Toward single particle reconstruction without particle picking: Breaking the detection limit

Tamir Bendory, Nicolas Boumal, William Leeb and Amit Singer

June 11, 2018

Abstract

Here comes the abstract

1 Introduction

[Revise–Cryo–electron microscopy (cryo–EM) is an innovative technology for single particle reconstruction (SPR) of macromolecules.] In a cryo–EM experiment, biological samples are rapidly frozen in a thin layer of vitreous ice. Within the ice, the molecules are randomly oriented and positioned. The microscope produces a 2-D tomographic image of the samples embedded in the ice, called a *micrograph*. Each micrograph contains tomographic projections of the samples at unknown locations and under unknown viewing directions. The goal is to construct 3-D models of the molecules from the micrographs.

The signal to noise ratio (SNR) of the projections in the micrographs is a function of two dominating factors. On the one hand, the SNR is a function of the electron dose. To keep radiation damage within acceptable bounds, the dose must be kept low, which leads to high noise levels. On the other hand, the SNR is a function of the molecule size. The smaller the molecules, the fewer detected electrons carry information about them.

All contemporary methods in the field split the reconstruction procedure into several stages. The first stage consists in extracting the various particle projections from the micrographs. This is called *particle picking*. Later stages aim to construct a 3-D model of the molecule from these projections. The quality of the reconstruction eventually hinges on the quality of the particle picking stage. Figure 1 illustrates how particle picking becomes increasingly challenging as the SNR degrades.

Crucially, it can be shown that reliable detection of individual particles is impossible below a certain critical SNR. This fact has been recognized early on by the cryo-EM community. In particular, in an influential paper from 1995, Henderson [28] investigates the following questions:

> *For the purposes of this review, I would like to ask the question: what is the smallest size of free-standing molecule whose structure can in principle be determined by phase-contrast electron microscopy? Given what has already been demonstrated in published work, this reduces to the question: what is the smallest size of molecule for which it is possible to determine from images of unstained molecules the five*

parameters needed to define accurately its orientation (three parameters) and position (two parameters) so that averaging can be performed?

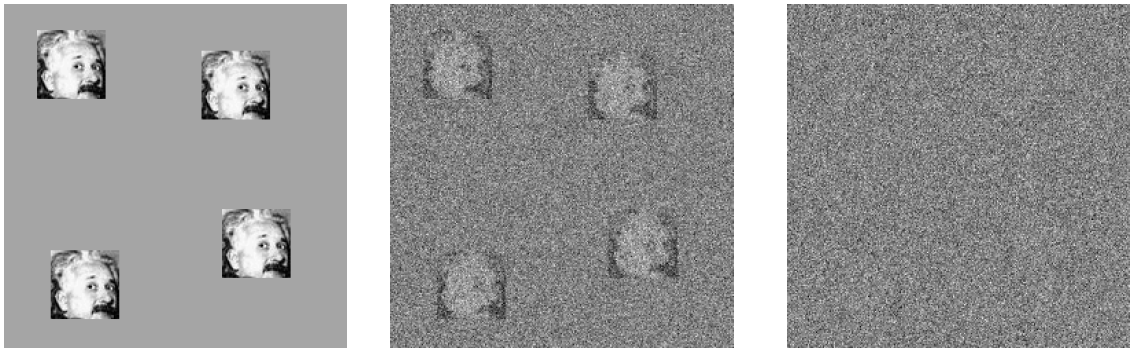
In that paper and in others that followed (e.g., [22]), it was established that particle picking is impossible for molecules below a certain weight. As a result, it is impossible to reconstruct such small molecules by any of the existing computational pipelines for single particle analysis in cryo–EM, as the particles themselves cannot be picked from the micrographs. This has motivated recent technical advances in the field, including the use of Volta phase plates [31, 35] and scaffolding cages [37]. Despite this progress, detecting small molecules in the micrographs (below ∼50 kDa) remains a challenge. We note that nuclear magnetic resonance (NMR) spectroscopy and X-ray crystallography are well suited to reconstruct small molecules. Yet, cryo–EM has a lot to offer even for molecules with already known structures obtained via NMR spectroscopy or X-ray crystallography, because these methods have limited ability to distinguish conformational variability. [Need a ref for this claim.]

In this paper, we argue that there is a gap between the two questions in the quoted excerpt above, and that one may be able to exploit it to design better reconstruction algorithms. Specifically, the impossibility of particle picking does not necessarily imply impossibility of particle reconstruction. Indeed, the aim is only to reconstruct the molecule: estimating the locations of the particles in the micrograph is merely a helpful intermediate stage when it can be done. Our main message is that the limits particle picking imposes on molecule size do not translate into limits on particle reconstruction.

As a proof of concept, we study a simplified model where an unknown image appears numerous times at unknown locations in several micrographs, each affected by additive Gaussian noise—see Figure 1 for an illustration. The goal is to estimate the planted image. The task is interesting in particular when the SNR is low enough that particle picking cannot be done reliably. A precise mathematical formulation of the model, including an extension where more than one planted images are to be recovered, is provided in Section 4. To be clear, we do not consider here many prominent features of real SPR experiments and do not aim to reconstruct any 3-D structure. Instead, we solve a simpler problem that we believe captures key elements of the SPR problem. We note that similar models emerge in spike sorting [34], passive radar [23] and system identification [38].

In order to recover the planted image, we use autocorrelation analysis. In a nutshell, we relate the autocorrelation functions of the micrographs to the autocorrelation functions of the planted image. For any noise level, these autocorrelations can be estimated to any desired accuracy, provided individual occurrences of the image are well separated and the image appears sufficiently many times in the micrographs. Importantly, there is no need to detect individual occurrences. The autocorrelations of the micrographs are straightforward to compute and require only one pass over the data. These directly yield estimates for the autocorrelations of the image. To estimate the image itself from its estimated autocorrelations, we solve a nonlinear inverse problem via least-squares or a phase retrieval algorithm; see Figure 2 for an illustration and Section 4 for details. As a side note, we mention that expectation-maximization (EM)—a popular framework in SPR—is intractable for this problem; see Appendix C for a discussion.

Another interesting feature of the described approach pertains to model bias, whose importance in cryo-EM was stressed by a number of authors [44, 47, 29, 46]. In the classical "Einstein from noise" experiment, multiple realizations of pure noise are aligned to a picture of Einstein using cross-correlation and then averaged. In [44], it was shown that the averaged



(a) $\sigma = 0$ (b) $\sigma = 0.5$ (c) $\sigma = 3$

Figure 1: Example of micrographs of size 250×250 with additive white Gaussian noise of variance σ^2 for increasing values of σ. Each micrograph contains the same four occurrences of a 50×50 image of Einstein. In panel (c), the noise level is such that it is very challenging to locate the occurrences of the planted image. In fact, it can be shown that at low SNR, reliable detection of individual image occurrences is impossible, even if the true image is known. By analogy to cryo-EM, this depicts a scenario where particle picking cannot be done.

noise rapidly becomes remarkably similar to the Einstein template. In the context of cryo–EM, this experiment exemplifies how prior assumptions about the particles may influence the reconstructed structure. This model bias is common to all particle picking methods based on template matching. In our approach, no templates are required, thus significantly reducing concerns about model bias.

2 Results

We conducted two experiments in the simplified image formation model described in the introduction:

1. The first experiment aims to recover a 2-D image from an increasing number of micrographs with high noise, similar to the rightmost panel of Figure 1. This is done using moments of second order, as these are sufficient to recover a 2-D image up to elementary symmetries;

2. The second experiment aims to recover three distinct 1-D signals from an increasing number of 1-D micrographs with high noise. For this task, it is necessary to use moments up to third order.

As outlined below, we find that it is indeed possible to recover accurate estimates of the ground truth signals from the highly corrupted micrographs, without particle picking. Furthermore, we find that the quality of estimation increases with the amount of data collected, despite the fact that particle picking remains challenging. The Methods section provides additional details. In the discussion section, we outline how the general approach could be extended to full 3-D SPR.

In the first experiment, we estimated Einstein's image of size 50×50 and mean zero from a growing number of micrographs, each of size 4096×4096 pixels. A micrograph contains,

on average, 700 occurrences of the target image at random locations. The latter are chosen so that two occurrences are always separated by at least 49 pixels. Thus, about 10% of each micrograph contains signal. The micrographs are contaminated with additive white Gaussian noise with standard deviation $\sigma = 3$ (this corresponds to $\mathsf{SNR} = 1/20$). This high noise level is illustrated in the right panel of Figure 1. In this first experiment, we assume knowledge of σ and of the total number of signal occurrences across all micrographs.

We compute the average autocorrelation of the micrographs (equivalently, the average of their power spectra). This is a particularly simple computation. In the methods section, we show how, owing to separation of the occurrences, a determined portion of the averaged autocorrelation allows to estimate the power spectrum of the unknown image itself. Mathematically, it is easy to show that the quality of this estimate improves steadily as the amount of data grows, regardless of noise level. Then, to estimate the target image, we resort to a standard phase retrieval algorithm called relaxed-reflect-reflect (RRR) [19]. RRR is initialized far away from the ground truth, and it iterates to produce the estimate, up to a reflection ambiguity.

