The Annals of Applied Statistics
2011, Vol. 5, No. 4, 2572–2602
DOI: 10.1214/11-AOAS479
© Institute of Mathematical Statistics, 2011

SPARSE APPROXIMATIONS OF PROTEIN STRUCTURE FROM NOISY RANDOM PROJECTIONS[1]

BY VICTOR M. PANARETOS AND KJELL KONIS

Ecole Polytechnique Fédérale de Lausanne

Single-particle electron microscopy is a modern technique that biophysicists employ to learn the structure of proteins. It yields data that consist of noisy random projections of the protein structure in random directions, with the added complication that the projection angles cannot be observed. In order to reconstruct a three-dimensional model, the projection directions need to be estimated by use of an ad-hoc starting estimate of the unknown particle. In this paper we propose a methodology that does not rely on knowledge of the projection angles, to construct an objective data-dependent low-resolution approximation of the unknown structure that can serve as such a starting estimate. The approach assumes that the protein admits a suitable sparse representation, and employs discrete L^1-regularization (LASSO) as well as notions from shape theory to tackle the peculiar challenges involved in the associated inverse problem. We illustrate the approach by application to the reconstruction of an E. coli protein component called the *Klenow fragment*.

1. Introduction. The structure of biological macromolecules is at the heart of the quest to understand life in purely physical terms, and thus is fundamental to any biophysical project. A key element in solving the structure of a protein is to be able to visualize the protein in three dimensions, both in terms of exterior shape as well as of interior variations. This is a challenging task given the microscopic scale of the structures we wish to access, which can be less than a nanometer wide. The mechanisms which enable us to gain structural information will typically provide indirect knowledge (posing *inverse problems*), which will have to be translated into initial structural terms in a mathematically sound way. Such mechanisms include X-ray crystallography and electron microscopy, among others [Drenth (1999), Glaeser et al. (2007)]. The electron microscope (Figure 1) in particular, is a powerful tool that possesses important advantages over its "competitors," such as high scattering power and the retainment of phase information [Chiu (1993), Henderson (2004)]. It allows the retrieval of sufficiently detailed three-dimensional representations to be able to deduce atomic-level representations of the macromolecules of interest: the chemical structure of the particle of interest (i.e., the amino acid sequence) can be fitted (docked) into the density map

Received March 2010; revised April 2011.

[1]Supported in part by a European Research Council Starting Grant Award.

Key words and phrases. Statistical tomography, electron microscopy, single particles, nearly black object, LASSO, deconvolution, Roman surface.

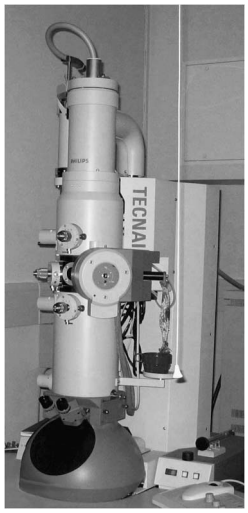


FIG. 1. *A transmission electron microscope at the MRC Laboratory of Molecular Biology, Cambridge, UK.*

produced via electron microscopy to obtain the complete three-dimensional folding of the particle [Glaeser et al. (2007)].

At the most basic level, the structure of a protein is determined by the spatial configuration of its constituent atoms. The negatively charged electrons of each atom create a field of electric potential surrounding the atom, and when combined, these electric potentials create a density of potential $\rho(x, y, z)$ in space (that can be thought of as a probability density function). This is called the *shielded Coulomb potential density*, and we usually think of a particle as being one and the same as its potential density (it is this density that we seek in order to dock the amino acid sequence of the particle and completely understand its structure). When placed under an electron beam, the potential ρ causes a reduction to the beam intensity due to electron scattering. If the beam is in the z-direction, then the *Abbe image formation theory* [Glaeser et al. (2007)] stipulates that the reduced intensity recorded is approximately given by

$$\int_{-\infty}^{+\infty} \rho(x, y, z)\, dz + \text{noise}.$$

Essentially, knowledge of the optical density recorded on the film corresponds to knowledge of the two-dimensional marginal density of ρ in the z-direction, except for some minor optical effects such as astigmatism, defocus, etc. If we were able to obtain multiple such measurements on the same particle from various beam directions, then determination of the three-dimensional density would amount to the solution of a noisy tomography problem with random projection angles. This type of problem is well understood and has been extensively studied in the statistical literature, both methodologically [e.g., Vardi, Shepp and Kaufman (1985);

Silverman et al. (1990); Green (1990); Jones and Silverman (1989)] and theoretically [e.g., Johnstone and Silverman (1990), O'Sullivan (1995)], principally in its positron emission tomography (PET) version.

It is impossible, however, to image the same particle under many angles because extended exposure to the electron beam will cause chemical bonds of the particle to break, and thus will alter the structure of the specimen. A means to surpass this difficulty is to crystallize many identical particles, and thus distribute the electron dose over multiple occurrences of the same structure, but the crystallization process is usually cumbersome, time-consuming, and unpredictably varying for different types of particles [Glaeser (1999)].

Single particle cryo-electron microscopy is a technique of electron microscopy that avoids the process of crystallization [e.g., Glaeser (1999)], and, as such, it is increasingly popular as a structure determination tool in structural biology. The idea is to image a large number of unconstrained particles in solution. The particles rotate and diffuse freely in solution, and are then rapidly vitrified, having assumed various different random orientations. After preliminary processing, the data yielded are essentially a number of noisy versions of the projected potential densities, at orientations both *random* and *unknown*.

Traditional tomographic techniques break down in this setting, as these crucially depend on the knowledge of the projection angles. In order to be able to put these techniques to use, biophysicists attempt to estimate the unobservable projection angles [Frank (1999)]. To this aim, they typically assume *a completely specified* low resolution form for the unknown density. This model often relies on knowledge on the structure of the particle gained either from other experiments or from an ad hoc examination of the projections by eye. In some cases, this model can be derived from the data using the so-called *projection-slice theorem* [Deans (1993)] for Radon transforms, but the success of this approach will depend on the level of noise in an image. Once such a prior is given, the unknown angles are considered as parameters to be estimated. When a set of angles is estimated, they are used in order to obtain a traditional tomographic reconstruction [Natterer (2001), Deans (1993)], and update the starting model [Frank (1999), Glaeser et al. (2007)]. The procedure is then iterated until it stabilizes. Broad classes of such *refinement* methods include the so-called *projection matching method* [Penczek, Grassucci and Frank (1994)] and *3D Radon transform method* [Rademacher (1994)]. In the first approach, the prior model is projected over a wide range of directions, obtaining so-called *re-projections*. Each data projection is then cross-correlated with each re-projection, and is assigned an angle corresponding to the angle of that re-projection which produced the highest cross-correlation. The 3D Radon approach is essentially equivalent, the only difference being that it focuses on the Radon transform rather than on the X-ray transform (which are of course very closely related). Variations of these approaches also exist that try to "integrate out" the angles rather than estimate them: treating them as unobservable random variables

(missing data) and using an approach based on the EM algorithm (EM here standing for Expectation–Maximization), initialized again by some prior model for the structure of the particle [Sigworth (1998), Bern, Chen and Wong (2005)]. Indeed, this latter approach draws interesting parallels to the methodology of Vardi, Shepp and Kaufman (1985) in the case of positron emission tomography. As already mentioned, though, what is common to any of these strategies is that they require a completely specified initial estimate for the structure. In cases where previous structural information is not available, the level of noise is relatively high, and a naked eye examination is either infeasible (e.g., when the particle has no symmetries) or would best be avoided, it is natural to seek approaches to obtaining "objective" initial models directly from the data, in order to then initialize approaches such as those mentioned above.

The purpose of this paper is to develop statistical tools that will enable the construction of a data-dependent starting model in the noisy setting encountered in practice. If the starting model is to depend only on the data at hand, its construction will have to bypass the unknown angles, thus requiring the approximate solution of a tomographic problem that has a second layer of ill-posedness. Nevertheless, it was seen in Panaretos (2009) that a consistent formal estimator for the *shape* of the particle may be constructed. However, the problem of the actual construction of an estimate in a practical situation still remained open, as the formal estimators introduced in Panaretos (2009) are only implicitly defined. Their construction requires the solution of further inverse problems, with severe instabilities due to the presence of noise, and the approximate nature of the modeling framework. In this paper we propose a framework for implementing estimators such as those proposed in Panaretos (2009) under sparsity constraints. Our approach combines L^1-regularization using Least Angles Regression with the special geometry of the sample space to yield a procedure applicable to actual electron microscope data. We illustrate the approach through an artificial example and also by application to noisy single particle projections of the so-called *Klenow fragment*, a large protein fragment that is produced during DNA polymerase interactions in E. coli. The paper is structured as follows. Section 2 provides a statistical formulation of the problem, and some relevant background. Section 3 introduces the modeling framework in which sparsity is to be understood. Our approach is presented in Section 4 and illustrated on an artificial sparse density. Finally, the method is applied to single Klenow particles, and an initial sparse approximation of the structure is obtained in Section 6. Some concluding remarks are made in Section 7.

2. Statistical formulation. From the statistical perspective, the problem can be phrased as follows. We wish to recover a compactly supported probability density function $\rho(\mathbf{x})$, $\mathbf{x} = (x_1, x_2, x_3)^\top \in \mathbb{R}^3$, given noisy discrete images of N random projections,

$$\breve{\rho}_n(x, y) = \int_{-\infty}^{+\infty} \rho(U_n \mathbf{x})\, dx_3, \qquad n = 1, \ldots, N,$$

where $\{U_n\}_{n=1}^N$ is a collection of i.i.d. random rotation matrices distributed according to normalized Haar measure on $\mathsf{SO}(3)$, the group of rotations in $\mathbb{R}^3$, that is,

$$U_n^\top U_n = I \qquad \text{a.s.}, \qquad \det(U_n) = 1 \qquad \text{a.s.}$$

and

$$WU_n \stackrel{d}{=} U_n V \qquad \forall W, V \in \mathsf{SO}(3).$$

The N discrete noisy profile images $\{P_n\}_{n=1}^N$ are obtained by sampling the projections $\{\breve{\rho}_n\}_{n=1}^N$ on a regular $T \times T$ lattice, subject to corruption by additive noise,

$$P_n(i, j) = \breve{\rho}_n(x_i, y_j) + \varepsilon_n(i, j), \qquad i, j = 1, \ldots, T.$$

It will be assumed that the noise arrays are independent, white and Gaussian, $\varepsilon_n(i, j,) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$. The (more or less) standard problem of tomography would be described by

$$\text{Recover } \rho(\mathbf{x}) \text{ given } \{(P_n, U_n)\}_{n=1}^N. \tag{2.1}$$

However, in the single particle setup, the rotations $\{U_n\}$ are unobservable, leading to the perturbed problem

$$\text{Recover } \rho(\mathbf{x}) \text{ given } \{P_n\}_{n=1}^N. \tag{2.2}$$

The difference between these two problems is fundamental. Every established technique for solving problem 2.1 (e.g., based on singular value decomposition, likelihood, smoothed backprojection and Fourier methods) crucially depends on the knowledge of the projection directions $\{U_n\}$. In the absence of these directions, the estimation problem is not even well defined: it is easy to see that the density $\rho(\mathbf{x})$ is unidentifiable, since any rotated/reflected version $\rho(Q\mathbf{x})$, with $Q^\top Q = I$, will generate data with the same distributional properties. Intuitively, this means that one cannot recover an exact coordinate system for the density. Although this is conceptually obvious, it can be a serious hurdle to statistical estimation: for example, if one wishes to parametrize the unknown density using a Fourier expansion, the Fourier coefficients will not be invariant to changes of the coordinate system.

