SIAM J. SCI. COMPUT.
Vol. 39, No. 6, pp. A3107–A3128

PROJECTED NONLINEAR LEAST SQUARES FOR EXPONENTIAL FITTING[*]

JEFFREY M. HOKANSON[†]

Abstract. The modern ability to collect vast quantities of data provides a challenge for parameter estimation. When posed as a nonlinear least squares problem fitting a model to data, the cost of each iteration grows linearly with the amount of data and with large data it can easily become too expensive to perform many iterations. Here we reduce the cost of each iteration by orthogonally projecting the data onto a low-dimensional subspace preserving the quality of the resulting parameter estimates. We provide results from both an optimization and a statistical perspective that show accurate parameter estimates are recovered when the subspace angles between this subspace and the range Jacobian of the model at the current iterate remain small. However, for this approach to reduce computational complexity, both the projected model and projected Jacobian must be computed inexpensively. This places a constraint on the pairs of models and subspaces for which this approach provides a computational speedup. Here we consider the exponential fitting problem projected onto the range of a Vandermonde matrix for which both the projected model and projected Jacobian can be computed in closed form using a generalized geometric sum formula. We further provide an inexpensive heuristic for choosing this Vandermonde matrix which ensures the subspace angles with the Jacobian remain small, and use this heuristic to update the subspace during optimization. Although the asymptotic cost still depends on the data dimension, the overall cost of solving this sequence of projected problems is significantly less expensive than the original.

Key words. exponential fitting, harmonic estimation, modal analysis, spectral analysis, parameter estimation, nonlinear least squares, dimension reduction, experimental design

AMS subject classifications. 11L03, 62K99, 65K10, 90C55

DOI. 10.1137/16M1084067

1. Introduction.

With the increasing prowess of data acquisition hardware and storage, collecting vast amounts of data has become trivial. This poses a challenge for parameter estimation problems where the sheer scale of data makes these problems expensive. Here we consider a nonlinear least squares parameter estimation problem [12] that seeks to fit a model $\mathbf{f}$ with q parameters $\boldsymbol{\theta} \in \mathbb{C}^q$ to (noisy) measurements $\widetilde{\mathbf{y}}$, yielding a (noisy) parameter estimate $\widetilde{\boldsymbol{\theta}}$ that minimizes the 2-norm mismatch:

$$(1) \qquad \widetilde{\boldsymbol{\theta}} := \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{C}^q} \|\mathbf{f}(\boldsymbol{\theta}) - \widetilde{\mathbf{y}}\|_2^2, \quad \text{where} \quad \mathbf{f} : \mathbb{C}^q \to \mathbb{C}^n, \quad \widetilde{\mathbf{y}} \in \mathbb{C}^n, \quad q \ll n.$$

With vast quantities of data, the asymptotic cost of solving this problem is dominated by the n-dependent steps in the optimization. For example, using either Gauss–Newton or Levenberg–Marquardt, each optimization step solves a least squares problem involving the Jacobian of $\mathbf{f}$, $\mathbf{J} : \mathbb{C}^q \to \mathbb{C}^{n\times q}$, at a cost of $\mathcal{O}(nq^2)$ operations. To reduce this cost, we propose replacing the full least squares problem (1) with a sequence of low-dimensional surrogate problems by projecting measurements $\widetilde{\mathbf{y}}$ onto

[*] Submitted to the journal's Methods and Algorithms for Scientific Computing section July 14, 2016; accepted for publication (in revised form) September 19, 2017; published electronically December 21, 2017.

http://www.siam.org/journals/sisc/39-6/M108406.html

Funding: This work was supported by NSF VIGRE grants DMS-0240058 and DMS-0739420 at Rice University and the Department of Defense, Defense Advanced Research Project Agency's program Enabling Quantification of Uncertainty in Physical Systems.

[†] Department of Computer Science, University of Colorado Boulder, Boulder, CO 80309 (Jeffrey.Hokanson@colorado.edu).

a sequence of subspaces $\{\mathcal{W}_\ell\}_\ell \subset \mathbb{C}^n$ with $m_\ell := \dim \mathcal{W}_\ell \ll n$:

$$\widetilde{\boldsymbol{\theta}}_{\mathcal{W}_\ell} := \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{C}^q} \|\mathbf{P}_{\mathcal{W}_\ell}[\mathbf{f}(\boldsymbol{\theta}) - \widetilde{\mathbf{y}}]\|_2^2 = \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{C}^q} \|\mathbf{W}_\ell^*\mathbf{f}(\boldsymbol{\theta}) - \mathbf{W}_\ell^*\widetilde{\mathbf{y}}\|_2^2, \quad \mathbf{P}_{\mathcal{W}_\ell} = \mathbf{W}_\ell\mathbf{W}_\ell^*, \tag{2}$$

where $\mathbf{P}_{\mathcal{W}_\ell}$ is an orthogonal projector onto $\mathcal{W}_\ell$ and $\mathbf{W}_\ell \in \mathbb{C}^{n\times m_\ell}$ is an orthonormal basis for $\mathcal{W}_\ell$. Although the total cost is still asymptotically n-dependent due to the multiplication $\mathbf{W}_\ell^*\widetilde{\mathbf{y}}$, each optimization step is cheaper since the projected Jacobian $\mathbf{W}_\ell^*\mathbf{J}(\boldsymbol{\theta})$ is smaller. However, for this computational speedup to be fully realized, the products $\mathbf{W}_\ell^*\mathbf{f}(\boldsymbol{\theta})$ and $\mathbf{W}_\ell^*\mathbf{J}(\boldsymbol{\theta})$ need to be formed without the expensive, n-dependent multiplication. Additionally, we must ensure that the final projected parameter estimate $\widetilde{\boldsymbol{\theta}}_{\mathcal{W}_\ell}$ remains a good estimate of the full parameter estimate using the original data, $\widetilde{\boldsymbol{\theta}}$. Here we do so by requiring the subspace angles between $\mathcal{W}_\ell$ and the Jacobian at the current iterate remain small. This requirement is justified by perspectives from both optimization and statistics. From an optimization perspective described in section 2, the accuracy of each optimization step depends on these subspace angles, and when the subspace angles go to zero, the projected parameter estimate $\widetilde{\boldsymbol{\theta}}_{\mathcal{W}_\ell}$ is equal to the full parameter estimate $\widetilde{\boldsymbol{\theta}}$. From a statistical perspective described in section 3, when measurements $\widetilde{\mathbf{y}}$ are contaminated by additive Gaussian noise, the covariance of projected parameter estimate $\widetilde{\boldsymbol{\theta}}_{\mathcal{W}_\ell}$ is larger than the covariance of full parameter estimate $\widetilde{\boldsymbol{\theta}}$ by an amount that scales with these subspace angles as measured by *efficiency*. Hence the subspace angles between $\mathcal{W}_\ell$ and the Jacobian at the current iterate determine the quality of our projected parameter estimate $\widetilde{\boldsymbol{\theta}}_{\mathcal{W}_\ell}$. The challenge in applying this projected nonlinear least squares approach to a specific problem is satisfying both criteria simultaneously: finding a sequence of subspaces $\{\mathcal{W}_\ell\}_\ell$ with orthogonal bases $\{\mathbf{W}_\ell\}_\ell$ where $\mathbf{W}_\ell^*\mathbf{f}(\boldsymbol{\theta})$ and $\mathbf{W}_\ell^*\mathbf{J}(\boldsymbol{\theta})$ can be formed inexpensively independently of n and where the subspace angles between $\mathcal{W}_\ell$ and the range of $\mathbf{J}(\boldsymbol{\theta}_k)$ for each iterate $\boldsymbol{\theta}_k$ remain small.

Here we consider the *exponential fitting problem* [27], also known as *modal analysis* [5], *harmonic estimation* [18], and *spectral analysis* [34], that seeks to approximate data $\widetilde{\mathbf{y}}$ as a sum of p complex exponentials with frequencies $\boldsymbol{\omega}$ and amplitudes $\mathbf{a}$, where

$$[\mathbf{f}([\boldsymbol{\omega}, \mathbf{a}])]_j = \sum_{k=1}^{p} a_k e^{j\omega_k}, \quad \boldsymbol{\omega}, \mathbf{a} \in \mathbb{C}^p; \quad \boldsymbol{\theta} = [\boldsymbol{\omega}, \mathbf{a}], \quad q = 2p. \tag{3}$$

There is an extensive body of literature on this problem, with a wide array of methods for recovering the frequencies $\boldsymbol{\omega}$: from classical approaches such as Prony's method [28], to subspace methods [13] such as HSVD [1], HTLS [40], and the matrix-pencil method [16], to parameter estimation approaches using optimization (such as ours) [12, 39], to more recent approaches based on ideas from sparse recovery [35], and many others described in reviews [17, 20, 27, 37]. We choose this problem due to the exploitable structure of the model function $\mathbf{f}([\boldsymbol{\omega}, \mathbf{a}])$ that allows us to obtain inexpensive inner products with subspaces that approximately contain the range of the Jacobian. Specifically, as the model function is the product of a Vandermonde matrix $\mathbf{V}(\boldsymbol{\omega})$ with the amplitudes $\mathbf{a}$,

$$\mathbf{f}([\boldsymbol{\omega}, \mathbf{a}]) = \mathbf{V}(\boldsymbol{\omega})\mathbf{a}, \qquad [\mathbf{V}(\boldsymbol{\omega})]_{j,k} = e^{j\omega_k}, \tag{4}$$

by projecting measurements onto the subspace $\mathcal{W}(\boldsymbol{\mu})$,

$$\mathcal{W}(\boldsymbol{\mu}) := \operatorname{Range} \mathbf{V}(\boldsymbol{\mu}) = \operatorname{Range} \mathbf{W}(\boldsymbol{\mu}), \qquad \mathbf{W}(\boldsymbol{\mu})^*\mathbf{W}(\boldsymbol{\mu}) = \mathbf{I}, \quad \boldsymbol{\mu} \in \mathbb{C}^m, \tag{5}$$

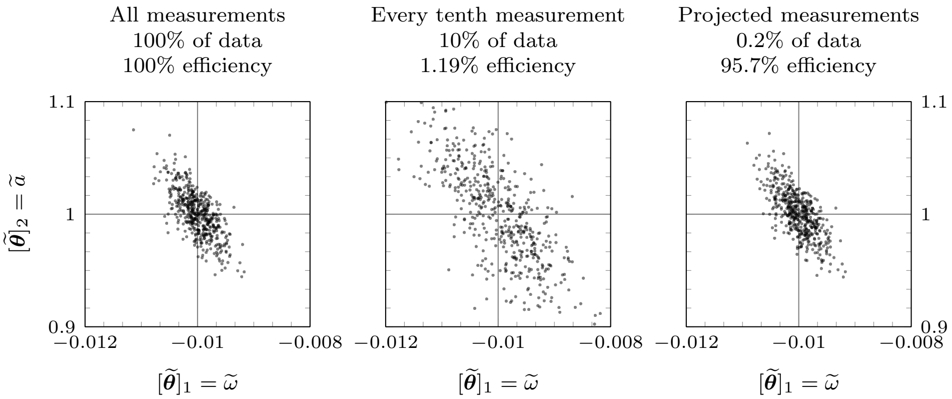

FIG. 1. *Parameter estimates from a toy exponential fitting problem with true parameters* $\widehat{\omega} = -0.01$ *and* $\widehat{a} = 1$ *using* $n = 1000$ *measurements contaminated with zero-mean Gaussian noise* $\mathbf{g}$ *with* $\operatorname{Cov}\mathbf{g} = 0.01\mathbf{I}$, $\widetilde{\mathbf{y}} = \mathbf{f}([-0.01, 1]) + \mathbf{g}$. *The parameter estimates on the left are computed by solving* (1) *with all measurements* $\widetilde{\mathbf{y}}$. *In the center, the parameter estimates are computed by solving* (2) *using a subspace that selects every* 10*th measurement. On the right, the parameter estimates are also computed by solving* (2), *but instead using the subspace* $\mathcal{W}(\boldsymbol{\mu})$ *where* $\boldsymbol{\mu} = [-0.008 \pm 0.0014i]$. *In each case, the resulting nonlinear least squares problem was solved using the MATLAB routine* `lsqnonlin`. *Efficiency, defined in section* 3, *quantifies how close the covariance of the projected problem resembles the full problem (left). As this example shows, the projected parameter estimate with well-chosen subspace yields parameter estimates almost identical to those of the full problem.*

we can inexpensively obtain the inner products $\mathbf{W}(\boldsymbol{\mu})^*\mathbf{f}([\boldsymbol{\omega}, \mathbf{a}])$ and $\mathbf{W}(\boldsymbol{\mu})^*\mathbf{J}([\boldsymbol{\omega}, \mathbf{a}])$ as described in section 4 using the geometric sum formula and its generalization given in Appendix B. Further, using a heuristic described in section 5, we can ensure the subspace angles between $\mathcal{W}(\boldsymbol{\mu})$ and $\operatorname{Range}\mathbf{J}([\boldsymbol{\omega}, \mathbf{a}])$ remain small by a careful selection of $\boldsymbol{\mu}$. The net result is a faster solution to the exponential fitting problem in the limit of large data, as illustrated by a magnetic resonance spectroscopy test case in section 7. Further, due to careful selection of the subspaces, the projected parameter estimate $\widetilde{\boldsymbol{\theta}}_{\mathcal{W}}$ remains close to the full parameter estimate $\widetilde{\boldsymbol{\theta}}$, as seen in Figure 1 for a toy problem and in Figure 4 for the magnetic resonance spectroscopy test case.

Projection is a recurring theme in applied mathematics, appearing in a variety of contexts from Galerkin projections for solving partial differential equations to the randomized projections that form the foundation of randomized numerical linear algebra. Our projection approach for solving a nonlinear least squares problem fits into this theme and it is not without precedent. Incremental methods, such as *incremental gradient* [3] and the *extended Kalman filter* [2], when applied to a nonlinear least squares problem can be interpreted as projecting onto a row (or set of rows) at each iteration, choosing the basis $\mathbf{W}_\ell = [\mathbf{I}]_{\cdot,\mathcal{I}_\ell}$, where $\mathcal{I}_\ell$ is the set of rows at the ℓth step. With this perspective, we note that both our method and incremental methods require the projected model $\mathbf{W}_\ell^*\mathbf{f}(\boldsymbol{\theta})$ and projected Jacobian $\mathbf{W}_\ell^*\mathbf{J}(\boldsymbol{\theta})$ to be formed inexpensively. Satisfying this requirement is straightforward for incremental methods when $\mathbf{f}(\boldsymbol{\theta})$ is defined entrywise, whereas in our case we must be careful to choose the orthonormal basis $\mathbf{W}_\ell$ such that these products can be, for example, evaluated in closed form. However, these methods differ in how the basis $\mathbf{W}_\ell$ is chosen. For incremental methods, the set of rows is typically chosen either deterministically by cycling through rows or by randomly selecting rows [8]; whereas in our case, we carefully choose the basis $\mathbf{W}_\ell$ such that our steps are accurate, and, when the data is contaminated by noise, our parameter estimates are precise.