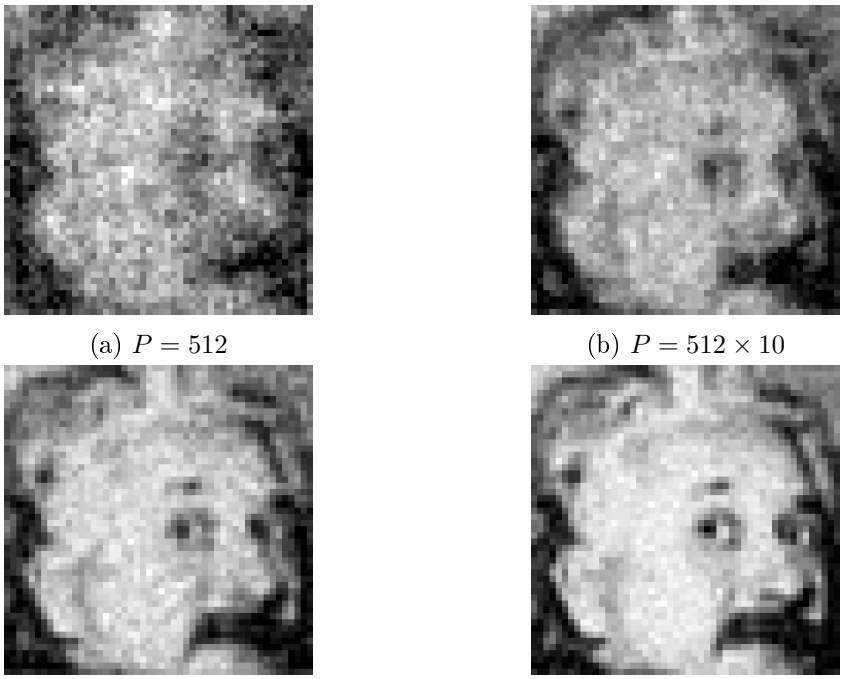
Figure 2 shows several estimated images for a growing number of micrographs, and a movie is available in [supplementary material]. Figure 3 presents the normalized recovery error as a function of the amount of data available. Error is measured as the ratio of the root mean square error (RMSE) to the norm of the ground truth (square root of the sum of squared pixel intensities.) This is computed after fixing elementary symmetries (see Methods.) As evidenced by these figures, the ground truth image can be estimated increasingly well from increasingly many micrographs, without particle picking.

In the second experiment, three 1-D signals, each of length $L = 21$, appear at random locations in one long 1-D signal, which we call a micrograph by analogy. Any two occurrences are separated by at least 20 entries. The signals appear respectively about 30, 20 and 10 million times in a micrograph of length 12.3 billion. The micrograph is then contaminated with additive white Gaussian noise. This results in an SNR of about 1/9, while about 10% of the micrograph contains signal. Neither the number of occurrences nor the noise level σ are known to the algorithm.

In the Methods section, we detail how autocorrelations of the micrograph can be used to estimate weighted averages of the autocorrelations of the target signals. The individual signals and their relative densities are then estimated from autocorrelations up to order three by solving a nonlinear least-squares problem.

Figure 4 shows how the estimates improve as we see a larger and larger fraction of the micrograph (that is, as more and more data becomes available.) As is clear from the picture, despite the high noise level which would make it very challenging to locate the individual signal occurrences, the signals can be estimated accurately given enough data. Furthermore, the [propensity[1]] of each signal can also be estimated.

[1]Might not be the right word; 'density' is not good because it might refer to the density of the particle for example, whereas here we mean to say the 'fraction of occurrences that come from a particular class'



(a) $P = 512$ (b) $P = 512 \times 10$

(c) $P = 512 \times 10^2$ (d) $P = 512 \times 10^3$

Figure 2: Recovery of Einstein from micrographs at noise level $\sigma = 3$ (see Figure 1(c)). Averaged autocorrelations of the micrographs allow to estimate the power spectrum of the target image. This does not require particle picking. A phase retrieval algorithm (RRR) produces the estimates here shown, initialized with an image of the physicist Isaac Newton. Estimates are obtained from P micrographs (growing across panels), each containing 700 image occurrences on average. [To add: initialization figure]

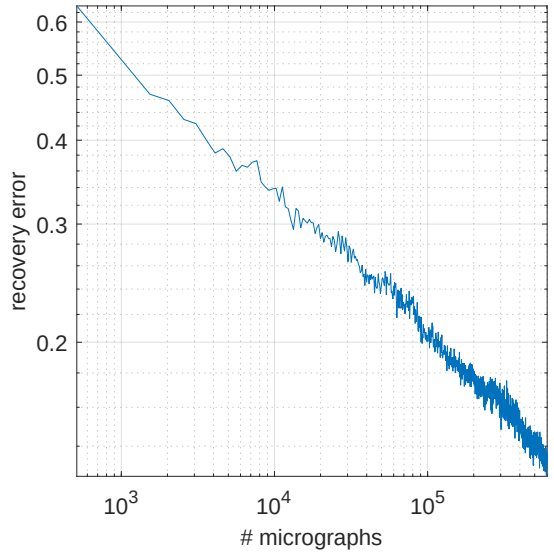


Figure 3: Relative root mean square error of the estimate of Einstein's image as a function of the number of observed micrographs (logarithmic scale along both axes.)

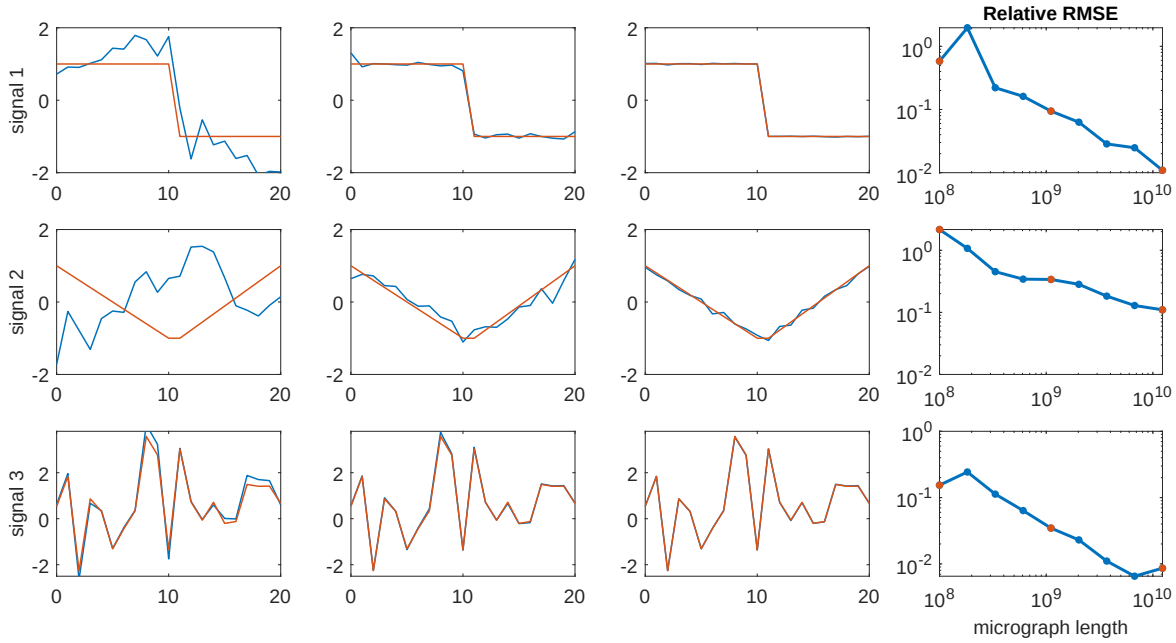


Figure 4: For the second experiment, each row shows, three times, one of the target signals (red), overlaid with an estimate (blue) obtained from a growing portion of the noisy micrograph (about 10^8, 10^9 and 10^{10} entries available to compute autocorrelations). The last column depicts evolution of the relative root mean square error in estimating each individual target signal. Signals 1 to 3 appear respectively about 30.0, 20.0 and 10.0 million times. With the whole micrograph available, the algorithm estimated those to be 29.8, 21.9 and 10.0 million, respectively.

3 Discussion

In the simplified model we examined, the aim is to estimate one, or possibly several, images from micrographs. Our strategy is to compute autocorrelation functions of the data and to relate these statistics to the unknown parameters. Recovering the parameters from the statistics reduces to solving a set of polynomial equations. Depending on the scenario, we did so using either a phase retrieval algorithm or a nonlinear least-squares algorithm.

The same general approach can, in principle, be applied directly to SPR from cryo-EM. Here, the micrographs contain numerous tomographic projections of molecules (possibly in different conformations) taken from unknown viewing directions. The aim is to estimate the 3-D volumes of the different conformations directly from micrographs. Each volume can be expanded linearly in a basis, so that the volume is characterized by its expansion coefficients. Since tomographic projection is a linear operation, autocorrelations of the micrographs (which can be estimated easily) are polynomial functions of the sought coefficients. Thus, autocorrelations of the micrographs provide a system of polynomial equations in the volume parameters, and the question becomes: are these equations sufficient to uniquely identify the volumes, and can we solve the system?

