Multi-target detection with application to cryo-electron microscopy

Tamir Bendory[a], Nicolas Boumal[b], William Leeb[c], Eitan Levin[a,b], and Amit Singer[a,b]

[a]The Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ, USA
[b]Department of Mathematics, Princeton University, Princeton, NJ, USA
[c]School of Mathematics, University of Minnesota, Minneapolis, MN, USA

June 17, 2019

Abstract

We consider the multi-target detection problem of recovering a set of signals that appear multiple times at unknown locations in a noisy measurement. In the low noise regime, one can estimate the signals by first detecting occurrences, then clustering and averaging them. In the high noise regime, however, neither detection nor clustering can be performed reliably, so that strategies along these lines are destined to fail. Notwithstanding, using autocorrelation analysis, we show that the impossibility of detecting and clustering signal occurrences in the presence of high noise does not necessarily preclude signal estimation. Specifically, to estimate the signals, we derive simple relations between the autocorrelations of the observation and those of the signals. These autocorrelations can be estimated accurately at any noise level given a sufficiently long measurement. To recover the signals from the observed autocorrelations, we solve a set of polynomial equations through nonlinear least-squares. We provide analysis regarding well-posedness of the task, and demonstrate numerically the effectiveness of the method in a variety of settings.

The main goal of this work is to provide theoretical and numerical support for a recently proposed framework to image 3-D structures of biological macromolecules using cryo-electron microscopy in extreme noise levels.

1 Introduction

We consider the *multi-target detection* problem of recovering a set of K signals that appear multiple times at unknown locations in a noisy measurement. Let $x_1, \ldots, x_K \in \mathbb{R}^L$

be the sought signals and let $y \in \mathbb{R}^N$ be the observed data, where we assume N is much larger than L. Let $s[i]$ count the number of signal occurrences whose first entry is positioned at $y[i]$. Each of those $s[i]$ signals is chosen among $x_1, \dots, x_K$ according to some (possibly unknown) distribution over $\{1, \dots, K\}$. If signal occurrences overlap, they interfere additively. With additive white Gaussian noise, the measurement model can be written as

$$y = \sum_{k=1}^{K} s_k * x_k + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_N), \tag{1.1}$$

where $*$ denotes linear convolution, and $s_k[i]$ indicates the number of occurrences of x_k starting at $y[i]$, so that $s = s_1 + \cdots + s_K$. Explicitly, with zero-based indexing,

$$y[i] = \sum_{k=1}^{K} \sum_{j=0}^{L-1} s_k[i-j] x_k[j] + \varepsilon[i].$$

The goal is to estimate $x_1, \dots, x_K$ from y. In parts of the paper, we focus on the case $K = 1$, called the *homogeneous* case; the case $K \geq 2$ is called *heterogeneous*. This idealized setup appears in several scientific applications, including structural biology [13] (as we detail below), spike sorting [32], passive radar [20], and system identification [37].

In the low noise regime (small σ), a valid strategy is to first detect the signal occurrences in y (that is, estimate s), cluster them (that is, separate s into $s_1, \dots, s_K$), and solve standard deconvolution problems. Crucially, we focus on the high noise regime, where *reliable detection of signal occurrences is impossible* [13, 3]. This limitation does not, however, preclude estimation of the signals $x_1, \dots, x_K$, as we show in this paper. In this setting, we refer to $s_1, \dots, s_K$ as *nuisance variables*: knowing them would certainly help, but we do not aim to estimate them.

In order to recover the signals in the high noise regime, we use autocorrelation analysis. At any noise level, the autocorrelations of the observation can be estimated to any desired accuracy for sufficiently large N. This computation is straightforward and requires only one pass over the data. The underlying principle is to relate the autocorrelations of the observation y to the autocorrelations of $x_1, \dots, x_K$.

Below we describe two generative models for s. In these models, the relationship between the autocorrelations of y and those of $x_1, \dots, x_K$ depends on $s_1, \dots, s_K$ only through their expected sums, that is, the expected total number of occurrences of each signal. To estimate the signals and occurrence counts from the computed autocorrelations, we solve a nonlinear least-squares problem as explained in Section 5.

The multi-target detection problem is an instance of *blind deconvolution*—a long-standing problem arising in a variety of engineering and scientific applications, such as astronomy, communication, image deblurring, system identification and optics; see [24, 41, 5, 2], to name a few. Different variants of the blind deconvolution problem have been thoroughly analyzed recently [4, 34, 33, 30, 35, 27]. In clear contrast to multi-target detection, these works focus on the low noise regime and aim to estimate both unknown signals (in our setting, this means estimating both x_k's and s_k's).

Models for target distribution

We consider two models for the distribution of signal occurrences in the observation, that is, for $s_1, \dots, s_K$.

The well-separated model. As a first setup, we allow any generative model for s which meets the following separation requirement: s is binary, and

$$\text{If } s[i] = 1 \text{ and } s[j] = 1 \text{ for } i \neq j, \text{ then } |i - j| \geq 2L - 1. \tag{1.2}$$

In words: the starting positions of any two occurrences must be separated by at least $2L - 1$ positions, so that their end points are necessarily separated by at least $L - 1$ signal-free (but still noisy) entries in the data. Furthermore, we require that the last signal occurrence in y is also followed by at least $L-1$ signal-free entries. This property ensures that correlating y with versions of itself shifted by at most $L-1$ entries does not involve correlating distinct signal occurrences. Once s is determined, for each position i such that $s[i] = 1$, one of the signals x_k is selected independently at random, and accordingly we set $s_k[i] = 1$. As a result, the only properties of $s_1, \dots, s_K$ that affect the autocorrelations of y (for shifts up to $L - 1$) are the total number of occurrences of the distinct signals: their individual and relative locations do not intervene. We detail this in Section 3.

The Poisson model. If the separation condition is violated, more knowledge about the location distribution is necessary to disentangle the autocorrelations of y. To that effect, we analyze a Poisson generative model.

Specifically, for each position i, the number $s[i]$ of signal occurrences starting at that position is drawn independently from a Poisson distribution with parameter γ/L, that is, $s[i] \overset{i.i.d.}{\sim} \text{Poisson}(\gamma/L)$ for some parameter $\gamma > 0$. Then, $s[i]$ is split into $s_1[i] + \cdots + s_K[i]$ by selecting $s[i]$ signals among $x_1, \dots, x_K$, independently at random following a fixed distribution over $\{1, \dots, K\}$. It is possible for more than one occurrence of the same x_k to start at position i. As for the well-separated model, the autocorrelations of y under this model depend weakly on s and $s_1, \dots, s_K$, essentially through the parameter γ: see Section 3.

Extensions

Extending the problem setup and autocorrelation analysis to signals in more than one dimension is straightforward: see the discussion in Section 4.4 and numerical experiments in Section 5.

Likewise, it is easy to extend the model to situations where the signal occurrences are sampled from a general distribution rather than from a finite set of choices $x_1, \dots, x_K$. The finite setup corresponds to the following distribution for signal occurrences:

$$x \sim \sum_{k=1}^{K} \pi_k \delta(x - x_k), \tag{1.3}$$

where $\delta(x - x_k)$ is a Dirac delta (a point mass) located at x_k, and $(\pi_1, \ldots, \pi_K)$ encodes the discrete distribution over $\{1, \ldots, K\}$. In the generalized setup, the goal is to estimate the distribution (possibly defined by a finite set of parameters). In particular, this allows for continuous distributions of targets. We adopt this perspective when deriving the autocorrelations in Section 3.

In the next section, we show how this flexibility allows us to model an important imaging problem in structural biology.

2 Connection with single-particle reconstruction via cryo-electron microscopy

Cryo-electron microscopy (cryo-EM) has recently joined X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy as a high-resolution structural method for biological macromolecules [18, 26, 10]. In a cryo–EM experiment, biological samples (e.g., macromolecules, viruses) are rapidly frozen in a thin layer of vitreous ice. The microscope produces 2-D tomographic images of the samples embedded in the ice, called *micrographs*. Each micrograph contains multiple tomographic projections of the samples at unknown locations and under unknown viewing directions. Importantly, the electron dose must be kept low to mitigate radiation damage, inducing high noise levels. The goal is to reconstruct 3-D models of the molecular structures from the micrographs. Since cryo-EM produces images of individual particles, it can elucidate multiple structures simultaneously. This is in clear contrast to X-ray and NMR, which aggregate information from an ensembles of particles.

Considering the extensions described in the previous section, we can phrase a simplified generative model for micrographs (the observation y) within our framework. Specifically, locations are chosen in the 2-D plane of the image, corresponding to s: this is where the molecules are fixed in the plane of the ice layer. At each selected location, a 2-D tomographic projection of a molecule (a signal occurrence) is added in the observation y. This signal is drawn from a probability distribution described by a discrete number of parameters which correspond to the sought 3-D structure, as follows.

The 3-D structure V (the target parameter) can be expanded into a linear combination of basis functions (for example, spherical harmonics for the spherical part and Bessel functions for the radial part): the coefficients of this expansion are the unknowns. Then, a rotation R_ω is applied to the volume (to model viewing directions) according to a (possibly unknown) distribution of ω over $SO(3)$ (the group of 3-D rotations). With tomographic projection denoted by P, x is a random signal, related to the distribution of ω through:

$$x = P(R_\omega V). \tag{2.1}$$

By further allowing V itself to be a random signal as well, this model can also encode a mixture of structures in the biological sample. This mixture might also be continuous, corresponding to continuous conformational variability.

All contemporary methods for single particle reconstruction using cryo-EM split the reconstruction procedure into two main stages. The first stage, called *particle picking*, detects and extracts the particle projections from the micrographs. Given the projections, the second stage aims to reconstruct a 3-D model of the molecular structure, usually using an expectation-maximization algorithm [40]. Crucially, reliable detection of individual particles is impossible in a highly noisy environment. This fact has been recognized early on by the cryo-EM community. Particularly, in [22, 19], it was reasoned that particle picking is impossible for molecules below a certain weight (below $\sim$50 kDa).

Even if particle picking is feasible, procedures may be affected by *model bias*. Particle picking algorithms are typically based on correlating the micrograph with several templates. Areas highly correlated with some of the templates are assumed to contain projections: these yield the picked particles. Clearly, the picked particles depend heavily on the chosen templates: different templates may lead to different picked particles, hence, which is more problematic, to different 3-D structure reconstructions. The cryo-EM community is well aware of this potential pitfall [45, 23, 44], which was notably exemplified by the "Einstein from noise" experiment [42].

A recent work of the authors suggests a methodology to bypass particle picking and reconstruct the 3-D structure directly from the micrograph [13]. Based on autocorrelation analysis, it was shown that—at least in principle—the limits high noise regimes impose on particle picking do not necessarily translate into limits on 3-D reconstruction. The main goal of the present paper is to provide theoretical and numerical support for this approach, in a simplified setting.

In the next section, we introduce autocorrelation analysis in detail, focusing on the multi-target detection model. We mention that similar ideas in cryo-EM can be traced back to a seminal paper of Zvi Kam [25]. Kam proposed autocorrelation analysis for 3-D reconstruction, under the assumption of picked, perfectly centered, particles. Kam's method has been extended and used in X-ray free electron lasers (XFEL) and cryo-EM [36, 28, 31, 46]. In order to investigate the computational and statistical properties of Kam's method, a series of papers have studied a simplified model, called *multi-reference alignment* [8, 14, 6, 39, 7, 1]. We follow the same line of research by considering the multi-target detection as an abstraction to the application of reconstructing 3-D structures directly from the micrograph [13].

