Appendices for the Paper
Subspace Fitting Meets Regression: The Effects of Supervision and Orthonormality Constraints on Double Descent of Generalization Errors

Yehuda Dar [1] **Paul Mayer** [1] **Lorenzo Luzi** [1] **Richard G. Baraniuk** [1]

Outline

These appendices support the main paper in the following ways. Appendix A provides proofs and various explanations to the statements provided in Section 3 of the main text. In particular, in Appendix Section A.5, we provide mathematical analysis and experimental justification for the claim regarding the *on average* decrease of the out-of-sample error $\mathcal{E}_{\text{out}}^{\text{unsup}}\left(\widehat{\mathbf{U}}_k\right)$ with the number of features p. In Appendix B we refer to Section 4 of the paper, prove the specific projection operators used in our projected gradient descent algorithms, and provide additional details on the experiments for the supervised settings. In Appendix C we elaborate on the semi-supervised subspace fitting method presented in Section 5 of the main text. Appendix D provides the details on the range of unsupervised problems with soft orthonormality constraints.

Note that the indexing of equations and figures in the Appendices below is prefixed with the letter of the corresponding Appendix. Other references correspond to the main paper.

A. Proofs and Explanations for Section 3

A.1. Explanation for Corollaries 3.1 and 3.2

One should note that any overparameterized subspace estimate $\widehat{\mathcal{U}}_k$ is induced by a rank-deficient sample covariance matrix $\widehat{\mathbf{C}}_{\mathbf{x},\mathcal{S}}^{(n)}$ of rank $\rho \triangleq \text{rank}\left\{\widehat{\mathbf{C}}_{\mathbf{x},\mathcal{S}}^{(n)}\right\} \leq n-1$. This is simply because $\widehat{\mathbf{C}}_{\mathbf{x},\mathcal{S}}^{(n)}$ is formed based on n centered samples of p-dimensional feature vectors where, as implied from the definition of overparameterization, $p > n$. This is also the case for rank-overparameterized subspace estimates (which are a particular type of overparameterized subspace estimates). However, the point that Corollary 3.1 emphasizes is that rank-overparameterized subspace estimates are guaranteed to be affected by the rank-deficiency of $\widehat{\mathbf{C}}_{\mathbf{x},\mathcal{S}}^{(n)}$. Accordingly, the construction provided in Corollary 3.2 shows that, due to the insufficient number of nonzero eigenvalues of $\widehat{\mathbf{C}}_{\mathbf{x},\mathcal{S}}^{(n)}$, a rank-overparameterized estimate has freedom in setting $k-\rho$ out of its k spanning orthonormal vectors.

A.2. Proof of Proposition 3.1

Since $p > n$ (due to overparameterization), the sample covariance matrix $\widehat{\mathbf{C}}_{\mathbf{x},\mathcal{S}}^{(n)}$ has size $p \times p$ and rank $\rho \triangleq \text{rank}\left\{\widehat{\mathbf{C}}_{\mathbf{x},\mathcal{S}}^{(n)}\right\} \leq n-1$. Hence, the eigendecomposition $\widehat{\mathbf{C}}_{\mathbf{x},\mathcal{S}}^{(n)} = \widehat{\mathbf{\Psi}}_{\mathcal{S}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Psi}}_{\mathcal{S}}^{*}$ corresponds to a $p \times p$ unitary matrix $\widehat{\mathbf{\Psi}}_{\mathcal{S}} \triangleq \left[\widehat{\boldsymbol{\psi}}_{\mathcal{S}}^{(1)}, \dots, \widehat{\boldsymbol{\psi}}_{\mathcal{S}}^{(p)}\right]$ and a diagonal matrix $\widehat{\mathbf{\Lambda}} \triangleq \text{diag}\left\{\widehat{\lambda}^{(1)}, \dots, \widehat{\lambda}^{(p)}\right\}$ with only ρ nonzero eigenvalues $\widehat{\lambda}^{(h_1)}, \dots, \widehat{\lambda}^{(h_\rho)}$, where $1 \leq h_1 < h_2 < \dots < h_\rho \leq p$. Therefore, the eigenvectors $\widehat{\boldsymbol{\psi}}_{\mathcal{S}}^{(h_1)}, \dots, \widehat{\boldsymbol{\psi}}_{\mathcal{S}}^{(h_\rho)}$ are those associated with the nonzero eigenvalues. Here $\mathbf{\Psi}^{*}$ denotes the conjugate transpose of the matrix $\mathbf{\Psi}$.

The subspace estimate is rank-overparameterized (recall Definition 3.2), thus, $p > n$ and $k \in \{n, \dots, p\}$. Then, $\widehat{\mathbf{U}}_{k,\mathcal{S}}$ is a $p \times k$ matrix with k orthonormal columns, where the first ρ of them satisfy $\widehat{\mathbf{u}}_{\mathcal{S}}^{(i)} = \widehat{\boldsymbol{\psi}}^{(h_i)}$ for $i = 1, ..., \rho$. The additional $k-\rho$ columns $\widehat{\mathbf{u}}_{\mathcal{S}}^{(\rho+1)}, \dots, \widehat{\mathbf{u}}_{\mathcal{S}}^{(k)}$ are chosen arbitrarily from the $p-\rho$ columns of $\widehat{\mathbf{\Psi}}_{\mathcal{S}}$ corresponding to zero eigenvalues. Namely, $\widehat{\mathbf{u}}_{\mathcal{S}}^{(\rho+i)} = \widehat{\boldsymbol{\psi}}^{(r_i)}$ for $i = 1, ..., k-\rho$ and $\{r_1, \dots, r_{k-\rho}\}$ is an arbitrary subset of $\{1, \dots, p\} \setminus \{h_1, \dots, h_\rho\}$. This construction satisfies the orthonormality demand for the k columns of $\widehat{\mathbf{U}}_{k,\mathcal{S}}$.

Here, the in-sample error of interest is (7), namely,

$$\mathcal{E}_{\text{in},\mathcal{S}}^{\text{unsup}}\left(\widehat{\mathbf{U}}_{k,\mathcal{S}}\right) = \text{Tr}\left\{\left(\mathbf{I}_p - \widehat{\mathbf{U}}_{k,\mathcal{S}}\widehat{\mathbf{U}}_{k,\mathcal{S}}^{*}\right)\widehat{\mathbf{\Psi}}_{\mathcal{S}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Psi}}_{\mathcal{S}}^{*}\left(\mathbf{I}_p - \widehat{\mathbf{U}}_{k,\mathcal{S}}\widehat{\mathbf{U}}_{k,\mathcal{S}}^{*}\right)^{*}\right\} \tag{A.1}$$

Note that, by the construction of $\widehat{\mathbf{U}}_{k,\mathcal{S}}$, the eigendecomposition of the $p \times p$ projection operator $\widehat{\mathbf{U}}_{k,\mathcal{S}}\widehat{\mathbf{U}}_{k,\mathcal{S}}^{*}$ satisfies

$$\widehat{\mathbf{U}}_{k,\mathcal{S}}\widehat{\mathbf{U}}_{k,\mathcal{S}}^{*} = \widehat{\mathbf{\Psi}}_{\mathcal{S}}\widehat{\mathbf{\Lambda}}_{\mathcal{S},\text{ind}[k]}\widehat{\mathbf{\Psi}}_{\mathcal{S}}^{*} \tag{A.2}$$

where $\widehat{\mathbf{\Psi}}_{\mathcal{S}}$ is the $p \times p$ unitary matrix that diagonalizes $\widehat{\mathbf{C}}_{\mathbf{x},\mathcal{S}}^{(n)}$, and $\widehat{\mathbf{\Lambda}}_{\mathcal{S},\text{ind}[k]}$ is a $p \times p$ diagonal matrix with ones at the coordinates

$\{(h_1,h_1),\dots,(h_\rho,h_\rho)\}\cup\{(r_1,r_1),\dots,(r_{k-\rho},r_{k-\rho})\}$ and zeros elsewhere. Therefore,

$$\begin{aligned}\mathcal{E}^{\text{unsup}}_{\text{in},\mathcal{S}}\left(\widehat{\mathbf{U}}_{k,\mathcal{S}}\right) &= \text{Tr}\left\{\widehat{\mathbf{\Psi}}_{\mathcal{S}}\left(\mathbf{I}_p-\widehat{\mathbf{\Lambda}}_{\mathcal{S},\text{ind}[k]}\right)\times\right.\\ &\qquad\left.\widehat{\mathbf{\Psi}}^*_{\mathcal{S}}\widehat{\mathbf{\Psi}}_{\mathcal{S}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Psi}}^*_{\mathcal{S}}\widehat{\mathbf{\Psi}}_{\mathcal{S}}\left(\mathbf{I}_p-\widehat{\mathbf{\Lambda}}_{\mathcal{S},\text{ind}[k]}\right)\widehat{\mathbf{\Psi}}^*_{\mathcal{S}}\right\}\\ &= \text{Tr}\left\{\left(\mathbf{I}_p-\widehat{\mathbf{\Lambda}}_{\mathcal{S},\text{ind}[k]}\right)\widehat{\mathbf{\Lambda}}\left(\mathbf{I}_p-\widehat{\mathbf{\Lambda}}_{\mathcal{S},\text{ind}[k]}\right)\right\}\\ &= \mathbf{0}\end{aligned}\tag{A.3}$$

This proves that a rank-overparameterized subspace estimate formed by the construction in Corollary 3.2 is $\mathcal{S}$-interpolating.

One can extend the last proof to the general form of rank-overparameterized subspace estimates, where the additional arbitrary $k-\rho$ orthonormal vectors can be any $(k-\rho)$-dimensional subspace of the $(p-\rho)$-dimensional null space of $\widehat{\mathbf{C}}^{(n)}_{\mathbf{x},\mathcal{S}}$.

