Dimensionality Reduction Using the Sparse Linear Model: Supplementary Material

Ioannis Gkioulekas
Harvard SEAS
Cambridge, MA 02138
igkiou@seas.harvard.edu

Todd Zickler
Harvard SEAS
Cambridge, MA 02138
zickler@seas.harvard.edu

Abstract

In this supplementary material, we report results from additional experiments, and provide complete proofs for the mathematical statements made in the paper. Specifically, we provide larger scale versions of the visualization results on the CMU PIE dataset presented in the paper, as well as the corresponding visualization when using PCA. We also perform further experiments report on facial images using the Extended Yale B. Furthermore, we present the derivation of the optimization problems for the general dictionary and sparse linear models and their closed-form solutions, presented in Sections 3.1 and 3 of the paper, respectively. For the nonlinear extension of Section 3.2, we discuss an alternative generalization of the sparse linear model than that discussed in the paper, and rigorously derive the optimization problem and representer theorem we used the kernel case. We use the same notation as in the paper. To avoid confusion, when referring to sections, figures, and equations in the main paper, we always explicitly indicate so.

1 Visualization of CMU PIE

In Figure 1 of the paper, we show two-dimensional projections of all samples in the CMU PIE dataset, as well as identity-averaged faces across the dataset for various illuminations, poses, and expressions, produced using LPP and the proposed method. In Figures 1-2, we show larger scale versions of these projections, so that details are better visible. Furthermore, in Figure 3, we show the corresponding projections produced by using PCA.

2 Experiments on Extended Yale B Dataset

We report additional experimental results on facial images in the linear case, this time using the Extended Yale B [1] dataset (specifically, the subset used in [2]). We pre-normalize face images to be unit-length vectors, and use the same settings for dictionary learning as for the experiments on CMU PIE.

We repeat the recognition experiments on facial images of Section 4 of the paper. Due to the much smaller size of the Extended Yale B dataset when compared to CMU PIE, we only consider the cases of 50 and 40 training samples for each of the 38 individuals in the dataset. Other than that, we use the same experimental setting as with CMU PIE. In Figure 4, we show the average recognition accuracy versus the number of projections achieved by various methods, and for different numbers of training samples.

The main conclusions drawn from the experiments on CMU PIE apply here as well, with the proposed method outperforming competing methods. Note, however, that in this case performance

1
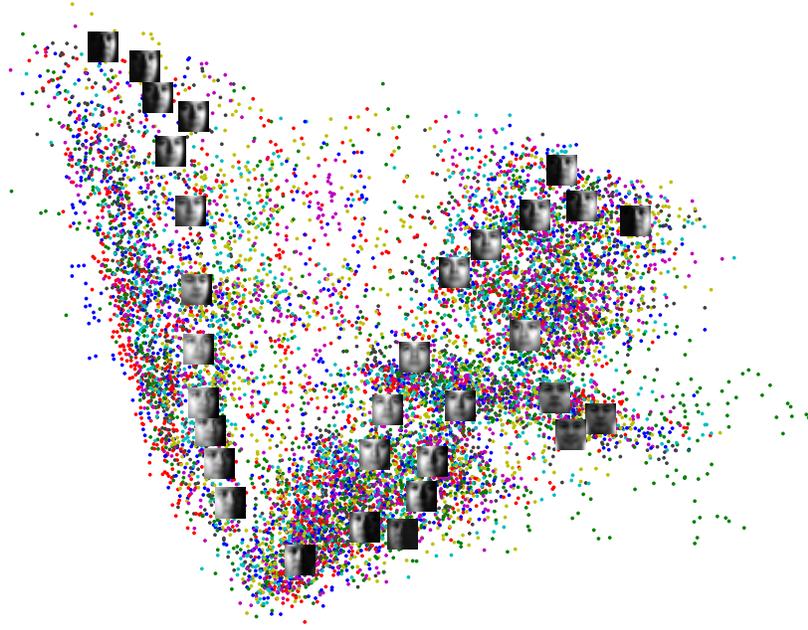
Figure 1: Two-dimensional projection of CMU PIE dataset, colored by identity, obtained using LPP. Shown at high resolution and at their respective projections are identity-averaged faces across the dataset for various illuminations, poses, and expressions. Insets show projections of samples from only two distinct identities. (Best viewed in color.)

for all methods deteriorates faster as the training samples per individual decrease, due to the much smaller dataset.

3 Derivation of optimization problem for the general and sparse linear models

We begin by deriving equation (10) of the paper for the more general dictionary model. We denote for convenience $\boldsymbol{E} = \boldsymbol{D}^T \boldsymbol{S} \boldsymbol{D} - \boldsymbol{I}$, and therefore equation (5) of the paper (which holds for the general dictionary model as well) becomes

$$\min_{\boldsymbol{L}_{M \times N}} \mathbb{E}_{\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2} \left[\left(\boldsymbol{a}_1^T \boldsymbol{E} \boldsymbol{a}_2 + \boldsymbol{\varepsilon}_1^T \boldsymbol{S} \boldsymbol{D} \boldsymbol{a}_2 + \boldsymbol{\varepsilon}_2^T \boldsymbol{S} \boldsymbol{D} \boldsymbol{a}_1 + \boldsymbol{\varepsilon}_1^T \boldsymbol{S} \boldsymbol{\varepsilon}_2 \right)^2 \right], \tag{1}$$