Nevertheless, the *shape* of the density ρ can potentially be recovered [Panaretos (2009)]. The shape of ρ, denoted $[\rho]$, encodes the totality of characteristics of ρ that are invariant with respect to the coordinate system

$$[\rho] = \{\rho(U\mathbf{x}) : U \in \mathsf{O}(3)\},$$

where we denote the group of orthogonal transformations in $\mathbb{R}^d$ by $\mathsf{O}(d)$. Furthermore, it was seen in the same paper that the shape of the projection $\breve{\rho}_n$, $[\breve{\rho}_n] = \{\breve{\rho}_n(U\mathbf{x}) : U \in \mathsf{O}(2)\}$, constitutes a sufficient statistic for $[\rho]$. Hence, identifiability combined with the sufficiency principle would lead one to consider estimating $[\rho]$ on the basis of estimators depending on the data solely through their shape characteristics $[\breve{\rho}_1], \ldots, [\breve{\rho}_n]$.

Unfortunately, any likelihood-type approach turns out to be completely intractable in this setup. However, the feasibility of extracting an estimator from the random projections, without any recourse to the angular component, suggests that one might consider techniques that yield inefficient estimators that are nevertheless "efficient enough" to serve as a starting model for an iterative procedure that estimates angles, conducts traditional tomography, and iterates until the reconstruction stabilizes. Less formally, one can set to obtain a *rough initial approximation* that nevertheless captures the essential features of the object that are required to obtain a first set of angular estimates. In the next section we formulate these approximations through a class of *sparse radial mixtures*. These provide, on the one hand, a means to fruitfully parametrize the notion of shape, and, on the other hand, a natural way to impose sparsity.

3. Sparse approximations by radial mixtures. The key to our approach is the realization that approximating the unknown density by a relatively simple object suffices, if the aim is to obtain a starting reconstruction. Indeed, ad-hoc starting models used by biophysicists often consist of collections of solid spheres. The class of approximate models that we shall be pursuing is that of *radial mixtures*,

$$\rho(\mathbf{x}) = \sum_{k=1}^{K} q_k \phi(\mathbf{x} - \boldsymbol{\mu}_k), \qquad K \geq 1, \qquad \{\boldsymbol{\mu}_k\} \subset \mathbb{R}^3, \qquad q_k > 0, \tag{3.1}$$
$$\sum_{i=1}^{K} q_k = 1$$

with $\phi(\cdot)$ a radial probability density function on $\mathbb{R}^3$ (e.g., an isotropic Gaussian density), that is,

$$\phi(\mathbf{y}) = \phi(U\mathbf{y}) \qquad \forall \mathbf{y} \in \mathbb{R}^3, U \in \mathsf{O}(3). \tag{3.2}$$

Radial mixtures comprise a flexible yet tractable class of models for density estimation [see, e.g., Hastie, Tibshirani and Friedman (2001), Chapter 6, Section 7]. The choice of this class is especially well suited to this problem, as it offers two technical advantages:

(1) Good behavior under rotation and projection: the rotated version of ρ according to $U \in \mathsf{SO}(3)$ is given by

$$(U\rho)(\mathbf{x}) = \rho(U^\top \mathbf{x}) = \sum_{k=1}^{K} q_k \phi(U^\top \mathbf{x} - \boldsymbol{\mu}_k) \overset{(3.2)}{=} \sum_{k=1}^{K} q_k \phi(\mathbf{x} - U\boldsymbol{\mu}_k),$$

that is, by a radial mixture of the same densities with the same mixing coefficients, but centered at the rotated location parameters $\{U\boldsymbol{\mu}_k\}$. The projected density at

orientation U will then be given by

$$\int_{-\infty}^{+\infty} (U\rho)(x_1, x_2, x_3)\, dx_3 = \sum_{k=1}^{K} q_k \int_{-\infty}^{+\infty} \phi(\mathbf{x} - U\boldsymbol{\mu}_k)\, dx_3$$
$$= \sum_{k=1}^{K} q_k \varphi\big(H(\mathbf{x} - U\boldsymbol{\mu}_k)\big),$$

where H is the identity matrix with its last row deleted, and φ is the (unique) two-dimensional marginal of ϕ,

$$H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \qquad \varphi(x_1, x_2) = \int_{-\infty}^{+\infty} \phi(x_1, x_2, x_3)\, dx_3.$$

(2) The possibility of a finite-dimensional parametrization of the shapes of ρ and of a projection $\breve{\rho}$ using the Gram matrix of the original and projected location parameters, respectively,

$$[\rho] = (\mathsf{Gram}(\{\boldsymbol{\mu}_k\}), \{q_k\}), \qquad [\breve{\rho}] = (\mathsf{Gram}(\{HU\boldsymbol{\mu}_k\}), \{q_k\}),$$

where for a collection of K vectors $\{\mathbf{w}_j\}_{j=1}^K$, $\mathsf{Gram}(\{\mathbf{w}_j\})$ represents the symmetric nonnegative matrix with (i, j)-element equal to $\langle \mathbf{w}_i, \mathbf{w}_j \rangle$.

In Kendall's Shape Theory, Gram matrices are employed as a coordinate system for the shape manifold induced by rigid motions [Kendall et al. (1999), Kendall and Le (2009)]. Note that if the vectors $\{\mathbf{w}_j\}_{j=1}^K$ are arranged as the columns of a $3 \times K$ matrix $\mathbf{W}$, then we may simply write $\mathsf{Gram}(W) = W^\top W$. This Gram matrix encodes all the invariant characteristics with respect to $\mathsf{O}(3)$ of the configuration $\{\mathbf{w}_j\}$, since it is invariant under orthogonal transformations of the generating vectors: for $B \in \mathsf{O}(3)$ we immediately see that $\mathsf{Gram}(BW) = W^\top B^\top BW = W^\top W = \mathsf{Gram}(W)$. Furthermore, given a Gram matrix of rank p, one can find K vectors in $\mathbb{R}^d$, $d \geq p$, with centroid zero whose pairwise inner products are given by that Gram matrix (in fact, the specification of such an ensemble amounts to merely solving nondegenerate lower triangular linear systems of equations). Therefore, for a given density ϕ (or projected density φ), the Gram matrices coupled with the corresponding mixing proportions comprise a complete description of the *shapes* of the original and projected densities, respectively.

A further importance of this parametrization is that it provides an interface with the finite-dimensional case, where projected shape is better understood [Panaretos (2006, 2008), Le and Barden (2010)]. In particular, it allows use of the following simple connection between projected shape and original shape:

THEOREM 3.1 [Panaretos (2009)]. *Let $\{\mathbf{w}_k\}_{k=1}^K$ be a configuration of K vectors in $\mathbb{R}^3$ and let U be a random element of $\mathsf{SO}(3)$ satisfying $WU \overset{d}{=} UV \overset{d}{=} U$, for any $W, V \in \mathsf{SO}(3)$. Then*

$$\mathbb{E}[\mathsf{Gram}(\{HU\mathbf{w}_k\}_{k=1}^K)] = \tfrac{2}{3}\mathsf{Gram}(\{\mathbf{w}_k\}_{k=1}^K),$$

where H is the 3×3 identity matrix with its last row deleted.

Based on this result, Panaretos (2009) proved, for known K and under the assumption that for $i \neq j$ we have $q_i \neq q_j$, that the hybrid maximum likelihood/method of moments estimator $\hat{\rho}(\mathbf{x})$,

$$\hat{\rho}(\mathbf{x}) = \sum_{k=1}^{K} \hat{q}_k \phi(\mathbf{x} - \hat{\boldsymbol{\mu}}_k) \tag{3.3}$$

is consistent modulo $\mathsf{O}(3)$, as the resolution of each image $T \times T$ and the number of projections N grows. Here, $\hat{\boldsymbol{\mu}}_k$ is any collection of K vectors in $\mathbb{R}^3$ with Gram matrix

$$\hat{G} = \frac{3}{2N} \sum_{n=1}^{N} \mathsf{Gram}(\{\widehat{HU_n \boldsymbol{\mu}_k}\}_{k=1}^{K}).$$

The $\{\hat{q}_k\}$ and $\{\widehat{HU_n \boldsymbol{\mu}_k}\}$ are maximum likelihood estimators of the common mixing proportions and the individual projected location parameters for each profile, stemming from the loglikelihood

$$\begin{aligned} &\ell(\{q_k\}, \{HU_n \boldsymbol{\mu}_k\}) \\ &\quad \propto -\frac{1}{NT^2} \sum_{n=1}^{N} \sum_{i=1}^{T} \sum_{j=1}^{T} \left\{ P_n(i, j) - \sum_{k=1}^{K} q_k \varphi\big(H(\mathbf{x} - U_n \boldsymbol{\mu}_k)\big) \right\}^2. \end{aligned} \tag{3.4}$$

The latter loglikelihood stems from the independence between projections and between pixels, and the Gaussian assumption on the noise. Notice that each of the vectors $HU_n \boldsymbol{\mu}_k$ is treated as a separate parameter.

4. Reconstruction of a nearly black protein. Although the latter development provides a consistent solution to the problem from a theoretical perspective, it does not provide a solution to the practical problem. The estimator defined formally as $\hat{\rho}$ cannot be readily constructed given a data set of projections, as it is implicitly defined through the likelihood equation (3.4). The optimization of the objective function given by the latter equation is a separate challenge of its own—not in terms of computational tractability, but in terms of accuracy and stability. Among the reasons for this is the dimension of the search space (which is $2KN + K$). This can to some extent be mitigated, if one is to obtain separate likelihood estimates within each projection image, breaking the overall problem into N independent problems, each with search space dimension $3K$ (and then seek a global estimate for the mixing proportions). But, more importantly, it is the highly nonlinear form of the objective function (3.4) and the possibility of parameters being almost unidentifiable (when projected means fall close to one another) that presents the most serious complications in a practical reconstruction. The objective

function admits multiple local optima that are in addition unstable to minor perturbations of the noise term (the search surface has multiple relatively flat peaks). This instability of the nonlinear likelihood function is a manifestation of an inherent ill-posedness, which is clearly revealed once we re-express the problem as a collection of N deconvolution problems to be solved given discrete data:

$$\{(\widehat{HU_n\boldsymbol{\mu}}_1, \ldots, \widehat{HU_n\boldsymbol{\mu}}_K); (\hat{q}_1, \ldots, \hat{q}_K)\}$$
$$:= \arg\min \left\| P_n(i, j) - \varphi(\mathbf{x}) * \sum_{k=1}^{K} q_K \delta(\mathbf{x} - HU_n\boldsymbol{\mu}_k) \right\|_2^2.$$

Here, δ denotes Dirac's delta function. In this format, the problem is seen to be a linear inverse problem in the unknown function $h(\mathbf{x}) := \sum_{k=1}^{K} q_K \delta(\mathbf{x} - HU_n\boldsymbol{\mu}_k)$. The solution of such a problem would require regularization through the imposition of some norm penalty on the function h that we wish to recover. This cannot work here because the unknown function to be recovered is a Dirac comb—which is not an element of L^2 and hence does not allow Hilbert space regularization methods. This is simply a different way of saying that the problem cannot be treated as a linear one: if we are interested in the locations themselves (the spikes), the problem is fundamentally nonlinear.