2. An optimization perspective.

In this section we provide three different results that inform the choice of subspace $\mathcal{W}_\ell$ from the perspective of optimization. Each of these results points to the key role played by the principal subspace angles between the subspace $\mathcal{W}_\ell$ and the range of the Jacobian at the current iterate. These principal subspace angles are defined as follows: if $\mathcal{A}$ and $\mathcal{B}$ are two subspaces of $\mathbb{C}^n$ and if $\mathbf{A} \in \mathbb{C}^{n\times m_a}$ and $\mathbf{B} \in \mathbb{C}^{n\times m_b}$ are orthonormal bases for $\mathcal{A}$ and $\mathcal{B}$, then the principal subspace angles $\phi_k(\mathcal{A},\mathcal{B})$ between $\mathcal{A}$ and $\mathcal{B}$ are [11, section 6.4.3]

$$\cos\phi_k(\mathcal{A},\mathcal{B}) := \sigma_k(\mathbf{A}^*\mathbf{B}), \quad 0 \le \phi_1(\mathcal{A},\mathcal{B}) \le \phi_2(\mathcal{A},\mathcal{B}) \le \cdots \le \phi_{\min\{m_a,m_b\}}(\mathcal{A},\mathcal{B}) \le \pi/2, \tag{6}$$

where $\sigma_k(\mathbf{X})$ is the kth singular value of $\mathbf{X}$ in descending order. Our first result in subsection 2.1 uses the first order necessary conditions to observe that the projected problem will have the same stationary points as the full problem when the subspace angles between the $\mathcal{W}$ and the range of the Jacobian at the stationary point are zero. Our second result in subsection 2.2 shows that the difference between the Gauss–Newton steps of the full and projected problems depends on the subspace angles between $\mathcal{W}_\ell$ and the range of the Jacobian at the current iterate. Our third result in subsection 2.3 interprets the Levenberg–Marquardt method applied to the projected problem as computing inexact steps of the Levenberg–Marquardt method applied to the full problem. We show that by making the subspace angles between $\mathcal{W}_\ell$ and the range of the Jacobian at the current iterate small, we can satisfy one of the conditions for the convergence of inexact Newton. Each result suggests that the subspace angles between $\mathcal{W}_\ell$ and the range of the Jacobian at the current iterate should be small.

2.1. First order optimality.

The first order necessary conditions for a point $\breve{\boldsymbol{\theta}}$ to be a local optimum require that the gradient of the objective function at this point be zero. In the context of nonlinear least squares, where the gradient of the full problem (1) is

$$\nabla_{\boldsymbol{\theta}} \|\mathbf{f}(\boldsymbol{\theta}) - \widetilde{\mathbf{y}}\|_2^2 = 2\,\mathbf{J}(\boldsymbol{\theta})^*\mathbf{r}(\boldsymbol{\theta}), \qquad \mathbf{r}(\boldsymbol{\theta}) := \mathbf{f}(\boldsymbol{\theta}) - \widetilde{\mathbf{y}}, \quad [\mathbf{J}(\boldsymbol{\theta})]_{\cdot,k} := \frac{\partial \mathbf{r}(\boldsymbol{\theta})}{\partial[\boldsymbol{\theta}]_k}, \tag{7}$$

a point $\breve{\boldsymbol{\theta}}$ satisfies the first order necessary conditions if

$$\mathbf{J}(\breve{\boldsymbol{\theta}})^*\mathbf{r}(\breve{\boldsymbol{\theta}}) = \mathbf{0}. \tag{8}$$

Similarly for the projected problem (2), a point $\breve{\boldsymbol{\theta}}_{\mathcal{W}}$ will satisfy the first order necessary conditions for the projected problem if

$$\mathbf{J}(\breve{\boldsymbol{\theta}}_{\mathcal{W}})^*\mathbf{P}_{\mathcal{W}}\mathbf{r}(\breve{\boldsymbol{\theta}}_{\mathcal{W}}) = \mathbf{0}. \tag{9}$$

To assess the quality of the projected problem, we ask, Under what conditions will $\breve{\boldsymbol{\theta}}_{\mathcal{W}}$ also satisfy the first order necessary conditions for the full problem (8)? There are two conditions under which this happens. The zero-residual case, where $\mathbf{r}(\breve{\boldsymbol{\theta}}_{\mathcal{W}}) = \mathbf{0}$, implies that measurements $\widetilde{\mathbf{y}}$ exactly fit the model $\mathbf{f}$. This situation makes the problem easy to solve, as any subspace $\mathcal{W}$ that yields a well-posed optimization problem can be used. The other, more general situation occurs when $\mathcal{W}$ contains the range of the Jacobian, as then $\mathbf{P}_{\mathcal{W}}\mathbf{J}(\breve{\boldsymbol{\theta}}_{\mathcal{W}}) = \mathbf{J}(\breve{\boldsymbol{\theta}}_{\mathcal{W}})$. This is equivalent to requiring all the subspace angles between $\mathcal{W}$ and $\operatorname{Range}\mathbf{J}(\breve{\boldsymbol{\theta}}_{\mathcal{W}})$ to be zero. The challenge with this condition is that it is black or white: either the $\mathcal{W}$ contains the range of the Jacobian or it does not. In the next two subsections we suggest other requirements on $\mathcal{W}$ that allow more shades of gray.

2.2. Proximity of steps. Another result that provides insight into the choice of subspace $\mathcal{W}$ comes from considering the Gauss–Newton step [26, section 10.3] for the full and projected problems at $\boldsymbol{\theta}$:

$$\mathbf{s} = -\mathbf{J}(\boldsymbol{\theta})^+\mathbf{r}(\boldsymbol{\theta}) \quad \text{(full)}, \qquad \mathbf{s}_{\mathcal{W}} = -[\mathbf{P}_{\mathcal{W}}\mathbf{J}(\boldsymbol{\theta})]^+\mathbf{P}_{\mathcal{W}}\mathbf{r}(\boldsymbol{\theta}) \quad \text{(projected)}, \tag{10}$$

where $\mathbf{A}^+$ denotes the pseudoinverse of $\mathbf{A}$ [11, section 5.5.2]. We bound the difference between these two steps using Lemma A.1 from Appendix A, removing terms involving the subspace angle between $\mathcal{W}$ and the span of $\mathbf{r}(\boldsymbol{\theta})$ using the upper bound of one on both the sine and cosine.

THEOREM 2.1 (Gauss–Newton step accuracy). *Let* $\mathbf{s}$ *and* $\mathbf{s}_{\mathcal{W}}$ *be the Gauss–Newton steps for the full and projected problems at* $\boldsymbol{\theta}$ *as given in* (10)*; then their mismatch is bounded by*

$$\|\mathbf{s} - \mathbf{s}_{\mathcal{W}}\|_2 \le \|\mathbf{J}(\boldsymbol{\theta})^+\|_2 \, \|\mathbf{r}(\boldsymbol{\theta})\|_2 \left[\sin\phi_q(\mathcal{W}, \mathcal{J}(\boldsymbol{\theta})) + \tan^2\phi_q(\mathcal{W}, \mathcal{J}(\boldsymbol{\theta}))\right], \tag{11}$$

where $\mathcal{J}(\boldsymbol{\theta}) := \operatorname{Range} \mathbf{J}(\boldsymbol{\theta})$.

Using this theorem, we can provide a heuristic for estimating the mismatch between the full and projected parameter estimates. Applying the Gauss–Newton method to the full problem starting at a stationary point of the projected problem $\check{\boldsymbol{\theta}}_{\mathcal{W}}$, we note the first step cannot move further than

$$\|\mathbf{s}\|_2 \le \|\mathbf{J}(\check{\boldsymbol{\theta}}_{\mathcal{W}})^+\|_2 \, \|\mathbf{r}(\check{\boldsymbol{\theta}}_{\mathcal{W}})\|_2 \left[\sin\phi_q(\mathcal{W}, \mathcal{J}(\check{\boldsymbol{\theta}}_{\mathcal{W}})) + \tan^2\phi_q(\mathcal{W}, \mathcal{J}(\check{\boldsymbol{\theta}}_{\mathcal{W}}))\right]. \tag{12}$$

Although multiple iterations of the Gauss–Newton method might be required to reach a stationary point of the full problem, if $\check{\boldsymbol{\theta}}_{\mathcal{W}}$ is sufficiently close to a stationary point $\check{\boldsymbol{\theta}}$ of the full problem, then we expect this first step to yield a good estimate; i.e., $\check{\boldsymbol{\theta}}_{\mathcal{W}} + \mathbf{s} \approx \check{\boldsymbol{\theta}}$. This suggests choosing subspaces $\mathcal{W}$ to minimize the largest subspace angle between $\mathcal{W}$ and the range of Jacobian at the stationary point of the projected problem $\mathcal{J}(\check{\boldsymbol{\theta}}_{\mathcal{W}})$ to ensure the full and projected parameter estimates are nearby.

2.3. Inexact Levenberg–Marquardt. A third and final result that provides insight into the choice of subspace $\mathcal{W}$ comes from considering steps of the Levenberg–Marquardt method [26, section 10.3] applied to the projected problem (2) as inexact steps of the Levenberg–Marquardt method applied to the full problem (1). For the full problem, the Levenberg–Marquardt method generates a sequence of iterates $\{\boldsymbol{\theta}_k\}_k$ starting from a given $\boldsymbol{\theta}_0$ using the rule

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \mathbf{s}_k, \qquad \mathbf{s}_k := \operatorname*{argmin}_{\mathbf{s}\in\mathbb{C}^q} \left\| \begin{bmatrix} \mathbf{J}(\boldsymbol{\theta}_k) \\ \lambda_k \mathbf{I} \end{bmatrix} \mathbf{s} + \begin{bmatrix} \mathbf{r}(\boldsymbol{\theta}_k) \\ \mathbf{0} \end{bmatrix} \right\|_2^2, \tag{13}$$

where λ_k has been chosen to enforce a trust region; see, e.g., [21, section 3.3.5]. Iterates of the projected problem $\{\widetilde{\boldsymbol{\theta}}_k\}_{k\ge 0}$ follow a similar update rule:

$$\widetilde{\boldsymbol{\theta}}_{k+1} = \widetilde{\boldsymbol{\theta}}_k + \widetilde{\mathbf{s}}_k, \qquad \widetilde{\mathbf{s}}_k := \operatorname*{argmin}_{\mathbf{s}\in\mathbb{C}^q} \left\| \begin{bmatrix} \mathbf{W}_k^*\mathbf{J}(\boldsymbol{\theta}_k) \\ \lambda_k \mathbf{I} \end{bmatrix} \mathbf{s} + \begin{bmatrix} \mathbf{W}_k^*\mathbf{r}(\boldsymbol{\theta}_k) \\ \mathbf{0} \end{bmatrix} \right\|_2^2, \tag{14}$$

where $\mathbf{W}_k$ is the orthonormal basis for the subspace applied at the kth step. Here we ask, Under what conditions on $\mathbf{W}_k$ does the sequence $\{\widetilde{\boldsymbol{\theta}}_k\}_k$ converge to the same point as $\{\boldsymbol{\theta}_k\}_k$? Although we are unable to prove the convergence of the projected iterates unless the subspace angles between $\mathcal{J}(\widetilde{\boldsymbol{\theta}}_k)$ and $\mathcal{W}_k$ go to zero, we invoke the convergence analysis of inexact Newton [4], and specific results for inexact Levenberg–Marquardt [41], to suggest a choice of $\mathcal{W}_k$. These convergence results require that the error in the step $\widetilde{\mathbf{s}}_k$ be bounded by a forcing sequence $\{\alpha_k\}_k$:

$$\frac{\|(\mathbf{J}(\widetilde{\boldsymbol{\theta}}_k)^*\mathbf{J}(\widetilde{\boldsymbol{\theta}}_k) + \lambda_k^2\mathbf{I})\widetilde{\mathbf{s}}_k + \mathbf{J}(\widetilde{\boldsymbol{\theta}}_k)^*\mathbf{r}(\widetilde{\boldsymbol{\theta}}_k)\|_2}{\|\mathbf{J}(\widetilde{\boldsymbol{\theta}}_k)^*\mathbf{r}(\widetilde{\boldsymbol{\theta}}_k)\|_2} \le \alpha_k < \alpha < 1. \tag{15}$$

Here we show that the quantity on the left can be bounded above in terms of the subspace angles and that this quantity can be made arbitrarily small.

To bound (15) we use a result from Appendix A applied to the augmented subspace, Jacobian, and residual in the least squares problem for $\widetilde{\mathbf{s}}_k$, (14):

$$\widehat{\mathcal{W}}_k := \operatorname{Range}\begin{bmatrix}\mathbf{W}_k & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}, \quad \widehat{\mathbf{J}}_k := \begin{bmatrix}\mathbf{J}(\widetilde{\boldsymbol{\theta}}_k)\\ \lambda_k\mathbf{I}\end{bmatrix}, \quad \widehat{\mathbf{r}}_k := \begin{bmatrix}\mathbf{r}(\boldsymbol{\theta}_k)\\ \mathbf{0}\end{bmatrix}.$$

Then applying Lemma A.2 to (14) and bounding $\cos\phi_1(\widehat{\mathcal{J}}_k, \widehat{\mathcal{W}}_k) \le 1$ yields

$$\frac{\|(\mathbf{J}(\widetilde{\boldsymbol{\theta}}_k)^*\mathbf{J}(\widetilde{\boldsymbol{\theta}}_k) + \lambda_k^2\mathbf{I})\widetilde{\mathbf{s}}_k + \mathbf{J}(\widetilde{\boldsymbol{\theta}}_k)^*\mathbf{r}(\widetilde{\boldsymbol{\theta}}_k)\|_2}{\|\mathbf{J}(\widetilde{\boldsymbol{\theta}}_k)^*\mathbf{r}(\widetilde{\boldsymbol{\theta}}_k)\|_2} \le \frac{\sin\phi_q(\widehat{\mathcal{W}}_k, \widehat{\mathcal{J}}_k)}{\cos^2\phi_q(\widehat{\mathcal{W}}_k, \widehat{\mathcal{J}}_k)}\,\frac{\|\widehat{\mathbf{J}}_k\|_2\|\mathbf{P}^\perp_{\widehat{\mathcal{J}}_k}\widehat{\mathbf{r}}_k\|_2}{\|\mathbf{J}(\widetilde{\boldsymbol{\theta}}_k)^*\mathbf{r}(\widetilde{\boldsymbol{\theta}}_k)\|_2}, \tag{16}$$

where $\widehat{\mathcal{J}}_k := \operatorname{Range}\widehat{\mathbf{J}}_k$, $\widehat{\mathcal{R}}_k := \operatorname{Range}\widehat{\mathbf{r}}_k$, and $\mathbf{P}^\perp_{\widehat{\mathcal{J}}_k}$ denotes the orthogonal projector onto the subspace perpendicular to $\widehat{\mathcal{J}}_k$. To obtain an expression in terms of subspace angles, we note the numerator can be written in terms of sines,

$$\|\mathbf{P}^\perp_{\widehat{\mathcal{J}}_k}\widehat{\mathbf{r}}_k\|_2 = \sin\phi_1(\widehat{\mathcal{J}}_k, \widehat{\mathcal{R}}_k)\|\widehat{\mathbf{r}}_k\|_2 = \sin\phi_1(\widehat{\mathcal{J}}_k, \widehat{\mathcal{R}}_k)\|\mathbf{r}(\widetilde{\boldsymbol{\theta}}_k)\|_2,$$

and similarly we can bound the denominator in terms of subspace angles,

$$\|\mathbf{J}(\widetilde{\boldsymbol{\theta}}_k)^*\mathbf{r}(\widetilde{\boldsymbol{\theta}}_k)\|_2 = \|\widehat{\mathbf{J}}_k^*\widehat{\mathbf{r}}_k\|_2 = \|\widehat{\mathbf{J}}_k^*\mathbf{P}_{\widehat{\mathcal{J}}_k}\widehat{\mathbf{r}}_k\|_2 \ge \sigma_q(\widehat{\mathbf{J}}_k)\cos\phi_1(\widehat{\mathcal{J}}_k, \widehat{\mathcal{R}}_k)\|\mathbf{r}(\widetilde{\boldsymbol{\theta}}_k)\|_2.$$