which, after expanding the square, can be written as

$$\begin{aligned}
\min_{\boldsymbol{L}_{M \times N}} \mathbb{E}_{\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2} \Big[& \left(\boldsymbol{a}_1^T \boldsymbol{E} \boldsymbol{a}_2 \right)^2 + \left(\boldsymbol{\varepsilon}_1^T \boldsymbol{S} \boldsymbol{D} \boldsymbol{a}_2 \right)^2 + \left(\boldsymbol{\varepsilon}_2^T \boldsymbol{S} \boldsymbol{D} \boldsymbol{a}_1 \right)^2 + \left(\boldsymbol{\varepsilon}_1^T \boldsymbol{S} \boldsymbol{\varepsilon}_2 \right)^2 \\
& + 2 \boldsymbol{a}_1^T \boldsymbol{E} \boldsymbol{a}_2 \boldsymbol{\varepsilon}_1^T \boldsymbol{S} \boldsymbol{D} \boldsymbol{a}_2 + 2 \boldsymbol{a}_1^T \boldsymbol{E} \boldsymbol{a}_2 \boldsymbol{\varepsilon}_2^T \boldsymbol{S} \boldsymbol{D} \boldsymbol{a}_1 + 2 \boldsymbol{a}_1^T \boldsymbol{E} \boldsymbol{a}_2 \boldsymbol{\varepsilon}_1^T \boldsymbol{S} \boldsymbol{\varepsilon}_2 \\
& + 2 \boldsymbol{\varepsilon}_1^T \boldsymbol{S} \boldsymbol{D} \boldsymbol{a}_2 \boldsymbol{\varepsilon}_2^T \boldsymbol{S} \boldsymbol{D} \boldsymbol{a}_1 + 2 \boldsymbol{\varepsilon}_1^T \boldsymbol{S} \boldsymbol{D} \boldsymbol{a}_2 \boldsymbol{\varepsilon}_1^T \boldsymbol{S} \boldsymbol{\varepsilon}_2 \\
& + 2 \boldsymbol{\varepsilon}_2^T \boldsymbol{S} \boldsymbol{D} \boldsymbol{a}_1 \boldsymbol{\varepsilon}_1^T \boldsymbol{S} \boldsymbol{\varepsilon}_2 \Big].
\end{aligned} \tag{2}$$

Due to the zero-mean and independence assumptions for $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2$, it is straightforward to show that the expectation of the summands corresponding to the cross-terms is equal to zero. Therefore, (2) can be reduced to

$$\begin{aligned}
\min_{\boldsymbol{L}_{M \times N}} \mathbb{E}_{\boldsymbol{a}_1, \boldsymbol{a}_2} \left[\left(\boldsymbol{a}_1^T \boldsymbol{E} \boldsymbol{a}_2 \right)^2 \right] + \mathbb{E}_{\boldsymbol{a}_2, \boldsymbol{\varepsilon}_1} \left[\left(\boldsymbol{\varepsilon}_1^T \boldsymbol{S} \boldsymbol{D} \boldsymbol{a}_2 \right)^2 \right] \\
+ \mathbb{E}_{\boldsymbol{a}_1, \boldsymbol{\varepsilon}_2} \left[\left(\boldsymbol{\varepsilon}_2^T \boldsymbol{S} \boldsymbol{D} \boldsymbol{a}_1 \right)^2 \right] + \mathbb{E}_{\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2} \left[\left(\boldsymbol{\varepsilon}_1^T \boldsymbol{S} \boldsymbol{\varepsilon}_2 \right)^2 \right].
\end{aligned} \tag{3}$$

2
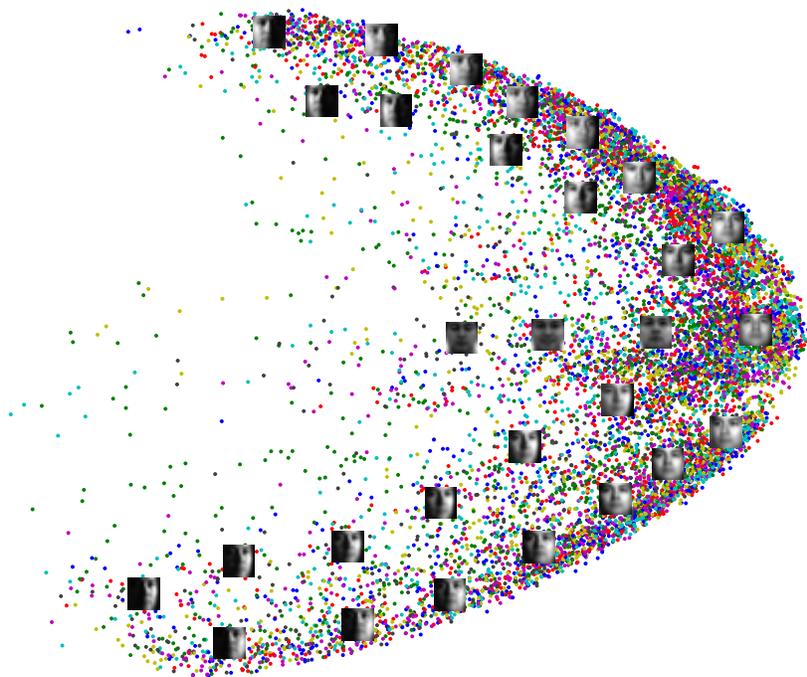
Figure 2: Two-dimensional projection of CMU PIE dataset, colored by identity, obtained using the proposed method. Shown at high resolution and at their respective projections are identity-averaged faces across the dataset for various illuminations, poses, and expressions. Insets show projections of samples from only two distinct identities. (Best viewed in color.)
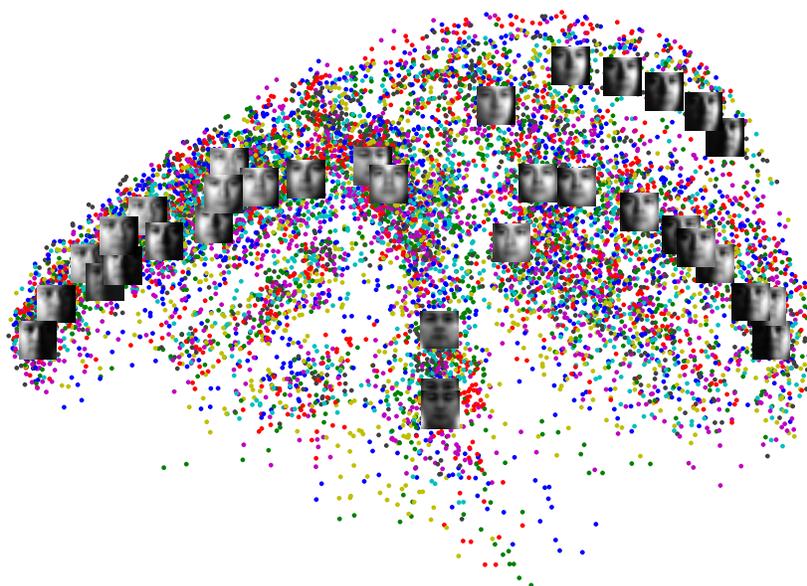
Figure 3: Two-dimensional projection of CMU PIE dataset, colored by identity, obtained using PCA. Shown at high resolution and at their respective projections are identity-averaged faces across the dataset for various illuminations, poses, and expressions. Insets show projections of samples from only two distinct identities. (Best viewed in color.)