Our basic idea to tackle this problem is to turn the drawback of "singularity" into an advantage by transforming the nonlinear problem into a linear problem through discretization of the solution search space. While the function we seek to recover is not well behaved when considering it as defined over a continuous domain, it reduces to a very simple object once thought of as a high-dimensional vector. This simplicity is reflected through *sparseness*.

In particular, suppose that we relax our search problem and ask to recover the image pixels that contain spikes, rather than the precise spike locations themselves. Choose a projection $n = n_0$ and omit the index for simplicity. Then, the problem can be approximately expressed via the following linear equation:

$$\mathcal{P}_{T^2\times 1} = \mathcal{X}_{T^2\times T^2}\beta_{T^2\times 1} + \varepsilon_{T^2\times 1}. \tag{4.1}$$

Here, $\mathcal{P}$ is the vectorized image, obtained by stacking the columns of the image matrix P. The matrix $\mathcal{X}$ is constructed as follows: the pth column of $\mathcal{X}$ is a vectorized (by column) image (which we call a *base profile*) generated by placing a single density φ at the center of the jth pixel. More precisely, let u_j be the center of the jth pixel. Then, the pth column of $\mathcal{X}$ is given by the vector

$$\{\varphi(u_j - u_p)\}_{j=1}^{T^2},$$

where j runs so that the pixels are arranged in column-major order. The parameter vector β is a $T^2 \times 1$ vector containing at most K nonzero entries:

$$\|\beta\|_0 \le K.$$

These entries reveal which pixels contain spikes. Since the entire density must be contained within the image boundaries, we a priori fix entries of β that correspond to pixels near the boundary to be zero (or, alternatively, drop the corresponding columns from the matrix $\mathcal{X}$). Finally, ε is an i.i.d. Gaussian mean-zero error vector.

Thus, in this discrete form, the problem has been reduced to a model selection problem in linear regression: we wish to recover the nonzero entries of β, which will reveal the approximate spike locations. The key observation, of course, is that β is *sparse*: we expect that $K \ll T^2$, so that we are attempting to recover a *nearly black object* in the terminology of Donoho et al. (1992). It therefore seems quite appropriate to employ a shrinkage estimator in this setting. There are various possibilities, but it is the LASSO [Tibshirani (1996)] that arises as the most natural one in the setting of this problem [see also Hastie, Tibshirani and Friedman (2001)]. Specifically, observe that the nonzero entries of β should be equal to the mixing proportions corresponding to the respective spikes (in case multiple spikes fall within the same pixel, then it would be the sum of the corresponding mixing proportions). Since the object in question is a probability density, we must have

$$\|\beta\|_1 = \sum_{i=1}^{T^2} |\beta_i| = \sum_{i=1}^{K} q_i = 1,$$

and we are naturally led to the following L^1-constrained least squares problem in β:

$$\min \|\mathcal{P} - \mathcal{X}\beta\|_2^2 \qquad \text{subject to } \|\beta\|_1 = 1. \tag{4.2}$$

Since the specifications of the problem determined the precise value for the L_1 penalty, there is even no need to perform cross-validation to determine the bandwidth parameter. In practice, of course, the total mass m of the density will not be precisely known, and may slightly differ from projection to projection. However, an approximate value $\hat{m}$ can be easily estimated. Therefore, one can employ the Least Angle Regression (LARS) algorithm [Efron et al. (2004)] to compute the whole LASSO path, and calibrate the results around a small neighborhood of the bandwidth corresponding to the approximate mass $\hat{m}$.

In order to illustrate the details (and effectiveness) of this discrete regularization approach, we revisit an artificial example presented in Panaretos (2009), where a three-dimensional mixture of four Gaussian kernels was to be recovered given its projections at randomly chosen directions. The pseudo-particle potential density in three dimensions was given by

$$\rho(u) = \sum_{k=1}^{4} \frac{q_k}{\sigma^3(\sqrt{2\pi})} \exp\left\{-\frac{(u-\boldsymbol{\mu}_k)^\top (u-\boldsymbol{\mu}_k)}{2\sigma^2}\right\}, \tag{4.3}$$

where $u = (u_x, u_y, u_z)^\top \in \mathbb{R}^3$, $\sigma = 0.46$, $q_1 = 0.18$, $q_2 = 0.26$, $q_3 = 0.21$, $q_4 = 0.35$, and with $\{\boldsymbol{\mu}_k\}$ given by $\boldsymbol{\mu}_1 = (0, 0.8, -0.3)^\top$, $\boldsymbol{\mu}_2 = (0.7, -0.4, -0.3)^\top$,

$\boldsymbol{\mu}_3 = (-0.7, -0.4, -0.3)^\top$, $\boldsymbol{\mu}_4 = (0, 0, 0.8)^\top$. The corresponding signal-to-noise level for the projections (understood as the ratio of the signal to the noise variance) was at the level of 61 : 1.

The method employed in Panaretos (2009) to perform the deconvolutions required for the construction of the estimator was a direct spectral approach based on results on Toeplitz forms [Grenander and Szegö (1958), Pisarenko (1973)]. The approach performed well on noiseless projections, but would fail completely even with very small amounts of noise. This effect is easily anticipated as the Toeplitz form approach amounts to an approximate discrete version of deconvolution by unregularized inversion of Fourier coefficients—which is bound to be highly unstable in the presence of noise.

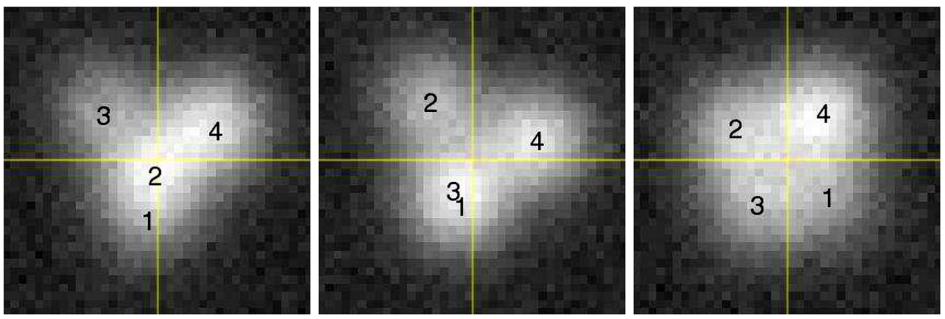
In order to produce a reconstruction in the presence of noise, we implement the LASSO deconvolution approach on the basis of $N = 150$ noisy random discrete profiles. The typical profile (Figure 2) is given by a discretized image $\mathcal{P} = \{P(i, j)\}$ defined as

$$P(i, j) = \sum_{k=1}^{K} q_k \varphi(u_{ij} | \tilde{\mu}_k, \sigma^2) + \varepsilon(i, j), \qquad i, j = 1, \ldots, 64, \tag{4.4}$$

where $\varepsilon(i, j)$ are i.i.d. Gaussian with mean 0 and standard deviation 10^{-4}; $\varphi(\cdot|\nu, \sigma^2)$ is a spherical bivariate Gaussian density with mean ν and variance σ^2; u_{ij} is the center of the (i, j)th image pixel; $\{\tilde{\mu}_k\}_{k=1}^K$ are the locations of the 4 (unobservable) projected means in that profile:

$$\tilde{\mu}_k := HU\boldsymbol{\mu}_k.$$

Since each image contains a region that is known a priori to be "empty" (i.e., does not contain a projected mean), we limit our interest to pixels in the complement of that region. Let $\mathcal{M}$ be the set of indices $p \in \{1, \ldots, T^2\}$ such that the pixel centers



(a) (b) (c)

FIG. 2. *Three random noisy profiles of the mixture* (4.3). *The digits indicate the true locations of the projected component means.*

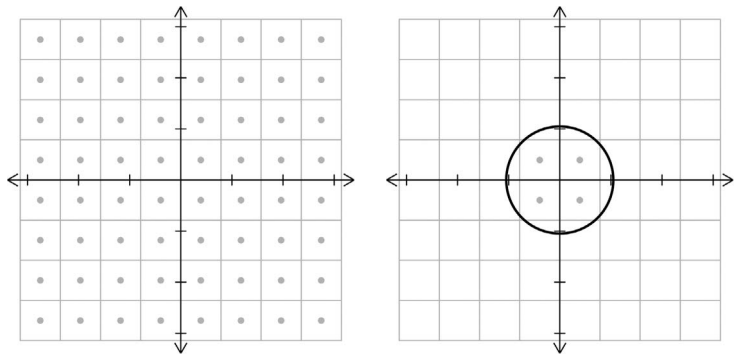


FIG. 3. *Illustration of the restriction of the support for a discrete profile.* (a) *A discrete profile with* $T = 8$. *The pixel centers* u_p *are denoted by gray dots.* (b) *The set of candidate means* $\mathcal{M}$ *for the convolution matrix* $\mathcal{X}$.

u_p satisfy $\|u_p\| < w$ (Figure 3). We call the elements of $\{u_p : p \in \mathcal{M}\}$ candidate means and build the convolution matrix $\mathcal{X}$ as

$$\mathcal{X}_{j,p} = \varphi(u_j | u_p, \sigma^2), \qquad j \in \{1, \ldots, T^2\},\ p \in \mathcal{M}.$$

The choice of the tuning parameter w is made so as to ensure that the base profiles integrate to (approximately) one. In the specific example, the choice $w = \pi/3$ is seen to be sufficient.

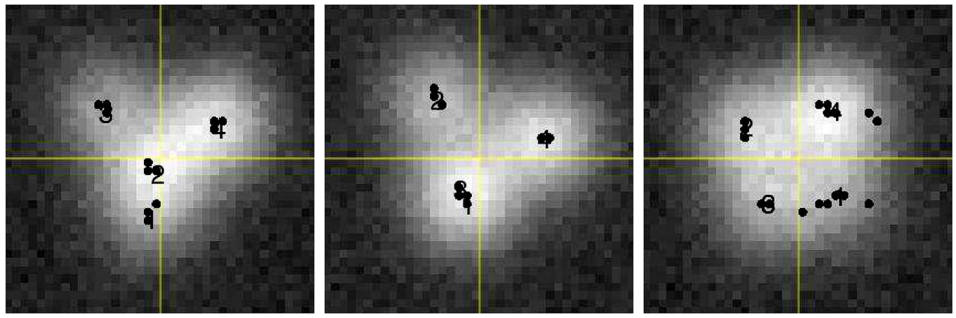
We used the LARS algorithm (in particular, the `lars` function in the lars package [Hastie and Efron (2011)] for the R Project for Statistical Computing [R Development Core Team (2011)]) in order to compute the complete regularization path (in t) for the LASSO problem

$$\min \|\mathcal{P} - \mathcal{X}\beta\|_2^2 \qquad \text{subject to } \|\beta\|_1 \le t, \tag{4.5}$$

and retained the parameter estimates $\hat{\beta}$ provided for t slightly less than the estimated mass $\hat{m}$ (to avoid overfitting). The latter was estimated by the average total intensity of the projections. Figure 4 depicts three characteristic noisy random profiles, along with the pixels the LASSO picked out as candidate locations for mean parameters. We denote the centers of these pixels as $\{u_p\}_{p \in \mathcal{A}}$, where

$$\mathcal{A} := \{p \in \mathcal{M} : \hat{\beta}_p \neq 0\}.$$

Since the true locations of the projected means will almost certainly not be contained in $\mathcal{M}$, and since the discrete representation of the convolution will only be approximate, a fit with precisely the right number K of nonzero parameters ($K = 4$ in this case) can rarely be achieved (i.e., $|\mathcal{A}| \neq K$). However, the nonzero parameters found by the LASSO will tend to bracket the locations of the projected means, as can be noticed in Figure 4. It therefore suffices to use a naive clustering rule to associate nonzero LASSO parameter estimates with projected means: if two pixels



(a) (b) (c)

FIG. 4. *The same three random profiles, with black dots indicating the locations of the nonzero LASSO coefficient estimates. The digits indicate the true locations of the projected component means.*

selected by the LASSO share either an edge or a corner, then they belong to the same cluster.

