad (29)$$

completing the proof. ∎

Equation (24) implies that $\overline{\rho}$ and Q are very closely related, even for uncentered datasets. The difference between $\overline{\rho}$ and Q is a positive semidefinite matrix M, which has several interesting properties, as stated in the following lemma.

Lemma 1: *For uncentered datasets, the $d \times d$ matrix $M = \boldsymbol{\mu}\boldsymbol{\mu}^\dagger$ has the following properties.*

(a) It is Hermitian and positive semidefinite.
(b) It is rank one.
(c) Its only nonzero eigenvalue is

$$\lambda_{\boldsymbol{\mu}} = \boldsymbol{\mu}^\dagger\boldsymbol{\mu} = \|\boldsymbol{\mu}\|^2 = \mu_1^*\mu_1 + \cdots + \mu_d^*\mu_d. \quad (30)$$

(d) The eigenvector associated with this nonzero eigenvalue is $v_{\boldsymbol{\mu}} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$.

Proof. The fact that it is Hermitian is obvious since $M^\dagger = M$. The fact that is rank one follows from the ability to write $M = \lambda_{\boldsymbol{\mu}}|\boldsymbol{\mu}\rangle\langle\boldsymbol{\mu}|$ as being proportional to a projector onto a pure quantum state $|\boldsymbol{\mu}\rangle$. This also implies that it is positive semidefinite, since it is proportional to a density matrix, with a positive proportionality constant. The fact the $v_{\boldsymbol{\mu}}$ is an eigenvector with eigenvalue $\lambda_{\boldsymbol{\mu}} = \boldsymbol{\mu}^\dagger\boldsymbol{\mu}$ can be directly verified:

$$M v_{\boldsymbol{\mu}} = \boldsymbol{\mu}\boldsymbol{\mu}^\dagger\boldsymbol{\mu}/\|\boldsymbol{\mu}\| = (\boldsymbol{\mu}^\dagger\boldsymbol{\mu})\boldsymbol{\mu}/\|\boldsymbol{\mu}\| = \lambda_{\boldsymbol{\mu}}v_{\boldsymbol{\mu}}. \quad (31)$$

This completes the proof. ∎

We are now in a position to relate the eigenvalues of $\overline{\rho}$ and Q. Namely, we formulate inequalities that bound the deviation of the eigenvalues of $\overline{\rho}$ from those of Q. The proof of the following result relies on Proposition 2, Lemma 1, and Weyl's theorem as stated in Appendix A.

Proposition 3: *Consider a classical dataset of pure states $D_{|\psi\rangle} = \{|\psi^{(i)}\rangle\}_{i=1}^{N}$, with the ensemble denoted as $E_{|\psi\rangle} = \{p^{(i)}, |\psi^{(i)}\rangle\}_{i=1}^{N}$. Let $\{r_j\}_{j=1}^{d}$ and $\{q_j\}_{j=1}^{d}$ respectively be the eigenvalues of $\overline{\rho}$ and Q listed in nondecreasing order. Then the eigenvalues of $\overline{\rho}$ are interlaced with the eigenvalues of Q as*

$$r_d \geqslant q_d \geqslant r_{d-1} \geqslant q_{d-1} \geqslant \cdots \geqslant r_1 \geqslant q_1. \quad (32)$$

In addition, for each j, the bound

$$q_j \leqslant r_j \leqslant q_j + \|\boldsymbol{\mu}\|^2 \qquad (33)$$

holds, where $\boldsymbol{\mu}$ is the vector of mean values.

Proof. Recall from Proposition 2 that we have $\overline{\rho} = Q + M$. Using this result, we can then apply Weyl's theorem for the eigenvalues of Hermitian matrices, noting that $\overline{\rho}$, Q, and M are all Hermitian. Appendix A gives the general statement of Weyl's theorem, which relates the eigenvalues of two Hermitian matrices A and B to those of $A + B$. We refer the reader to Ref. [25] for additional details on Weyl's theorem. In Appendix A, we specialize Weyl's theorem to the case where A is an arbitrary Hermitian matrix and B is a rank-one Hermitian matrix whose only nonzero eigenvalue is $\lambda_B > 0$, to obtain

$$\lambda_j(A + B) \leqslant \lambda_j(A) + \lambda_B \quad \text{for } j = 1, \ldots, d, \qquad (34)$$

$$\lambda_j(A + B) \leqslant \lambda_{j+1}(A) \quad \text{for } j = 1, \ldots, d-1, \qquad (35)$$

$$\lambda_j(A) \leqslant \lambda_j(A + B) \quad \text{for } j = 1, \ldots, d. \qquad (36)$$

In this set of inequalities, the respective eigenvalues of A and $A + B$ are denoted $\{\lambda_j(A)\}_{j=1}^{d}$ and $\{\lambda_j(A + B)\}_{j=1}^{d}$, and these eigenvalues are listed in nondecreasing order. Let us apply the above inequalities by choosing $A = Q$, $B = M$, and $A + B = \overline{\rho}$. In this case, $\lambda_j(A + B)$ becomes r_j, and $\lambda_j(A)$ becomes q_j. Moreover, by invoking Lemma 1, the nonzero eigenvalue of B, λ_B, corresponds to $\|\boldsymbol{\mu}\|^2$. Hence, the above inequalities become

$$r_j \leqslant q_j + \|\boldsymbol{\mu}\|^2 \quad \text{for } j = 1, \ldots, d, \qquad (37)$$

$$r_j \leqslant q_{j+1} \quad \text{for } j = 1, \ldots, d-1, \qquad (38)$$

$$q_j \leqslant r_j \quad \text{for } j = 1, \ldots, d. \qquad (39)$$

Combining Eqs. (38) and (39) gives the result in Eq. (32). Also, combining Eqs. (37) and (39) gives the result in Eq. (33). ∎

Proposition 3 implies that the eigenvalues of $\overline{\rho}$ and Q can never deviate too much from each other. For example, for the jth eigenvalue, the deviation $\delta_j := r_j - q_j$ is upper bounded as

$$\delta_j \leqslant \min\{\|\boldsymbol{\mu}\|^2, (q_{j+1} - q_j)\}. \tag{40}$$

In this sense, the eigenvalues of $\overline{\rho}$ form a good approximation for the eigenvalues of Q.

We remark that Proposition 3 is more general than the main result in Ref. [22]. Specifically, Proposition 3 holds for arbitrary probability distributions $\{p^{(i)}\}$ and for complex random variables $y_j^{(i)}$, whereas Cadima and Jolliffe [22] restricted to uniform probability distributions and real random variables.

We also note that the proof technique used to prove the main result in Ref. [22] is only valid for uniform probability distributions. Therefore, to prove Proposition 3, we could not simply use the proof technique in Ref. [22]. Rather, we required a novel proof technique, and this involved using Weyl's theorem.

Let us now relate the eigenvectors of $\overline{\rho}$ and Q. In general, for a matrix A and a state $|\psi\rangle$, we can quantify how far $|\psi\rangle$ is from being an eigenvector of A, as follows. Let us define the unnormalized state

$$|\delta_{A,|\psi\rangle}\rangle = A|\psi\rangle - \langle\psi|A|\psi\rangle|\psi\rangle. \tag{41}$$

Note that $|\delta_{A,|\psi\rangle}\rangle$ is the zero vector whenever $|\psi\rangle$ is an eigenvector of A. The norm of the $|\delta_{A,|\psi\rangle}\rangle$ vector quantifies how far $|\psi\rangle$ is from being an eigenvector of A. Therefore we define the eigenvector error as

$$e_{A,|\psi\rangle} = \langle\delta_{A,|\psi\rangle}|\delta_{A,|\psi\rangle}\rangle. \tag{42}$$

Consider the following proposition that quantifies the eigenvector error.

Proposition 4: *Consider a classical dataset of pure states $D_{|\psi\rangle} = \{|\psi^{(i)}\rangle\}_{i=1}^N$, with the ensemble denoted as $E_{|\psi\rangle} = \{p^{(i)}, |\psi^{(i)}\rangle\}_{i=1}^N$. Let $|Q_j\rangle$ be an eigenvector of the covariance matrix Q. For this state, the eigenvector error for matrix $\overline{\rho}$ is*

$$e_{\overline{\rho},|Q_j\rangle} = \|\boldsymbol{\mu}\|^4|\langle v_{\boldsymbol{\mu}}|Q_j\rangle|^2(1 - |\langle v_{\boldsymbol{\mu}}|Q_j\rangle|^2). \tag{43}$$

Similarly, let $|R_j\rangle$ be an eigenvector of the ensemble average density matrix $\overline{\rho}$. For this state, the eigenvector error for matrix Q is given by the same expression:

$$e_{Q,|R_j\rangle} = \|\boldsymbol{\mu}\|^4|\langle v_{\boldsymbol{\mu}}|R_j\rangle|^2(1 - |\langle v_{\boldsymbol{\mu}}|R_j\rangle|^2). \tag{44}$$

Proof. This follows from a direct calculation. One can write

$$\begin{aligned}|\delta_{\overline{\rho},Q_j}\rangle &= \overline{\rho}|Q_j\rangle - \langle Q_j|\overline{\rho}||Q_j\rangle \\ &= \overline{\rho}|Q_j\rangle - |Q_j\rangle\langle Q_j|\overline{\rho}| \\ &= \Pi^{\perp}\overline{\rho}|Q_j\rangle, \end{aligned} \tag{45}$$

where $\Pi^{\perp} = \mathbb{1} - |Q_j\rangle\langle Q_j|$ is the projector onto the orthogonal complement of $|Q_j\rangle$. Next we use $\overline{\rho} = Q + M$ to write

$$\begin{aligned}|\delta_{\overline{\rho},Q_j}\rangle &= \Pi^{\perp}(Q + M)|Q_j\rangle \\ &= \Pi^{\perp}M|Q_j\rangle, \end{aligned} \tag{46}$$

which follows from the fact that $\Pi^{\perp}Q|Q_j\rangle = 0$, since $|Q_j\rangle$ is an eigenvector of Q. The eigenvector error is then

$$\begin{aligned}e_{\overline{\rho},|Q_j\rangle} &= \langle Q_j|M\Pi^{\perp}M|Q_j\rangle \\ &= \langle Q_j|M^2|Q_j\rangle - \langle Q_j|M|Q_j\rangle\langle Q_j|M|Q_j\rangle. \end{aligned} \tag{47}$$

Combining this final expression with $M = \|\boldsymbol{\mu}\|^2|v_{\boldsymbol{\mu}}\rangle\langle v_{\boldsymbol{\mu}}|$ then gives the desired result in Eq. (43). The above proof can be rewritten analogously to derive Eq. (44). ∎

From this proposition, we see that an eigenvector for $\overline{\rho}$ (Q) is an eigenvector of Q ($\overline{\rho}$) if and only if at least one of the following conditions is satisfied.