3 Autocorrelation analysis

In what follows, we consider autocorrelations of both the observation y and of the signal occurrences in y. As per our discussion of extensions, the signal occurrences may be sampled from a discrete set $\{x_1, \dots, x_K\}$ (as in (1.3)), or from a more general distribution. Accordingly, we define autocorrelations broadly for a random signal z of length M. For our purposes, this will be applied both to signal occurrences (of length L) and to y (of length N).

For a random signal $z \in \mathbb{R}^M$, the autocorrelation of order $q = 1, 2, \ldots$ is given for any integer shifts $\ell_1, \ldots, \ell_{q-1}$ by

$$a_z^q[\ell_1, \ldots, \ell_{q-1}] = \mathbb{E}_z \left\{ \frac{1}{M} \sum_{i=-\infty}^{\infty} z[i] z[i+\ell_1] \cdots z[i+\ell_{q-1}] \right\}, \tag{3.1}$$

where the expectation is taken with respect to the distribution of z. Indexing out of bounds is zero-padded, that is, $z[i] = 0$ for i out of the range $0, \ldots, M-1$. Explicitly, the first-, second- and third-order autocorrelations are given by:

$$\begin{aligned} a_z^1 &= \mathbb{E}_z \left\{ \frac{1}{M} \sum_{i=0}^{M-1} z[i] \right\}, \\ a_z^2[\ell] &= \mathbb{E}_z \left\{ \frac{1}{M} \sum_{i=\max\{0,-\ell\}}^{M-1+\min\{0,-\ell\}} z[i] z[i+\ell] \right\}, \\ a_z^3[\ell_1, \ell_2] &= \mathbb{E}_z \left\{ \frac{1}{M} \sum_{i=\max\{0,-\ell_1,-\ell_2\}}^{M-1+\min\{0,-\ell_1,-\ell_2\}} z[i] z[i+\ell_1] z[i+\ell_2] \right\}. \end{aligned} \tag{3.2}$$

Since autocorrelations depend only on the differences between indices, they obey the following symmetries:

$$a_z^2[\ell] = a_z^2[-\ell],$$

and

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

In particular, for x sampled from $\{x_1, \ldots, x_K\}$ with probabilities $(\pi_1, \ldots, \pi_K)$ as in (1.3), the autocorrelations of x are given in explicit form in terms of those of the deterministic signals $x_1, \ldots, x_K$ as:

$$a_x^q = \sum_{k=1}^{K} \pi_k a_{x_k}^q. \tag{3.3}$$

Explicit expressions of the autocorrelations for the more involved model of cryo-EM (2.1) are given in [13].

We are given one observation (one realization) of y. Thus, we cannot compute the autocorrelations of y exactly as they involve taking an expectation against the distribution of y. However, by the law of large numbers, as N grows to infinity the empirical autocorrelations of y almost surely (a.s.) converge to the actual (population) autocorrelations of $y \in \mathbb{R}^N$, that is,

$$\lim_{N \to \infty} \frac{1}{N} \sum_{i=-\infty}^{\infty} y[i] y[i+\ell_1] \cdots y[i+\ell_{q-1}] \overset{a.s.}{=} a_y^q[\ell_1, \ldots, \ell_{q-1}]. \tag{3.4}$$

This provides a concrete means of estimating the quantities a_y^q. In the remainder of this section, we relate the observables a_y^q to the unknowns a_x^q, first under the well-separated model, then under the Poisson model.

3.1 Autocorrelations under the well-separated model

Under the separation condition (1.2), the relation between autocorrelations of the observation y and those of x is particularly simple, as we now show. It is useful to introduce some notation: let $|s| = \sum_i s[i]$ denote the number of signal occurrences in y, and let

$$\gamma = \frac{|s|L}{N}. \tag{3.5}$$

This γ is the fraction of entries of y occupied by signal occurrences. The separation condition imposes $\gamma \le \frac{L}{2L-1} \approx 1/2$. For the heterogeneous model (1.1), one can define

$$\gamma_k = \frac{|s_k|L}{N}, \tag{3.6}$$

where $|s_k| = \sum_i s_k[i]$ so that $\gamma = \sum_k \gamma_k$.

Owing to the separation condition, when correlating y with shifted versions of itself for shifts in $0, \dots, L-1$, any given occurrence of x in y is only ever correlated with itself, and never with another occurrence. As a result, the autocorrelations of y depend on the corresponding autocorrelations of x, the noise level σ and the density γ (which is a weak dependence on the support signal s). Specifically, we show the following identities in Appendix A:

$$a_y^1 = \gamma a_x^1, \tag{3.7}$$

$$a_y^2[\ell] = \gamma a_x^2[\ell] + \sigma^2 \delta[\ell], \tag{3.8}$$

$$a_y^3[\ell_1, \ell_2] = \gamma a_x^3[\ell_1, \ell_2] + \sigma^2 \gamma a_x^1\big(\delta[\ell_1] + \delta[\ell_2] + \delta[\ell_1 - \ell_2]\big), \tag{3.9}$$

where $\delta[0] = 1$ and $\delta[\ell \neq 0] = 0$, and indices ℓ, ℓ_1, ℓ_2 are in the range $0 \le \ell \le L-1$. Terms proportional to σ^2 are due to noise. If σ is known, they can be handled easily. If σ is unknown, one can either estimate it form the data, or one can ignore the few entries of the autocorrelations that are affected by σ—one in a_y^2 and $3L-2$ in a_y^3, a relatively small number in both cases.

We show in Section 4.1 that x, γ and σ can be identified uniquely from the observed autocorrelations for the homogeneous case $K = 1$ (all signal occurrences are the same).

3.2 Autocorrelations under the Poisson model

In this section, we give the expressions for the autocorrelations of the observed signal y under the Poisson model. We first note that the expected total number of signals occurring in y is equal to $(N - L + 1)\gamma/L$, and consequently the total lengths of all signals (including overlaps) divided by N is equal to

$$\frac{N-L+1}{N} \cdot \frac{\gamma}{L} \cdot L \quad \overset{N\to\infty}{\longrightarrow} \quad \gamma. \tag{3.10}$$

Therefore, in the large N limit, similarly to the role of γ in the well-separated model (3.5), the parameter γ may be interpreted as the signal density.

In Appendix B, we prove the following expressions for the autocorrelations of y under the Poisson model. As in the well-separated model, the observed autocorrelations do not depend on individual occurrences of the signal, but only on the distribution of x itself, and the parameters γ and σ:

$$a_y^1 = \gamma a_x^1, \tag{3.11}$$

$$a_y^2[\ell] = \gamma a_x^2[\ell] + \sigma^2 \delta[\ell] + (\gamma a_x^1)^2, \tag{3.12}$$

$$\begin{aligned} a_y^3[\ell_1, \ell_2] = {} & \gamma a_x^3[\ell_1, \ell_2] + \sigma^2 \gamma a_x^1 \big(\delta[\ell_1] + \delta[\ell_2] + \delta[\ell_1 - \ell_2]\big) \\ & + (\gamma a_x^1)^3 + \gamma a_x^1 \cdot (\gamma a_x^2[\ell_1] + \gamma a_x^2[\ell_2] + \gamma a_x^2[\ell_2 - \ell_1]). \end{aligned} \tag{3.13}$$

Note that from those autocorrelations, it is easy to retrieve the autocorrelations of the well-separated model for all entries unaffected by σ. If σ is known, then it is true for all entries.

In the homogeneous case ($K = 1$), we show in Section 4.2 that a_y^1, a_y^2 and a_y^3 identify uniquely the signal x, the Poisson parameter γ, and the noise level σ for generic x.

4 Theory

We begin this section by showing that, in the homogeneous case ($K = 1$), under both the well-separated and the Poisson models, the first three observed autocorrelations identify the (deterministic) signal x, the density parameter γ, and the noise level σ^2. Then, for the heterogeneous case ($K \geq 2$), we bound from above the number K of signals that can be recovered from those autocorrelations as a function of the signal length L. Finally, we briefly discuss multi-dimensional signals.

Before that, we start by showing that in the homogeneous case a deterministic signal $z \in \mathbb{R}^L$ is identified uniquely by its second and the third autocorrelations. Indeed, assuming $z[0]$ and $z[L-1]$ are nonzero, we can recover z explicitly by

$$z[k] = \frac{z[0]z[k]z[L-1]}{z[0]z[L-1]} = \frac{a_z^3[k, L-1]}{a_z^2[L-1]}, \tag{4.1}$$

for $k = 0, \ldots, L-1$. If $z[0]$ or $z[L-1]$ are equal to zero, then $a_z^2[L-1] = 0$ and we can use that indication to shrink L. This proves the following useful fact:

Proposition 4.1. *A deterministic signal $z \in \mathbb{R}^L$ is determined uniquely by a_z^2 and a_z^3.*

Note that the procedure described in (4.1) is not numerically stable: if $z[0]$ or $z[L-1]$ are close to 0, recovery of z is sensitive to errors in the autocorrelations. In practice, we recover z by fitting it to its autocorrelations using a nonconvex least-squares procedure, which is empirically robust to additive noise. In prior work, we have observed similar phenomena for the related problem of multi-reference alignment [14, 16, 1].

4.1 Guarantees for the homogeneous well-separated model

The observed moments a_y^1, a_y^2 and a_y^3 under the well-separated model do not immediately yield the autocorrelations of the signal x; rather, the two are related by the noise level σ and the signal density γ. We will show, however, that all parameters—x, γ and σ^2—are still identified uniquely by the observed moments of y.

First, we observe that if the noise level σ is known, generally, one can estimate γ from the first two moments of the micrograph. The other direction is true as well: if $\gamma > 0$ is known, then one can estimate σ^2. The proof is provided in Appendix C.

Proposition 4.2. *Assume the separation condition (1.2) holds and $K = 1$ (all signal occurrences in y are identical, up to noise). If the mean of x is nonzero, then*

$$\gamma = \frac{L(a_y^1)^2}{a_y^2[0] + 2\sum_{\ell=1}^{L-1} a_y^2[\ell] - \sigma^2}, \tag{4.2}$$

meaning γ can be determined from σ (and vice versa) using the observables a_y^1, a_y^2.

Using third-order autocorrelation information of y, both the ratio γ and the noise σ can be determined simultaneously. For the following results, when we say that a result holds for a "generic" signal x, we mean that the set of signals which cannot be determined by these measurements has Lebesgue measure zero. In particular, this means that we can recover almost all signals with the given measurements. The proof is provided in Appendix D.