A.3. Proof of Proposition 3.2

Let us denote the eigenvalues of the true covariance matrix, $\mathbf{C}_{\mathbf{x}}$, as $\lambda^{(1)},\dots,\lambda^{(d)}$. The covariance matrix of the p-dimensional feature vectors is denoted as $\mathbf{C}_{\mathbf{x},\mathcal{S}}$, and its eigendecomposition satisfies $\mathbf{C}_{\mathbf{x},\mathcal{S}}=\mathbf{\Psi}_{\mathcal{S}}\mathbf{\Lambda}_{\mathcal{S}}\mathbf{\Psi}^*_{\mathcal{S}}$ where $\mathbf{\Psi}_{\mathcal{S}}$ is a $p\times p$ unitary matrix, and $\mathbf{\Lambda}_{\mathcal{S}}=\text{diag}\left\{\lambda^{(1)}_{\mathcal{S}},\dots,\lambda^{(p)}_{\mathcal{S}}\right\}$ is a diagonal matrix containing the eigenvalues of $\mathbf{C}_{\mathbf{x},\mathcal{S}}$. Similar to the construction in (A.2) we have here $\widehat{\mathbf{U}}_{k,\mathcal{S}}\widehat{\mathbf{U}}^*_{k,\mathcal{S}}=\widehat{\mathbf{\Psi}}_{\mathcal{S}}\widehat{\mathbf{\Lambda}}_{\mathcal{S},\text{ind}[k]}\widehat{\mathbf{\Psi}}^*_{\mathcal{S}}$, where $\widehat{\mathbf{\Lambda}}_{\mathcal{S},\text{ind}[k]}$ is a diagonal matrix with ones at the main-diagonal coordinates corresponding to columns of $\widehat{\mathbf{\Psi}}_{\mathcal{S}}$ chosen to define $\widehat{\mathbf{U}}_{k,\mathcal{S}}$ and zeros elsewhere. Then, the expression for the unsupervised out-of-sample error is developed as follows.

$$\begin{aligned}\mathcal{E}^{\text{unsup}}_{\text{out}}\left(\widehat{\mathbf{U}}_k\right) &= \mathbb{E}\left\|\left(\mathbf{I}_d-\widehat{\mathbf{U}}_k\widehat{\mathbf{U}}^*_k\right)\mathbf{x}_{\text{test}}\right\|_2^2\\ &= \mathbb{E}\left\|\mathbf{x}_{\text{test}}\right\|_2^2-\mathbb{E}\left\|\widehat{\mathbf{U}}_k\widehat{\mathbf{U}}^*_k\mathbf{x}_{\text{test}}\right\|_2^2\\ &= \text{Tr}\left\{\mathbf{C}_{\mathbf{x}}\right\}-\mathbb{E}\left\|\widehat{\mathbf{U}}_{k,\mathcal{S}}\widehat{\mathbf{U}}^*_{k,\mathcal{S}}\mathbf{x}_{\text{test},\mathcal{S}}\right\|_2^2\\ &= \text{Tr}\left\{\mathbf{C}_{\mathbf{x}}\right\}-\text{Tr}\left\{\widehat{\mathbf{U}}_{k,\mathcal{S}}\widehat{\mathbf{U}}^*_{k,\mathcal{S}}\mathbf{C}_{\mathbf{x},\mathcal{S}}\widehat{\mathbf{U}}_{k,\mathcal{S}}\widehat{\mathbf{U}}^*_{k,\mathcal{S}}\right\}\\ &= \text{Tr}\left\{\mathbf{C}_{\mathbf{x}}\right\}\\ &\quad-\text{Tr}\left\{\widehat{\mathbf{\Psi}}_{\mathcal{S}}\widehat{\mathbf{\Lambda}}_{\mathcal{S},\text{ind}[k]}\widehat{\mathbf{\Psi}}^*_{\mathcal{S}}\mathbf{\Psi}_{\mathcal{S}}\mathbf{\Lambda}_{\mathcal{S}}\mathbf{\Psi}^*_{\mathcal{S}}\widehat{\mathbf{\Psi}}_{\mathcal{S}}\widehat{\mathbf{\Lambda}}_{\mathcal{S},\text{ind}[k]}\widehat{\mathbf{\Psi}}^*_{\mathcal{S}}\right\}\\ &= \text{Tr}\left\{\mathbf{C}_{\mathbf{x}}\right\}-\text{Tr}\left\{\widehat{\mathbf{\Lambda}}_{\mathcal{S},\text{ind}[k]}\widehat{\mathbf{\Psi}}^*_{\mathcal{S}}\mathbf{\Psi}_{\mathcal{S}}\mathbf{\Lambda}_{\mathcal{S}}\mathbf{\Psi}^*_{\mathcal{S}}\widehat{\mathbf{\Psi}}_{\mathcal{S}}\right\}\\ &= \sum_{i=1}^{d}\lambda^{(i)}-\sum_{i\in\mathcal{S}}\lambda^{(i)}_{\mathcal{S},\text{ind}[k]}\sum_{j=1}^{p}\lambda^{(j)}_{\mathcal{S}}\left|\left\langle\boldsymbol{\psi}^{(j)}_{\mathcal{S}},\widehat{\boldsymbol{\psi}}^{(i)}_{\mathcal{S}}\right\rangle\right|^2\\ &= \sum_{i=1}^{d}\lambda^{(i)}-\sum_{i\in\widehat{\mathcal{S}}^{(k)}_{\text{max}}}\sum_{j=1}^{p}\lambda^{(j)}_{\mathcal{S}}\left|\left\langle\boldsymbol{\psi}^{(j)}_{\mathcal{S}},\widehat{\boldsymbol{\psi}}^{(i)}_{\mathcal{S}}\right\rangle\right|^2\end{aligned}\tag{A.4}$$

where $\widehat{\mathcal{S}}^{(k)}_{\text{max}}\subset\{1,\dots,p\}$ is the set of k indices corresponding to the columns of $\widehat{\mathbf{\Psi}}_{\mathcal{S}}$ used for the construction of $\widehat{\mathbf{U}}_{k,\mathcal{S}}$. This means that the indices in $\widehat{\mathcal{S}}^{(k)}_{\text{max}}$ correspond to the k maximal eigenvalues of $\widehat{\mathbf{C}}^{(n)}_{\mathbf{x},\mathcal{S}}$. If $k>\rho$, then $k-\rho$ of the indices in $\widehat{\mathcal{S}}^{(k)}_{\text{max}}$ correspond to zero eigenvalues.

A.4. Proof of Proposition 3.3

The error expression provided in Proposition 3.2 has the property that

$$\begin{aligned}&\mathcal{E}^{\text{unsup}}_{\text{out}}\left(\widehat{\mathbf{U}}_{k+1}\right)=\\ &\mathcal{E}^{\text{unsup}}_{\text{out}}\left(\widehat{\mathbf{U}}_k\right)-\sum_{j=1}^{p}\lambda^{(j)}_{\mathcal{S}}\left|\left\langle\boldsymbol{\psi}^{(j)}_{\mathcal{S}},\widehat{\boldsymbol{\psi}}^{(i_{\text{added}})}_{\mathcal{S}}\right\rangle\right|^2\end{aligned}\tag{A.5}$$

where $i_{\text{added}}\in\{1,\dots,p\}\setminus\widehat{\mathcal{S}}^{(k)}_{\text{max}}$ is the index of the column of $\widehat{\mathbf{\Psi}}_{\mathcal{S}}$ that is joined to $\widehat{\mathbf{U}}_k$ as the $(k+1)$-th column that yields $\widehat{\mathbf{U}}_{k+1}$. Note that $\lambda^{(j)}_{\mathcal{S}}\geq 0$ for any j, as these are eigenvalues of a covariance matrix. Hence,

$$\sum_{j=1}^{p}\lambda^{(j)}_{\mathcal{S}}\left|\left\langle\boldsymbol{\psi}^{(j)}_{\mathcal{S}},\widehat{\boldsymbol{\psi}}^{(i_{\text{added}})}_{\mathcal{S}}\right\rangle\right|^2\geq 0.\tag{A.6}$$

This implies that $\mathcal{E}^{\text{unsup}}_{\text{out}}\left(\widehat{\mathbf{U}}_{k+1}\right)\leq\mathcal{E}^{\text{unsup}}_{\text{out}}\left(\widehat{\mathbf{U}}_k\right)$, proving that the unsupervised out-of-sample error is monotonic decreasing in k (for a subspace construction that is sequential in k as described above).

A.5. On the Monotonic Decrease of $\mathcal{E}^{\text{unsup}}_{\text{out}}\left(\widehat{\mathbf{U}}_k\right)$ with p

We now justify our statement regarding the monotonic decrease of $\mathcal{E}^{\text{unsup}}_{\text{out}}\left(\widehat{\mathbf{U}}_k\right)$ as the number of features, p, increases (and k is kept fixed).

