MERL – A MITSUBISHI ELECTRIC RESEARCH LABORATORY
http://www.merl.com

Subspace mappings for image sequences

Matthew Brand

TR-2002-25 May 2002

Abstract

We consider the use of low-dimensional linear subspace models to infer one high-dimensional signal from another, for example, predicting an image sequence from a related image sequence. In the memoryless case the subspaces are found by rank-constrained division, and inference is an inexpensive sequence of projections. In the finite-memory case, the subspaces form a linear dynamical system that is identified via factorization, and inference is Kalman filtering. In both cases we give novel closed-form solutions for all parameters, with optimality properties for truncated subspaces. Our factorization is related to the subspace methods that revolutionized stochastic system identification methods in the last decade, but we offer tight finite-data approximations and direct estimates of the system parameters without explicit computation of the subspace. Applications are made to view-mapping and synthesis of video textures.

First circulated summer 2001.

Proceedings, Statistical Methods in Video Processing, SMVP, Copenhagen 2002.

Subspace mappings for image sequences

Matthew Brand
Mitsubishi Electric Research Labs, Cambridge, MA 02139 USA

Abstract

We consider the use of low-dimensional linear subspace models to infer one high-dimensional signal from another, for example, predicting an image sequence from a related image sequence. In the memoryless case the subspaces are found by rank-constrained division, and inference is an inexpensive sequence of projections. In the finite-memory case, the subspaces form a linear dynamical system that is identified via factorization, and inference is Kalman filtering. In both cases we give novel closed-form solutions for all parameters, with optimality properties for truncated subspaces. Our factorization is related to the subspace methods [8, 1] that revolutionized stochastic system identification methods in the last decade, but we offer tight finite-data approximations and direct estimates of the system parameters without explicit computation of the subspace. Applications are made to view-mapping and controlled synthesis of video textures. We demonstrate both analytically and empirically that our factorizations provide more accurate reconstructions of estimation data and predictions of held-out test-data.

1 Introduction

A common problem in film and video post-production is the generation of natural-looking texture sequences for texture-mapping dynamic objects. We consider three scenarios for generating such textures: 1] High-dimensional input, for example, mapping a scene with nonrigid motion to another viewpoint (without geometric information). 2] Low-dimensional input, for example, changing the texture map of a face as a function of geometry to simulate BRDF changes that accompany skin deformations. 3] No inputs at all, for example, synthesizing video of dynamic scenes of hard-to-model phenomena such as turbulence. We treat these all as *linear* signal-mapping problems with gaussian noise, for which simple probabilistic models are remarkably effective but optimal (or numerically tractable) parameter estimation is still an open problem.

2 Rank-constrained division

Consider two datasets $\mathbf{Y}_{p\times n}$ and $\mathbf{Z}_{q\times n}$ that are high-dimensional with $n \ll \min(p,q)$. For example, each column $\mathbf{y} \in \mathbf{Y}$ and $\mathbf{z} \in \mathbf{Z}$ could be a vectorized image, with p and q being $O(10^{5+})$. We would like to estimate an invertible linear mapping $\mathbf{M}$ from one image set to the other, maximizing $p(\mathbf{z}|\mathbf{y}) = \mathcal{N}(\mathbf{z}|\mathbf{M}\mathbf{y}, \mathbf{M}\Sigma_{\mathbf{y}}\mathbf{M}^\top)$ and $p(\mathbf{y}|\mathbf{z}) = \mathcal{N}(\mathbf{y}|\mathbf{M}^{-1}\mathbf{y}, \mathbf{M}^{-1}\Sigma_{\mathbf{y}}\mathbf{M}^{-\top})$. Under the assumption of white gaussian noise in $\mathbf{Y}$ and $\mathbf{Z}$, finding $\mathbf{M}$ becomes a least-squares problem. Due to the large size of images, the obvious solution $\mathbf{z} \leftarrow \mathbf{M}\mathbf{y};\ \mathbf{M} = \mathbf{Z}/\mathbf{Y}$ is underconstrained and computationally impractical ($O(npq)$ time, $O(pq)$ space). This motivates a reduced-rank approach, to suppress measurement noise and improve numerical conditioning. For example, one might eigen-code the two datasets and estimate a matrix $\mathbf{M}$ that maps between the eigen-codes. This is suboptimal because eigen-codes optimize within-set reconstruction, not between-set reconstruction. This has been approached as a (non-concave) gradient ascent problem [4], but if the mapping is to be invertible, there is an optimal, closed-form solution that maximizes the *joint* probability of both datasets. This is accomplished by eigen-coding $[\mathbf{Y}\atop\mathbf{Z}]$, then using QR-decompositions to extract orthogonal subspaces from the rank-r (truncated) joint eigen-coding:

