The Geometry of Estimation in High Dimensions
White Paper – draft not for circulation

Haim Bar, Lisa Goldberg, Alec Kercheval, Alex Shkolnik

March 23, 2024

1. Introduction

More than two decades ago, at the American Mathematical Society meeting *Mathematical Challenges of the 21st Century,* David Donoho[1] hypothesized that exploiting the "blessing of dimensionality" may be key to solving "important structural problems in data analysis". Since then, a collection of results that includes our own recent work has supported this prediction, with theoretical developments that leverage the geometry of high dimensional data in insightful ways. In this white paper, we develop elementary but novel illustrations that navigate the curses and exploit the blessings of dimensionality in a setting of statistical estimation. We are guided by geometrical intuition and supply a brief account of new and compelling applications of our ideas.

1.1. Back to the future – high dimension and low sample size. The literature on asymptotic estimators falls into distinct regimes, LH, HH and HL, which are important to delineate. The regimes are named for the relative sizes of the number of variables/dimensions p and the number of observations/samples n, respectively. The LH regime, (low dimension and high sample size) that is hundreds years old, corresponds to a fixed p with n growing to infinity. For example, by the law of large numbers (LLN), the sample covariance is a consistent estimator of the population counterpart in the LH regime. The HH regime (high dimension and high sample size), which was steered by modern problems and an outsized influence of random matrix theory beginning in the 1960s, lets both n and p tend to infinity together, most often in fixed proportion. Here, the consistency can no longer rely on the LLN and some underlying model structure (or "signal") must be incorporated. For example, the consistency of the sample covariance matrix estimator discussed above, now depends on the magnitude of the population eigenvalues (e.g., Paul (2007)). The most recent and lesser known asymptotic framework is the HL regime (high dimension and low sample size) as popularized by Hall et al. (2005). This includes the case when the number of observations n is fixed but the dimension p tends to infinity. Such a perspective on data is often neglected but has tremendous advantages, which are twofold. (1) The HL regime is highly relevant for real-world data because a small sample size is often imposed by experimental constraints, or by the lack of long-range stationarity of time series. (2) The HL regime allows for a simpler theoretical treatment that can more readily use tools like the LLN and the more general framework of the concentration of measure.

This white paper takes a "*back to the future*" approach to the asymptotic analysis of data. That is, we use our seasoned understanding of basic tools like the LLN and the concentration of spherical volume/measure in high dimensional spaces to develop estimators that are tuned for data problems of today and the future. We keep our examples to a finite sample n and make use of the high dimensional limit $p \to \infty$ to drive our results.

1.2. Research roadmap – an example, an estimator and applications. We keep to a simplified analysis in the content below, but mention the known and the potential generalizations along the way. Section 2 presents a simple yet widely used optimization problem and points to a puzzle that arises in many practical settings. To this end, the HL regime yields clarity in a way the LH and HH frameworks do not. Section 3 revisits the well-known James-Stein estimator (circa 1960s) and introduces a "Geometric LLN" as a way of understanding the "blessings of dimensionality". This estimator resolves the optimization puzzle posed in Section 2 and prompts the use of concentration of measure as a key tool in a full research program. Section 4 provides a glimpse at further work through the lens of applications.

Throughout, $\langle x, y\rangle$ denotes the standard inner product of $x, y \in \mathbb{R}^m$ and $|x| = \sqrt{\langle x, x\rangle}$.

2. A simple, motivating example on optimization theory

One animating theme concerns estimation error and its interaction with optimization programs, which are routine in modern statistical problems in engineering and science. Questions that arise in this context include the following: "*Does optimization amplify or reduce statistical estimation errors in a model? How does one leverage that information if it is known? Which components of the model should be estimated more precisely, and which can afford less accuracy?*" To motivate the study of the interplay between the optimization and model estimation, we will consider a quadratic function in p variables, as a simple but important example.