The following definitions and notations will be useful in the current discussion. Consider a set $\mathcal{S}_p\triangleq\{s_1,...,s_p\}$ of $p<d$ coordinates $1\leq s_1<s_2<\dots<s_p\leq d$. In addition, $\mathcal{S}_{p+1}\triangleq\mathcal{S}_p\cup\{s_{p+1}\}$ is a set of $p+1$ coordinates that is formed by adding a new coordinate $s_{p+1}\in\{1,\dots,d\}\setminus\mathcal{S}_p$ to $\mathcal{S}_p$. We also denote here the out-of-sample errors of interest with explicit indications of the underlying sets of coordinates: $\mathcal{E}^{\text{unsup}}_{\text{out}}\left(\widehat{\mathbf{U}}_k;\mathcal{S}_p\right)$ and $\mathcal{E}^{\text{unsup}}_{\text{out}}\left(\widehat{\mathbf{U}}_k;\mathcal{S}_{p+1}\right)$ are the errors induced by forming subspace estimates based on $\mathcal{S}_p$ and $\mathcal{S}_{p+1}$, respectively. Now, our goal is to justify the claim that

$$\mathcal{E}^{\text{unsup}}_{\text{out}}\left(\widehat{\mathbf{U}}_k;\mathcal{S}_p\right)\geq\mathcal{E}^{\text{unsup}}_{\text{out}}\left(\widehat{\mathbf{U}}_k;\mathcal{S}_{p+1}\right).\tag{A.7}$$

Using the error expression in (A.4), we translate the inequality (A.7) into

$$\begin{aligned}&\sum_{i\in\widehat{\mathcal{S}}^{(k)}_{p,\max}} \sum_{j=1}^{p} \lambda^{(j)}_{\mathcal{S}_p} \left|\left\langle \boldsymbol{\psi}^{(j)}_{\mathcal{S}_p}, \widehat{\boldsymbol{\psi}}^{(i)}_{\mathcal{S}_p}\right\rangle\right|^2 \\ &\leq \sum_{i\in\widehat{\mathcal{S}}^{(k)}_{p+1,\max}} \sum_{j=1}^{p+1} \lambda^{(j)}_{\mathcal{S}_{p+1}} \left|\left\langle \boldsymbol{\psi}^{(j)}_{\mathcal{S}_{p+1}}, \widehat{\boldsymbol{\psi}}^{(i)}_{\mathcal{S}_{p+1}}\right\rangle\right|^2.\end{aligned} \tag{A.8}$$

Here, the covariance matrix of the p-feature vector induced by $\mathcal{S}_p$ is $\mathbf{C}_{\mathbf{x},\mathcal{S}_p} \triangleq \mathbb{E}\{\mathbf{x}_{\mathcal{S}_p}\mathbf{x}_{\mathcal{S}_p}^T\}$, and its eigenvalues and eigenvectors are $\left\{\lambda^{(j)}_{\mathcal{S}_p}\right\}_{j=1}^{p}$ and $\left\{\boldsymbol{\psi}^{(j)}_{\mathcal{S}_p}\right\}_{j=1}^{p}$, respectively. Similarly, the covariance matrix of the $(p+1)$-feature vector stemming from $\mathcal{S}_{p+1}$ is $\mathbf{C}_{\mathbf{x},\mathcal{S}_{p+1}} \triangleq \mathbb{E}\{\mathbf{x}_{\mathcal{S}_{p+1}}\mathbf{x}_{\mathcal{S}_{p+1}}^T\}$, and its eigenvalues and eigenvectors are $\left\{\lambda^{(j)}_{\mathcal{S}_{p+1}}\right\}_{j=1}^{p+1}$ and $\left\{\boldsymbol{\psi}^{(j)}_{\mathcal{S}_{p+1}}\right\}_{j=1}^{p+1}$, respectively. To distinguish between the various origins of $\widehat{\mathcal{S}}^{(k)}_{\max}$, we define here the notation of $\widehat{\mathcal{S}}^{(k)}_{p,\max}$ as the set of k coordinates utilized based on the p-dimensional sample covariance matrix. Correspondingly, the set $\widehat{\mathcal{S}}^{(k)}_{p+1,\max}$ includes k coordinates selected based on the $(p+1)$-dimensional sample covariance matrix.

For a start, note that the sums in (A.8) are over non-negative elements. Moreover, the inner summation on the right-hand side of (A.8) is over $p+1$ terms, whereas its counterpart sum on the left-hand side is over p terms. However, the eigenvalues and eigenvectors in the two sides of (A.8) are different, as will be explained next.

The $p\times p$ covariance matrix $\mathbf{C}_{\mathbf{x},\mathcal{S}_p}$ is a principal submatrix of $\mathbf{C}_{\mathbf{x},\mathcal{S}_{p+1}}$, which is the covariance matrix of the $(p+1)$-feature vector induced by $\mathcal{S}_{p+1}$. This can be easily observed by defining the $p\times(p+1)$ matrix $\mathbf{Q}$ such that $\mathbf{x}_{\mathcal{S}_p} = \mathbf{Q}\mathbf{x}_{\mathcal{S}_{p+1}}$; namely, $\mathbf{Q}$ deletes the single feature added to create $\mathbf{x}_{\mathcal{S}_{p+1}}$ from $\mathbf{x}_{\mathcal{S}_p}$. Then,

$$\begin{aligned}\mathbf{C}_{\mathbf{x},\mathcal{S}_p} &= \mathbb{E}\{\left(\mathbf{Q}\mathbf{x}_{\mathcal{S}_{p+1}}\right)\left(\mathbf{Q}\mathbf{x}_{\mathcal{S}_{p+1}}\right)^T\}\\ &= \mathbf{Q}\mathbb{E}\{\mathbf{x}_{\mathcal{S}_{p+1}}\mathbf{x}_{\mathcal{S}_{p+1}}^T\}\mathbf{Q}^T\\ &= \mathbf{Q}\mathbf{C}_{\mathbf{x},\mathcal{S}_{p+1}}\mathbf{Q}^T.\end{aligned} \tag{A.9}$$

This shows that the matrix $\mathbf{C}_{\mathbf{x},\mathcal{S}_p}$ can be obtained from $\mathbf{C}_{\mathbf{x},\mathcal{S}_{p+1}}$ by deletion of the row and column (having the same index) corresponding to the added feature. Thus, $\mathbf{C}_{\mathbf{x},\mathcal{S}_p}$ is a principal submatrix of $\mathbf{C}_{\mathbf{x},\mathcal{S}_{p+1}}$. This relation between the symmetric matrices $\mathbf{C}_{\mathbf{x},\mathcal{S}_p}$ and $\mathbf{C}_{\mathbf{x},\mathcal{S}_{p+1}}$, lets us apply Cauchy's interlacing theorem for eigenvalues of Hermitian matrices (Hwang, 2004) to obtain

$$\begin{aligned}\lambda^{(\mathrm{sort}[p+1])}_{\mathcal{S}_{p+1}} \leq \lambda^{(\mathrm{sort}[p])}_{\mathcal{S}_p} \leq \lambda^{(\mathrm{sort}[p])}_{\mathcal{S}_{p+1}} \leq \lambda^{(\mathrm{sort}[p-1])}_{\mathcal{S}_p} \leq \dots\\ \dots \leq \lambda^{(\mathrm{sort}[2])}_{\mathcal{S}_{p+1}} \leq \lambda^{(\mathrm{sort}[1])}_{\mathcal{S}_p} \leq \lambda^{(\mathrm{sort}[1])}_{\mathcal{S}_{p+1}}\end{aligned} \tag{A.10}$$

where the eigenvalues of each of the matrices are referred to in a sorted order, namely,

$$\lambda^{(\mathrm{sort}[p+1])}_{\mathcal{S}_{p+1}} \leq \lambda^{(\mathrm{sort}[p])}_{\mathcal{S}_{p+1}} \leq \dots \leq \lambda^{(\mathrm{sort}[2])}_{\mathcal{S}_{p+1}} \leq \lambda^{(\mathrm{sort}[1])}_{\mathcal{S}_{p+1}} \tag{A.11}$$

are the sorted eigenvalues of $\mathbf{C}_{\mathbf{x},\mathcal{S}_{p+1}}$, and

$$\lambda^{(\mathrm{sort}[p])}_{\mathcal{S}_p} \leq \lambda^{(\mathrm{sort}[p-1])}_{\mathcal{S}_p} \leq \dots \leq \lambda^{(\mathrm{sort}[2])}_{\mathcal{S}_p} \leq \lambda^{(\mathrm{sort}[1])}_{\mathcal{S}_p} \tag{A.12}$$

are the sorted eigenvalues of $\mathbf{C}_{\mathbf{x},\mathcal{S}_p}$.

The interlaced structure of the eigenvalue inequalities in (A.10) provides an interesting aspect to the analysis of the desired inequality in (A.8). To see this, we rearrange (A.8) to rely on the sorted indexing of (A.11)-(A.12) and change the order of the nested summations, namely, the inequality under question (A.8) becomes

$$\sum_{j=1}^{p} \alpha_p^{(j)} \lambda^{(\mathrm{sort}[j])}_{\mathcal{S}_p} \leq \sum_{j=1}^{p+1} \alpha_{p+1}^{(j)} \lambda^{(\mathrm{sort}[j])}_{\mathcal{S}_{p+1}} \tag{A.13}$$

where

$$\begin{aligned}&\alpha_p^{(j)} \triangleq \sum_{i\in\widehat{\mathcal{S}}^{(k)}_{p,\max}} \left|\left\langle \boldsymbol{\psi}^{(\mathrm{sort}[j])}_{\mathcal{S}_p}, \widehat{\boldsymbol{\psi}}^{(i)}_{\mathcal{S}_p}\right\rangle\right|^2\\ &\text{for } j=1,\dots,p, \text{ and}\\ &\alpha_{p+1}^{(j)} \triangleq \sum_{i\in\widehat{\mathcal{S}}^{(k)}_{p+1,\max}} \left|\left\langle \boldsymbol{\psi}^{(\mathrm{sort}[j])}_{\mathcal{S}_{p+1}}, \widehat{\boldsymbol{\psi}}^{(i)}_{\mathcal{S}_{p+1}}\right\rangle\right|^2\\ &\text{for } j=1,\dots,p+1.\end{aligned} \tag{A.14}$$

The value of $\alpha_p^{(j)}$ reflects the quality of approximating the true eigenvector $\boldsymbol{\psi}^{(\mathrm{sort}[j])}_{\mathcal{S}_p}$ by the set of k sample eigenvectors $\left\{\widehat{\boldsymbol{\psi}}^{(i)}_{\mathcal{S}_p}\right\}_{i\in\widehat{\mathcal{S}}^{(k)}_{p,\max}}$. The value of $\alpha_{p+1}^{(j)}$ has a similar meaning (with respect to $\mathcal{S}_{p+1}$).