$$[\mathbf{U}_Y\atop\mathbf{U}_Z]\mathbf{S}\mathbf{V}^\top \overset{\text{SVD}_r}{\longleftarrow} [\mathbf{Y}\atop\mathbf{Z}] \tag{1}$$

$$\mathbf{Q}_Y\mathbf{R}_Y \overset{\text{QR}}{\longleftarrow} \mathbf{U}_Y \tag{2}$$

$$\mathbf{Q}_Z\mathbf{R}_Z \overset{\text{QR}}{\longleftarrow} \mathbf{U}_Z \tag{3}$$

$$\mathbf{F} \leftarrow \mathbf{R}_Z/\mathbf{R}_Y \tag{4}$$

Here the QR-decompositions are motivated by the fact that $\mathbf{U}_Z$, $\mathbf{U}_Y$ are *not* orthogonal. The optimal rank-r approximation to $\mathbf{Z}/\mathbf{Y}$ is

$$\mathbf{M} = \mathbf{Q}_Z\mathbf{F}\mathbf{Q}_Y^\top = \mathbf{U}_Z\mathbf{U}_Y^{-1} = \mathbf{U}_Z\mathbf{S}\mathbf{V}^\top\mathbf{V}\mathbf{S}^{-1}\mathbf{U}_Y^{-1}. \tag{5}$$

Since the SVD in equation 1 gives the optimal variance-preserving rank-r approximation of $\mathbf{Y}$ and $\mathbf{Z}$ together, and equations 2-4 are exact, equation 5 is the optimal invertible[1] mapping, in the sense that $\mathbf{M}$ minimizes both $||\mathbf{Z} - \mathbf{Q}_Z\mathbf{F}\mathbf{Q}_Y^\top\mathbf{Y}||_F$ and $||\mathbf{Y} - \mathbf{Q}_Y\mathbf{F}^{-1}\mathbf{Q}_Z^\top\mathbf{Z}||_F$. For asymmetric error, one scales $\mathbf{Z}$ prior to the SVD and inversely scales $\mathbf{F}$ after. $\mathbf{Q}_Z$ and $\mathbf{Q}_Y$ are orthogonal subspaces of $\mathbf{Z}$ and $\mathbf{Y}$, respectively, and upper-triangular $\mathbf{F}$ maps from projections onto one subspace to projections onto the other. Therefore the actual mapping is computationally inexpensive ($O(r(p+q+r)) \ll O(pq)$) if one orders operations: $\mathbf{z} \leftarrow \mathbf{Q}_Z(\mathbf{F}(\mathbf{Q}_Y^\top\mathbf{y}))$.

Application to view mapping: We recorded a rotating and clenching hand from an ultrawide-baseline stereo camera, obtaining two short synchronized sequences of the hand viewed from front

[1] Strictly speaking, $\mathbf{F}$ is merely pseudo-invertible iff rank($\mathbf{U}_Y$) $< r$ or rank($\mathbf{U}_Z$) $< r$. We assume rank($[\mathbf{Y}\atop\mathbf{Z}]$)$\geq$max(rank($\mathbf{Y}$),rank($\mathbf{Z}$)) and min(rank($\mathbf{Y}$),rank($\mathbf{Z}$))$\geq r$. Under such assumptions, it can be shown that for high dimensional data, the pathological case has vanishing probability.

Figure 1: Reconstructing the front view of an articulating hand from a side view, using held-out test image-pairs. All images have been identically contrast-enhanced to highlight the hand. 1ST ROW: Test-set source images, side view. 2ND ROW: Reconstructions made via least-squares mapping between dual 20-dimensional eigen-codings of the source and target training image sets can be quite blurry. **3RD ROW**: Predictions made by a **20-dimensional rank-constrained division** are sharp and virtually indistinguishable from the best possible reconstruction using using the entire training set (4TH ROW). LAST ROW: Test-set target images.

and side. The goal is to reconstruct one view (the target) from the other (the source). The sequences were split evenly into train and test image sets, each containing 200 image-pairs. We compared a 20-dimensional mapping obtained via rank-constrained division with 1] a least-squares mapping between 20-dimensional eigen-codings of the training-set source and target images, and 2] the best possible reconstruction of each test-set target image as a linear combination of all the training-set target images (which gives the lowest reconstruction error one can attain from any linear mapping derived from the training set). Figure 1 shows visual results. Figure 2LEFT makes a numerical comparison of the reconstruction errors, confirming that the rank-constrained division makes much more accurate predictions than a mapping between eigen-codes. In fact its mean-squared error is quite close to the lower bound.