3
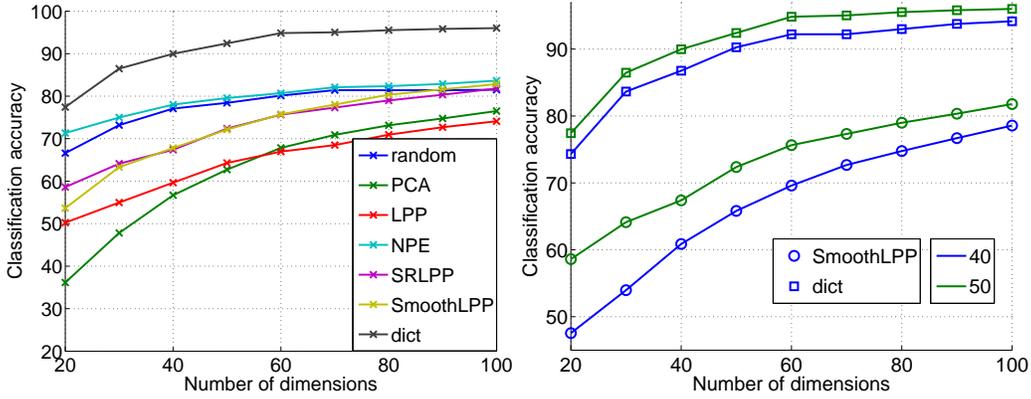
Figure 4: Results on Extended YaleB. Left: Classification accuracy of various methods, for different values of the number of projections M, using 50 training samples per individual. Right: Classification accuracy of proposed and best alternative method, for different values of the number of training samples per individual (color) and projections M.

We consider the first term in (3), $\mathbb{E}_{\boldsymbol{a}_1,\boldsymbol{a}_2}\left[\left(\boldsymbol{a}_1^T\boldsymbol{E}\boldsymbol{a}_2\right)^2\right]$. If we perform the vector-matrix-vector multiplication and expand the square, the resulting scalar is equal to a linear combination of terms of the form $a_{1i}a_{1k}a_{2j}a_{2l}$, for $i,j,k,l=1,\ldots,K$, where $\boldsymbol{a}_1=(a_{11},\ldots,a_{1K})$ and similarly for $\boldsymbol{a}_2$, and with constant coefficients that depend on the entries of the matrix $\boldsymbol{E}$. From the independence and zero-mean assumptions for the components of the vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$, the expectation of all the terms with $i \neq k$ or $j \neq l$ is equal to zero. Therefore, we have that

$$\mathbb{E}_{\boldsymbol{a}_1,\boldsymbol{a}_2}\left[\left(\boldsymbol{a}_1^T\boldsymbol{E}\boldsymbol{a}_2\right)^2\right] = \sum_{i=1}^{K}\sum_{j=1}^{K}c_{ij}a_{1i}^2a_{2j}^2, \qquad (4)$$

and by evaluating the vector-matrix-vector product analytically, it is easy to see that

$$c_{ij} = (\boldsymbol{E})_{ij}^2. \qquad (5)$$

Using the fact that, for any matrix $\boldsymbol{M}$, its Frobenius norm can be written as $\|M\|_F^2 = \sum_i\sum_j(\boldsymbol{M})_{ij}^2$, along with equations (4) and (5), we get directly that

$$\mathbb{E}_{\boldsymbol{a}_1,\boldsymbol{a}_2}\left[\left(\boldsymbol{a}_1^T\boldsymbol{E}\boldsymbol{a}_2\right)^2\right] = \left\|\boldsymbol{E}\sqrt{\boldsymbol{W}_1}\right\|_F^2, \quad (\boldsymbol{W}_1)_{ij} = \mathbb{E}\left[a_{1i}^2a_{2j}^2\right]. \qquad (6)$$