[1] Donoho (2000); see also Gorban et al. (2016), Gorban & Tyukin (2018).

For constants $a_0, a_1 \in \mathbb{R}$, a vector $\zeta \in \mathbb{R}^p$, and a symmetric positive definite (SPD) matrix Σ, let

$$Q(x) = a_0 + a_1 \langle x, \zeta \rangle - \frac{1}{2} \langle x, \Sigma x \rangle \qquad (x \in \mathbb{R}^p). \tag{1}$$

The problem of maximizing (1) is encountered in many classical contexts within statistics, probability and operations research. Examples are least-squares regression, maximum *a posteriori* estimation, trust-region methods, utility maximization problems, saddle point approximations, and Legendre-Fenchel transforms in moderate/large deviations theory. Specific applications include the Markowitz selection problem in finance, robust beamforming in signal processing, and optimal fingerprinting in climate science. Since additional linear equality constraints may be put into an unconstrained Lagrangian form (1), this setting is more general than it first appears.

The maximizer of $Q(\cdot)$ is given by $a_1 \Sigma^{-1}\zeta$, with a corresponding maximum value

$$\max_{x \in \mathbb{R}^p} Q(x) = a_0 + \frac{a_1^2 \mu_p^2}{2} \qquad \left(\mu_p^2 = \langle \zeta, \Sigma^{-1}\zeta \rangle\right), \tag{2}$$

but, in practice, the maximum is computed with an estimate $\widehat{\Sigma}$ replacing an unknown Σ. This "plug-in" step is known to yield a perplexing computational problem in practice. In essence, optimization chases the errors in $\widehat{\Sigma}$ to produce systematic bias in the computed maximum. We show this bias is further amplified by a high dimension.

Consider the high dimensional limit $p \to \infty$ and a sequence of SPD matrices $\Sigma = \Sigma_{p\times p}$, with some fixed number K of eigenvalues that grow in p and with all remaining ones bounded in $(0, \infty)$. This is a well known "spiked" covariance model where K eigenvalues of Σ grow with p to represent the strongest sources of correlation. To illustrate the impact of estimation on (2), let $\hat{x}$ be the maximizer of $\hat{Q}(\cdot)$, defined by replacing Σ with some estimator $\widehat{\Sigma}$ with the same eigenvalue properties. This leads to an estimated maximum of the true objective in (1),

$$Q(\hat{x}) = a_0 + a_1\langle \hat{x}, \zeta \rangle - \frac{1}{2}\langle \hat{x}, \Sigma \hat{x} \rangle = a_0 + \frac{a_1^2 \hat{\mu}_p^2}{2} D_p \tag{3}$$

where $\hat{\mu}_p^2 = \langle \zeta, \widehat{\Sigma}^{-1}\zeta \rangle$ and D_p is a discrepancy (relative to (2)) that can grow rapidly as the dimension increases. Unless $\widehat{\Sigma}$ is fine-tuned in a particular way (see below), D_p always tends to $-\infty$ as $p \to \infty$.[2] Assuming the natural pervasiveness condition that $|\zeta|^2 = \sum_{i=1}^p \zeta_i^2$ tends to infinity, we obtain the following puzzling behaviour.[3]

As $p \to \infty$, the estimated maximum $Q(\hat{x})$ tends to $-\infty$ while the true maximum (2) tends to $+\infty$.

By scaling the constants a_0 and a_1, one can arrive at an alternative pairs of limits, but practical scalings preserve the large disparity between the true and estimated objective values in (2) and (3). To avoid this pathological behaviour, the $\widehat{\Sigma}$ must be fine-tuned in a very precise way. In particular, consider the spectral decompositions

$$\widehat{\Sigma} = \textstyle\sum_\eta \eta\eta^\top = \sum_{(\vartheta^2, h)} \vartheta^2 hh^\top, \qquad \Sigma = \sum_\beta \beta\beta^\top = \sum_{(\lambda^2, b)} \lambda^2 bb^\top \tag{4}$$

where the sums are over the p eigenvalue/eigenvector pairs (e.g., $\Sigma b = \lambda^2 b$ and $\lambda^2 = \langle \beta, \beta \rangle$). A natural question is which of the p eigenpairs (λ^2, b) must be estimated well to preclude the undesirable behaviour above.