3 Linear Dynamical Systems

When the target system has some memory (internal dynamics) that must be accounted for in the mapping or prediction, the appropriate linear model is a time-invariant linear dynamical system (LDS), also known as a auto-regressive moving average (ARMA) or a Kalman filter (KF). An LDS is defined by the difference equations

$$\mathbf{x}(t+1) \leftarrow \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{e}(t) + \gamma(t) \tag{6}$$

$$\mathbf{y}(t) \leftarrow \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{e}(t) + \psi(t) \tag{7}$$

where the noise variables are normally distributed as $\gamma \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_\gamma)$ and $\psi \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_\psi)$. In matrix form:

$$\overline{\mathbf{X}} = \mathbf{A}\underline{\mathbf{X}} + \mathbf{B}\underline{\mathbf{E}} + \Gamma \tag{8}$$

$$\mathbf{Y} = \mathbf{C}\mathbf{X} + \mathbf{D}\mathbf{E} + \Psi \tag{9}$$

where $\underline{\mathbf{X}} \doteq \mathbf{X}_{(:,1:T-1)} = [\mathbf{x}(1), \mathbf{x}(2), ..., \mathbf{x}(T-1)]$ and $\overline{\mathbf{X}} \doteq \mathbf{X}_{(:,2:T)} = [\mathbf{x}(2), \mathbf{x}(3), ..., \mathbf{x}(T)]$. We will use the under/overline convention in the paper to denote "before" and "after" data, or equivalently, "past" and "future." Our goal is to solve for the system matrix $\mathbf{A}$, output matrix $\mathbf{C}$ and noise matrices Ψ and Γ given a record of outputs $\mathbf{Y}$. It is assumed that the outputs exhibit wide-sense stationarity. When an input signal $\mathbf{E}$ is given, we will also solve for input matrix $\mathbf{B}$ and feed-through matrix $\mathbf{D}$. It is assumed that $\text{rank}(\mathbf{E}) < T$, otherwise there is no need for hidden state $\mathbf{X}$ since $\mathbf{Y}$ could be reconstructed perfectly from $\mathbf{Y} = \mathbf{D}\mathbf{E}$ using equations 1–5 to solve for $\mathbf{D}$. Preferably, $\text{rank}(\mathbf{E}) \ll T$.

There is a large literature of subspace methods for LDS system identification surveyed recently in [1]; these have recently been adapted to high-dimensional output-only problems in [6]. Here we introduce a new but related solution for both inputs and outputs, advancing replacing the prior art's weak asymptotic properties to stronger finite-data guarantees of optimality. An additional appeal of our approach is that our factorization (equations 19-23 below) has a clear and simple derivation.

Without loss of generality, we assume that the data is translated to the origin: $\langle \mathbf{y}(t) \rangle_t = 0$. We combine equations 8 and 9 to exploit the problem's temporal shift-invariance, obtaining

$$\overline{\mathbf{Y}} = \mathbf{C}\mathbf{A}\underline{\mathbf{X}} + \mathbf{C}\mathbf{B}\underline{\mathbf{E}} + \mathbf{D}\overline{\mathbf{E}} \tag{10}$$

$$= \mathbf{C}\mathbf{A}\mathbf{C}^\dagger(\underline{\mathbf{Y}} - \mathbf{D}\underline{\mathbf{E}}) + \mathbf{C}\mathbf{B}\underline{\mathbf{E}} + \mathbf{D}\overline{\mathbf{E}} \tag{11}$$

where † denotes pseudo-inverse. Rearranging into matrices,

$$\begin{bmatrix} \mathbf{I} & -\mathbf{D} \end{bmatrix} \begin{bmatrix} \overline{\mathbf{Y}} \\ \overline{\mathbf{E}} \end{bmatrix} = \begin{bmatrix} \mathbf{C}\mathbf{A}\mathbf{C}^\dagger & \mathbf{C}(\mathbf{B} - \mathbf{A}\mathbf{C}^\dagger\mathbf{D}) \end{bmatrix} \begin{bmatrix} \underline{\mathbf{Y}} \\ \underline{\mathbf{E}} \end{bmatrix}. \tag{12}$$