Using derivations analogous to the above, it is easy to also prove that

$$\mathbb{E}_{\boldsymbol{a}_2,\boldsymbol{\varepsilon}_1}\left[\left(\boldsymbol{\varepsilon}_1^T\boldsymbol{S}\boldsymbol{D}\boldsymbol{a}_2\right)^2\right] + \mathbb{E}_{\boldsymbol{a}_1,\boldsymbol{\varepsilon}_2}\left[\left(\boldsymbol{\varepsilon}_2^T\boldsymbol{S}\boldsymbol{D}\boldsymbol{a}_1\right)^2\right] = \left\|(\boldsymbol{S}\boldsymbol{D})\odot\sqrt{\boldsymbol{W}_2}\right\|_F^2, \qquad (7)$$

$$\mathbb{E}_{\boldsymbol{\varepsilon}_1,\boldsymbol{\varepsilon}_2}\left[\left(\boldsymbol{\varepsilon}_1^T\boldsymbol{S}\boldsymbol{\varepsilon}_2\right)^2\right] = \left\|\boldsymbol{S}\odot\sqrt{\boldsymbol{W}_3}\right\|_F^2, \qquad (8)$$

with

$$(\boldsymbol{W}_2)_{ij} = \mathbb{E}\left[\varepsilon_{1i}^2a_{2j}^2\right] + \mathbb{E}\left[\varepsilon_{2i}^2a_{1j}^2\right], \qquad (9)$$

$$(\boldsymbol{W}_3)_{ij} = \mathbb{E}\left[\varepsilon_{1i}^2\varepsilon_{2j}^2\right]. \qquad (10)$$

Combining the above, we obtain equation (10) of the paper.

In the case of the sparse linear model, from the assumption that the components of $\boldsymbol{a}$ and $\boldsymbol{\varepsilon}$ are i.i.d. Laplace and Gaussian respectively, we have for all i and j,

$$\mathbb{E}\left[a_{1i}^2a_{2j}^2\right] = \mathbb{E}\left[a_{1i}^2\right]\mathbb{E}\left[a_{2j}^2\right] = 4\tau^4, \qquad (11)$$

$$\mathbb{E}\left[\varepsilon_{1i}^2a_{2j}^2\right] = \mathbb{E}\left[\varepsilon_{1i}^2\right]\mathbb{E}\left[a_{2j}^2\right] = 2\sigma^2\tau^2, \qquad (12)$$

$$\mathbb{E}\left[a_{1i}^2\varepsilon_{2j}^2\right] = \mathbb{E}\left[a_{1i}^2\right]\mathbb{E}\left[\varepsilon_{2j}^2\right] = 2\sigma^2\tau^2, \qquad (13)$$

$$\mathbb{E}\left[\varepsilon_{1i}^2\varepsilon_{2j}^2\right] = \mathbb{E}\left[\varepsilon_{1i}^2\right]\mathbb{E}\left[\varepsilon_{2j}^2\right] = \sigma^4. \qquad (14)$$

Using these, equation (10) of the paper is simplified into equation (6) of the paper, which is the optimization problem we consider for the case of the sparse linear model.

4

4 Derivation of solution for the sparse linear model case

We introduce some notation. The singular value decomposition of L is

$$L = U\Sigma Z^T, \tag{15}$$

where U is a $M \times M$ orthogonal matrix, Σ is an invertible $M \times M$ diagonal matrix, and Z is a $N \times M$ matrix with orthonormal columns, that is $Z^T Z = I$. Similarly, the eigendecomposition of the positive semidefinite matrix DD^T is

$$DD^T = V\Lambda V^T, \tag{16}$$

where V is a $N \times N$ orthogonal matrix, and Λ is a $N \times N$ diagonal matrix.

We re-write below for convenience the objective function of the optimization problem of equation (6) of the paper,

$$\mathcal{F}(L) = 4\tau^4 \left\| D^T L^T L D - I \right\|_F^2 + 4\tau^2\sigma^2 \left\| L^T L D \right\|_F^2 + \sigma^4 \left\| L^T L \right\|_F^2. \tag{17}$$

We can obtain the stationary points of (17) by solving

$$\frac{\partial \mathcal{F}}{\partial L} = 0. \tag{18}$$

Using standard matrix differentiation tools, we can write the derivative of $\mathcal{F}$ with respect to L as

$$\frac{\partial \mathcal{F}}{\partial L} = 4L \left(4\tau^4 DD^T L^T L DD^T - 4\tau^4 DD^T + 4\sigma^2\tau^2 L^T L DD^T + \sigma^4 L^T L \right). \tag{19}$$

From (18) and (19), we firstly obtain the trivial solution $L = 0$. All other solutions of (18) correspond to the case when all of the columns of the matrix

$$4\tau^4 DD^T L^T L DD^T - 4\tau^4 DD^T + 4\sigma^2\tau^2 L^T L DD^T + \sigma^4 L^T L \tag{20}$$

belong to in the null space of L. From basic linear algebra, this condition can be equivalently expressed as

$$Z^T \left(4\tau^4 DD^T L^T L DD^T - 4\tau^4 DD^T + 4\sigma^2\tau^2 L^T L DD^T + \sigma^4 L^T L \right) = 0. \tag{21}$$