(a) The eigenvector is orthogonal to the mean vector.
(b) The eigenvector is colinear with the mean vector.
(c) The mean vector is the zero vector (i.e, the dataset is centered).

Hence, the only way for an eigenvector to not be shared between $\overline{\rho}$ and Q is if the dataset is uncentered and the eigenvector has partial (but not complete) overlap with the mean vector. Because of Proposition 4, one can see that it is often the case that eigenvectors are approximately shared between the two matrices, $\overline{\rho}$ and Q.

In addition, Eq. (44) can be used as a diagnostic tool for "PCA without centering." Specifically, this equation can be used to verify the quality of an eigenvector obtained from diagonalizing $\overline{\rho}$. One can simply calculate the overlap of that eigenvector with the mean vector and then use Eq. (44) to quantify the eigenvalue error for the true covariance matrix Q.

Now let us consider both the eigenvalues and eigenvectors. The following proposition gives a sufficient condition for both the eigenvalues and eigenvectors of $\overline{\rho}$ and Q to perfectly match. As we will see in our numerical implementations, this sufficient condition is often satisfied. Namely, it is often the case that the first principal component of $\overline{\rho}$ (i.e., the eigenvector associated with the

largest eigenvalue) is very close to being the normalized mean vector. Hence, the following proposition is relevant to practical scenarios of interest.

Proposition 5: *Consider a classical dataset of pure states $D_{|\psi\rangle} = \{|\psi^{(i)}\rangle\}_{i=1}^N$, with the ensemble denoted as $E_{|\psi\rangle} = \{p^{(i)}, |\psi^{(i)}\rangle\}_{i=1}^N$. If one of the eigenvectors of $\overline{\rho}$ or Q is the normalized mean vector $\boldsymbol{\mu}/\|\boldsymbol{\mu}\|$ then the following conditions are satisfied.*

    (a) Matrices $\overline{\rho}$ and Q have a common set of eigenvectors. Any set of eigenvectors for $\overline{\rho}$ is also a valid set of eigenvectors for Q, and vice versa.

    (b) The $d-1$ eigenvalues that are not associated with $\boldsymbol{\mu}/\|\boldsymbol{\mu}\|$ are shared. Let $r_{\hat{j}}$ and $q_{\hat{j}}$ be the eigenvalues associated with $\boldsymbol{\mu}/\|\boldsymbol{\mu}\|$; then $r_{\hat{j}} - q_{\hat{j}} = \|\boldsymbol{\mu}\|^2$.

Proof. Denote the normalized mean vector as the quantum state $|v_{\boldsymbol{\mu}}\rangle = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$. Consider a spectral decomposition of Q given by

$$Q = \sum_{j \neq \hat{j}} q_j |Q_j\rangle\langle Q_j| + q_{\hat{j}} |v_{\boldsymbol{\mu}}\rangle\langle v_{\boldsymbol{\mu}}|, \tag{48}$$

where the $|Q_j\rangle$ states are orthogonal eigenvectors (and also orthogonal to $|v_{\boldsymbol{\mu}}\rangle$). Then, we invoke the equation $\overline{\rho} = Q + M$ to obtain

$$\begin{aligned}\overline{\rho} &= Q + \|\boldsymbol{\mu}\|^2 |v_{\boldsymbol{\mu}}\rangle\langle v_{\boldsymbol{\mu}}| \\ &= \sum_{j \neq \hat{j}} q_j |Q_j\rangle\langle Q_j| + (q_{\hat{j}} + \|\boldsymbol{\mu}\|^2)|v_{\boldsymbol{\mu}}\rangle\langle v_{\boldsymbol{\mu}}|. \end{aligned} \tag{49}$$

Note that this is also a spectral decomposition of $\overline{\rho}$. Hence, we see that $\overline{\rho}$ inherits the same eigenvectors as those of Q, and all of the eigenvalues are also the same except for the $\hat{j}$th eigenvalue, which is shifted by $\|\boldsymbol{\mu}\|^2$. Finally, note that one can apply the exact same argument in the reverse direction, where one first starts with a spectral decomposition of $\overline{\rho}$, and then one derives the corresponding spectral decomposition of Q. ∎

Thus far, we have related the eigenvalues and eigenvectors of $\overline{\rho}$ and Q. We remark that one can also relate the diagonal elements of $\overline{\rho}$ and Q, and we formally state this in Appendix B.

B. Results for quantum datasets

Classical datapoints are affected by application of a global phase, such as multiplying by a minus sign as in Fig. 4. One can see in this case that the minus sign inverts the color of the image, from a white digit to a black digit.

However, quantum states are invariant under application of a global phase to the state vector. This can be seen from
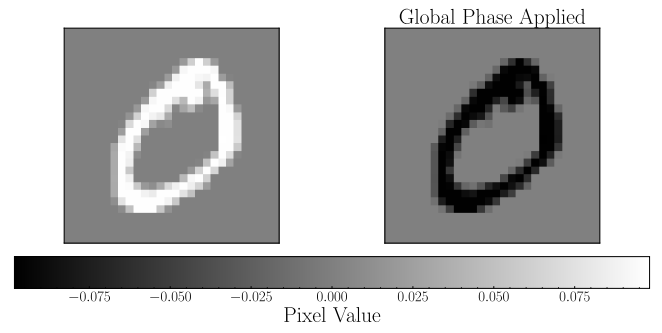

FIG. 4. Impact of global phase on classical data. Unlike quantum states, classical datapoints are impacted by applying a global phase. Here we give a visual demonstration of this, where multiplying by a minus sign inverts the color of a handwritten digit from white to black.

the fact that the density matrix is unaffected by a global phase applied to the state vector:

$$\begin{aligned}|\psi\rangle &\to e^{i\phi}|\psi\rangle, \\ |\psi\rangle\langle\psi| &\to e^{i\phi}e^{-i\phi}|\psi\rangle\langle\psi| = |\psi\rangle\langle\psi|.\end{aligned} \tag{50}$$

This global phase symmetry implies that global phases have no physical effect or physical manifestation. We now discuss how global phase symmetry can allow us to assume that all datasets that are prepared on quantum devices admit a description that is centered. More precisely, there are multiple statevector descriptions for a given dataset of density matrices, and there always exist some statevector descriptions that are centered.

We first state the following lemma.

Lemma 2: *Consider a quantum dataset of pure states. In this case, the global phase information is lost. Hence, the dataset can be described by an ensemble of density matrices $E_\rho = \{p^{(i)}, |\psi^{(i)}\rangle\langle\psi^{(i)}|\}_{i=1}^N$. Then, there always exists a statevector ensemble $E_{|\tilde{\psi}\rangle} = \{\tilde{p}^{(j)}, |\tilde{\psi}^{(j)}\rangle\}_{j=1}^{\tilde{N}}$ that satisfies the following conditions.*

    (a) Ensemble $E_{|\tilde{\psi}\rangle}$ physically corresponds to the aforementioned E_ρ, in the sense that applying the outer product mapping leads to $\mathcal{P}(E_{|\tilde{\psi}\rangle}) = E_\rho$. In other words, $E_{|\tilde{\psi}\rangle}$ is an unraveling of E_ρ.

    (b) Ensemble $E_{|\tilde{\psi}\rangle}$ is centered, i.e., the mean value of all features is zero.

Proof. We prove this lemma by construction, i.e., by constructing an $E_{|\tilde{\psi}\rangle}$ that satisfies the required conditions.

Let $D_\rho = \{|\psi^{(i)}\rangle\langle\psi^{(i)}|\}_{i=1}^N$, and let $D_{|\psi\rangle} = \{|\psi^{(i)}\rangle\}_{i=1}^N$ be a particular unraveling of D_ρ.

Now let us consider a symmetrized dataset composed of $\tilde{N} = 2N$ datapoints that involves appending the set

$(-D_{|\psi\rangle})$ onto the end of the set $D_{|\psi\rangle}$. We write the overall dataset as

$$D_{|\tilde{\psi}\rangle} = \{|\psi^{(1)}\rangle, \ldots, |\psi^{(N)}\rangle, -|\psi^{(1)}\rangle, \ldots, -|\psi^{(N)}\rangle\}. \quad (51)$$

For the probabilities associated with this dataset, we use

$$\tilde{P} = \tfrac{1}{2}\{p^{(1)}, \ldots, p^{(N)}, p^{(1)}, \ldots, p^{(N)}\}. \quad (52)$$

Using the previous two equations, we can write the overall ensemble as

$$E_{|\tilde{\psi}\rangle} = \left\{ \left(\frac{p^{(1)}}{2}, |\psi^{(1)}\rangle\right), \ldots, \left(\frac{p^{(N)}}{2}, |\psi^{(N)}\rangle\right), \right.$$
$$\left. \left(\frac{p^{(1)}}{2}, -|\psi^{(1)}\rangle\right), \ldots, \left(\frac{p^{(N)}}{2}, -|\psi^{(N)}\rangle\right) \right\}. \quad (53)$$

Let us note that this ensemble, $E_{|\tilde{\psi}\rangle}$, is centered. One can see this by computing the mean values for each feature as

$$\mu_j = \sum_{i=1}^{N} \frac{p^{(i)}}{2} y_j^{(i)} + \sum_{i=1}^{N} \frac{p^{(i)}}{2} (-y_j^{(i)}) = 0. \quad (54)$$

Here we used the fact that the feature values for $-|\psi^{(i)}\rangle$ are the negatives of the feature values for $|\psi^{(i)}\rangle$. Hence we have shown that $E_{|\tilde{\psi}\rangle}$ is centered.

