SIAM J. SCI. COMPUT.
Vol. 42, No. 5, pp. A3447–A3461

A DATA-DRIVEN MCMILLAN DEGREE LOWER BOUND*

JEFFREY M. HOKANSON†

Abstract. In the context of linear time-invariant systems, the McMillan degree prescribes the smallest possible dimension of a system that reproduces the observed dynamics. When these observations take the form of impulse response measurements where the system evolves without input from an unknown initial condition, a result of Ho and Kalman reveals the McMillan degree as the rank of a Hankel matrix built from these measurements. Unfortunately, using this result in experimental practice is challenging as measurements are invariably contaminated by noise and hence the Hankel matrix will almost surely be full rank. Hence practitioners estimate the rank of this matrix—and thus the McMillan degree—by manually setting a threshold separating large singular values corresponding to the nonzero singular values of the noise-free Hankel matrix and small singular values corresponding to perturbation of zero singular values of the noise-free Hankel matrix. Here we replace this manual threshold with a threshold guided by Weyl's theorem. Specifically, assuming measurements are perturbed by additive Gaussian noise we construct a probabilistic upper bound on how much the singular values of the noise-free Hankel matrix can be perturbed; this provides a conservative threshold for estimating the rank and hence the McMillan degree. This result follows from a new probabilistic bound on the 2-norm of a random Hankel matrix with normally distributed entries. Unlike existing results for random Hankel matrices, this bound features no unknown constants and, moreover, is within a small factor of the empirically observed bound. This bound on the McMillan degree provides an inexpensive alternative to more general model order selection techniques such as the Akaike information criteria.

Key words. McMillan degree, Hankel matrix, model order selection, random matrix

AMS subject classifications. 15B52, 60B20, 62B10, 70J10, 93E12

DOI. 10.1137/18M1194481

1. Introduction.

Here we consider discrete-time, linear time-invariant dynamical systems that map an input $\mathbf{u} \in \ell_\infty(\mathbb{N})$ to an output $\mathbf{y} \in \ell_\infty(\mathbb{N})$. Such systems are uniquely defined via their impulse response $\mathbf{h} \in \ell_1(\mathbb{N})$ through a discrete convolution

$$\mathbf{y} = \mathbf{h} * \mathbf{u}, \quad \text{where} \quad [\mathbf{h} * \mathbf{u}]_k \coloneqq \sum_{j=0}^{k} h_j u_{k-j}. \tag{1.1}$$

In *system identification* [11], the goal is to recover the system described by $\mathbf{h}$ through observations of pairs of inputs $\mathbf{u}$ and outputs $\mathbf{y}$. Rather than recovering $\mathbf{h}$ explicitly, typically one recovers a *state-space model* instead. State-space models take the form

$$\left\{ \begin{aligned} \mathbf{x}_j &= \mathbf{A}\mathbf{x}_{j-1} + \mathbf{b}u_j, \quad \mathbf{x}_{-1} = \mathbf{0} \\ y_j &= \mathbf{c}^* \mathbf{x}_j \end{aligned} \right\}, \quad \text{where} \quad \mathbf{x}_j, \mathbf{b}, \mathbf{c} \in \mathbb{C}^q, \ \mathbf{A} \in \mathbb{C}^{q \times q}. \tag{1.2}$$

An important hyperparameter for many system identification algorithms is the dimension of the state-space q in (1.2). However, this dimension is not unique. For

*Submitted to the journal's Methods and Algorithms for Scientific Computing section August 20, 2018; accepted for publication (in revised form) July 27, 2020; published electronically October 27, 2020.

https://doi.org/10.1137/18M1194481

Funding: This work was supported by the NSF VIGRE grants DMS-0240058 and DMS-0739420 and by the DARPA program Enabling Quantification of Uncertainty in Physical Systems (EQUiPS).

†Department of Computer Science, University of Colorado Boulder, Boulder, CO 80309 USA (Jeffrey.Hokanson@colorado.edu, http://www.hokanson.us).

example, the impulse response of the system in (1.2) is $\{\mathbf{c}^*\mathbf{A}^k\mathbf{b}\}_{k=0}^{\infty}$; an enlarged system with $\mathbf{A} \to \left[\begin{smallmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \star\end{smallmatrix}\right]$, $\mathbf{b} \to \left[\begin{smallmatrix}\mathbf{b}\\ \mathbf{0}\end{smallmatrix}\right]$, and $\mathbf{c} \to \left[\begin{smallmatrix}\mathbf{c}\\ \mathbf{0}\end{smallmatrix}\right]$ has the same impulse response. As identifying smaller systems requires fewer parameters and less computation, we ask: what is the smallest possible state-space system whose impulse response is $\mathbf{h}$? This is the *McMillan degree* [25, Rem. 6.7.4],

$$\mathcal{M}(\mathbf{h}) \coloneqq \min_{q\in\mathbb{N}} q \;\text{ s.t. }\; \exists\, \mathbf{b},\mathbf{c}\in\mathbb{C}^q,\; \mathbf{A}\in\mathbb{C}^{q\times q} \text{ with } h_k = \mathbf{c}^*\mathbf{A}^k\mathbf{b} \;\; \forall k\in\mathbb{N}, \tag{1.3}$$

named in honor of McMillan's pioneering work on this subject [18, 19]. Remarkably, the McMillan degree can be computed without explicitly recovering a *minimal realization* with matrices $\mathbf{A}$, $\mathbf{b}$, and $\mathbf{c}$ appearing in the optimizer of (1.3).

THEOREM 1.1 (Ho and Kalman [12, Thm. 2, Cor.]). *Let $\mathbf{H}_\infty$ denote the infinite Hankel matrix built from $\mathbf{h}\in\ell_1(\mathbb{N})$, then*

$$\mathcal{M}(\mathbf{h}) = \operatorname{rank}(\mathbf{H}_\infty) \coloneqq \sup_{n\in\mathbb{N}} \operatorname{rank}(\mathbf{H}_{n,n}), \quad \textit{where} \quad \mathbf{H}_\infty \coloneqq \begin{bmatrix} h_0 & h_1 & h_2 & \cdots \\ h_1 & h_2 & h_3 & \cdots \\ h_2 & h_3 & h_4 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{1.4}$$

and $\mathbf{H}_{m,n}\in\mathbb{C}^{m\times n}$ denotes the $m\times n$ leading principal submatrix of $\mathbf{H}_\infty$.

At first glance, it would appear straightforward to apply Ho and Kalman's result to estimate the McMillan degree in experimental practice. By providing an impulse response input, $\mathbf{u} = \{1,0,0,0,\ldots\}$, we can directly measure the impulse response $\mathbf{h}$ in the output $\mathbf{y}$. However, this poses two challenges. One is that we necessarily only measure finite data and hence cannot build the infinite-dimensional matrix $\mathbf{H}_\infty$. Instead we can only construct a lower bound on the McMillan degree from the rank of $\mathbf{H}_{m,n}$. A more substantive challenge is that experimental measurements are invariably contaminated with noise. So rather than measuring h_k, we can only measure a noisy version $\widetilde{h}_k$. If we build the analogous Hankel matrix $\widetilde{\mathbf{H}}_{m,n}\in\mathbb{C}^{m\times n}$ from $\widetilde{h}_k$,

$$\mathbf{H}_{m,n} \coloneqq \begin{bmatrix} h_0 & h_1 & \cdots & h_{n-1} \\ h_1 & h_2 & \cdots & h_n \\ \vdots & & & \vdots \\ h_{m-1} & h_m & \cdots & h_{m+n-2} \end{bmatrix}, \quad \widetilde{\mathbf{H}}_{m,n} \coloneqq \begin{bmatrix} \widetilde{h}_0 & \widetilde{h}_1 & \cdots & \widetilde{h}_{n-1} \\ \widetilde{h}_1 & \widetilde{h}_2 & \cdots & \widetilde{h}_n \\ \vdots & & & \vdots \\ \widetilde{h}_{m-1} & \widetilde{h}_m & \cdots & \widetilde{h}_{m+n-2} \end{bmatrix}, \tag{1.5}$$

then even if $\mathbf{H}_{m,n}$ is low rank, $\widetilde{\mathbf{H}}_{m,n}$ may be, and likely is, full rank. Thus we cannot naively apply Ho and Kalman's theorem to compute the McMillan degree.