Remarkably, one needs to pay attention to only a few parts of the spectral decomposition of Σ.

> *To remedy the asymptotics $Q(\hat{x}) \to -\infty$, we need only attend to the estimation accuracy of the span of the K eigenvectors of Σ for which the eigenvalues λ^2 grow with p. The accuracy of all eigenvalue estimates as well as the estimates of all remaining eigenvectors do not determine (3) asymptotically!*

We see that remarkably few of the $p \times p$ parameters of $\Sigma_{p\times p}$ must be estimated well. This analysis motivates the study of more accurate estimation of eigenvectors, or vectors more generally, as we do in the next section. Lastly, we note the subtle role played by the HL regime, which allowed us to "freeze" the finite sample bias in $\widehat{\Sigma}$, and take the high dimensional limit $p \to \infty$ to reveal yet another manifestation of the curse of dimensionality.

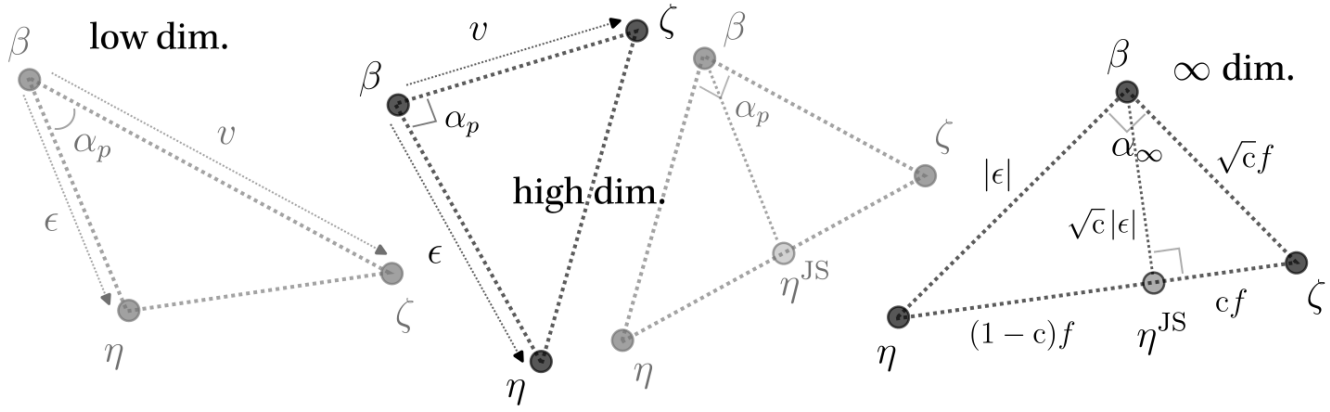
3. The Geometric LLN and the blessings of dimensionality

The "curse of dimensionality" has been an impediment in many areas of applied mathematics. Yet, in many high dimensional settings, we can leverage the statistical properties of noisy data sets to counter its effects. We now illustrate this on an elementary signal/noise separation problem, casting basic concepts such as the law of large numbers (LLN) in geometric terms. The latter yields the necessary intuition to construct the James-Stein estimator that concludes this section by showing us how to correct high dimensional vector estimates. Due to the structure of Hilbert spaces (where the inner product is generalized as $\langle H, B \rangle = \mathrm{tr}(H^\top B)$ for matrices H, B) these results generalize to correct entire subspaces. This is particular relevant for dimension reduction tools such as principal component analysis.

[2] The rate is determined exclusively by the growing eigenvalues of $\Sigma_{p\times p}$.