Let $\{\mathcal{C}_k\}_{k=1}^K$ denote the K clusters comprising $\mathcal{A}$, so that $\mathcal{A} = \biguplus_{k=1}^K \mathcal{C}_k$, where $\biguplus$ denotes a disjoint union. Then, the estimates of the locations of the projected means are computed by taking a weighted average of the locations of the nonzero lasso parameter estimates in each cluster using the parameter estimates as the relative weights,

$$\widehat{HU\boldsymbol{\mu}_k} = \frac{\sum_{p\in\mathcal{C}_k} \hat{\beta}_p \boldsymbol{u}_p}{\sum_{p\in\mathcal{C}_k} \beta_p}.$$

Further, the sum of the LASSO parameter estimates in each cluster provides an initial estimate of the mixing weight associated with that cluster. Once all of the mixing weights have been initially estimated, the final estimates are achieved by scaling the initial mixing weights so that they sum to $\hat{m}$, the estimated total mass of the particle,

$$\hat{q}_k := \frac{\sum_{p\in\mathcal{C}_k} \hat{\beta}_p}{\sum_{p\in\mathcal{A}} \hat{\beta}_p} \hat{m}.$$

Note here that this is the estimate of the mixing proportions stemming from a single profile (these will later be combined to produce a global estimate). To mitigate any bias in the mixing coefficients estimates introduced by choosing a constraint parameter less than 1, one may allow the constraint parameter to increase so long as the number of clusters remains constant. That is, the constraint parameter is increased either until it is equal to 1 or until any further increase would spawn a new cluster. Often, this results in one or more additional nonzero lasso parameter estimates joining the current clusters.

Once a set of estimated mixing proportions and mean locations has been obtained for each projection, these are used in order to construct the hybrid estimator (3.3). An intermediate step required is building the estimated Gram matrices

for each projection consistently. That is, we should ensure to the extent possible that estimated location parameters that correspond to the same three-dimensional mean should share the same index. For this reason, within each profile, the estimated location and mixing parameters are relabeled according to the ascending ordering of their mixing proportions (which were assumed to be distinct); that is, the indices are assigned so that

$$\hat{q}_1 < \hat{q}_2 < \cdots < \hat{q}_K .$$

Once the labels have been consistently assigned to the location estimates within each profile, one may obtain a single profile likelihood estimate for the mixing proportions, by solving the ordinary least squares problem obtained when plugging the estimated location parameters into the loglikelihood (3.4).

Finally, the estimated Gram matrices and the single set of estimated mixing proportions are used to construct the hybrid estimator (3.3). Figure 5 shows the original pseudo-particle in contrast with the estimated version. The reconstruction was based on 53 of the 150 profiles in our sample, for which four clusters were more or less clearly identifiable. In the majority of these profiles, the 4 clusters correspond to the component means. However, from time-to-time, one of the clusters was a false positive. In these cases, the smallest mixing weight was far smaller than typical for the sample. To further filter these profiles out, we rejected profiles with left-outlying mixing proportions (left outlying values were omitted when calculating the mixing weights in Table 1).

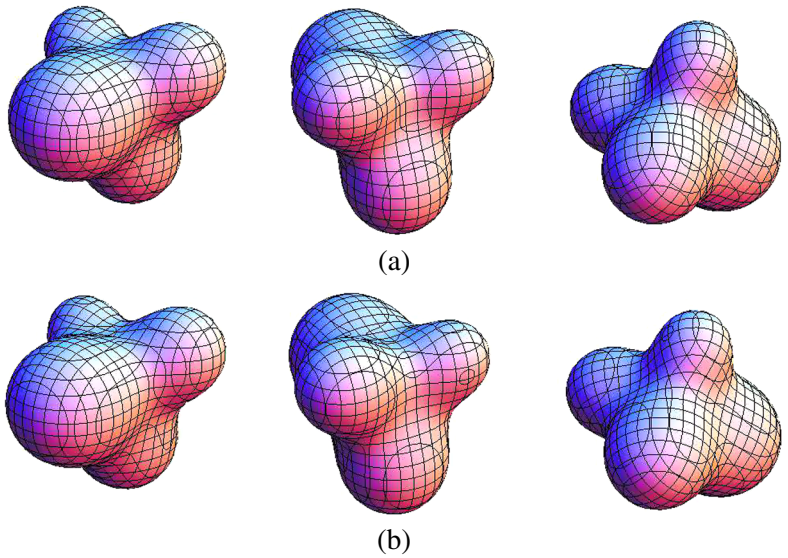


FIG. 5. *The actual pseudo-particle density and the estimated density.* (a) *A level surface of the true pyramid density from* 3 *different vantage points*. (b) *Corresponding level surfaces of the estimated pyramid density.*

TABLE 1
Estimated mixing weights

	$k=1$	$k=2$	$k=3$	$k=4$
True mixing weights: $\{q_k\}$	0.180	0.210	0.260	0.350
Estimate mixing weights: $\{\hat{q}_k\}$	0.170	0.210	0.263	0.357

The estimated Gram matrix and corresponding estimated location parameters are contrasted below with the true values (the estimated locations of components in 3 dimensions $\hat{\mu}$ can be computed by solving a simple system of linear equations):

$$G = \begin{pmatrix} 0.681 & -0.227 & -0.227 & -0.227 \\ -0.227 & 0.726 & -0.254 & -0.244 \\ -0.227 & -0.254 & 0.726 & -0.244 \\ -0.227 & -0.244 & -0.244 & 0.716 \end{pmatrix},$$

$$\hat{G} = \begin{pmatrix} 0.696 & -0.176 & -0.279 & -0.241 \\ -0.176 & 0.660 & -0.247 & -0.237 \\ -0.279 & -0.247 & 0.736 & -0.209 \\ -0.241 & -0.237 & -0.209 & 0.687 \end{pmatrix},$$

$$(\boldsymbol{\mu}_1 \quad \boldsymbol{\mu}_2 \quad \boldsymbol{\mu}_3 \quad \boldsymbol{\mu}_4) = \begin{pmatrix} 0.825 & -0.275 & -0.275 & -0.275 \\ 0.000 & 0.806 & -0.409 & -0.397 \\ 0.000 & 0.000 & 0.695 & -0.695 \end{pmatrix},$$

$$(\hat{\boldsymbol{\mu}}_1 \quad \hat{\boldsymbol{\mu}}_2 \quad \hat{\boldsymbol{\mu}}_3 \quad \hat{\boldsymbol{\mu}}_4) = \begin{pmatrix} 0.834 & -0.211 & -0.335 & -0.289 \\ 0.000 & 0.784 & -0.405 & -0.380 \\ 0.000 & 0.000 & 0.678 & -0.678 \end{pmatrix}.$$

Interestingly enough, the reconstruction procedure was not severely affected by higher levels of noise contamination. Even when the noise variance was increased by two orders of magnitude, leading to a 1 : 1 signal-to-noise ratio, the reconstructed version of the particle was not significantly perturbed (see Figure 6). Since it is the deconvolution step that is the most ill-posed aspect of our approach, this can be largely attributed to a noteworthy degree of stability exhibited by the LASSO as a means for deconvolution.

5. More on the geometry of the problem. The implementation of the LASSO based hybrid estimator to the almost black pseudo-particle of the previous section brings to the surface two potential issues that might arise when implementing the procedure to actual proteins (as will be done in Section 6). We consider these in the next two paragraphs.

5.1. *Using fewer projections.* The first point relates to the usability of all the profiles available. It was seen that several projections were not used because the

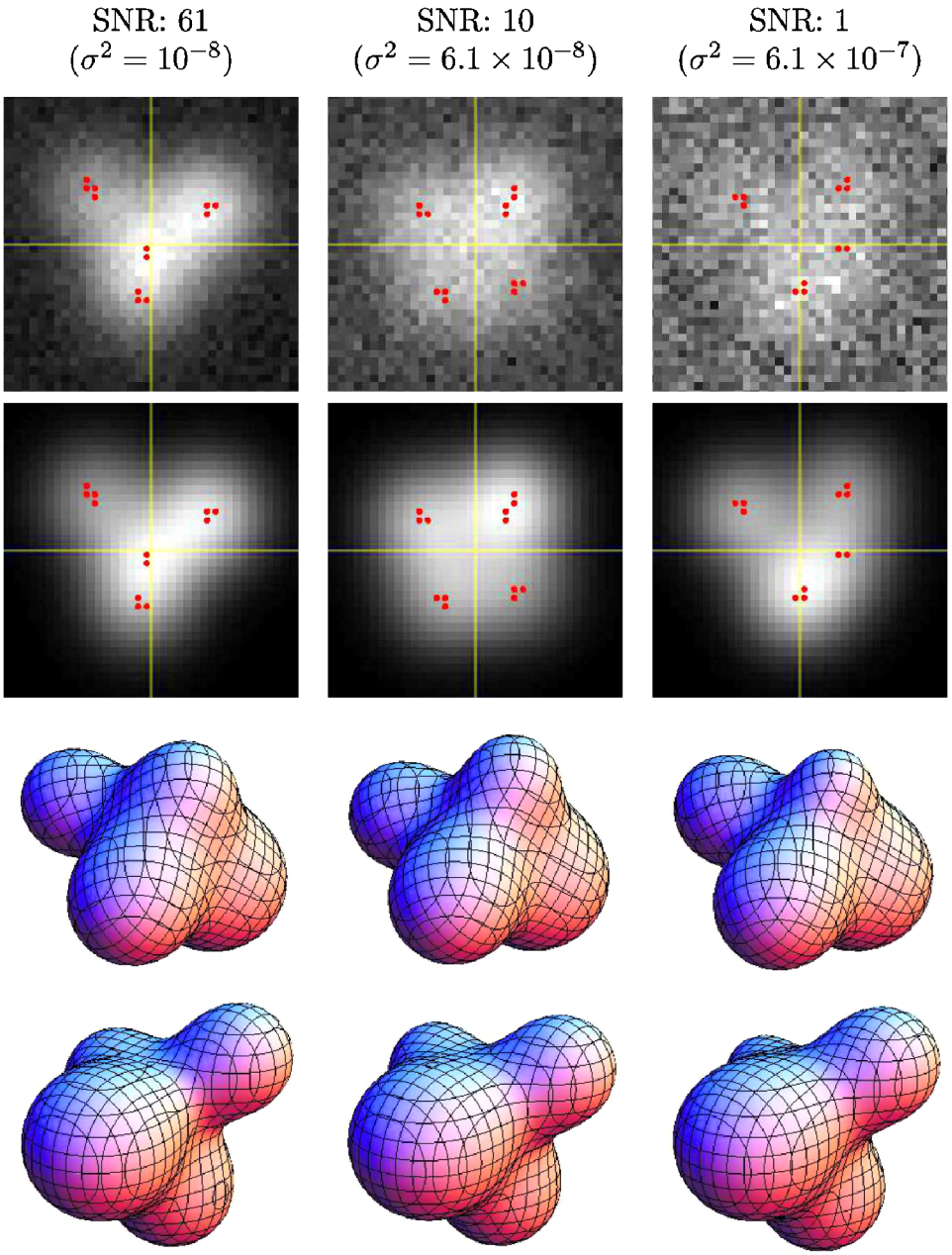


FIG. 6. *Reconstructions under different signal-to-noise scenarios. Each column corresponds to a different noise level. The first row presents a typical profile along with the candidate mean positions obtained via the LASSO. The second row presents the same profiles without any noise, and the corresponding candidate mean positions obtained via the LASSO. The last two rows present two different viewpoints of the final reconstruction obtained.*

viewing angles that they represented caused problems in the construction of the estimator. Nevertheless, the estimator constructed seemed to be rather successful. It is therefore natural to wonder if one could do with far fewer projections. This will become especially relevant in practical situations where a number of projections might not present well-identifiable mixture means. The answer is in the affirmative, that is, one can typically use a very small number of projections, as is illustrated by the next lemma.