Now we just need to show that $E_{|\tilde{\psi}\rangle}$ is an unraveling of E_ρ. In other words we need to show that $\mathcal{P}(E_{|\tilde{\psi}\rangle}) = E_\rho$, where $\mathcal{P}$ is the outer product mapping. Applying $\mathcal{P}$ gives

$$\mathcal{P}(E_{|\tilde{\psi}\rangle})$$
$$= \left\{ \left(\frac{p^{(1)}}{2}, |\psi^{(1)}\rangle\langle\psi^{(1)}|\right), \ldots, \left(\frac{p^{(N)}}{2}, |\psi^{(N)}\rangle\langle\psi^{(N)}|\right), \right.$$
$$\left. \left(\frac{p^{(1)}}{2}, |\psi^{(1)}\rangle\langle\psi^{(1)}|\right), \ldots, \left(\frac{p^{(N)}}{2}, |\psi^{(N)}\rangle\langle\psi^{(N)}|\right) \right\}. \quad (55)$$

Note that there are N datapoints that appear twice in this ensemble with the same probability. These redundant datapoints can be aggregated, with their probabilities summed. After we aggregate these datapoints, we see that the ensemble $\mathcal{P}(E_{|\tilde{\psi}\rangle})$ is equivalent to the ensemble $E_\rho = \{p^{(i)}, |\psi^{(i)}\rangle\langle\psi^{(i)}|\}_{i=1}^{N}$. This proves the desired result. ∎

In what follows, it will help if we use more explicit notation. We use $\overline{\rho}(E_\rho)$, instead of $\overline{\rho}$, to indicate that the ensemble average density matrix is a function of the ensemble E_ρ in Eq. (18). We also use $Q(E_{|\psi\rangle})$, instead of Q, to indicate that the covariance matrix is a function of the statevector ensemble $E_{|\psi\rangle}$ in Eq. (17).

With the previous lemma in hand, we can state the following proposition for quantum datasets.

Proposition 6: *Consider a quantum dataset of pure states. In this case, the global phase information is lost. Hence, the dataset can be described by an ensemble of density matrices $E_\rho = \{p^{(i)}, |\psi^{(i)}\rangle\langle\psi^{(i)}|\}_{i=1}^{N}$. Then, there always exists a statevector ensemble $E_{|\tilde{\psi}\rangle} = \{\tilde{p}^{(j)}, |\tilde{\psi}^{(j)}\rangle\}_{j=1}^{\tilde{N}}$ that satisfies the following conditions.*

(a) *Ensemble $E_{|\tilde{\psi}\rangle}$ physically corresponds to the aforementioned E_ρ, in the sense that applying the outer product mapping leads to $\mathcal{P}(E_{|\tilde{\psi}\rangle}) = E_\rho$. In other words, $E_{|\tilde{\psi}\rangle}$ is an unraveling of E_ρ.*

(b) *The covariance matrix $Q(E_{|\tilde{\psi}\rangle})$ for $E_{|\tilde{\psi}\rangle}$ is equal to the ensemble average density matrix for E_ρ:*

$$Q(E_{|\tilde{\psi}\rangle}) = \overline{\rho}(E_\rho). \quad (56)$$

Proof. We prove this by constructing an $E_{|\tilde{\psi}\rangle}$ that satisfies the required conditions. Specifically, we use the same $E_{|\tilde{\psi}\rangle}$ used to prove the previous lemma, with $E_{|\tilde{\psi}\rangle}$ given in Eq. (53).

For this choice of $E_{|\tilde{\psi}\rangle}$, we already showed that it satisfies the first condition, i.e., that it is an unraveling of E_ρ [see Eq. (55)].

Hence we just need to show the second condition. The second condition essentially follows by combining Proposition 1 and Lemma 2. In more detail, in Lemma 2 we showed that $E_{|\tilde{\psi}\rangle}$ is centered; see Eq. (54). Hence, $\boldsymbol{\mu} = \mathbf{0}$ for $E_{|\tilde{\psi}\rangle}$. Next we apply Proposition 1 to see that

$$Q(E_{|\tilde{\psi}\rangle}) = \overline{\rho}(E_{|\tilde{\psi}\rangle}), \quad (57)$$

where $\overline{\rho}(E_{|\tilde{\psi}\rangle})$ is the ensemble average density matrix for $\overline{\rho}(E_{|\tilde{\psi}\rangle})$. One can also see this by applying Eq. (24) and setting $\boldsymbol{\mu} = \mathbf{0}$. Finally, one can note that the ensemble average density matrix is the same for $E_{|\tilde{\psi}\rangle}$ and E_ρ,

$$\overline{\rho}(E_{|\tilde{\psi}\rangle}) = \overline{\rho}(E_\rho). \quad (58)$$

Combining the two previous equations gives the desired result. ∎

C. Implications of theoretical results

The implications of our theoretical results are as follows. If you have any dataset of pure states then their ensemble average density matrix $\overline{\rho}$ can always be interpreted as a covariance matrix for an appropriately symmetrized (or centered) version of the dataset.

If the original dataset of pure states lives on a quantum device then their global phase information has already been erased, and hence one can always interpret this dataset as being centered. This means that the action of extracting the principal eigenvectors of $\overline{\rho}$, for any pure-state dataset living on a quantum device, can always be interpreted as performing PCA on this dataset. This result is extremely useful in light of the fact that quantum PCA

for quantum datasets could lead to exponential quantum speedup [12,26]. We therefore believe that Proposition 6 will have technological importance in the quest for quantum advantage, as it provides a simple means for preparing the covariance matrix for quantum datasets.

With that said, for classical datasets (living on classical devices), the global phase information is retained. The action of encoding this dataset into quantum states and preparing $\overline{\rho}$ on a quantum device will necessarily destroy the global phase information. In this case, extracting the principal eigenvectors of $\overline{\rho}$ will only correspond to PCA for a slightly different dataset, i.e., a symmetrized version of the original dataset. This idea was depicted in Fig. 3.

Nevertheless, Propositions 3–5 show that the spectrum obtained from $\overline{\rho}$ can be quite similar to that obtained from standard PCA. Proposition 4 is a conceptually novel formulation, while Propositions 3 and 5 generalize the results in Ref. [22] to the cases of complex random variables and nonuniform probability distributions over datapoints. (This generalization was nontrivial and required a different proof technique than that used in Ref. [22].) In particular, we find that the spectrum of $\overline{\rho}$ is essentially identical to that of the covariance matrix whenever one eigenvector is colinear with the mean vector $\boldsymbol{\mu}$. This condition is often satisfied by the first principal eigenvector of $\overline{\rho}$ [22]. One can see this geometrically in Fig. 3(b), where the principal axis of the ellipse is approximately colinear with the mean vector.

We further investigate these issues, for both classical and quantum datasets, in our numerical implementations in Sec. VI.

V. ANALYSIS OF THE SAMPLING OVERHEAD

Here we detail the sampling cost incurred by our approach when being used as a subroutine for two near-term algorithms that can be used to diagonalize a density matrix. Namely, we explore the application of our approach within the variational quantum state diagonalization (VQSD) algorithm [4] and the variational quantum state eigensolver (VQSE) algorithm [5] for performing quantum PCA. In addition, we briefly remark on the relevance of our approach to the original quantum PCA algorithm [3] in Sec. V C.

In both VQSD and VQSE a cost function is minimized in order to find a circuit that diagonalizes a quantum state ρ. This circuit can then be used to return the approximate eigenvalues and eigenvectors of the state. Naturally, in our analysis, we assume that the state to be diagonalized is the ensemble average density matrix, i.e., $\rho = \overline{\rho}$. After all, we connect $\overline{\rho}$ to the covariance matrix in our work here, and PCA involves diagonalizing the covariance matrix.

In VQSD, the cost function takes the form

$$\mathcal{C}_{\mathrm{VQSD}}(U(\boldsymbol{\alpha}), \rho) = \mathrm{Tr}(\rho^2) - \mathrm{Tr}(\mathcal{Z}[(U(\boldsymbol{\alpha})\rho U(\boldsymbol{\alpha})^{\dagger})^2]),$$
(59)

where $\mathcal{Z}(\rho)$ is a quantum channel that dephases ρ in the standard basis. The cost function vanishes under the condition $\tilde{\rho} = \mathcal{Z}(\tilde{\rho})$, meaning that $\tilde{\rho}$ is diagonal in the standard basis. Two copies of state ρ are required to compute the terms in the VQSD cost. The first term can be computed using the destructive SWAP test, whereas the second term can be evaluated using the diagonalized inner product (DIP) test.

For VQSE, the cost function has the form

$$\mathcal{C}_{\mathrm{VQSE}}(V(\boldsymbol{\theta}), \rho) = \mathrm{Tr}(V(\boldsymbol{\theta})\rho V^{\dagger}(\boldsymbol{\theta})H),$$
(60)

where H is a Hamiltonian that is nondegenerate over its m-lowest-energy levels, assuming that one wishes to extract the m-largest eigenvalues of ρ. While this leaves much freedom in choosing H, one possible form is

$$H = \mathbb{1} - \sum_{i=1}^{m} q_i |\mathbf{e}_i\rangle\langle\mathbf{e}_i|$$
(61)

with $q_i > 0$ (such that $q_i > q_{i+1}$) and the $|\mathbf{e}_i\rangle$ orthogonal states in the standard basis. In this case only one copy of the state ρ is needed to evaluate the cost function.