Combining these two results yields the upper bound

$$\frac{\|(\mathbf{J}(\widetilde{\boldsymbol{\theta}}_k)^*\mathbf{J}(\widetilde{\boldsymbol{\theta}}_k)+\lambda_k^2\mathbf{I})\widetilde{\mathbf{s}}_k+\mathbf{J}(\widetilde{\boldsymbol{\theta}}_k)^*\mathbf{r}(\widetilde{\boldsymbol{\theta}}_k)\|_2}{\|\mathbf{J}(\widetilde{\boldsymbol{\theta}}_k)^*\mathbf{r}(\widetilde{\boldsymbol{\theta}}_k)\|_2} \le \frac{\sin\phi_q(\widehat{\mathcal{W}}_k, \widehat{\mathcal{J}}_k)\tan\phi_1(\widehat{\mathcal{J}}_k, \widehat{\mathcal{R}}_k)}{\cos^2\phi_q(\widehat{\mathcal{W}}_k, \widehat{\mathcal{J}}_k)}\,\frac{\sigma_1(\widehat{\mathbf{J}}_k)}{\sigma_q(\widehat{\mathbf{J}}_k)}. \tag{17}$$

This result again confirms the centrality of the subspace angles between $\mathcal{W}$ and the range of the Jacobian, although in this result, it is the augmented subspace $\widehat{\mathcal{W}}_k$ and augmented Jacobian $\widehat{\mathcal{J}}_k$. By controlling the subspace $\mathcal{W}_k$, we can ensure that the bound in (17) is smaller than one so that step $\widetilde{\mathbf{s}}_k$ obeys the bound required by inexact Newton (15). This suggests that the Levenberg–Marquardt method applied to the projected problem makes progress toward solving the full problem. However, this result cannot be used online as it requires evaluating the full residual to compute $\phi_1(\widehat{\mathcal{J}}_k, \widehat{\mathcal{R}}_k)$. Nor can we use this result to guarantee convergence since Wright and Holt's convergence result for inexact Levenberg–Marquardt [41, Thm. 5] requires an additional sufficient decrease condition. The projected problem is unlikely to satisfy this additional constraint since the projected problem converges to different stationary points unless the subspace angles between $\mathcal{W}_k$ and $\mathcal{J}(\widetilde{\boldsymbol{\theta}}_k)$ go to zero as $k \to \infty$. This prompts the statistical approach we use in the next section to answer the question, How close are the projected parameter estimates to the full parameter estimates?

3. A statistical perspective.

One setting in which the nonlinear least squares problem (1) can arise is when measurements $\widetilde{\mathbf{y}}$ are the sum of $\mathbf{f}$ evaluated at some true parameters $\widehat{\boldsymbol{\theta}} \in \mathbb{C}^q$ plus Gaussian random noise $\mathbf{g}$ with zero mean and covariance $\epsilon^2\mathbf{I}$; $\widetilde{\mathbf{y}} = \mathbf{f}(\widehat{\boldsymbol{\theta}}) + \mathbf{g}$. Then the nonlinear least squares problem,

$$\widetilde{\boldsymbol{\theta}}(\mathbf{g}) := \operatorname*{argmin}_{\boldsymbol{\theta}} \|\mathbf{f}(\boldsymbol{\theta}) - (\mathbf{f}(\widehat{\boldsymbol{\theta}}) + \mathbf{g})\|_2, \tag{18}$$

yields the *maximum likelihood estimate* $\widetilde{\boldsymbol{\theta}}$ of $\widehat{\boldsymbol{\theta}}$ [32, section 2.1]. This estimate has a number of beneficial features. In the limit of large data or small noise, the estimator $\widetilde{\boldsymbol{\theta}}$ is unbiased and obtains the *Cramér–Rao lower bound*, namely, $\widetilde{\boldsymbol{\theta}}$ has the smallest possible covariance of any unbiased estimator of $\widehat{\boldsymbol{\theta}}$ [30, section 6.3]. Hence, the corresponding projected parameter estimate

$$\widetilde{\boldsymbol{\theta}}_{\mathcal{W}}(\mathbf{g}) := \operatorname*{argmin}_{\boldsymbol{\theta}} \|\mathbf{P}_{\mathcal{W}}[\ \mathbf{f}(\boldsymbol{\theta}) - (\mathbf{f}(\widehat{\boldsymbol{\theta}}) + \mathbf{g})\ ]\|_2 \tag{19}$$

must have a larger covariance. By using the inexpensive projected parameter estimate $\widetilde{\boldsymbol{\theta}}_{\mathcal{W}}$ as an alternative to full parameter estimate $\widetilde{\boldsymbol{\theta}}$, we are following in a tradition that dates back to Fisher [7, section 8]. Fisher quantified the loss of precision incurred by a particular scalar estimator by the *efficiency*: the ratio of the minimum covariance to the estimator's covariance. For our vector valued estimates, the covariance is a positive definite matrix in $\mathbb{C}^{q\times q}$, and so to obtain a scalar value for the efficiency ratio, we follow the lead of experimental design [33, section 2.1], [24, section 1.4] and consider the determinant of the covariance matrix, which leads to the *D-efficiency*

$$\widehat{\eta}(\mathcal{W}) := \frac{\det \operatorname{Cov} \widetilde{\boldsymbol{\theta}}}{\det \operatorname{Cov} \widetilde{\boldsymbol{\theta}}_{\mathcal{W}}} \in [0,1]. \tag{20}$$

In choosing our subspace $\mathcal{W}$, our goal will be to make the efficiency as large as possible so that the covariances of the estimates for $\widetilde{\boldsymbol{\theta}}_{\mathcal{W}}$ and $\widetilde{\boldsymbol{\theta}}$ are similar. However, as $\widetilde{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}_{\mathcal{W}}$ are both nonlinear functions of the noise $\mathbf{g}$, we cannot compute a closed form expression for the efficiency. Instead, following a standard approach for nonlinear experimental design [6, section 1.4], we linearize the parameter estimates $\widetilde{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}_{\mathcal{W}}$ about $\widehat{\boldsymbol{\theta}}$, yielding the *linearized D-efficiency* as derived in subsection 3.1. The main result of this section will be to connect linearized efficiency to the subspace angles between $\mathcal{W}$ and the range of the Jacobian at $\widehat{\boldsymbol{\theta}}$, $\mathcal{J}(\widehat{\boldsymbol{\theta}})$:

$$\eta(\mathcal{W}, \mathcal{J}(\widehat{\boldsymbol{\theta}})) := \frac{\det[\mathbf{J}(\widehat{\boldsymbol{\theta}})^*\mathbf{J}(\widehat{\boldsymbol{\theta}})]^{-1}}{\det[\mathbf{J}(\widehat{\boldsymbol{\theta}})^*\mathbf{P}_{\mathcal{W}}\mathbf{J}(\widehat{\boldsymbol{\theta}})]^{-1}} = \prod_{k=1}^{q} \cos^2 \phi_q(\mathcal{W}, \mathcal{J}(\widehat{\boldsymbol{\theta}})) \approx \frac{\det \operatorname{Cov} \widetilde{\boldsymbol{\theta}}}{\det \operatorname{Cov} \widetilde{\boldsymbol{\theta}}_{\mathcal{W}}}. \tag{21}$$

We will refer to η as simply the efficiency, and following Fisher we will say a subspace $\mathcal{W}$ is 95% efficient for $\widehat{\boldsymbol{\theta}}$ if $\eta(\mathcal{W}, \mathcal{J}(\widehat{\boldsymbol{\theta}})) = 0.95$. Later in section 5 we will design subspaces for the exponential fitting problem with the goal of obtaining a target efficiency. Further, note that maximizing efficiency corresponds to minimizing the covariance of $\widetilde{\boldsymbol{\theta}}_{\mathcal{W}}$, and hence selecting the subspace $\mathcal{W}$ is similar to experimental design [6, 33, 24], albeit where the design is happening after the data has been collected.

3.1. Linearized efficiency.

Here we briefly derive the linearized estimate of $\widetilde{\boldsymbol{\theta}}_{\mathcal{W}}$ and the corresponding linearized covariance; cf. [32, section 12.2.6]. In the limit of small noise $\mathbf{g}$, we expand $\widetilde{\boldsymbol{\theta}}_{\mathcal{W}}$ about the true parameters $\widehat{\boldsymbol{\theta}}$:

$$\widetilde{\boldsymbol{\theta}}_{\mathcal{W}}(\mathbf{g}) = \widehat{\boldsymbol{\theta}} + [\mathbf{W}^*\mathbf{J}(\widehat{\boldsymbol{\theta}})]^+\mathbf{W}^*\mathbf{g} + \mathcal{O}(\|\mathbf{g}\|_2^2). \tag{22}$$

Applying this first order estimate in the covariance, we have

$$\begin{aligned}\operatorname{Cov}\widetilde{\boldsymbol{\theta}}_{\mathcal{W}} &= \mathbb{E}_{\mathbf{g}}[(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}_{\mathcal{W}}(\mathbf{g}))(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}_{\mathcal{W}}(\mathbf{g}))^*] \\ &\approx \mathbb{E}_{\mathbf{g}}\left[[\mathbf{W}^*\mathbf{J}(\widehat{\boldsymbol{\theta}})]^{+*}\mathbf{W}^*\mathbf{g}\mathbf{g}^*\mathbf{W}[\mathbf{W}^*\mathbf{J}(\widehat{\boldsymbol{\theta}})]^{+}\right] = \epsilon^2[\mathbf{J}(\widehat{\boldsymbol{\theta}})^*\mathbf{W}\mathbf{W}^*\mathbf{J}(\widehat{\boldsymbol{\theta}})]^{-1}\end{aligned} \tag{23}$$

when $\operatorname{Cov}\mathbf{g} = \epsilon^2\mathbf{I}$. To obtain the D-linearized efficiency (21), we replace $\operatorname{Cov}\widetilde{\boldsymbol{\theta}}$ and $\operatorname{Cov}\widetilde{\boldsymbol{\theta}}_{\mathcal{W}}$ with the estimate above.

3.2. Relating efficiency to subspace angles. As with the optimization perspective, a good subspace from a statistical perspective will have small subspace angles between $\mathcal{W}$ and the Jacobian $\mathcal{J}(\widehat{\boldsymbol{\theta}})$. The following theorem establishes this connection.

THEOREM 3.1. *If $\mathcal{W}$ is an m-dimensional subspace of $\mathbb{C}^n$ with orthonormal basis $\mathbf{W}$ and $\mathbf{J} \in \mathbb{C}^{n\times q}$, where $m \geq q$ with $\mathcal{J} := \operatorname{Range}\mathbf{J}$, then*

$$\eta(\mathcal{W},\mathcal{J}) := \frac{\det([\mathbf{J}^*\mathbf{J}]^{-1})}{\det([\mathbf{J}^*\mathbf{W}\mathbf{W}^*\mathbf{J}]^{-1})} = \prod_{k=1}^{q}\cos^2\phi_k\left(\mathcal{W},\mathcal{J}\right), \tag{24}$$

where $\phi_k(\mathcal{A},\mathcal{B})$ is the kth principal angle between $\mathcal{A},\mathcal{B} \subset \mathbb{C}^n$, as defined in (6)*.*

Proof. Let $\mathbf{J} = \mathbf{Q}\mathbf{T}$ be the short-form QR-factorization of $\mathbf{J}$, where $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}$. Then using the multiplicative property of the determinant [14, section 0.3.5],

$$\begin{aligned}\eta(\mathcal{W},\mathcal{J}) = \frac{\det(\mathbf{J}^*\mathbf{W}\mathbf{W}^*\mathbf{J})}{\det(\mathbf{J}^*\mathbf{J})} &= \frac{\det(\mathbf{Q}^*\mathbf{W}\mathbf{W}^*\mathbf{Q})\det(\mathbf{T})\det(\mathbf{T}^*)}{\det(\mathbf{Q}^*\mathbf{Q})\det(\mathbf{T})\det(\mathbf{T}^*)} \\ &= \det(\mathbf{Q}^*\mathbf{W}\mathbf{W}^*\mathbf{Q}) = \prod_{k=1}^{q}\sigma_k(\mathbf{W}^*\mathbf{Q})^2 = \prod_{k=1}^{q}\cos^2\phi_k(\mathcal{W},\mathcal{J}). \qquad \square\end{aligned}$$

3.3. Properties of efficiency for projected problems. We conclude this section with three results about the linearized D-efficiency that aid in our construction of subspaces for exponential fitting in section 5. There, our approach will be to precompute a finite number of subspaces for a single exponential and combine these to produce subspaces for multiple exponentials.

The first result proves an intuitive fact: by enlarging the subspace $\mathcal{W}$, the efficiency will not decrease. The following theorem establishes this result, making use of the partial ordering of positive definite matrices; namely, $\mathbf{A} \succeq \mathbf{B}$ if $\mathbf{A} - \mathbf{B}$ is positive definite [14, section 7.7].

THEOREM 3.2. *If $\mathcal{W}_1 \subseteq \mathcal{W}_2$ and $\mathcal{J}$ are subspaces of $\mathbb{C}^n$, then $\eta(\mathcal{W}_1,\mathcal{J}) \leq \eta(\mathcal{W}_2,\mathcal{J})$.*

Proof. Let $\mathbf{W}_1$ and $\mathbf{W}_2$ be orthonormal bases for $\mathcal{W}_1$ and $\mathcal{W}_2$, and let $\mathbf{Q}$ be an orthonormal basis for $\mathcal{J}$. Then $\mathbf{W}_1\mathbf{W}_1^* \preceq \mathbf{W}_2\mathbf{W}_2$ and by [14, Cor. 7.7.4],

$$\eta(\mathcal{W}_1,\mathcal{J}) = \det(\mathbf{Q}^*\mathbf{W}_1\mathbf{W}_1^*\mathbf{Q}) \leq \det(\mathbf{Q}^*\mathbf{W}_2\mathbf{W}_2^*\mathbf{Q}) = \eta(\mathcal{W}_2,\mathcal{J}). \qquad \square$$

Since our subspaces for multiple exponentials will be built from a union of subspaces for each exponential, this second result shows that the union satisfies a necessary (but not sufficient) condition for the combined subspace to have the same efficiency as each component subspace had for a single exponential.

THEOREM 3.3. *If $\mathcal{W}$, $\{\mathcal{J}_k\}_k$ are subspaces of $\mathbb{C}^n$, $\mathcal{J} = \bigcup_k \mathcal{J}_k$, and the dimension of $\mathcal{W}$ exceeds $\mathcal{J}$, then $\eta(\mathcal{W},\mathcal{J}) \leq \min_k \eta(\mathcal{W},\mathcal{J}_k)$.*

Proof. Let $\mathbf{Q} = [\mathbf{Q}_1, \mathbf{Q}_2]$ be an orthonormal basis for $\mathcal{J}$, where $\mathbf{Q}_1 \in \mathbb{C}^{n\times q_1}$ and $\mathbf{Q}_2 \in \mathbb{C}^{n\times q_2}$, such that $\mathbf{Q}_1 \in \mathbb{C}^{n\times q_1}$ is a basis for $\mathcal{J}_\ell$, and let $\mathbf{W}$ be an orthonormal basis for $\mathcal{W}$. Then as $\sigma_k(\mathbf{W}^*\mathbf{Q}) \le 1$,

$$\eta(\mathcal{W}, \mathcal{J}) = \prod_{k=1}^{q_1+q_2} \sigma_k(\mathbf{W}^*\mathbf{Q})^2 \le \prod_{k=1}^{q_1} \sigma_{q_2+k}(\mathbf{W}^*\mathbf{Q})^2 \le \prod_{k=1}^{q_1} \sigma_k(\mathbf{W}^*\mathbf{Q}_1)^2 = \eta(\mathcal{W}, \mathcal{J}_\ell),$$

where the second inequality follows from deleting the last q_2 columns of $\mathbf{W}^*\mathbf{Q}$ and applying [15, Cor. 3.1.3]. The result follows by repeating this process for each $\mathcal{J}_\ell$. □

The final result provides a lower bound on the efficiency for a nearby Jacobian. This bound is used in subsection 5.1 to convert a check for efficiency over a continuous set of subspaces $\mathcal{J}(\boldsymbol{\theta})$ into a check over a discrete set.