1.1. Lower bound. Weyl's theorem provides a way to use Ho and Kalman's theorem to obtain a lower bound for the McMillan degree. Recall that the rank of any matrix $\mathbf{A}\in\mathbb{C}^{m\times n}$ is the number of nonzero singular values; denoting the kth singular value of $\mathbf{A}$ as $\sigma_k(\mathbf{A})$

$$\operatorname{rank}(\mathbf{A}) = \sum_{k=1}^{\min(m,n)} \mathbb{I}[\sigma_k(\mathbf{A})], \qquad \mathbb{I}[\alpha] \coloneqq \begin{cases} 0, & \alpha < 0, \\ 1, & \alpha \geq 0, \end{cases} \tag{1.6}$$

where $\mathbb{I}$ is the indicator function. Using Weyl's theorem [14, Cor. 7.3.8] provides a

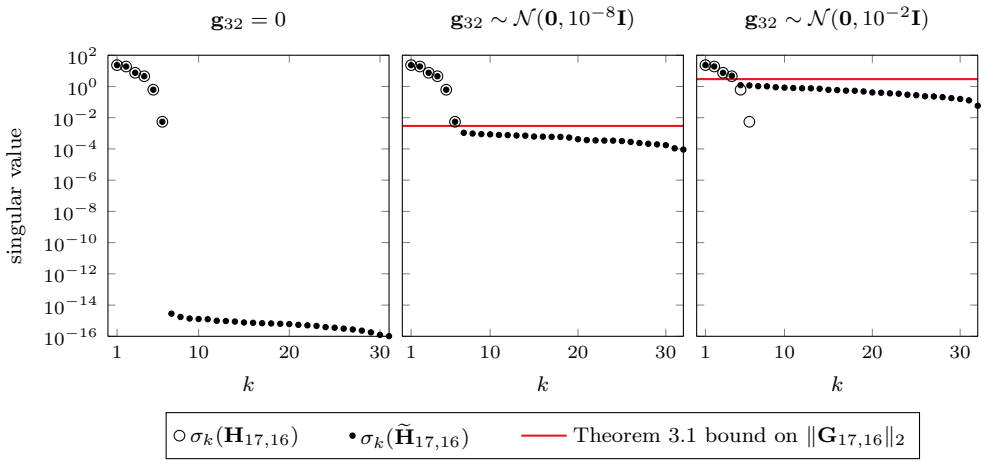

FIG. 1. *An example estimating the McMillan degree using the singular values of* $\widetilde{\mathbf{H}}_{17,16}$. *Here the true impulse response* $\mathbf{h}$ *was generated from a real state-space system with McMillan degree six with* $\mathbf{A} = \mathrm{diag}[0.9 - 0.4i, 0.9 + 0.4i, 0.9 + 0.2i, 0.9 - 0.2i, 0.7, 0.60]$, $\mathbf{b} = \mathbf{1}$, *and* $\mathbf{c} = \mathbf{1}$. *On the left, even with no noise all the singular values of* $\widetilde{\mathbf{H}}_{17,16}$ *are nonzero as a result of finite precision computation, but it is easy to identify a threshold for computing the rank of* $\widetilde{\mathbf{H}}_{17,16}$. *In the middle with a moderate amount of noise, the magnitude of the trailing singular values has increased, but visually we can still identify a threshold for identfying the McMillan degree. Note that the bound given in Theorem* 3.1 *matches visual intuition. On the right with a significant amount of noise the bound underestimates the McMillan degree as four. This is as expected as our result only provides a* lower bound *on the McMillan degree.*

bound connecting the singular values of $\mathbf{H}_{m,n}$ and $\widetilde{\mathbf{H}}_{m,n}$:

$$|\sigma_k(\widetilde{\mathbf{H}}_{m,n}) - \sigma_k(\mathbf{H}_{m,n})| \leq \|\widetilde{\mathbf{H}}_{m,n} - \mathbf{H}_{m,n}\|_2. \tag{1.7}$$

This provides a lower bound on the singular values of $\mathbf{H}_{m,n}$:

$$\sigma_k(\mathbf{H}_{m,n}) \geq \sigma_k(\widetilde{\mathbf{H}}_{m,n}) - \|\widetilde{\mathbf{H}}_{m,n} - \mathbf{H}_{m,n}\|_2. \tag{1.8}$$

Combining this result, (1.6), and Theorem 1.1 provides a lower bound on the McMillan degree

$$\mathcal{M}(\mathbf{h}) \geq \mathrm{rank}(\mathbf{H}_{m,n}) \geq \sum_{k=1}^{\min(m,n)} \mathbb{I}[\sigma_k(\widetilde{\mathbf{H}}_{m,n}) - \|\widetilde{\mathbf{H}}_{m,n} - \mathbf{H}_{m,n}\|_2]. \tag{1.9}$$

Although this bound requires computing a quantity we cannot measure—namely, the threshold $\|\widetilde{\mathbf{H}}_{m,n} - \mathbf{H}_{m,n}\|_2$—if this threshold is sufficiently small, we can visually identify an appropriate approximate threshold as illustrated in Figure 1; see, e.g., [24, subsec. 3.5]. This is necessary even with exact data in $\mathbf{H}_{m,n}$ (rounding to floating point accuracy) as computing the SVD using standard, backward stable algorithms implies we recover the singular values of a nearby $\widetilde{\mathbf{H}}_{m,n}$, not those of $\mathbf{H}_{m,n}$.

1.2. Bounding noise. In order to make the lower bound on the McMillan degree in (1.9) practical, we must estimate the threshold $\|\widetilde{\mathbf{H}}_{m,n} - \mathbf{H}_{m,n}\|_2$. Here we make the assumption that the noise in $\widetilde{h}_k$ is additive and independent of h_k so that $\widetilde{h}_k = h_k + g_k$.

Thus the threshold is the 2-norm of a structured random matrix $\mathbf{G}_{m,n}$:

$$\mathbf{G}_{m,n} \coloneqq \begin{bmatrix} g_0 & g_1 & \cdots & g_{n-1} \\ g_1 & g_2 & \cdots & g_n \\ \vdots & & & \vdots \\ g_{m-1} & g_m & \cdots & g_{m+n-2} \end{bmatrix} = \widetilde{\mathbf{H}}_{m,n} - \mathbf{H}_{m,n}. \tag{1.10}$$

In this paper we construct a probabilistic upper bound on $\|\mathbf{G}_{m,n}\|_2$ in Theorem 3.1 under the assumption that $\{g_k\}_{k=0}^{\infty}$ samples two variants of Gaussian random noise.

1.2.1. Real Gaussian. The first case considers real-valued Gaussian random noise. Here we denote the first N entries of $\{g_k\}_{k=0}^{\infty}$ as the vector $\mathbf{g}_N \in \mathbb{R}^N$ and assume $\mathbf{g}_N$ samples a real-valued multivariate normal distribution with mean zero and covariance $\mathbf{\Sigma} \in \mathbb{R}^{N\times N}$, denoted $\mathbf{g}_N \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a symmetric positive definite (SPD) matrix; $\mathbf{g}_N$ has probability density

$$p(\mathbf{g}_N) = (2\pi)^{-\frac{N}{2}} (\det \mathbf{\Sigma})^{-\frac{1}{2}} \exp[-\tfrac{1}{2}\mathbf{g}_N^{\top}\mathbf{\Sigma}^{-1}\mathbf{g}_N]. \tag{1.11}$$

1.2.2. Complex Gaussian. The second case considers complex-valued Gaussian random noise, again denoting the first N entries of $\{g_k\}_{k=0}^{\infty}$ as $\mathbf{g}_N \in \mathbb{C}^N$. Complex normal distributions require more care to define than their real counterparts. One approach is to describe $\mathbf{g}_N$ in terms of its real and imaginary parts,

$$\begin{bmatrix} \operatorname{Re}\mathbf{g}_N \\ \operatorname{Im}\mathbf{g}_N \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{12}^{\top} & \mathbf{\Sigma}_{22} \end{bmatrix} \right), \quad \text{where } \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{12}^{\top} & \mathbf{\Sigma}_{22} \end{bmatrix} \text{ is SPD.} \tag{1.12}$$

Instead we follow Schreier and Scharf [23] and characterize $\mathbf{g}_N$ via an augmented complex vector $\underline{\mathbf{g}}_N \in \mathbb{C}^{2N}$ containing $\mathbf{g}_n$ and its conjugate $\overline{\mathbf{g}}_N$ [23, sec. 2.1][1]

$$\underline{\mathbf{g}}_N \coloneqq \begin{bmatrix} \mathbf{g}_N \\ \overline{\mathbf{g}}_N \end{bmatrix} = \begin{bmatrix} \mathbf{I} & i\mathbf{I} \\ \mathbf{I} & -i\mathbf{I} \end{bmatrix} \begin{bmatrix} \operatorname{Re}\mathbf{g}_N \\ \operatorname{Im}\mathbf{g}_N \end{bmatrix}. \tag{1.13}$$