Note that $\alpha_p^{(j)}$ and $\alpha_{p+1}^{(j)}$ are values in the range $[0,1]$. However, since (A.14) depends on the true and sample eigenvectors of covariance matrices and their submatrices, its characterization is very complex. To generally understand the difficulty in the mathematical analysis of (A.14), one can examine the study of the eigenvalue-eigenvector relations provided in (Denton et al., 2019) that, although being simpler than our case, leads to intricate expressions that are under current research.

The above analysis leads us to choose an empirical approach for justifying our statement on the decay of the out-of-sample error $\mathcal{E}_{\mathrm{out}}^{\mathrm{unsup}}\left(\widehat{\mathbf{U}}_k;\mathcal{S}_p\right)$ with the increase in the number of features p. The experiment settings, referring to the data model provided in Section 2 of the main text, are as follows. The data vectors are of dimension $d=128$

and only $n = 70$ examples are given. The linear subspace in the noisy linear data model is of dimension $m = 40$, which is also the number of columns of $\mathbf{U}_m$. Each of the experiments below consider one of the following structures for columns of $\mathbf{U}_m$:

- The first $m = 40$ normalized columns of the $d \times d$ *Hadamard* matrix (these normalized columns are, by definition, orthonormnal).
- $m = 40$ *random* orthonormal vectors that are a subset of the left singular vectors of a $d \times d$ Gaussian matrix of i.i.d. components $\mathcal{N}(0, 1)$.

These Hadamard and random constructions are *global* in the sense that they are defined using all the d coordinates of the feature space. However, unlike the random form, the Hadamard form has a deterministic structure. In all the settings $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$, but we consider two different levels of noise (that is represented by the variable ϵ in the data model (1)): $\sigma_\epsilon = 0.1$ and $\sigma_\epsilon = 0.5$.

For a start, we exemplify the evolution of the eigenvalues $\left\{\lambda_{\mathcal{S}_p}^{(\mathrm{sort}[j])}\right\}_{j=1}^{p}$ with p. We consider three different settings as described in the caption of Fig. A.1. Figures A.1a, A.1d, A.1g clearly show the monotonic increase explained by the application of Cauchy's interlacing theorem in (A.10). The corresponding behavior of $\left\{\alpha_p^{(j)}\right\}_{j=1}^{p}$ (see Figures A.1b, A.1e, A.1h) is indeed intricate as mentioned above. Specifically, Fig. A.1e shows the effect of an increased noise level. Fig. A.1h demonstrates the consequence of estimating a subspace of an incorrect dimension. Despite the complex behavior of $\left\{\alpha_p^{(j)}\right\}_{j=1}^{p}$, Figures A.1c, A.1f, A.1i present that the resulting out-of-sample errors monotonically decrease *on average* (where $\mathcal{S}_p$ is uniformly chosen at random) with the increase in p (see solid blue curves in Figs. A.1c, A.1f, A.1i). This is explained next.

We now proceed to the empirical results that explain the decay of the out-of-sample error $\mathcal{E}_{\mathrm{out}}^{\mathrm{unsup}}\left(\widehat{\mathbf{U}}_k; \mathcal{S}_p\right)$ with the increase in p. Figures A.1c, A.1f present the evolution of the out-of-sample error for estimated subspaces of dimension $k = m$ (i.e., the true subspace dimension is known) and Fig. A.1i corresponds to $k = 10 < m$ (namely, an incorrect dimension). Each figure contains two curves: the dotted red curves present the sequence of errors $\mathcal{E}_{\mathrm{out}}^{\mathrm{unsup}}\left(\widehat{\mathbf{U}}_k; \mathcal{S}_p\right)$ induced by a single sequential construction of $\mathcal{S}_p$; the solid blue curves show the sequence of *averages* over the errors $\mathcal{E}_{\mathrm{out}}^{\mathrm{unsup}}\left(\widehat{\mathbf{U}}_k; \mathcal{S}_p\right)$ induced by 500 different (and uniformly chosen at random) sequential constructions of $\mathcal{S}_p$.

Figures A.1c, A.1f, A.1i show that, *on average*, adding features is beneficial and reduces $\mathcal{E}_{\mathrm{out}}^{\mathrm{unsup}}\left(\widehat{\mathbf{U}}_k; \mathcal{S}_p\right)$. However, for a *specific and arbitrary* order of adding features, there is no guarantee that each added feature is indeed useful (for example, see the dotted red curve in Fig. A.1i that does not exhibit a monotonic decreasing trend). The results also show that the deviation from monotonicity is larger for higher noise levels and/or significant differences between the dimensions of the estimated and true subspaces. Corresponding experiments for the *random* subspace setting, are provided in Fig. A.2 and further support the findings of the Hadamard case discussed above.

The results in Figures A.1c,A.1f,A.1i are only for several values of k. Therefore, we also present results for the entire range possible for the dimension of the subspace estimate, i.e., $k = 1, \ldots, d$. This extensive set of experiments is provided in Fig. A.3 in a summarized form described as follows. We again use the notation emphasizing the dependency of the error on p, namely, $\mathcal{E}_{\mathrm{out}}^{\mathrm{unsup}}\left(\widehat{\mathbf{U}}_k; \mathcal{S}_p\right)$. For the various settings, we are interested in assessing the monotonic decrease of the error curve of $\mathcal{E}_{\mathrm{out}}^{\mathrm{unsup}}\left(\widehat{\mathbf{U}}_k; \mathcal{S}_p\right)$ over the (discrete) range of $p = k, \ldots, d$. Hence, we evaluate the monotonicity of the discrete sequence $\left\{\mathcal{E}_{\mathrm{out}}^{\mathrm{unsup}}\left(\widehat{\mathbf{U}}_k; \mathcal{S}_j\right)\right\}_{j=k}^{d}$ by computing the relative number of feature additions that reduced (or kept) the error. Namely, this metric is defined as

$$\eta\left(\left\{\mathcal{E}_{\mathrm{out}}^{\mathrm{unsup}}\left(\widehat{\mathbf{U}}_k; \mathcal{S}_j\right)\right\}_{j=k}^{d}\right) \triangleq \frac{\sum_{j=k+1}^{d} \mathbb{I}\left\{\mathcal{E}_{\mathrm{out}}^{\mathrm{unsup}}\left(\widehat{\mathbf{U}}_k; \mathcal{S}_j\right) - \mathcal{E}_{\mathrm{out}}^{\mathrm{unsup}}\left(\widehat{\mathbf{U}}_k; \mathcal{S}_{j-1}\right) \leq 0\right\}}{d - k} \tag{A.15}$$

where $\mathbb{I}\{\cdot\}$ is an indicator function returning 1 if the condition is applied on is true and 0 otherwise. Essentially, the metric (A.15) summarizes the monotonicity of an entire error curve into a single value in the range $[0, 1]$. An error curve with $\eta\left(\left\{\mathcal{E}_{\mathrm{out}}^{\mathrm{unsup}}\left(\widehat{\mathbf{U}}_k; \mathcal{S}_j\right)\right\}_{j=k}^{d}\right) = 1$ is monotonic decreasing over the entire range of p.

In Fig. A.3 we exhibit the values of the monotonicity metric for a variety of settings, including subspaces in the Hadamard and random forms (note that the horizontal axes represent the dimension of the subspace estimate). Each subfigure includes two curves: the dotted red curves present the monotonicity metric values induced by individual sequential constructions of $\mathcal{S}_p$; the solid blue curves show the monotonicity metric values obtained for curves of errors *averaged* over 500 experiments differing in their sequential constructions of $\mathcal{S}_p$. Clearly, specific orders of adding features do not necessarily yield error curves that are purely monotonically decreasing. However, the *averaged* error curves are monotonic decreasing over the entire range of p (and this is the case for any k; see blue-colored curves in Fig. A.3). We take the results of these and numerous similar simulations with other parameter settings as strong experimental evidence that, on average, $\mathcal{E}_{\text{out}}^{\text{unsup}}\left(\widehat{\mathbf{U}}_k;\mathcal{S}_p\right)$ decays with the increase in p.

B. Proofs and Additional Details for Section 4

B.1. On the Singular Values of Rectangular, Tall Matrices with Orthonormal Columns

A tall, rectangular matrix $\mathbf{W}\in\mathbb{R}^{p\times m}$ (where $p\geq m$) has orthonormal columns if and only if all of its singular values equal 1. This is proved next.

Consider a real matrix $\mathbf{W}\in\mathbb{R}^{p\times m}$ (where $p\geq m$) with orthonormal columns. Then, the corresponding SVD is $\mathbf{W}=\mathbf{\Omega}\mathbf{\Sigma}\mathbf{\Theta}^T$, where $\mathbf{\Omega}$ and $\mathbf{\Theta}$ are $p\times p$ and $m\times m$ real orthonormal matrices, respectively, and $\mathbf{\Sigma}$ is a $p\times m$ real diagonal matrix with m singular values $\left\{\sigma_i\left(\mathbf{W}\right)\right\}_{i=1}^{m}$ on its main diagonal. Since $\mathbf{W}$ has orthonormal columns, we can write $\mathbf{W}^T\mathbf{W}=\mathbf{I}_m$. Using the SVD form we get that

$$\mathbf{I}_m=\left(\mathbf{\Omega}\mathbf{\Sigma}\mathbf{\Theta}^T\right)^T\mathbf{\Omega}\mathbf{\Sigma}\mathbf{\Theta}^T=\mathbf{\Theta}\mathbf{\Sigma}^T\mathbf{\Sigma}\mathbf{\Theta}^T \tag{B.1}$$

and this can be translated into

$$\mathbf{I}_m=\mathbf{\Sigma}^T\mathbf{\Sigma}. \tag{B.2}$$

Since singular values are, by definition, non-negative real values, then Eq. (B.2) implies that $\sigma_i\left(\mathbf{W}\right)=1$ for $i=1,\ldots,m$. This proves the left-to-right direction of the statement.