THEOREM 3.4. *If $\mathcal{W}$, $\mathcal{J}_1$, and $\mathcal{J}_2$ are subspaces of $\mathbb{C}^n$ and $\mathcal{J}_1$ and $\mathcal{J}_2$ have the same dimension, then $\eta(\mathcal{W}, \mathcal{J}_2)\eta(\mathcal{J}_1, \mathcal{J}_2) \le \eta(\mathcal{W}, \mathcal{J}_1)$.*

Proof. Let $\mathbf{Q}_1$ and $\mathbf{Q}_2$ be orthonormal bases for $\mathcal{J}_1$ and $\mathcal{J}_2$. As $\mathbf{P}_\mathcal{W} \succeq \mathbf{P}_{\mathcal{J}_2}\mathbf{P}_\mathcal{W}\mathbf{P}_{\mathcal{J}_2}$, we obtain the lower bound after application of [14, Cor. 7.7.4]

$$\begin{aligned}\eta(\mathcal{W}, \mathcal{J}_1) = \det(\mathbf{Q}_1^*\mathbf{P}_\mathcal{W}\mathbf{Q}_1) &\ge \det(\mathbf{Q}_1^*\mathbf{Q}_2\mathbf{Q}_2^*\mathbf{P}_\mathcal{W}\mathbf{Q}_2\mathbf{Q}_2^*\mathbf{Q}_1) \\ &= \det(\mathbf{Q}_2^*\mathbf{P}_\mathcal{W}\mathbf{Q}_2)\det(\mathbf{Q}_1^*\mathbf{Q}_2\mathbf{Q}_2^*\mathbf{Q}_1) = \eta(\mathcal{W}, \mathcal{J}_2)\eta(\mathcal{J}_1, \mathcal{J}_2). \quad \square\end{aligned}$$

With these general results from the two preceding sections complete, we now turn to the specifics of the exponential fitting problem.

4. Fast inner products for exponential fitting.

In the two preceding sections, we have argued that subspaces $\mathcal{W}$ should be chosen such that the subspace angles between $\mathcal{W}$ and the range of the Jacobian are small. Now, in this section we turn to the specific problem of selecting a family of subspaces $\mathcal{W}$ for the exponential fitting problem that not only satisfy this requirement, but also have orthonormal bases $\mathbf{W}$ such that projected model $\mathbf{W}^*\mathbf{f}([\boldsymbol{\omega}, \mathbf{a}])$ and projected Jacobian $\mathbf{W}^*\mathbf{J}([\boldsymbol{\omega}, \mathbf{a}])$ can be inexpensively computed in fewer than $\mathcal{O}(n)$ operations. For the exponential fitting problem we chose the subspace $\mathcal{W}(\boldsymbol{\mu})$ parameterized by $\boldsymbol{\mu} \in \mathbb{C}^m$ with corresponding orthonormal basis $\mathbf{W}(\boldsymbol{\mu}) \in \mathbb{C}^{n\times m}$:

$$(25) \qquad \mathcal{W}(\boldsymbol{\mu}) := \operatorname{Range}\mathbf{V}(\boldsymbol{\mu}), \qquad \mathbf{W}(\boldsymbol{\mu}) := \mathbf{V}(\boldsymbol{\mu})\mathbf{R}(\boldsymbol{\mu})^{-1}, \qquad \mathbf{R}(\boldsymbol{\mu}) \in \mathbb{C}^{m\times m},$$

where $\mathbf{V}(\boldsymbol{\mu}) \in \mathbb{C}^{n\times m}$ is the Vandermonde matrix $[\mathbf{V}(\boldsymbol{\mu})]_{j,k} = e^{j\mu_k}$ and $\mathbf{R}(\boldsymbol{\mu})$ is constructed as described in subsection 4.3 such that $\mathbf{W}(\boldsymbol{\mu})$ has orthonormal columns. We call the parameters $\boldsymbol{\mu}$ *interpolation points* since if the entries of $\boldsymbol{\omega}$ are a subset of the entries of $\boldsymbol{\mu}$, then the projected model interpolates the full model:

$$(26) \qquad \boldsymbol{\omega} \subset \boldsymbol{\mu} \quad \Longrightarrow \quad \mathbf{P}_{\mathcal{W}(\boldsymbol{\mu})}\mathbf{f}([\boldsymbol{\omega}, \mathbf{a}]) = \mathbf{P}_{\operatorname{Range}\mathbf{V}(\boldsymbol{\mu})}\mathbf{V}(\boldsymbol{\omega})\mathbf{a} = \mathbf{V}(\boldsymbol{\omega})\mathbf{a} = \mathbf{f}([\boldsymbol{\omega}, \mathbf{a}]).$$

In this section we show how to inexpensively compute the product of $\mathbf{W}(\boldsymbol{\mu})$ with the exponential fitting model $\mathbf{f}([\boldsymbol{\omega}, \mathbf{a}])$ and Jacobian $\mathbf{J}([\boldsymbol{\omega}, \mathbf{a}])$,

$$(27) \quad \mathbf{J}([\boldsymbol{\omega}, \mathbf{a}]) := \begin{bmatrix}\mathbf{V}'(\boldsymbol{\omega})\operatorname{diag}(\mathbf{a}) & \mathbf{V}(\boldsymbol{\omega})\end{bmatrix}, \qquad [\mathbf{V}'(\boldsymbol{\omega})]_{j,k} = \frac{\partial}{\partial \omega_k}[\mathbf{V}(\boldsymbol{\omega})]_{j,k} = je^{j\omega_k}.$$

Examining the products $\mathbf{W}(\boldsymbol{\mu})^*\mathbf{f}([\boldsymbol{\omega}, \mathbf{a}])$ and $\mathbf{W}(\boldsymbol{\mu})^*\mathbf{J}([\boldsymbol{\omega}, \mathbf{a}])$,

$$(28) \qquad \mathbf{W}(\boldsymbol{\mu})^*\mathbf{f}([\boldsymbol{\omega}, \mathbf{a}]) = \mathbf{R}(\boldsymbol{\mu})^{-*}\mathbf{V}(\boldsymbol{\mu})^*\mathbf{V}(\boldsymbol{\omega})\mathbf{a},$$

$$(29) \qquad \mathbf{W}(\boldsymbol{\mu})^*\mathbf{J}([\boldsymbol{\omega}, \mathbf{a}]) = \begin{bmatrix}\mathbf{R}(\boldsymbol{\mu})^{-*}\mathbf{V}(\boldsymbol{\mu})^*\mathbf{V}'(\boldsymbol{\omega})\operatorname{diag}(\mathbf{a}) & \mathbf{R}(\boldsymbol{\mu})^{-*}\mathbf{V}(\boldsymbol{\mu})^*\mathbf{V}(\boldsymbol{\omega})\end{bmatrix},$$

reveals two matrix multiplications of size n, $\mathbf{V}(\boldsymbol{\mu})^*\mathbf{V}(\boldsymbol{\omega})$ and $\mathbf{V}(\boldsymbol{\mu})^*\mathbf{V}'(\boldsymbol{\omega})$, that need to be inexpensively computed. Here we use the geometric sum formula and generalization provided by Theorem B.2 in Appendix B to compute the entries of these products in closed form. Unfortunately, these formulas exhibit catastrophic cancellation in finite precision arithmetic, necessitating careful modifications to obtain high relative accuracy as described in subsection 4.1 for $\mathbf{V}(\boldsymbol{\mu})^*\mathbf{V}(\boldsymbol{\omega})$ and in subsection 4.2 for $\mathbf{V}(\boldsymbol{\mu})^*\mathbf{V}'(\boldsymbol{\omega})$. Additionally, we discuss how to compute $\mathbf{R}(\boldsymbol{\mu})$ inexpensively from $\mathbf{V}(\boldsymbol{\mu})^*\mathbf{V}(\boldsymbol{\mu})$ in subsection 4.3. The choice of interpolation points $\boldsymbol{\mu}$ is later discussed in section 5 and combined with these results yields our algorithm for exponential fitting described in section 6.

4.1. Geometric sum. Each entry in the product of two Vandermonde matrices $\mathbf{V}(\boldsymbol{\mu})^*\mathbf{V}(\boldsymbol{\omega})$ is a geometric sum and hence has a closed form expression via the *geometric sum formula*:

$$(30) \qquad [\mathbf{V}(\boldsymbol{\mu})^*\mathbf{V}(\boldsymbol{\omega})]_{j,k} = \sum_{\ell=0}^{n-1} e^{\overline{\mu}_j \ell} e^{\omega_k \ell} = \begin{cases} \dfrac{1-e^{n(\overline{\mu}_j+\omega_k)}}{1-e^{\overline{\mu}_j+\omega_k}}, & e^{\overline{\mu}_j+\omega_k} \neq 1; \\ n, & e^{\overline{\mu}_j+\omega_k} = 1. \end{cases}$$

In finite precision arithmetic this formula exhibits catastrophic cancellation when $e^{\overline{\mu}_j+\omega_k} \approx 1$. Fortunately, many standard libraries provide the special function `expm1` that evaluates $e^x - 1$ to high relative accuracy. However, even with this special function, there is still a removable discontinuity at $e^{\overline{\mu}_j+\omega_k} = 1$. Hence in floating point we patch this function using a two-term Taylor series expansion around $e^{\overline{\mu}_j+\omega_k} = 1$:

$$(31) \quad [\mathbf{V}(\boldsymbol{\mu})^*\mathbf{V}(\boldsymbol{\omega})]_{j,k} = \begin{cases} \dfrac{\texttt{expm1}(n(\overline{\mu}_j+\omega_k))}{\texttt{expm1}(\overline{\mu}_j+\omega_k)}, & |\,\texttt{expm1}(\overline{\mu}_j+\omega_k)| > 10^{-15}; \\ n(1+(n-1)(\overline{\mu}_j+\omega_k)/2), & |\,\texttt{expm1}(\overline{\mu}_j+\omega_k)| \leq 10^{-15}. \end{cases}$$

In our numerical experiments, this expression has a relative accuracy of $\sim 10^{-16}$ when compared to a 500-digit reference evaluation of (30) using `mpmath` [19].

4.2. Geometric sum derivative. Entries of the product $\mathbf{V}'(\boldsymbol{\mu})^*\mathbf{V}(\boldsymbol{\omega})$ are no longer a geometric sum, but a *generalized geometric sum* that has a closed form expression given by Theorem B.2 in Appendix B:

$$(32) \qquad \begin{aligned} [\mathbf{V}(\boldsymbol{\mu})^*\mathbf{V}'(\boldsymbol{\omega})]_{j,k} &= \sum_{\ell=0}^{n-1} \ell e^{\overline{\mu}_j \ell} e^{\omega_k \ell} \\ &= \begin{cases} \dfrac{-ne^{n(\overline{\mu}_j+\omega_k)}}{1-e^{\overline{\mu}_j+\omega_k}} + \dfrac{e^{\omega_k+\overline{\mu_j}}(1-e^{n(\omega_k+\overline{\mu_j})})}{(1-e^{\omega_k+\overline{\mu_j}})^2}, & e^{\omega_k+\overline{\mu_j}} \neq 1; \\ n(n-1)/2, & e^{\omega_k+\overline{\mu_j}} = 1. \end{cases} \end{aligned}$$

As with the geometric sum formula, this expression also exhibits catastrophic cancellation, but this can no longer be fixed using standard special functions. Instead, we derive a more accurate expression in floating point arithmetic by rearranging the expression in the first case and using a Taylor series about $e^{\overline{\mu}_j+\omega_k} = 1$ in the second. Defining $\delta_{j,k} := \overline{\mu}_j + \omega_k \in \mathbb{R} \times [-\pi/2, \pi/2)i$ (removing periodicity in the imaginary part), the first case of (32) can be rearranged to yield

$$(33) \qquad \frac{-ne^{n\delta_{j,k}}}{1-e^{\delta_{j,k}}} + \frac{e^{\delta_{j,k}}(1-e^{n\delta_{j,k}})}{(1-e^{\delta_{j,k}})^2} = \frac{1-e^{n\delta_{j,k}}}{1-e^{\delta_{j,k}}}\left[\frac{e^{\delta_{j,k}}}{1-e^{\delta_{j,k}}} - \frac{ne^{n\delta_{j,k}}}{1-e^{n\delta_{j,k}}}\right].$$

Although the expression on the right displays even worse catastrophic cancellation than the expression on the left, the first term can be computed using `expm1`, and the expression inside the brackets has a rapidly converging Taylor series:

$$\frac{e^{\delta}}{1-e^{\delta}} - \frac{ne^{n\delta}}{1-e^{n\delta}} = \frac{n-1}{2} + \frac{(n^2-1)\delta}{12} - \frac{(n^4-1)\delta^3}{720} + \frac{(n^6-1)\delta^5}{30240} - \frac{(n^8-1)\delta^7}{1209600} + \frac{(n^{10}-1)\delta^9}{47900160} - \frac{691(n^{12}-1)\delta^{11}}{1307674368000} + \mathcal{O}(n^{14}\delta^{13}).$$

Calling the first seven terms of this expansion the special function $\texttt{expdiff}(n, \delta)$, we then evaluate the product $\mathbf{V}(\boldsymbol{\mu})^*\mathbf{V}'(\boldsymbol{\omega})$ in finite precision arithmetic using

$$[\mathbf{V}(\boldsymbol{\mu})^*\mathbf{V}'(\boldsymbol{\omega})]_{j,k} = \begin{cases} \dfrac{ne^{n\delta_{j,k}}\,\texttt{expm1}(\delta_{j,k}) - e^{\delta_{j,k}}\,\texttt{expm1}(\delta_{j,k}n)}{[\texttt{expm1}(\delta_{j,k})]^2}, & |\delta_{j,k}| > 0.5/n; \\ \dfrac{\texttt{expm1}(n\delta_{j,k})}{\texttt{expm1}(\delta_{j,k})}\,\texttt{expdiff}(n, \delta_{j,k}), & 0 < |\delta_{j,k}| \le 0.5/n; \\ n(n-1)/2, & \delta_{j,k} = 0. \end{cases} \tag{34}$$

In our numerical experiments, this expression has a relative accuracy of $\sim 10^{-15}$ when compared to a 500-digit reference evaluation of (33) using `mpmath`.