Now consider the covariance of $\underline{\mathbf{g}}_N$. As the expected value of $\underline{\mathbf{g}}_N$ is zero, the covariance matrix is simply the expected value of the outer product $\underline{\mathbf{g}}_N\underline{\mathbf{g}}_N^*$ (cf. [23, sec. 2.2])

$$\begin{aligned} \mathbb{E}(\underline{\mathbf{g}}_N\underline{\mathbf{g}}_N^*) &= \mathbb{E}\left(\begin{bmatrix} \mathbf{I} & i\mathbf{I} \\ \mathbf{I} & -i\mathbf{I} \end{bmatrix} \begin{bmatrix} \operatorname{Re}\mathbf{g}_N \\ \operatorname{Im}\mathbf{g}_N \end{bmatrix} \begin{bmatrix} \operatorname{Re}\mathbf{g}_N \\ \operatorname{Im}\mathbf{g}_N \end{bmatrix}^* \begin{bmatrix} \mathbf{I} & i\mathbf{I} \\ \mathbf{I} & -i\mathbf{I} \end{bmatrix}^* \right) \\ &= \begin{bmatrix} \mathbf{\Sigma}_{11} + \mathbf{\Sigma}_{22} - i\mathbf{\Sigma}_{12} + i\mathbf{\Sigma}_{12}^{\top} & \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{22} + i\mathbf{\Sigma}_{12} + i\mathbf{\Sigma}_{12}^{\top} \\ \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{22} - i\mathbf{\Sigma}_{12} - i\mathbf{\Sigma}_{12}^{\top} & \mathbf{\Sigma}_{11} + \mathbf{\Sigma}_{22} + i\mathbf{\Sigma}_{12} - i\mathbf{\Sigma}_{12}^{\top} \end{bmatrix} = \begin{bmatrix} \mathbf{\Gamma} & \widetilde{\mathbf{\Gamma}} \\ \widetilde{\mathbf{\Gamma}}^* & \overline{\mathbf{\Gamma}} \end{bmatrix}. \end{aligned} \tag{1.14}$$

In contrast with real normal distributions, which are completely described by their mean and covariance, describing $\mathbf{g}_N$ requires the mean, the Hermitian *covariance matrix* $\mathbf{\Gamma} \in \mathbb{C}^{N\times N}$, and the symmetric *complementary covariance matrix* $\widetilde{\mathbf{\Gamma}} \in \mathbb{C}^{N\times N}$.

There is a case where specifying a complex normal distribution simplifies and the resulting random variable acts similar to the real case. For a generic complex random variable $\mathbf{z}$ we say the following:

- $\mathbf{z}$ is *proper* if the complementary covariance $\mathbb{E}[(\mathbf{z}-\mathbb{E}[\mathbf{z}])(\mathbf{z}-\mathbb{E}[\mathbf{z}])^{\top}]$ is zero [23, Def. 2.1];

[1] We denote the conjugate of $\mathbf{g}$ by $\overline{\mathbf{g}}$ and the complex conjugate transpose of $\mathbf{g}$ by $\mathbf{g}^*$, whereas Schreier and Scharf denote the conjugate of $\mathbf{g}$ by $\mathbf{g}^*$ and the complex conjugate transpose of $\mathbf{g}$ by $\mathbf{g}^{\mathrm{H}}$.

- $\mathbf{z}$ is *circular* if the probability density of any complex rotation $e^{i\theta}\mathbf{z}$ for $\theta \in [0, 2\pi)$ is identical to that of $\mathbf{z}$ [23, Def. 2.4] (this requires $\mathbb{E}[\mathbf{z}] = \mathbf{0}$).

For a complex normally distributed random variable $\mathbf{z}$ with zero mean, $\mathbf{z}$ is proper if and only if $\mathbf{z}$ is circular [23, Res. 2.11]. Hence circular Gaussian random variables are completely specified by their covariance $\mathbf{\Gamma} \in \mathbb{C}^{N\times N}$. Here we exclusively consider circular Gaussian random variables, denoted $\mathbf{g}_N \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Gamma})$, where where $\mathbf{\Gamma} \in \mathbb{C}^{N\times N}$ is Hermitian positive definite; then $\mathbf{g}_N \in \mathbb{C}^N$ has probability density [23, Res. 2.5]

$$p(\mathbf{g}_N) = \pi^{-n} \det(\mathbf{\Gamma}_N)^{-1} \exp[-\mathbf{g}_N^* \mathbf{\Gamma}^{-1} \mathbf{g}_N]. \tag{1.15}$$

1.3. Related problems. Estimating the McMillan degree via this Hankel matrix approach is closely related to many problems in system identification and signal processing [8]. For example, given a complex sinusoidal signal

$$y(t) = \sum_{k=1}^{q} \alpha_k e^{\omega_k t} \quad \alpha_k, \omega_k \in \mathbb{C}, \tag{1.16}$$

we can compute the number of components q by considering the McMillan degree of the sequence $\{y(\delta j)\}_{j=0}^{\infty}$ for some time step $\delta > 0$ (cf. [8, subsec. 2.3]).

1.4. Contributions. Here we develop a new probabilistic upper bound on the 2-norm of a random Hankel matrix $\mathbf{G}_{m,n}$ in Theorem 3.1 based on a circulant embedding. Unlike existing results for random Hankel matrices (summarized in subsection 2.2) we are able to obtain an upper bound with a fixed probability with no unknown constants matching existing asymptotic rate results. Combined with (1.9), this upper bound on $\|\mathbf{G}_{m,n}\|_2$ allows us to obtain a lower bound on the McMillan degree of noisy measurements of the impulse response $\widetilde{\mathbf{h}}$, extending Ho and Kalman's result for noisy data. As illustrated in section 6, this bound provides a practical estimate of the McMillan degree. Replacing this probabilistic upper bound on $\|\mathbf{G}_{m,n}\|_2$ with an empirical estimate as described in section 5 provides an even sharper estimate. Finally, estimating the McMillan degree based on the singular values of a Hankel matrix compares favorably to *model selection* approaches such as the Akaike information criteria (AIC). Model selection requires identifying a minimal realization for each potential McMillan degree, a process that is both expensive and prone to identify an unrepresentative local minimum far from the global minimizer. By using our Hankel matrix approach for estimating the McMillan degree we avoid this expense and complication.

2. Background. Estimating the McMillan degree touches on four distinct domains: fast Hankel-vector products, structured random matrices, heuristics from engineering practice, and model order selection. In the following, we briefly review relevant results from each domain.

2.1. Fast Hankel matrix-vector products. Although $\widetilde{\mathbf{H}}_{m,n} \in \mathbb{C}^{m\times n}$ is a dense matrix, we can exploit the Hankel structure to provide fast matrix-vector products [22, sec. 3.4] and hence accelerate the computation of the SVD. One approach for fast Hankel vector products is to recognize a Hankel matrix can be embedded inside a circulant matrix, which in turn can be diagonalized by the discrete Fourier transform (DFT) matrix. This allows the product $\widetilde{\mathbf{H}}_{m,n}\mathbf{x}$ to be computed using only $\mathcal{O}(N \log N)$ operations where $N = m + n - 1$, rather than the $\mathcal{O}(mn)$ operations normally required. These inexpensive inner products can then accelerate the computation of the SVD when using an iterative eigensolver like ARPACK [16], with the leading k singular values being computed in approximately $\mathcal{O}(kN \log N)$ operations.

2.2. Structured random matrices. The spectral properties of structured random matrices have only started to be explored in the past two decades. The distribution of the singular values of a random Hankel matrix (and hence the 2-norm) was posed as an open problem in a 1999 paper by Bai [3]. Byrc, Dembo, and Jiang in 2006 were the first to establish the limiting spectral distribution for Hankel matrices with independent and identically distributed (i.i.d.) Gaussian entries [5]. The next year, Meckes provided bounds on the distribution of the 2-norm under weaker assumptions that entries are uniformly sub-Gaussian, independent, but not necessarily identically distributed [20]. Combining Meckes' Theorems 1 and 3 we know the growth rate of $\mathbb{E}\|\mathbf{G}_{n,n}\|_2$ as a function of n; assuming the entries of $\mathbf{G}_{n,n}$ are i.i.d. Gaussian random variables with zero mean and unit variance, then there exists $0 < c_1 < c_2$ such that

$$c_1\sqrt{n\log n} \leq \mathbb{E}\|\mathbf{G}_{n,n}\|_2 \leq c_2\sqrt{n\log n} \quad \forall n > 0. \tag{2.1}$$

Similar results were established under even weaker constraints for the distribution of the entries by Adamczak [1] and Nekrutkin [21]; the latter also treated nonsquare Hankel matrices. Note that although our results require a more restrictive assumption that entries of $\mathbf{G}_{n,n}$ sample a multivariate Gaussian distribution, we provide a different result: a computable probabilistic upper bound on $\|\mathbf{G}_{n,n}\|_2$.