We first consider the case where every state in the dataset of interest is prepared deterministically and the statistics of the probability distribution are reproduced in a classical postprocessing step. Then we consider the case where we prepare each state by sampling from the distribution of the dataset.

Although the cost functions noted above are global and hence can have trainability issues [27], we consider them here due to their simplicity, and we note that the analysis we present can easily be extended to local versions of these cost functions.

A. Deterministic state preparation

Let us consider a deterministic state preparation as follows. One can imagine rewriting the cost functions in Eqs. (59) and (60) in terms of the states in the ensemble. In others words, we expand the density matrix as in Eq. (22) as $\rho = \overline{\rho} = \sum_{i=1}^{N} p^{(i)} |\psi^{(i)}\rangle\langle\psi^{(i)}|$, and we insert this expression into the cost functions. Now the cost functions are written entirely in terms of the states $|\psi^{(i)}\rangle$ that compose the ensemble. Hence one can estimate these cost functions using state preparation circuits for the individual $|\psi^{(i)}\rangle$ states.

When considering VQSE, the situation is very simple, as only one copy of the state ρ is necessary to evaluate the cost function. Therefore, one only needs to compute the output from $\mathcal{O}(N)$ circuits at each cost function evaluation. Hence, the number of state preparation circuits is linear in N, which is a relatively minor overhead.

For the case of VQSD, the cost function is quadratic in ρ. Hence, the circuits required to evaluate each term in the

cost function require two copies of ρ. When expanding the cost function in terms of the states that decompose ρ [as in Eq. (22)], there are N^2 terms in the expansion. Hence, each of the two terms of the VQSD cost in Eq. (59) requires N^2 state preparation circuits, as the first term is evaluated with the destructive SWAP test and the second term is evaluated with the DIP test. Therefore, the overall VQSD cost function can be evaluated using $\mathcal{O}(N^2)$ state preparation circuits.

In summary, only a small amount of overhead in N is required to integrate our method into the VQSE and VQSD algorithms. Namely, we require linear overhead for VQSE and quadratic overhead for VQSD.

B. Sampling from the dataset

While the overhead with the deterministic approach is small, this overhead can be reduced even further via random sampling with M samples. Namely, the scaling with the number of states per cost function evaluation can be improved by sampling from the distribution of states making up the dataset, rather than preparing each state deterministically and classically combining the results.

The value of the cost function is a scalar quantity. Therefore, we can explore the number of samples necessary to obtain a good estimate of the cost function using Hoeffding's inequality. We first explore how this applies in the case of VQSD. In this case the cost function consists of two terms that both have a nonlinear dependence on ρ. However, one can bring each term together and consider them to be one observable evaluation. We show that using Hoeffding's inequality allows one to bound the deviation of the observable from its true value with the number of samples and therefore control the overall error in the cost function estimation.

Proposition 7: *Suppose that we estimate $C_{\text{VQSD}}(U(\boldsymbol{\alpha}), \rho)$ for some unitary $U(\boldsymbol{\alpha})$ by randomly sampling from the dataset $\{|\psi^{(i)}\rangle\}$ that decomposes ρ. To ensure that the deviation from the true cost function value $C_{\text{VQSD}}(U(\boldsymbol{\alpha}), \rho)$ is smaller than ε with probability $(1 - \delta)$, it suffices for the number of samples to be*

$$M = \frac{9 \log(2/\delta)}{2\varepsilon^2}. \tag{62}$$

Proof. As previously stated, the two terms in the VQSD cost function can be exactly evaluated using a destructive SWAP test and a so-called DIP test for each term [4]. To perform the SWAP test, one measures the expectation value of the SWAP operator. For the DIP test, one evolves the state by a controlled-NOT ladder and measures the projector onto the all zero state on the first system. Therefore, the cost

function can be written as

$$
\begin{aligned}
C_{\text{VQSD}}(U(\boldsymbol{\alpha}), \rho) = &\, \text{Tr}((\rho \otimes \rho)\text{SWAP}) \\
&- \text{Tr}[(\rho \otimes \rho)U^{\dagger}(\boldsymbol{\alpha})U_D^{\dagger}(|0\rangle\langle 0| \otimes \mathbb{1}) \\
&\times U_D U(\boldsymbol{\alpha})],
\end{aligned} \tag{63}
$$

where we have used the cyclic property of the trace. We can bring these two terms together and write the cost function as

$$C_{\text{VQSD}}(U(\boldsymbol{\alpha}), \rho) = \text{Tr}((\rho \otimes \rho)H), \tag{64}$$

where $H = \text{SWAP} - U(\boldsymbol{\alpha})U_D^{\dagger}(|0\rangle\langle 0| \otimes \mathbb{1})U_D U(\boldsymbol{\alpha})$. Using $\rho = \sum_i^N p^{(i)}|\psi^{(i)}\rangle\langle\psi^{(i)}|$ then leads to the expression

$$
\begin{aligned}
&C_{\text{VQSD}}(U(\boldsymbol{\alpha}), \rho) \\
&= \sum_{i,j=1}^N p^{(i)}p^{(j)}\text{Tr}((|\psi^{(i)}\rangle\langle\psi^{(i)}| \otimes |\psi^{(j)}\rangle\langle\psi^{(j)}|)H).
\end{aligned} \tag{65}
$$

Relabeling indices from pairs of i, j to k and rewriting the above equation gives

$$C_{\text{VQSD}}(U(\boldsymbol{\alpha}), \rho) = \sum_{k=1}^{N^2} p_k \text{Tr}(|\tilde{\psi}^{(k)}\rangle\langle\tilde{\psi}^{(k)}|H). \tag{66}$$

Therefore, we can consider sampling from the set of states $\tilde{\rho} = \sum_{k=1}^{N^2} p^{(k)}|\tilde{\psi}^{(k)}\rangle\langle\tilde{\psi}^{(k)}|$ and estimating the value of the cost function over the sampled states. First, consider the quantity

$$S_M = \frac{1}{M}\sum_{m=1}^M X^{(m)}, \tag{67}$$

where

$$X^{(m)} = \text{Tr}(|\tilde{\psi}^{(m)}\rangle\langle\tilde{\psi}^{(m)}|H). \tag{68}$$

Let us note that every term $X^{(m)} \in [-2, 1]$. This follows from the fact that the spectrum of SWAP is in $[-1, 1]$ while the spectrum of $|0\rangle\langle 0| \otimes \mathbb{1}$ is in $[0, 1]$, and hence the overall spectrum of H is in $[-2, 1]$. Furthermore, noting that $\mathbb{E}(S_M) = C_{\text{VQSD}}(U(\boldsymbol{\alpha}), \rho)$, we can apply Hoeffding's inequality,

$$\mathcal{P}(|S_M - C_{\text{VQSD}}(U(\boldsymbol{\alpha}), \rho)| \geqslant \varepsilon) \leqslant 2 \exp\left(\frac{-2\varepsilon^2 M}{9}\right). \tag{69}$$

The probability of the estimate S_M deviating from the true value $C_{\text{VQSD}}(U(\boldsymbol{\alpha}), \rho)$ by more than some value ε

decreases exponentially with the number of samples M. Defining this probability as δ leads to an estimate of the lower bound on the number of samples to bound ε with confidence $1 - \delta$:

$$M = \frac{9 \log(2/\delta)}{2\varepsilon^2}. \tag{70}$$

This completes the proof. ∎

A similar approach can be taken when considering VQSE. In that case the cost function only consists of the evaluation of one observable and one copy of the state ρ. This makes the analysis simpler, but the conclusion is the same. Furthermore, $|C_{\text{VQSE}}| \leqslant \|H\|_\infty$ with H given by Eq. (61), and this Hamiltonian norm ends up appearing in the exponent in Hoeffding's inequality. This leads to a similar proposition for the case of VQSE.

Proposition 8: *Suppose that we estimate $C_{\text{VQSE}}(U(\boldsymbol{\alpha}), \rho)$ for some unitary $U(\boldsymbol{\alpha})$ by randomly sampling from the dataset $\{|\psi^{(i)}\rangle\}$ that decomposes ρ. To ensure that the deviation from the true cost function value $C_{\text{VQSE}}(U(\boldsymbol{\alpha}), \rho)$ is smaller than ε with probability $(1 - \delta)$, it suffices for the number of samples to be*

$$M = \frac{2\|H\|_\infty^2 \log(2/\delta)}{\varepsilon^2} \tag{71}$$

with H given by Eq. (61).

One key feature of this result is the absence of N, the total number of states in the dataset. Therefore, we can conclude that the number of samples necessary in order to obtain a good cost function estimate does not scale with the number of datapoints that make up the dataset. Sampling to obtain good cost function estimates would then in turn lead to a successful optimization, and therefore accurate determination of the eigenvalues and eigenvectors corresponding to the principal components.

The above analysis can also be extended to the number of shots needed to obtain an accurate cost function value. If one considers the limiting case where only one shot is used for the evaluation of the cost function for each state, the same argument given above applies. This leads to the conclusion that only $9 \log(2/\delta)/2\varepsilon^2$ shots are necessary to compute the cost function with accuracy ε and confidence δ for the case of C_{VQSD}. Each shot is taken while also randomly sampling from the dataset $\{|\psi^{(i)}\rangle\}$. A similar result follows for C_{VQSE}.

Overall, when looking at using our method to prepare the covariance matrix as a subroutine in VQSD and VQSE, we find that the sampling overhead necessary to still obtain an optimization close to the exact case is favorable. Indeed, we find essentially no dependence on the total size of the dataset.

C. Remark about the original quantum PCA algorithm

We emphasize that we focus on VQSD and VQSE instead of the original quantum PCA algorithm [3] because the latter is a long-term algorithm and we are aiming at a near-term approach to quantum PCA. The simplicity of our method is ideally suited to noisy intermediate-scale quantum (NISQ) devices. Hence, our method partners well with other NISQ algorithms (for state diagonalization).