4.3. Orthogonalization. Finally we need to inexpensively compute the matrix $\mathbf{R}(\boldsymbol{\mu})$ such that $\mathbf{V}(\boldsymbol{\mu})\mathbf{R}(\boldsymbol{\mu})^{-1}$ has orthonormal columns. One approach would be to simply take the QR-factorization of $\mathbf{V}(\boldsymbol{\mu})$, but this has an $\mathcal{O}(n)$-dependent cost. Instead, our approach is to form $\mathbf{V}(\boldsymbol{\omega})^*\mathbf{V}(\boldsymbol{\omega})$ using (30) and take either its Cholesky decomposition or its eigendecomposition to compute $\mathbf{R}(\boldsymbol{\mu})$:

$$\mathbf{V}(\boldsymbol{\mu})^*\mathbf{V}(\boldsymbol{\mu}) = \mathbf{R}(\boldsymbol{\mu})\mathbf{R}(\boldsymbol{\mu})^* \qquad \text{(Cholesky)},$$
$$\mathbf{V}(\boldsymbol{\mu})^*\mathbf{V}(\boldsymbol{\mu}) = \mathbf{U}(\boldsymbol{\mu})\boldsymbol{\Lambda}(\boldsymbol{\mu})\mathbf{U}(\boldsymbol{\mu})^* \qquad \text{(eigendecomposition)}, \quad \mathbf{R}(\boldsymbol{\mu}) = \mathbf{U}(\boldsymbol{\mu})\boldsymbol{\Lambda}(\boldsymbol{\mu})^{1/2}.$$

Although the Cholesky decomposition should be preferred since $\mathbf{V}(\boldsymbol{\mu})^*\mathbf{V}(\boldsymbol{\mu})$ is positive definite provided each of the $\{e^{\mu_j}\}_j$ is distinct, in finite precision arithmetic this product can have small negative eigenvalues. Instead we compute $\mathbf{R}(\boldsymbol{\mu})^{-1}$ using the eigendecomposition, truncating the small ($< 10^{-14}$) eigenvalues.

5. A subspace for exponential fitting.

With the results of the previous section, we can now inexpensively project the model and Jacobian onto the subspace $\mathcal{W}(\boldsymbol{\mu})$. However, this leaves one question: how do we choose the interpolation points $\boldsymbol{\mu}$ such that the subspace angles between $\mathcal{W}(\boldsymbol{\mu})$ and the exponential fitting Jacobian,

$$\mathbf{J}([\boldsymbol{\omega}, \mathbf{a}]) = \begin{bmatrix}\mathbf{V}'(\boldsymbol{\omega})\operatorname{diag}(\mathbf{a}) & \mathbf{V}(\boldsymbol{\omega})\end{bmatrix}, \qquad [\mathbf{V}'(\boldsymbol{\omega})]_{j,k} = \frac{\partial}{\partial\omega_k}[\mathbf{V}(\boldsymbol{\omega})]_{j,k} = je^{j\omega_k}, \tag{35}$$

are small? Immediately we note that these subspace angles do not depend on $\mathbf{a}$ (if any entry of $\mathbf{a}$ was zero, we would instead fit fewer exponentials); hence we define

$$\mathcal{J}(\boldsymbol{\omega}) := \operatorname{Range}\begin{bmatrix}\mathbf{V}'(\boldsymbol{\omega}) & \mathbf{V}(\boldsymbol{\omega})\end{bmatrix}. \tag{36}$$

Structurally, it may seem impossible to have small subspace angles between $\mathcal{W}(\boldsymbol{\mu})$ and $\mathcal{J}(\boldsymbol{\omega})$ since $\mathcal{W}(\boldsymbol{\mu}) = \operatorname{Range}\mathbf{V}(\boldsymbol{\mu})$ does not contain any columns from $\mathbf{V}'$. However, since columns of $\mathbf{V}'$ are the derivatives of the columns of $\mathbf{V}$, we can approximate the range of $\mathbf{V}'$ using small, finite difference steps δ_- and δ_+:

$$\mathbf{V}(\omega) \approx \frac{\mathbf{V}(\omega+\delta_+) + \mathbf{V}(\omega+\delta_-)}{2}, \qquad \mathbf{V}'(\omega) \approx \frac{\mathbf{V}(\omega+\delta_+) - \mathbf{V}(\omega+\delta_-)}{\delta_+ - \delta_-}. \tag{37}$$

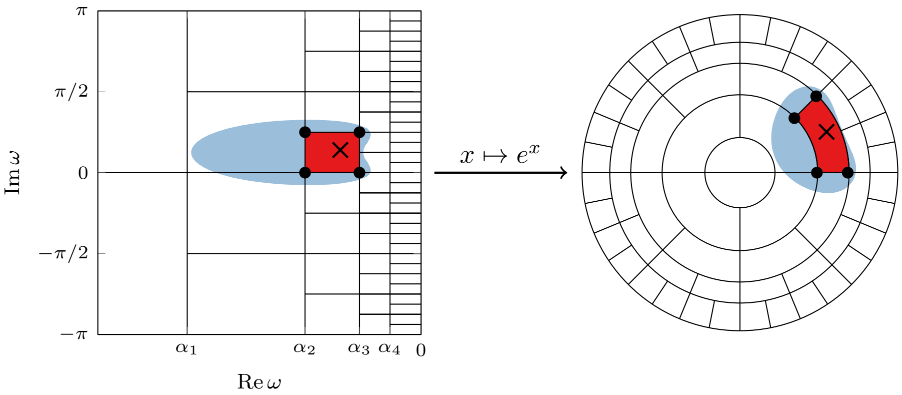

FIG. 2. *An illustration of the box partition of the parameter space* $\boldsymbol{\omega} \in (-\infty, 0] \times [-\pi, \pi)i$ *for exponential fitting. For a given exponential with frequency* ω *denoted by* ×*, we select the box containing it, denoted in red, and the corresponding four interpolation points at the corners, denoted by* •*. The blue shaded region shows the set of* ω *where the efficiency at* ω *using this subspace is at least* 95%*. This region includes the entire box and extends outward. The figure on the right shows the same features under the exponential map, exposing the periodicity of the parameter space. (Color available online.)*

Hence, for an appropriate choice of interpolation points, the subspace angles between $\mathcal{W}(\boldsymbol{\mu})$ and $\mathcal{J}(\boldsymbol{\omega})$ are small; for example, the interpolation points $\delta_{\pm}(\omega) = 0.8\operatorname{Re}\omega \pm \max\{-0.52\operatorname{Re}\omega, 1.39/n\}$ yield a subspace with 95% efficiency for any $\omega \in (-\infty, 0] \times [-\pi, \pi)i$. Here our approach for selecting interpolation points is to divide the parameter space for a single exponential with frequency $\omega \in (-\infty, 0] \times [-\pi, \pi)$ into a series of boxes as shown in Figure 2. These boxes have been constructed such that when the corners of the box containing ω are taking as interpolation points, the efficiency of this subspace is at least 95% and hence the subspace angles between $\mathcal{W}(\boldsymbol{\mu})$ and $\mathcal{J}(\boldsymbol{\omega})$ are small. Then, for multiple exponentials, we simply combine the subspaces generated by this heuristic, justified by Theorem 3.3 that this combination is a necessary condition for the combined subspace to also have at least 95% efficiency. There are several advantages to this box partition approach. By limiting ourselves to a finite number of interpolation points, we can frequently reuse the multiplication $\mathbf{V}(\mu_j)^*\widetilde{\mathbf{y}}$ as the subspace updates. Moreover, constructing this subspace is inexpensive, allowing frequent updates during optimization. In the remainder of this section we first discuss a practical algorithm for building this box partition for any target efficiency and then give the coordinates for box partition with 95% efficiency.

5.1. Building the partition. Our goal in constructing the box partition is to guarantee that the target efficiency η_{target} is obtained for every single exponential with frequency ω inside the box. This is an expensive task, so we build our boxes to simplify the verification process. As shown in Figure 2, our box partition consists of a series of stacks where each stack has twice as many boxes as the one on its left, evenly dividing the imaginary component of the parameter space. Then, as efficiency depends on $\delta_{j,k} = \overline{\mu}_j + \omega_k$ (cf. (30) and (32)), simultaneously shifting the imaginary parts of μ_j and ω_k does not change $\delta_{j,k}$ and hence verifying that one box in the stack obtains the target efficiency for each ω inside establishes the same for the remaining boxes in the stack. With this construction, there is only one set of free parameters: the real coordinates of each box $\{\alpha_\ell\}_{\ell\geq 0}$. We choose these α_ℓ, starting from $\alpha_0 = -\infty$, by making α_ℓ as large as possible while still obtaining the target efficiency inside each box:

$$\begin{aligned}\alpha_\ell = \underset{\alpha > \alpha_{\ell-1}}{\text{maximize}}\ \alpha \quad \text{such that} \quad \eta_{\min} &\le \underset{\omega\in[\alpha_{\ell-1},\alpha]\times[0,2\pi/2^\ell]i}{\text{minimize}} \eta(\mathcal{W}(\boldsymbol{\mu}),\mathcal{J}(\omega)),\\ \text{where} \quad \boldsymbol{\mu} &= \begin{bmatrix}\alpha_{\ell-1} & \alpha_{\ell-1}+2\pi/2^\ell i & \alpha & \alpha+2\pi/2^\ell i\end{bmatrix}.\end{aligned} \tag{38}$$

This is a challenging nested optimization problem over a continuous set of ω, so we invoke Theorem 3.4 to construct an auxiliary grid of exponentials where obtaining a slightly higher efficiency at these discrete points guarantees the target efficiency is reached for any ω inside the box.

To construct this auxiliary grid, we specify a series of real parts $a_j \in \mathbb{R}$ and imaginary spacings $b_j \in \mathbb{R}_+$ that define the grid points $z_{j,k} := a_j + ikb_j$. To specify a_j and b_j with a grid efficiency of η_{grid}, starting from $a_0 = \alpha_\ell$ we solve the single variable finding root problem that yields a_j and b_j,

$$\eta(\mathcal{J}(a_j),\mathcal{J}(a+(1+1i)c)) = \eta_{\text{grid}} \quad \Rightarrow \quad a_{j+1} := a_j + c, \quad b_{j+1} := c, \tag{39}$$

setting $b_0 = b_1$. Then, invoking Theorem 3.4, we have the bound

$$\underset{\omega\in[\alpha_{\ell-1},\alpha]\times[0,2\pi/2^\ell]i}{\text{minimize}} \eta(\mathcal{W}(\boldsymbol{\mu}),\mathcal{J}(\omega)) \le \underset{\substack{j,k\in\mathbb{Z}_+\\ z_{j,k}\in[\alpha_{\ell-1},\alpha]\times[0,2\pi/2^\ell]i}}{\text{minimize}} \eta_{\text{grid}}\cdot\eta(\mathcal{W}(\boldsymbol{\mu}),\mathcal{J}(z_{j,k})). \tag{40}$$

Substituting this bound in (38) replaces the inner optimization with finding the minimum over a discrete set, simplifying the problem. Further, since the accuracy of the efficiency computation is limited by the grid, we restrict the maximization over α to the discrete set of grid points a_j.

5.2. The 95% efficiency partition. Here we provide the coordinates for a box partition with a target efficiency of 95% constructed using $\eta_{\text{grid}} = 0.99999$ in Table 1 for multiple values of n. In practice, we restrict our interpolation points to the closed left half plane and hence set the first α_ℓ greater than zero to zero. Although this choice of target efficiency was arbitrary, it does make the rightmost interpolation points correspond to the nth roots of unity that appear in the discrete Fourier transform (DFT).

Although Table 1 only displays the coordinates for several values of n, two patterns emerge that allow us to estimate the box partition for any n. First note that the values for α_ℓ when $n \ne \infty$ match those for $n = \infty$ for all but the last two, which are always larger. Hence the values of α_ℓ for $n = \infty$ are a lower bound on those for arbitrary n. The other pattern is that after the first five, the α_ℓ for $n = \infty$ shrink exponentially with

$$\alpha_\ell \approx -2.9720 \cdot 2^{-\ell}, \quad \ell \ge 5. \tag{41}$$

These two patterns allow us to pick the box partition using α_ℓ from the $n = \infty$ case, extending this sequence using the approximation above for larger values of ℓ.

When n is not a power of two, the box partition will no longer have the nth roots of unity available as interpolation points. Due to their connection with the DFT, we prefer to keep nth roots of unity available and thus modify the construction of the box partition. For an n that is not a power of two, everything remains the same except when ω is in the rightmost stack, $\operatorname{Re}\omega \in (\alpha_{\widehat{\ell}}, 0]$. In this case we no longer use boxes, but pick the two closest interpolation points with $\operatorname{Re}\mu = \alpha_{\widehat{\ell}}$ from the stack to the left and the two closest nth roots of unity where $\operatorname{Re}\mu = 0$. Although we are no longer able to guarantee 95% efficiency for exponentials in this range, this heuristic still provides a high efficiency subspace in practice.

Table 1

The real coordinates α_ℓ of the box partition determined by solving (38) with $\eta_{grid} = 0.99999$. With this discretization, these values are accurate to approximately three digits.

ℓ	$n = 16$	$n = 256$	$n = 1024$	$n = 2^{20}$	$n = \infty$
1	$-1.421 \cdot 10^{0}$	$-1.421 \cdot 10^{0}$	$-1.421 \cdot 10^{0}$	$-1.421 \cdot 10^{0}$	$-1.421 \cdot 10^{0}$
2	$-6.667 \cdot 10^{-1}$	$-6.667 \cdot 10^{-1}$	$-6.667 \cdot 10^{-1}$	$-6.667 \cdot 10^{-1}$	$-6.667 \cdot 10^{-1}$
3	$-3.480 \cdot 10^{-1}$	$-3.529 \cdot 10^{-1}$	$-3.529 \cdot 10^{-1}$	$-3.529 \cdot 10^{-1}$	$-3.529 \cdot 10^{-1}$
4	$8.121 \cdot 10^{-2}$	$-1.819 \cdot 10^{-1}$	$-1.819 \cdot 10^{-1}$	$-1.819 \cdot 10^{-1}$	$-1.819 \cdot 10^{-1}$
5		$-9.198 \cdot 10^{-2}$	$-9.198 \cdot 10^{-2}$	$-9.198 \cdot 10^{-2}$	$-9.198 \cdot 10^{-2}$
6		$-4.617 \cdot 10^{-2}$	$-4.617 \cdot 10^{-2}$	$-4.617 \cdot 10^{-2}$	$-4.617 \cdot 10^{-2}$
7		$-2.294 \cdot 10^{-2}$	$-2.313 \cdot 10^{-2}$	$-2.313 \cdot 10^{-2}$	$-2.313 \cdot 10^{-2}$
8		$3.552 \cdot 10^{-3}$	$-1.157 \cdot 10^{-2}$	$-1.157 \cdot 10^{-2}$	$-1.157 \cdot 10^{-2}$
9			$-5.748 \cdot 10^{-3}$	$-5.782 \cdot 10^{-3}$	$-5.782 \cdot 10^{-3}$
10			$8.713 \cdot 10^{-4}$	$-2.891 \cdot 10^{-3}$	$-2.891 \cdot 10^{-3}$
11				$-1.445 \cdot 10^{-3}$	$-1.445 \cdot 10^{-3}$
12				$-7.227 \cdot 10^{-4}$	$-7.227 \cdot 10^{-4}$
13				$-3.613 \cdot 10^{-4}$	$-3.613 \cdot 10^{-4}$
14				$-1.807 \cdot 10^{-4}$	$-1.807 \cdot 10^{-4}$
15				$-9.033 \cdot 10^{-5}$	$-9.033 \cdot 10^{-5}$
16				$-4.516 \cdot 10^{-5}$	$-4.516 \cdot 10^{-5}$
17				$-2.258 \cdot 10^{-5}$	$-2.258 \cdot 10^{-5}$
18				$-1.129 \cdot 10^{-5}$	$-1.129 \cdot 10^{-5}$
19				$-5.611 \cdot 10^{-6}$	$-5.645 \cdot 10^{-6}$
20				$8.605 \cdot 10^{-7}$	$-2.822 \cdot 10^{-6}$

6. A projected exponential fitting algorithm.

Equipped with subspace $\mathcal{W}(\boldsymbol{\mu})$ for which the projected model $\mathbf{W}(\boldsymbol{\mu})^*\mathbf{f}([\boldsymbol{\omega}, \mathbf{a}])$ and Jacobian $\mathbf{W}(\boldsymbol{\mu})^*\mathbf{J}([\boldsymbol{\omega}, \mathbf{a}])$ can be inexpensively computed as described in section 4 and combined with the heuristic from section 5 to pick interpolation points $\boldsymbol{\mu}$ so that the subspace angles between $\mathcal{W}(\boldsymbol{\mu})$ and the range of the Jacobian $\mathcal{J}(\boldsymbol{\omega})$ are small, we now construct an algorithm to solve the exponential fitting problem using projected nonlinear least squares. Our basic approach is to solve a sequence of projected nonlinear least squares problems using Levenberg–Marquardt and infrequently update the subspace during the course of optimization. In this section we describe several important details for this algorithm. First, in subsection 6.1 we show how *variable projection* [9, 10] can be used to implicitly solve for the amplitudes $\mathbf{a}$ revealing an optimization problem over frequencies $\boldsymbol{\omega}$ alone; then in subsection 6.2 we discuss the details of how to update the subspace; and finally in subsection 6.3 we show how to obtain initial estimates of the frequencies $\boldsymbol{\omega}$ to enable a fair comparison with subspace based methods which do not require initial estimates.