We decouple the constraints on $\mathbf{A}, \mathbf{C}$ from those on $\mathbf{B}, \mathbf{D}$ by projecting the problem onto the subspace orthogonal to that of the inputs. To do so, we QR-decompose $\mathbf{Q}_{\overline{\mathbf{E}}}\mathbf{R}_{\overline{\mathbf{E}}} \overset{\text{QR}}{\longleftarrow} \overline{\mathbf{E}}^\top$ into orthogonal $\mathbf{Q}_{\overline{\mathbf{E}}}$ and upper-triangular $\mathbf{R}_{\overline{\mathbf{E}}}$, (similarly $\mathbf{Q}_{\underline{\mathbf{E}}}\mathbf{R}_{\underline{\mathbf{E}}} \overset{\text{QR}}{\longleftarrow} \underline{\mathbf{E}}^\top$) and post-multiply both sides of equation 12 by $\mathbf{E}^\perp \doteq (\mathbf{I} - \mathbf{Q}_{\overline{\mathbf{E}}}\mathbf{Q}_{\overline{\mathbf{E}}}^\top)(\mathbf{I} - \mathbf{Q}_{\underline{\mathbf{E}}}\mathbf{Q}_{\underline{\mathbf{E}}}^\top)$ to obtain

$$\left(\overline{\mathbf{Y}}\mathbf{E}^\perp\right) = \mathbf{C}\mathbf{A}\mathbf{C}^\dagger\left(\underline{\mathbf{Y}}\mathbf{E}^\perp\right) \tag{13}$$

$\mathbf{Y}\mathbf{E}^\perp$ is the component of the output signal that cannot be directly regressed onto the input signal, and therefore must be explained in terms of hidden state. By making this our objective function, we are essentially optimizing the fidelity of the model w.r.t. frame-to-frame dynamics—the ability of the system to produce realistic output sequences with *or without* inputs. This is good for synthesizing video textures but not necessarily optimal for process control, where there is more of an emphasis on modeling the time-delayed response of a system to inputs.

We will concentrate on the case of high-dimensional data, e.g., when the right and left data matrices have d rows and $T-1$ columns, with $d \gg T$. In such cases one uses a lossless dimensionality reduction by computing an orthogonal basis $\mathbf{L}_{d\times T-1}$ ($T-1$ because the data is zero-mean) and an encoding $\mathbf{Z} \doteq \mathbf{L}^\top\mathbf{Y}$. In particular, we assume that $\mathbf{L}$ are the eigenvectors of the data's covariance matrix, and the columns of $\mathbf{Z}$ are the "eigen-codes" of each sample, with the rows of $\mathbf{Z}$ sorted by decreasing variance. $\mathbf{L}$ could be computed as the left singular vectors of the data, and $\mathbf{Z}$ could be computed as the product of the singular values and the right singular vectors. For low-dimensional data, $\mathbf{L} = \mathbf{I}$ and $\mathbf{Z} = \mathbf{Y}$.

$\mathbf{E}^\perp$ need not be explicitly computed to obtain $\underline{\mathbf{Z}}\mathbf{E}^\perp$ and $\overline{\mathbf{Z}}\mathbf{E}^\perp$. Instead, the projections of equation 13 are efficiently calculated in QR-decomposed form as $\underline{\mathbf{Z}}\mathbf{E}^\perp = (\underline{\mathbf{Q}}[\underline{\mathbf{R}}, \underline{\mathbf{M}}])^\top$ from

$$\begin{bmatrix}\mathbf{Q}_{\underline{\mathbf{E}}} & \mathbf{Q}_{\overline{\underline{\mathbf{E}}}} & \underline{\mathbf{Q}}\end{bmatrix}\begin{bmatrix}\mathbf{R}_{\underline{\mathbf{E}}} & \mathbf{M}_a & \mathbf{M}_b & \mathbf{M}_d \\ 0 & \mathbf{R}_{\overline{\underline{\mathbf{E}}}} & \mathbf{M}_c & \mathbf{M}_e \\ 0 & 0 & \underline{\mathbf{R}} & \underline{\mathbf{M}}\end{bmatrix} \overset{\text{QR}}{\longleftarrow} \begin{bmatrix}\underline{\mathbf{E}} \\ \overline{\underline{\mathbf{E}}} \\ \underline{\mathbf{Z}}\end{bmatrix}^\top, \tag{14}$$

and similarly $\overline{\mathbf{Z}}\mathbf{E}^\perp = (\overline{\mathbf{Q}}[\overline{\mathbf{R}}, \overline{\mathbf{M}}])^\top$. In the case where r.h.s. equation 14 has more columns than rows, $\underline{\mathbf{M}}$ and $\overline{\mathbf{M}}$ will be non-empty matrices containing components of $\underline{\mathbf{Z}}\mathbf{E}^\perp$ and $\overline{\mathbf{Z}}\mathbf{E}^\perp$ that are linearly dependent on the columns of $\underline{\mathbf{R}}, \overline{\mathbf{R}}$ [2]. This gives a first-order factorization; higher-order solutions are available by putting extra rows containing $\underline{\underline{\mathbf{E}}}, \overline{\underline{\mathbf{E}}}, \underline{\underline{\mathbf{Z}}}$ etc., into equation 14.