By multiplying the above from the right with $\left(DD^T \right)^{-1} Z\Sigma^{-2}$, with some basic matrix manipulation and using the orthogonality properties of the matrices involved, the above becomes

$$4\tau^4 Z^T V\Lambda V^T Z + \sigma^4 \Sigma^2 Z^T V\Lambda^{-1}V^T Z\Sigma^{-2} = 4\tau^4 \Sigma^{-2} - 4\sigma^2\tau^2 I. \tag{22}$$

We notice that the right-hand part of (18) is always symmetric. For the left-hand part of (18) to be symmetric, it is easy to check that one of the following conditions must be true: Σ must be the identify matrix; or the off-diagonal elements of the matrix $Z^t V$ must be zero. The first condition can be rejected, as it is easy to check that then it is not possible for both (18) and the requirement $Z^T Z = I$ to be true. The second condition, and the fact that V is orthogonal, imply that Z is formed from any M columns of V, that is $Z = V_M$. Then, (22) becomes

$$4\tau^4 \Lambda_M + \sigma^4 \Sigma^2 \Lambda_M^{-1}\Sigma^{-2} = 4\tau^4 \Sigma^{-2} - 4\sigma^2\tau^2 I, \tag{23}$$

where Λ_M is the diagonal matrix formed from the M eigenvalues of DD^T corresponding to the M columns of V used to form Z. This is an equation involving only diagonal matrices and can be solved trivially for its diagonal elements. Combining the above, we obtain that the solutions of (18), and therefore the stationary points of (17), are either $L = 0$, or of the form (up to an arbitrary rotation corresponding to the matrix U in (15))

$$L = \text{diag}\left(f\left(\boldsymbol{\lambda}_M \right) \right) V_M^T. \tag{24}$$

where $\boldsymbol{\lambda}_M = (\lambda_1, \ldots, \lambda_M)$ is a $M \times 1$ vector composed of any subset of M eigenvalues of the $N \times N$ matrix DD^T, V_M is the $N \times M$ matrix with the corresponding eigenvectors as columns, $f(\cdot)$ is a function applied element-wise to the vector $\boldsymbol{\lambda}_M$, equal to (as obtained by solving (23))

$$f(\lambda_i) = \sqrt{\frac{4\tau^4 \lambda_i}{\sigma^4 + 4\tau^2\sigma^2\lambda_i + 4\tau^4\lambda_i^2}}, \tag{25}$$

5

and $\operatorname{diag}\left(f\left(\boldsymbol{\lambda}_M\right)\right)$ is the $M\times M$ diagonal matrix with $f\left(\boldsymbol{\lambda}_M\right)$ as its diagonal. Substituting the above solution in (17), and using (15), (16) and the orthogonality properties of the matrices involved, it is easy to see that the objective function evaluated at the stationary points is equal to

$$\mathcal{F}\left(\boldsymbol{L}\right) = 4\tau^4 \left\| \boldsymbol{\Lambda}_M^{\frac{1}{2}} \operatorname{diag}\left(f\left(\boldsymbol{\lambda}_M\right)\right)^2 \boldsymbol{\Lambda}_M^{\frac{1}{2}} - \boldsymbol{I} \right\|_F^2$$
$$+ 4\tau^2\sigma^2 \left\| \operatorname{diag}\left(f\left(\boldsymbol{\lambda}_M\right)\right)^2 \boldsymbol{\Lambda}_M^{\frac{1}{2}} \right\|_F^2 + \sigma^4 \left\| \operatorname{diag}\left(f\left(\boldsymbol{\lambda}_M\right)\right)^2 \right\|_F^2 . \tag{26}$$

From the definition of the Frobenius norm, and some algebraic manipulation, we can rewrite the above as

$$\mathcal{F}\left(\boldsymbol{L}\right) = \sum_{i=1}^{M} h\left(\lambda_i\right), \tag{27}$$

where

$$h\left(\lambda_i\right) = \frac{4\sigma^2\tau^4\left(\sigma^2 + 4\lambda_i\tau^2\right)}{\left(\sigma^2 + 2\lambda_i\tau^2\right)^2}. \tag{28}$$

It is easy to see that $h\left(\lambda\right)$ is a strictly decreasing function of λ for all positive values of σ and τ. Consequently, the stationary point where the objective function $\mathcal{F}\left(\boldsymbol{L}\right)$ has the lowest value is obtained when the M largest eigenvalues are selected in (24), therefore arriving at the solution presented in equation (8) of the paper. We also see that for $\sigma = 0$, the objective function has the same value for all stationary points, which is the reason for the solution ambiguity we discuss in the paper with regards to the noiseless case (see equation (9) of the paper).

5 Alternative model for the kernel case

We briefly discuss here an extension of the sparse linear model to the kernel case different from the one used in Section 3.2 of the paper. We denote by $\mathcal{C}$ the subspace spanned by the atoms of dictionary $\mathcal{D}$,

$$\mathcal{C} \equiv \operatorname{span}\left\{ \tilde{d}_i, i = 1, \ldots, K \right\}. \tag{29}$$

$\mathcal{C}$ has finite dimension $d \leq K$, and thus is closed. Therefore, it has an orthogonal complement $\mathcal{C}^\perp$ and we can write $\mathcal{H} = \mathcal{C} \oplus \mathcal{C}^\perp$. We also denote by $P_\mathcal{C}$ the orthogonal projection to $\mathcal{C}$.