2.3. Heuristics for estimating McMillan degree. Although rigorous estimates of the 2-norm of a random Hankel matrix have only been available for the past two decades, many authors in the 1970s, 1980s, and 1990s recognized that the singular values of $\widetilde{\mathbf{H}}_{m,n}$ could be used to infer the McMillan degree. For example, in 1985 Juang and Pappa suggested picking a threshold manually to separate singular values into those associated with $\mathbf{H}_{m,n}$ and those associated with noise [15, p. 622]—a process that as illustrated in Figure 1 sometimes yields an obvious choice, but that sometimes can be misleading. This manual selection approach also appears in more recent work using matrices related to $\mathbf{H}_{m,n}$; see, e.g., [17, sec. 16.3], [29], and [28]. Other authors have attempted to provide estimates of $\|\mathbf{G}_{m,n}\|_2$ to select this threshold in (1.9). For example, Holt and Antill bounded the norm of a Hankel matrix by its Frobenius norm [13, eq. (19)]. Assuming $\mathbf{g}_{2n-1} \sim \mathcal{N}(\mathbf{0}, \epsilon\mathbf{I})$,

$$\|\mathbf{G}_{n,n}\|_2 \leq \|\mathbf{G}_{n,n}\|_{\mathrm{F}} \quad \Longrightarrow \quad \mathbb{E}[\|\mathbf{G}_{n,n}\|_2] \leq \mathbb{E}[\|\mathbf{G}_{n,n}\|_{\mathrm{F}}] = \sqrt{n^2\mathbb{E}[g_0^2]} = n\epsilon. \tag{2.2}$$

However, this bound is far too conservative: from (2.1) we know $\|\mathbf{G}_{n,n}\|_2$ grows with n like $\mathcal{O}(\sqrt{n\log n})$, whereas this bound is $\mathcal{O}(n)$. Another threshold that has been suggested when $\mathbf{g}_{2n-1} \sim \mathcal{N}(\mathbf{0}, \epsilon\mathbf{I})$ is $\epsilon\sqrt{n}$; see, e.g., [9, eq. (4.3)] and [26, sec. IV.C]. This is based on the expected value of $\widetilde{\mathbf{H}}_{n,n}^*\widetilde{\mathbf{H}}_{n,n}$

$$\begin{aligned}\mathbb{E}[\widetilde{\mathbf{H}}_{n,n}^*\widetilde{\mathbf{H}}_{n,n}] &= \mathbf{H}_{n,n}^*\mathbf{H}_{n,n} + \mathbb{E}[\mathbf{G}_{n,n}^*\mathbf{H}_{n,n}] + \mathbb{E}[\mathbf{H}_{n,n}^*\mathbf{G}_{n,n}] + \mathbb{E}[\mathbf{G}_{n,n}^*\mathbf{G}_{n,n}] \\ &= \mathbf{H}_{n,n}^*\mathbf{H}_{n,n} + \epsilon^2 n\mathbf{I},\end{aligned} \tag{2.3}$$

whose eigenvalues are all shifted upward by $\epsilon^2 n$; hence the singular values of the matrix square root of $\mathbb{E}[\widetilde{\mathbf{H}}_{n,n}^*\widetilde{\mathbf{H}}_{n,n}]$ are shifted upward by $\epsilon\sqrt{n}$. However, this threshold makes a mistake interchanging expectation and the eigenvalues: the eigenvalues of $\mathbb{E}[\widetilde{\mathbf{H}}_{n,n}^*\widetilde{\mathbf{H}}_{n,n}]$ are *not* the expected eigenvalues of $\widetilde{\mathbf{H}}_{n,n}^*\widetilde{\mathbf{H}}_{n,n}$. The result is a threshold that is too permissive as it grows like $\mathcal{O}(\sqrt{n})$ whereas we should expect $\mathcal{O}(\sqrt{n\log n})$.

2.4. Model selection. Model selection provides an alternative perspective on estimating the McMillan degree using generic statistical tools for selecting the most

parsimonious model among a set of candidate models. In the context of estimating the McMillan degree, the candidate models are realizations consisting of matrices $\mathbf{A} \in \mathbb{C}^{q\times q}$ and vectors $\mathbf{b}, \mathbf{c} \in \mathbb{C}^q$ for differing dimensions q. There are a large number of different criteria for selecting the most parsimonious model (see, e.g., [6]). Here we focus on information theoretic approaches which score candidate models on both likelihood and number of parameters. The AIC [2] is one such popular model selection criteria where the score of each model is proportional to the number of free parameters minus the log-likelihood. In our context, for either real $\mathbf{g}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_n)$ or complex circular $\mathbf{g}_n \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_n)$ Gaussian random noise, the AIC score for a model of degree q is

$$\text{AIC}(q) \propto 2 \min_{\substack{\mathbf{A}\in\mathbb{C}^{q\times q}\\ \mathbf{b},\mathbf{c}\in\mathbb{C}^q}} \left\| \mathbf{\Sigma}^{-\frac{1}{2}} \left(\begin{bmatrix} \widetilde{h}_0 \\ \widetilde{h}_1 \\ \vdots \\ \widetilde{h}_{n-1} \end{bmatrix} - \begin{bmatrix} \mathbf{c}^*\mathbf{A}^0\mathbf{b} \\ \mathbf{c}^*\mathbf{A}^1\mathbf{b} \\ \vdots \\ \mathbf{c}^*\mathbf{A}^{n-1}\mathbf{b} \end{bmatrix} \right) \right\|_2^2 + 4q + \text{constant}. \tag{2.4}$$

The second term in the AIC encodes the number of real degrees of freedom in the model. Although $\mathbf{A}$, $\mathbf{b}$, and $\mathbf{c}$ have a collective $q^2 + 2q$ degrees of freedom, there are only effectively $4q$ degrees of freedom. Without loss of generality we can assume $\mathbf{c} = \mathbf{1}$ and that $\mathbf{A}$ is diagonal as nondiagonalizable matrices are nowhere dense in $\mathbb{C}^{q\times q}$ [10, p. 2739]; this leaves $2q$ complex parameters or $4q$ real parameters. Then the AIC selects the q minimizing AIC(q). The challenge with this approach is its expense: for each candidate McMillan degree a minimal realization $\{\mathbf{A}, \mathbf{b}, \mathbf{c}\}$ must be constructed.

3. Random Hankel matrix 2-norm bound. We now establish our main result: a probabilistic upper bound on the 2-norm of a random Hankel matrix whose entries are drawn from a multivariate normal distribution.

THEOREM 3.1. *Suppose $\mathbf{g}_N \in \mathbb{C}^N$ is a random variable and $\mathbf{G}_{m,n} \in \mathbb{C}^{m\times n}$ is a Hankel matrix constructed from $\mathbf{g}_N$ as in* (1.10)*, where $N = m + n - 1$, then*

$$\|\mathbf{G}_{m,n}\|_2 \le \alpha\sqrt{N} \text{ with probability } p(\alpha), \tag{3.1}$$

where $p(\alpha)$ depends on the distribution of $\mathbf{g}_N$:

$$\text{if } \mathbf{g}_N \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \text{ then} \quad p(\alpha) = \begin{cases} \text{erf}(\alpha/2)\ (1 - e^{-\alpha^2/2})^{(N-1)/2}, & N \text{ odd}, \\ \text{erf}(\alpha/2)^2 (1 - e^{-\alpha^2/2})^{N/2-1}, & N \text{ even}; \end{cases} \tag{3.2a}$$

$$\text{if } \mathbf{g}_N \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}), \text{ then} \quad p(\alpha) = (1 - e^{-\alpha^2/2})^N; \tag{3.2b}$$

$$\text{if } \mathbf{g}_N \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \text{ then} \quad p(\alpha) = \gamma(N/2, \alpha^2/(2\|\mathbf{\Sigma}^{\frac{1}{2}}\|_2^2))/\Gamma(N/2); \tag{3.2c}$$

$$\text{if } \mathbf{g}_N \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}), \text{ then} \quad p(\alpha) = \gamma(N, \alpha^2/\|\mathbf{\Sigma}^{\frac{1}{2}}\|_2^2)/\Gamma(N), \tag{3.2d}$$

where Γ denotes the Gamma function, $\Gamma(s) \coloneqq \int_0^\infty t^{s-1}e^{-t}\,\mathrm{d}t$, γ is the lower incomplete gamma function, $\gamma(s, x) \coloneqq \int_0^x t^{s-1}e^{-t}\,\mathrm{d}t$, and erf *is the error function,* $\text{erf}(x) \coloneqq 2\pi^{-1/2}\int_0^x e^{-t^2}\,\mathrm{d}t$.

We are able to state this result for any rectangular Hankel matrix $\mathbf{G}_{m,n}$ as the first component of the proof—a circulant embedding to obtain a bound in terms of the DFT of $\mathbf{g}_N$—only depends on the number variables generating the Hankel matrix. The second component then takes this bound and generates a probabilistic upper bound assuming a particular distribution for $\mathbf{g}_N$.