We now make a change of variable to work in the eigen-coded data space, making the constraint equation

$$\{\overline{\mathbf{Z}}\mathbf{E}^\perp\} = \mathbf{C}'\mathbf{A}\mathbf{C}'^\dagger\{\underline{\mathbf{Z}}\mathbf{E}^\perp\}, \tag{15}$$

where $\mathbf{C}' \doteq \mathbf{L}^\top\mathbf{C}$ and we brace $\{\underline{\mathbf{Z}}\mathbf{E}^\perp\}$ to indicate that we are only working with the nonredundant columns $\underline{\mathbf{Q}}\underline{\mathbf{R}} \subset \underline{\mathbf{Q}}[\underline{\mathbf{R}}, \underline{\mathbf{M}}] = (\underline{\mathbf{Z}}\mathbf{E}^\perp)^\top$. Equation 15 has an obvious solution with $\mathbf{A}$ in diagonal form via the eigen-decomposition: $\{\overline{\mathbf{Z}}\mathbf{E}^\perp\}/\{\underline{\mathbf{Z}}\mathbf{E}^\perp\} = \mathbf{C}'\mathbf{A}\mathbf{C}'^\dagger$. However, the eigen-solution has some undesirable properties: The estimated LDS is typically complex-valued and unbalanced in the sense of Moore [5]: estimated parameters of unbalanced systems are sensitive to small perturbations in the data. We seek an LDS that is real-valued, balanced, and has the SVD'S optimal variance-preserving property for any truncation of its dimensionality.

We now develop a well-behaved factorization by setting up a pair of orthogonal Procrustes relating the subspaces of "before" and "after" observations.

3.1 Factorization

We begin with the QR-decompositions $\overline{\mathbf{Q}\mathbf{R}} \overset{\text{QR}}{\longleftarrow} \{\overline{\mathbf{Z}}\mathbf{E}^\perp\}^\top$ and similarly $\underline{\mathbf{Q}\mathbf{R}} \overset{\text{QR}}{\longleftarrow} \{\underline{\mathbf{Z}}\mathbf{E}^\perp\}^\top$ calculated above. Each column of $\mathbf{Q}$ is a time-series of the normalized amplitude of one of the signal's orthogonal modes of variation. This essentially decomposes the signal into a set of oscillators whose couplings determine the dynamics of the signal (see figure 3). Our goal is to model these couplings by finding a system matrix $\mathbf{A}$ that relates the past $\underline{\mathbf{Q}}$ and the future $\overline{\mathbf{Q}}$. The main complication is that $\underline{\mathbf{Q}}$ and $\overline{\mathbf{Q}}$ are not mutually consistent models of the past and future because they decompose the signal with regard to inconsistent $\underline{\mathbf{R}} \neq \overline{\mathbf{R}}$.

To minimize this inconsistency, let $\mathbf{J}$ be an orthogonal matrix that minimizes $\|\mathbf{J}^\top\overline{\mathbf{R}} - \mathbf{J}\underline{\mathbf{R}}\|_F$ (Frobenius norm). Since rotations preserve the Frobenius norm, $\mathbf{J}^2$ is the solution to the orthogonal Procrustes problem $\mathbf{J} = \arg\min_{\mathbf{J}} \|\mathbf{J}(\mathbf{J}^\top\overline{\mathbf{R}} - \mathbf{J}\underline{\mathbf{R}})\|_F = \arg\min_{\mathbf{J}} \|\overline{\mathbf{R}} - \mathbf{J}^2\underline{\mathbf{R}}\|_F$, obtainable via SVD:

$$\mathbf{U}_\mathbf{J}\mathbf{S}_\mathbf{J}\mathbf{V}_\mathbf{J}^\top \overset{\text{SVD}}{\longleftarrow} \overline{\mathbf{R}}\underline{\mathbf{R}}^\top \tag{16}$$

$$\mathbf{J} \leftarrow (\mathbf{U}_\mathbf{J}\mathbf{V}_\mathbf{J}^\top)^{1/2} \tag{17}$$