We show in Appendix F that the number of polynomial equations provided by the third-order autocorrelations is of the same order as the number of coefficients required to describe one volume [at resolution comparable to that of the particle projections—remove?]. This hints that it may be possible to reconstruct one or even several distinct volumes from these equations directly. Additional parameters in these equations could encode an unknown viewing direction distribution and an unknown conformation distribution, which could then also be estimated. Crucially, the outlined approach involves no particle picking, hence a fortiori no viewing direction estimation or conformation clustering. As a result, it may not be limited to large molecules in the same way that particle picking approaches are. Concerns for model bias would also greatly be reduced.

Of course, we recognize that significant challenges lay ahead for the implementation of the proposed approach to 3-D reconstruction directly from the micrographs. We discuss a few now.

One possible concern is that the numerical experiments conducted here suggest a large amount of data may be necessary.[2] Recent trends in high-throughput cryo-EM technology [?] give hope that this may be a lesser concern in the long term. Still, large amounts of data also imply large amounts of computations. On this front, we note that computing autocorrelations of low orders can be done efficiently on CPUs and GPUs, and in parallel across micrographs. It can even be done in streaming mode, as only one look at each micrograph is necessary. The output of this data processing stage is a succinct summary in the form of autocorrelation estimates: its size is a function of the resolution, not a function of the number of observed micrographs. Subsequent steps, which involve solving the system of polynomial equations, scale only in the size of that summary. Of course, an important question then is whether the equations can be solved meaningfully in practice. The proof-of-concept experiments above suggest they might.

Beyond data acquisition and computational challenges, there are modeling issues to con-

[2] Whether or not this large amount of data would be necessary for any method to succeed given the unfavorable SNR is an interesting research question.

sider. As stated, our approach relies on two core assumptions that are not necessarily verified in SPR experiments. First, we assume an additive white noise model, while in practice the noise may be structured or signal dependent. To address this point, it may be necessary to investigate better noise models and to extend the autocorrelation analysis accordingly. Second, we assume that any two signal occurrences are sufficiently separated, and we use this assumption to derive algebraic relations between autocorrelations of the micrographs and autocorrelations of the target signals. Perhaps this separation could be induced by careful experimental design [?]. Alternatively, if the signals are not well separated, one can introduce new parameters which encode the distribution of the spacing between occurrences. Here as well, relations between autocorrelation functions of the data and of the signals can be derived.

[Do we want to note here that EM used to be perceived as computationally out of reach back when it was proposed [cite Fred], yet is now the de facto standard method as exemplified by RELION? Do we also want to stress that models have been refined over decades? Both points are rather defensive in nature.]

[Should we discuss CTF? Where, and to what extent?]

[Where and how do we cite Kam? Fred?]

4 Methods

We derive algebraic relations between the autocorrelation functions of the micrographs and the autocorrelation functions of the target signals. For ease of exposition, we do so in the 1-D case. Extension to the 2-D case is straightforward. We also give additional technical details regarding the two experiments presented in Section 2.

Let $x_1, \dots, x_K \in \mathbb{R}^L$ be the sought signals and let $y \in \mathbb{R}^N$ be the observed micrograph (notice that it is equivalent to think of the data as being one long micrograph or multiple shorter micrographs concatenated into one.) The forward model (or "image" formation model) is as follows. For each target signal x_k, an unknown binary signal $s_k \in \{0,1\}^{N-L+1}$ indicates (with 1's) the starting positions of all occurrences of x_k in y, so that, with additive white Gaussian noise:

$$y = \sum_{k=1}^{K} x_k * s_k + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_N), \tag{4.1}$$

where $*$ denotes linear convolution. The binary signals obey the following property:

$$\text{If } s_k[i] = 1 \text{ and } s_{k'}[j] = 1 \text{ but } (k,i) \neq (k',j), \text{ then } |i-j| \geq 2L-1. \tag{4.2}$$

In words: the starting positions of any two occurrences (be it of the same signal or of two different signals) must be separated by at least $2L-1$ positions, so that their end points are necessarily separated by at least $L-1$ signal-free entries in the micrograph.

From y, we aim to recover $x_1, \dots x_K$, and possibly also the number of occurrences of each. In contrast, particle picking is the task of estimating the binary signal $s = s_1 + \cdots + s_K$, while clustering is the task of separating s into its components $s_1, \dots, s_K$. Neither of the latter can be performed reliably if σ is large (that is, at low SNR.)

Considering (4.1), one can think of this inverse problem as a mixture of blind deconvolution problems between binary signals and the target signals. Related literature is briefly surveyed in Appendix B.

Aperiodic autocorrelation functions In general, for a signal z of length m, the autocorrelation of order $q \geq 1$ is given for any integer shifts $\ell_1, \ldots, \ell_{q-1}$ by

$$a_z^q[\ell_1, \ldots, \ell_{k-1}] = \frac{1}{m} \sum_{i=-\infty}^{+\infty} z[i]z[i+\ell_1] \cdots z[i+\ell_{q-1}], \tag{4.3}$$

where indexing of z out of the range $0, \ldots, m-1$ is zero-padded. For our purposes, this will be applied both to x_k's (each of length L) and to y (of length N). Explicitly, the first-, second- and third-order autocorrelations are given by

$$\begin{aligned} a_z^1 &= \frac{1}{m} \sum_{i=0}^{m-1} z[i], \\ a_z^2[\ell] &= \frac{1}{m} \sum_{i=\max\{0,-\ell\}}^{m-1+\min\{0,-\ell\}} z[i]z[i+\ell], \\ a_z^3[\ell_1, \ell_2] &= \frac{1}{m} \sum_{i=\max\{0,-\ell_1,-\ell_2\}}^{m-1+\min\{0,-\ell_1,-\ell_2\}} z[i]z[i+\ell_1]z[i+\ell_2]. \end{aligned} \tag{4.4}$$

The autocorrelation functions have symmetries. Specifically, $a_z^2[\ell] = a_z^2[-\ell]$, and

$$a_z^3[\ell_1, \ell_2] = a_z^3[\ell_2, \ell_1] = a_z^3[-\ell_1, \ell_2 - \ell_1].$$

Estimating the autocorrelation function of a single signal. For the special case $K = 1$ where a single signal $x = x_1$ must be recovered, the relation between autocorrelations of the micrograph and those of x is particularly simple, so that we treat it first. It is useful to introduce some notation: let M denote the number of occurrences of x in y, and let

$$\gamma = \frac{ML}{N} \tag{4.5}$$

denote the density of x in y (that is, the fraction of entries of y occupied by occurrences of x.) The spacing constraint (4.2) imposes $\gamma \leq \frac{L}{2L-1} \approx 1/2$.

One simple observation is that the first-order autocorrelation of y (its mean) is independent of the locations of x. Since the noise is independent of the signal, the mathematical expectation of a_y^1 is easily seen to be:[3]

$$\mathbb{E}\{a_y^1\} = \gamma a_x^1.$$

We consider the asymptotic regime where $M, N \to \infty$, while γ remains constant (we see an increasingly large micrograph, containing increasingly many signal occurrences, with constant signal density.) In that regime, the law of large numbers gives meaning to the following statement:

$$\lim_{N \to \infty} a_y^1 = \gamma a_x^1.$$

[3]We did not fully specify a random generating model for the location vector s. The expectation is still well defined specifically because the quantity under consideration is independent of s under the assumptions.

Thus, given enough data, if γ is known, we can estimate a_x^1 from y. (We show later how to estimate γ as well.)

The spacing constraint (4.2) gives rise to more powerful observations. Consider the second-order autocorrelation in particular: $a_y^2[\ell]$ computes the correlation between y and a copy of y shifted by ℓ entries. Considering ℓ only in the range $0, \dots, L-1$, one can see that any given occurrence of x in y is only ever correlated with itself (with the same shift ℓ), and never with another occurrence. As a result,

$$\lim_{N\to\infty} a_y^2[\ell] = \gamma a_x^2[\ell] + \sigma^2 \delta[\ell]$$

for $\ell = 0, \dots, L-1$, where δ denotes the Kronecker delta function. The last part captures the autocorrelation of the noise. Notice that, even if σ is unknown, entries $\ell = 1, \dots, L-1$ still provide useful information about a_x^2. Along the same lines, one can establish a relation for third-order autocorrelations:

$$\lim_{N\to\infty} a_y^3[\ell_1, \ell_2] = \gamma a_x^3[\ell_1, \ell_2] + \sigma^2 \gamma a_x^1 \cdot \big(\delta[\ell_1, 0] + \delta[0, \ell_2] + \delta[\ell_1, \ell_2]\big), \tag{4.6}$$

for $\ell_1, \ell_2 = 0, \dots, L-1$. Here too, few entries are affected by σ in the limit. Detailed derivations for identities in this and the next part are given in Appendix A.