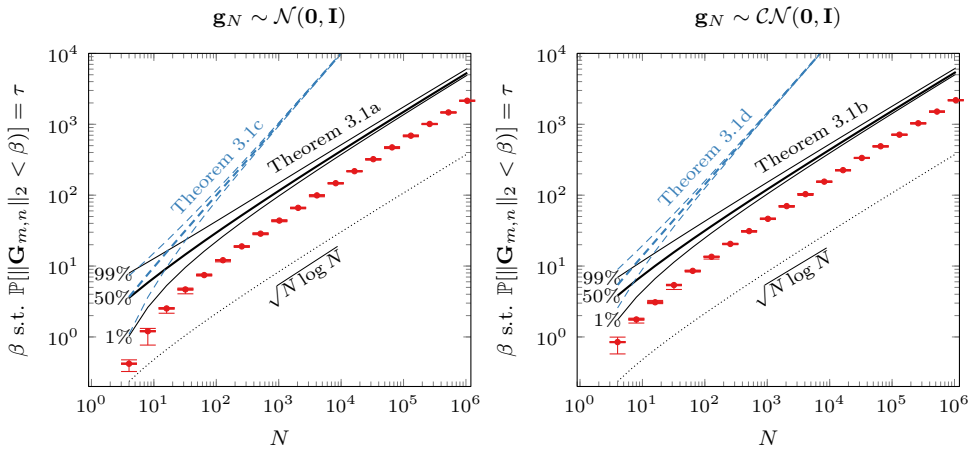

FIG. 2. *A comparison of the upper bounds from Theorem* 3.1 *for* $\|\mathbf{G}_{m,n}\|_2$ *for increasing* N*, where* $m = \lceil \frac{N-1}{2} \rceil$ *and* $n = \lfloor \frac{N-1}{2} \rfloor$*. The solid curves show the bounds from Theorem* 3.1 *which hold when the covariance matrix of* $\mathbf{g}_N$ *is the identity matrix; the dashed lines show the bound allowing other covariance matrices. The curves show the* 1*st,* 50*th, and* 99*th percentiles. The red bars show an empirical estimate of* $\|\mathbf{G}_{m,n}\|_2$ *based on* 10^3 *Monte Carlo samples, showing the bound* β *that holds with probability* τ*; the bars similarly correspond to the* 1*st,* 50*th, and* 99*th percentiles.*

3.1. Asymptotic growth. Before proving this result, we ask: does $\|\mathbf{G}_{n,n}\|_2$ grow at the same rate as $n \to \infty$ as the bound provided by Meckes [20, Thm. 3], namely, $\mathcal{O}(n \log n)$? This is true for the circular complex normal case (3.2b). For a fixed probability $\tau \in (0, 1)$, then the α satisfying $\tau = p(\alpha)$ is

$$\alpha = \sqrt{-2\log(1 - \tau^{\frac{1}{N}})} = \sqrt{2\log N - \log(\log \tau)^2 + \mathcal{O}(N^{-1})} = \mathcal{O}(\log N) \tag{3.3}$$

as $N \to \infty$. Here we used a Taylor expansion of the exponential in $\tau^{\frac{1}{N}} = \exp[\log[\tau^{\frac{1}{N}}]]$ to obtain this estimate. Hence in (3.1), $\|\mathbf{G}_{n,n}\|_2 = \mathcal{O}(\sqrt{N \log N}) = \mathcal{O}(n \log n)$ with probability τ when $\mathbf{g}_N \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$.

Figure 2 compares an empirical estimate of the distribution of $\|\mathbf{G}_{n,n}\|_2$ to the bounds provided by Theorem 3.1. We observe that both the real and circular complex normal distribution bounds in (3.2a) and (3.2b) match the expected asymptotic growth rate of $\mathcal{O}(\sqrt{n \log n})$. Moreover, for these two cases, the bound is only approximately 2.5 times larger than the empirical estimate.

3.2. Circulant embedding bound. The first step in establishing Theorem 3.1 bounds $\|\mathbf{G}_{m,n}\|_2$ by embedding $\mathbf{G}_{m,n}$ inside a circulant matrix. This circulant matrix is diagonalized by the DFT matrix allowing us to obtain its 2-norm. Although this circulant embedding technique has long been used for fast Hankel matrix-vector products [22, sec. 3.4], this is, to the best of our knowledge, the first time this embedding technique has been used to obtain bounds on the norm of a Hankel matrix.

LEMMA 3.2. *Suppose* $\mathbf{g}_N$ *and* $\mathbf{G}_{m,n}$ *are defined as in Theorem* 3.1*, then*

$$\|\mathbf{G}_{m,n}\|_2 \leq \sqrt{N} \|\mathbf{F}_N \mathbf{g}_N\|_\infty, \tag{3.4}$$

where $[\mathbf{F}_N]_{j,k} = N^{-\frac{1}{2}} e^{-2\pi i jk/N}$ *is the DFT matrix.*

Proof. Let $\mathbf{C}_N \in \mathbb{C}^{N\times N}$ be a circulant matrix [14, sec. 0.9.6] whose first column is $\mathbf{g}_N$ and recalling $N = m + n - 1$,

$$\mathbf{C}_N = \begin{bmatrix} g_0 & g_{m+n-2} & \cdots & g_{n-1} & g_{n-2} & \cdots & g_1 \\ g_1 & g_0 & \cdots & g_n & g_{n-1} & \cdots & g_2 \\ \vdots & & \ddots & \vdots & \vdots & & \vdots \\ \boxed{\begin{matrix} g_{m-1} & g_{m-2} & \cdots & g_0 \\ g_m & g_{m-1} & \cdots & g_1 \\ \vdots & & & \vdots \\ g_{m+n-2} & g_{m+n-3} & \cdots & g_{n-1} \end{matrix}} & & & & \begin{matrix} g_{m+n-2} & \cdots & g_m \\ g_0 & \cdots & g_{m+1} \\ \vdots & \ddots & \vdots \\ g_{n-2} & \cdots & g_0 \end{matrix} \end{bmatrix}. \tag{3.5}$$

Note the Hankel matrix $\mathbf{G}_{m,n} \in \mathbb{C}^{m\times n}$ appears in the boxed region of $\mathbf{C}_N$ with reversed columns. Hence the multiplication $\mathbf{G}_{m,n}\mathbf{x}_n$ can be written as

$$\mathbf{G}_{m,n}\mathbf{x}_n = \begin{bmatrix}\mathbf{0} & \mathbf{I}_m\end{bmatrix} \mathbf{C}_N \begin{bmatrix}\mathbf{J}_n \\ \mathbf{0}\end{bmatrix} \mathbf{x}_n, \tag{3.6}$$

where $\mathbf{I}_m \in \mathbb{C}^{m\times m}$ is the identity matrix and $\mathbf{J}_m \in \mathbb{C}^{n\times n}$ is the identity matrix with columns reversed. Then, as the matrix 2-norm is induced by the vector 2-norm,

$$\|\mathbf{G}_{m,n}\|_2 \coloneqq \max_{\mathbf{x}_n\in\mathbb{C}^n\setminus\{0\}} \frac{\|\mathbf{G}_{m,n}\mathbf{x}_n\|_2}{\|\mathbf{x}_n\|_2} = \max_{\mathbf{x}_n\in\mathbb{C}^n\setminus\{0\}} \frac{\left\|\begin{bmatrix}\mathbf{0} & \mathbf{I}_m\end{bmatrix} \mathbf{C}_N \begin{bmatrix}\mathbf{J}_n \\ \mathbf{0}\end{bmatrix} \mathbf{x}_n\right\|_2}{\|\mathbf{x}_n\|_2} \tag{3.7}$$

$$\leq \max_{\mathbf{x}_n\in\mathbb{C}^n\setminus\{0\}} \frac{\left\|\mathbf{C}_N \begin{bmatrix}\mathbf{J}_n \\ \mathbf{0}\end{bmatrix} \mathbf{x}_n\right\|_2}{\|\mathbf{x}_n\|_2} \leq \max_{\mathbf{y}_N\in\mathbb{C}^N\setminus\{0\}} \frac{\|\mathbf{C}_N\mathbf{y}_N\|_2}{\|\mathbf{y}_N\|_2} = \|\mathbf{C}_N\|_2. \tag{3.8}$$

Finally, to bound the norm of $\mathbf{C}_N$ we note that since $\mathbf{C}_N$ is a circulant matrix, it has spectral decomposition [22, eq. (3.27)],

$$\mathbf{C}_N = \mathbf{F}_N^*\mathbf{\Lambda}_N\mathbf{F}_N, \qquad \mathbf{\Lambda}_N = \sqrt{N}\,\mathrm{diag}(\mathbf{F}_N\mathbf{g}_N), \tag{3.9}$$

and then, as the 2-norm is unitarily invariant,

$$\|\mathbf{C}_N\|_2 = \|\mathbf{F}_N^*\mathbf{\Lambda}_N\mathbf{F}_N\|_2 = \|\mathbf{\Lambda}_N\|_2 = \sqrt{N}\|\mathbf{F}_N\mathbf{g}_N\|_\infty. \tag{3.10}$$

∎

3.3. Bounds on noise. We now seek to bound $\|\mathbf{F}_N\mathbf{g}_N\|_\infty$ for four different distributions associated with $\mathbf{g}_N$, corresponding to the four cases in (3.2). Here we denote the probability of an expression being true by $\mathbb{P}$; e.g., the probability of $z \leq \alpha$ being true for some random variable z is $\mathbb{P}[z \leq \alpha] \coloneqq \mathbb{E}[\mathbb{I}[\alpha - z]]$.