We use $\mathbf{J}$ to define revised QR-decompositions, $(\overline{\mathbf{Q}}\mathbf{J})(\mathbf{J}^\top\overline{\mathbf{R}}) = \{\overline{\mathbf{Z}}\mathbf{E}^\perp\}^\top$, $(\underline{\mathbf{Q}}\mathbf{J}^\top)(\mathbf{J}\underline{\mathbf{R}}) = \{\underline{\mathbf{Z}}\mathbf{E}^\perp\}^\top$. Define $\underline{\mathbf{W}} \doteq (\underline{\mathbf{Q}}\mathbf{J}^\top)$, $\overline{\mathbf{W}} \doteq (\overline{\mathbf{Q}}\mathbf{J})$, $\mathbf{R}_\mathbf{J} \doteq (\mathbf{J}^\top\overline{\mathbf{R}} + \mathbf{J}\underline{\mathbf{R}})/2$, and $\mathbf{M}_\mathbf{J} \doteq (\mathbf{J}^\top\overline{\mathbf{M}} + \mathbf{J}\underline{\mathbf{M}})/2$. $\underline{\mathbf{W}}$ and $\overline{\mathbf{W}}$ are subspaces of the past and future that are *maximally consistent* in that they give minimum Frobenius-error reconstructions of the data with regard to $\mathbf{R}_\mathbf{J}$. (Perfect consistency is impossible unless the data spans an integral cycle of a periodic signal.)

To factor $\mathbf{A}$ and $\mathbf{C}$ we set up an orthogonal Procrustes problem seeking the rotation that takes the subspace of the past into the subspace of the future. This begins with the SVD

$$\mathbf{U}\mathbf{S}\mathbf{V}^\top \overset{\text{SVD}_r}{\longleftarrow} \overline{\mathbf{W}}^\top\underline{\mathbf{W}}^{-\top} = (\overline{\mathbf{Q}}\mathbf{J})^\top(\underline{\mathbf{Q}}\mathbf{J}^\top)^{-\top} = \mathbf{J}^\top\overline{\mathbf{Q}}^\top\underline{\mathbf{Q}}\mathbf{J}^\top, \tag{18}$$

where r is the order of the system and the matrices $\overline{\mathbf{W}}$ and $\underline{\mathbf{W}}$ have been truncated to the first k columns (corresponding to the k dominant eigenmodes in the original data). This is partly motivated as noise-suppression, and partly motivated by the fact that the amount of information in $\mathbf{A}$ will be determined by the subspace angle between $\overline{\mathbf{W}}$ and $\underline{\mathbf{W}}$. If they are square, the subspace angle is zero, meaning that the low-variance components of the signal have obscured the relationship between past and future.

Now we expand the SVD into the desired factorization:

[2] In the no-inputs case, there are more efficient ways to calculate $\overline{\mathbf{Q}}$ and $\underline{\mathbf{Q}}$, including treating $\mathbf{Z}$ as a QR decomposition and down-dating it to remove the first or last column while preserving row-orthogonality (see [3, §12.5]).

$$\{\overline{\mathbf{Z}}\mathbf{E}^{\perp}\}/\{\underline{\mathbf{Z}}\mathbf{E}^{\perp}\} = \overline{\mathbf{R}}^{\top}\overline{\mathbf{Q}}^{\top}\underline{\mathbf{Q}}\underline{\mathbf{R}}^{-\top} \tag{19}$$

$$= \overline{\mathbf{R}}^{\top}(\mathbf{J}\mathbf{J}^{\top})\overline{\mathbf{Q}}^{\top}\underline{\mathbf{Q}}(\mathbf{J}^{\top}\mathbf{J})\underline{\mathbf{R}}^{-\top} \tag{20}$$

$$= \overline{\mathbf{R}}^{\top}\mathbf{J}(\mathbf{U}\mathbf{S}\mathbf{V}^{\top})\mathbf{J}\underline{\mathbf{R}}^{-\top} \tag{21}$$

$$= \overline{\mathbf{R}}^{\top}\mathbf{J}(\mathbf{V}\sqrt{\mathbf{S}^{-1}}\sqrt{\mathbf{S}}\mathbf{V}^{\top})(\mathbf{U}\sqrt{\mathbf{S}}\sqrt{\mathbf{S}}\mathbf{V}^{\top})\mathbf{J}\underline{\mathbf{R}}^{-\top} \tag{22}$$

$$\approx \underbrace{\mathbf{R}_{\mathbf{J}}^{\top}\mathbf{V}\sqrt{\mathbf{S}^{-1}}}_{\mathbf{C}'}\underbrace{\sqrt{\mathbf{S}}\mathbf{V}^{\top}\mathbf{U}\sqrt{\mathbf{S}}}_{\mathbf{A}}\underbrace{\sqrt{\mathbf{S}}\mathbf{V}^{\top}\mathbf{R}_{\mathbf{J}}^{-\top}}_{\mathbf{C}'^{\dagger}} \tag{23}$$

Equation 23 becomes an equality as $T \to \infty$ or for finite T if the past and future subspaces are perfectly consistent in the sense described above; otherwise the approximation has minimal Frobenius error. Similarly, equation 22 requires $r = T - 1$ for equality, but preserves maximal variance for $r \leq k$.