LEMMA 5.1. *Let H_1, H_2, H_3 be projection matrices of rank 2 acting on $\mathbb{R}^3$ and $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$ be an ensemble of K vectors in $\mathbb{R}^3$. If the ranges of $I - H_1$, $I - H_2$, $I - H_3$ are pairwise orthogonal, then*

$$\mathsf{Gram}(\{\boldsymbol{\mu}_k\}_{k=1}^K) = \frac{1}{2}\sum_{i=1}^3 \mathsf{Gram}(\{H_i\boldsymbol{\mu}_k\}_{k=1}^K).$$

PROOF. Since the rank of the projection matrices involved is 2, we may find unit vectors $\{\mathbf{e}_i\}_{i=1}^3$ such that

$$H_i = (I - \mathbf{e}_i\mathbf{e}_i^\top), \qquad i = 1, 2, 3.$$

Furthermore, since the images of $I - H_1, I - H_2$, and $I - H_3$ are pairwise orthogonal, it must also be that $\{\mathbf{e}_i\}_{i=1}^3$ be pairwise orthogonal, thus constituting an orthonormal basis for $\mathbb{R}^3$. Letting V denote the $3 \times K$ matrix with $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$ as its columns, it follows that

$$\begin{aligned}\sum_{i=1}^3 \mathsf{Gram}(\{H_i\boldsymbol{\mu}_k\}_{k=1}^K) &= \sum_{i=1}^3 V^\top H_i^\top H_i V = V^\top\left(\sum_{i=1}^3 H_i\right)V \\ &= 3V^\top V - V^\top(\mathbf{e}_1\mathbf{e}_1^\top + \mathbf{e}_2\mathbf{e}_2^\top + \mathbf{e}_3\mathbf{e}_3^\top)V = 2V^\top V \\ &= 2\mathsf{Gram}(\{\boldsymbol{\mu}_k\}_{k=1}^K). \qquad \square\end{aligned}$$

Lemma 5.1 allows us to heuristically reinterpret the estimator $\hat{G}$ of the Gram component given by

$$\hat{G} = \frac{3}{N}\frac{1}{2}\sum_{n=1}^N \mathsf{Gram}(\{\widehat{HU_n\boldsymbol{\mu}_k}\}_{k=1}^K)$$

by thinking of it as grouping the data into $N/3$ triads of nearly orthogonal views, forming an estimator within each triad using Lemma 5.1, and then averaging these $N/3$ estimators.

It follows that, in principle, only a few random projections at unknown angles suffice to reconstruct a Gram matrix—provided that we can arrange them in groups that represent views that carry information from relatively different viewpoints. In practice, one cannot know whether projection angles are indeed orthogonal, since

they are unknown. However, one can try to identify classes of profiles that appear to be carrying significantly different profile information, and use these as a proxy. The procedure is illustrated in the next section.

5.2. *Consistent construction of Gram matrices*. The second issue that became apparent from the pseudo-particle example has to do with the consistent construction of the Gram matrices across different profiles. This construction hinges on the assumption that the mixing proportions are distinct. The formula defining the pseudo-particle guaranteed that the mixing proportions were indeed distinct, and allowed us to successfully construct the estimator. In practice, it is natural to expect situations where the mixing proportions for certain components are not significantly different, leading to instabilities in the construction of the estimated Gram matrices. To address this problem, we can take advantage of the special geometry of the problem and, in particular, the fact that the projections of a radial basis function cannot lie in a totally arbitrary subspace of the set of 2D radial basis functions: the locus of projections is highly constrained, a fact that may be exploited in order to assign mixing proportions in a way that attempts to respect these constraints. The constraints on the radial basis functions induce corresponding constraints on the *support* of the projected Gram matrices, forcing this support to depend crucially on the three-dimensional original Gram matrix (i.e., we are dealing with a *nonregular* problem). Specifically, a projected Gram matrix cannot be any arbitrary nonnegative definite symmetric matrix. The locus of possible projected Gram matrices $\mathcal{G}$ comprises a smooth surface in $\mathbb{R}^{k\times k}$ of (intrinsic) dimension 2. The idea is therefore that an arbitrary permutation of the entries of a projected Gram matrix induces an abrupt change in its location relative to $\mathcal{G}$, typically mapping it far from $\mathcal{G}$. In principle, we should thus be able to choose an arrangement of the entries of a projected Gram matrix so as to make it "closest" to the locus of "allowed" Gram matrices.

To be more precise, if V is any $3 \times k$ matrix such that $G = V^\top V$, then the projected Gram matrix at direction given by $\mathbf{e} = (e_1, e_2, e_3)^\top \in \mathbb{S}^2$ is defined as

$$G(e) = V^\top (I - \mathbf{e}\mathbf{e}^\top) V.$$

As e ranges over the unit sphere, the matrix $I - \mathbf{e}\mathbf{e}^\top$ ranges over the real projective space (the sphere with antipodal points identified). This real projective space can be visualized in three dimensions as the *Roman surface* (see Figure 7), using the mapping $\mathbb{S}^2 \ni (e_1, e_2, e_3)^\top \mapsto (e_2e_3, e_1e_3, e_1e_2)^\top$ [Apery (1987)]. The effect of pre-multiplying by $V^\top$ and post-multiplying by V is to stretch this Roman surface according to the singular values of V, rotate it by its left singular vectors and finally shift it [much like a full column rank $d \times n$ matrix transforms the sphere $\mathbb{S}^{n-1}$ into an $(n-1)$-dimensional ellipsoid in $\mathbb{R}^d$]. This can be seen directly by using Kronecker products:

$$\operatorname{vec}\{G(\mathbf{e})\} = \operatorname{vec}\{V^\top (I - \mathbf{e}\mathbf{e}^\top) V\} = (V^\top \otimes V^\top) \operatorname{vec}\{(I - \mathbf{e}\mathbf{e}^\top)\}.$$

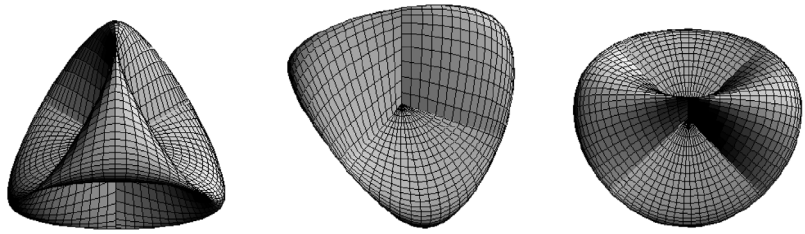


FIG. 7. *The Roman surface from three different vantage points.*

In practice, the estimated Gram matrices will not lie precisely on the locus $\mathcal{G}$ since their construction is subject to error. However, we expect that they should lie close to this surface. Therefore, given a Gram matrix that is determined up to a permutation of its entries, one can select the arrangement of its entries so as to minimize its distance from the underlying locus $\mathcal{G}$. Of course, the exact locus $\mathcal{G}$ will not be known in practice, as it is in bijective correspondence with the unknown three-dimensional Gram matrix G. However, an initial estimate of the surface $\mathcal{G}$ can be constructed using those projected Gram matrices for which correspondences are known. The procedure is illustrated in the next section, where we construct a sparse initial model for the potential density of a real biological particle.

6. Application: Sparse approximation of a Klenow fragment. We now turn to demonstrate our approach on noisy projections of an actual biological particle called the *Klenow fragment*. The Klenow fragment is a large protein fragment that is produced in E. coli when DNA polymerase reacts with certain enzymes [Klenow and Henningsen (1970)]. The data set we will consider consists of 250 noisy projections of the actual known structure of the particle, produced in silico, mimicking the behavior of the electron microscope, and kindly provided by Professor Andres Leschziner, Harvard University [for a detailed description of the data generation methodology, see Leschziner and Nogales (2006), Sections 2.1 and 2.2]. A sample of twelve of these projections is depicted in Figure 8. The projection signal-to-noise ratio is at the level of 3 : 1.

6.1. *Identifiability and blind deconvolution.* A brief visual inspection of these projections should make it immediately clear that, unlike the synthetic particle example treated earlier, the Klenow fragment does not fit precisely within the sparse radial mixture framework. However, it is also apparent that if it is a coarse first order approximation that we are interested in, then the sparse radial model is quite reasonable. Nevertheless, the approximate nature of this representation will have certain implications:

(1) The isotropic density function on which the radial representation is based, is unknown. In essence, this means that the deconvolution problem at hand is a blind deconvolution problem, as the point spread function itself is poorly determined.

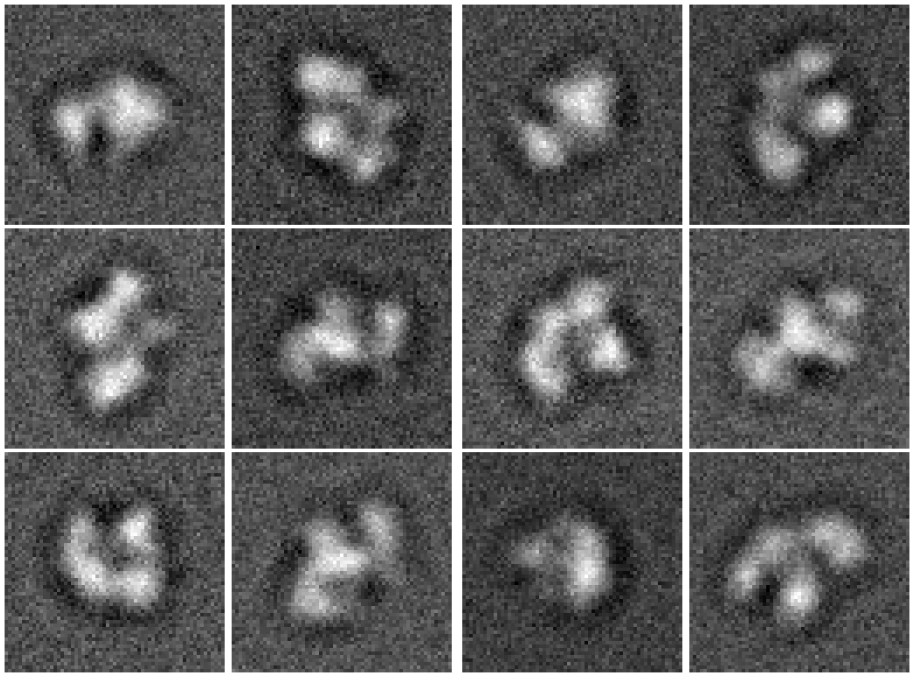


FIG. 8. *A sample of twelve projections from the Klenow fragment data set.*

Fortunately, we will see that the discrete deconvolution approach based on the LASSO remains successful even when the convolution matrix is approximate.