LEMMA 3.3. *Suppose* $\mathbf{g}_N \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$ *and* $\alpha \geq 0$, *then*

$$\mathbb{P}[\,\|\mathbf{F}_N\mathbf{g}_N\|_\infty \leq \alpha] = (1 - e^{-\alpha^2/2})^N. \tag{3.11}$$

Proof. We begin by characterizing $\mathbf{F}_N\mathbf{g}_N$. Note $\mathbb{E}[\mathbf{F}_N\mathbf{g}_N] = \mathbf{0}$ and hence the covariance and complementary covariance matrices are

$$\mathbb{E}[\mathbf{F}_N\mathbf{g}_N\mathbf{g}_N^*\mathbf{F}_N^*] = \mathbf{F}_N\mathbf{I}\mathbf{F}_N^* = \mathbf{I}, \qquad \mathbb{E}[\mathbf{F}_N\mathbf{g}_N\mathbf{g}_N^\top\mathbf{F}_N^\top] = \mathbf{F}_N\mathbf{0}\mathbf{F}_N^\top = \mathbf{0}, \tag{3.12}$$

where the second statement follows as $\mathbf{g}_N$ is circular. Hence $\mathbf{F}_N\mathbf{g}_N$ is a circular Gaussian random variable with $\mathbf{F}_N\mathbf{g}_N \sim \mathcal{CN}(\mathbf{0},\mathbf{I})$; (cf. [23, subsec. 2.3.1]). As such, the k entry of $\mathbf{F}_N\mathbf{g}_N$ is independent of the ℓth entry when $k \neq \ell$ and hence

$$\mathbb{P}[\,\|\mathbf{F}_N\mathbf{g}_N]\|_\infty \le \alpha] = \mathbb{P}[\max_k |\mathbf{e}_k^*\mathbf{F}_N\mathbf{g}_N| \le \alpha] = \prod_{k=0}^{N-1} \mathbb{P}[\,|\mathbf{e}_k^*\mathbf{F}_N\mathbf{g}_N| \le \alpha], \tag{3.13}$$

where $\mathbf{e}_k$ denotes the kth column of the identity. Note $\mathbf{e}_k^*\mathbf{F}_N\mathbf{g}_N$ has the distribution

$$\mathbf{e}_k^*\mathbf{F}_N\mathbf{g}_N \sim \mathcal{CN}(0,1). \tag{3.14}$$

Hence $|\mathbf{e}_k^*\mathbf{F}_N\mathbf{g}_N|$ follows a Rayleigh distribution (i.e., χ_2, a χ-distribution with two degrees of freedom) with the cumulative density function [23, eq. (2.74)]

$$\mathbb{P}[\,|\mathbf{e}_k^*\mathbf{F}_N\mathbf{g}_N| \le \alpha] = 1 - e^{-\alpha^2/2}. \tag{3.15}$$

Combining this with (3.13) provides the desired probability. ∎

The analogous result for real Gaussian random variables $\mathbf{g}_N \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ requires additional care as the entries of $\mathbf{F}_N\mathbf{g}_N$ are no longer independent—half of the entries are conjugates of the other half.

LEMMA 3.4. *Suppose* $\mathbf{g}_N \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ *and* $\alpha \ge 0$, *then*

$$\mathbb{P}[\,\|\mathbf{F}_N\mathbf{g}_N\|_\infty \le \alpha] = \begin{cases} \operatorname{erf}(\alpha/2)\ (1-e^{-\alpha^2/2})^{(N-1)/2}, & N \text{ odd}, \\ \operatorname{erf}(\alpha/2)^2(1-e^{-\alpha^2/2})^{N/2-1}, & N \text{ even}. \end{cases} \tag{3.16}$$

Proof. First, we write the real random variable $\mathbf{g}_N$ as a function of the complex circular normal variable $\mathbf{z}_N \sim \mathcal{CN}(\mathbf{0},\mathbf{I})$:

$$\mathbf{g}_N = 2^{-\frac{1}{2}}(\mathbf{z}_N + \overline{\mathbf{z}_N}) \sim \mathcal{N}(\mathbf{0},\mathbf{I}). \tag{3.17}$$

Then, defining $\mathbf{w}_N := \mathbf{F}_N\mathbf{z}_N$,

$$\mathbf{F}_N\mathbf{g}_N = 2^{-\frac{1}{2}}(\mathbf{F}_N\mathbf{z}_N + \mathbf{F}_N\overline{\mathbf{z}_N}) = 2^{-\frac{1}{2}}(\mathbf{w}_N + \mathbf{F}_N\mathbf{F}_N^\top\overline{\mathbf{w}_N}). \tag{3.18}$$

Above, the matrix $\mathbf{F}_N\mathbf{F}_N^\top$ has the form

$$\mathbf{F}_N\mathbf{F}_N^\top = \begin{bmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & \mathbf{J}_{N-1} \end{bmatrix} \in \mathbb{R}^{N\times N}, \tag{3.19}$$

where $\mathbf{J}_{N-1}$ is the reversed identity matrix. Thus, the entries of $\mathbf{F}_N\mathbf{g}_N$ are

$$\mathbf{e}_k^*\mathbf{F}_N\mathbf{g}_N = 2^{-\frac{1}{2}}\mathbf{e}_k^*\left(\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{N-1} \end{bmatrix} + \begin{bmatrix} \overline{w_0} \\ \overline{w_{N-1}} \\ \vdots \\ \overline{w_1} \end{bmatrix}\right) = \begin{cases} 2^{\frac{1}{2}}\operatorname{Re}[w_0], & k=0, \\ 2^{-\frac{1}{2}}(w_k + \overline{w_{N-k}}), & k \neq 0. \end{cases} \tag{3.20}$$

Then since $\mathbf{w}_N = \mathbf{F}_N\mathbf{z}_N \sim \mathcal{CN}(\mathbf{0},\mathbf{I})$, each entry of $\mathbf{F}_N\mathbf{g}_N$ is distributed like

$$\mathbf{e}_k^*\mathbf{F}_N\mathbf{g}_N \sim \begin{cases} \mathcal{N}(0,2), & k=0 \text{ or } k=N/2, \\ \mathcal{CN}(0,1) & \text{otherwise} \end{cases} \tag{3.21}$$

with cumulative density functions

$$\mathbb{P}[\,|\mathbf{e}_k^*\mathbf{F}_N\mathbf{g}_N| \le \alpha] = \begin{cases} \operatorname{erf}(\alpha/2), & k = 0 \text{ or } N/2, \\ 1 - e^{-\alpha^2/2} & \text{otherwise.} \end{cases} \tag{3.22}$$

Then since the first $\lfloor N/2 \rfloor$ entries of $\mathbf{F}_N\mathbf{g}_N$ are independent of each other and the remaining are fully determined by this first half (cf. (3.20)), we have

$$\mathbb{P}[\,\|\mathbf{F}_N\mathbf{g}_N\|_\infty \le \alpha] = \mathbb{P}\left[\max_{k=0,\dots,\lfloor N/2 \rfloor} |\mathbf{e}_k^*\mathbf{F}_N\mathbf{g}_N| \le \alpha\right] = \prod_{k=0}^{\lfloor N/2 \rfloor} \mathbb{P}[\,|\mathbf{e}_k^*\mathbf{F}_N\mathbf{g}_N| \le \alpha]. \tag{3.23}$$

Then using the entrywise expression (3.22) we obtain the desired bound. ∎

Unfortunately we have been unable to find satisfying bounds on $\|\mathbf{F}_N\mathbf{g}_N\|_\infty$ when $\mathbf{g}_N$ has covariance matrix that is not an identity matrix. Suppose $\mathbf{g}_N \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma})$ for some Hermitian positive definite $\mathbf{\Sigma}$; then $\mathbf{F}_N\mathbf{g}_N \sim \mathcal{CN}(\mathbf{0}, \mathbf{F}_N\mathbf{\Sigma}\mathbf{F}_N^*)$. As the entries of $\mathbf{F}_N\mathbf{g}_N$ are now correlated, we cannot separate the probabily of the max into the product of probabilities (3.13). The following two bounds provide guideance in this case, using the equivalence of finite-dimensional norms and the fact $\mathbf{F}_N$ is a unitary matrix:

$$\|\mathbf{F}_N\mathbf{g}_N\|_\infty \le \|\mathbf{F}_N\mathbf{g}_N\|_2 = \|\mathbf{g}_N\|_2. \tag{3.24}$$

Although this provides a bound, as evidenced in Figure 2, it does not achieve the expected asymptotic growth rate of $\mathcal{O}(\sqrt{N \log N})$ as $N \to \infty$.