$\mathbf{A}$ has interesting structure: $\mathbf{V}^{\top}\mathbf{U}$ is the solution to the orthogonal Procrustes problem optimally rotating $\underline{\mathbf{W}}$ into $\overline{\mathbf{W}}$, while $\mathrm{diag}(\mathbf{S})$ are the canonical correlations—cosines of the angles between corresponding columns in the past and future subspaces.

The parameter $\mathbf{C}'$ only gives the component of the output that is orthogonal to the inputs. To compute the full output matrix, we must recover the (redundant) components of $\mathbf{Z}$ that were discarded in the original orthogonal projection of equation 14:

$$\mathbf{C} = \mathbf{L}[\mathbf{R}_{\mathbf{J}},\ \mathbf{M}_{\mathbf{J}}]^{\top}\mathbf{V}\mathbf{S}^{-1/2}. \tag{24}$$

The input parameters can then be solved from equation 11.

3.2 Properties

Although not strictly necessary, the hidden state $\mathbf{X}$ is easy to calculate, and it affords an opportunity to study the factorization (equation 23):

$$\underline{\mathbf{X}} \doteq \mathbf{C}'^{\dagger}\{\underline{\mathbf{Z}}\mathbf{E}^{\perp}\} = \sqrt{\mathbf{S}}\mathbf{V}^{\top}\mathbf{J}\underline{\mathbf{Q}}^{\top} \tag{25}$$

$$= \sqrt{\mathbf{S}}\mathbf{V}^{\top}(\mathbf{V}\mathbf{S}^{-1}\mathbf{U}^{\top})(\mathbf{J}^{\top}\overline{\mathbf{Q}}^{\top}\underline{\mathbf{Q}}\mathbf{J}^{\top})\mathbf{J}\underline{\mathbf{Q}}^{\top} \tag{26}$$

$$= \sqrt{\mathbf{S}^{-1}}\mathbf{U}^{\top}\mathbf{J}^{\top}\overline{\mathbf{Q}}^{\top}\underline{\mathbf{Q}}\mathbf{Q}^{\top} \tag{27}$$

$$= \sqrt{\mathbf{S}^{-1}}\mathbf{U}^{\top}\overline{\mathbf{W}}^{\top}\underline{\mathbf{Q}}\mathbf{Q}^{\top} \tag{28}$$

$$\overline{\mathbf{X}} \doteq \mathbf{C}'^{\dagger}\{\overline{\mathbf{Z}}\mathbf{E}^{\perp}\} = \sqrt{\mathbf{S}}\mathbf{V}^{\top}\mathbf{J}^{\top}\overline{\mathbf{Q}}^{\top} \tag{29}$$

$$= \sqrt{\mathbf{S}}\mathbf{V}^{\top}(\mathbf{V}\mathbf{S}^{-1}\mathbf{U}^{\top})(\mathbf{J}^{\top}\overline{\mathbf{Q}}^{\top}\underline{\mathbf{Q}}\mathbf{J}^{\top})\mathbf{J}^{\top}\overline{\mathbf{Q}}^{\top} \tag{30}$$

$$= \sqrt{\mathbf{S}^{-1}}\mathbf{U}^{\top}\mathbf{J}^{\top}\overline{\mathbf{Q}}^{\top}\underline{\mathbf{Q}}\mathbf{J}^{\top}\mathbf{J}^{\top}\overline{\mathbf{Q}}^{\top} \tag{31}$$

$$= \sqrt{\mathbf{S}^{-1}}\mathbf{U}^{\top}\overline{\mathbf{W}}^{\top}\underline{\mathbf{W}}\overline{\mathbf{W}}^{\top} \tag{32}$$

Since we have two overlapping estimates of the state $\mathbf{X}$, we set $\mathbf{X} = \left[\underline{\mathbf{X}}_{(:,1:t)},\ \underline{\mathbf{X}}_{(:,t:T)}\right]$ for any $1 \leq t \leq T$. In what follows it is convenient to set $t = T$. In the limit of infinite data, $\underline{\mathbf{X}}$ and $\overline{\mathbf{X}}$ become perfectly consistent ($\lim_{T\to\infty} \|\underline{\mathbf{X}}_{(:,2:T)} - \overline{\mathbf{X}}_{(:,1:T-1)}\|_F = 0$). In practice we found that the subspace angle between the two estimates is vanishingly small, due largely to $\mathbf{J}$.