6.1. Variable projection.

The key insight behind variable projection originated in a Ph.D. thesis by Scolnik on the exponential fitting problem [31]. Recognizing the optimal linear parameters $\mathbf{a}$ are given by the pseudoinverse for a fixed $\boldsymbol{\omega}$, $\mathbf{a} = \mathbf{V}(\boldsymbol{\omega})^+\widetilde{\mathbf{y}}$, allows the residual to be stated as a function of $\boldsymbol{\omega}$ alone:

$$\mathbf{r}([\boldsymbol{\omega}, \mathbf{a}]) = \mathbf{f}([\boldsymbol{\omega}, \mathbf{a}]) - \widetilde{\mathbf{y}} = \mathbf{V}(\boldsymbol{\omega})\mathbf{a} - \widetilde{\mathbf{y}} \Rightarrow \left[\mathbf{V}(\boldsymbol{\omega})\mathbf{V}(\boldsymbol{\omega})^+ - \mathbf{I}\right]\widetilde{\mathbf{y}} = \mathbf{P}^\perp_{\mathbf{V}(\boldsymbol{\omega})}\widetilde{\mathbf{y}}, \tag{42}$$

where $\mathbf{P}^\perp_{\mathbf{V}(\boldsymbol{\omega})}$ is the orthogonal projector onto the subspace perpendicular to the range of $\mathbf{V}(\boldsymbol{\omega})$. This allows us to define an equivalent optimization problem over $\boldsymbol{\omega}$ alone:

$$\underset{\boldsymbol{\omega} \in \mathbb{C}^p}{\text{minimize}} \|\widehat{\mathbf{r}}(\boldsymbol{\omega})\|_2^2, \qquad \widehat{\mathbf{r}}(\boldsymbol{\omega}) := \mathbf{P}^\perp_{\mathbf{V}(\boldsymbol{\omega})}\widetilde{\mathbf{y}}. \tag{43}$$

Golub and Pereyra [10] provide the Jacobian for this variable projection residual $\widehat{\mathbf{r}}$,

$$[\widehat{\mathbf{J}}(\boldsymbol{\omega})]_{\cdot,k} := -\left[\mathbf{P}^{\perp}_{\mathbf{V}(\boldsymbol{\omega})}\frac{\partial \mathbf{V}(\boldsymbol{\omega})}{\partial \omega_k}\mathbf{V}(\boldsymbol{\omega})^{-} + \mathbf{V}(\boldsymbol{\omega})^{-*}\left(\frac{\partial \mathbf{V}(\boldsymbol{\omega})}{\partial \omega_k}\right)^{*}\mathbf{P}^{\perp}_{\mathbf{V}(\boldsymbol{\omega})}\right]\widetilde{\mathbf{y}}, \tag{44}$$

and we can further exploit the structure of the exponential fitting problem to reveal a simple expression for this Jacobian. Defining the short-form QR-factorization of $\mathbf{V}(\boldsymbol{\omega})$, $\mathbf{V}(\boldsymbol{\omega}) = \mathbf{Q}(\boldsymbol{\omega})\mathbf{T}(\boldsymbol{\omega})$, this Jacobian becomes

$$\widehat{\mathbf{J}}(\boldsymbol{\omega}) = [\mathbf{I} - \mathbf{Q}(\boldsymbol{\omega})\mathbf{Q}(\boldsymbol{\omega})^*]\,\mathbf{V}'(\boldsymbol{\omega})\operatorname{diag}(\mathbf{V}(\boldsymbol{\omega})^{+}\widetilde{\mathbf{y}}) - \mathbf{Q}(\boldsymbol{\omega})\mathbf{T}(\boldsymbol{\omega})^{-*}\operatorname{diag}(\mathbf{V}'(\boldsymbol{\omega})^*\widehat{\mathbf{r}}(\boldsymbol{\omega})). \tag{45}$$

With the linear parameters removed, the Levenberg–Marquardt method can then be applied to the variable projection residual $\widehat{\mathbf{r}}(\boldsymbol{\omega})$ and Jacobian $\widehat{\mathbf{J}}(\boldsymbol{\omega})$. The same expressions also apply to the projected problem upon making the substitutions: $\mathbf{V}(\boldsymbol{\omega}) \to \mathbf{W}(\boldsymbol{\mu})^*\mathbf{V}(\boldsymbol{\omega})$, $\mathbf{V}'(\boldsymbol{\omega}) \to \mathbf{W}(\boldsymbol{\mu})^*\mathbf{V}'(\boldsymbol{\omega})$, and $\widetilde{\mathbf{y}} \to \mathbf{W}(\boldsymbol{\mu})^*\widetilde{\mathbf{y}}$.

6.2. Updating subspaces. As the analysis in section 2 suggests that the subspace angles between $\mathcal{W}(\boldsymbol{\mu})$ and $\mathcal{J}(\boldsymbol{\omega})$ need to remain small, we repeatedly update the interpolation points during the course of optimization. Here we use the efficiency based heuristic described in section 5 to pick interpolation points, as a high efficiency ensures small subspace angles between $\mathcal{W}(\boldsymbol{\mu})$ and $\mathcal{J}(\boldsymbol{\omega})$. However, rather than discarding interpolation points no longer required by this heuristic, we preserve them, continually expanding the subspace. This is necessary to prevent the optimization algorithm from entering a cycle.

6.3. Initialization. A final issue concerns how we provide the initial values of the optimization algorithm. Subspace based methods do not require these initial values, and so to provide a fair comparison we use a simple initialization heuristic. It is well known that peaks in the DFT of a signal, $\mathbf{F}_n^*\widetilde{\mathbf{y}}$, where $[\mathbf{F}_n]_{j,k} = n^{-1/2}e^{2\pi i jk/n}$, correspond to the frequencies present [34]. This forms the foundation of many initialization approaches. For example, in magnetic resonance spectroscopy these peaks can be identified manually [38, section 3.3] to initialize an optimization algorithm. Here, we pick the initial estimate iteratively. Starting with the first exponential, we set $\omega_1 = 2\pi i\widehat{k}/n$, where $\widehat{k}$ is the largest entry in $\mathbf{F}_n^*\widetilde{\mathbf{y}}$. Then after the optimization algorithm has terminated, we initialize ω_2 based on the largest entry of the residual $\mathbf{F}_n^*\widehat{\mathbf{r}}(\boldsymbol{\omega})$. This process repeats until the desired number of exponentials have been recovered. This approach is similar to that of Macleod [23], but we optimize all the frequencies $\boldsymbol{\omega}$ at each step.

7. A numerical example.

To demonstrate the effectiveness of projected nonlinear least squares for exponential fitting, we apply the algorithm described in section 6 to a magnetic resonance spectroscopy test problem from [39, Table 1]; see, e.g., [12, section 12.4] for a discussion of the underlying physics. This example describes a continuous complex signal $y(t)$ consisting of eleven exponentials:

$$\begin{aligned} &y(t) = \sum_{k=1}^{11} a_k e^{135 i\pi/180} e^{(2i\pi f_k - d_k)t}, \qquad \text{where} \\ &\begin{array}{llrrrrrrrrrrrl} \mathbf{a} & = [& 75 & 150 & 75 & 150 & 150 & 150 & 150 & 150 & 1400 & 60 & 500 &], \\ \mathbf{f} & = [& -86 & -70 & -54 & 152 & 168 & 292 & 308 & 360 & 440 & 490 & 530 &], \\ \mathbf{d} & = [& 50 & 50 & 50 & 50 & 50 & 50 & 50 & 25 & 285.7 & 25 & 200 &], \end{array} \end{aligned} \tag{46}$$

from which we construct measurements $\widetilde{\mathbf{y}} \in \mathbb{C}^n$ by sampling $y(t)$ uniformly in time and contaminating these with independent and identically distributed additive Gaussian

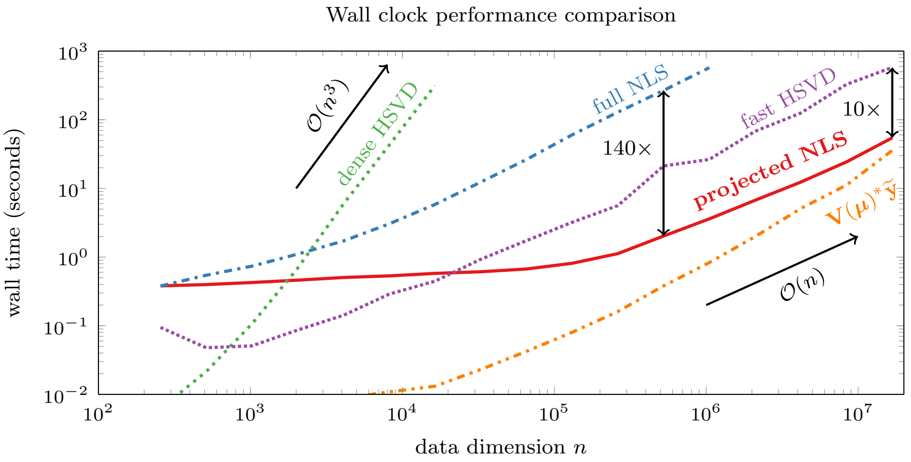

FIG. 3. *The median wall clock time from from ten runs for four different exponential fitting algorithms implemented in MATLAB* 2016b*, applied to data from* (46)*, and running on a* 2013 *Mac Pro with a* 3.5 *GHz* 6*-Core Intel Xeon E*5 *and* 16 *GB of RAM clocked at* 1866 *MHz. We also show the time taken to form* $\mathbf{V}(\boldsymbol{\mu})^*\widetilde{\mathbf{y}}$ *for* $\boldsymbol{\mu} \in \mathbb{C}^{44}$*, which is a lower bound on the time taken by our projected nonlinear least squares algorithm; this is approximately the time required to check the first order necessary conditions.*

noise $\mathbf{g}$ with $\operatorname{Cov}\mathbf{g} = \mathbf{I}$ according to the formula

$$[\widetilde{\mathbf{y}}]_k = y(\delta(n)k) + 15[\mathbf{g}]_k, \qquad \delta(n) := \frac{256}{3n} \cdot 10^{-3}. \tag{47}$$

This allows us to scale the original problem which took $n = 256$ by increasing the sample rate δ. In this section we consider three algorithms applied to this exponential fitting problem: conventional nonlinear least squares, our projected nonlinear least squares, and HSVD [1] as a representative of subspace based methods due to its simple implementation. We present our results using two different implementations of HSVD: an implementation using dense linear algebra, and fast implementation using an $\mathcal{O}(n \log n)$ Hankel matrix-vector product and an iterative SVD algorithm following [22]. Our goal is to compare these algorithms on two metrics: the wall clock time taken to solve the exponential fitting problem and the precision of the resulting parameter estimates. A MATLAB implementation of our projected nonlinear least squares algorithm for exponential fitting, the two HSVD implementations described, and code to construct these examples are provided at https://github.com/jeffrey-hokanson/ExpFit. In these implementations we use tight convergence tolerances: 10^{-16} for both residual norm and solution change in the MATLAB nonlinear least squares solver `lsqnonlin`, and 10^{-16} for the Ritz residual in the MATLAB routine `eigs` used in the fast HSVD implementation.

7.1. Timing. As these three algorithms use different paradigms for solving the exponential fitting problem, we compare their performance using total wall clock time. Figure 3 shows the time taken by each algorithm when applied to the magnetic resonance spectroscopy test problem given in (46). Asymptotically, the time taken by the dense HSVD implementation scales like $\mathcal{O}(n^3)$ due to the dense SVD, whereas the fast HSVD implementation scales like $\mathcal{O}(n \log n)$ due to the use of the fast Fourier transform (FFT) to compute the Hankel matrix-vector product. The cost of both nonlinear least squares approaches also scales like $\mathcal{O}(n \log n)$ due to their use of the

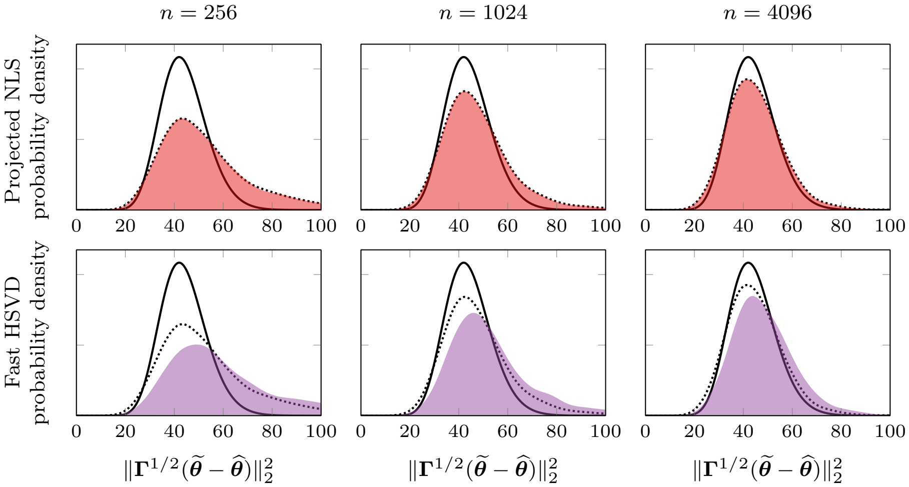

Fig. 4. *The density of standardized error in the parameter estimate* $\widetilde{\boldsymbol{\theta}}$, $\mathbf{\Gamma}^{1/2}(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}})$ *as described in subsection* 7.2. *Thick curves show the expected distribution of the* 2*-norm of the standardized error, and the filled regions show the empirically determined density from* 4000 *realizations of each method for each* n. *The dotted lines shows the density of the standardized error in the full nonlinear least squares parameter estimate.*

FFT in the initialization heuristic. However, although these last three algorithms each have the same asymptotic rate, their constants are different. In the limit of large data, the projected nonlinear least squares implementation is fastest, but for small data, the repeated initialization of the optimization algorithm dominates the cost. It is possible that a more careful implementation could avoid this cost and bring the wall clock time for projected nonlinear least squares closer to, or perhaps faster than, the fast HSVD implementation in the limit of small data.