Denote by the set $\{\tilde{e}_i, i = 1, \ldots, d\}$ the orthonormal basis of $\mathcal{C}$, and by the set $\{\tilde{e}_i, i = d+1, \ldots\}$ its extension to the rest of the space $\mathcal{H}$. Then, we assume a probabilistic model where samples are generated from equation (12) of the paper, and under the same assumptions for $\boldsymbol{a}$ as in the paper. We also still assume that ε is a Gaussian process over $\mathbb{R}^N$ with sample patchs in $\mathcal{H}$. However, we assume that the covariance operator of ε is defined as follows,

$$\langle \tilde{e}_i, C_{\tilde{\varepsilon}} \tilde{e}_j \rangle_\mathcal{H} = \sigma^2,\ 1 \leq i, j \leq d, \tag{30}$$
$$\langle \tilde{e}_i, C_{\tilde{\varepsilon}} \tilde{e}_j \rangle_\mathcal{H} = 0,\ i > d \text{ or } j > d. \tag{31}$$

As the set $\{\tilde{e}_i, i = 1, \ldots\}$ is an orthonormal basis of $\mathcal{H}$, the above offers a full characterization of the operator $C_{\tilde{\varepsilon}}$. It is easy to construct such an operator, and therefore the Gaussian process $\tilde{\varepsilon}$, even on infinite dimensional spaces. Equivalently, the operator acts as the identity (times a constant) for the projections of signals in the subspace $\mathcal{C}$, and as the zero operator for the residuals. As a consequence, the above model allows for Gaussian noise of zero-mean and variance σ^2 along all of the dimensions of the subspace $\mathcal{C}$, but does not allow for noise along any other dimension of $\mathcal{H}$ (or alternatively, noise there has zero variance). If we denote by $\left\{\boldsymbol{b}_i \in \mathbb{R}^d, i = 1, \ldots, K\right\}$ the coordinates of the dictionary elements with respect to the subspace basis $\{\tilde{e}_i, i = 1, \ldots, d\}$, then all functions generated by the above model will also belong to $\mathcal{C}$, and their coordinates with respect to the above basis will be

$$c_j = \sum_{i=1}^{K} b_{ij} a_i + \epsilon_j,\ j = 1, \ldots, d, \tag{32}$$

where $\epsilon_j,\ j = 1, \ldots, d$ are Gaussian random variables with mean zero and variance σ^2. Then, for such functions, it is easy to see that MAP estimation of $\boldsymbol{a}$ reduces to the kernelized lasso of equation (13) of the paper.

6

We note that, for any $f \in \mathcal{H}$, the optimization problem (13) of the paper can be written equivalently as

$$\min_{\boldsymbol{a} \in \mathbb{R}^K} \frac{1}{2\sigma^2} \|f - P_{\mathcal{C}} f\|_{\mathcal{H}}^2 + \frac{1}{2\sigma^2} \|P_{\mathcal{C}} f - \mathcal{D}\boldsymbol{a}\|_{\mathcal{H}}^2 + \frac{1}{\tau} \|\boldsymbol{a}\|_1. \tag{33}$$

As the part $\frac{1}{2\sigma^2} \|f - P_{\mathcal{C}} f\|_{\mathcal{H}}^2$ does not depend on $\boldsymbol{a}$, the above is exactly equivalent to

$$\min_{\boldsymbol{a} \in \mathbb{R}^K} \frac{1}{2\sigma^2} \|P_{\mathcal{C}} f - \mathcal{D}\boldsymbol{a}\|_{\mathcal{H}}^2 + \frac{1}{\tau} \|\boldsymbol{a}\|_1, \tag{34}$$

and as $P_{\mathcal{C}} f \in \mathcal{C}$, the above optimization problem can be rewritten as MAP estimation of the co-ordinates (32) and be explained probabilistically by the model we introduced above. Therefore, the above model can be used for the projections on $\mathcal{C}$ of all functions in $\mathcal{H}$, and then kernel lasso becomes MAP estimation of their projection's coordinates in some basis. The term $f - P_{\mathcal{C}} f$ is not explained by the above model, and in fact any non-zero such component occurs only with zero probability. The fact that $\frac{1}{2\sigma^2} \|f - P_{\mathcal{C}} f\|_{\mathcal{H}}^2$ cannot be the likelihood of some distribution for noise is exactly the problem we run into when trying to extend the model we used in the paper to the infinite dimensional case. However, if one is willing to "discard" the component $f - P_{\mathcal{C}} f$ for all samples arising in practice, this alternative model can be applied to the infinite dimensional case for all signals.

6 Derivation of solution for the nonlinear case

Firstly, we derive equation (14) of the paper. We have

$$\delta p^2 = \left((\mathcal{V}\Phi x_1)^T (\mathcal{V}\Phi x_2) - \boldsymbol{a}_1^T \boldsymbol{a}_2 \right)^2 \tag{35}$$

$$= \left(\langle \mathcal{V}\Phi x_1, \mathcal{V}\Phi x_2 \rangle_{\mathbb{R}^M} - \boldsymbol{a}_1^T \boldsymbol{a}_2 \right)^2 \tag{36}$$

$$= \left(\langle \Phi x_1, \mathcal{V}^* \mathcal{V}\Phi x_2 \rangle_{\mathcal{H}} - \boldsymbol{a}_1^T \boldsymbol{a}_2 \right)^2 \tag{37}$$

$$= \left(\langle \Phi x_1, \mathcal{S}\Phi x_2 \rangle_{\mathcal{H}} - \boldsymbol{a}_1^T \boldsymbol{a}_2 \right)^2, \tag{38}$$

where we have used that $\mathcal{S} = \mathcal{V}^* \mathcal{V}$. Then, using equation (12) of the paper, the above becomes