Estimating the autocorrelation function of multiple signals. Returning to the general case $K \geq 1$, let $M_1, \dots, M_K$ denote the number of occurrences of signals $x_1, \dots, x_K$ respectively, and define

$$\gamma_k = \frac{M_k L}{N}, \qquad\qquad \gamma = \sum_{k=1}^{K} \gamma_k. \tag{4.7}$$

As above, we consider the asymptotic regime where $M_1, \dots, M_K, N \to \infty$ while preserving the ratios γ_k constant. Still under the spacing constraint (4.2), similarly to the developments above, one can estimate a mixture of the autocorrelations of the K target signals from the autocorrelations of the micrograph:

$$\begin{aligned}
\lim_{N\to\infty} a_y^1 &= \sum_{k=1}^{K} \gamma_k a_{x_k}^1, \\
\lim_{N\to\infty} a_y^2[\ell] &= \sum_{k=1}^{K} \gamma_k a_{x_k}^2[\ell] + \sigma^2 \delta[\ell], \\
\lim_{N\to\infty} a_y^3[\ell_1, \ell_2] &= \sum_{k=1}^{K} \gamma_k a_{x_k}^3[\ell_1, \ell_2] + \sigma^2 \left(\sum_{k=1}^{K} \gamma_k a_{x_k}^1 \right) (\delta[\ell_1, 0] + \delta[0, \ell_2] + \delta[\ell_1, \ell_2]),
\end{aligned} \tag{4.8}$$

where $\ell, \ell_1, \ell_2 = 0, \dots, L-1$. The left hand side is straightforward to estimate from data: it provides a succinct summary of it. The right hand side involves polynomial functions of unknowns $\gamma_1, \dots, \gamma_K, x_1, \dots, x_K$, and possibly σ^2. The task is to solve these polynomial equations in a robust way.

Numerical experiment with three 1-D signals. For the 1-D experiment depicted in Figure 4, we fix $K = 3$ signals of length $L = 21$. Following the forward model described at the beginning of this section, we generate an observation y of length $12.3 \cdot 10^9$. Each of the three signals appears, respectively (and approximately), $30.0 \cdot 10^6$, $20.0 \cdot 10^6$ and $10.0 \cdot 10^6$ times in y for a total of exactly $60 \cdot 10^6$ occurrences, such that at least $L - 1$ zeros separate any two occurrences of any signals. This is done by randomly selecting $60 \cdot 10^6$ placements in y, one at a time with an accept/reject rule based on the separation constraint and locations picked so far. For each placement, one of the three signals is picked at random according to the proportions $1/2, 1/3, 1/6$. Then, i.i.d. Gaussian noise with mean zero and standard deviation $\sigma = 3$ is added, to form the observed y. The resulting SNR of y is about $1/9$.

This is enough noise to make cross-correlations of y even with the true signals display peaks at essentially random locations, uninformative of the actual locations of the signal occurrences. Thus, we contend that it would be difficult for any algorithm to locate the signal occurrences, let alone to classify them according to which signal appears where.

Given the observation y, we proceed to compute the autocorrelations. The first-order autocorrelation is straightforward. For second-order autocorrelations, notice from equation (4.8) that $a_y^2[\ell]$ suffers no noise-induced bias for ℓ in 1 to $L-1$. Thus, we omit $\ell = 0$, which has the practical effect that we need not know σ to make sense of the computed quantities. Likewise, for third-order autocorrelations, $a_y^3[\ell_1, \ell_2]$ for $0 \leq \ell_1, \ell_2 \leq L - 1$ such that $\ell_2 \leq \ell_1$ includes all relevant entries for our purpose (this accounts for symmetries), and we further exclude any such that ℓ_1, ℓ_2 or $\ell_1 - \ell_2$ are zero to avoid the need to estimate σ—there are $\frac{(L-1)(L-2)}{2}$ remaining entries. We have

$$1 + (L-1) + \frac{(L-1)(L-2)}{2} = \frac{1}{2}L(L-1) + 1$$

coefficients in total. Since we aim to estimate KL parameters (for the K signals of length L) plus K parameters (for the densities γ_k), an absolute upper bound on K (simply to ensure we have at least as many equations as we have unknowns) is

$$K(L+1) \leq \frac{1}{2}L(L-1) + 1.$$

Thus, $(L-1)/2$ (up to a small approximation) is an absolute upper limit on K. [We may want to cite MRA literature here.]

In practice, the autocorrelations are computed on disjoint segments of y of length $100 \cdot 10^6$ and added up, without correction for the junction points. Segments are handled sequentially on a GPU, as GPUs are particularly well suited to execute simple instructions across large vectors of data. If multiple GPUs are available, segments can of course be handled in parallel.

Having computed the moments of interest, we now estimate signals $x_1, \ldots, x_K$ and coefficients $\gamma_1, \ldots, \gamma_K$ which agree with the data. We choose to do so by running an optimization

algorithm on the following nonlinear least-squares problem:

$$\min_{\substack{\hat{x}_1,\ldots,\hat{x}_K\in\mathbb{R}^W\\ \hat{\gamma}_1,\ldots,\hat{\gamma}_K>0}} w_1\left(a_y^1-\sum_{k=1}^K\hat{\gamma}_k a_{\hat{x}_k}^1\right)^2+w_2\sum_{\ell=1}^{L-1}\left(a_y^2[\ell]-\sum_{k=1}^K\hat{\gamma}_k a_{\hat{x}_k}^2[\ell]\right)^2+ \\ w_3\sum_{\substack{2\le\ell_1\le L-1\\ 1\le\ell_2\le\ell_1-1}}\left(a_y^3[\ell_1,\ell_2]-\sum_{k=1}^K\hat{\gamma}_k a_{\hat{x}_k}^3[\ell_1,\ell_2]\right)^2, \quad (4.9)$$

where $W \geq L$ is the length of the sought signals and the weights are set to $w_1 = 1/2, w_2 = 1/2n_2, w_3 = 1/2n_3$, where n_2, n_3 are the number of moments used: $n_2 = L-1$, $n_3 = \frac{(L-1)(L-2)}{2}$ (weights could also be set in accordance with variance estimates as in [16]).

Setting $W = L$ (as is a priori desired) is problematic because the above optimization problems appears to have numerous poor local optimizers. Thus, we first run the optimization with $W = 2L - 1$. This problem appears to have few poor local optima, perhaps because the additional degrees of freedom allow for more escape directions. Since we hope the signals estimated this way correspond to the true signals zero-padded to length W, we extract from each one a subsignal of length L that has largest ℓ_2-norm. This estimator is then used as initial iterate for (4.9), this time with $W = L$. We find that this procedure is reliable for a wide range of experimental parameters. To solve (4.9), we run the trust-region method implemented in Manopt [15], which allows to treat the positivity constraints on coefficients $\hat{\gamma}_k$. Notice that the cost function is a polynomial in the variables, so that it is straightforward to compute it and its derivatives.

Numerical experiment with 2-D image. For the 2-D experiment shown in Figures 2 and 3, we generate P micrographs of size 4096×4096 pixels. In each micrograph, we place Einstein's image (of zero mean) of size 50×50 in random locations, while preserving the separation condition (4.2). This is done by randomly selecting 4000 placements in the micrograph, one at a time with an accept/reject rule based on the separation constraint and locations picked so far. On average, 700 images are placed in each micrograph. Then, i.i.d. Gaussian noise with standard deviation $\sigma = 3$ is added, inducing an SNR of approximately 1/20. An example of a micrograph's excerpt is presented in the right panel of Figure 1.

In this experiment, we assume we know the noise level σ and the total number of occurrences of the target image across all micrographs. In stark contrast with the 1-D setup, the second-order autocorrelation determines almost any target image uniquely, up to reflection through the origin [25] (see also [10] for a review). This is because the second-order autocorrelations correspond to the Fourier magnitudes of the signal through the 2-D Fourier transform. Therefore, we estimate the signal's Fourier magnitudes (or power spectrum) from the Fourier magnitudes of the micrographs, at the cost of one 2-D fast Fourier transform (FFT) per micrograph. These can be computed highly efficiently and in parallel.

To recover the target image from the estimated power spectrum, we use a standard phase retrieval algorithm called relaxed-reflect-reflect (RRR). This algorithm iterates the map

$$z \leftarrow z + \beta(P_2(2P_1(z) - z) - P_1(z))$$

on an image z of size $2L \times 2L$. We set the parameter β to 1. The map is designed so that, if the estimated power spectrum is exact, then fixed points contain Einstein's image in the

upper-left corner of size $L \times L$, possibly reflected through its origin, and zeros elsewhere. The operator $P_2(z)$ combines the Fourier phases of the current estimation z with the estimated Fourier magnitudes. The operator $P_1(z)$ zeros out all entries of z outside the $L \times L$ upper-left corner.

In order to compare the performance in multiple cases and at different noise levels, the algorithm is stopped after a fixed number of iterations (1000) and the iterate with the smallest error compared to the ground truth (up to the reflection ambiguity) is chosen as the solution. While this cannot be done in practice (since we do not have access to the ground truth to determine which iterate is best), this procedure enables us to compare a large number of instances in different noise environments. [Note the last two sentences!]

References

[1] Emmanuel Abbe, Tamir Bendory, William Leeb, João Pereira, Nir Sharon, and Amit Singer. Multireference alignment is easier with an aperiodic translation distribution. *arXiv preprint arXiv:1710.02793*, 2017.

[2] Karim Abed-Meraim, Wanzhi Qiu, and Yingbo Hua. Blind system identification. *Proceedings of the IEEE*, 85(8):1310–1322, 1997.