Proposition 4.3. *Assume $L \geq 3$, $K = 1$ and assume that the separation condition (1.2) holds. Then, the observed autocorrelations 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 $\gamma > \frac{1}{4}$, then this holds for any signal x with nonzero mean.*

From Propositions 4.1 and 4.3 we deduce the following:

Corollary 4.4. *Assume $L \geq 3$ and $K = 1$. Under the separation condition (1.2), the signal x, the ratio γ, and the noise level σ are determined from the first three autocorrelation functions of y if either the signal x is generic, or x has nonzero mean and $\gamma > \frac{1}{4}$.*

4.2 Guarantees for the homogeneous Poisson model

Similarly to the homogeneous well-separated model, the observed autocorrelations under the Poisson model identify x, γ and σ^2 uniquely. Opposed to Proposition 4.3, these quantities can be computed explicitly. In Appendix E we prove the following result:

Proposition 4.5. *Under the homogeneous Poisson model, the signal x, the noise level σ^2 and the Poisson parameter γ are identified uniquely from the observed autocorrelations. In particular,*

$$\gamma = L\frac{(\gamma a_y^1)(a_y^2[1] - (a_y^1)^2)}{\sum_{\ell=0}^{L-1}\gamma a_x^3[1,\ell] + \sum_{\ell=2}^{L-1}\gamma a_x^3[\ell,\ell+1]}, \tag{4.3}$$

$$\sigma^2 = a_y^2[0] + 2\sum_{\ell=1}^{L-1} a_y^2[\ell] - \frac{L(a_y^1)^2}{\gamma} - (2L-1)(a_y^1)^2, \tag{4.4}$$

where γa_x^3 is indeed observable since

$$\gamma a_x^3[\ell_1,\ell_2] = a_y^3[\ell_1,\ell_2] - (a_y^1)^3 - a_y^1\cdot\big(a_y^2[\ell_1] + a_y^2[\ell_2] + a_y^2[\ell_1-\ell_2] - 3(a_y^1)^2\big). \tag{4.5}$$

4.3 Elementary limitations of the heterogeneous case

In the heterogeneous model (1.1), the unknowns are K signals of length L, together with their densities $\gamma_1,\dots,\gamma_K$ (3.6) (equivalently: the distribution π and overall density γ) and possibly the noise level σ. To estimate these parameters, we must collect at least as many independent equations. Within our framework, polynomial equations are provided by the observable autocorrelations, which correspond to mixed autocorrelations of the unknowns as per (3.3). In this section, following [16], we count how many equations the first three autocorrelations may provide in the best case (discounting symmetries). This leads to a straightforward information-theoretic upper bound on the number K of signals which can be estimated, as a function of L. This is only an upper bound, though a bound of the same type was shown to be tight in a similar setting [7]. The counting is based on the autocorrelations of the well-separated model. For entries independent of σ, the autocorrelations of the Poisson process contain the same information as those of the well-separated model. If σ is known, this holds true for all entries; see Section 3.2.

The first-order autocorrelation a_y^1 (3.7) provides one equation. For second-order autocorrelations $a_y^2[\ell]$ (3.8), if σ is known we obtain L equations with ℓ ranging from 0 to $L-1$. If σ is unknown, we may disregard $a_y^2[0]$ (the only entry affected by σ) and still collect $L-1$ equations. Similarly, for third-order autocorrelations, $a_y^3[\ell_1,\ell_2]$ (3.9) with $0\le\ell_1,\ell_2\le L-1$ such that $\ell_2\le\ell_1$ includes all relevant entries for our purpose (this accounts for symmetries), providing $\frac{(L+1)(L+2)}{2}-2$ equations in total. If we further exclude any entries 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.

Hence, if σ is known we collect

$$1 + L + \frac{(L+1)(L+2)}{2} - 2 = \frac{1}{2}L(L+5)$$

equations, while if it is unknown and we choose not to estimate it, then we collect

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

equations in total. Of course, there may be redundancy in these equations: we aim only to provide an upper bound.

Since we aim to estimate KL parameters for the K signals of length L, plus K parameters for the densities γ_k, there are $K(L+1)$ unknowns. As a result, an absolute upper bound on K such that the estimation problem may be solvable is

$$K \leq \frac{L(L+5)}{2(L+1)}$$

for the case of σ known, and

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

for the case of σ unknown and not estimated. Overall, this indicates that, at best, approximately $L/2$ signals and their densities can be recovered from the first three mixed autocorrelations. Based on related results in [7], we expect that as many as $L/2$ signals can indeed be estimated, though possibly not with computationally tractable estimators.

4.4 Autocorrelations in higher dimensions

Autocorrelations in d dimensions are defined for $\ell_1, \ldots, \ell_{q-1} \in \mathbb{Z}^d$ as

$$a_z^q[\ell_1, \ldots, \ell_{q-1}] = \mathbb{E}\left\{ \frac{1}{L^d} \sum_{i \in \mathbb{Z}^d} z[i] z[i+\ell_1] \cdots z[i+\ell_{q-1}] \right\}. \tag{4.6}$$

Interestingly, for the homogeneous case ($K = 1$) in dimensions greater than one, almost all signals are determined uniquely from their second-order autocorrelation, up to two symmetries: sign (or phase for complex signals) and reflection through the origin (with conjugation in the complex case) [21]. If the mean of the signal is available and non-zero, the sign symmetry can be resolved. However, determining the reflection symmetry still requires additional information, beyond the second-order autocorrelation. The case of 1-D signals is fundamentally different: generally there are 2^{L-2} signals with the same second-order autocorrelation (after eliminating symmetries) [11, 12].

This uniqueness result for two and three-dimensional signals is the basis of a popular imaging technique called coherent diffraction imaging (CDI). In CDI, an object is illuminated with a coherent wave, and the far field diffraction pattern is measured, corresponding to the object's Fourier magnitude [38, 43]. If the diffraction pattern is over-sampled by at least twice the Nyquist frequency, the data is equivalent to the signal's second-order autocorrelation. In this context, the computational problem of recovering the signal from its second-order autocorrelation is usually referred to as *phase retrieval* or the *phase problem*. However, for 2-D images it has been shown recently that the problem is ill-conditioned unless the support of the image is known exactly [9]. That is, there exist other images whose second-order autocorrelations agree up to machine precision.

5 Algorithms and numerical experiments

In this section, we present three numerical experiments. In the first two experiments, we consider the heterogeneous model (1.1). First, with a fixed noise level, we show how the estimation quality improves as the length of the observation grows. Second, we explore how many signals can be recovered as a function of their length, in an infinite data regime where effects of the noise have been averaged out. In the last experiment, we extend our model to 2-D signals. We run the experiments on a shared computer with 144 logical CPUs of type Intel(R) Xeon(R) CPU E7-8880 v3 @ 2.30GHz and 755Gb of RAM; we use at most 72 of these CPUs. The code for all experiments is available at https://github.com/PrincetonUniversity/BreakingDetectionLimit.

In this section, to assess the quality of our reconstruction, we compare against the ground truth. In a realistic setting, the ground truth is of course not available, which raises the question of how one can validate the results. A common technique used in cryo-EM is to split the data in two halves, produce two independent reconstructions based on these halves, then to compare the two reconstructions. It is clear how the same technique can be applied to our setting.

5.1 Experiment 1

For the experiment depicted in Figure 1, we fix $K = 3$ signals of length $L = 21$: see the three red signals in the first column. The first signal's actual support has length 11 (rather than 21), which allows us to simulate the situation in which the support of the signal is overestimated.

Following (1.1), 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 according to (1.2). 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 (1.2) and locations picked so far. For each placement, one of the three signals is picked at random according to the proportions $\pi = (1/2, 1/3, 1/6)$. Then, i.i.d. Gaussian noise with mean zero and standard deviation $\sigma = 3$ is added, to form the observed y.

Visually, the noise dominates the signal to the point that it is challenging to detect occurrences. More precisely, the cross-correlations of y even with the true signals presents 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 cluster them according to which signal appears where.

Our aim is to investigate how accurately we can estimate the signals as a function of the observation length. To this end, we consider a growing part of the observation y. For each length, we compute autocorrelations on that part, then we go on to estimate the signals from these "mixed" quantities. In practice, the autocorrelations are computed

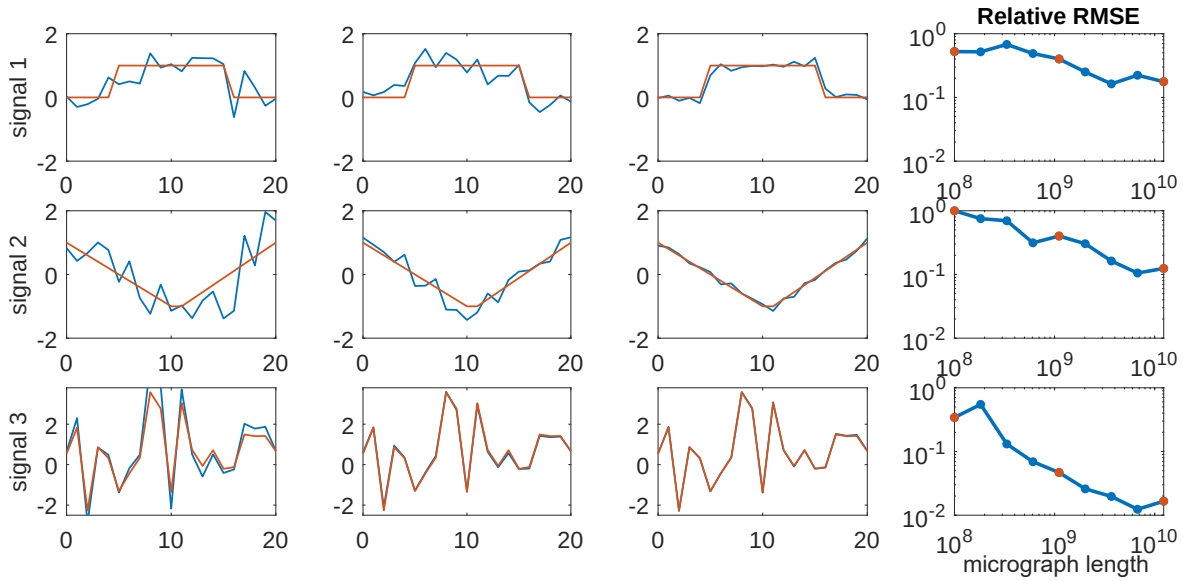


Figure 1: Experiment described in Section 5.1. For a fixed noise level $\sigma = 3$ and a fixed set of $K = 3$ signals of length $L = 21$, an observation y of length $N = 1.23 \cdot 10^{10}$ is generated according to (1.1) in the well-separated model, with fixed occurrence probabilities. Each row corresponds to one of the signals. The last column shows evolution of the relative root mean squared error in estimating each signal, as a longer and longer subset of y is observed. Red dots mark the three snapshots that are illustrated in columns 1–3: red signals are the ground truth and blue signals are the estimators.

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 be handled in parallel.

Having computed the autocorrelations of interest, we 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:

$$\min_{\substack{\hat{x}_1,\ldots,\hat{x}_K \in \mathbb{R}^W \\ \hat{\gamma}_1,\ldots,\hat{\gamma}_K > 0}} w_1 \left(a_y^1 - \sum_{k=1}^{K} \hat{\gamma}_k a_{\hat{x}_k}^1 \right)^2 + w_2 \sum_{\ell=1}^{L-1} \left(a_y^2[\ell] - \sum_{k=1}^{K} \hat{\gamma}_k a_{\hat{x}_k}^2[\ell] \right)^2 + w_3 \sum_{\substack{2 \le \ell_1 \le L-1 \\ 1 \le \ell_2 \le \ell_1 - 1}} \left(a_y^3[\ell_1, \ell_2] - \sum_{k=1}^{K} \hat{\gamma}_k a_{\hat{x}_k}^3[\ell_1, \ell_2] \right)^2, \quad (5.1)$$

where $W \ge L$ 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 coefficients used for each autocorrelation order: $n_2 = L - 1$, $n_3 = \frac{(L-1)(L-2)}{2}$ (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 problem 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 on either side 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 (5.1), this time with $W = L$. We find that this procedure is reliable for a wide range of experimental parameters. To solve (5.1), we run the trust-region method implemented in Manopt [15] from 10 different random initial guesses and keep the one with lowest cost function value. Manopt 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.