(2) It is likely that only a subset of the projections will be usable, because several of the projections may involve projected means that lie close to one another, hence pushing to the limit of unidentifiability.

(3) The mixing proportions corresponding to the best fitting radial representation have no guarantee of being well separated. Therefore, we will need to make use of the special geometry of the problem, as the estimated mixing weights will not be sufficient for labeling the components.

We begin by applying the LASSO deconvolution procedure to each of the 250 profiles. Since the isotropic density for the expansion is unknown, we employ a Gaussian kernel using $\sigma = 0.224$—a value chosen experimentally (and which will later be refined). Interestingly, we observed that employing different kernels (or even different scale parameters) did not significantly influence the results, provided that σ was not too large. Even though the point spread function was more or less arbitrarily selected, the LASSO deconvolution procedure produced highly sensible output (some examples are shown in Figure 9), providing evidence to the effect that the procedure is relatively robust to perturbations of the point spread function, provided that it remains isotropic and relatively concentrated.

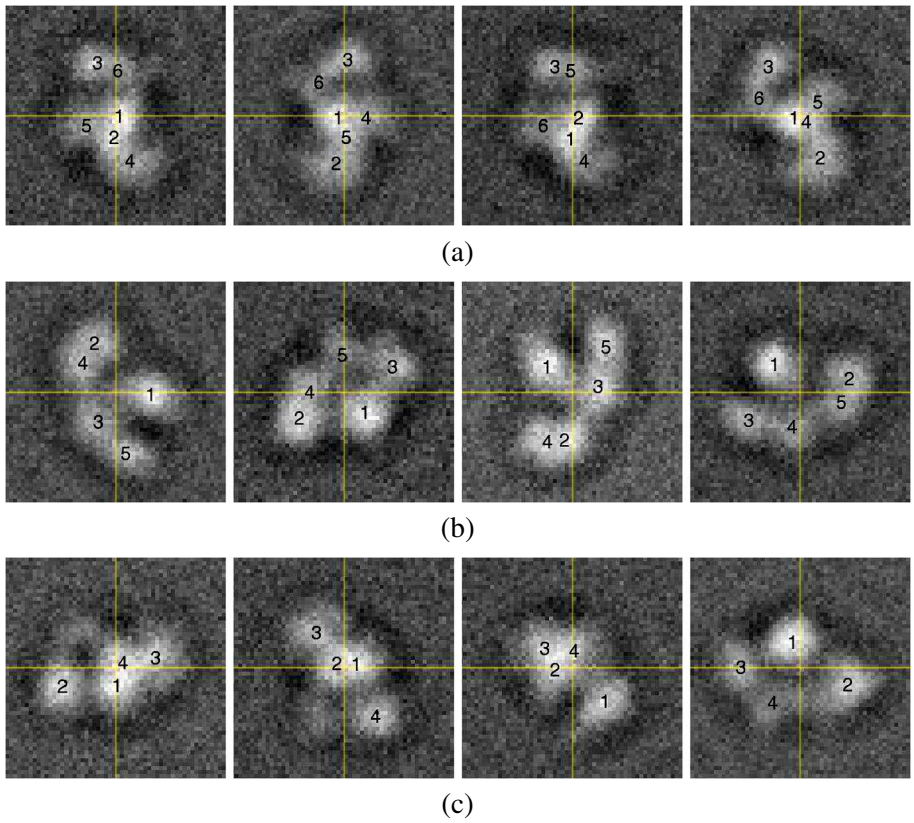


FIG. 9. *Three classes of* 4 *profiles. The labeling of the components was obtained by taking the mixing weights in descending order.* (a) *Class* 1: *profiles with* 6 *identifiable projected component means*. (b) *Class* 2: *profiles with* 5 *identifiable projected component means*. (c) *Class* 3: *profiles with* 4 *identifiable projected component means*.

Since several profiles fell into the "almost unidentifiable" regime, we selected three classes of profiles where the location parameters seemed well identified and that comprised relatively different viewpoints of the particle. The classes were constructed by choosing a *generating* profile and then selecting additional profiles that appeared to be reflections or rotations of the generating profile. Class 1 consisted of profiles where the LASSO deconvolution procedure identified 6 component means. Classes 2 and 3 consisted of profiles where the LASSO deconvolution procedure identified, respectively, 5 and 4 component means. The three classes are shown in Figure 9. Our experience showed that only very few particles are actually required to obtain a good reconstruction and so we limited class membership to four particles per class (in the pseudo-particle example, we observed that as few as a dozen could be used to produce an excellent reconstruction).

The next steps require determining the correct labeling of the components within each class relative to the generating profile, and consistently combining the

Gram matrix estimates from each class to obtain an overall estimate of the Gram matrix. To describe these steps, we use the notation $\hat{\mu}_k^{(i\cdot j)}$ and $\hat{q}_k^{(i\cdot j)}$ to denote, respectively, the estimated mean and mixing weight of the kth component in the jth profile of class i. Additionally, we use $\hat{\mu}^{(i\cdot j)}$ to denote the matrix with columns $\hat{\mu}_k^{(i\cdot j)}$ and $\hat{q}^{(i\cdot j)}$ to denote the vector with elements $\hat{q}_k^{(i\cdot j)}$, $k = 1, \ldots, K_i$.

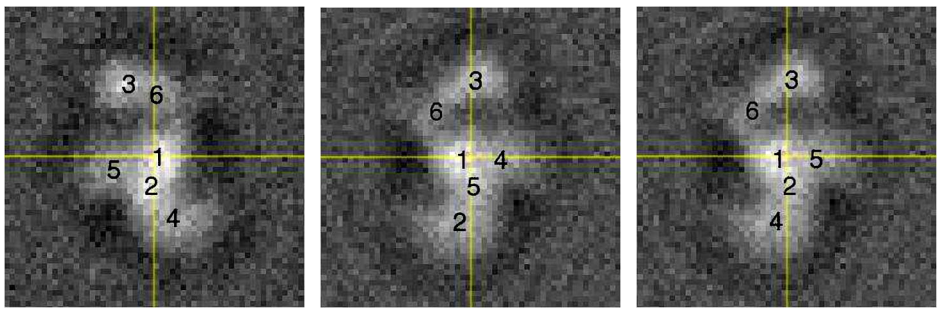
6.2. *Labeling the projected component means within a class of profiles.* Since each class consists of profiles that are assumed to be approximately rotations or reflections (plus some *small* perturbation) of the generating profile, the Gram matrix generated by any member of the class should be *close* to the Gram matrix generated by the generating profile when the corresponding components have the same labels. This suggests the following Procrustean algorithm for determining the correspondences between the projected component means in a candidate profile and those in the generating profile:

(1) Make a list of all possible labelings of the components in the candidate profile.

(2) For each labeling l, compute the quantity $d_l = \|G_R - G_l\|_F$ where G_R is the Gram matrix generated by the reference profile, G_l is the Gram matrix generated by the candidate profile with labeling l and $\|\cdot\|_F$ is the Frobenius matrix norm.

(3) Choose the labeling corresponding to the smallest d_l.

An example is shown in Figure 10. The correspondences within each class can now be obtained by applying these steps, in turn, to each candidate profile in the class.



(a) (b) (c)

FIG. 10. *The target profile is roughly a reflection of the reference profile. The initial labeling of the components in the target profile was obtained from the estimated mixing weights and does not agree with the reference profile. However, the alignment algorithm finds the correct correspondences.* (a) *Reference*; (b) *target*: *before*; (c) *target*: *after.*

6.3. *Estimating the Gram matrix.* We begin by using Theorem 3.1 to produce an initial estimate of the Gram matrix for Class 1:

$$\tilde{G}_1 = \frac{3}{2 \cdot 4} \sum_{j=1}^{4} \mathsf{Gram}\big(\hat{\mu}^{(1 \cdot j)}\big). \tag{6.1}$$

It should be noted that while each individual Gram matrix within this class encodes an ensemble that is intrinsically two-dimensional (i.e., has rank 2), the Gram matrix obtained by the averaging procedure does not necessarily encode an ensemble that can be imbedded into a two-dimensional plane (i.e., the averaged matrix has rank higher than 2). This provides some intuition on the workings of the inversion procedure: if all the projections within a class were identical, the average would be exactly of rank 2, so that the averaging provides no three-dimensional information. However, none of the class members represent precisely the same orientation. These minor perturbations provide some three-dimensional information, even though not dramatic: the resulting matrix might no longer be of rank 2, but its third singular value will be relatively small- the three-dimensional ensemble it encodes is almost two dimensional. The intuition is that when Gram matrices from further classes are added in (representing significantly different orientations), the ensemble generated by the averaged Gram matrix becomes "more three-dimensional."

In fact, there is no guarantee that a Gram matrix formed by averaging several rank-2 Gram matrices will have rank 3: the rank may actually end up being higher. For this reason, we further make a rank 3 approximation of $\tilde{G}_1$ using its singular value decomposition. Let

$$\tilde{G}_1 = U_1 D_1 V_1^{\top}$$

be the singular value decomposition of $\tilde{G}_1$ and define

$$\hat{G}_1 = U_1' D_1' V_1'^{\top},$$

where U_1' and V_1' are, respectively, the first 3 columns of U_1 and V_1 and D' is a diagonal matrix containing the first 3 singular values of $\tilde{G}_1$.

In class 2, we have $K_2 = K_1 - 1 = 5$, hence, we assume that one of the identified means in class 2 has multiplicity 2 (i.e., we assume there are in fact 6 components in the true density and that the projections of two of them fall sufficiently close in the profiles in class 2 that the LASSO deconvolution approach identifies them as a single component). Also, we note that the largest component is sufficiently distinct that it can be used reliably to identify the first component.

Following from these assumptions and in order to obtain a 6×6 Gram matrix "compatible" with $\hat{G}_1$, we consider ensembles of means of the form

$$\big[\hat{\mu}^{(2 \cdot j)}\ \hat{\mu}_k^{(2 \cdot j)}\big]$$

for $k = 1, \ldots, 5$ and $j = 1, \ldots, 4$. The idea is to use the geometrical properties introduced in the previous section, to choose that candidate Gram matrix which lies *closest* to the locus of projected Gram matrices. The latter is unknown, but we may approximate it by the locus generated by $\hat{G}_1$, which constitutes itself an estimator of the unknown three-dimensional Gram matrix. Let $\mathcal{P} = \{P_l\}$ for $l \in \mathcal{L}$ be the set (with index l) of all 6×6 permutation matrices that leave the first row unchanged. We then build the set of candidate Gram matrices with elements

$$G_{lk} = \frac{3}{2 \cdot 4} \sum_{j=1}^{4} \mathsf{Gram}\big([\hat{\mu}^{(2 \cdot j)} \ \hat{\mu}_k^{(2 \cdot j)}] P_l\big) \tag{6.2}$$

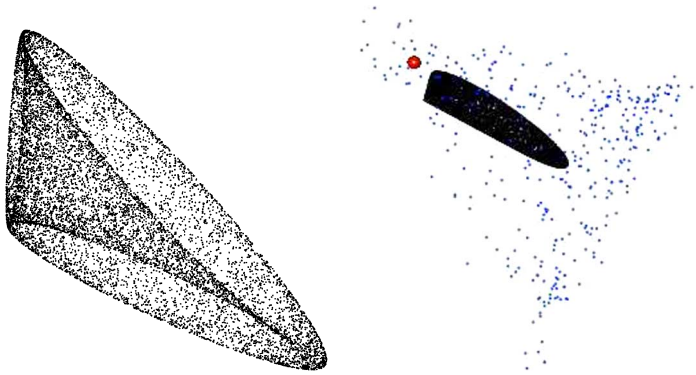
for all combinations of $l \in \mathcal{L}$ and $k \in \{1, \ldots, 5\}$.