Nevertheless, one can apply our approach (of preparing the ensemble average density matrix $\overline{\rho}$) in the context of the original quantum PCA algorithm as well. Recall that the algorithm of Lloyd *et al.* [3] algorithm uses c copies of $\overline{\rho}$ in order to approximately implement the $e^{-i\overline{\rho}t}$ gate in the context of quantum phase estimation (QPE). Here, one also feeds $\overline{\rho}$ as the input state into the QPE circuit. Their algorithm requires $c \in \mathcal{O}(1/\epsilon^3)$ copies of $\overline{\rho}$ in order to extract its eigenvalues and eigenvectors with accuracy ϵ.

One can employ a deterministic approach to prepare $\overline{\rho}$, as discussed in Sec. V A. Operationally speaking, this approach requires the simultaneous decomposition of both the input state $\overline{\rho}$ and the copies of $\overline{\rho}$ that are used to approximate the $e^{-i\overline{\rho}t}$ gate implemented in the QPE circuit. Let us now discuss how this decomposition leads to the expected (i.e., correct) output, due to the linearity of the operations involved. In this setting, one would: (1) choose a state from the dataset (i.e., from the decomposition of $\overline{\rho}$), (2) feed this state into the QPE circuit, (3) estimate observable expectation values on the resulting state (to characterize the eigenvalues and eigenvectors), and (4) average these observable expectation values over all of the states in the dataset. By the linearity of the QPE unitary and of the trace, the average of the expectation values is the same as the expectation value of the average, and hence this method reproduces the correct results. In addition, as previously mentioned, one would also need to choose states from the dataset for each ancilla system that is used to approximate the $e^{-i\overline{\rho}t}$ gate using exponential SWAP gates followed by a partial trace over the ancilla. Once again, due to the linearity of the exponential SWAP gates and of the partial trace, averaging over the dataset gives the correct result.

In this deterministic setting, one would need N state preparations for each copy of $\overline{\rho}$, i.e., for the input system to QPE and for the ancilla systems used to approximate the $e^{-i\overline{\rho}t}$ gate. If a total of c copies of $\overline{\rho}$ are employed then one would potentially need N^c state preparations with this approach.

We remark that a random sampling approach, instead of a deterministic approach, to preparing each copy of $\overline{\rho}$ could alternatively be used. See Sec. V B for further discussion of this approach. Although a detailed analysis of random sampling in this context is beyond the scope of this work, we do believe that random sampling would likely significantly reduce the number of state preparations required,

as compared to the deterministic case. This intuition arises from Hoeffding's bound, which implies that an estimator will concentrate about its mean value as one increases the number of samples, and this concentration guarantee is essentially independent of the dataset size N.

VI. NUMERICAL IMPLEMENTATIONS

Here we perform PCA (i.e., diagonalizing Q) and simulate quantum PCA (i.e., diagonalizing $\overline{\rho}$) on two different datasets. The first dataset is classical and is the famous MNIST dataset of handwritten digits. The second dataset is the quantum dataset of molecular ground states for various interatomic distances.

A. MNIST implementation

From the MNIST dataset, we randomly select 5000 instances of each digit from 0 to 9. We vectorize the 28×28 gray-scaled images, leading to vectors $\mathbf{y}^{(i)}$. This gives $d = 784$ features for standard PCA, and for our simulation of quantum PCA, we trivially embed the data in a $(d = 2^{10} = 1024)$-dimensional feature space corresponding to the Hilbert space of ten qubits. Each vector is normalized such that $(\mathbf{y}^{(i)})^T \mathbf{y}^{(i)} = 1$ This results in a dataset $D = \{\mathbf{y}^{(i)}\}_{i=1}^N$ with $N = 50\,000$. We assume a uniform probability distribution so that the ensemble is $E = \{1/N, \mathbf{y}^{(i)}\}$.

In our numerics, we perform PCA and we simulate quantum PCA for this dataset. We then compare the principal components produced by each approach. For standard PCA, we diagonalize the covariance matrix Q formed from the vectors in D. The top n principal components are the eigenvectors corresponding to the n largest eigenvalues. For our simulation of quantum PCA, we prepare $\bar{\rho}$ as outlined in the text above. We then diagonalize $\bar{\rho}$ to give the quantum principal components.

For verification purposes, we also perform standard PCA on a symmetrized dataset

$$\tilde{D} = D \cup (-D) = \{\mathbf{y}^{(i)}\}_{i=1}^N \cup \{-\mathbf{y}^{(i)}\}_{i=1}^N. \quad (72)$$

We verify that we obtain the same spectrum from performing PCA with this symmetrized dataset as we obtain with our simulation of quantum PCA. This is expected as both scenarios effectively involve a symmetrized dataset.

1. Eigenvalues and eigenvectors

We now discuss the spectrum that we obtained in the two cases. In Fig. 5(a) we show the eigenvectors associated with the first five principal components for PCA and the first six for quantum PCA. We see an intriguing correspondence between the eigenvectors obtained via standard and quantum PCA. Specifically, $|Q_{d-j}\rangle$ appears to be very similar to $|R_{d-j-1}\rangle$. Indeed, one can see in
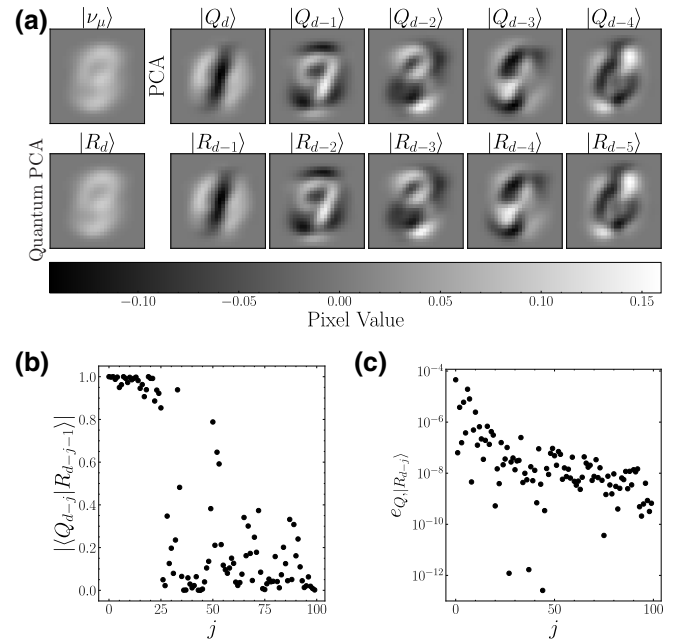

FIG. 5. Eigenvectors for the MNIST dataset of handwritten digits. We perform standard PCA (diagonalizing Q) and quantum PCA (diagonalizing $\overline{\rho}$). (a) The images of the first five principal components for PCA and the first six for quantum PCA are shown. The first principal component for quantum PCA is close to (has a large overlap with) the mean vector $|v_\mu\rangle$, which is also shown. Also, one can visually see that $|Q_{d-j}\rangle$ is quite similar to $|R_{d-j-1}\rangle$ for these eigenvectors. (b) We plot the magnitude of the overlap of the principal components calculated from PCA and quantum PCA, namely, $|Q_{d-j}\rangle$ and $|R_{d-j-1}\rangle$. This overlap remains large when j is less than 25. (c) Finally, we show the eigenvector error, which quantifies how far the eigenvectors of $\overline{\rho}$ are from being eigenvectors of Q.

Fig. 5(b) that the overlap $|\langle Q_{d-j}|R_{d-j-1}\rangle|^2$ is close to one for $j < 25$. For larger j, this overlap is smaller; however, this does not mean that the $|R_{d-j-1}\rangle$ states are not approximate eigenvectors of Q. To clarify this point, we also plot the eigenvector error in Fig. 5(c). For all values of j shown, the eigenvector error is below 10^{-4}. This suggests that the eigenvectors of $\overline{\rho}$ serve as approximate eigenvectors for Q for this dataset.

We see a similar pattern in the eigenvalues, shown in Fig. 6. Excluding the largest eigenvalue from quantum PCA, the eigenvalues match very well for the two methods. More specifically, r_{d-j-1} is very close to q_{d-j}, and this is shown in more detail in the inset of Fig. 6.

The close correspondence between both the eigenvectors and the eigenvalues for the two methods is likely a consequence of Proposition 5. The assumption in Proposition 5 is almost satisfied, i.e., one of the eigenvectors of $\overline{\rho}$ is close to the mean vector. Specifically, the first principal component of $\overline{\rho}$, which is $|R_d\rangle$ and is displayed in Fig. 5(a), has an infidelity of roughly 2.8×10^{-4} with the
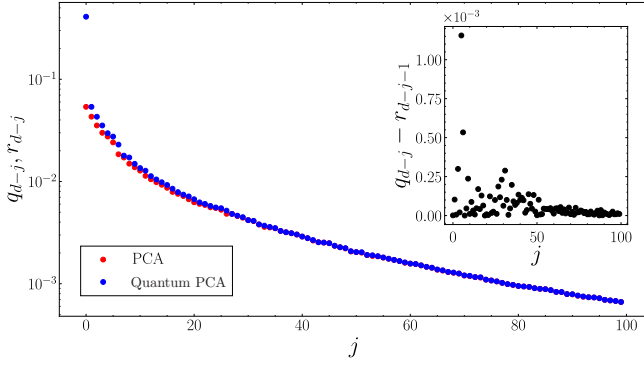
FIG. 6. Eigenvalues for the MNIST dataset of handwritten digits. We perform standard PCA (diagonalizing Q) and quantum PCA (diagonalizing $\overline{\rho}$). The eigenvalues calculated using PCA and quantum PCA are plotted. Note that q_{d-j} is quite similar to r_{d-j-1} for these principal components. In the inset we plot the difference between the shifted eigenvalues q_{d-j} and r_{d-j-1} calculated using PCA and quantum PCA, respectively.

mean vector $|v_\mu\rangle$. The $|R_d\rangle$ eigenvector appears to capture the bias of the dataset, i.e., the fact that the dataset is biased towards white colored images. The fact that $|R_d\rangle$ is close to the mean vector suggests that the spectral decomposition in Eq. (49) is almost valid. Hence, the spectra of $\overline{\rho}$ and Q are close to matching.