To asses reconstruction quality, we report the relative root mean squared error between the estimated signals and the ground truth (up to permutation of the K signals.) For the first signal (with the overestimated support), we also translationally aligned the estimate with the ground truth because its estimation is only possible up to shift.

In Figure 1, we find that the signals can be recovered with good accuracy despite the noise levels which seemingly hinders location and clustering. We also note that the amount of data required to produce these good estimations is large. Furthermore, as illustrated here and as we have observed in numerous experiments, signals with more variations (such as the third signal in this experiment which was generated once from a Gaussian distribution) are easier to estimate accurately than more regular signals (in this case, despite the fact that the third signal occurs less frequently than the others). This phenomenon has been also observed in multi-reference alignment [39, Section 3.2].

5.2 Experiment 2

In this second experiment, presented in Figure 2, we investigate how many distinct signals $x_1, \ldots, x_K$ can be estimated from mixed autocorrelations (3.3). In order to do so, we consider a setup where the mixed autocorrelations are known perfectly. This corresponds to the limit of an infinitely long observation y with fixed density γ and fixed noise level (that may be arbitrarily high). The specific value of σ is immaterial since we only consider autocorrelations that are unaffected by noise bias. Furthermore, we assume uniform occurrence distribution γ_k (known to the algorithm), and known density γ (which is then irrelevant as it only induces a global scaling of the autocorrelations of y).

To produce Figure 2, we consider each pair (K, L) in turn, with $K = 1, 2, 3, \ldots, 10$ and $L = 5, 10, 15, \ldots, 100$. For each pair, we generate K random normal signals of length L, once. The perfect mixed autocorrelations are computed. They are then provided to the inversion algorithm described in Section 5.1, together with the knowledge that signals occur with equal probability, as well as the correct density γ. The algo-

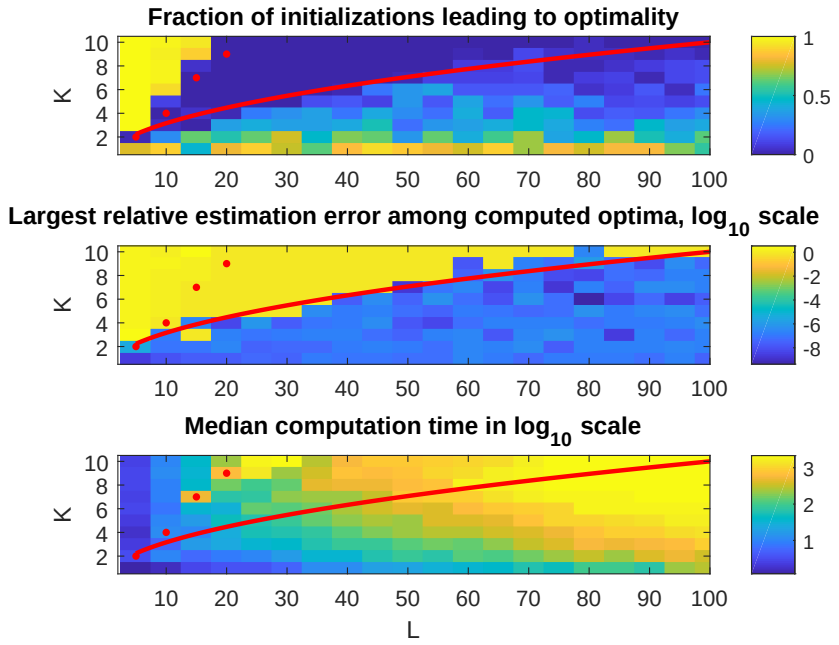


Figure 2: In the $N \to \infty$ regime (access to exact autocorrelations, excluding biased entries) and with known uniform densities, it seems that K up to $\sqrt{L}$ (red curve) i.i.d. Gaussian signals of length L can be recovered from the known moments. CPU time is in seconds. Strictly above red dots, recovery is impossible because the number of unknowns exceeds the number of computed autocorrelations; see Section 4.3. Similarly to [16, Fig. 4.1], this experiment suggests a possible statistical-computational gap.

rithm is initialized 50 times with an independent random initial guess, also following a normal distribution. For each run, we record three metrics:

1. Whether the optimization algorithm managed to produce a solution with cost function value below 10^{-16}: this assesses whether optimization succeeded.

2. The relative root mean squared error between the estimated signals and the ground truth (up to permutation of the K signals.)

3. The computation time in seconds (keeping in mind that the 50 runs are done in parallel on the same, shared computer, so that this is more of a qualitative assessment.)

These metrics are summarized and presented in Figure 2 as three panels.

1. Panel 1 shows for each pair (K, L) which fraction of the 50 runs reached optimality (between 0 and 1).

2. For each pair (K, L), any estimator produced by the optimization algorithm such that the cost function value is close to zero must be considered a valid estimator,

since it agrees almost perfectly with the data. For all of those, we compute the error compared to the ground truth. Panel 2 shows the largest such error, on a log scale in base 10. (If optimization never succeeded for that pair, we report a relative error of 1.) A large value means that, among all near global optima of the optimization problem (if any), at least one was a poor estimator. A small value indicates all computed near global optima were good estimators.

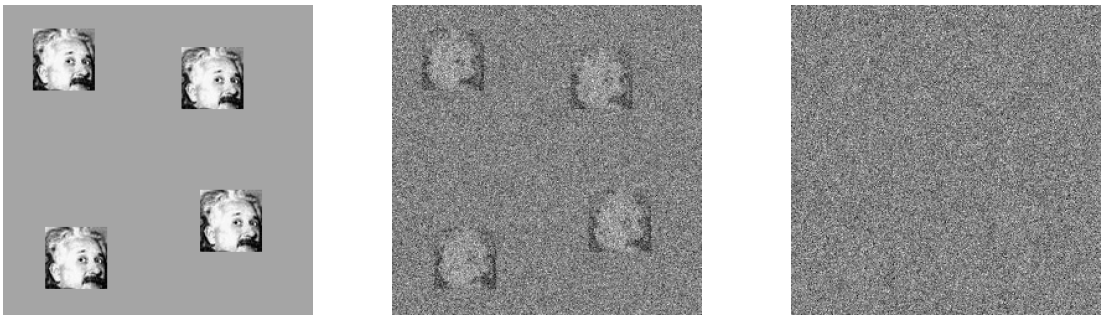
3. Median $\log_{10}$ computation time (where the median is computed after taking the log of the CPU times, in base 10.)

Following [16], overlaid on the panels we trace the red curve $K = \sqrt{L}$ as well as red dots which are computed from the considerations in Section 4.3 (adapted to the fact that the γ_k's are known). We observe that strictly above the red dots the optimization problem appears to be easy to solve (despite non-convexity), yet, as predicted, the corresponding estimators are not informative since there is not enough information in the computed quantities compared to the number of parameters. On the other hand, below the (empirical) red curve, the optimization problem is sometimes solved to optimality (although it may take more than one random initialization to get one successful run), and the corresponding estimators are accurate. In between the red curve and the red dots, the optimization problem appears to be particularly challenging: we essentially never produce a global optimum, hence we also do not have an estimator. This experiment suggests a possible computational-statistical gap in the area between the red curve and the red dots, where it is possible that the signals could be estimated, but perhaps not with a computationally efficient procedure. Similar results were observed for the multi-reference problem [16, 47, 7].

5.3 Experiment 3

Autocorrelation analysis can be carried out in dimensions greater than one. In the following experiment, we estimate a 50-by-50 pixel grayscale picture of Einstein with mean zero from a growing number of observations y. Each observation is of size 4096×4096 pixels and contains 700 occurrences on average at random locations, while maintaining the separation condition (1.2) in each axis separately. The observations are contaminated with additive white Gaussian noise with standard deviation $\sigma = 3$, illustrated in Figure 3.

We compute the average second-order autocorrelation of the observations. This is a particularly simple computation which can be efficiently executed with a fast Fourier transform (FFT), in parallel over the numerous observations. We assume the number of signal occurrences (akin to the density γ) and the standard deviation of the noise, σ, are known. Given those quantities, the second-order autocorrelation of the image can be easily deduced from (3.8). As explained in Section 4.4, an image is determined uniquely form its second-order autocorrelations. Then, to estimate the target image, we use a standard phase retrieval algorithm called relaxed-reflect-reflect (RRR) [17], initialized randomly.



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

Figure 3: Example of observations for the 2-D experiment (of size 250×250) with additive white Gaussian noise of variance σ^2 for increasing values of σ. Each observation contains the same four occurrences of a 50×50 image of Einstein. In panel (c), the noise level is such that it is challenging to detect the planted images.

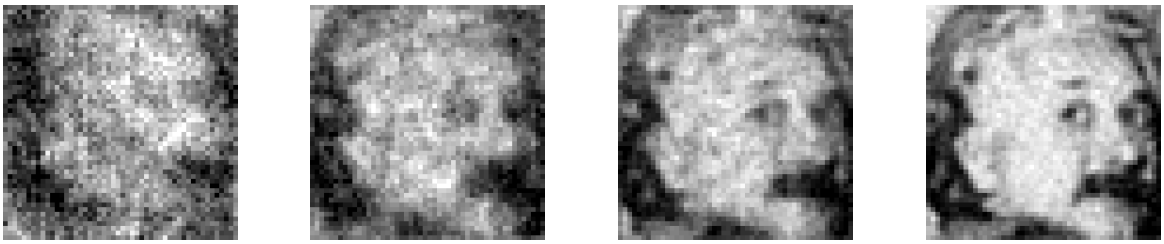


Figure 4: Recovery of Einstein from observations at noise level $\sigma = 3$ (see Figure 3(c)). Averaged autocorrelations of the data allow to estimate the power spectrum of the target image. This does not require locating the signal occurrences. The RRR algorithm produces the estimates, obtained from $2 \times 10^2, 2 \times 10^3, 2 \times 10^4$ and 2×10^5 observations (growing across panels from left to right).

Relative error is measured as the ratio of the root mean square error to the norm of the ground truth (square root of the sum of squared pixel intensities). Figure 4 shows several estimated images for a growing number of observations. Figure 5 presents the normalized recovery error as a function of the amount of data available. This is computed after fixing the reflection symmetries (see Section 4.4).

As evidenced by these figures, the ground truth image can be estimated increasingly well from increasingly many observations, without the need to locate the signal occurrences.