[3] Cecilia Aguerrebere, Mauricio Delbracio, Alberto Bartesaghi, and Guillermo Sapiro. Fundamental limits in multi-image alignment. *IEEE Transactions on Signal Processing*, 64(21):5707–5722, 2016.

[4] Christophe Andrieu, Éric Barat, and Arnaud Doucet. Bayesian deconvolution of noisy filtered point processes. *IEEE Transactions on Signal Processing*, 49(1):134–146, 2001.

[5] GR Ayers and J Christopher Dainty. Iterative blind deconvolution method and its applications. *Optics letters*, 13(7):547–549, 1988.

[6] Jean-Marc Azais, Yohann De Castro, and Fabrice Gamboa. Spike detection from inaccurate samplings. *Applied and Computational Harmonic Analysis*, 38(2):177–195, 2015.

[7] Heinz H. Bauschke, Patrick L. Combettes, and D.Russell Luke. Finding best approximation pairs relative to two closed convex sets in Hilbert spaces. *Journal of Approximation Theory*, 127(2):178–192, 2004.

[8] Robert Beinert and Gerlind Plonka. Ambiguities in one-dimensional discrete phase retrieval from fourier magnitudes. *Journal of Fourier Analysis and Applications*, 21(6):1169–1198, 2015.

[9] Tamir Bendory. Robust recovery of positive stream of pulses. *IEEE Transactions on Signal Processing*, 65(8):2114–2122, 2017.

[10] Tamir Bendory, Robert Beinert, and Yonina C Eldar. Fourier phase retrieval: Uniqueness and algorithms. In *Compressed Sensing and its Applications*, pages 55–91. Springer, 2017.

[11] Tamir Bendory, Nicolas Boumal, Chao Ma, Zhizhen Zhao, and Amit Singer. Bispectrum inversion with application to multireference alignment. *arXiv preprint arXiv:1705.00641*, 2017.

[12] Tamir Bendory, Shai Dekel, and Arie Feuer. Robust recovery of stream of pulses using convex optimization. *Journal of Mathematical Analysis and Applications*, 442(2):511–536, 2016.

[13] Albert Benveniste, Maurice Goursat, and Gabriel Ruget. Robust identification of a nonminimum phase system: Blind adjustment of a linear equalizer in data communications. *IEEE Transactions on Automatic Control*, 25(3):385–399, 1980.

[14] Brett Bernstein and Carlos Fernandez-Granda. Deconvolution of point sources: A sampling theorem and robustness guarantees. *arXiv preprint arXiv:1707.00808*, 2017.

[15] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15:1455–1459, 2014.

[16] Nicolas Boumal, Tamir Bendory, Roy R Lederman, and Amit Singer. Heterogeneous multireference alignment: a single pass approach. *arXiv preprint arXiv:1710.02590*, 2017.

[17] Olivier Cappé, Arnaud Doucet, Marc Lavielle, and Eric Moulines. Simulation-based methods for blind maximum-likelihood filter identification. *Signal processing*, 73(1-2):3–25, 1999.

[18] Quentin Denoyelle, Vincent Duval, and Gabriel Peyré. Support recovery for sparse super-resolution of positive measures. *Journal of Fourier Analysis and Applications*, 23(5):1153–1194, 2017.

[19] Veit Elser. Matrix product constraints by projection methods. *Journal of Global Optimization*, 68(2):329–355, 2017.

[20] Joachim Frank and Terence Wagenknecht. Automatic selection of molecular images from electron micrographs. *Ultramicroscopy*, 12(3):169–175, 1983.

[21] Georgios B Giannakis and Jerry M Mendel. Identification of nonminimum phase systems using higher order statistics. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3):360–377, 1989.

[22] Robert M Glaeser. Electron crystallography: present excitement, a nod to the past, anticipating the future. *Journal of structural biology*, 128(1):3–14, 1999.

[23] Sandeep Gogineni, Pawan Setlur, Muralidhar Rangaswamy, and Raj Rao Nadakuditi. Passive radar detection with noisy reference channel using principal subspace similarity. *IEEE Transactions on Aerospace and Electronic Systems*, 2017.

[24] George Harauz and Amelia Fong-Lochovsky. Automatic selection of macromolecules from electron micrographs by component labelling and symbolic processing. *Ultramicroscopy*, 31(4):333–344, 1989.

[25] MHMH Hayes. The reconstruction of a multidimensional sequence from the phase or magnitude of its fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 30(2):140–154, 1982.

[26] Monson H Hayes and James H McClellan. Reducible polynomials in more than one variable. *Proceedings of the IEEE*, 70(2):197–198, 1982.

[27] Ayelet Heimowitz, Amit Singer, et al. Apple picker: Automatic particle picking, a low-effort cryo-em framework. *arXiv preprint arXiv:1802.00469*, 2018.

[28] Richard Henderson. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Quarterly reviews of biophysics*, 28(2):171–193, 1995.

[29] Richard Henderson. Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. *Proceedings of the National Academy of Sciences*, 110(45):18037–18041, 2013.

[30] Stuart M Jefferies and Julian C Christou. Restoration of astronomical images by iterative blind deconvolution. *The Astrophysical Journal*, 415:862, 1993.

[31] Maryam Khoshouei, Mazdak Radjainia, Wolfgang Baumeister, and Radostin Danev. Cryo-EM structure of haemoglobin at 3.2 å determined with the Volta phase plate. *Nature communications*, 8:16099, 2017.

[32] John Kormylo and J Mendel. Identifiability of nonminimum phase linear stochastic systems. *IEEE transactions on automatic control*, 28(12):1081–1090, 1983.

[33] Robert Langlois, Jesper Pallesen, Jordan T Ash, Danny Nam Ho, John L Rubinstein, and Joachim Frank. Automated particle picking for low-contrast macromolecules in cryo-electron microscopy. *Journal of structural biology*, 186(1):1–7, 2014.

[34] Michael S Lewicki. A review of methods for spike sorting: the detection and classification of neural action potentials. *Network: Computation in Neural Systems*, 9(4):R53–R78, 1998.

[35] Yi-Lynn Liang, Maryam Khoshouei, Mazdak Radjainia, Yan Zhang, Alisa Glukhova, Jeffrey Tarrasch, David M Thal, Sebastian GB Furness, George Christopoulos, Thomas Coudrat, et al. Phase-plate cryo-EM structure of a class B GPCR–G-protein complex. *Nature*, 546(7656):118, 2017.

[36] KS Lii, M Rosenblatt, et al. Deconvolution and estimation of transfer function phase and coefficients for nongaussian linear processes. *The annals of statistics*, 10(4):1195–1208, 1982.

[37] Yuxi Liu, Shane Gonen, Tamir Gonen, and Todd O. Yeates. Near-atomic cryo-EM imaging of a small protein displayed on a designed scaffolding system. *Proceedings of the National Academy of Sciences*, 2018.

[38] Lennart Ljung. System identification. In *Signal analysis and prediction*, pages 163–173. Springer, 1998.

[39] Toshihiko Ogura and Chikara Sato. Automatic particle pickup method using a neural network has high accuracy by applying an initial weight derived from eigenimages: a new

reference free method for single-particle analysis. *Journal of structural biology*, 145(1-2):63–75, 2004.

[40] Amelia Perry, Jonathan Weed, Afonso Bandeira, Philippe Rigollet, and Amit Singer. The sample complexity of multi-reference alignment. *arXiv preprint arXiv:1707.00943*, 2017.

[41] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[42] Sjors HW Scheres. Semi-automated selection of cryo-em particles in relion-1.3. *Journal of structural biology*, 189(2):114–122, 2015.

[43] Ofir Shalvi and Ehud Weinstein. New criteria for blind deconvolution of nonminimum phase systems (channels). *IEEE Transactions on information theory*, 36(2):312–321, 1990.

[44] Maxim Shatsky, Richard J Hall, Steven E Brenner, and Robert M Glaeser. A method for the alignment of heterogeneous macromolecules from electron microscopy. *Journal of structural biology*, 166(1):67–78, 2009.

[45] Jitendra Tugnait. Identification of nonminimum phase linear stochastic systems. In *The 23rd IEEE Conference on Decision and Control*, number 23, pages 342–347, 1984.

[46] Marin van Heel. Finding trimeric HIV-1 envelope glycoproteins in random noise. *Proceedings of the National Academy of Sciences*, 110(45):E4175–E4177, 2013.

[47] Marin van Heel, Michael Schatz, and Elena Orlova. Correlation functions revisited. *Ultramicroscopy*, 46(1–4):307–316, 1992.

[48] NR Voss, CK Yoshioka, M Radermacher, CS Potter, and B Carragher. Dog picker and tiltpicker: software tools to facilitate particle selection in single particle electron microscopy. *Journal of structural biology*, 166(2):205–213, 2009.

[49] Alex Wein. *Statistical Estimation in the Presence of Group Actions*. PhD thesis, 2018.

[50] Yanan Zhu, Qi Ouyang, and Youdong Mao. A deep learning approach to single-particle recognition in cryo-electron microscopy. *arXiv preprint arXiv:1605.05543*, 2016.

A Proof of (4.8)

Throughout the proof, we consider the case of one signal $K = 1$. The extension to $K > 1$ is straightforward by averaging the contributions of all signal with appropriate weights; see [16].