$$\delta p^2 = \left(\langle \mathcal{D}\boldsymbol{a}_1 + \tilde{\varepsilon}_1, \mathcal{S}(\mathcal{D}\boldsymbol{a}_2 + \tilde{\varepsilon}_2) \rangle_{\mathcal{H}} - \boldsymbol{a}_1^T \boldsymbol{a}_2 \right)^2 \tag{39}$$

$$= \Big(\langle \mathcal{D}\boldsymbol{a}_1, \mathcal{S}(\mathcal{D}\boldsymbol{a}_2) \rangle_{\mathcal{H}} + \langle \tilde{\varepsilon}_1, \mathcal{S}(\mathcal{D}\boldsymbol{a}_2) \rangle_{\mathcal{H}} + \langle \mathcal{D}\boldsymbol{a}_2, \mathcal{S}\tilde{\varepsilon}_2 \rangle_{\mathcal{H}}$$

$$+ \langle \tilde{\varepsilon}_1, \mathcal{S}\tilde{\varepsilon}_2 \rangle_{\mathcal{H}} - \boldsymbol{a}_1^T \boldsymbol{a}_2 \Big)^2 \tag{40}$$

$$= \Big(\sum_{i=1}^{K} \sum_{j=1}^{K} a_{1i} a_{2j} \left(\left\langle \tilde{d}_i, \mathcal{S}\tilde{d}_j \right\rangle_{\mathcal{H}} - \delta_{ij} \right) + \sum_{j=1}^{K} a_{2j} \left\langle \tilde{\varepsilon}_1, \mathcal{S}\tilde{d}_j \right\rangle_{\mathcal{H}}$$

$$+ \sum_{i=1}^{K} a_{1i} \left\langle \tilde{d}_i, \mathcal{S}\tilde{\varepsilon}_2 \right\rangle_{\mathcal{H}} + \langle \tilde{\varepsilon}_1, \mathcal{S}\tilde{\varepsilon}_2 \rangle_{\mathcal{H}} - \boldsymbol{a}_1^T \boldsymbol{a}_2 \Big)^2. \tag{41}$$

where δ_{ij} is the Kronecker delta. From the equivalence of Gaussian processes with sample paths on Hilbert spaces and Gaussian measures [3,4], we have that

$$\mathbb{E} \left[\left\langle \tilde{\varepsilon}_1, \tilde{f} \right\rangle_{\mathcal{H}} \right] = 0, \forall \tilde{f} \in \mathcal{H}, \tag{42}$$

and similarly for $\tilde{\varepsilon}_2$. Furthermore, from the properties of the covariance operator $C_{\tilde{\varepsilon}_1} = C_{\tilde{\varepsilon}_2} = \sigma^2 \mathcal{I}$, we have that

$$\mathbb{E} \left[\left\langle \tilde{\varepsilon}_1, \tilde{f} \right\rangle_{\mathcal{H}} \langle \tilde{\varepsilon}_1, \tilde{g} \rangle_{\mathcal{H}} \right] = \left\langle C_{\tilde{\varepsilon}_1} \tilde{f}, \tilde{g} \right\rangle_{\mathcal{H}} = \sigma^2 \left\langle \tilde{f}, \tilde{g} \right\rangle_{\mathcal{H}}, \forall \tilde{f}, \tilde{g} \in \mathcal{H}, \tag{43}$$

and similarly for $\tilde{\varepsilon}_2$. If in (41) we expand the square and take the expectation, then using (42), (43), and an analysis similar to that presented before for deriving equation (10) of the paper, we arrive at

$$\mathbb{E}\left[\delta p^2\right] = 4\tau^4 \sum_{i=1}^{K} \sum_{i=1}^{K} \left(\left\langle \tilde{d}_i, \mathcal{S}\tilde{d}_j \right\rangle_{\mathcal{H}} - \delta_{ij} \right)^2 + 4\tau^2 \sigma^2 \sum_{i=1}^{K} \left\langle \mathcal{S}\tilde{d}_i, \mathcal{S}\tilde{d}_i \right\rangle_{\mathcal{H}} + \|\mathcal{S}\|_{HS}^2, \tag{44}$$

7

which corresponds to equation (14) in the paper.

We now need to prove that the minimizer of (44) over the set of compact, positive semi-definite, self-adjoint, and linear operators $\mathcal{S}$ of rank M has the form of equation (15) of the paper. For this purpose, we use and extend the representer theorem presented in [5]. The first and third term of the objective function (44) along with the rank constraint correspond to the conditions of Theorem 3 of [5]. Here, we have additionally the constraint that $\mathcal{S}$ be self-adjoint, and also the second term of the objective function that violates the conditions of that theorem. Extending the theorem to the case when $\mathcal{S}$ is also required to be positive semi-definite is straightforward. In order to handle terms of the form $\left\langle \mathcal{S}\tilde{d}_i, \mathcal{S}\tilde{d}_i \right\rangle_{\mathcal{H}}$, for $i = 1, \ldots, K$, note that due to the Hilbert-Schmidt norm term in the objective function (44), its minimizer $\mathcal{S}$ has finite Hilbert-Schmidt norm and thus is a Hilbert-Schmidt operator. Therefore, we can consider its decomposition as $\mathcal{S} = \mathcal{S}^S + \mathcal{S}^\perp$, where $\mathcal{S}^S$ is the projection of $\mathcal{S}$ onto the linear span of $\left\{ \tilde{d}_i \otimes \tilde{d}_j, \, i, j = 1, \ldots, K \right\}$,

$$\mathcal{S}^S = \sum_{i=1}^{K} \sum_{j=1}^{K} \gamma_{ij} \tilde{d}_i \otimes \tilde{d}_j, \tag{45}$$

and $\mathcal{S}^\perp$ is orthogonal to each element of the above set,

$$\left\langle \tilde{d}_i, \mathcal{S}^\perp \tilde{d}_j \right\rangle_{\mathcal{H}} = 0, \, i, j = 1, \ldots, K. \tag{46}$$