6 Summary

This paper suggests a computational framework for estimation under extreme noise levels. The crux of the method lies in the distinction between parameters of interest (the signals) and nuisance variables (parameters associated with individual signal oc-

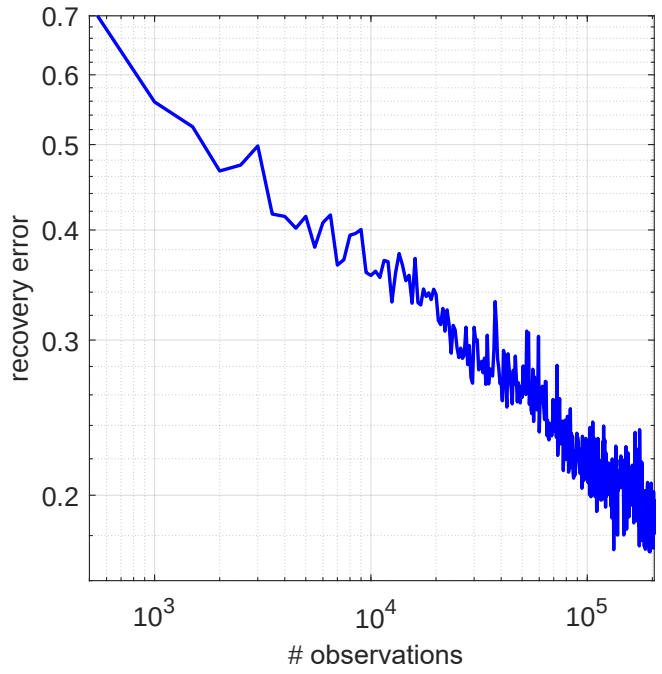


Figure 5: Relative error curve for Experiment 3 in Figure 4. Each observation contains about 700 image occurrences at unknown locations.

currences, such as location and class). In part through theory and in part through numerical experiments, we show that estimating the signals is possible even when they cannot be detected in the data. The method consists of two steps. First, we estimate the autocorrelations of the observation. A key feature is that, for any noise level, these autocorrelations can be estimated to any desired accuracy given sufficiently rich observations. Second, we recover the signals from the autocorrelations. This recovery entails solving a system of low-order polynomial equations. While solving such systems is hard in general, we found that in the homogeneous case we can solve them explicitly, and in both the homogeneous and heterogeneous cases we can solve them with reasonable robustness through non-convex optimization, in a wide regime of parameters.

In addition, autocorrelation analysis provides a flexible framework to extend the multi-target detection model by relating the expected autocorrelations of the data with the signals, and all parameters necessary to describe the generative model. For instance, a follow-up paper relaxes the separation condition (1.2) and allows an arbitrary spacing of targets, as long as the signal occurrences do not overlap [29]. In that case, the autocorrelations of the data are functions of the signal and the unknown target distribution. In a similar fashion, one may include more realistic noise models, the effect of a blurring kernel (e.g., the point spread function of the microscope) and so on.

The prime motivation of this paper emanates from challenges in small particle reconstruction using cryo-EM. Small particles induce such low signal-to-noise ratio in the

micrograph that particle picking—the first step in any current cryo-EM reconstruction algorithm—is impossible. The main message of our recent report [13] is that particle picking is merely a means to an end (although admittedly of key usefulness when it can be done): the locations and classes of individual particle projections are nuisance variables. The ultimate goal is only to estimate the 3-D structures. To this end, we used autocorrelation analysis to estimate the structure directly from the micrograph, without particle picking. In order to gain better understanding of the method, this paper focuses on an abstraction of cryo-EM—the multi-target detection model. Our next goal is to reconsider the full cryo-EM model, both from theoretical and algorithmic perspectives. In particular, the numerical results in [13] suggest that the achieved resolution using autocorrelations up to third order is limited by ill-conditioning of the system of polynomial equations. Higher resolution may require computing higher-order autocorrelations, which would increase the sample complexity and computational complexity of the algorithm. Despite the challenges, we believe that this approach may ultimately offer a way to reconstruct 3-D structures that are too small for current algorithmic pipelines.

Acknowledgment

The authors thank Ayelet Heimowitz, Joe Kileel, Ti-Yen Lan and Amit Moscovich for helpful discussions. The research is supported in parts by Award Number R01GM090200 from the NIGMS, FA9550-17-1-0291 from AFOSR, Simons Foundation Math+X Investigator Award, the Moore Foundation Data-Driven Discovery Investigator Award, and NSF BIGDATA Award IIS-1837992. NB is partially supported by NSF award DMS-1719558.

References

[1] Emmanuel Abbe, Tamir Bendory, William Leeb, João M Pereira, Nir Sharon, and Amit Singer. Multireference alignment is easier with an aperiodic translation distribution. *IEEE Transactions on Information Theory*, 65(6):3565–3584, 2019.

[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] Ali Ahmed, Benjamin Recht, and Justin Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2014.

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

[6] Afonso Bandeira, Philippe Rigollet, and Jonathan Weed. Optimal rates of estimation for multi-reference alignment. *arXiv preprint arXiv:1702.08546*, 2017.

[7] Afonso S Bandeira, Ben Blum-Smith, Joe Kileel, Amelia Perry, Jonathan Weed, and Alexander S Wein. Estimation under group actions: recovering orbits from invariants. *arXiv preprint arXiv:1712.10163*, 2017.

[8] Afonso S Bandeira, Moses Charikar, Amit Singer, and Andy Zhu. Multireference alignment using semidefinite programming. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 459–470. ACM, 2014.

[9] Alexander Barnett, Charles L Epstein, Leslie Greengard, and Jeremy Magland. Geometry of the phase retrieval problem. *arXiv preprint arXiv:1808.10747*, 2018.

[10] Alberto Bartesaghi, Alan Merk, Soojay Banerjee, Doreen Matthies, Xiongwu Wu, Jacqueline LS Milne, and Sriram Subramaniam. 2.2 Å resolution cryo-EM structure of β-galactosidase in complex with a cell-permeant inhibitor. *Science*, 348(6239):1147–1151, 2015.

[11] 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.

[12] 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.

[13] Tamir Bendory, Nicolas Boumal, William Leeb, Eitan Levin, and Amit Singer. Toward single particle reconstruction without particle picking: Breaking the detection limit. *arXiv preprint arXiv:1810.00226*, 2018.

[14] Tamir Bendory, Nicolas Boumal, Chao Ma, Zhizhen Zhao, and Amit Singer. Bispectrum inversion with application to multireference alignment. *IEEE Transactions on Signal Processing*, 66(4):1037–1050, 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. In *Information Sciences and Systems (CISS), 2018 52nd Annual Conference on*, pages 1–6. IEEE, 2018.

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

[18] Joachim Frank. *Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state*. Oxford University Press, 2006.

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

[20] 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*, 54(1):18–36, 2018.

[21] M. 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.

[22] 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.

[23] 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.

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

[25] Zvi Kam. The reconstruction of structure from electron micrographs of randomly oriented particles. *Journal of Theoretical Biology*, 82(1):15–39, 1980.

[26] Werner Kühlbrandt. The resolution revolution. *Science*, 343(6178):1443–1444, 2014.

[27] Han-Wen Kuo, Yenson Lau, Yuqian Zhang, and John Wright. Geometry and symmetry in short-and-sparse deconvolution. *arXiv preprint arXiv:1901.00256*, 2019.

[28] Ruslan P Kurta, Jeffrey J Donatelli, Chun Hong Yoon, Peter Berntsen, Johan Bielecki, Benedikt J Daurer, Hasan DeMirci, Petra Fromme, Max Felix Hantke, Filipe RNC Maia, et al. Correlations in scattered X-ray laser pulses reveal nanoscale structural features of viruses. *Physical review letters*, 119(15):158102, 2017.

[29] Ti-Yen Lan, Tamir Bendory, Nicolas Boumal, and Amit Singer. Multi-target detection with an arbitrary spacing distribution. *arXiv preprint arXiv:1905.03176*, 2019.

[30] Kiryung Lee, Yanjun Li, Marius Junge, and Yoram Bresler. Blind recovery of sparse signals from subsampled convolution. *IEEE Transactions on Information Theory*, 63(2):802–821, 2017.

[31] Eitan Levin, Tamir Bendory, Nicolas Boumal, Joe Kileel, and Amit Singer. 3D ab initio modeling in cryo-EM by autocorrelation analysis. In *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pages 1569–1573. IEEE, 2018.

[32] 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.

[33] Xiaodong Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Applied and Computational Harmonic Analysis*, 2018.

[34] Yanjun Li, Kiryung Lee, and Yoram Bresler. Identifiability in blind deconvolution with subspace or sparsity constraints. *IEEE Transactions on Information Theory*, 62(7):4266–4275, 2016.

[35] Shuyang Ling and Thomas Strohmer. Blind deconvolution meets blind demixing: Algorithms and performance bounds. *IEEE Transactions on Information Theory*, 63(7):4497–4520, 2017.

[36] Haiguang Liu, Billy K Poon, Dilano K Saldin, John CH Spence, and Peter H Zwart. Three-dimensional single-particle imaging using angular correlations from X-ray laser data. *Acta Crystallographica Section A: Foundations of Crystallography*, 69(4):365–373, 2013.

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

[38] Jianwei Miao, Pambos Charalambous, Janos Kirz, and David Sayre. Extending the methodology of X-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens. *Nature*, 400(6742):342, 1999.

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

[40] Sjors HW Scheres. RELION: implementation of a bayesian approach to cryo-EM structure determination. *Journal of structural biology*, 180(3):519–530, 2012.

[41] 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.

[42] 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.

[43] Yoav Shechtman, Yonina C Eldar, Oren Cohen, Henry Nicholas Chapman, Jianwei Miao, and Mordechai Segev. Phase retrieval with application to optical imaging: a contemporary overview. *IEEE signal processing magazine*, 32(3):87–109, 2015.

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

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

[46] Benjamin von Ardenne, Martin Mechelke, and Helmut Grubmüller. Structure determination from single molecule X-ray scattering with three photons per image. *Nature communications*, 9(1):2375, 2018.

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

A Autocorrelations in the well-separated model

Let $x_{(1)}, \dots, x_{(|s|)}$ denote the (independent) realizations of the random signal x in the observation y, starting at (deterministic) positions $s_{(1)}, \dots, s_{(|s|)}$. Let I_{ij} be the indicator variable for whether position i is in the support of occurrence j, that is, it is one if i is in $\{s_{(j)}, \dots, s_{(j)} + L - 1\}$, and zero otherwise. Then,

$$y[i] = \sum_{j=1}^{|s|} I_{ij} x_{(j)}[i - s_{(j)}] + \varepsilon[i]. \tag{A.1}$$

This gives a simple expression for the first autocorrelation of y. Indeed,

$$a_y^1 = \mathbb{E}_y \left\{ \frac{1}{N} \sum_{i=0}^{N-1} y[i] \right\} \tag{A.2}$$

$$= \frac{1}{N} \mathbb{E}_{x_{(1)},\dots,x_{(|s|)},\varepsilon} \left\{ \sum_{i=0}^{N-1} \sum_{j=1}^{|s|} I_{ij} x_{(j)}[i - s_{(j)}] + \varepsilon[i] \right\}. \tag{A.3}$$

Now switch the sums over i and j, and observe that I_{ij} is zero unless $i = s_{(j)} + t$ for t in the range $0, \dots, L-1$. Hence,

$$a_y^1 = \frac{1}{N} \sum_{j=1}^{|s|} \mathbb{E}_{x_{(j)}} \left\{ \sum_{t=0}^{L-1} x_{(j)}[t] \right\} + \frac{1}{N} \mathbb{E}_{\varepsilon} \left\{ \sum_{i=0}^{N-1} \varepsilon[i] \right\}. \tag{A.4}$$