7.2. Precision. In addition to providing faster performance than fast HSVD for large data, the projected nonlinear least squares approach also yields more precise parameter estimates. Considering the same magnetic resonance spectroscopy example, we seek to quantify the precision of our parameter estimates. In the limit of small noise, the error in the parameter estimate $\widetilde{\boldsymbol{\theta}} = [\widetilde{\boldsymbol{\omega}}, \widetilde{\mathbf{a}}]$ relative to the true parameters $\widehat{\boldsymbol{\theta}} = [\widehat{\boldsymbol{\omega}}, \widehat{\mathbf{a}}]$ is normally distributed with zero mean and covariance:

$$\mathbf{\Gamma} := \mathbf{J}(\widehat{\boldsymbol{\theta}})^* \mathbb{E}_{\mathbf{g}}[\mathbf{g}\mathbf{g}^*] \mathbf{J}(\widehat{\boldsymbol{\theta}}) = \mathbf{J}(\widehat{\boldsymbol{\theta}})^* \mathbf{J}(\widehat{\boldsymbol{\theta}}) \cdot \mathbb{E}_{\mathbf{g}}[\|\mathbf{g}\|_2^2] \approx \operatorname{Cov} \widetilde{\boldsymbol{\theta}}. \tag{48}$$

If $\mathbf{\Gamma}$ is actually the covariance of $\widetilde{\boldsymbol{\theta}}$, then $\mathbf{\Gamma}^{-1/2}(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}})$ is normally distributed with zero mean and unit variance. Hence, the norm of the mismatch $\|\mathbf{\Gamma}^{-1/2}(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}})\|_2^2$ follows a χ^2 distribution with 44 degrees of freedom. As seen in Figure 4, the distribution of error of the projected nonlinear least squares problem approximately matches that of the full problem and approaches the desired χ^2 distribution as n becomes large. However, HSVD provides less precise parameter estimates, a result that follows from the analysis of Rao [29].

8. Discussion. In this paper we have shown that by solving a sequence of projected nonlinear least squares problems we can substantially improve the run time performance with a negligible loss of accuracy for solving the exponential fitting problem when compared to both conventional nonlinear least squares and HSVD, a typical subspace based approach. For the exponential fitting problem, there are still several open questions. Is there a better choice for selecting interpolation points than our box partition? Are there better subspaces, such as one that includes not only $\mathbf{V}(\boldsymbol{\mu})$, but $\mathbf{V}'(\boldsymbol{\mu})$ as well? More generally: Can we provide conditions that guarantee the convergence of the series of projected problems? Can we bound the error incurred by the projection in a deterministic sense? Finally, we ask, What were the key features that allowed the projected approach to work for the exponential fitting problem and could this approach be applied to other problems? One key feature was that projected model $\mathbf{W}_\ell^*\mathbf{f}(\boldsymbol{\theta})$ and Jacobian $\mathbf{W}_\ell^*\mathbf{J}(\boldsymbol{\theta})$ could be computed inexpensively. The other key feature was that we were able to generate subspaces $\mathcal{W}_\ell$ such that the subspace angles between $\mathcal{W}_\ell$ and the range of the Jacobian $\mathcal{J}(\boldsymbol{\theta})$ remained small. These two requirements limit the applicability of these results to specific pairs of models $\mathbf{f}(\boldsymbol{\theta})$ and subspaces $\mathcal{W}_\ell$, but for those problems that satisfy these requirements the projected nonlinear least squares approach presents a way to improve performance.

Appendix A. Projected least squares error bounds.

Here we provide two lemmas used in section 2 related to the accuracy of a projected least squares problem.

LEMMA A.1. *Let $\mathbf{A} \in \mathbb{C}^{n\times q}$ have full column rank, and let $\mathbf{b} \in \mathbb{C}^n$ with respective range $\mathcal{A}$ and span $\mathcal{B}$. Let $\mathcal{W}$ be an m-dimensional subspace of $\mathbb{C}^n$ where $m \ge q$, and let $\mathbf{P}_\mathcal{W}$ be the orthogonal projector onto $\mathcal{W}$ where $\mathbf{P}_\mathcal{W}\mathbf{A}$ has full column rank. If $\mathbf{x}$ is the minimizer of $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2$ and $\mathbf{y}$ is the minimizer of $\|\mathbf{P}_\mathcal{W}(\mathbf{A}\mathbf{y}-\mathbf{b})\|_2$, then*

$$\|\mathbf{x}-\mathbf{y}\|_2 \le \|\mathbf{A}^+\|_2\|\mathbf{b}\|_2\big[\sin\phi_q(\mathcal{W},\mathcal{A})\sin\phi_1(\mathcal{W},\mathcal{B})+\tan^2\phi_q(\mathcal{W},\mathcal{A})\cos\phi_1(\mathcal{W},\mathcal{B})\big]. \tag{49}$$

Proof. Using the pseudoinverse, we write $\mathbf{x}$ and $\mathbf{y}$ as

$$\mathbf{x} = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{b}, \tag{50}$$

$$\mathbf{y} = (\mathbf{A}^*\mathbf{P}_\mathcal{W}\mathbf{A})^{-1}\mathbf{A}^*\mathbf{P}_\mathcal{W}\mathbf{b}. \tag{51}$$

Inserting the decomposition of the identity $\mathbf{I} = \mathbf{P}_\mathcal{W} + \mathbf{P}_\mathcal{W}^\perp$ before $\mathbf{b}$ in $\mathbf{x}$,

$$\mathbf{x} = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{P}_\mathcal{W}\mathbf{b} + (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{P}_\mathcal{W}^\perp\mathbf{b}, \tag{52}$$

we then note the difference between $\mathbf{x}$ and $\mathbf{y}$ is

$$\mathbf{x}-\mathbf{y} = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{P}_\mathcal{W}^\perp\mathbf{b} + \left[(\mathbf{A}^*\mathbf{A})^{-1} - (\mathbf{A}^*\mathbf{P}_\mathcal{W}\mathbf{A})^{-1}\right]\mathbf{A}^*\mathbf{P}_\mathcal{W}\mathbf{b}. \tag{53}$$

Replacing $\mathbf{A}$ with its short-form SVD, $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^*$, and $\mathbf{P}_\mathcal{W}$ with $\mathbf{W}\mathbf{W}^*$, where $\mathbf{W}$ is an orthonormal basis for $\mathcal{W}$, we have

$$\mathbf{x}-\mathbf{y} = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^*\mathbf{P}_\mathcal{W}^\perp\mathbf{b} - \mathbf{V}\boldsymbol{\Sigma}^{-1}\left[(\mathbf{U}^*\mathbf{W}\mathbf{W}^*\mathbf{U})^{-1} - \mathbf{I}\right]\boldsymbol{\Sigma}^{-1}\mathbf{U}^*\mathbf{W}\mathbf{W}^*\mathbf{b},$$
$$\|\mathbf{x}-\mathbf{y}\|_2 \le \|\boldsymbol{\Sigma}^{-1}\|_2\left(\|\mathbf{P}_\mathcal{W}^\perp\mathbf{U}\|_2\|\mathbf{P}_\mathcal{W}^\perp\mathbf{b}\|_2 + \|(\mathbf{U}^*\mathbf{W}\mathbf{W}^*\mathbf{U})^{-1} - \mathbf{I}\|_2\|\mathbf{U}^*\mathbf{W}\|_2\|\mathbf{W}^*\mathbf{b}\|_2\right).$$

Then invoking the subspace angle identities, $\|\mathbf{P}_\mathcal{W}^\perp\mathbf{b}\|_2 = \sin\phi_1(\mathcal{W},\mathcal{B})\|\mathbf{b}\|_2$, $\|\mathbf{W}^*\mathbf{b}\|_2 = \cos\phi_1(\mathcal{W},\mathcal{B})\|\mathbf{b}\|_2$, $\|\mathbf{U}^*\mathbf{W}\|_2 = \cos\phi_1(\mathcal{W},\mathcal{A})$, and $\|\mathbf{P}_\mathcal{W}^\perp\mathbf{U}\|_2 = \sin\phi_q(\mathcal{A},\mathcal{W})$,

$$\begin{aligned}\|\mathbf{x}-\mathbf{y}\|_2 \le \|\boldsymbol{\Sigma}^{-1}\|_2\|\mathbf{b}\|_2\big(&\sin\phi_q(\mathcal{W},\mathcal{A})\sin\phi_1(\mathcal{W},\mathcal{B})\\ &+\|(\mathbf{U}^*\mathbf{W}\mathbf{W}^*\mathbf{U})^{-1} - \mathbf{I}\|_2\cos\phi_1(\mathcal{W},\mathcal{A})\cos\phi_1(\mathcal{W},\mathcal{B})\big).\end{aligned} \tag{54}$$

To bound $\|(\mathbf{U}^*\mathbf{W}\mathbf{W}^*\mathbf{U})^{-1} - \mathbf{I}\|_2$, we note that as $\mathbf{U}^*\mathbf{W}\mathbf{W}^*\mathbf{U}$ is positive semidefinite, there exists an $\alpha \ge 0$ such that $\mathbf{U}^*\mathbf{W}\mathbf{W}^*\mathbf{U} \succeq \alpha^2\mathbf{I}$. This implies

$$\lambda_k(\mathbf{U}^*\mathbf{W}\mathbf{W}^*\mathbf{U}) - \alpha^2 \ge 0 \quad \Rightarrow \quad \sigma_k(\mathbf{W}^*\mathbf{U})^2 - \alpha^2 \ge 0 \quad \forall k \in 1,\dots,q, \tag{55}$$

where $\lambda_k(\mathbf{X})$ is the kth eigenvalue in descending order of $\mathbf{X}$. The largest α satisfying this inequality is $\alpha = \sigma_q(\mathbf{W}^*\mathbf{U}) = \cos\phi_q(\mathcal{W},\mathcal{A})$. Invoking [14, Cor. 7.7.4], $\mathbf{U}^*\mathbf{W}\mathbf{W}^*\mathbf{U} \succeq \alpha^2\mathbf{I}$ implies $(\mathbf{U}^*\mathbf{W}\mathbf{W}^*\mathbf{U})^{-1} \preceq \alpha^{-2}\mathbf{I}$ and hence

$$(\mathbf{U}^*\mathbf{W}\mathbf{W}^*\mathbf{U})^{-1} - \mathbf{I} \preceq \alpha^{-2}\mathbf{I} - \mathbf{I} = (\alpha^{-2}-1)\mathbf{I}. \tag{56}$$

Upon taking the norm, we have

$$\|(\mathbf{U}^*\mathbf{W}\mathbf{W}^*\mathbf{U})^{-1} - \mathbf{I}\|_2 \le (\alpha^{-2}-1). \tag{57}$$

Thus $\alpha^{-2}-1 = \sec^2\phi_q(\mathcal{W},\mathcal{A})$, and invoking trigonometric identities, $\sec^2\phi_q(\mathcal{W},\mathcal{A}) - 1 = \tan^2\phi_q(\mathcal{W},\mathcal{A})$; hence

$$\begin{aligned}\|\mathbf{x}-\mathbf{y}\|_2 \le \|\mathbf{\Sigma}^{-1}\|_2\|\mathbf{b}\|_2\big(&\sin\phi_q(\mathcal{W},\mathcal{A})\sin\phi_1(\mathcal{W},\mathcal{B})\\ &+ \tan^2\phi_q(\mathcal{W},\mathcal{A})\cos\phi_1(\mathcal{W},\mathcal{A})\cos\phi_1(\mathcal{W},\mathcal{B})\big).\end{aligned} \tag{58}$$

By applying the upper bound $\cos\phi_1(\mathcal{W},\mathcal{A}) \le 1$ and noting $\|\mathbf{\Sigma}^{-1}\|_2 = \|\mathbf{A}^+\|_2$, we obtain the desired bound. □

Lemma A.2. *In the same setting as Lemma* A.1*,*

$$\|\mathbf{A}^*\mathbf{A}\mathbf{y} - \mathbf{A}^*\mathbf{b}\|_2 \le \frac{\cos\phi_1(\mathcal{A},\mathcal{W})\sin\phi_q(\mathcal{A},\mathcal{W})}{\cos^2\phi_q(\mathcal{A},\mathcal{W})}\|\mathbf{A}\|_2\|\mathbf{P}_{\mathcal{A}}^\perp\mathbf{b}\|_2. \tag{59}$$

Proof. Using the pseudoinverse,

$$\mathbf{A}^*\mathbf{A}\mathbf{y} - \mathbf{A}^*\mathbf{b} = \mathbf{A}^*\mathbf{A}(\mathbf{A}^*\mathbf{P}_{\mathcal{W}}\mathbf{A})^{-1}\mathbf{A}^*\mathbf{P}_{\mathcal{W}}\mathbf{b} - \mathbf{A}^*\mathbf{b}. \tag{60}$$

Then, inserting the decomposition of the identity $\mathbf{I} = \mathbf{P}_{\mathcal{A}} + \mathbf{P}_{\mathcal{A}}^\perp$ between $\mathbf{P}_{\mathcal{W}}$ and $\mathbf{b}$,

$$\mathbf{A}^*\mathbf{A}\mathbf{y} - \mathbf{b} = \mathbf{A}^*\mathbf{A}(\mathbf{A}^*\mathbf{P}_{\mathcal{W}}\mathbf{A})^{-1}\mathbf{A}^*\mathbf{P}_{\mathcal{W}}(\mathbf{P}_{\mathcal{A}} + \mathbf{P}_{\mathcal{A}}^\perp)\mathbf{b} - \mathbf{A}^*\mathbf{b}. \tag{61}$$

After expanding the first term on the right, the $\mathbf{P}_{\mathcal{A}}$ component is $\mathbf{A}^*\mathbf{b}$, i.e.,

$$\mathbf{A}^*\mathbf{A}(\mathbf{A}^*\mathbf{P}_{\mathcal{W}}\mathbf{A})^{-1}\mathbf{A}^*\mathbf{P}_{\mathcal{W}}\mathbf{P}_{\mathcal{A}}\mathbf{b} = \mathbf{A}^*\mathbf{A}(\mathbf{A}^*\mathbf{P}_{\mathcal{W}}\mathbf{A})^{-1}\mathbf{A}^*\mathbf{P}_{\mathcal{W}}\mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{b} = \mathbf{A}^*\mathbf{b}, \tag{62}$$

and hence cancels $\mathbf{A}^*\mathbf{b}$ leaving one term:

$$\mathbf{A}^*\mathbf{A}\mathbf{y} - \mathbf{A}^*\mathbf{b} = \mathbf{A}^*\mathbf{A}(\mathbf{A}^*\mathbf{P}_{\mathcal{W}}\mathbf{A})^{-1}\mathbf{A}^*\mathbf{P}_{\mathcal{W}}\mathbf{P}_{\mathcal{A}}^\perp\mathbf{b}. \tag{63}$$

Next, we define the oblique projector above $\mathbf{X} := \mathbf{A}(\mathbf{A}^*\mathbf{P}_{\mathcal{W}}\mathbf{A})^{-1}\mathbf{A}^*\mathbf{P}_{\mathcal{W}}$ in terms of the SVD of $\mathbf{A}$. If $\mathbf{A}$ has a full and reduced SVD