The second direction is proved as follows. Consider a real matrix $\mathbf{W}\in\mathbb{R}^{p\times m}$ (where $p\geq m$) with SVD $\mathbf{W}=\mathbf{\Omega}\mathbf{\Sigma}\mathbf{\Theta}^T$, where $\mathbf{\Omega}$ and $\mathbf{\Theta}$ are $p\times p$ and $m\times m$ real orthonormal matrices, respectively, and $\mathbf{\Sigma}$ is a $p\times m$ real diagonal matrix with m singular values $\sigma_i\left(\mathbf{W}\right)=1$ for $i=1,\ldots,m$ on its main diagonal. This means that $\mathbf{\Sigma}^T\mathbf{\Sigma}=\mathbf{I}_m$. Then,

$$\begin{aligned}\mathbf{W}^T\mathbf{W}&=\left(\mathbf{\Omega}\mathbf{\Sigma}\mathbf{\Theta}^T\right)^T\mathbf{\Omega}\mathbf{\Sigma}\mathbf{\Theta}^T\\&=\mathbf{\Theta}\mathbf{\Sigma}^T\mathbf{\Sigma}\mathbf{\Theta}^T=\mathbf{\Theta}\mathbf{\Theta}^T=\mathbf{I}_m\end{aligned} \tag{B.3}$$

implying that the columns of $\mathbf{W}$ are orthonormal. This completes the proof of the entire statement.

B.2. The Hard Orthonormality-Constraints Projection Operator T_{hard}

The operator projecting onto the hard orthonormality constraints was defined in Section 4.1 as follows. Consider a matrix $\mathbf{W}^{(\text{in})}\in\mathbb{R}^{p\times m}$ (where $p\geq m$), with the SVD $\mathbf{W}^{(\text{in})}=\mathbf{\Omega}\mathbf{\Sigma}^{(\text{in})}\mathbf{\Theta}^T$, where $\mathbf{\Omega}$ and $\mathbf{\Theta}$ are $p\times p$ and $m\times m$ real orthonormal matrices, respectively, and $\mathbf{\Sigma}^{(\text{in})}$ is a $p\times m$ real diagonal matrix with m singular values $\left\{\sigma_i\left(\mathbf{W}^{(\text{in})}\right)\right\}_{i=1}^{m}$ on its main diagonal. Then, projecting $\mathbf{W}^{(\text{in})}$ onto the hard-orthonormality constraint via

$$\mathbf{W}^{(\text{out})}=\underset{\mathbf{W}\in\mathbb{R}^{p\times m}:\ \mathbf{W}^T\mathbf{W}=\mathbf{I}_m}{\arg\min}\left\|\mathbf{W}-\mathbf{W}^{(\text{in})}\right\|_F^2 \tag{B.4}$$

induces the mapping $\mathbf{W}^{(\text{out})}\triangleq T_{\text{hard}}\left(\mathbf{W}^{(\text{in})}\right)$, where $\mathbf{W}^{(\text{out})}=\mathbf{\Omega}\mathbf{\Sigma}^{(\text{out})}\mathbf{\Theta}^T$ and the singular values along the main diagonal of $\mathbf{\Sigma}^{(\text{out})}$ are $\sigma_i\left(\mathbf{W}^{(\text{out})}\right)=1$ for $i=1,\ldots,m$. A relevant proof is available in (Kahan, 2011) and also in a more general form in (Keller, 1975).

B.3. The Soft Orthonormality-Constraints Projection Operator T_α

The projection of a given matrix $\mathbf{W}^{(\text{in})}\in\mathbb{R}^{p\times m}$ (where $p\geq m$) was defined in the main paper (see Eq. (16)) as follows. Consider the SVD $\mathbf{W}^{(\text{in})}=\mathbf{\Omega}\mathbf{\Sigma}^{(\text{in})}\mathbf{\Theta}^T$, where $\mathbf{\Omega}$ and $\mathbf{\Theta}$ are $p\times p$ and $m\times m$ real orthonormal matrices, respectively, and $\mathbf{\Sigma}^{(\text{in})}$ is a $p\times m$ real diagonal matrix with m singular values $\left\{\sigma_i\left(\mathbf{W}^{(\text{in})}\right)\right\}_{i=1}^{m}$ on its main diagonal. Then, the projection of $\mathbf{W}^{(\text{in})}$ on the soft-orthonormality constraints is defined in its basic form as

$$\begin{aligned}\mathbf{W}^{(\text{out})}&=\underset{\mathbf{W}\in\mathbb{R}^{p\times m}}{\arg\min}\left\|\mathbf{W}-\mathbf{W}^{(\text{in})}\right\|_F^2\\&\text{subject to }\left|\sigma_i^2\left(\mathbf{W}\right)-1\right|\leq\alpha\ \text{ for } i=1,...,m\end{aligned} \tag{B.5}$$

is equivalent to the thresholding mapping $\mathbf{W}^{(\text{out})}\triangleq T_\alpha\left(\mathbf{W}^{(\text{in})}\right)$ where $\mathbf{W}^{(\text{out})}=\mathbf{\Omega}\mathbf{\Sigma}^{(\text{out})}\mathbf{\Theta}^T$ and the singular values along the main diagonal of $\mathbf{\Sigma}^{(\text{out})}$ are

$$\sigma_i\left(\mathbf{W}^{(\text{out})}\right)= \tag{B.6}$$
$$\begin{cases}\sigma_i\left(\mathbf{W}^{(\text{in})}\right), & \text{if } \sigma_i\left(\mathbf{W}^{(\text{in})}\right)\in\left[\tau_\alpha^{\text{low}},\tau_\alpha^{\text{high}}\right]\\ \tau_\alpha^{\text{low}}, & \text{if } \sigma_i\left(\mathbf{W}^{(\text{in})}\right)<\tau_\alpha^{\text{low}}\\ \tau_\alpha^{\text{high}}, & \text{if } \sigma_i\left(\mathbf{W}^{(\text{in})}\right)>\tau_\alpha^{\text{high}}\end{cases}$$

for $i=1,...,m$, where the threshold levels are defined by $\tau_\alpha^{\text{low}}\triangleq\sqrt{\max\left\{0,1-\alpha\right\}}$ and $\tau_\alpha^{\text{high}}\triangleq\sqrt{1+\alpha}$. Also recall that singular values are non-negative by their definition.

The relation between (B.6) and (B.5) is based on the extension of the case of strict orthonormality constraints that was explained above and proved in (Kahan, 2011).

B.4. The Algorithm for Supervised Subspace Fitting with Soft Orthonormality Constraints

We present here the explicit form of the method proposed in Section 4.3 for supervised subspace fitting with soft orthonormality constraints, i.e., the numerical procedure to address the problem in (15). We utilize the projected gradient descent technique to obtain the procedure outlined in Algorithm B.1.

Similar to Algorithm 1, we initialize the iterative process by setting $\mathbf{W}^{(i=0)}$ by projecting the closed-form solution of the

Algorithm B.1 Supervised Subspace Fitting via Projected Gradient Descent: **Soft** Orthonormality Constraints

```
Input: a dataset $\mathcal{D}_{\mathcal{S}}^{\text{sup}} = \left\{ \left( \mathbf{x}_{\mathcal{S}}^{(\ell)}, \mathbf{z}^{(\ell)} \right) \right\}_{\ell=1}^{n}$, a coordinate subset $\mathcal{S}$, and a threshold level $\alpha \geq 0$
Initialize $\mathbf{W}^{(t=0)} = T_{\alpha} \left( \left( \mathbf{Z}\mathbf{X}_{\mathcal{S}}^{+} \right)^{T} \right)$, $t = 0$
repeat
    $t \leftarrow t + 1$
    $\mathbf{Y}^{(t)} = \mathbf{W}^{(t-1)} - \mu \mathbf{X}_{\mathcal{S}} \left( \left( \mathbf{W}^{(t-1)} \right)^{T} \mathbf{X}_{\mathcal{S}} - \mathbf{Z} \right)^{T}$
    $\mathbf{W}^{(t)} = T_{\alpha} \left( \mathbf{Y}^{(t)} \right)$
until stopping criterion is satisfied
Set $\widehat{\mathbf{U}}_{m,\mathcal{S}} = \mathbf{W}^{(t)}$
Create $\widehat{\mathbf{U}}_{m}$ based on $\widehat{\mathbf{U}}_{m,\mathcal{S}}$ and zeros at rows corresponding to $\mathcal{S}_{c}$
Output: $\widehat{\mathbf{U}}_{m}$
```

unconstrained supervised problem onto the orthonormality constraint of interest (here using the operator T_{α}). The gradient step size μ is updated in each iteration based on a simple line search mechanism that scales the former step size by finding the best within a set of update factors. This line search approach was also used in the implementation of Algorithm 1.

One can also implement the proposed Algorithms without the line search mechanism and instead set a fixed gradient step size based on the worst case gradient direction induced by the quadratic cost functions examined in this paper.

B.5. Additional Details on the Experiments in Section 4 (Supervised Settings)

In Section 4 of the main paper we present fully-supervised subspace fitting problems that are categorized into three types: strict orthonormally constrained (Section 4.1), unconstrained (essentially, a regression problem form, see Section 4.2), and soft orthonotmally constrained (Section 4.3). The empirical measurements of the out-of-sample errors of the various supervised settings are provided together in Fig. 3b (in the main text). We here elaborate on the settings of the experiments presented in Fig. 3.

Since the problems are supervised, then the dimension m of the true subspace is known. Accordingly, the results are only for estimation of m-dimensional representations. As usual, the data model is based on (1). Here the dimension of the entire space is $d = 64$, the true subspace dimension is $m = 20$, the number of examples is $n = 32$, and the noise in the model corresponds to $\sigma_{\epsilon} = 0.5$. Each of the curves in Fig. 3b presents the values $\mathcal{E}_{\text{out}}^{\text{unsup}} \left(\widehat{\mathbf{U}}_{k}; \mathcal{S}_{p} \right)$ versus p, which is the number of features used for the actual learning. The increase in p refers to a sequential extension of $\mathcal{S}_{p}$ to include additional coordinates of features to be utilized.