[3] Writing $\mu_p^2 = (\zeta, \Sigma^{-1}\zeta) = \langle z, \Sigma^{-1} z \rangle |\zeta|^2$ for $z = \zeta/|\zeta|$, we see that $\langle z, \Sigma^{-1} z \rangle$ remains bounded in $(0, \infty)$ provided z is eventually (in p) not entirely contained in the span of eigenvectors of Σ that have eigenvalues growing with the dimension.

Figure 1. ***Left panel:*** *A finite sample observation η of the unknown $\beta \in \mathbb{R}^p$ is illustrated in low dimension, where η is the only useful estimator of β.* ***Middle panel:*** *In high dimension, the Geometric* LLN *takes hold: the vector $v = \zeta - \beta$ is, with high probability, at an approximate right angle α_p to the error vector $\epsilon = \eta - \beta$. We see the* JS *estimator η^{JS} that, aided by the Geometric* LLN*, yields an improved estimator of β with high probability.* ***Right panel:*** *In the limit $p \to \infty$, α_p converges to $\alpha_\infty = 90°$ despite finite sample error. The* JS *estimator yields an improvement factor of $\sqrt{\text{c}}$ relative to $|\epsilon| = |\eta - \beta|$. Limit distances between all points (at scale $1/\sqrt{p}$) are as shown with $f = \frac{|\epsilon|}{\sqrt{1-\text{c}}}$. The* c *is random in $(0, 1)$ as it depends on the finite sample but is computable solely from data if we have at least one additional observation of β that is independent of η.*



Consider a p-vector $\eta = (\eta_1, \ldots, \eta_p) \in \mathbb{R}^p$, an estimate of an unknown parameter $\beta \in \mathbb{R}^p$. As p grows, these p-dimensional vectors may be viewed as the first p entries of an infinite sequence $\{\beta_j\}_{j \geq 1}$. Then,

$$\text{(Noise)} \qquad \epsilon = \eta - \beta$$

is a p-vector representing the residual noise term (i.e., ϵ_i is the error in ith variable). In a LH regime where the number of samples/measurements/observations n tends to infinity, we may obtain $\epsilon \to 0$, yielding a consistent estimator η for p fixed. In stark contrast, in the HL framework, the finite sample error is left intact. But noise, or pure randomness, is still associated with LLN behaviour. In particular, suppose we introduce a sequence (independent of above) of numbers $\{\zeta_j\}_{j \geq 1}$ with associated vector $\zeta \in \mathbb{R}^p$ and let α_p denote the angle between $v = \zeta - \beta$ and the vector ϵ (see Figure 1). With this geometry in mind, under a diverse set of conditions,[4] we state the following LLN.

$$\text{(Geometric LLN)} \qquad \langle \eta - \beta, \zeta - \beta \rangle / p = \langle \epsilon, v \rangle / p = \frac{1}{p} \sum_{i=1}^{p} \epsilon_i v_i \to 0 \qquad (\text{angle } \alpha_p \to 90° \text{ as } p \to \infty). \tag{5}$$

The Geometric LLN begins to take hold in higher dimensions with limit realized as the angle $\alpha_\infty = 90°$ in the right panel of Figure 1. The mode of convergence may be taken in any appropriate sense (e.g., almost surely). This represents a geometrical interpretation of an elementary LLN principle:

Noise is orthogonal to (or independent of) any high dimensional vector not corrupted by it.

The blessing of dimensionality now arrives in the form of the James-Stein (JS) estimator η^{JS} (see right panel of Figure 1), which finds an optimal convex combination of η and ζ that reduces the error $|\epsilon|$. Let,

$$\eta^{\text{JS}} = \zeta + \text{c}\,(\eta - \zeta), \qquad \text{c} = 1 - \frac{\nu^2}{|\eta - \zeta|^2} \tag{6}$$

where ν^2 is any p-asymptotically consistent estimate of $|\epsilon|^2 = |\eta - \beta|^2$, which may be obtained with just $n = 2$ weakly dependent observations.[5] Write $f_p \sim g_p$ to stand for the statement that $\lim_{p \to \infty} f_p / g_p = 1$.