As $\mathcal{S}$ is compact and self-adjoint, $\mathcal{S}^S$ and $\mathcal{S}^\perp$ are also compact and self-adjoint. Therefore, from the spectral theorem, $\mathcal{S}^\perp$ admits an eigendecomposition

$$\mathcal{S}^\perp = \sum_k \lambda_k \tilde{v}_k \otimes \tilde{v}_k, \tag{47}$$

where $\tilde{v}_k \in \mathcal{H}$ is an orthonormal set and λ_k are positive. Combining (47) with (46) for $i = j$, $i = 1, \ldots, K$, we obtain

$$\left\langle \tilde{v}_k, \tilde{d} \right\rangle_{\mathcal{H}} = 0, \, \forall i = 1, \ldots, K, \, \forall k \text{ such that } \lambda_k \neq 0. \tag{48}$$

For each $i = 1, \ldots, K$, we have

$$\left\langle \mathcal{S}\tilde{d}_i, \mathcal{S}\tilde{d}_i \right\rangle_{\mathcal{H}} = \left\langle \tilde{d}_i, \mathcal{S}\mathcal{S}\tilde{d}_i \right\rangle_{\mathcal{H}} \tag{49}$$

$$= \left\langle \tilde{d}_i, \left(\mathcal{S}^S + \mathcal{S}^\perp \right) \left(\mathcal{S}^S + \mathcal{S}^\perp \right) \tilde{d}_i \right\rangle_{\mathcal{H}} \tag{50}$$

$$= \left\langle \tilde{d}_i, \mathcal{S}^S \mathcal{S}^S \tilde{d}_i \right\rangle_{\mathcal{H}} + \left\langle \tilde{d}_i, \mathcal{S}^S \mathcal{S}^\perp \tilde{d}_i \right\rangle_{\mathcal{H}}$$
$$+ \left\langle \tilde{d}_i, \mathcal{S}^\perp \mathcal{S}^S \tilde{d}_i \right\rangle_{\mathcal{H}} + \left\langle \tilde{d}_i, \mathcal{S}^\perp \mathcal{S}^\perp \tilde{d}_i \right\rangle_{\mathcal{H}}. \tag{51}$$

where we have used linearity and self-adjointness. However, using (45), (47) and (48), it is easy to see that the terms involving $\mathcal{S}^S \mathcal{S}^\perp$, $\mathcal{S}^\perp \mathcal{S}^S$ and $\mathcal{S}^\perp \mathcal{S}^\perp$ are equal to zero. Therefore we conclude that

$$\left\langle \mathcal{S}\tilde{d}_i, \mathcal{S}\tilde{d}_i \right\rangle_{\mathcal{H}} = \left\langle \mathcal{S}^S \tilde{d}_i, \mathcal{S}^S \tilde{d}_i \right\rangle_{\mathcal{H}}. \tag{52}$$

From (52) and equation (20) of the proof of Theorem 3 of [5], and using the full rank assumption for $\mathbf{K}_{\mathcal{DD}}$, we can deduce that the minimizer of (44) can be written in the form

$$\mathcal{S} = \sum_{i=1}^{K} \sum_{j=1}^{K} \gamma_{ij} \tilde{d}_i \otimes \tilde{d}_j. \tag{53}$$

The matrix $\boldsymbol{\gamma}$ formed by the coefficients γ_{ij} is related to the matrix $\boldsymbol{\alpha}$ of Theorem 3 of [5] through the equation (again assuming that $\mathbf{K}_{\mathcal{DD}}$ has full rank)

$$\boldsymbol{\alpha} = \mathbf{K}_{\mathcal{DD}}^{\frac{1}{2}} \boldsymbol{\gamma} \mathbf{K}_{\mathcal{DD}}^{\frac{1}{2}} \tag{54}$$

as proved in the proof of Theorem 3 of [5] (see fifth equation in page 824) and applied to our case where the set of elements in the optimizer is $\left\{ \tilde{d}_i \otimes \tilde{d}_j, \, i, j = 1, \ldots, K \right\}$. The eigenvalues of $\boldsymbol{\alpha}$ are the same of those of $\mathcal{S}$, as shown in the proof of Theorem 3 of [5], and therefore $\boldsymbol{\alpha}$ is positive semi-definite and of rank M. Under the assumption that the matrix $\mathbf{K}_{\mathcal{DD}}$ is full rank, we deduce that $\boldsymbol{\gamma}$ is also positive-semidefinite and of rank M. Combining this with (53), we conclude that the minimizer of (44) has the form of equation (15) of the paper.

8

References

[1] K.C. Lee, J. Ho, and D.J. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *PAMI*, 2005.

[2] D. Cai, X. He, Y. Hu, J. Han, and T. Huang. Learning a spatially smooth subspace for face recognition. *CVPR*, 2007.

[3] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer, 2004.

[4] V.I. Bogachev. *Gaussian measures*. AMS, 1998.

[5] J. Abernethy, F. Bach, T. Evgeniou, and J.P. Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. *JMLR*, 2009.

9