The results in Fig. 3b present smooth curves by conducting the corresponding experiments 10 times with different sequential constructions of $\mathcal{S}_{p}$ and then averaging the induced errors. We present in Fig. B.1 the corresponding non-smooth curves by conducting these experiments for a specific (but arbitrary) order of adding features (i.e., without averaging over multiple experiments).

Clearly, for the less orthonormally constrained settings (see the upper curves in Fig. 3b), the shape of the error curves resemble the double descent behavior, where the peak of each of these curves is obtained for $p = n - 1$ (the minus 1 is due to the centering of the n examples given). Importantly, after reaching the peak values, the out-of-sample errors start to decrease as the number of features increases and eventually achieving significantly lower error values than in the underparameterized range (i.e., for $p < n - 1$). This exemplifies the benefits of overparameterization in subspace fitting problems that are fully supervised and may have soft orthonormality constraints.

The settings that are nearly or (completely) orthonormally constrained (see the lower curves in Fig. 3b) present trends of decrease over the entire range of p. This may resemble the results presented above for unsupervised and strictly constrained subspace fitting. While these errors do not follow the double descent trend, they do exhibit the benefits of overparameterization even when the problem includes strict (or nearly strict) orthonormality constraints.

C. Additional Details for Section 5: The Algorithm for Semi-Supervised Subspace Fitting

Section 5 established an approach for semi-supervised subspace fitting with a flexible level of orthonormality constraints. The basic optimization problem is presented in (18) and does not have a closed-form solution. The following extends the details provided in the main text about the numerical procedure for addressing (18) using the projected gradient descent technique. Recall that in this semi-supervised setting there are two datasets in use: a supervised set of examples $\widetilde{\mathcal{D}}_{\mathcal{S}}^{\text{sup}} = \left\{ \left(\mathbf{x}_{\mathcal{S}}^{(\ell)}, \mathbf{z}^{(\ell)} \right) \right\}_{\ell=1}^{n^{\text{sup}}}$, and an unsupervised set of examples $\widetilde{\mathcal{D}}_{\mathcal{S}}^{\text{unsup}} = \left\{ \mathbf{x}_{\mathcal{S}}^{(\ell)} \right\}_{\ell=n^{\text{sup}}+1}^{n}$.

The proposed method is presented in Algorithm C.1. In contrast to Algorithms 1 and B.1 that address fully supervised settings, in the semi-supervised case the evolving solution is initialized to a random matrix, which contains i.i.d. Gaussian components with zero mean and variance $1/p$, that is projected onto the relevant orthonormality constraint (via the operator T_{α} that for $\alpha = 0$ is equivalent to T_{hard}). The data from the unsupervised exam-

Algorithm C.1 Semi-Supervised Subspace Fitting via Projected Gradient Descent (Soft Orthonormality Constraints)

```
Input: datasets $\widetilde{\mathcal{D}}_{\mathcal{S}}^{\text{sup}} = \left\{ \left( \mathbf{x}_{\mathcal{S}}^{(\ell)}, \mathbf{z}^{(\ell)} \right) \right\}_{\ell=1}^{n^{\text{sup}}}$ and $\widetilde{\mathcal{D}}_{\mathcal{S}}^{\text{unsup}} = \left\{ \mathbf{x}_{\mathcal{S}}^{(\ell)} \right\}_{\ell=n^{\text{sup}}+1}^{n}$, a coordinate subset $\mathcal{S}$, and a threshold level $\alpha \geq 0$
Initialize $\mathbf{W}^{(t=0)} = T_{\alpha}(\mathbf{H})$ where $\mathbf{H}$ is a $p \times m$ random Gaussian matrix of i.i.d. components $\mathcal{N}(0, 1/p)$, $t = 0$
repeat
    $t \leftarrow t + 1$
    $\mathbf{Y}^{(t)} = \mathbf{W}^{(t-1)} - \mu \cdot G^{\text{semisup}}\left(\mathbf{W}^{(t-1)}\right)$
    $\mathbf{W}^{(t)} = T_{\alpha}\left(\mathbf{Y}^{(t)}\right)$
until stopping criterion is satisfied
Set $\widehat{\mathbf{U}}_{m,\mathcal{S}} = \mathbf{W}^{(t)}$
Create $\widehat{\mathbf{U}}_{m}$ based on $\widehat{\mathbf{U}}_{m,\mathcal{S}}$ and zeros at rows corresponding to $\mathcal{S}_c$
Output: $\widehat{\mathbf{U}}_{m}$
```

ples, $\mathbf{X}_{\mathcal{S}}^{\text{unsup}} \triangleq \left[\mathbf{x}_{\mathcal{S}}^{(n^{\text{sup}}+1)}, \dots, \mathbf{x}_{\mathcal{S}}^{n} \right]$, is used in conjunction with the supervised examples in the gradient descent steps throughout the iterations of the algorithm.

Since (18) extends (15) only with respect to the optimization cost, then Algorithm C.1 simply extends Algorithm B.1 by updating the gradient used in the descent stage of the t^{th} iteration with

$$\begin{aligned} G^{\text{semisup}}\left(\mathbf{W}^{(t)}\right) &\triangleq \mathbf{X}_{\mathcal{S}}^{\text{sup}} \left(\left(\mathbf{W}^{(t)}\right)^{T} \mathbf{X}_{\mathcal{S}}^{\text{sup}} - \mathbf{Z}^{\text{sup}} \right)^{T} \\ &- 2\mathbf{X}_{\mathcal{S}}^{\text{unsup}} \left(\mathbf{X}_{\mathcal{S}}^{\text{unsup}}\right)^{T} \mathbf{W}^{(t)} \\ &+ \mathbf{X}_{\mathcal{S}}^{\text{unsup}} \left(\mathbf{X}_{\mathcal{S}}^{\text{unsup}}\right)^{T} \mathbf{W}^{(t)} \left(\mathbf{W}^{(t)}\right)^{T} \mathbf{W}^{(t)} \\ &+ \mathbf{W}^{(t)} \left(\mathbf{W}^{(t)}\right)^{T} \mathbf{X}_{\mathcal{S}}^{\text{unsup}} \left(\mathbf{X}_{\mathcal{S}}^{\text{unsup}}\right)^{T} \mathbf{W}^{(t)} \end{aligned} \tag{C.1}$$

that was obtained by differentiation of the semi-supervised cost function of (18), $\left\| \mathbf{Z}^{\text{sup}} - \mathbf{W}^{T} \mathbf{X}_{\mathcal{S}}^{\text{sup}} \right\|_F^2 + \left\| \left(\mathbf{I}_p - \mathbf{W}\mathbf{W}^{T} \right) \mathbf{X}_{\mathcal{S}}^{\text{unsup}} \right\|_F^2$, with respect to $\mathbf{W}$.

The gradient step size μ is updated in each iteration based on a simple line search approach that was described above for Algorithm B.1.

The error curves in Figures 4a and 4b are smooth due to averaging over 25 experiments with different sequential orders of adding coordinates to $\mathcal{S}$. In Figures C.1a and C.1b we provide the corresponding error curves obtained from a single experiment (i.e., for a single order of adding coordinates to $\mathcal{S}$).

D. Unsupervised Subspace Fitting with Soft Orthonormality Constraints

In Section 3.1 we defined the unsupervised form of the linear subspace fitting problem that included a strict orthonormality constraint and solved it via PCA. Now, we can define the corresponding range of unsupervised problems with flexible levels of orthonormality constraints, namely,

$$\begin{aligned} &\widehat{\mathbf{U}}_{m,\mathcal{S}} = \underset{\mathbf{W} \in \mathbb{R}^{p \times m}}{\arg\min} \left\| \left(\mathbf{I}_p - \mathbf{W}\mathbf{W}^{T} \right) \mathbf{X}_{\mathcal{S}} \right\|_F^2 \\ &\text{subject to } \left| \sigma_i^2 \left(\mathbf{W} \right) - 1 \right| \leq \alpha \ \text{ for } i = 1, ..., m, \end{aligned} \tag{D.1}$$

where we assume that m is known, $\mathbf{X}_{\mathcal{S}} \triangleq \left[\mathbf{x}_{\mathcal{S}}^{(1)}, \dots, \mathbf{x}_{\mathcal{S}}^{(n)} \right]$ is the data matrix corresponding to the (unsupervised) dataset that was considered in Section 3, and α determines the orthonormality constraint level. The optimization cost in (D.1) reflects the unsupervised aspect of the problem. We address (D.1) using the projected gradient descent method and get the process described in Algorithm D.1. As before, the soft orthonormality constraints induce a projection stage that uses the the soft-threshold projection T_{α} from (B.6). Importantly, unlike (B.6) we set the lower threshold to $\tau_{\alpha}^{\text{low}} \triangleq \sqrt{\max\left\{10^{-16}, 1 - \alpha\right\}}$ that avoids clipping of singular values to zero, and the upper threshold remains the same, i.e., $\tau_{\alpha}^{\text{high}} \triangleq \sqrt{1 + \alpha}$. Avoiding clipping the singular values to zero is important for maintaining the full rank of the evolving solution matrix throughout the (projected) gradient descent process. Unlike the supervised and semi-supervised settings, we empirically found that avoiding clipping singular values to zero is a crucial aspect in the unsupervised problems when optimized via projected gradient descent. The gradient descent step (in the t^{th} iteration) is based on the gradient of the unsupervised cost of (D.1), i.e.,

$$\begin{aligned} G^{\text{unsup}}\left(\mathbf{W}^{(t)}\right) \triangleq &-2\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{T}\mathbf{W}^{(t)} \\ &+ \mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{T}\mathbf{W}^{(t)} \left(\mathbf{W}^{(t)}\right)^{T} \mathbf{W}^{(t)} \\ &+ \mathbf{W}^{(t)} \left(\mathbf{W}^{(t)}\right)^{T} \mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{T}\mathbf{W}^{(t)}. \end{aligned} \tag{D.2}$$

Note that due to the unsupervised form of the problem we cannot initialize the process using the unconstrained linear regression solution (as we did in the Algorithms developed above for the fully supervised settings with soft orthonormality constraints). Therefore, the initialization in Algorithm D.1 sets $\mathbf{W}^{(i=0)}$ to a $p \times m$ matrix with i.i.d. Gaussian entries $\mathcal{N}(0, 1/p)$.