Theorem 1. *If both $|\beta|^2/p$ and $|\epsilon|^2/p$ are bounded in $(0, \infty)$ in the limit as $p \to \infty$ and the Geometric* LLN *holds, then $\sqrt{\text{c}}$ is eventually in the interval $(0, 1)$ and,*

$$|\eta^{\text{JS}} - \beta| \sim \sqrt{\text{c}}\,|\eta - \beta| = \sqrt{\text{c}}\,|\epsilon|. \tag{7}$$

The mode of convergence above is inherited from the Geometric LLN and both sides in (7), including the improvement factor $\sqrt{\text{c}}$, are random due to the finite sample error that persists in the high dimensional limit. Standard presentations of the JS estimator (James & Stein (1961), Efron & Morris (1977), Brown & Zhao (2012)) in the literature assume p and n are fixed and that ϵ is Gaussian to produce bounds on $|\eta^{\text{JS}} - \beta|^2$ in expectation. Our

[4] For example, if the $\{\epsilon_i\}$ are mean-zero, uncorrelated, independent of the $\{v_i\}$ and for $\sigma_k^2 = \text{VAR}(\epsilon_k)$ we have $\sum_{k \geq 1} (\sigma_k v_k)^2 \log(k)/k^2$, then the Geometric LLN is a consequence of the Radamacher strong LLN in an almost sure sense. Numerous other statements are possible.

[5] e.g., letting H be a $p \times 2$ matrix, with first column η and the second column independent copy of η, we have $\nu^2 = \lambda_{\min}(H^\top H) \sim |\epsilon|^2$.

asymptotic version of the JS estimator relies on geometrical arguments using the high dimensional limit to drop the distributional assumptions. This theory extends to much more complicated settings, e.g., eigenvectors of random matrices as well as subspaces of such vectors that are often used as low dimensional representations. Significant progress in this direction appears in a number of recent works by the present authors, including Shkolnik (2022), Goldberg et al. (2020, 2022), Goldberg & Kercheval (2023), Goldberg et al. (2023), Gurdogan & Kercheval (2022).

There is more magic to the η^{JS} in Figure 1. To appreciate this and the use of ζ from (1) for the construction in (6), consider $\Sigma = \beta\beta^\top + \Gamma$ for a $p \times p$ sequence of matrices $\Gamma = \Gamma_{p\times p}$ with eigenvalues bounded in $(0, \infty)$ and $\Gamma\beta = 0$. This is a spiked model provided $\lambda_p^2 = \langle \beta, \beta \rangle = \sum_{i=1}^p \beta_i^2 \to \infty$ as $p \to \infty$ corresponding to $K = 1$ spike. Suppose our estimator of Σ used to estimate the objective (3) is taken as $\widehat{\Sigma} = \eta\eta^\top + \hat{\gamma}^2 \mathrm{I}$ for some sequence $\hat{\gamma} = \hat{\gamma}_p \in \mathbb{R}$ bounded away from zero. The simplicity of the estimate $\hat{\gamma}^2\mathrm{I}$ of Γ reinforces the fact that estimation of eigenvalues is of secondary importance with respect to the large p behaviour of (3). In this setting, $h = \eta/|\eta|$ is the eigenvector of $\widehat{\Sigma}$ with the largest eigenvalue $\mathfrak{s}_p^2 = \langle \eta, \eta \rangle \to \infty$. It estimates the true eigenvector $b = \beta/|\beta|$ per (4).