LEMMA 3.5. *Suppose* $\mathbf{g}_N \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$*, where* $\mathbf{\Sigma} \in \mathbb{R}^{N\times N}$ *is SPD and* $\alpha \ge 0$*, then*

$$\mathbb{P}[\,\|\mathbf{F}_N\mathbf{g}_N\|_\infty \le \alpha\|\mathbf{\Sigma}^{\frac{1}{2}}\|_2\,] = 1 - \Gamma(N/2)^{-1}\gamma(N/2, \alpha^2/2). \tag{3.25}$$

Proof. Writing $\mathbf{g}_N = \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{w}_N$, where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then invoking (3.24),

$$\|\mathbf{F}_N\mathbf{g}_N\|_\infty = \|\mathbf{F}_N\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{w}_N\|_\infty \le \|\mathbf{F}_N\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{w}_N\|_2 \le \|\mathbf{\Sigma}^{\frac{1}{2}}\|_2\|\mathbf{w}_N\|_2. \tag{3.26}$$

The term $\|\mathbf{w}_N\|_2$ samples a χ-distribution with n degrees of freedom and the result follows from this density's cumulative distribution. ∎

The proof for the complex case is identical except that the χ-distribution has a total of $2N$ degrees of freedom, with half coming from the real part and half from the imaginary part.

LEMMA 3.6. *Suppose* $\mathbf{g}_N \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$*, where* $\mathbf{\Sigma} \in \mathbb{C}^{N\times N}$ *is a Hermitian positive definite matrix and* $\alpha \ge 0$*, then*

$$\mathbb{P}[\,\|\mathbf{F}_N\mathbf{g}_N\|_\infty \le \alpha\|\mathbf{\Sigma}^{\frac{1}{2}}\|_2\,] = 1 - \Gamma(N)^{-1}\gamma(N, \alpha^2/2). \tag{3.27}$$

4. McMillan degree lower bound. Having provided the bound on the norm of a random Hankel matrix $\|\mathbf{G}_{m,n}\|_2$ in Theorem 3.1, we now formally state the McMillan degree lower bound based on this result.

THEOREM 4.1. *Suppose* $\mathbf{h} \in \ell_1(\mathbb{N})$ *is the impulse response of a system. Given noisy measurements* $\widetilde{h}_k = h_k + g_k$ *constructed into a Hankel matrix* $\widetilde{\mathbf{H}}_{m,n} \in \mathbb{C}^{m\times n}$*,*

the McMillan degree of $\mathbf{h}$ *is bounded below by*

$$\mathcal{M}(\mathbf{h}) \geq \sum_{k=1}^{\min(m,n)} \mathbb{I}[\,\sigma_k(\widetilde{\mathbf{H}}_{m,n}) - \alpha\sqrt{m+n-1}\,] \quad \textit{with probablity} \quad p(\alpha), \tag{4.1}$$

where $p(\alpha)$ *depends on the distribution of* $\mathbf{g}_{m+n-1}$ *as given in* (3.2).

Proof. From (1.9) and (1.10)

$$\mathcal{M}(\mathbf{h}) \geq \sum_{k=1}^{\min(m,n)} \mathbb{I}[\,\sigma_k(\widetilde{\mathbf{H}}_{m,n}) - \|\mathbf{G}_{m,n}\|_2]. \tag{4.2}$$

From Theorem 3.1 we obtain a probabilistic upper bound on $\|\mathbf{G}_{m,n}\|_2$, which in turn provides a probabilistic lower bound on $\mathcal{M}(\mathbf{h})$. ∎

5. Empirical bound. Before continuing to the numerical experiments, we note that we need not necessarily rely on an exact, probabilistic upper bound of $\|\mathbf{G}_{m,n}\|_2$. Instead, as it is inexpensive to compute the 2-norm of a Hankel matrix (see subsection 2.1), we can instead sample many realizations of the noise to estimate the cumulative density function associated with the 2-norm of this Hankel matrix. The advantage of this approach is it provides sharper estimates of $\|\mathbf{G}_{m,n}\|_2$ and is applicable to a wider variety of distributions of g_k. However, because this is an empirical estimate, we cannot provide the guarantees as in Theorem 4.1.

6. Numerical examples. Here we provide two examples of our McMillan degree lower bound: one with complex valued data with a system known McMillan degree and another with real data with a highly reducible system. In these examples we compute the AIC score using HSVD [4] to estimate the (approximate) optimal model parameters of each candidate McMillan degree. Following the principles of reproducible research, code for constructing these examples is available at https://github.com/jeffrey-hokanson/McMillanDegree.

6.1. Complex valued data. This test problem from magnetic resonance spectroscopy [27, Tab. 1] considers the sum of eleven complex exponentials.

$$h_k = \sum_{k=1}^{11} a_k e^{135i\pi/180} e^{(2i\pi f_k - d_k)j\delta}, \tag{6.1}$$

where $\delta = \frac{1}{3} \times 10^{-3}$ and parameters

$$\begin{array}{rlcccccccccccc} & \mathbf{a} & = [& 75 & 150 & 75 & 150 & 150 & 150 & 150 & 150 & 1400 & 60 & 500 &] \\ (6.2) & \mathbf{f} & = [& -86 & -70 & -54 & 152 & 168 & 292 & 308 & 360 & 440 & 490 & 530 &] \\ & \mathbf{d} & = [& 50 & 50 & 50 & 50 & 50 & 50 & 50 & 25 & 285.7 & 25 & 200 &]. \end{array}$$

To simulate detector noise, we add complex circular Gaussian random noise $\mathbf{g} \sim \mathcal{CN}(\mathbf{0}, 15^2\,\mathbf{I})$. In this example we use a total of $N = 256$ measurements.

Figure 3 illustrates how different approaches estimate the number of complex exponentials that are present in $\widetilde{\mathbf{h}}_N$; recall from subsection 1.3 that determining the number of exponentials is equivalent to determining the McMillan degree. As expected, the McMillan degree lower bound in Theorem 4.1 provides a lower bound on McMillan degree. By using an empirical estimate of $\|\mathbf{G}_{m,n}\|_2$ as described in 5 we obtain a sharper and frequently correct estimate of the McMillan degree. This suggests

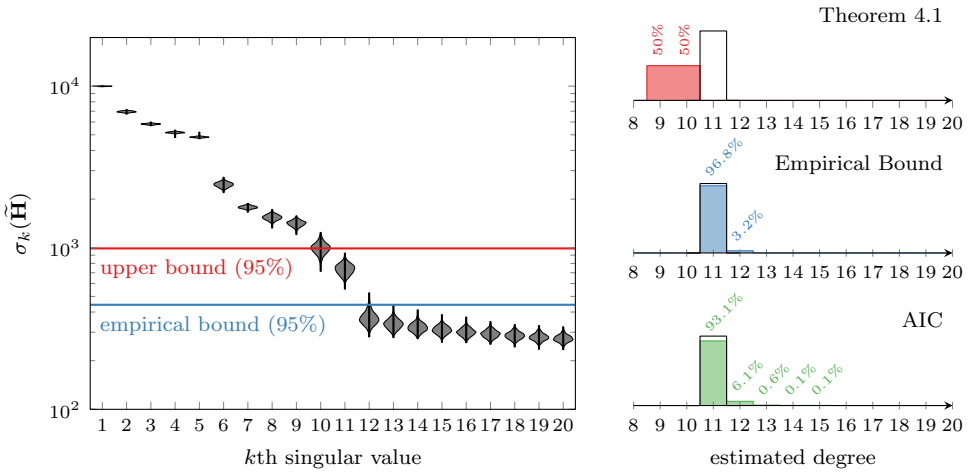

FIG. 3. *The application of our bounds and the AIC to estimate the number of complex exponentials embedded in complex Gaussian noise as described in subsection* 6.1. *The left plot shows the distribution of the first twenty singular values of* $\widetilde{\mathbf{H}}_{129,128}$ *constructed from* 1000 *realizations, where the frequency is denoted by the width of the shaded region and the range is denoted by the vertical black bar. The three right plots show the estimated model order using different techniques and the true model order of eleven is denoted by the hollow black rectangle.*

that most of the loss of accuracy in our bound occurs mainly in the embedding step (Lemma 3.2), not in the use Weyl's theorem for the singular values Theorem 4.1. The AIC performs well in this case, but requires more computation as each potential degree we must find a minimal realization.