The empirical results obtained using Algorithm D.1 for a range of α values from zero (strictly constrained) to infinity (unconstrained) showed that all the respective solutions accurately follow the PCA solution obtained for the unsupervised problem with a strict orthonormality constraint (i.e., the solution obtained in Section 3 for $k = m$).

Algorithm D.1 Unsupervised Subspace Fitting via Projected Gradient Descent (Soft Orthonormality Constraints)

```
**Input:** a dataset $\mathcal{D}_{\mathcal{S}} = \left\{\mathbf{x}_{\mathcal{S}}^{(\ell)}\right\}_{\ell=1}^{n}$, a coordinate subset $\mathcal{S}$, $m$, and a threshold level $\alpha \geq 0$
**Initialize** $\mathbf{W}^{(t=0)} = T_\alpha\left(\mathbf{H}\right)$ where $\mathbf{H}$ is a $p \times m$ random Gaussian matrix of i.i.d. components $\mathcal{N}(0, 1/p)$, $t = 0$
**repeat**
    $t \leftarrow t + 1$
    $\mathbf{Y}^{(t)} = \mathbf{W}^{(t-1)} - \mu \cdot G^{\mathrm{unsup}}\left(\mathbf{W}^{(t-1)}\right)$
    $\mathbf{W}^{(t)} = T_\alpha\left(\mathbf{Y}^{(t)}\right)$
**until** stopping criterion is satisfied
Set $\widehat{\mathbf{U}}_{m,\mathcal{S}} = \mathbf{W}^{(t)}$
Create $\widehat{\mathbf{U}}_m$ based on $\widehat{\mathbf{U}}_{m,\mathcal{S}}$ and zeros at rows corresponding to $\mathcal{S}_c$
**Output:** $\widehat{\mathbf{U}}_m$
```

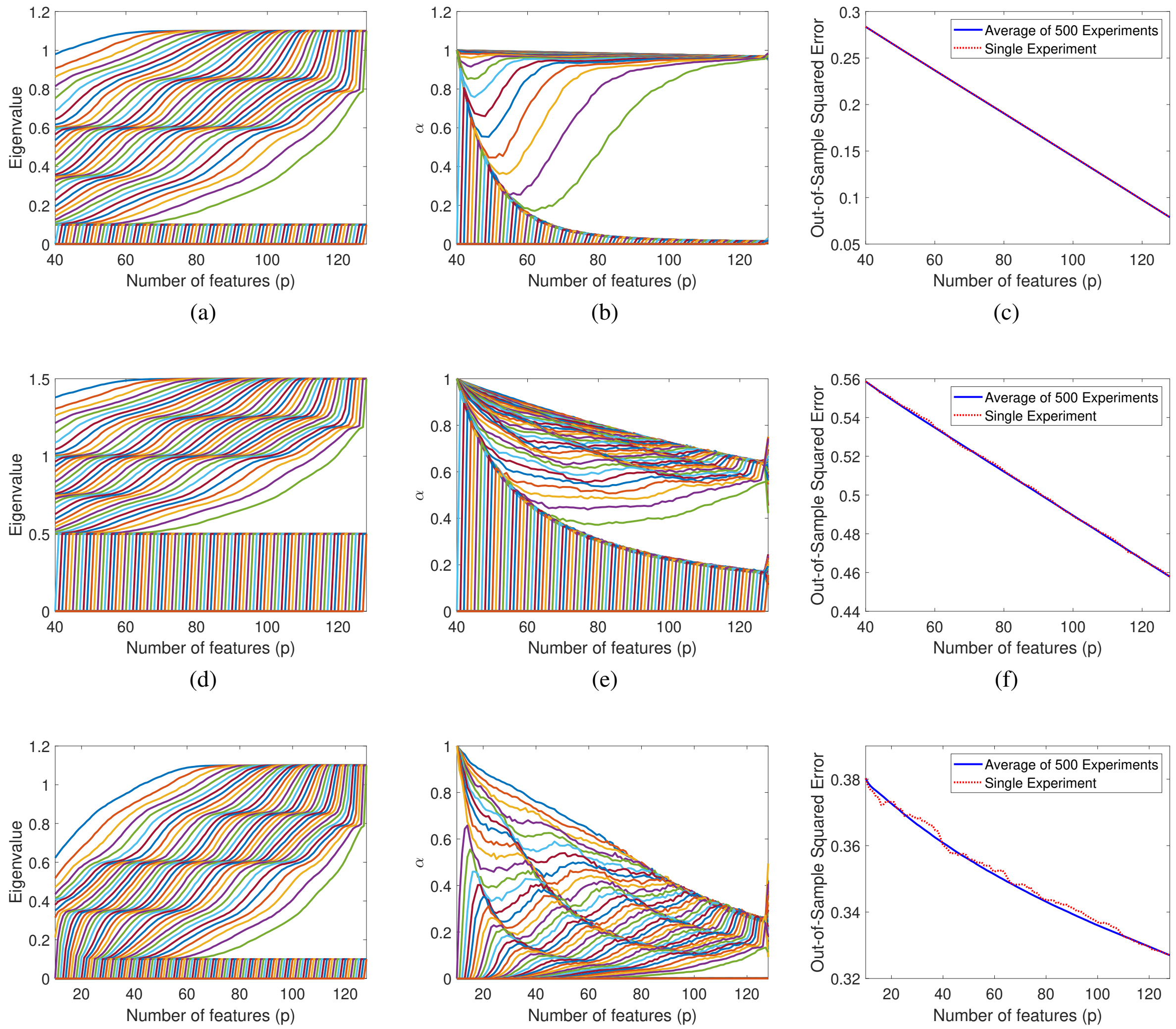
References

Denton, P. B., Parke, S. J., Tao, T., and Zhang, X. Eigenvectors from eigenvalues: A survey of a basic identity in linear algebra. *arXiv preprint arXiv:1908.03795*, 2019.

Hwang, S.-G. Cauchy's interlace theorem for eigenvalues of Hermitian matrices. *The American Mathematical Monthly*, 111(2):157–159, 2004.

Kahan, W. The nearest orthogonal or unitary matrix, August 2011. ”URL: https://people.eecs.berkeley.edu/~wkahan/Math128/NearestQ.pdf. Last visited on 2020/02/06”.

Keller, J. B. Closest unitary, orthogonal and Hermitian operators to a given operator. *Mathematics Magazine*, 48 (4):192–197, 1975.