We will let the number of instances of the signal M grow with N, and write $M = M_N$ to emphasize this. We assume M_N grows proportionally with N, and define:

$$\gamma = \lim_{N\to\infty} \frac{M_N L}{N} < 1. \tag{A.1}$$

We will assume that $M_N = \Omega(N)$, so that $\gamma > 0$. In the sequel, we will suppress the explicit dependence of M on N for notational convenience.

We start by considering the mean of the data:

$$a_y^1 = \frac{1}{N}\sum_{i=0}^{N-1} y[i] = \frac{1}{N/L}\sum_{j=0}^{M-1}\frac{1}{L}\sum_{i=0}^{L-1} x[i] + \underbrace{\frac{1}{N}\sum_{i=0}^{N-1}\varepsilon[i]}_{\text{noise term}} \xrightarrow{a.s.} \gamma a_x^1, \tag{A.2}$$

where the noise term converges to zero almost surely (a.s.) by the strong law of large numbers.

We proceed with the (second-order) autocorrelation for fixed $\ell \in [0, \ldots, L-1]$. We can compute:

$$\begin{aligned} a_y^2[\ell] &= \frac{1}{N}\sum_{i=0}^{N-1-\ell} y[i]y[i+\ell] \\ &= \underbrace{\frac{1}{N}\sum_{j=1}^{M}\sum_{i=0}^{L-\ell-1} x[i]x[i+\ell]}_{\text{signal term}} + \underbrace{\frac{1}{N}\sum_{i=0}^{N-1-\ell}\varepsilon[i]\varepsilon[i+\ell]}_{\text{noise term}} + \underbrace{\frac{[1,2?]}{N}\sum_{j=1}^{M}\sum_{i=0}^{L-1} x[i]\varepsilon[s_j+i+\ell]}_{\text{cross-term}}. \end{aligned} \tag{A.3}$$

The cross-terms are linear in the noise, and are easily shown to vanish almost surely in the limit $N \to \infty$, by the strong law of large numbers. As for the signal term, we break it into M different sums, each containing one copy of the signal. This gives:

$$\frac{1}{N}\sum_{j=1}^{M}\sum_{i=0}^{L-\ell-1} x[i]x[i+\ell] = \frac{ML}{N}\frac{1}{L}\sum_{i=0}^{L-\ell-1} x[i]x[i+\ell] \xrightarrow{N\to\infty} \gamma a_x^2[\ell]. \tag{A.4}$$

We next analyze the pure noise term. When $\ell \neq 0$, we can break the noise term into a sum of independent terms:

$$\frac{1}{N}\sum_{i=0}^{N-1-\ell}\varepsilon[i]\varepsilon[i+\ell] = \frac{1}{\ell}\sum_{i=0}^{\ell-1}\frac{1}{N/\ell}\sum_{j=0}^{N/\ell-1}\varepsilon[j\ell+i]\varepsilon[(j+1)\ell+i]. \tag{A.5}$$

Each sum $\frac{1}{N/\ell}\sum_{j=0}^{N/\ell-1}\varepsilon[j\ell+i]\varepsilon[(j+1)\ell+i]$ is an average of N/ℓ independent terms with expectation zero, and thus converges to zero almost surely as $N \to \infty$. If $\ell = 0$, then we have:

$$\frac{1}{N}\sum_{i=0}^{N-1}\varepsilon^2[i] \xrightarrow{a.s.} \sigma^2. \tag{A.6}$$

We now analyze the third-order autocorrelation. Let us fix $\ell_1 \geq \ell_2 \geq 0$. We have:

$$\begin{aligned}
a_y^3[\ell_1,\ell_2] &= \frac{1}{N}\sum_{i=0}^{N-1-\ell_1} y[i]y[i+\ell_1]y[i+\ell_2] \\
&= \underbrace{\frac{ML}{N}\frac{1}{M}\sum_{j=1}^{M}\frac{1}{L}\sum_{i=0}^{L-1-\ell_1} x[i]x[i+\ell_1]x[i+\ell_2]}_{(1)} + \underbrace{\frac{1}{N}\sum_{i=0}^{N-1-\ell_1}\varepsilon[i]\varepsilon[i+\ell_1]\varepsilon[i+\ell_2]}_{(2)} \\
&+ \underbrace{\frac{1}{N}\sum_{j=1}^{M}\sum_{i=0}^{L-1} x[i]\varepsilon[s_j+i+\ell_1]\varepsilon[s_j+i+\ell_2]}_{(3)} + \underbrace{\frac{1}{N}\sum_{j=1}^{M}\sum_{i=0}^{L-1}\varepsilon[s_j+i-\ell_1]x[i]\varepsilon[s_j+i+\ell_2-\ell_1]}_{(4)} \\
&+ \underbrace{\frac{1}{N}\sum_{j=1}^{M}\sum_{i=0}^{L-1}\varepsilon[s_j+i-\ell_2]\varepsilon[s_j+i+\ell_1-\ell_2]x[i]}_{(5)} + \underbrace{\frac{1}{N}\sum_{j=1}^{M}\sum_{i=0}^{L-\ell_1+\ell_2-1}\varepsilon[s_j+i]x[i+\ell_1-\ell_2]x[i]}_{(6)} \\
&+ \underbrace{\frac{1}{N}\sum_{j=1}^{M}\sum_{i=0}^{L-\ell_2-1} x[i]\varepsilon[s_j+i+\ell_1]x[s_j+i+\ell_2]}_{(7)} + \underbrace{\frac{1}{N}\sum_{j=1}^{M}\sum_{i=0}^{L-\ell_1-1} x[i]x[i+\ell_1]\varepsilon[s_j+i+\ell_2]}_{(8)}.
\end{aligned} \tag{A.7}$$

Terms (6), (7) and (8) are linear in ε, and can easily be shown to converge to 0 almost surely by the law of large numbers, by similar arguments as used previously. Term (1) converges to $\gamma a_x^3[\ell_1,\ell_2]$ almost surely, for the same reasons as (A.4). To deal with terms (2)–(5), we must distinguish between different values of ℓ_1 and ℓ_2.

Case 1: $0 < \ell_2 < \ell_1$. Here, all summands with elements of ε involve products of distinct entries, which have expected value 0. Consequently, the usual argument shows that terms (2)–(5) all converge to 0 almost surely as $N \to \infty$.

Case 2: $0 = \ell_2 < \ell_1$. Term (2) is an average of products of the form $\varepsilon[i]^2\varepsilon[i+\ell_1]$, which have mean zero; consequently, term (2) converges to 0 almost surely. The same argument as for Case 1 shows that (3) and (5) also converge to 0. For term (4), we write:

$$\begin{aligned}
&\frac{1}{N}\sum_{j=1}^{M}\sum_{i=0}^{L-1}\varepsilon[s_j+i-\ell_1]x[i]\varepsilon[s_j+i+\ell_2-\ell_1] \\
&= \frac{ML}{N}\frac{1}{L}\sum_{i=0}^{L-1} x[i]\frac{1}{M}\sum_{j=1}^{M}\varepsilon[s_j+i-\ell_1]^2 \xrightarrow{N\to\infty} \gamma\frac{1}{L}\sum_{i=0}^{L-1} x[i]\sigma^2 = \gamma a_x^1\sigma^2.
\end{aligned} \tag{A.8}$$

Case 3: $0 < \ell_2 = \ell_1$. An argument nearly identical to that for Case 2 shows that terms (2), (4) and (5) converge to 0, while term (3) converges to $\gamma a_x^1\sigma^2$.

Case 4: $0 = \ell_2 = \ell_1$. The same argument as for term (4) in Case 2 shows that terms (3), (4) and (5) all converge to $\gamma a_x^1\sigma^2$. Term (2) is an average of $\varepsilon[i]^3$, which is mean zero; consequently, it converges to 0.

This completes the proof of (4.8).

B Related work

System identification. For $K = 1$, our problem can be interpreted as a special case of the system identification problem. Similarly to (4.1), the system identification forward model takes the form $y = x * w + \varepsilon$, where x is the unknown signal (the "system"), w is an unknown, random, input sequence, and ε is an additive noise. The goal of this problem is to estimate x, usually referred to as "identifying the system." The question of identifiability of x under this observation model is addressed for certain Gaussian and non-Gaussian w in [13, 32]. In the special case where $w \in \{0,1\}^N$, satisfying the spacing requirement (4.2), we obtain our model in the case of a single signal ($K = 1$). The same observation model is used for blind deconvolution, a longstanding problem arising in a variety of engineering and scientific applications such as astronomy, communication, image deblurring, system identification and optics; see [30, 43, 5, 2], just to name a few.

Likelihood-based methods. Likelihood-based methods estimate x as the maximizer of some function $f(x|y)$, where f is derived from the likelihood function of x given the observed signal y. For example, f may be the likelihood itself, or a related function with a similar form (leading to the class of "quasi-likelihood" methods). If some prior is assumed on x, then $f(x|y)$ can be taken to be the posterior distribution of x given the data; this is the simplest form of Bayesian inference.

Optimizing the function $f(x|y)$ exactly is often intractable, and thus heuristic methods are used instead. One proposed technique is to use Markov Chain Monte Carlo (MCMC) [17]. Another paper considers parameterized models for multiple distinct signals, as in our framework ($K > 1$) [4]. Their proposed solution is an MCMC algorithm tailored for their specific parametrized problem.