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^* = \begin{bmatrix}\mathbf{U}_1 & \mathbf{U}_2\end{bmatrix}\begin{bmatrix}\mathbf{\Sigma}_1\\ \mathbf{0}\end{bmatrix}\mathbf{V}^* = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}^*, \tag{64}$$

then this oblique projector is

$$\begin{aligned}\mathbf{X} &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*(\mathbf{V}^*\mathbf{\Sigma}^*\mathbf{U}\mathbf{W}\mathbf{W}^*\mathbf{U}\mathbf{\Sigma}\mathbf{V}^*)^{-1}\mathbf{V}\mathbf{\Sigma}^*\mathbf{U}^*\mathbf{W}\mathbf{W}^*\\ &= \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}^*\mathbf{V}(\mathbf{\Sigma}_1\mathbf{U}_1^*\mathbf{W}\mathbf{W}^*\mathbf{U}_1\mathbf{\Sigma}_1)^{-1}\mathbf{V}^*\mathbf{V}\mathbf{\Sigma}_1^*\mathbf{U}_1^*\mathbf{W}\mathbf{W}^*\\ &= \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{\Sigma}_1^{-1}(\mathbf{U}_1^*\mathbf{W}\mathbf{W}^*\mathbf{U}_1)^{-1}\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_1^*\mathbf{U}_1^*\mathbf{W}\mathbf{W}^*\\ &= \mathbf{U}_1(\mathbf{U}_1^*\mathbf{W}\mathbf{W}^*\mathbf{U}_1)^{-1}\mathbf{U}_1^*\mathbf{W}\mathbf{W}^*.\end{aligned}$$

Inserting this result into the expression for $\mathbf{A}^*\mathbf{A}\mathbf{y} - \mathbf{A}^*\mathbf{b}$, we obtain the bound

$$\begin{aligned}\|\mathbf{A}^*\mathbf{A}\mathbf{y} - \mathbf{A}^*\mathbf{b}\|_2 &= \|\mathbf{A}^*\mathbf{U}_1(\mathbf{U}_1^*\mathbf{W}\mathbf{W}^*\mathbf{U}_1)^{-1}\mathbf{U}_1^*\mathbf{W}\mathbf{W}^*\mathbf{U}_2\mathbf{U}_2^*\mathbf{b}\|_2 \\ &\le \|\mathbf{A}\|_2\|(\mathbf{U}_1^*\mathbf{W}\mathbf{W}^*\mathbf{U}_1)^{-1}\|_2\|\mathbf{U}_1^*\mathbf{W}\|_2\|\mathbf{W}^*\mathbf{U}_2\|_2\|\mathbf{U}_2^*\mathbf{b}\|_2 \\ &= \sigma_q(\mathbf{U}_1^*\mathbf{W})^{-2}\sigma_1(\mathbf{U}_1^*\mathbf{W})\sigma_1(\mathbf{W}^*\mathbf{U}_2)\|\mathbf{A}\|_2\|\mathbf{P}_{\mathcal{A}}^{\perp}\mathbf{b}\|_2 \\ &= \frac{\cos\phi_1(\mathcal{A},\mathcal{W})\sin\phi_q(\mathcal{A},\mathcal{W})}{\cos^2\phi_q(\mathcal{A},\mathcal{W})}\|\mathbf{A}\|_2\|\mathbf{P}_{\mathcal{A}}^{\perp}\mathbf{b}\|_2. \qquad \square\end{aligned}$$

Appendix B. Generalized geometric sum formula. A critical component for our algorithm is the ability to compute in closed form the *generalized geometric sum*,

$$\sum_{k=n_1}^{n_2-1} k^p e^{\delta k}, \tag{65}$$

where $\delta \in \mathbb{C}$, and p, n_1, n_2 are nonnegative integers. The standard geometric sum formula provides a closed form expression when $p = 0$, and when $e^\delta = 1$, this sum can be written in terms of Bernoulli polynomials [25, eq. (24.4.9)]. The following lemma establishes the remaining case when $e^\delta \neq 1$ and $p > 0$.

LEMMA B.1. *Let $\delta \in \mathbb{C}$ with $e^\delta \neq 1$, $p, n_1, n_2 \in \mathbb{Z}_+$, where $0 \le n_1 \le n_2$; then*

$$\sum_{k=n_1}^{n_2-1} k^p e^{\delta k} = \sum_{\ell=0}^{p} \frac{\chi_{n_1}(p,\ell)e^{\delta(n_1+\ell)} - \chi_{n_2}(p,\ell)e^{\delta(n_2+\ell)}}{(1-e^\delta)^{\ell+1}}, \tag{66}$$

where $\chi_n(p,\ell)$ is given by the recurrence

$$\chi_n(p+1,\ell) = (n+\ell)\chi_n(p,\ell) + k\,\chi_n(p,\ell-1); \qquad \chi_n(0,\ell) = \delta_{\ell,0}, \quad p,\ell \ge 0. \tag{67}$$

Proof. Multiplying each term of the geometric sum by k^p corresponds to a pth derivative with respect to δ of each entry. Since this is a finite sum, we pull the derivative outside the sum, yielding

$$\sum_{k=n_1}^{n_2-1} k^p e^{\delta k} = \sum_{k=n_1}^{n_2-1} \frac{\partial^p}{\partial \delta^p} e^{\delta k} = \frac{\partial^p}{\partial \delta^p} \sum_{k=n_1}^{n_2-1} e^{\delta k} = \frac{\partial^p}{\partial \delta^p} \frac{e^{\delta n_1} - e^{\delta n_2}}{1 - e^\delta}. \tag{68}$$

To obtain an explicit formula for the derivative on the right, we show by induction

$$\frac{\partial^p}{\partial \delta^p} \frac{e^{n\delta}}{1-e^\delta} = \sum_{\ell=0}^{p} \chi_n(p,\ell)\frac{e^{(n+\ell)\delta}}{(1-e^\delta)^{\ell+1}}. \tag{69}$$

The base case $p = 0$ holds as $\chi_n(0,0) = 1$. The inductive step follows by taking the derivative of each side:

$$\begin{aligned}\frac{\partial}{\partial \delta}\sum_{\ell=0}^{p} \chi_n(p,\ell)\frac{e^{(n+\ell)\delta}}{(1-e^\delta)^{\ell+1}} &= \sum_{\ell=0}^{p+1}\left[\chi_n(p,\ell)(n+\ell) + \chi_n(p,\ell-1)\ell\right]\frac{e^{(n+\ell)\delta}}{(1-e^\delta)^{\ell+1}} \\ &= \sum_{\ell=0}^{p+1} \chi_n(p+1,\ell)\frac{e^{(n+\ell)\delta}}{(1-e^\delta)^{\ell+1}}.\end{aligned}$$

Subtracting (69) evaluated at $n = n_2$ from (69) evaluated at $n = n_1$ yields (66). $\square$

With this lemma, we now state the generalized geometric sum formula.

THEOREM B.2 (generalized geometric sum formula). *Given $\delta \in \mathbb{C}$, $p \in \mathbb{Z}_+$, and $n_1, n_2 \in \mathbb{Z}$ with $0 \leq n_1 < n_2$ integers, then*

$$\sum_{k=n_1}^{n_2-1} k^p e^{\delta k} = \begin{cases} \displaystyle\sum_{\ell=0}^{p} \frac{\chi_{n_1}(p,\ell)e^{(n_1+\ell)\delta} - \chi_{n_2}(p,\ell)e^{(n_2+\ell)\delta}}{(1-e^\delta)^{\ell+1}}, & e^\delta \neq 1; \\ \displaystyle\frac{B_{p+1}(n_2) - B_{p+1}(n_1)}{p+1}, & e^\delta = 1; \end{cases} \tag{70}$$

where B_p is the pth Bernoulli polynomial.

Acknowledgments. I would like to thank Mark Embree, Caleb Magruder, and Paul Constantine for their help in preparing this manuscript; Christopher Beattie, the Einstein Foundation of Berlin, Volker Mehrmann, and TU Berlin for hosting me during the writing of this manuscript; my Ph.D. committee: Steve Cox, Thanos Antoulas, and Matthias Heinkenschloss; and the anonymous reviewers for their thoughtful comments during the review of this manuscript.

REFERENCES

[1] H. Barkhuijsen, R. de Beer, and D. van Ormondt, *Improved algorithm for noniterative time-domain model fitting to exponentially damped magnetic resonance signals*, J. Magn. Reson., 73 (1987), pp. 553–557, https://doi.org/10.1016/0022-2364(87)90023-0.

[2] D. P. Bertsekas, *Incremental least squares methods and the extended Kalman filter*, SIAM J. Optim., 6 (1996), pp. 807–822, https://doi.org/10.1137/S1052623494268522.

[3] D. P. Bertsekas, *A new class of incremental gradient methods for least squares problems*, SIAM J. Optim., 7 (1997), pp. 913–926, https://doi.org/10.1137/S1052623495287022.

[4] R. S. Dembo, S. C. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408, https://doi.org/10.1137/0719025.

[5] D. J. Ewins, *Modal Testing: Theory and Practice*, Research Studies Press, Letchworth, Hertfordshire, England, 1984.

[6] V. V. Fedorov, *Theory of Optimal Experiments*, Academic Press, New York, 1972.

[7] R. A. Fisher, *On the mathematical foundations of theoretical statistics*, Philos. Trans. R. Soc. Lond. Ser. A, 222 (1922), pp. 309–368.

[8] M. P. Friedlander and M. Schmidt, *Hybrid deterministic-stochastic methods for data fitting*, SIAM J. Sci. Comput., 34 (2012), pp. A1380–A1405, https://doi.org/10.1137/110830629.

[9] G. Golub and V. Pereyra, *Separable nonlinear least squares: The variable projection method and its applications*, Inverse Problems, 19 (2003), pp. R1–R26, https://doi.org/10.1088/0266-5611/19/2/201.

[10] G. H. Golub and V. Pereyra, *The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate*, SIAM J. Numer. Anal., 10 (1973), pp. 413–432, https://doi.org/10.1137/0710036.

[11] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed., Johns Hopkins University Press, Baltimore, 2013.

[12] P. C. Hansen, V. Pereyra, and G. Scherer, *Least Squares Data Fitting with Applications*, Johns Hopkins University Press, Baltimore, 2013.

[13] B. L. Ho and R. E. Kalman, *Effective construction of linear state-variable models from input/output functions*, Regelungstechnik, 14 (1966), pp. 545–592, https://doi.org/10.1524/auto.1966.14.112.545.

[14] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[15] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[16] Y. Hua and T. K. Sarkar, *Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise*, IEEE Trans. Acoust. Speech Signal Process., 38 (1990), pp. 814–824, https://doi.org/10.1109/29.56027.

[17] A. A. Istratov and O. F. Vyvenko, *Exponential analysis in physical phenomena*, Rev. Sci. Instrum., 70 (1999), pp. 1233–1257, https://doi.org/10.1063/1.1149581.

[18] S. K. Jain and S. N. Singh, *Harmonics estimation in emerging power system: Key issues and challenges*, Electr. Power Syst. Res., 81 (2011), pp. 1754–1766, https://doi.org/10.1016/j.epsr.2011.05.004.

[19] F. Johansson et al., *mpmath: A Python library for arbitrary-precision floating-point arithmetic (version* 0.19*)*, 2014, http://mpmath.org.

[20] D. W. Kammler and R. J. McGlinn, *A bibliography for approximation with exponential sums*, J. Comput. Appl. Math., 4 (1978), pp. 167–173, https://doi.org/10.1016/0771-050X(78)90042-6.

[21] C. T. Kelley, *Iterative Methods for Optimization*, SIAM, Philadelphia, 1999, https://doi.org/10.1137/1.9781611970920.

[22] T. Laudadio, N. Mastronardi, L. Vanhamme, P. Van Hecke, and S. Van Huffel, *Improved Lanczos algorithms for blackbox MRS data quantitation*, J. Magn. Reson., 157 (2002), pp. 292–297, https://doi.org/10.1006/jmre.2002.2593.

[23] M. D. Macleod, *Fast nearly ML estimation of the parameters of real or complex single tones or resolved multiple tones*, IEEE Trans. Signal Process., 46 (1998), pp. 141–148, https://doi.org/10.1109/78.651200.

[24] V. B. Melas, *Functional Approach to Optimal Experimental Design*, Springer, New York, 2006.

[25] National Institute of Standards and Technology, *Digital Library of Mathematical Functions*, version 1.0.16, http://dlmf.nist.gov/, 2017.

[26] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, 2006.

[27] V. Pereyra and G. Scherer, *Exponential Data Fitting and Its Applications*, Bentham Science Publishers, Emirate of Sharjah, United Arab Emirates, 2010.

[28] R. Prony, *Essai expérimental et analytique sur les lois de la dilatabilité et sur celles de la force expansive de la vapeur de l'eau et de la vapeur de l'alkool, à différentes températures*, J. de l'Ecole Polytechnique, 1 (1795), pp. 24–76; English translation in [37, App. A].

[29] B. D. Rao, *Perturbation analysis of an SVD-based linear prediction method for estimating the frequencies of multiple sinusoids*, IEEE Trans. Acoust. Speech Signal Process., 36 (1988), pp. 1026–1035, https://doi.org/10.1109/29.1626.

[30] P. J. Schreier and L. L. Scharf, *Statistical Signal Processing of Complex-Valued Data: Theory of Improper and Noncircular Signals*, Cambridge University Press, Cambridge, UK, 2010.

[31] H. D. Scolnik, *On the Solution of Nonlinear Least Squares Problems*, Ph.D. thesis, University of Zurich, Zurich, 1970.

[32] G. A. F. Seber and C. J. Wild, *Nonlinear Regression*, John Wiley & Sons, New York, 1989.

[33] S. D. Silvey, *Optimal Design*, Chapman & Hall, London, 1980.

[34] P. Stoica and R. Moses, *Introduction to Spectral Analysis*, Prentice Hall, Upper Saddle River, NJ, 1997.

[35] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, *Compressed sensing off the grid*, IEEE Trans. Inform. Theory, 59 (2013), pp. 7465–7490, https://doi.org/10.1109/TIT.2013.2277451.

[36] D. Vandevoorde, *A Fast Exponential Decomposition Algorithm and Its Applications to Structured Matrices*, Ph.D. thesis, Rensselaer Polytechnic Institute, Troy, NY, 1996.

[37] L. Vanhamme, T. Sudin, P. Van Hecke, and S. Van Huffel, *MR spectroscopy quantitation: A review of time-domain methods*, NMR Biomed., 14 (2001), pp. 233–246, https://doi.org/10.1002/nbm.695.

[38] L. Vanhamme, A. van den Boogaart, and S. Van Huffel, *Fast and accurate parameter estimation of noisy complex exponentials with use of prior knowledge*, in Proceedings EUSIPCO-96, IEEE, Piscataway, NJ, 1996, pp. 1–4.

[39] L. Vanhamme, A. van den Boogaart, and S. Van Huffel, *Improved method for accurate and efficient quantification of MRS data with use of prior knowledge*, J. Magn. Reson., 129 (1997), pp. 35–43, https://doi.org/10.1006/jmre.1997.1244.

[40] S. Van Huffel, H. Chen, C. Decanniere, and P. Van Hecke, *Algorithm for time-domain NMR data fitting based on total least squares*, J. Magn. Reson. Ser. A, 110 (1994), pp. 228–237, https://doi.org/10.1006/jmra.1994.1209.

[41] S. J. Wright and J. N. Holt, *An inexact Levenberg-Marquardt method for large sparse nonlinear least squares*, J. Austral. Math. Soc. Ser. B, 26 (1985), pp. 387–403, https://doi.org/10.1017/S0334270000004604.