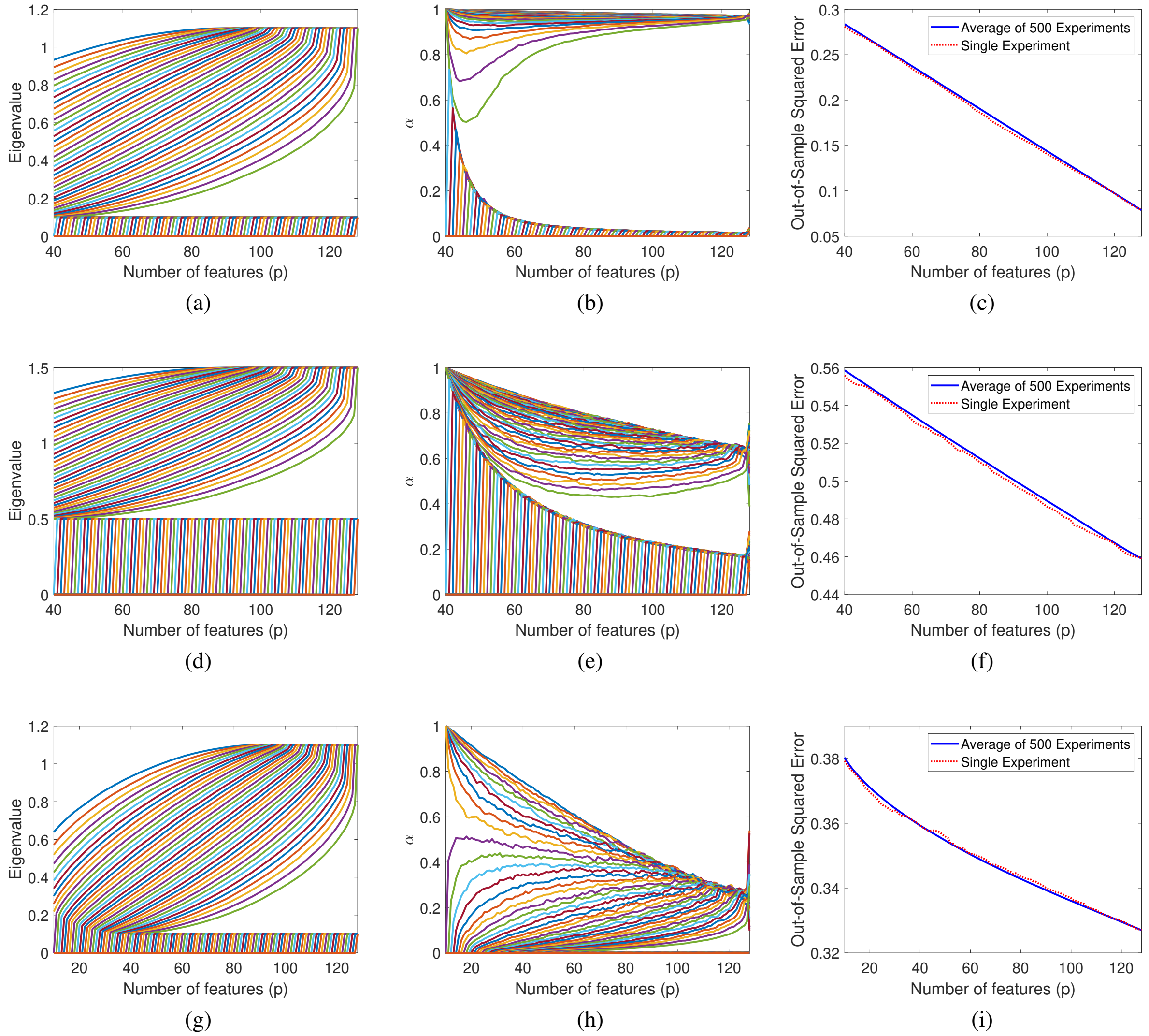


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure A.1. Empirical demonstration of the evolution of the components in (A.13) and the corresponding out-of-sample error $\mathcal{E}_{\text{out}}^{\text{unsup}}\left(\widehat{\mathbf{U}}_k; \mathcal{S}_p\right)$, and their evolution with the increase in the number of features p. Each line of subfigures corresponds to a different experimental setting, yet, for all of them the true subspace is of the **Hadamard** form, $d = 128$, $m = 40$, and $n = 70$. The first line of subfigures considers $k = m = 40$ and a noise level of $\sigma_\epsilon = 0.1$. The second line of subfigures corresponds to $k = m = 40$ and a noise level of $\sigma_\epsilon = 0.5$. The third line of subfigures corresponds to $k = 10$ and a noise level of $\sigma_\epsilon = 0.1$. (a), (d) and (g) present the sorted eigenvalues $\lambda_{\mathcal{S}_p}^{(\text{sort}[j])}$ of the true covariance matrices corresponding to p-feature vectors (each of the curves corresponds to another value of j). (b), (e) and (h) show the (sorted) coefficients $\alpha_p^{(j)}$ defined in (A.14). (c), (f) and (i) exhibit the out-of-sample error $\mathcal{E}_{\text{out}}^{\text{unsup}}\left(\widehat{\mathbf{U}}_k; \mathcal{S}_p\right)$ for a single instance of sequential increase of $\mathcal{S}_p$ (dotted red line) and for average over 500 different orders of sequentially increasing $\mathcal{S}_p$ (solid blue line).

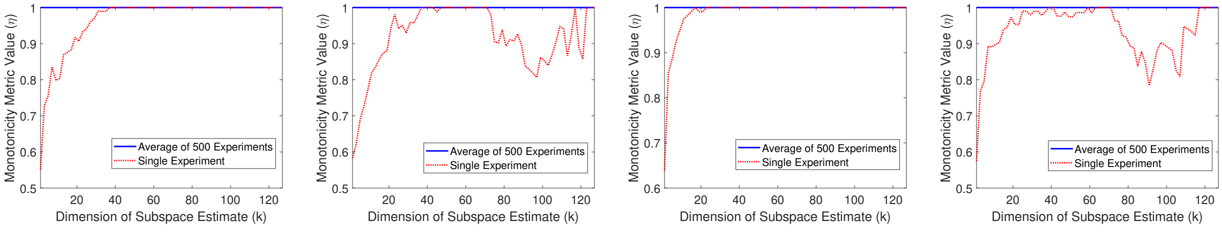


(a) (b) (c)

(d) (e) (f)

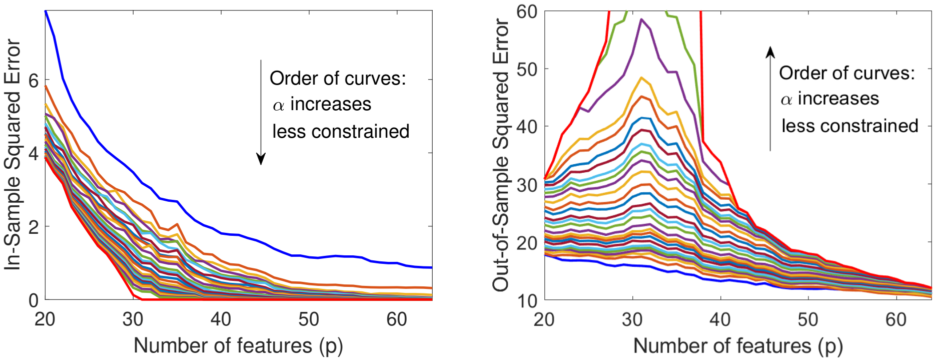
(g) (h) (i)

Figure A.2. Empirical demonstration of the evolution of the components in (A.13) and the corresponding out-of-sample error $\mathcal{E}_{\rm out}^{\rm unsup}\left(\widehat{\mathbf{U}}_k;\mathcal{S}_p\right)$, and their evolution with the increase in the number of features p. Each line of subfigures corresponds to a different experimental setting, yet, for all of them the true subspace is of the **random** form, $d=128$, $m=40$, and $n=70$. The first line of subfigures considers $k=m=40$ and a noise level of $\sigma_\epsilon=0.1$. The second line of subfigures corresponds to $k=m=40$ and a noise level of $\sigma_\epsilon=0.5$. The third line of subfigures corresponds to $k=10$ and a noise level of $\sigma_\epsilon=0.1$. (a), (d) and (g) present the sorted eigenvalues $\lambda_{\mathcal{S}_p}^{({\rm sort}[j])}$ of the true covariance matrices corresponding to p-feature vectors (each of the curves corresponds to another value of j). (b), (e) and (h) show the (sorted) coefficients $\alpha_p^{(j)}$ defined in (A.14). (c), (f) and (i) exhibit the out-of-sample error $\mathcal{E}_{\rm out}^{\rm unsup}\left(\widehat{\mathbf{U}}_k;\mathcal{S}_p\right)$ for a single instance of sequential increase of $\mathcal{S}_p$ (dotted red line) and for average over 500 different orders of sequentially increasing $\mathcal{S}_p$ (solid blue line).



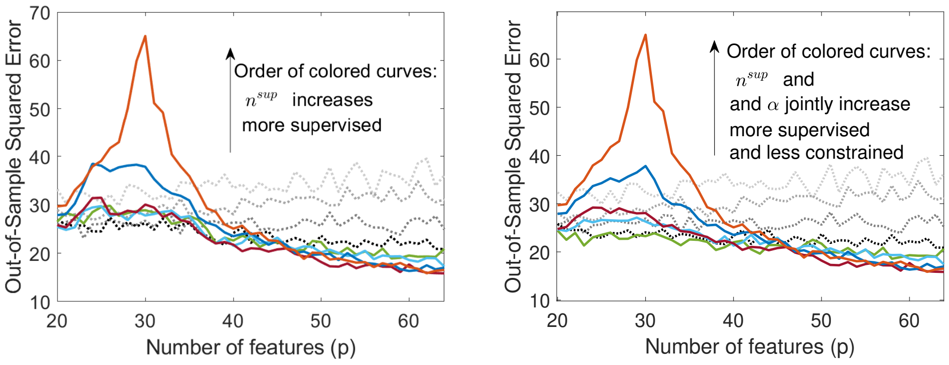
(a) (b) (c) (d)

Figure A.3. Empirical evaluation of the monotonicity metric, defined in (A.15), versus the estimated subspace dimension. All the evaluated settings correspond to $d = 128$, $m = 40$, and $n = 70$. The results in (a) and (b) are for the Hadamard case with noise levels $\sigma_\epsilon = 0.1$ and $\sigma_\epsilon = 0.5$, respectively. The results in (c) and (d) are for the Random subspace construction with noise levels $\sigma_\epsilon = 0.1$ and $\sigma_\epsilon = 0.5$, respectively. The dotted red curves obtained for a single instance of sequential increase of $\mathcal{S}_p$, and the solid blue curves are monotonicity evaluations based on the average out-of-sample errors obtained from 500 different orders of sequentially increasing $\mathcal{S}_p$.



(a) (b)

Figure B.1. The **(a) in-sample** errors $\mathcal{E}_{\text{in}}^{\text{sup}}\left(\widehat{\mathbf{U}}_m\right)$ and **(b) out-of-sample** errors $\mathcal{E}_{\text{out}}^{\text{sup}}\left(\widehat{\mathbf{U}}_m\right)$ of fully-supervised learning versus the number of parameters p. The errors correspond to a single sequential construction of $\mathcal{S}_p$. Here $d = 64$, $m = 20$, $n = 32$, and $\sigma_\epsilon = 0.5$. Each curve presents the results for a different level α of orthonormality constraints. The results here correspond to problems located along the yellow-colored border line in Fig. 1.



(a) (b)

Figure C.1. The out-of-sample errors, $\mathcal{E}_{\text{out}}^{\text{sup}}\left(\widehat{\mathbf{U}}_m\right)$ versus the number of parameters p. The errors correspond to a single experiment with a single sequential order of adding coordinates to $\mathcal{S}$. Here $d = 64$, $m = 20$ and $n = 32$. **(a) Unconstrained** settings ($\alpha \to \infty$): Each curve presents the results for a different supervision level, $n^{\text{sup}} \in \{0, 4, 8, 12, 16, 20, 24, 28, n = 32\}$. **(b)** Problems residing at the supervision-orthonormality plane along the **diagonal trajectory** connecting the standard subspace fitting and the pure regression. Each curve presents the results for a different pair of supervision and orthonormality constraint levels that jointly increase. For better visibility, the curves corresponding to $n^{\text{sup}} \in \{0, 4, 8, 12\}$ are gray dotted-lines.