Our measure of affinity to the locus generated by $\hat{G}_1$ is the Euclidean distance between the candidate Gram matrix and this locus: the stretched Roman surface generated by $\hat{G}_1$. In practice, this distance is computed by randomly sampling a set of points on the Roman surface, then taking the minimum distance between each of these points and the candidate Gram matrix (Figure 11). Of course, the induced distribution on the Roman surface will no longer be uniform, but it is not necessary that it be. All that is required is a relatively good coverage of the surface. We use a sample of 1,000 points on the perturbed Roman surface, defined as

$$S_n = V_1'(I - u_n u_n^\top) V_1'^\top, \qquad n = 1, \ldots, 1{,}000,$$

with $\{u_n\}$ being unit random vectors and $(V_1')^\top V_1' = \hat{G}_1$. We then define the distance

$$d(G_{lk}) = \min_n \|G_{lk} - S_n\|_F.$$



(a) (b)

FIG. 11. (a) *Sampled points on the Roman surface generated by* $\hat{G}_1$. (b) *Point cloud of all possible permuted Gram matrices for class* 2, *relative to the Roman surface generated by* $\hat{G}_1$. *The red point corresponds to the point of minimum distance.*

Taking $\tilde{l}$ and $\tilde{k}$ such that $d(G_{\tilde{l}\tilde{k}})$ is minimum, we proceed to compute an initial estimate of the Gram matrix from the profiles in classes 1 and 2 as

$$\tilde{G}_{12} = \frac{3}{2\cdot 8}\left[\sum_{j=1}^{4} \mathsf{Gram}(\hat{\mu}^{(1\cdot j)}) + \sum_{j=1}^{4} \mathsf{Gram}([\hat{\mu}^{(2\cdot j)}\ \hat{\mu}_{\tilde{k}}^{(2\cdot j)}]P_{\tilde{l}})\right]. \tag{6.3}$$

The final estimate $\hat{G}_{12}$ is obtained by making a rank 3 approximation of $\tilde{G}_{12}$ using the singular value decomposition.

Finally, we proceed analogously for the remaining class. In class 3, we have $K_3 = K_1 - 2 = 4$, hence, we assume that either one of the identified means has multiplicity 3, or two identified means have multiplicity 2. Again, to obtain a 6×6 Gram matrix, we consider ensembles of the form

$$[\hat{\mu}^{(3\cdot j)}\ \hat{\mu}_{k_1}^{(3\cdot j)}\ \hat{\mu}_{k_2}^{(3\cdot j)}],$$

where $k_1 \le k_2 \in \{1,\dots,4\}$. The overall Gram matrix estimate is again computed by generating a set of candidate Gram matrices and taking the configuration that yields the smallest distance to the stretched Roman surface generated by $\hat{G}_{12}$, then by updating the Gram matrix estimate as above.

As a by-product of estimating the Gram matrix, we now know where to place each of the 6 component means in any given profile, and which component means correspond to which from profile to profile. Consequently, we may estimate the mixing weights and the tuning parameter σ^2 using linear regression. Given a candidate value for σ^2, we can construct N convolution matrices $\{\mathcal{X}_{n,\sigma^2}\}_{n=1}^{N}$ (corresponding to the N profiles) as described in Section 4. We thus obtain N linear regression problems, one for each projection. By stacking the corresponding convolution matrices into a single $N^2\times 6$ matrix, we obtain a single regression for the 6 common mixing weights, and estimate the latter by ordinary least squares. The procedure can be performed for different choices of σ^2 on a prespecified grid, retaining the set of mixing weight estimates that correspond to the regression with the best fit. The estimate for σ^2 thus obtained for the Klenow fragment was $\hat{\sigma}^2 = 0.0571$.

The sparse reconstruction produced by employing the estimated Gram matrix and mixing coefficients is depicted in Figure 12. In order to appreciate the "fit" of the sparse reconstruction to the data, we construct noisy projections from the reconstruction and contrast them to several typical projections of the Klenow fragment. The projections of the reconstructed model are constructed so as to mimic the effects that the microscope induces on the profiles (astigmatism, defocus, contrast transfer function effects) and so as to be characterized by a signal-to-noise ratio similar to that of the actual projections. A sample of such contrasts is given in Figure 13. We observe that, even though rather sparse, the reconstructed density is able to capture the main features of the projections quite successfully. This hints that the reconstructed density can be appropriate to use as a starting model.

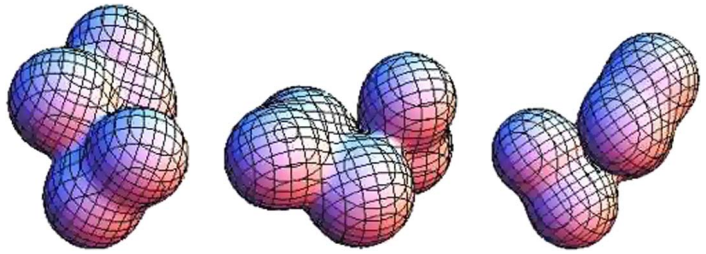


FIG. 12. *Estimated density for the Klenow fragment*. (a) *View point* 1; (b) *view point* 2; (c) *view point* 3.

What is especially important is that the data produced by our reconstruction seems to be highly consistent with the actual Klenow data even for viewing angles that were *not used in the reconstruction*; in fact, this remains true even for viewing angles that fall in the unidentifiability regime. Figure 14 shows the corresponding residual deviation heat-maps. We observe that underestimation (corresponding to yellow/orange regions) occurs in the regions between the components of the Gaussian mixture—evidently, there is mass there that cannot be captured by the Gaussian mixture. There are also some regions where overestimation occurs (darker green regions), mostly close to the center of blob-like components of the particle profiles, principally due to the fact that the mixing components of the Gaussian mixture will obviously have relatively different higher-order concentration characteristics from the blob-like components of the actual particle. For example, in Figure 14(f), we observe slight overestimation of the density at locations corresponding to the center of the blob-like components of the actual profile. These two are the only systematic patterns that appear in the residuals, and are evidently attributed to the bias introduced from our regularization via the Gaussian mixture model employed.

7. Concluding remarks. Despite the severely ill-posed nature of noisy random tomography, this paper demonstrates that it is practically feasible to obtain useful three-dimensional structural information on a protein given only noisy projections at random and unknown angles. The approach proposed to this effect rests on two basic elements: the imposition of a certain degree of sparsity on the required reconstruction, and the exploitation of the special geometry that is intrinsic to tomography data and provides valuable information. Though the sparsity assumption will typically lead to a relatively coarse-grained approximation to the protein under investigation, this is precisely what is required: a low-resolution starting model that can be used as a reference structure to iteratively recover the unknown angles to then produce a high-resolution reconstruction based on traditional nonparametric tomographic techniques (once the projection angles have been estimated, it is no longer necessary to maintain the mixture model).

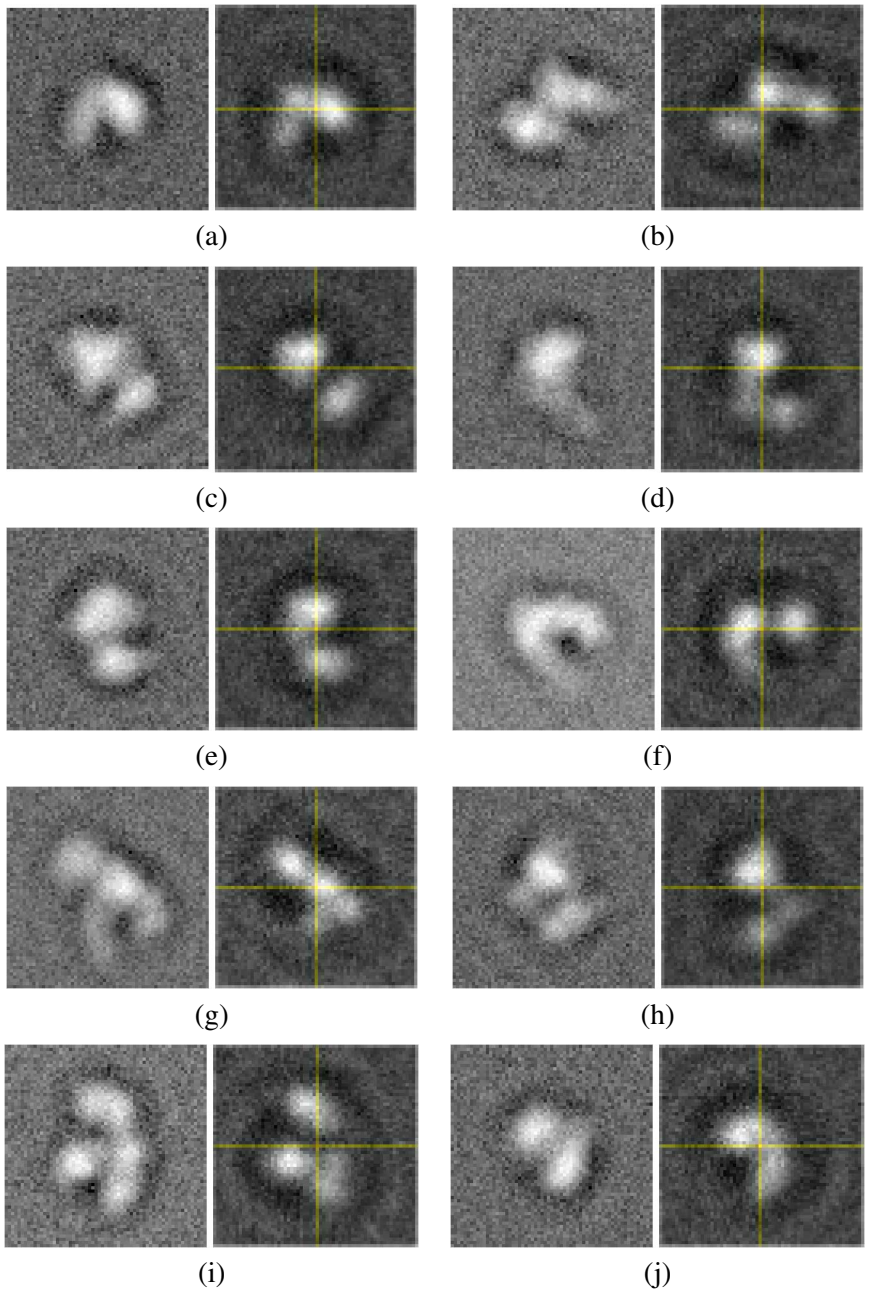


FIG. 13. *Ten pairs of projections. Each pair contains an actual Klenow fragment projection* (*left*) *coupled with a projection from our sparse approximation* (*right*).