Now, the discrepancy D_p between the true and estimated objectives (2) and (3), may be shown to obey

$$D_p = -(\lambda_p^2/\hat{\gamma}^2)\, \mathscr{E}_p^2(h) + O(1) \qquad \Big(\text{where } \mathscr{E}_p(h) = \frac{\langle b, z\rangle - \langle h, b\rangle\langle h, z\rangle}{1 - \langle h, z\rangle^2} \text{ and } z = \frac{\zeta}{|\zeta|}\Big).$$

We call $\mathscr{E}_p(h)$ the *quadratic optimization bias* of the estimate η which depends only on its direction. The discrepancy D_p drifts to $-\infty$ unless the optimization bias $\mathscr{E}_p(h)$ tends to zero, which typically does not happen. But remarkably, $\mathscr{E}_p(h^{\text{JS}}) \to 0$ for $h^{\text{JS}} = \eta^{\text{JS}}/|\eta^{\text{JS}}|$ under the assumptions of Theorem 1 and yet, $|h^{\text{JS}} - b| \not\to 0$ as $p \to \infty$. The points h, h^{JS}, b and z are the points in Figure 1 now embedded into a unit sphere in $\mathbb{R}^p$, the surface measure of which remains finite as p grows. This facilitates the study of the concentration of measure on the sphere as a refinement of the Geometric LLN. This concentration now drives all the geometry and supplies us with even more insight, including the rate of convergence of the discrepancy D_p equipped with a JS correction.

4. A glimpse at the applications

Optimization. The most immediate application of the example described in Section 2 concerns "black-box" function optimization. Indeed, multivariate quadratic optimization forms a key step for more general (black-box) minimization techniques such as trust-region methods (e.g., Maggiar et al. (2018)). In the latter setting, the matrix Σ in (1) corresponds to a Hessian matrix and $\widehat{\Sigma}$ is its approximation. Errors in $\widehat{\Sigma}$ arise in a variety of ways. Numerical, approximation and sampling errors are all common, especially in high dimensions. Moreover, the optimization of noisy functions is a highly relevant and important area of study. In parallel, derivative-free optimization (DFO) is and an active area with many unexplored aspects (Scheinberg (2022)). Its relation to Section 3 is almost immediate since the gradient is a vector (e.g., $\beta \in \mathbb{R}^p$), that is estimated by averaging n samples (i.e., function evaluations) of a p dimensional finite difference approximation ($\eta \in \mathbb{R}^p$). The JS estimator corrects the finite sample error in gradient approximations provided the dimension is sufficiently large (and the number of function evaluations, small).

Machine learning. An intriguing application of our ideas is to neural networks, which are learned functions parametrized by collections of weight matrices (one for each layer of the network). These weight matrices are outputs of the training process and, for simplicity, we discuss the input layer of a neural network, with weight matrix $W = W_{p\times n}$ where p is the dimension and n is the number of neurons. A recent empirical study by Martin & Mahoney (2021) of many state-of-the-art neural networks reveals that the symmetric matrix $\widehat{\Sigma} = WW^\top$ has properties akin to the spiked covariance model of Section 2. In particular, $\widehat{\Sigma} = HH^\top + G$ where $HH^\top = \sum_{(\mathfrak{s}^2, h)} \mathfrak{s}^2 hh^\top$ where the sum is over K largest eigenvalues $\mathfrak{s}^2$ of $\widehat{\Sigma}$ (i.e., spikes) that separate themselves from those of G in magnitude. For example, a (stripped down) version of the popular AlexNet neural network exhibits $K = 9$ spikes, with the eigenvalues of G distributed according to the Marchenko-Pastur (MP) law (Martin & Mahoney 2021, Figure 8). The MP distribution is synonymous with noise, i.e., that which obeys a Geometric LLN and thus amenable to a JS type corrections of Section 3. The HL regime is very appropriate here, as the number of neurons n may (or even should) be taken to be much smaller than the dimension p. Our extension of the JS estimator to the K-dimensional subspace $H = H_{p\times K}$ corrects the signal component of the weight matrix W of the neural network. Correcting for the noise inherent in the training process may be hypothesized to improve the performance of the network.