In special cases, including the case where w is binary, expectation maximization (EM) has been used [17]. The EM method for discrete w is based upon a certain "forward-backward" procedure used in hidden Markov models [41]. However, the complexity of this procedure is nonlinear in N, and therefore its usage is limited for big data sets. Indeed, on each iteration of EM, a probability must be assigned to any feasible combination of positions for the current signal estimate in M locations on the grid $\{1, \ldots, N\}$. In total, even when excluding forbidden combinations due to the spacing constraint, there are $O(N^M)$ such combinations, and the problem becomes computationally intractable when M grows with N and N is large.

Because likelihood methods are computationally expensive, methods based on recovery from moments, which are akin to our method, have also been previously used for system identification. Methods based on the third- and fourth-order moments are described and analyzed in [36, 21, 45].

C Theory

The impossibility of detection in low SNR. If $K = 1$ and x is known, then the locations s_i can be estimated via linear programming in the high SNR regime [6, 18, 12, 9, 14]. However, in the low SNR regime, estimating the binary sparse signal s is impossible. To see this, suppose that an oracle provides us M windows of length $W > L$, each containing exactly one copy of x. Suppose too that the oracle provides us with x itself. That is to say, we get a series of

windows of length W, each one containing a signal x at an unknown location; and our only task is to estimate the locations. This is an easier problem than detecting the support of s. Nevertheless, even this simpler problem is impossible in the low SNR regime [3]. Consequently, detecting the nonzero values of s is impossible in low SNR.

[NB: This is an important point for us. As stated, it is imprecise. I like the reduction idea to M windows, but then two things: (1) if we cite Aguerrebere, we need to be specific what theorem we reference and we need to check the assumptions, and (b) in any case, we need to be precise regarding what "impossible" means – I'm guessing the result is a bit more subtle, saying something like: the probability of type I / type II error is at least this much. Also, someone might oppose that we don't need to get them *all* right, and we don't need to get them *exactly* right.. To address such concerns, we need to be specific. After reading the Aguerberre paper quite carefully, one of the important points in there is that they assume a continuous model, so that shifts are continuous (that's why they can do CRBs – CRBs are not defined for discrete problems; of course, you could define them with some work (I assume), but that's not what they did, so just referencing them doesn't cut it.]

[NB: Here is a suggestion: we could further simplify the problem by having the oracle also state, for each window, that the shift is either of two possibly shifts (with 50/50 prior.) Then, knowing all of this and also knowing x, deciding which shift it is is a matter of deciding which Gaussian a vector came from, where both Gaussians have $\sigma^2 I$ variance and known mean. So, the distance test (which is equivalent to the likelihood ratio test) is optimal (that ought to be a Neyman-Pearson result). The probability of mistake is then at least $\frac{1}{2}\operatorname{erfc}\left(\frac{\|x_{shift1}-x_{shift2}\|}{2\sigma}\right)$, where x_{shift1} is a signal of the length of the given window with x placed at a certain position, and zeros elsewhere. We can then argue that (if there is a positive lower bound on the norms up there), there always exists a sigma such that the probability of mistake is this incredibly simpler problem is still arbitrarily close to 1/2 (random chance.) having said all of that, we can then also reference Aguereberre (carefully) and make the point that, in their specific setup, they show getting a lot of images indeed helps, but not beyond a certain error level which has to do with sigma, so that for large sigma, however much data we get, we can't estimate shifts (in their setting) accurately.]

Estimating a signal from its third-order autocorrelation. A one-dimensional signal is determined uniquely by its second- and third-order autocorrelations. Indeed, since $z[0]$ and $z[L-1]$ are non-zero by definition, we have the formula:

$$z[k] = \frac{z[0]z[k]z[L-1]}{z[0]z[L-1]} = \frac{a_z^3[k, L-1]}{a_z^2[L-1]}. \tag{C.1}$$

In particular, we have proven the following proposition:

Proposition C.1. *Let $z \in \mathbb{R}^L$ and suppose that $z[0]$ and $z[L-1]$ are nonzero. Then z is determined uniquely from a_z^2 and a_z^3.*

Some remarks are in order. First, formula (C.1) is not numerically stable if $z[0]$ and/or $z[L-1]$ are close to 0. In practice, we recover z by fitting it to its autocorrelations using a nonconvex least-squares procedures, which is empirically more robust to additive noise; we have seen similar phenomena for related problems [11, 16].

Second, if the spacing condition (4.2) holds, then the length of the signal can be determined from the autocorrelations. In particular, if (4.2) holds for some spacing $W \geq L$, then $a_z^2[i] = 0$ for all $i > L - 1$.

Note too that the second-order autocorrelation is not by itself sufficient to determine the signal uniquely [8, 10]. However, for dimensions greater than one, almost all signals are determined uniquely up to sign (phase for the complex signals) and reflection through the origin (with conjugation in the complex case) [25, 26]. The sign ambiguity can be resolved by the mean of the signal if it is not zero. However, determining the reflection symmetry still requires additional information, beyond the second-order autocorrelation.

Identifiability of parameters from the moments of y. The observed moments a_y^1, a_y^2 and a_y^3 of y do not immediately give the moments of the signal x, as seen by formula (4.6); rather, the two are related by the noise level σ and the ratio $\gamma = \lim_{N\to\infty} ML/N$, where $M = M_N$ grows with N. We will show, however, that x is still identifiable from the observed moments of y. In general, we say a parameter is "identifiable" if its value is uniquely determined in the limit $N \to \infty$.

First, we observe that if the noise level σ is known, one can estimate γ from the first two moments of the observed vector y.

Proposition C.2. *Let $K = 1$ and $\sigma > 0$ be fixed. If the mean of x is nonzero, then*

$$\gamma = \lim_{N\to\infty} \frac{(a_y^1)^2}{\sum_{j=0}^{L-1} a_y^2[j] - \sigma^2} \quad a.s.$$

Proof. The proof follows from plugging the explicit expressions of (4.6) into the right hand side of the equality. □

Using third-order autocorrelation information of y, both the ratio γ and the noise σ are identifiable. For the following results, when we say that a result holds for a "generic" signal x, we mean that it holds for all x inside a set $\Omega \subset \mathbb{R}^L$, whose complement $\mathbb{R}^L \setminus \Omega$ has Lebesgue measure zero.

Proposition C.3. *Let $K = 1$, and $\sigma > 0$ be fixed. Then, a_y^1, a_y^2 and a_y^3 determine the ratio γ and noise level σ uniquely for a generic signal x. If $\gamma \geq \frac{1}{4L(L-1)}$, then this holds for any signal x with nonzero mean.*

Proof. See Appendix D. □

From Propositions C.1 and C.3 we can directly deduce the following:

Corollary C.4. *Let $K = 1$ and $\sigma > 0$ be fixed. Then the signal x, the ratio γ, and the noise level σ are identifiable from the first three autocorrelation functions of y if:*

- *Either the signal x is generic; or*
- *Both $x[0]$ and $x[L-1]$ are nonzero, x has nonzero mean, and $\gamma \geq \frac{1}{4L(L-1)}$.*

Open theoretical questions. Our method of estimating x uses the third-order moments of the observations. These empirical moments are used to obtain consistent estimators of population parameters related to the the mean and second- and third-order autocorrelations of x, to which we fit the signal x. Consequently, the number of signal occurrences M should grow at least as fast as $1/\mathsf{SNR}^3$ to achieve a constant estimation error in the low SNR regime. In the related problem of multireference alignment [40, 1], this is optimal in the low SNR regime; we conjecture that the same is true for our problem.

Another interesting question is how many signals $x_1, \dots, x_K$ can be demixed from their mixed autocorrelation functions. In [16], it was empirically observed that $K \sim \sqrt{L}$ signals can be estimated simultaneously from their mixed second- and third-order autocorrelations, using the least-squares procedure. In [49] [TKTK: add reference to Alex Wein's thesis, or put personal correspondence], this result is shown theoretically for a different, and much less efficient, algorithm. In our current setting, the additional parameters γ and σ make the problem more challenging; however, we conjecture that the number of estimable signals still grows like $\sqrt{L}$.

D Proof of Proposition C.3

We will prove that both σ and γ are identifiable from the observed first three moments of y. For convenience, we will work with $\beta = \gamma/L$ rather than γ itself. We will construct two quadratic equations satisfied by β from observed quantities, independent of σ. Then, we will show that these equations are independent, and hence that β is uniquely defined. Given β, we can estimate σ using Proposition C.2.

Throughout the proof, it is important to distinguish between observed and unobserved values. We denote the observed values by E_i or a_y^1, a_y^2, a_y^3, while using F_i for functions of the signal's autocorrelations.