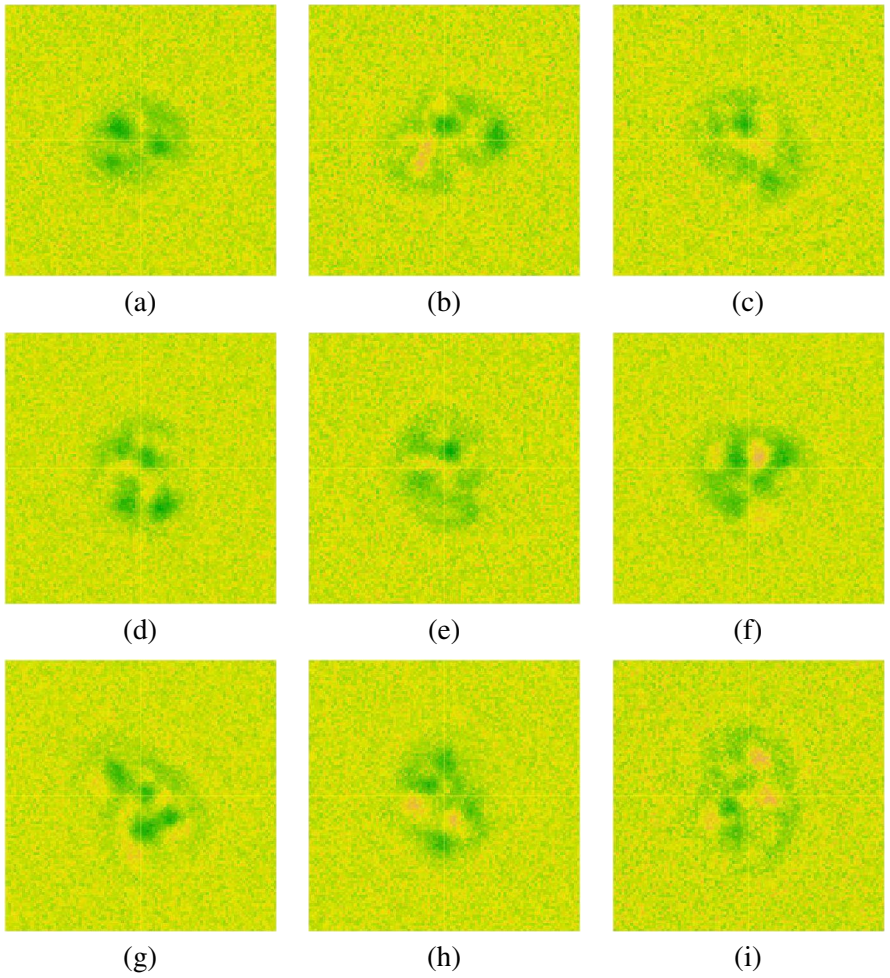


FIG. 14. *Heat-maps of residual deviations of the fitted projections from the actual projections. Each of the subfigures* (a)–(i) *is generated from the corresponding pairs* (a)–(i) *in Figure* 13.

While it had previously been theoretically demonstrated in Panaretos (2009) that it is feasible to reconstruct a three-dimensional object in this setting (up to an orthogonal transformation), obtaining an explicit reconstruction in practice remained elusive. By employing a radial basis representation of the unknown protein, the problem of structure determination was reduced to the problem of recovering the Euclidean shape of the ensemble of location parameters of the radial functions and the associated mixing coefficients. This was done in two steps: nonlinear deconvolution and shape averaging. In the deconvolution step, the projected location parameters had to be identified within the noisy projections. Since the nature of the radial expansion representation is approximate, the deconvolution problem was blind.

To tackle this problem, our approach relaxed the nonlinear deconvolution problem into a linear problem by considering its discretized version with an approximately chosen point spread function. When seen in this setting, the problem falls precisely in the framework of sparse model selection. Since the object to be recovered is a density function, the LASSO arose as the most natural technique to attack the problem, with an L^1 penalty corresponding to a requirement on the total mass of the density to be recovered. Despite the fact that the exact point spread function was unknown, it was seen that the LASSO performed extremely satisfactorily where other solution approaches break down. This was true both in the setting of artificial examples, as well as in the setting of protein data. Once the projections of the location parameters had been deconvolved, the averaging step was carried out. This required the recovery of the correspondences between location parameters in different projections. To this effect, our approach exploited the nonregularity of the tomography problem: it was seen that the Gram matrices of the $k \times k$ projected components are constrained to lie in a smooth two-dimensional subset of $\mathbb{R}^{k \times k}$, which was identified as a deformed Roman surface. This was then exploited in order to choose consistent correspondences across projections.

The methodology was applied with success both to projection data arising from an artificial example, as well as to projections of an actual protein component, the Klenow fragment. Especially in the latter case, it was seen that the sparse reconstruction recovered from the noisy projection data can very well serve as a starting model, since its typical projections are highly similar with those of the projection of the true structure (Figure 13). It is therefore likely that our approach will provide a useful means to obtaining objective data-dependent starting models in the context of single particle electron microscopy. From the purely statistical perspective, the use of the LASSO in the setting of double blind deconvolution can be of independent interest when seen in the context of estimation of mixtures of scale-location densities of an unknown family.

Acknowledgments. We wish to thank Professor R. M. Glaeser for many useful interactions on the problem of initial model determination. Our thanks also go to Dr. Richard Hall for his kind help with electron microscopy software, and to Professor Andres Leschziner for sharing the Klenow fragment data set. We are also thankful to two anonymous referees for their constructive comments and suggestions.

REFERENCES

APERY, F. (1987). *Models of the Real Projective Plane*: *Computer Graphics of Steiner and Boy Surfaces*. Vieweg, Braunschweig. MR0986729

BERN, M., CHEN, J. and WONG, H. (2005). Avoiding local optima in single particle reconstruction. In *Research in Computational Molecular Biology* (S. Miyano, J. Mesirov, S. Kasif, S. Istrail, P. Pevzner and M. Waterman, eds.). *Lecture Notes in Computer Science* **3500** 118–132. Springer, Berlin. MR2304663

CHIU, W. (1993). What does electron cryomicroscopy provide that X-ray crystallography and NMR spectroscopy cannot? *Annu. Rev. Bioph. Biom.* **22** 233–255.

DEANS, S. R. (1993). *The Radon Transform and Some of Its Applications*. Krieger, Malabar, FL. MR1274701

DONOHO, D. L., JOHNSTONE, I. M., HOCH, J. C. and STERN, A. S. (1992). Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B* **54** 41–81. MR1157714

DRENTH, J. (1999). *Principles of Protein X-Ray Crystallography*. Springer, New York.

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–451. MR2060166

FRANK, J. (1999). *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Academic Press, San Diego.

GLAESER, R. M. (1999). Review: Electron crystallography: Present excitement, a nod to the past, anticipating the future. *J. Struct. Biol.* **128** 3–14.

GLAESER, R. M., CHIU, W., FRANK, J., DEROSIER, D., BAUMEISTER, W. and DOWNING, K. (2007). *Electron Crystallography of Biological Macromolecules*. Oxford Univ. Press, Oxford.

GREEN, P. J. (1990). Bayesian reconstructions from emission tomography data using a modified EM algorithm. *IEEE T. Med. Imaging* **9** 84–93.

GRENANDER, U. and SZEGÖ, G. (1958). *Toeplitz Forms and Their Applications*. Univ. California Press, Berkeley. MR0094840

HASTIE, T. and EFRON, B. (2011). lars: Least angle regression, Lasso and forward stagewise. R package version 0.9-8.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*. Springer, New York. MR1851606

HENDERSON, R. (2004). Realizing the potential of electron cryo-microscopy. *Q. Rev. Biophys.* **37** 3–13.

JOHNSTONE, I. M. and SILVERMAN, B. W. (1990). Speed of estimation in positron emission tomography and related inverse problems. *Ann. Statist.* **18** 251–280. MR1041393

JONES, M. C. and SILVERMAN, B. W. (1989). An orthogonal series density estimation approach to reconstructing positron emission tomography images. *J. Appl. Stat.* **16** 177–191.

KENDALL, W. S. and LE, H. (2009). Statistical shape theory. In *New Perspectives in Stochastic Geometry* (W. S. Kendall and I. S. Molchanov, eds.). Oxford Univ. Press, Oxford. MR2654683

KENDALL, D. G., BARDEN, D., CARNE, T. K. and LE, H. (1999). *Shape and Shape Theory*. Wiley, Chichester. MR1891212

KLENOW, H. and HENNINGSEN, I. (1970). Selective elimination of the exonuclease activity of the deoxyribonucleic acid polymerase from escherichia coli B by limited proteolysis. *Proc. Natl. Acad. Sci. USA* **65** 168–175.

LE, H. and BARDEN, D. (2010). On the induced distribution of the shape of the projection of a randomly rotated configuration. *Adv. in Appl. Probab.* **42** 331–346. MR2675105

LESCHZINER, A. E. and NOGALES, E. (2006). The orthogonal tilt reconstruction method: An approach to generating single-class volumes with no missing cone for ab initio reconstruction of asymmetric particles. *J. Struct. Biol.* **153** 284–299.

NATTERER, F. (2001). *The Mathematics of Computerized Tomography*. *Classics in Applied Mathematics* **32**. SIAM, Philadelphia, PA. MR1847845

O'SULLIVAN, F. (1995). A study of least squares and maximum likelihood for image reconstruction in positron emission tomography. *Ann. Statist.* **23** 1267–1300. MR1353506

PANARETOS, V. M. (2006). The diffusion of radon shape. *Adv. in Appl. Probab.* **38** 320–335. MR2264947

PANARETOS, V. M. (2008). Representation of radon shape diffusions via hyperspherical Brownian motion. *Math. Proc. Cambridge Philos. Soc.* **145** 457–470. MR2442137

PANARETOS, V. M. (2009). On random tomography with unobservable projection angles. *Ann. Statist.* **37** 3272–3306. MR2549560

PENCZEK, P., GRASSUCCI, R. A. and FRANK, J. (1994). The ribosome at improved resolution: New techniques for merging and orientation refinement in 3D cryoelectron microscopy of biological particles. *Ultramicroscopy* **40** 33–53.

PISARENKO, V. F. (1973). The retrieval of harmonics from a covariance function. *Geophys. J. R. Astr. S.* **33** 347–366.

R DEVELOPMENT CORE TEAM (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RADEMACHER, M. (1994). Three dimensional reconstruction from random projections: orientational alignment via Radon transforms. *Ultramicroscopy* **17** 117–126.

SIGWORTH, F. J. (1998). A maximum-likelihood approach to single-particle image refinement. *J. Struct. Biol.* **122** 328–339.

SILVERMAN, B. W., JONES, M. C., WILSON, J. D. and NYCHKA, D. W. (1990). A smoothed EM approach to indirect estimation problems with particular reference to stereology and emission tomography. *J. Roy. Statist. Soc. Ser. B* **52** 271–324. MR1064419

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

VARDI, Y., SHEPP, L. A. and KAUFMAN, L. (1985). A statistical model for positron emission tomography (with discussion). *J. Amer. Statist. Assoc.* **80** 8–37. MR0786595

SECTION DE MATHÉMATIQUES
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
EPFL STATION 8, 1015 LAUSANNE
SWITZERLAND
E-MAIL: victor.panaretos@epfl.ch
kjell.konis@epfl.ch