Since the noise has zero mean and $x_{(1)}, \ldots, x_{(|s|)}$ are identically distributed, we further find:

$$a_y^1 = \frac{|s|L}{N} a_x^1 = \gamma a_x^1. \tag{A.5}$$

To address the second-order moments, we resort to the separation condition (1.2). First, consider this expression:

$$\begin{aligned}
N \cdot a_y^2[\ell] &= \mathbb{E}_y \left\{ \sum_{i=0}^{N-\ell-1} y[i]y[i+\ell] \right\} \\
&= \sum_{i=0}^{N-\ell-1} \mathbb{E}_{x_{(1)},\ldots,x_{(|s|)},\varepsilon} \left\{ \left(\sum_{j=1}^{|s|} I_{ij} x_{(j)}[i - s_{(j)}] + \varepsilon[i] \right) \cdot \right. \\
&\qquad \left. \left(\sum_{j'=1}^{|s|} I_{i+\ell,j'} x_{(j')}[i + \ell - s_{(j')}] + \varepsilon[i+\ell] \right) \right\} \\
&= \sum_{i=0}^{N-\ell-1} \mathbb{E}_{x_{(1)},\ldots,x_{(|s|)},\varepsilon} \left\{ \sum_{j=1}^{|s|} \sum_{j'=1}^{|s|} I_{ij} I_{i+\ell,j'} x_{(j)}[i - s_{(j)}] x_{(j')}[i + \ell - s_{(j')}] \right. \\
&\qquad + \sum_{j=1}^{|s|} I_{ij} x_{(j)}[i - s_{(j)}] \varepsilon[i+\ell] \\
&\qquad + \sum_{j'=1}^{|s|} I_{i+\ell,j'} x_{(j')}[i + \ell - s_{(j')}] \varepsilon[i] \\
&\qquad \left. + \varepsilon[i]\varepsilon[i+\ell] \right\}.
\end{aligned}$$

The cross-terms vanish in expectation since ε is zero mean and independent from the signal occurrences. The last term vanishes in expectation unless $\ell = 0$ since distinct entries of ε are independent. For $\ell = 0$, $\mathbb{E}\{\varepsilon[i]^2\} = \sigma^2$. Finally, using the separation property, observe that if $I_{ij} I_{i+\ell,j'}$ is nonzero, then it is equal to one, $j = j'$ and $i = s_{(j)} + t$ for some t in $0, \ldots, L - \ell - 1$. Then, switch the order of summations to get

$$N \cdot a_y^2[\ell] = \sum_{j=1}^{|s|} \mathbb{E}_{x_{(j)}} \left\{ \sum_{t=0}^{L-\ell-1} x_{(j)}[t] x_{(j)}[t+\ell] \right\} + (N - \ell)\sigma^2 \delta[\ell], \tag{A.6}$$

where $\delta[0] = 1$ and $\delta[\ell \neq 0] = 0$. Since each $x_{(j)}$ is distributed as x, they all have the same autocorrelations as x and we finally get

$$a_y^2[\ell] = \gamma a_x^2[\ell] + \frac{N - \ell}{N} \sigma^2 \delta[\ell] = \gamma a_x^2[\ell] + \sigma^2 \delta[\ell]. \tag{A.7}$$

We now turn to the third-order autocorrelations. These involve the sum

$$\sum_{i=0}^{N-\max(\ell_1,\ell_2)-1} y[i]y[i+\ell_1]y[i+\ell_2]. \tag{A.8}$$

Using (A.1), we find that this quantity can be expressed as a sum of eight terms:

1. $\sum_i \sum_{j,j',j''=1}^{|s|} I_{ij} I_{i+\ell_1,j'} I_{i+\ell_2,j''} x_{(j)}[i-s_{(j)}] x_{(j')}[i+\ell_1-s_{(j')}] x_{(j'')}[i+\ell_2-s_{(j'')}]$
2. $\sum_i \sum_{j,j'=1}^{|s|} I_{ij} I_{i+\ell_1,j'} x_{(j)}[i-s_{(j)}] x_{(j')}[i+\ell_1-s_{(j')}] \varepsilon[i+\ell_2]$
3. $\sum_i \sum_{j,j''=1}^{|s|} I_{ij} I_{i+\ell_2,j''} x_{(j)}[i-s_{(j)}] \varepsilon[i+\ell_1] x_{(j'')}[i+\ell_2-s_{(j'')}]$
4. $\sum_i \sum_{j',j''=1}^{|s|} I_{i+\ell_1,j'} I_{i+\ell_2,j''} \varepsilon[i] x_{(j')}[i+\ell_1-s_{(j')}] x_{(j'')}[i+\ell_2-s_{(j'')}]$
5. $\sum_i \sum_{j=1}^{|s|} I_{ij} x_{(j)}[i-s_{(j)}] \varepsilon[i+\ell_1] \varepsilon[i+\ell_2]$
6. $\sum_i \sum_{j'=1}^{|s|} I_{i+\ell_1,j'} \varepsilon[i] x_{(j')}[i+\ell_1-s_{(j')}] \varepsilon[i+\ell_2]$
7. $\sum_i \sum_{j''=1}^{|s|} I_{i+\ell_2,j''} \varepsilon[i] \varepsilon[i+\ell_1] x_{(j'')}[i+\ell_2-s_{(j'')}]$
8. $\sum_i \varepsilon[i]\varepsilon[i+\ell_1]\varepsilon[i+\ell_2]$

Terms 2–4 and 8 vanish in expectation since odd moments of centered Gaussian variables are zero. For the first term, we use the fact that the separation condition implies

$$I_{ij} I_{i+\ell_1,j'} I_{i+\ell_2,j''} = 1 \iff j = j' = j'' \text{ and } i = s_{(j)} + t \text{ with } t \in \{0, \dots L - \max(\ell_1,\ell_2) - 1\}. \tag{A.9}$$

(Otherwise, the product of indicators is zero.) This allows to reduce the summations over j, j', j'' to a single sum over j. Then, switching the order of summation with i, we get that the first term is equal to

$$\sum_{j=1}^{|s|} \sum_{t=0}^{L-\max(\ell_1,\ell_2)-1} x_{(j)}[t] x_{(j)}[t+\ell_1] x_{(j)}[t+\ell_2]. \tag{A.10}$$

In expectation over the realizations $x_{(j)}$, using again that they are i.i.d. with the same distribution as x, this first term yields $|s| L a_x^3[\ell_1, \ell_2]$. Now consider the fifth term. Taking expectation against ε yields

$$\sum_{i=0}^{N-\max(\ell_1,\ell_2)-1} \sum_{j=1}^{|s|} I_{ij} x_{(j)}[i-s_{(j)}] \sigma^2 \delta[\ell_1 - \ell_2]. \tag{A.11}$$

Switch the order of summation over i and j again to get

$$\sigma^2 \delta[\ell_1 - \ell_2] \sum_{j=1}^{|s|} \sum_{t=0}^{L-1} x_{(j)}[t]. \tag{A.12}$$

Now taking expectation against the signal occurrences yields $|s| L \sigma^2 a_x^1 \delta[\ell_1 - \ell_2]$. A similar reasoning for terms 6 and 7 yields this final formula for the third-order autocorrelations of y:

$$a_y^3[\ell_1, \ell_2] = \gamma a_x^3[\ell_1, \ell_2] + \gamma \sigma^2 a_x^1 \left(\delta[\ell_1] + \delta[\ell_2] + \delta[\ell_1 - \ell_2] \right). \tag{A.13}$$

B Autocorrelations in the Poisson model

We will denote by m_l the moment tensors of x:

$$m_1[i] = \mathbb{E} x[i], \quad 0 \leq i \leq L-1, \tag{B.1}$$

$$m_2[i,j] = \mathbb{E} x[i] x[j], \quad 0 \leq i,j \leq L-1, \tag{B.2}$$

$$m_3[i,j,k] = \mathbb{E} x[i] x[j] x[k], \quad 0 \leq i,j,k \leq L-1. \tag{B.3}$$

We obtain the autocorrelations a_x^l of x by averaging over a slice of the moment tensors:

$$a_x^1 = \frac{1}{L} \sum_{i=0}^{L-1} m_1[i], \tag{B.4}$$

$$a_x^2[\ell] = \frac{1}{L} \sum_{i=0}^{L-1} m_2[i, i+\ell], \tag{B.5}$$

and

$$a_x^3[\ell_1, \ell_2] = \frac{1}{L} \sum_{i=0}^{L-1} m_3[i, i+\ell_1, i+\ell_2]. \tag{B.6}$$

We will make repeated use of the following elementary lemma:

Lemma B.1. *If $z \sim Poisson(\lambda)$, then*

$$\mathbb{E} \binom{z}{k} = \frac{\lambda^k}{k!}. \tag{B.7}$$

B.1 Computing a_y^1

We will first condition on the vector s of locations of the subsignals in y, and then average over s. We will denote by $x^i_{(1)}, \ldots, x^i_{(s[i])}$ the random vectors starting in $y[i]$. We have:

$$\mathbb{E}[y[i]|s] = \sum_{j=0}^{L-1} \sum_{k=1}^{s[i-j]} \mathbb{E}x^{i-j}_{(k)}[j] = \sum_{k=1}^{s[i-j]} La^1_x = s[i-j]La^1_x. \tag{B.8}$$

Now taking expectations over s and using $\mathbb{E}s[i-j] = \gamma/L$ we get:

$$\mathbb{E}y[i] = \gamma a^1_x. \tag{B.9}$$

Consequently,

$$a^1_y = \frac{1}{N}\sum_{i=1}^{n} \mathbb{E}y[i] = \gamma a^1_x. \tag{B.10}$$

B.2 Computing a_y^2

First we will consider the noise-free case, where $\sigma = 0$. We will condition on s first, and then take the expectation over s. Fix $i_1 \neq i_2$, and let $\ell = i_2 - i_1$. Then:

$$y[i_1]y[i_2] = \sum_{j_1=0}^{L-1}\sum_{j_2=0}^{L-1}\sum_{k_1=1}^{s[i_1-j_1]}\sum_{k_2=1}^{s[i_2-j_2]} x^{i_1-j_1}_{(k_1)}[j_1]x^{i_2-j_2}_{(k_2)}[j_2]. \tag{B.11}$$

We break up the double sum over j_1 and j_2 into two terms: one where $j_2 \neq j_1 + \ell$, and one where $j_2 = j_1 + \ell$ or equivalently $i_1 - j_1 = i_2 - j_2$. In the first case, all the terms are independent, and so the expectation factors. In the second case, when $k_1 \neq k_2$ we have independence, but otherwise not. This gives (all expectations are conditional on s):

$$\begin{aligned}
\mathbb{E}y[i_1]y[i_2] &= \sum_{j_1=0}^{L-1}\sum_{j_2=0}^{L-1}\sum_{k_1=1}^{s[i_1-j_1]}\sum_{k_2=1}^{s[i_2-j_2]} \mathbb{E}x^{i_1-j_1}_{(k_1)}[j_1]x^{i_2-j_2}_{(k_2)}[j_2] \\
&= \sum_{j_1-j_2\neq\ell}\sum_{k_1}\sum_{k_2} \mathbb{E}x^{i_1-j_1}_{(k_1)}[j_1]x^{i_2-j_2}_{(k_2)}[j_2] \\
&\quad + \sum_{j_1=0}^{L-1}\sum_{k_1\neq k_2} \mathbb{E}x^{i_1-j_1}_{(k_1)}[j_1]x^{i_1-j_1}_{(k_2)}[j_1+\ell] \\
&\quad + \sum_{j_1=0}^{L-1}\sum_{k_1=1}^{s[i_1-j_1]} \mathbb{E}x^{i_1-j_1}_{(k_1)}[j_1]x^{i_1-j_1}_{(k_1)}[j_1+\ell] \\
&= \sum_{j_1-j_2\neq\ell} s[i_1-j_1]s[i_2-j_2]m_1[j_1]m_1[j_2]
\end{aligned}$$

$$+\sum_{j_1=0}^{L-1} s[i_1-j_1](s[i_1-j_1]-1)m_1[j_1]m_1[j_1+\ell]$$
$$+\sum_{j_1=0}^{L-1} s[i_1-j_1]m_2[j_1,j_1+\ell]. \tag{B.12}$$