2. Principal component projections

In order to assess how the two methods for calculating the principal components perform, we construct a projected image by projecting each image onto a reduced subspace

$$|\hat{y}_n^{(i)}\rangle = \sum_{j=0}^{n-1} \langle \chi_{d-j}|y^{(i)}\rangle |\chi_{d-j}\rangle, \qquad (73)$$

where $|\hat{y}_n^{(i)}\rangle$ is the projected vector using n principal components $|\chi_{d-j}\rangle$ calculated from either PCA or quantum PCA. We can explore the infidelity between the projected vector and the original as

$$I = 1 - |\langle \hat{y}_{N_c}^{(i)}|y^{(i)}\rangle|. \qquad (74)$$

In Fig. 7(a) we show the median and 90% interval of the infidelity over every image as a function of n for the classical and quantum cases. The performance of PCA and quantum PCA is similar with roughly 100 components necessary for a median infidelity of 0.1. Therefore, despite the differences in the calculated eigenvalues and eigenvectors, both methods can be used to accurately compress the data on average.

In Fig. 7(b) we show how one projected image in the MNIST dataset appears visually as we increase the number of principal components used. There is a noticeable
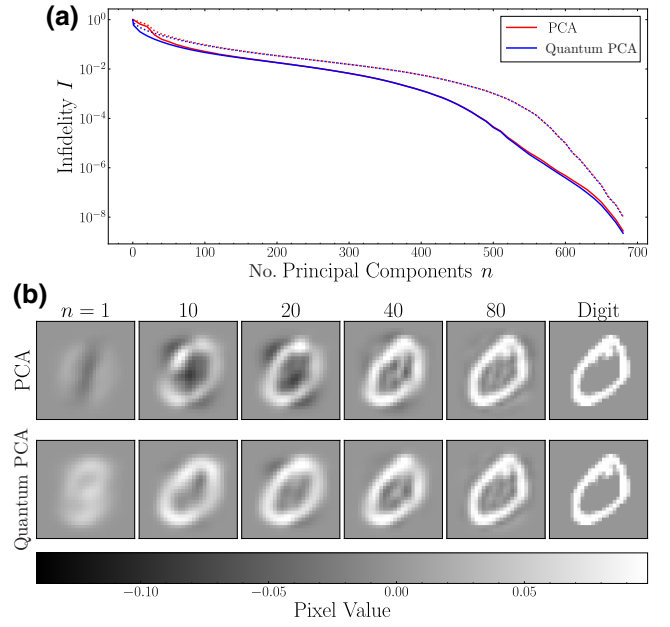

FIG. 7. Projection onto n principal components for PCA and quantum PCA of the MNIST dataset. (a) The plot shows the median (solid) and upper 90% interval (dotted) infidelity (I) values between the projected state constructed using n principal components calculated from PCA and quantum PCA. The median and upper interval are calculated over all 50 000 digits. (b) We show how the images are reproduced when using $n = 1$ to $n = 80$ principal components from PCA and quantum PCA in the upper and lower rows, respectively.

difference in the $n = 1$ case, due to $|R_d\rangle$ being quite different from $|Q_d\rangle$. However, by $n = 80$, the two projected images appear visually similar and also close to the true image. It therefore appears that quantum PCA is a successful surrogate for PCA for this dataset.

B. Molecular ground-state implementation

We simulate the task of performing quantum PCA on molecular ground states of the H_2 molecule in the 6-31g basis and the BeH_2 in the Slater-type orbital, three Gaussian basis. These implementations require system sizes of 8 qubits and 14 qubits, respectively, corresponding to feature space dimensions of $d = 256$ and $d = 16\,384$. The ground state is calculated by classically simulating the variational quantum eigensolver (VQE) [28] for 401 equally spaced interatomic distances $r \in [0.3, 2.3]$.

We find that the principal components and eigenvalues calculated using standard PCA and quantum PCA are almost identical in this case. This is due to the fact that our simulations of VQE produce ground states with random global phases, and this naturally results in a dataset that is approximately centered, especially as one increases the number of datapoints. Hence, in this case, Proposition 1 applies, and diagonalizing $\overline{\rho}$ is equivalent to diagonalizing
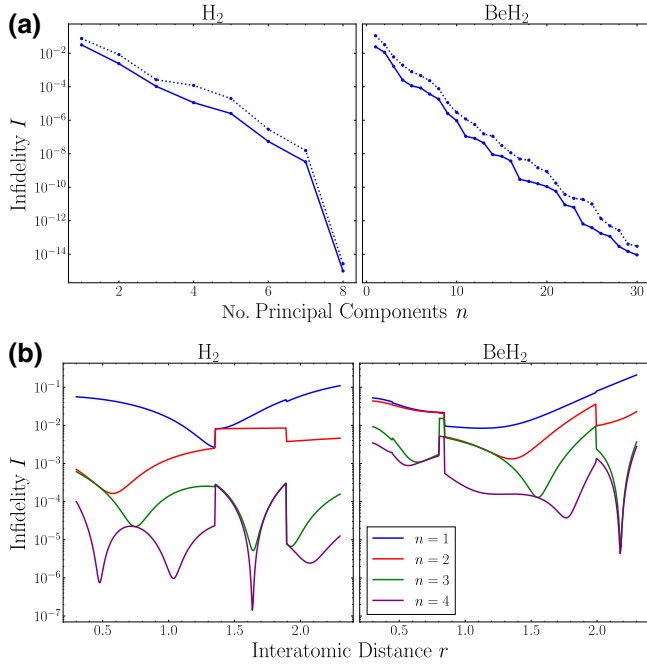
(a)

(b)

FIG. 8. Quantum PCA performed on ground states of molecules at different interatomic distances. (a) The plot shows the infidelity (I) between the original state and the projected state constructed using n principal components calculated by simulating quantum PCA. The median (solid line) and upper 90% interval (dotted line) are calculated over all 401 interatomic distances r. (b) The infidelity is plotted as a function of interatomic distance r for projected states calculated with $n = 1, 2, 3, 4$ principal components.

Q. (This point would be irrelevant if we assume that the dataset is truly generated on a quantum device, rather than through our classical simulations, since global phase is unphysical and then Proposition 6 would apply.)

Recall from Eq. (74) that the infidelity quantifies the inability to recover the original data from the compressed data. In Fig. 8(a) we show the median and 90% interval of the infidelity over every ground state as a function of n for the two molecules. The projected ground states have a very low infidelity with the actual ground states for low numbers of principal components, especially in the case of H_2. Therefore, there exists an accurate, more efficient representation of these states, which can be calculated with quantum PCA. In Fig. 8(b) we show the infidelity I as a function of the interatomic distance r for projected states calculated with $n = 1, 2, 3, 4$ principal components for both molecules. One can see that the curves monotonically decrease with n.

VII. DISCUSSION

Recently, there has been renewed interest in quantum algorithms for PCA. The history here is quite interesting. The quantum PCA algorithm of Lloyd *et al.* was proposed in 2014 and highlighted a potential exponential speedup over classical algorithms [3], and this also led to some near-term proposals for quantum PCA [4–7]. However, a spooky paper posted on Halloween of 2018 by Tang presented a "dequantized" classical algorithm that achieved the same asymptotic scaling as quantum PCA [11], suggesting that quantum speedup for quantum PCA was a false promise. But this was not the end of the story. In December 2021, a team from Caltech and Google posted two papers [12,26] arguing that exponential quantum speedup is possible for quantum PCA, and even for near-term algorithms, since (as they argued) dequantized classical algorithms are artificially given too much power via their mode of access to quantum state amplitudes. As a consequence of this latest work, we are left with the exciting possibility that exponential quantum speedup remains possible for quantum PCA, particularly for analysis of quantum data. It is worth remarking that Tang's dequantization results still apply to the analysis of classical data [12], although this does not preclude the possibility of modest (e.g., constant factor) quantum speedups for classical data analysis [13].

This interesting historical trajectory for quantum PCA makes our work even more important. A crucial piece of the puzzle was missing in this field, and that was a method for preparing the covariance matrix, given a dataset of quantum states. We have made major strides towards filling this gap, by proposing the ensemble average density matrix as a surrogate for the covariance matrix. We argued that this surrogate was equal to the covariance matrix for arbitrary quantum datasets or for centered classical datasets. Given the potential for exponential speedup with quantum PCA for quantum datasets [12,26], our results for these datasets are technologically important. Therefore we placed significant emphasis on PCA for quantum datasets in this work, providing a detailed discussion of complex random variables (see Sec. II) and showcasing an implementation for molecular ground states (see Sec. VI).

On the other hand, for uncentered classical datasets, we showed that diagonalizing this surrogate matrix corresponds to "PCA without centering," or PCA for a symmetrized version of the dataset as shown in Fig. 3. We nevertheless derived results that bound the deviation of the spectrum obtained with our method from the true PCA spectrum, indicating a close correspondence with true PCA. (At the mathematical level, our results here generalize and extend those of Ref. [22].) Hence, we believe that our method will be useful even for uncentered classical datasets.

We remark that a recent experimental quantum PCA implementation [29] employed a nonscalable method for covariance matrix preparation. They classically optimized over circuits to prepare the covariance matrix, which has exponential scaling with problem size. Moreover, that approach would be unnatural for quantum datasets, since

one would have to first readout the quantum states, leading to an additional source of exponential scaling. Thus, our approach of preparing $\overline{\rho}$ fills an important gap in the literature, towards scalable covariance matrix preparation, especially for quantum datasets [30].

Natural future work would be to actually implement our method on real quantum hardware. Combining our method for preparing the covariance matrix with other near-term methods for extracting the spectrum [4–7] would lead to a near-term approach for quantum PCA. Indeed, the nice feature of our method is how simple and easy to implement it is on near-term quantum hardware.