Recall that $a_y^1 = \beta(\mathbf{1}^T x)$ and and $a_y^2[0] = \beta\|x\|^2 + \sigma^2$, where $\mathbf{1} \in \mathbb{R}^L$ stands for vector of ones. Taking the product:

$$\begin{aligned} E_1 := a_y^1 a_y^2[0] &= (\beta(\mathbf{1}^T x))(\beta\|x\|^2 + \sigma^2) \\ &= \sigma^2 a_y^1 + \beta^2 F_1, \end{aligned} \tag{D.1}$$

where $F_1 := a_x^3[0,0] + \sum_{j=1}^{L-1}(a_x^3[j,j] + a_x^3[0,j])$. The terms of F_1 can be also estimated from a_y^3, while taking the scaling and bias terms into account:

$$E_2 := \beta F_1 + (2L+1)\sigma^2 a_y^1. \tag{D.2}$$

Therefore, from (D.1) and (D.2) we get

$$E_2\beta - (2L+1)\sigma^2\beta a_y^1 = E_1 - \sigma^2 a_y^1. \tag{D.3}$$

Let $a_y^2 := \sum_{j=0}^{L-1} a_y^2[j]$ and recall from Proposition C.2:

$$\sigma^2 = a_y^2 - (a_y^1)^2/(\beta L). \tag{D.4}$$

Plugging into (D.3) and rearranging we get

$$\mathcal{A}\beta^2 + \mathcal{B}\beta + \mathcal{C} = 0, \tag{D.5}$$

where

$$\begin{aligned}\mathcal{A} &= E_2 - (2L+1)a_y^1 a_y^2,\\ \mathcal{B} &= -E_1 + \frac{2L+1}{L}(a_y^1)^3 + a_y^1 a_y^2,\\ \mathcal{C} &= -(a_y^1)^3/L.\end{aligned}$$

Importantly, these coefficients are observable quantities.

We are now proceeding to derive the second quadratic equation. We notice that

$$E_3 = \frac{1}{L}(a_y^1)^3 = \frac{1}{L}\beta^3(\mathbf{1}^T x)^3 = \frac{1}{L}\beta^3 F_2, \tag{D.6}$$

where

$$F_2 = a_x^3[0,0] + 3\sum_{j=1}^{L-1} a_x^3[j,j] + 3\sum_{j=1}^{L-1} a_x^3[0,j] + 6\sum_{1\le i<j\le L-1} a_x^3[i,j].$$

On the other hand, from a_y^3 we can directly estimate F_2 up to scale and bias

$$E_4 = \beta F_2 + (6L-3)\sigma^2 a_y^1. \tag{D.7}$$

Taking the ratio:

$$\frac{E_4}{E_3} = \frac{L}{\beta^2} + \frac{(6L-3)L\sigma^2 a_y^1}{E_3},$$

we conclude:

$$\sigma^2 = \frac{E_4}{a_y^1 L(6L-3)} - \frac{E_3}{\beta^2 a_y^1(6L-3)}.$$

Using (D.4) and rearranging we get the second quadratic:

$$\mathcal{D}\beta^2 + \mathcal{E}\beta + \mathcal{F} = 0, \tag{D.8}$$

where

$$\begin{aligned}\mathcal{D} &= a_y^2 - \frac{E_4}{a_y^1 L(6L-3)},\\ \mathcal{E} &= -(a_y^1)^2/L,\\ \mathcal{F} &= \frac{E_3}{a_y^1(6L-3)}.\end{aligned}$$

To complete the proof, we need to show that the two quadratic equations (D.5) and (D.8) are independent. To this end, it is enough to show that the ratio between the coefficients is not the same. From (D.5) and (D.1), we have

$$\begin{aligned}\frac{\mathcal{B}}{\mathcal{C}} &= \frac{LE_1 - (2L+1)(a_y^1)^3 - La_y^1 a_y^2}{(a_y^1)^3}\\ &= \frac{La_y^2[0] - (2L+1)(a_y^1)^2 - La_y^2}{(a_y^1)^2}.\end{aligned}$$

In addition, using (D.6)

$$\frac{\mathcal{E}}{\mathcal{F}} = \frac{(3-6L)(a_y^1)^3}{LE_3} = 3-6L.$$

Now, suppose that the quadratics are dependent. Then, $\frac{\mathcal{B}}{\mathcal{C}} = \frac{\mathcal{E}}{\mathcal{F}}$, or,

$$La_y^2[0] - (2L+1)(a_y^1)^2 - La_y^2 = (a_y^1)^2(3-6L)$$

Rearranging the equation and writing in terms of x we get

$$4(L-1)\beta(a_x^1)^2 - \sum_{i=1}^{L-1} a_x^2[i] = 0. \tag{D.9}$$

For generic x, this polynomial equation is not satisfied. Therefore, the equations are independent. More than that, for any nonzero x, $(a_x^1)^2 > \sum_{i=1}^{L-1} a_x^2[i]$. Therefore, if $4(L-1)\beta \geq 1$, or,

$$\beta \geq \frac{1}{4(L-1)},$$

the condition (D.9) cannot be satisfied for any signal.

E Stuff

[Might go to related work section]

Many automatic and semi-automatic methods for particle picking have been proposed, based on edge detection, template matching and deep learning; see for instance [24, 39, 50, 20, 42, 27]. However, most of these procedures are prone to *model bias*. For instance, in the popular framework of RELION [42], the user manually marks hundreds of spots on the micrograph, believed to contain projections. Therefore, the algorithm's performance depends on the prior assumptions of the users about the particle's structure; the same holds true for deep learning based approaches which require constructing labeled sets of data. Other methods use disks or differences of Gaussians as templates [33, 48]. Nowadays, it also still popular to pick particles manually. This method, while it exploits the researcher's experience, is both tedious and subject to model bias.

F The autocorrelation functions in cryo–EM

[The most up to date version is the one from the patent – to update here too.]

The 3-D Fourier transform of an L-bandlimited 3-D volume (e.g., particle) can be expanded by spherical harmonics:

$$\hat{V}(k,\theta,\phi) = \sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} A_{\ell,m}(k) Y_\ell^m(\theta,\phi), \tag{F.1}$$

where (θ,ϕ) are two angles on the sphere, k is the radial coordinate, $Y_\ell^m(\theta,\phi)$ is the spherical harmonic of degree ℓ and order m and $A_{\ell,m}(k)$ are the associated spherical harmonics coefficients, to be estimated. A rotation of the volume by $\omega \in SO(3)$ can be described using the

Wigner D-function $D^{\ell}_{m,m'}$:

$$\begin{aligned}(R_\omega \hat{V})(k,\theta,\phi) &= \sum_{\ell=0}^{L}\sum_{m=-\ell}^{\ell} A_{\ell,m}(k)(R_\omega Y_\ell^m)(\theta,\phi) \\ &= \sum_{\ell=0}^{L}\sum_{m=-\ell}^{\ell} A_{\ell,m}(k)\sum_{m'=-\ell}^{m=\ell} D^{\ell}_{m,m'}(\omega)Y_\ell^{m'}(\theta,\phi).\end{aligned} \tag{F.2}$$

By the Fourier slice theorem, each cryo–EM measurement (projection) is equivalent (through 2-D Fourier transform) to the slice of $\hat{V}$, associated with $\theta = \pi/2$, after $\hat{V}$ was rotated by $\omega \in SO(3)$. Explicitly, the Fourier transform of a projection from the viewing direction ω is related to the spherical harmonic coefficients of the object through:

$$P_\omega(k,\phi) = \sum_{\ell=0}^{L}\sum_{m=-\ell}^{\ell} A_{\ell,m}(k)\sum_{m'=-\ell}^{m=\ell} D^{\ell}_{m,m'}(\omega)Y_\ell^{m'}(\pi/2,\phi). \tag{F.3}$$

Next, we want to relate the projections P_ω with the autocorrelation functions computed from the micrograph. We assume the projections are sufficiently separated so that the (Δ_x, Δ_y) entry of the second-order autocorrelation of the micrograph is proportional to:

$$M_2(\Delta_x,\Delta_y) \propto \sum_{n=1}^{N}\sum_{x,y} P_n(x,y)P_n(x+\Delta_x, y+\Delta_y) + \text{bias}, \tag{F.4}$$

where P_n denotes the nth projection. The assumption here is that (Δ_x, Δ_y) are small enough so that, in computing the auto-correlation, points (x, y) and $(x + \Delta_x, y + \Delta_y)$ do not touch distinct particles. By taking $n \to \infty$ and assuming uniform distribution of viewing directions, we get

$$M_2(\Delta_x,\Delta_y) \propto \sum_{x,y}\int_\omega P_\omega(x,y)P_\omega(x+\Delta_x, y+\Delta_y)d\omega. \tag{F.5}$$

In the same way and under the same conditions, the third moment is given by

$$M_3(\Delta_x^1,\Delta_y^1;\Delta_x^2,\Delta_y^2) \propto \sum_{x,y}\int_\omega P_\omega(x,y)P_\omega(x+\Delta_x^1, y+\Delta_y^1)P_\omega(x+\Delta_x^2, y+\Delta_y^2)d\omega + \text{bias}. \tag{F.6}$$

In order to determine the particle, by (F.1) one needs to estimate on the order of L^3 spherical harmonics coefficients. If the pixel size is proportional to $1/L$ (to match the volume's resolution), then M_3 provides on the order of L^4 equations involving triple products of P_ω. Since P_ω depends (after coordinate transformation) linearly in the spherical harmonic coefficients through (F.3), this means we have a system of $\sim L^4$ cubic equations in the $\sim L^3$ sought parameters. Importantly, the coefficients of these equations can be estimated from the micrographs directly, without particle picking stage.