Now take expectations over the Poisson random variables, using Lemma B.1:

$$\begin{aligned}
\mathbb{E}y[i_1]y[i_2] =& \sum_{j_1-j_2\neq\ell} \mathbb{E}s[i_1-j_1]s[i_2-j_2]m_1[j_1]m_1[j_2] \\
&+\sum_{j_1=0}^{L-1} \mathbb{E}s[i_1-j_1](s[i_1-j_1]-1)m_1[j_1]m_1[j_1+\ell] \\
&+\sum_{j_1=0}^{L-1} \mathbb{E}s[i_1-j_1]m_2[j_1,j_1+\ell] \\
=&\frac{1}{L^2}\sum_{j_1-j_2\neq\ell} \gamma^2 m_1[j_1]m_1[j_2] + \frac{1}{L^2}\sum_{j_1=0}^{L-1} \gamma^2 m_1[j_1]m_1[j_1+\ell] \\
&+\frac{1}{L}\sum_{j_1=0}^{L-1} \gamma m_2[j_1,j_1+\ell] \\
=&\left(\frac{\gamma}{L}\sum_{j=0}^{L-1} m_1[j]\right)^2 + \frac{\gamma}{L}\sum_{j=0}^{L-1} m_2[j,j+\ell] \\
=&(\gamma a_x^1)^2 + \gamma a_x^2[\ell].
\end{aligned} \tag{B.13}$$

For positive σ, we observe that any terms linear in the noise vanish in expectation. Denoting by x^* the clean signal component of length N, so $y = x^* + \varepsilon$, we have:

$$\mathbb{E}y[i_1]y[i_2] = \mathbb{E}x^*[i_1]x^*[i_2] + \mathbb{E}\varepsilon[i_1]\varepsilon[i_2] = (\gamma a_x^1)^2 + \gamma a_x^2[\ell] + \sigma^2\delta[i_1-i_2]. \tag{B.14}$$

We conclude by averaging over i_1 and i_2 with a fixed value of $\ell = i_2 - i_1$.

B.3 Computing a_x^3

We will first assume $\sigma = 0$. For three distinct i_1, i_2 and i_3, we let $\ell_1 = i_2 - i_1$ and $\ell_2 = i_3 - i_1$. We have:

$$y[i_1]y[i_2]y[i_3] = \sum_{j_1=0}^{L-1}\sum_{j_2=0}^{L-1}\sum_{j_3=0}^{L-1}\sum_{k_1=1}^{s[i_1-j_1]}\sum_{k_2=1}^{s[i_2-j_2]}\sum_{k_3=1}^{s[i_3-j_3]} x_{(k_1)}^{i_1-j_1}[j_1]x_{(k_2)}^{i_2-j_2}[j_2]x_{(k_3)}^{i_3-j_3}[j_3]. \tag{B.15}$$

We will break up the outer three sums into disjoint sums with the following ranges of indices:

1. $j_2 = j_1 + \ell_1$ and $j_3 = j_2 + \ell_2 - \ell_1$.
2. $j_2 = j_1 + \ell_1$ and $j_3 \neq j_2 + \ell_2 - \ell_1$.
3. $j_2 \neq j_1 + \ell_1$ and $j_3 = j_1 + \ell_2$.
4. $j_2 \neq j_1 + \ell_1$ and $j_3 \neq j_1 + \ell_2$ and $j_3 = j_2 + \ell_2 - \ell_1$.
5. $j_2 \neq j_1 + \ell_1$ and $j_3 \neq j_1 + \ell_2$ and $j_3 \neq j_2 + \ell_2 - \ell_1$.

For Case 1, we have $\ell \equiv i_1 - j_1 = i_2 - j_2 = i_3 - j_3$. We further break up the sum:

$$\begin{aligned}
&\sum_{j=0}^{L-1} \sum_{k_1=1}^{s[\ell]} \sum_{k_2=1}^{s[\ell]} \sum_{k_3=1}^{s[\ell]} x^\ell_{(k_1)}[j] x^\ell_{(k_2)}[j+\ell_1] x^\ell_{(k_3)}[j+\ell_2] \\
&= \underbrace{\sum_{j=0}^{L-1} \sum_{k_i \text{distinct}} x^\ell_{(k_1)}[j] x^\ell_{(k_2)}[j+\ell_1] x^\ell_{(k_3)}[j+\ell_2]}_{\text{(a)}} \\
&\quad + \underbrace{\sum_{j=0}^{L-1} \sum_{k_1=k_2\neq k_3} x^\ell_{(k_1)}[j] x^\ell_{(k_2)}[j+\ell_1] x^\ell_{(k_3)}[j+\ell_2]}_{\text{(b)}} \\
&\quad + \underbrace{\sum_{j=0}^{L-1} \sum_{k_1=k_3\neq k_2} x^\ell_{(k_1)}[j] x^\ell_{(k_2)}[j+\ell_1] x^\ell_{(k_3)}[j+\ell_2]}_{\text{(c)}} \\
&\quad + \underbrace{\sum_{j=0}^{L-1} \sum_{k_2=k_3\neq k_1} x^\ell_{(k_1)}[j] x^\ell_{(k_2)}[j+\ell_1] x^\ell_{(k_3)}[j+\ell_2]}_{\text{(d)}} \\
&\quad + \underbrace{\sum_{j=0}^{L-1} \sum_{k_1=k_2=k_3} x^\ell_{(k_1)}[j] x^\ell_{(k_2)}[j+\ell_1] x^\ell_{(k_3)}[j+\ell_2]}_{\text{(e)}} .
\end{aligned} \tag{B.16}$$

For term (a), the expectation conditional on s is:

$$\sum_{j=0}^{L-1} s[\ell](s[\ell]-1)(s[\ell]-2) m_1[j] m_1[j+\ell_1] m_1[j+\ell_2]. \tag{B.17}$$

Using Lemma B.1, the unconditional expectation of (a) is then:

$$\frac{\gamma^3}{L^3} \sum_{j=0}^{L-1} m_1[j] m_1[j+\ell_1] m_1[j+\ell_2]. \tag{B.18}$$

For term (b), the expectation conditional on s is:

$$\sum_{j=0}^{L-1} s[\ell](s[\ell]-1)m_2[j, j+\ell_1]m_1[j+\ell_2] \tag{B.19}$$

and then again using Lemma B.1 we get the expected value:

$$\frac{\gamma^2}{L^2}\sum_{j=0}^{L-1} m_2[j, j+\ell_1]m_1[j+\ell_2]. \tag{B.20}$$

Similarly, the expected values of terms (c) and (d) are:

$$\frac{\gamma^2}{L^2}\sum_{j=0}^{L-1} m_2[j, j+\ell_2]m_1[j+\ell_1]. \tag{B.21}$$

and

$$\frac{\gamma^2}{L^2}\sum_{j=0}^{L-1} m_2[j+\ell_1, j+\ell_2]m_1[j]. \tag{B.22}$$

Finally, the expected value of term (e) is easily shown to be:

$$\frac{\gamma}{L}\sum_{j=0}^{L-1} m_3[j, j+\ell_1, j+\ell_2]. \tag{B.23}$$

This concludes the computation for Case 1.

Moving onto Case 2, we have $\Delta_1 \equiv i_1 - j_1 = i_2 - j_2$, and also define $\Delta_2 \equiv i_3 - j_3$. By definition, $\Delta_1 \neq \Delta_2$. The sum is:

$$\begin{aligned}
&\sum_{j_1=0}^{L-1}\sum_{j_3\neq j_1+\ell_2}\sum_{1\leq k_1,k_2\leq s[\Delta_1]}\sum_{k_3=1}^{s[\Delta_2]} x_{(k_1)}^{\Delta_1}[j_1]x_{(k_2)}^{\Delta_1}[j_1+\ell_1]x_{(k_3)}^{\Delta_2}[j_3] \\
=&\sum_{j_1=0}^{L-1}\sum_{j_3\neq j_1+\ell_2}\sum_{k_3=1}^{s[\Delta_2]}\Bigg\{\sum_{1\leq k_1\neq k_2\leq s[\Delta_1]} x_{(k_1)}^{\Delta_1}[j_1]x_{(k_2)}^{\Delta_1}[j_1+\ell_1]x_{(k_3)}^{\Delta_2}[j_3] \\
&+\sum_{k_1=1}^{s[\Delta_1]} x_{(k_1)}^{\Delta_1}[j_1]x_{(k_1)}^{\Delta_1}[j_1+\ell_1]x_{(k_3)}^{\Delta_2}[j_3]\Bigg\}.
\end{aligned} \tag{B.24}$$

Taking expectations conditional on s, we then get:

$$\sum_{j_1=0}^{L-1}\sum_{j_3\neq j_1+\ell_2}\Bigg(s[\Delta_1](s[\Delta_1]-1)s[\Delta_2]m_1[j_1]m_1[j_1+\ell_1]m_1[j_3]$$

$$+ s[\Delta_1]s[\Delta_2]m_2[j_1, j_1+\ell_1]m_1[j_3]\Big). \tag{B.25}$$

Taking expectations over s and using Lemma B.1 then gives:

$$\frac{\gamma^3}{L^3}\sum_{j_1=0}^{L-1}\sum_{j_3\neq j_1+\ell_2} m_1[j_1]m_1[j_1+\ell_1]m_1[j_3] \tag{B.26}$$

$$+\frac{\gamma^2}{L^2}\sum_{j_1=0}^{L-1}\sum_{j_3\neq j_1+\ell_2} m_2[j_1, j_1+\ell_1]m_1[j_3]. \tag{B.27}$$

Similarly, Cases 3 and 4 give the expressions:

$$\frac{\gamma^3}{L^3}\sum_{j_1=0}^{L-1}\sum_{j_2\neq j_1+\ell_1} m_1[j_1]m_1[j_1+\ell_2]m_1[j_2] \tag{B.28}$$

$$+\frac{\gamma^2}{L^2}\sum_{j_1=0}^{L-1}\sum_{j_2\neq j_1+\ell_1} m_2[j_1, j_1+\ell_2]m_1[j_2] \tag{B.29}$$

and

$$\frac{\gamma^3}{L^3}\sum_{j_2=0}^{L-1}\sum_{j_1\neq j_2} m_1[j_1]m_1[j_2+\ell_1]m_1[j_2+\ell_2] \tag{B.30}$$

$$+\frac{\gamma^2}{L^2}\sum_{j_2=0}^{L-1}\sum_{j_1\neq j_2} m_2[j_2+\ell_1, j_2+\ell_2]m_1[j_1]. \tag{B.31}$$