6.2. Real valued data. As a second example, we consider the clamped beam model from the SLICOT benchmarks for model reduction [7]. This considers a beam where the input is a force applied at the free boundary and the output is the displacement at this boundary. Although originally a continuous time model, we can convert this to a discrete-time system in the form of (1.2) using the matrix exponential

$$\begin{aligned} \mathbf{x}_j &= e^{\mathbf{A}\delta}\mathbf{x}_{j-1} + \mathbf{b}u_j, \quad \text{where} \quad \mathbf{b}, \mathbf{c}, \mathbf{x}_j \in \mathbb{R}^{348}, \ \mathbf{A} \in \mathbb{R}^{348\times 348}, \\ y_j &= \mathbf{c}^\top \mathbf{x}_j. \end{aligned} \tag{6.3}$$

Here we take the time step $\delta = 0.1$ and use $N = 2^{13} = 8192$ measurements to which we add real Gaussian noise with $\mathbf{g}_N \sim \mathcal{N}(\mathbf{0}, 10^{-2}\mathbf{I})$. Although this example has a McMillan degree of 348, corresponding to the dimension of $\mathbf{A}$, it is highly reducible and the singular values of $\mathbf{H}_{m,n}$ decay rapidly. This simulates real systems which may have components that cannot be resolved due to noise.

Figure 4 illustrates different approaches for estimating the McMillan degree of this system. Unlike the previous example, we have no hope of estimating the true McMillan degree of $\mathbf{A}$, as even in the absence of noise only 105 singular values of $\mathbf{H}_{4097,4096}$ exceed 10^{-10}. With the addition of noise we obtain a McMillan degree lower bound of 8 using Theorem 4.1 and 12 using the empirical estimate. Both of these are smaller than than the McMillan degree estimate provided by the AIC.

7. Conclusion. Here we have established an upper bound on the norm of a random Hankel matrix with no unknown constants in Theorem 3.1 and used this result to construct a lower bound on the McMillan degree from noisy impulse response

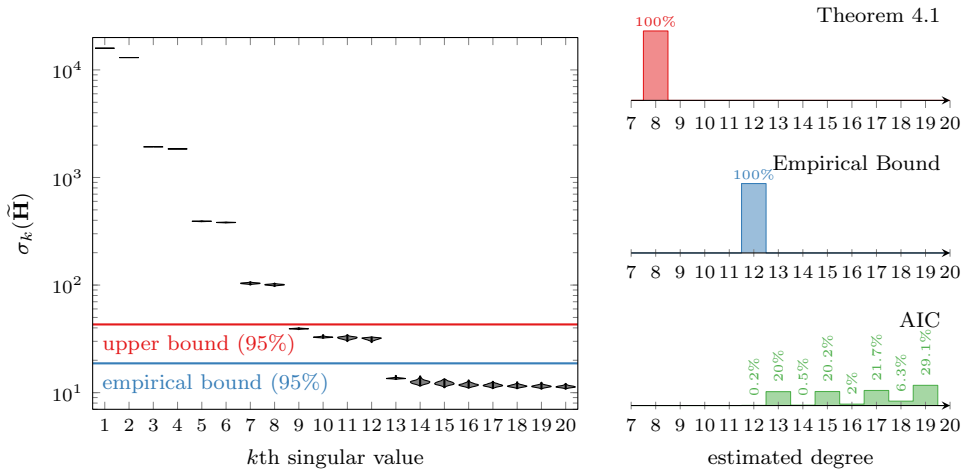

FIG. 4. *The application of our bounds and the AIC to estimate the McMillan degree of the beam model described in subsection* 6.2. *The left plot shows the distribution of singular values of* $\widetilde{\mathbf{H}}_{4097,4096}$ *over one thousand realizations as in Figure* 3. *The three plots on the right show the estimated model order using different techniques.*

measurements in Theorem 4.1. As the examples in section 6 illustrate, this bound provides a useful lower bound on the McMillan degree that can be applied to both modal analysis and system identification. However, in engineering practice, we expect the empirically determined bound on $\|\mathbf{G}_{m,n}\|_2$ to be more useful. It provides a sharper bound and is easy to compute without knowledge of the underlying noise distribution by using measurements of the system with no input.

Acknowledgments. I would like to thank Mark Embree for his support during my PhD where this result originated, Paul Martin for his feedback on an early draft of this manuscript, and the anonymous reviewers for their help refining this manuscript.

REFERENCES

[1] R. Adamczak, *A few remarks on the operator norm of random Toeplitz matrices*, J. Theoret. Probab., 23 (2010), pp. 85–108, https://doi.org/10.1007/s10959-008-0201-7.

[2] H. Akaike, *A new look at the statistical model identification*, IEEE Trans. Automat. Control, 19 (1974), https://doi.org/10.1109/TAC.1974.1100705.

[3] Z. D. Bai, *Methodologies in the spectral analysis of large dimensional random matrices, a review*, Statist. Sinica, 9 (1999), pp. 611–677.

[4] 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.

[5] W. Bryc, A. Dembo, and T. Jiang, *Spectral measure of large random Hankel, Markov and Toeplitz matrices*, Ann. Probab., 34 (2006), pp. 1–38, https://doi.org/10.1214/009117905000000495.

[6] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference*, Springer, New York, 2002, https://doi.org/10.1007/b97636.

[7] Y. Chahlaoui and P. V. Dooren, *A Collection of Benchmark Examples for Model Reduction of Linear Time Invariant Dynamical Systems*, Technical Report 2, SLICOT, 2002.

[8] M. T. Chu and M. M. Lin, *On the finite rank and finite-dimensional representation of bounded semi-infinite hankel operators*, IMA J. Numer. Anal., 35 (2015), pp. 1256–1276, https://doi.org/10.1093/imanum/dru001.

[9] P. de Groen and B. de Moor, *The fit of a sum of exponentials to noisy data*, J. Comput. Appl. Math., 20 (1987), pp. 175–187, https://doi.org/10.1016/0377-0427(87)90135-x.
[10] P. V. Dooren, K. A. Gallivan, and P.-A. Absil, *$\mathcal{H}_2$-optimal model reduction with higher-order poles*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2738–2753, https://doi.org/10.1137/080731591.
[11] G. C. Goodwin and R. L. Payne, *Dynamic System Identification: Experiment Design and Data Analysis*, Academic Press, New York, 1977, https://doi.org/10.1016/s0076-5392(08)x6188-0.
[12] 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.
[13] J. N. Holt and R. J. Antill, *Determining the number of terms in a Prony algorithm exponential fit*, Math. Biosci., 36 (1977), pp. 319–332, https://doi.org/10.1016/0025-5564(77)90054-2.
[14] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985, https://doi.org/10.1017/CBO9780511810817.
[15] J.-N. Juang and R. S. Pappa, *An eigensystem realization algorithm for modal parameter identification and model reduction*, J. Guid. Control Dynam., 8 (1985), pp. 620–627, https://doi.org/10.2514/3.20031.
[16] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK User's Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, SIAM, Philadelphia, 1998, https://doi.org/10.1137/1.9780898719628.
[17] L. Ljung, *System Identification: Theory for the User*, 2nd ed., Prentice Hall Inform. System Sci. Ser., Prentice Hall, Upper Saddle River, NJ, 1999.
[18] B. McMillan, *Introduction to formal realizability theory* I, Bell Syst. Tech. J., 31 (1952), pp. 217–279, https://doi.org/10.1002/j.1538-7305.1952.tb01383.x.
[19] B. McMillan, *Introduction to formal realizability theory* II, Bell Syst. Tech. J., 31 (1952), pp. 541–600, https://doi.org/10.1002/j.1538-7305.1952.tb01396.x.
[20] M. W. Meckes, *On the spectral norm of a random Toeplitz matrix*, Electron. Commun. Probab., 12 (2007), pp. 315–325, https://doi.org/10.1214/ecp.v12-1313.
[21] V. V. Nekrutkin, *Remark on the norm of random Hankel matrices*, Vestnik St. Petersburg Univ. Math., 46 (2013), pp. 189–192, https://doi.org/10.3103/s106345411304002x.
[22] M. K. Ng, *Iterative Methods for Toeplitz Systems*, Oxford University Press, New York, 2004.
[23] P. J. Schreier and L. L. Scharf, *Statistical Signal Processing of Complex-Valued Data: Theory of Improper and Noncircular Signals*, Cambridge University Press, Cambridge, 2010, https://doi.org/10.1017/CBO9780511815911.
[24] B. D. Schutter, *Minimal state-space realization in linear system theory: An overview*, J. Comput. Appl. Math., 121 (2000), pp. 331–354, https://doi.org/10.1016/S0377-0427(00)00341-1.
[25] E. D. Sontag, *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, Springer, New York, 1998, https://doi.org/10.1007/978-1-4612-0577-7.
[26] A.-J. Van Der Veen, E. F. Deprettere, and A. L. Swindlehurst, *Subspace-based signal analysis using singular value decomposition*, Proc. IEEE, 81 (1993), pp. 1277–1308, https://doi.org/10.1109/5.237536.
[27] 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.
[28] S. Yang and H. Li, *Estimating the number of harmonics using enhanced matrix*, IEEE Signal Proc. Lett., 14 (2007), pp. 137–140, https://doi.org/10.1109/lsp.2006.882095.
[29] Y. Zhang, Z. Zhang, X. Xu, and H. Hua, *Modal parameter identification using response data only*, J. Sound Vib., 282 (2005), pp. 367–380, https://doi.org/10.1016/j.jsv.2004.02.012.