APPENDIX A: WEYL'S THEOREM FOR EIGENVALUES

For our purposes, we need a theorem from Weyl about the eigenvalues of Hermitian matrices. This theorem is sometimes called Weyl's inequality. A detailed discussion and proof of this theorem can be found in the textbook of Horn and Johnson [25]. We repeat the theorem here, as it is stated in Ref. [25].

Lemma 3: *Let A and B be Hermitian $d \times d$ matrices. Let the respective eigenvalues of A, B, and $A + B$ be $\{\lambda_j(A)\}_{j=1}^d$, $\{\lambda_j(B)\}_{j=1}^d$, and $\{\lambda_j(A + B)\}_{j=1}^d$. Suppose that the eigenvalues in these sets are listed in nondecreasing order. Then*

$$\lambda_j(A + B) \leqslant \lambda_{j+k}(A) + \lambda_{d-k}(B) \tag{A1}$$

for each $j = 1, \ldots, d$ and $k = 0, 1, \ldots, d - j$. In addition,

$$\lambda_{j-k+1}(A) + \lambda_k(B) \leqslant \lambda_j(A + B) \tag{A2}$$

for each $j = 1, \ldots, d$ and $k = 1, \ldots, j$.

By setting $k = 0$ and $k = 1$ in Eq. (A1) and by setting $k = 1$ in Eq. (A2), we obtain the following corollary.

Lemma 4: *Let A and B be Hermitian $d \times d$ matrices. Let the respective eigenvalues of A, B, and $A + B$ be*

$\{\lambda_j(A)\}_{j=1}^d$, $\{\lambda_j(B)\}_{j=1}^d$, *and* $\{\lambda_j(A + B)\}_{j=1}^d$. *Suppose that the eigenvalues in these sets are listed in nondecreasing order. Then*

$$\lambda_j(A + B) \leqslant \lambda_j(A) + \lambda_d(B) \quad for\ j = 1, \ldots, d, \tag{A3}$$

$$\lambda_j(A + B) \leqslant \lambda_{j+1}(A) + \lambda_{d-1}(B)$$
$$for\ j = 1, \ldots, d - 1, \tag{A4}$$

$$\lambda_j(A) + \lambda_1(B) \leqslant \lambda_j(A + B) \quad for\ j = 1, \ldots, d. \tag{A5}$$

Finally, by specializing the result even further to the case where B is rank one with a positive eigenvalue, we obtain the following result.

Lemma 5: *Let A and B be Hermitian $d \times d$ matrices. Let the respective eigenvalues of A, B, and $A + B$ be $\{\lambda_j(A)\}_{j=1}^d$, $\{\lambda_j(B)\}_{j=1}^d$, and $\{\lambda_j(A + B)\}_{j=1}^d$. Suppose that the eigenvalues in these sets are listed in nondecreasing order. Suppose that B is rank one and that the only nonzero eigenvalue of B is $\lambda_B > 0$. Then*

$$\lambda_j(A + B) \leqslant \lambda_j(A) + \lambda_B \quad for\ j = 1, \ldots, d, \tag{A6}$$

$$\lambda_j(A + B) \leqslant \lambda_{j+1}(A) \quad for\ j = 1, \ldots, d - 1, \tag{A7}$$

$$\lambda_j(A) \leqslant \lambda_j(A + B) \quad for\ j = 1, \ldots, d. \tag{A8}$$

APPENDIX B: RELATING THE DIAGONAL ELEMENTS OF $\overline{\rho}$ AND Q

In the main text, we related the eigenvalues and eigenvectors of $\overline{\rho}$ and Q. In this appendix, we note that one can also bound the deviation of the diagonal elements of $\overline{\rho}$ and Q. This is different from bounding the deviation of the ordered eigenvalues, and hence the following proposition is distinct from Proposition 3. In fact, there is a simple equation that relates the diagonal elements of $\overline{\rho}$ and Q, as follows.

Proposition 9: *Consider a classical dataset of pure states $D_{|\psi\rangle} = \{|\psi^{(i)}\rangle\}_{i=1}^N$, with the ensemble denoted as $E_{|\psi\rangle} = \{p^{(i)}, |\psi^{(i)}\rangle\}_{i=1}^N$. Let $|\phi\rangle$ be any quantum state, and let $r_{|\phi\rangle} = \langle\phi|\overline{\rho}|$ and $q_{|\phi\rangle} = \langle\phi|Q|$. Then*

$$q_{|\phi\rangle} \leqslant r_{|\phi\rangle} \leqslant q_{|\phi\rangle} + \|\boldsymbol{\mu}\|^2. \tag{B1}$$

More specifically, the equation

$$r_{|\phi\rangle} = q_{|\phi\rangle} + \|\boldsymbol{\mu}\|^2 |\langle v_{\boldsymbol{\mu}}|\phi\rangle|^2 \tag{B2}$$

holds, where $\boldsymbol{\mu}$ is the mean vector and $|v_{\boldsymbol{\mu}}\rangle = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$ is the normalized mean vector.

Proof. The inequality follows from the equation in the proposition. The equation follows from $\overline{\rho} = Q + M$ and $M = \|\boldsymbol{\mu}\|^2 |v_{\boldsymbol{\mu}}\rangle\langle v_{\boldsymbol{\mu}}|$, which gives the desired result through

$$\langle\phi|\overline{\rho}|\phi\rangle = \langle\phi|Q|\phi\rangle + \langle\phi|M|\phi\rangle. \tag{B3}$$

∎

APPENDIX C: EXTENSION TO MIXED-STATE DATASETS

We now discuss how the above results can be extended to mixed-state datasets. For mixed-state datasets, there exist straightforward mathematical generalizations of the previous results. The conceptual interpretation of such results is less obvious. Nevertheless, we discuss this below.

1. Mixed-state dataset

Consider a dataset of mixed states, which we denote as

$$D_\rho = \{\rho^{(i)}\}_{i=1}^N. \tag{C1}$$

We assume a probability distribution over datapoints given by $P = \{p^{(i)}\}$. Hence, one can define the corresponding ensemble for this dataset as

$$E_\rho = \{p^{(i)}, \rho^{(i)}\}_{i=1}^N. \tag{C2}$$

For this ensemble, the ensemble average density matrix is

$$\overline{\rho} = \sum_{i=1}^N p^{(i)}\rho^{(i)}. \tag{C3}$$

2. Effective pure-state dataset

Let us note that each mixed-state datapoint can be decomposed as a convex combination of pure states:

$$\rho^{(i)} = \sum_{m=1}^{R_i} s^{(i,m)}|\psi^{(i,m)}\rangle\langle\psi^{(i,m)}|. \tag{C4}$$

Hence, we can rewrite the ensemble average density matrix as

$$\begin{aligned}\overline{\rho} &= \sum_{i=1}^N p^{(i)}\rho^{(i)} \\ &= \sum_{i=1}^N \sum_{m=1}^{R_i} p^{(i)}s^{(i,m)}|\psi^{(i,m)}\rangle\langle\psi^{(i,m)}|. \end{aligned} \tag{C5}$$

We can introduce an index $k = (i, m)$ and then we have

$$\overline{\rho} = \sum_{k=1}^{\widehat{N}} \widehat{p}^{(k)}|\psi^{(k)}\rangle\langle\psi^{(k)}|, \tag{C6}$$

where

$$\widehat{N} = \sum_{i=1}^N R_i \quad \text{and} \quad \widehat{p}^{(k)} = p^{(i)}s^{(i,m)}. \tag{C7}$$

The above equation suggests that we could interpret

$$\widehat{D}_\rho = \{|\psi^{(k)}\rangle\langle\psi^{(k)}|\}_{k=1}^{\widehat{N}} \tag{C8}$$

as an effective dataset, with pure-state datapoints. The corresponding effective ensemble is then

$$\widehat{E}_\rho = \{\widehat{p}^{(k)}, |\psi^{(k)}\rangle\langle\psi^{(k)}|\}_{k=1}^{\widehat{N}}. \tag{C9}$$

With these definitions in hand, we can now see how our main results, for pure-state datasets, can generalize to mixed-state datasets. If we are willing to reinterpret the dataset as $\widehat{D}_\rho$ and the ensemble as $\widehat{E}_\rho$ then all our results can be extended to this case. The idea is that we relate $\overline{\rho}$ to the covariance matrix for an unraveling of the ensemble $\widehat{E}_\rho$. Consequently, the conceptual interpretation of diagonalizing $\overline{\rho}$ is that it corresponds to performing PCA on a dataset composed of the pure states that decompose the mixed states in the original dataset. This interpretation is made precise in Proposition 10 below.

3. Theoretical results

Mixed-state datapoints naturally lack a global phase, since global phases are only relevant for statevector datasets. Therefore, in the context of mixed-state datasets, it seems natural to consider our results above for the case where global phase information is not important, or not relevant. Such results were given in Lemma 2 and Proposition 6. Hence, in what follows we state analogs of Lemma 2 and Proposition 6 for mixed-state datasets.

Lemma 6: *Consider an ensemble of mixed states* $E_\rho = \{p^{(i)}, \rho^{(i)}\}_{i=1}^N$. *Let* $\widehat{E}_\rho = \{\widehat{p}^{(k)}, |\psi^{(k)}\rangle\langle\psi^{(k)}|\}_{k=1}^{\widehat{N}}$ *be the corresponding effective ensemble of pure states, as given in Eqs. (C7) and (C9). Then, there always exists a statevector ensemble* $E_{|\tilde{\psi}\rangle} = \{\tilde{p}^{(j)}, |\tilde{\psi}^{(j)}\rangle\}_{j=1}^{\widehat{N}}$ *that satisfies the following conditions.*

(a) Ensemble $E_{|\tilde{\psi}\rangle}$ *physically corresponds to the aforementioned* $\widehat{E}_\rho$, *in the sense that applying the outer product mapping leads to* $\mathcal{P}(E_{|\tilde{\psi}\rangle}) = \widehat{E}_\rho$. *In other words,* $E_{|\tilde{\psi}\rangle}$ *is an unraveling of* $\widehat{E}_\rho$.
(b) Ensemble $E_{|\tilde{\psi}\rangle}$ *is centered, i.e., the mean value of all features is zero.*

Proof. The proof is essentially the same as the proof of Lemma 2. In fact, one can view this as a corollary of Lemma 2, where one applies Lemma 2 to ensemble $\widehat{E}_\rho$.