Finally, in Case 5 we have $i_1 - j_1$, $i_2 - j_2$, and $i_3 - j_3$ are all pairwise distinct. Consequently, the x variables are always independent, and the expectation conditional on s (letting $\Delta_q = i_q - j_q$, $q = 1, 2, 3$),

$$\sum_{j_1,j_2,j_3} s[\Delta_1]s[\Delta_2]s[\Delta_3]m_1[j_1]m_1[j_2]m_1[j_3]; \tag{B.32}$$

since the $s[\Delta_q]$'s are pairwise independent, the expectation over s then yields:

$$\frac{\gamma^3}{L^3}\sum_{j_1,j_2,j_3} m_1[j_1]m_1[j_2]m_1[j_3]. \tag{B.33}$$

Now we add all the terms from Cases 1 to 5. Expressions (B.18), (B.26), (B.28), (B.30), and (B.33) sum to the expression:

$$(\gamma a_x^1)^3. \tag{B.34}$$

Expressions (B.20), (B.21), (B.22), (B.27),(B.29), and (B.31) sum to the expression:

$$\gamma a_x^1 \cdot (\gamma a_x^2[\ell_1] + \gamma a_x^2[\ell_2] + \gamma a_x^2[\ell_2 - \ell_1]). \tag{B.35}$$

Finally, expression (B.23) is simply:

$$\gamma a_x^3[\ell_1, \ell_2]. \tag{B.36}$$

Now when $\sigma > 0$, we write $y = x^* + \varepsilon$, and

$$\begin{aligned}
\mathbb{E}y[i_1]y[i_2]y[i_3] &= \mathbb{E}x^*[i_1]x^*[i_2]x^*[i_3] + \mathbb{E}x^*[i_1]\varepsilon[i_2]\varepsilon[i_3] \\
&\quad + \mathbb{E}\varepsilon[i_1]x^*[i_2]\varepsilon[i_3] + \mathbb{E}\varepsilon[i_1]\varepsilon[i_2]x^*[i_3] + \mathbb{E}\varepsilon[i_1]\varepsilon[i_2]\varepsilon[i_3] \\
&= \mathbb{E}x^*[i_1]x^*[i_2]x^*[i_3] + \gamma a_x^1 \cdot \sigma^2\delta[i_2 - i_3] \\
&\quad + \gamma a_x^1 \cdot \sigma^2\delta[i_3 - i_1] + \gamma a_x^1 \cdot \sigma^2\delta[i_2 - i_1].
\end{aligned} \tag{B.37}$$

We conclude by averaging over all i_1, i_2, and i_3 with fixed values of $\ell_1 = i_2 - i_1$ and $\ell_2 = i_3 - i_1$.

C Proof of Proposition 4.2

Refer to equations (3.7)–(3.9) for expressions relating the moments of y and those of x, and the parameters γ and σ. First, note that

$$(a_y^1)^2 = \frac{\gamma^2}{L^2} \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} x[i]x[j].$$

Similarly, using $a_x^2[-\ell] = a_x^2[\ell]$:

$$\begin{aligned}
2\sum_{\ell=1}^{L-1} a_y^2[\ell] &= \gamma \sum_{\ell=1}^{L-1} \left(a_x^2[\ell] + a_x^2[-\ell]\right) \\
&= \frac{\gamma}{L} \sum_{i=0}^{L-1} \sum_{\ell=1}^{L-1} x[i] \left(x[i+\ell] + x[i-\ell]\right) \\
&= \frac{\gamma}{L} \sum_{i=0}^{L-1} \sum_{j=0, j\neq i}^{L-1} x[i]x[j],
\end{aligned}$$

where the last equality is obtained by noting that, given the summation bounds, the set of pairs $(i, i \pm \ell)$ and (i, j) are the same over the valid range $\{0, \ldots, L-1\}^2$. To conclude, notice that $a_y^2[0] = \frac{\gamma}{L}\sum_{i=0}^{L-1} x[i]^2 + \sigma^2$ and combine.

D Proof of Proposition 4.3

We prove that both σ and γ are identifiable from the observed first three moments of y. For convenience, we work with $\beta = \gamma/L$ rather than γ itself. To this end, we construct two quadratic equations satisfied by β and whose coefficients can be computed from observable quantities. Then, we show that these equations are independent, and hence that β is uniquely defined. Given β, we can estimate σ using Proposition 4.2.

Throughout the proof, it is important to distinguish between observed and unobserved values. We denote the observed values by E_i and a_y^1, a_y^2, a_y^3. We use F_i to denote functions of the signal's autocorrelations (which are not directly observable).

Recall that $a_y^1 = \beta(\mathbf{1}^T x)$ and $a_y^2[0] = \beta\|x\|^2 + \sigma^2$, where $\mathbf{1} \in \mathbb{R}^L$ is the vector of all-ones and $\|x\| = \sqrt{x[0]^2 + \cdots + x[L-1]^2}$ is the 2-norm. Consider the product E_1:

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

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 also be estimated from a_y^3, while taking the scaling and bias terms into account. This yields another observable, E_2:

$$E_2 = a_y^3[0,0] + \sum_{j=1}^{L-1}(a_y^3[j,j] + a_y^3[0,j]) = L\beta F_1 + (2L+1)\sigma^2 a_y^1. \tag{D.2}$$

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 $E_3 = a_y^2[0] + 2\sum_{\ell=1}^{L-1} a_y^2[\ell]$; recall from Proposition 4.2:

$$\sigma^2 = E_3 - (a_y^1)^2/\beta. \tag{D.4}$$

Plugging into (D.3) and rearranging, we get a first quadratic equation in β,

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

where

$$\begin{aligned}
\mathcal{A} &= E_2 - (2L+1)a_y^1 E_3, \\
\mathcal{B} &= -E_1 + (2L+1)(a_y^1)^3 + a_y^1 E_3, \\
\mathcal{C} &= -(a_y^1)^3.
\end{aligned}$$

Importantly, these coefficients are observable quantities. As we assume throughout this proof that x has nonzero mean, $a_y^1 \neq 0$ and we conclude that this equation is non-trivial.

Next, we derive the second quadratic equation for β. We notice that

$$E_4 = \frac{1}{L}(a_y^1)^3 = \frac{1}{L}\beta^3(\mathbf{1}^T x)^3 = \beta^3 F_2, \tag{D.6}$$

where $F_2 = \frac{1}{L}(\mathbf{1}^T x)^3$, and we can work out that:

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

Once again, F_2 can be estimated from a_y^3, taking bias and scaling into account:

$$E_5 = a_y^3[0,0] + 3\sum_{j=1}^{L-1}\left(a_y^3[j,j] + a_y^3[0,j]\right) + 6\sum_{1\le i<j\le L-1} a_y^3[i,j] = L\beta F_2 + (6L-3)\sigma^2 a_y^1. \tag{D.7}$$

Consider the following ratio:

$$\frac{E_5}{E_4} = \frac{L}{\beta^2} + \frac{(6L-3)\sigma^2 a_y^1}{E_4}.$$

From the latter, we deduce:

$$\sigma^2 = \frac{E_5}{a_y^1(6L-3)} - \frac{LE_4}{\beta^2 a_y^1(6L-3)}.$$

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

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

where

$$\begin{aligned}\mathcal{D} &= E_3 - \frac{E_5}{a_y^1(6L-3)},\\ \mathcal{E} &= -(a_y^1)^2,\\ \mathcal{F} &= \frac{LE_4}{a_y^1(6L-3)}.\end{aligned}$$

It is also non-trivial since $E_4 \neq 0$.

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 ratios between coefficients differ. From (D.5) and (D.1), we have:

$$\frac{\mathcal{B}}{\mathcal{C}} = \frac{E_1 - (2L+1)(a_y^1)^3 - a_y^1 E_3}{(a_y^1)^3} = \frac{a_y^2[0] - (2L+1)(a_y^1)^2 - E_3}{(a_y^1)^2}.$$

In addition, using (D.6),

$$\frac{\mathcal{E}}{\mathcal{F}} = \frac{(3-6L)(a_y^1)^3}{LE_4} = 3 - 6L.$$

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

$$a_y^2[0] - (2L+1)(a_y^1)^2 - E_3 = (a_y^1)^2(3-6L).$$

Rewriting the identity in terms of x and dividing by β we get:

$$4(L-1)\beta(\mathbf{1}^\top x)^2 - (\mathbf{1}^\top x)^2 + \|x\|^2 = 0. \tag{D.9}$$

For generic x, this polynomial equation is not satisfied so that the quadratic equations are independent. Furthermore, from the inequality $L\|x\|^2 \geq (\mathbf{1}^\top x)^2$ it follows immediately that the equations must be independent so long as

$$\beta > \frac{1}{4L}.$$

E Proof of Proposition 4.5

We first note that γa_x^1, $\gamma a_x^2[\ell] + \sigma^2\delta[\ell]$, and $\gamma a_x^3[\ell_1, \ell_2]$ can be computed directly form the observed autocorrelations (3.11), (3.12) and (3.13). Indeed, recovering γa_x^1 and $\gamma a_x^2[\ell] + \sigma^2\ell[\ell]$ is immediate from (3.11) and (3.12), and $\gamma a_x^3[\ell_1, \ell_2]$ then follows from

$$\gamma a_x^3[\ell_1, \ell_2] = a_y^3[\ell_1, \ell_2] - (a_y^1)^3 - a_y^1 \cdot \big(a_y^2[\ell_1] + a_y^2[\ell_2] + a_y^2[\ell_1 - \ell_2] - 3(a_y^1)^2\big). \tag{E.1}$$

Let us assume that x is generic. Indeed, we observe the product:

$$\begin{aligned}
L^2(\gamma a_x^1)(\gamma a_x^2[1]) &= \gamma^2 \left(\sum_i x[i]\right)\left(\sum_j x[j]x[j+1]\right) \\
&= \gamma^2 \sum_j \sum_i x[i]x[j]x[j+1] \\
&= \gamma^2 \sum_j \sum_\ell x[j+\ell]x[j]x[j+1] \\
&= \gamma^2 \sum_{\ell \geq 0} \sum_j x[j+\ell]x[j]x[j+1] + \gamma^2 \sum_{\ell < 0} \sum_j x[j+\ell]x[j]x[j+1] \\
&= \gamma^2 \sum_{\ell=0}^{L-1} a_x^3[1,\ell] + \gamma^2 \sum_{\ell > 0} \sum_j x[j-\ell]x[j]x[j+1] \\
&= \gamma^2 \sum_{\ell=0}^{L-1} a_x^3[1,\ell] + \gamma^2 \sum_{\ell > 0} \sum_j x[j]x[j+\ell]x[j+\ell+1] \\
&= \gamma \left(\sum_{\ell=0}^{L-1} \gamma a_x^3[1,\ell] + \sum_{\ell=1}^{L-2} \gamma a_x^3[\ell, \ell+1] \right).
\end{aligned} \tag{E.2}$$

Since we also observe $\sum_{\ell=0}^{L-1} \gamma a_x^3[1,\ell] + \sum_{\ell=1}^{L-2} \gamma a_x^3[\ell,\ell+1]$, we can form the ratio and solve for γ.

$$\gamma = L\frac{(\gamma a_x^1)(\gamma a_x^2[1])}{\sum_{\ell=0}^{L-1}\gamma a_x^3[1,\ell] + \sum_{\ell=2}^{L-1}\gamma a_x^3[\ell,\ell+1]}, \tag{E.3}$$

Then, similarly to Appendix C, one can solve for σ^2:

$$\sigma^2 = a_y^2[0] + 2\sum_{\ell=1}^{L-1} a_y^2[\ell] - \frac{L(a_y^1)^2}{\gamma} - (2L-1)(a_y^1)^2. \tag{E.4}$$

Once γ and σ were computed, one can recover x from a_x^2 and a_x^3 by Proposition 4.1.