Beamforming. Suppose that a signal of interest, corrupted by interfering signals and noise, is collected at a series of sensors. The "beamforming" problem is to weigh the information collected at the sensors to maximize the signal-to-noise ratio. One popular approach to solving this problem is Capon (or minimum variance) beamforming, introduced in (Capon (1969)). This and related approaches are equivalent to maximizing (1) (treated as the unconstrained Lagrangian), where Σ is taken to be the covariance matrix of signals observed at p sensors. Spike models for Σ (similar to those described in Section 2) for beamforming have been proposed in Yang et al. (2018) and elsewhere and analyzed in the HH regime. These approaches almost universally correct only the eigenvalues in the estimate $\widehat{\Sigma}$ and take the number of samples (called "snapshots") large, which ignores the non-stationarity of the observations (Cox (2002)). Indeed, our thorough pass through the highly active literature on this problem finds frequent mention of the fact that eigenvector estimates in $\widehat{\Sigma}$ carry material statistical errors, yet few approaches address this issue in any direct way. One rare example is in Quijano & Zurk (2015) where a "pruning" of sample

eigenvectors is investigated. Our HL regime analysis of the JS estimator is applicable to eigenvectors, has rigorous theoretical guarantees, and addresses directly the quadratic maximization aspect of beamforming (i.e., (3)).

Graphical models. Models that employ vertices and edges to impose graph structure, have a rich mathematical history. Coupled with the ideas of Section 2 on quadratic optimization, they form an abundant ground for a diverse set of applications and analyses. For example, taking Σ in (1) as a graph adjacency matrix and adding simple bound constraints on x in (1) leads to solutions/approximations of graph properties such as the maximum clique and maximum independent set (e.g., Hager & Hungerford (2015)). While Σ is not directly interpreted as a spiked covariance matrix in a graph theory setting, its eigenvalue properties are similar (e.g., see the celebrated Cheeger inequality). When the observation of the graph structure is noisy an estimated adjacency matrix $\widehat{\Sigma}$ must be used. In the direction of spectral methods (e.g. (4)), the leading eigenvector of a suitably normalized adjacency is the stationary distribution of a random walk on a graph. Thus, accurate estimation of this eigenvector in noisy settings is highly relevant for classical algorithms such as Markov Chain Monte Carlo and PageRank. Moreover, our recent work on graphical models flips the roles of the sample size n and the dimension p to infer the zeros in an adjacency matrix by using the distributional properties of the angle between two vectors. This approach facilitates a test for presence of edges between a large number of p vertices even when n is relatively small (Bar & Wells (2023)).

Dynamical systems. Many physical systems in the real-world are inherently low-dimensional and are governed by some state variable $x \in \mathbb{R}^n$ solving $\dot{x} = f(x, u)$ for some input $u \in \mathbb{R}^m$ and function f. An observable in such model is a $y = g(x, u)$ for some function g. Learning f and g becomes the main task given N observations of $y \in \mathbb{R}^p$, with p the problem dimension. It is well-known that, regardless of the magnitude of p, the maximal information that may be extracted is limited by the low dimension of the underlying state space $\dim(x) = n$ (e.g., (Wright & Ma 2022, Section 1.1.1)). Here, the perspective that flips the roles of the number of variables and observations, as highlighted earlier also offers benefits. Indeed, the HL regime once again becomes an appropriate framework that can leverage a large space of size $p \times N$ that has an intrinsic low rank n. Moreover, spectral methods (like the decompositions used in (4)) are often used to estimate linear dynamical systems, or linear approximations of nonlinear dynamics, making our subspace corrections that generalize the JS estimator of Section 3 pertinent.

References

Bar, H. & Wells, M. T. (2023), 'On graphical models and convex geometry', *Computational Statistics & Data Analysis* **187**, 107800.

Brown, L. D. & Zhao, L. H. (2012), 'A geometrical explanation of stein shrinkage', *Statistical Science* pp. 24–30.