A statevector ensemble that satisfies the two criteria stated in the lemma is

$$E_{|\tilde\psi\rangle} = \left\{ \left(\frac{\widehat{p}^{(1)}}{2}, |\psi^{(1)}\rangle\right), \ldots, \left(\frac{\widehat{p}^{(\widehat{N})}}{2}, |\psi^{(\widehat{N})}\rangle\right), \right.$$
$$\left. \left(\frac{\widehat{p}^{(1)}}{2}, -|\psi^{(1)}\rangle\right), \ldots, \left(\frac{\widehat{p}^{(\widehat{N})}}{2}, -|\psi^{(\widehat{N})}\rangle\right) \right\}.$$
(C10)

Note that this ensemble is centered, due to its symmetric nature.

In addition, one can see that $\mathcal{P}(E_{|\tilde\psi\rangle}) = \widehat{E}_\rho$, and hence $E_{|\tilde\psi\rangle}$ is an unraveling of $\widehat{E}_\rho$. This follows because $\mathcal{P}(E_{|\tilde\psi\rangle})$ has $\widehat{N}$ datapoints that appear twice in the ensemble, and these redundant datapoints can be aggregated to give $E_{|\tilde\psi\rangle}$. ∎

With the previous lemma in hand, we can state the following proposition, which is our main result for mixed-state datasets. Note that the following proposition generalizes the result in Proposition 6.

Proposition 10: *Consider an ensemble of mixed states $E_\rho = \{p^{(i)}, \rho^{(i)}\}_{i=1}^N$. Let $\widehat{E}_\rho = \{\widehat{p}^{(k)}, |\psi^{(k)}\rangle\langle\psi^{(k)}|\}_{k=1}^{\widehat{N}}$ be the corresponding effective ensemble of pure states, as given in Eqs. (C7) and (C9). Then, there always exists a statevector ensemble $E_{|\tilde\psi\rangle} = \{\tilde{p}^{(j)}, |\tilde\psi^{(j)}\rangle\}_{j=1}^{\widetilde{N}}$ that satisfies the following conditions.*

- *(a) Ensemble $E_{|\tilde\psi\rangle}$ physically corresponds to the aforementioned $\widehat{E}_\rho$, in the sense that applying the outer product mapping leads to $\mathcal{P}(E_{|\tilde\psi\rangle}) = \widehat{E}_\rho$. In other words, $E_{|\tilde\psi\rangle}$ is an unraveling of $\widehat{E}_\rho$.*
- *(b) The covariance matrix $Q(E_{|\tilde\psi\rangle})$ for $E_{|\tilde\psi\rangle}$ is equal to the ensemble average density matrix for E_ρ:*

$$Q(E_{|\tilde\psi\rangle}) = \overline{\rho}(E_\rho). \tag{C11}$$

Proof. We prove this by constructing an $E_{|\tilde\psi\rangle}$ that satisfies the required conditions. Specifically, we use the ensemble in Eq. (C10). Of course, we already showed above that this ensemble is an unraveling of $\widehat{E}_\rho$. So we just need to prove the second condition. Let us first note that $E_{|\tilde\psi\rangle}$ is centered, as discussed above. Hence we can apply Proposition 1 to see that

$$Q(E_{|\tilde\psi\rangle}) = \overline{\rho}(E_{|\tilde\psi\rangle}). \tag{C12}$$

Next we note that $\overline{\rho}(E_{|\tilde\psi\rangle}) = \overline{\rho}(\widehat{E}_\rho)$, since $E_{|\tilde\psi\rangle}$ is an unraveling of $\widehat{E}_\rho$. Finally, it is clear that $\overline{\rho}(\widehat{E}_\rho) = \overline{\rho}(E_\rho)$, which follows from Eq. (C5). This completes the proof. ∎

[1] Ian T. Jolliffe and Jorge Cadima, Principal component analysis: A review and recent developments, Philos. Trans. R. Soc. A: Math., Phys. Eng. Sci. **374**, 20150202 (2016).

[2] Karl Pearson, LIII. On lines and planes of closest fit to systems of points in space, Mag. J. Sci. **2**, 559 (1901).

[3] Seth Lloyd, Masoud Mohseni, and Patrick Rebentrost, Quantum principal component analysis, Nat. Phys. **10**, 631 (2014).

[4] Ryan LaRose, Arkin Tikku, Étude O'Neel-Judy, Lukasz Cincio, and Patrick J. Coles, Variational quantum state diagonalization, npj Quantum Inf. **5**, 1 (2019).

[5] M. Cerezo, Kunal Sharma, Andrew Arrasmith, and Patrick J. Coles, Variational quantum state eigensolver (2020), arXiv preprint ArXiv:2004.01372.

[6] Guillaume Verdon, Jacob Marks, Sasha Nanda, Stefan Leichenauer, and Jack Hidary, Quantum Hamiltonian-based models and the variational quantum thermalizer algorithm (2019), arXiv preprint ArXiv:1910.02071.

[7] Nic Ezzell, Zoë Holmes, and Patrick J. Coles, The quantum low-rank approximation problem (2022), arXiv preprint ArXiv:2203.00811.

[8] M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles, Variational quantum algorithms, Nat. Rev. Phys. **3**, 625 (2021).

[9] M. Cerezo, Alexander Poremba, Lukasz Cincio, and Patrick J. Coles, Variational quantum fidelity estimation, Quantum **4**, 248 (2020).

[10] Hsin-Yuan Huang, Michael Broughton, Jordan Cotler, Sitan Chen, Jerry Li Masoud Mohseni, Hartmut Neven, Ryan Babbush, Richard Kueng, John Preskill, *et al.*, Quantum advantage in learning from experiments, Science **376**, 1182 (2022).

[11] Ewin Tang, Quantum Principal Component Analysis Only Achieves an Exponential Speedup because of Its State Preparation Assumptions, Phys. Rev. Lett. **127**, 060503 (2021).

[12] Jordan Cotler, Hsin-Yuan Huang, and Jarrod R. McClean, Revisiting dequantization and quantum advantage in learning tasks (2021), arXiv preprint ArXiv:2112.00811.

[13] Juan Miguel Arrazola, Alain Delgado, Bhaskar Roy Bardhan, and Seth Lloyd, Quantum-inspired algorithms in practice, Quantum **4**, 307 (2020).

[14] Lov K. Grover, Synthesis of Quantum Superpositions by Quantum Computation, Phys. Rev. Lett. **85**, 1334 (2000).

[15] Lov Grover and Terry Rudolph, Creating superpositions that correspond to efficiently integrable probability distributions (2002), arXiv preprint ArXiv:quant-ph/0208112.

[16] Martin Plesch and Časlav Brukner, Quantum-state preparation with universal gate decompositions, Phys. Rev. A **83**, 032302 (2011).

[17] Maria Schuld and Francesco Petruccione, *Supervised Learning with Quantum Computers*, Vol. 17 (Springer, Cham, Switzerland, 2018).

[18] Yuval R. Sanders, Guang Hao Low, Artur Scherer, and Dominic W. Berry, Black-Box Quantum State Preparation without Arithmetic, Phys. Rev. Lett. **122**, 020502 (2019).

[19] Kouhei Nakaji, Shumpei Uno, Yohichi Suzuki, Rudy Raymond, Tamiya Onodera, Tomoki Tanaka, Hiroyuki Tezuka,

Naoki Mitsuda, and Naoki Yamamoto, Approximate amplitude encoding in shallow parameterized quantum circuits and its application to financial market indicator (2021), arXiv preprint ArXiv:2103.13211.

[20] Gabriel Marin-Sanchez, Javier Gonzalez-Conde, and Mikel Sanz, Quantum algorithms for approximate function loading (2021), arXiv preprint ArXiv:2111.07933.

[21] Christa Zoufal, Aurélien Lucchi, and Stefan Woerner, Quantum generative adversarial networks for learning and loading random distributions, npj Quantum Inf. **5**, 1 (2019).

[22] Jorge Cadima and Ian Jolliffe, On relationships between uncentred and column-centred principal component analysis, Pak. J. Stat. **25**, 473 (2009).

[23] Michael A. Nielsen and Isaac L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, 2000).

[24] The open-quantum-system literature refers to a stochastic statevector time evolution as an unraveling of the master equation for the density matrix.

[25] Roger A. Horn and Charles R. Johnson, *Matrix Analysis* (Cambridge University Press, Cambridge, 2012).

[26] Hsin-Yuan Huang, Michael Broughton, Jordan Cotler, Sitan Chen, Jerry Li, Masoud Mohseni, Hartmut Neven, Ryan Babbush, Richard Kueng, John Preskill, and Jarrod R. McClean, Quantum advantage in learning from experiments (2021), arXiv preprint ArXiv:2112.00778.

[27] M. Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J. Coles, Cost function dependent barren plateaus in shallow parametrized quantum circuits, Nat. Commun. **12**, 1 (2021).

[28] Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J. Love, Alán Aspuru-Guzik, and Jeremy L. O'Brien, A variational eigenvalue solver on a photonic quantum processor, Nat. Commun. **5**, 1 (2014).

[29] Tao Xin, Liangyu Che, Cheng Xi, Amandeep Singh, Xinfang Nie, Jun Li, Ying Dong, and Dawei Lu, Experimental Quantum Principal Component Analysis via Parametrized Quantum Circuits, Phys. Rev. Lett. **126**, 110502 (2021).

[30] For classical datasets, the scaling of our method will be determined by the scaling of the amplitude encoding step, which is still an active area of research [14–21].