Capon, J. (1969), 'High-resolution frequency-wavenumber spectrum analysis', *Proceedings of the IEEE* **57**(8), 1408–1418.

Cox, H. (2002), Adaptive beamforming in non-stationary environments, *in* 'Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers, 2002.', Vol. 1, IEEE, pp. 431–438.

Donoho, D. (2000), High-dimensional data analysis: the curses and blessings of dimensionality, Technical report, Department of Statistics, Stanford University.

Efron, B. & Morris, C. (1977), 'Stein's paradox in statistics', *Scientific American* **236**(5), 119–127.

Goldberg, L. R., Gurdogan, H. & Kercheval, A. (2023), Portfolio optimization via strategy-specific eigenvector shrinkage. CDAR working paper.

Goldberg, L. R. & Kercheval, A. N. (2023), 'James-Stein for the leading eigenvector', *Proceedings of the National Academy of Sciences* **120**(2).

Goldberg, L. R., Papacinicolau, A., Shkolnik, A. & Ulucam, S. (2020), 'Better betas', *The Journal of Portfolio Management* **47**(1), 119–136.

Goldberg, L. R., Papanicolau, A. & Shkolnik, A. (2022), 'The dispersion bias', *SIAM Journal of Financial Mathematics* **13**(2), 521–550.

Gorban, A. N. & Tyukin, I. Y. (2018), 'Blessing of dimensionality: mathematical foundations of the statistical physics of data', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Scienes* **376**(2118).

Gorban, A. N., Tyukin, I. Y. & Romanenko, I. (2016), 'The blessing of dimensionality: Separation theorems in the thermodynamic limit', *IFAC-PapersOnLine* **49**(24), 64–69.

Gurdogan, H. & Kercheval, A. (2022), 'Multi anchor point shrinkage for the sample covariance matrix', *SIAM Journal on Financial Mathematics* **13**(3), 1112–1143.

Hager, W. W. & Hungerford, J. T. (2015), 'Continuous quadratic programming formulations of optimization problems on graphs', *European Journal of Operational Research* **240**(2), 328–337.

Hall, P., Marron, J. & Neeman, A. (2005), 'Geometric representation of high dimension, low sample size data', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(3), 427–444.

James, W. & Stein, C. (1961), Estimation with quadratic loss, *in* 'Proc. Fourth Berkeley Symp. Math. Stat. Prob.', pp. 361–397.

Maggiar, A., Wachter, A., Dolinskaya, I. S. & Staum, J. (2018), 'A derivative-free trust-region algorithm for the optimization of functions smoothed via gaussian convolution using adaptive multiple importance sampling', *SIAM Journal on Optimization* **28**(2), 1478–1507.

Martin, C. & Mahoney, M. (2021), 'Implicit self-regularization in deep neural networks: Evidence from random matrix theorey and implications for learning', *Jounral of Machine Learning Research* **22**, 1–73.

Paul, D. (2007), 'Asymptotics of sample eigenstructure for a large dimensional spiked covariance model', *Statistica Sinica* **17**, 1617–1642.

Quijano, J. E. & Zurk, L. M. (2015), 'Eigenvector pruning method for high resolution beamforming', *The Journal of the Acoustical Society of America* **138**(4), 2152–2160.

Scheinberg, K. (2022), 'Finite difference gradient approximation: To randomize or not?', *INFORMS Journal on Computing* **34**(5), 2384–2388.

Shkolnik, A. (2022), 'James-Stein estimation of the first principal component', *Stat* **11**(1).

Wright, J. & Ma, Y. (2022), *High-dimensional data analysis with low-dimensional models: Principles, computation, and applications*, Cambridge University Press.

Yang, L., McKay, M. R. & Couillet, R. (2018), 'High-dimensional mvdr beamforming: Optimized solutions based on spiked random matrix models', *IEEE Transactions on Signal Processing* **66**(7), 1933–1947.