Using above the expressions for $\mathbf{X}$, we find the factorization has the following properties: The scatters of the past and future state estimates, $\underline{\mathbf{X}}\underline{\mathbf{X}}^{\top} = \overline{\mathbf{X}}\overline{\mathbf{X}}^{\top} = \mathbf{S}$, are diagonal and equal, indicating that the estimated LDS is *balanced* in the sense of [5] and therefore insensitive to small data perturbations[3]. The system residual (before considering inputs),

$$\mathbf{G} \doteq \overline{\mathbf{X}} - \mathbf{A}\underline{\mathbf{X}} = \sqrt{\mathbf{S}}\mathbf{V}^{\top}\mathbf{J}^{\top}(\overline{\mathbf{Q}}^{\top} - \overline{\mathbf{Q}}^{\top}\underline{\mathbf{Q}}\underline{\mathbf{Q}}^{\top}) = \overline{\mathbf{X}}(\mathbf{I} - \underline{\mathbf{Q}}\underline{\mathbf{Q}}^{\top}), \tag{33}$$

is the component of the future that is orthogonal to the past. For high-dimensional data, equation 33 shows that for any dimensionality-reduction of the system by truncating columns of $\mathbf{Q}$, the error associated with system matrix $\mathbf{A}$ is the projection of the state onto the subspace of the discarded columns. Since the retained columns are almost exactly the dominant eigenvectors of the original data (with equality at $t \gg d$), the optimal truncation property of the original SVD carries through to our parameter estimates.

The (eigen-coded) output residual (before considering inputs),

$$\{\underline{\mathbf{Z}}\mathbf{E}^{\perp}\} - \mathbf{C}'\underline{\mathbf{X}} = (\mathbf{J}\underline{\mathbf{R}} - \mathbf{R}_{\mathbf{J}})^{\top}\underline{\mathbf{Q}}^{\top} = (\mathbf{J}\underline{\mathbf{R}} - \mathbf{J}^{\top}\overline{\mathbf{R}})^{\top}\underline{\mathbf{Q}}^{\top}/2, \tag{34}$$

is essentially the residual (if any) after $\underline{\mathbf{R}}$ is rotated into $\overline{\mathbf{R}}$.

Substituting the system matrix $\mathbf{A}$ and output matrix $\mathbf{C}'$ into equations 8–9 gives the output residual (before considering inputs)

$$\mathbf{H} \doteq (\mathbf{Y} - \mathbf{C}\mathbf{X}) \tag{35}$$

$$= (\mathbf{Y} - \mathbf{L}[\mathbf{R}_{\mathbf{J}},\ \mathbf{M}_{\mathbf{J}}]^{\top}[\underline{\mathbf{W}}_{(:,1:t)}^{\top}, \overline{\mathbf{W}}_{(:,t:T)}^{\top}]), \quad 1 \leq t \leq T, \tag{36}$$

where we use $\mathbf{Y}$ instead of $\mathbf{Z}$ to recover information that was lost in the projection to the subspace orthogonal to inputs $\mathbf{E}$. One may use the system residual $\mathbf{G}$ and the output residual $\mathbf{H}$ to estimate the input and feed-through matrices

$$\mathbf{B} = (\overline{\mathbf{X}} - \mathbf{A}\underline{\mathbf{X}})\underline{\mathbf{E}}^{\dagger} = \mathbf{G}\underline{\mathbf{E}}^{\dagger} \tag{37}$$

$$\mathbf{D} = (\mathbf{Y} - \mathbf{C}\mathbf{X})\mathbf{E}^{\dagger} = \mathbf{H}\mathbf{E}^{\dagger}. \tag{38}$$

The residuals (after considering inputs) are

$$\Gamma = \overline{\mathbf{X}} - \mathbf{A}\underline{\mathbf{X}} - \mathbf{B}\underline{\mathbf{E}} = (\overline{\mathbf{X}} - \mathbf{A}\underline{\mathbf{X}})(\mathbf{I} - \underline{\mathbf{E}}^{\dagger}\underline{\mathbf{E}}) = \mathbf{G}(\mathbf{I} - \mathbf{Q}_{\underline{\mathbf{E}}}\mathbf{Q}_{\underline{\mathbf{E}}}^{\top}), \tag{39}$$

$$\Psi = \mathbf{Y} - \mathbf{C}\mathbf{X} - \mathbf{D}\mathbf{E} = \mathbf{H}(\mathbf{I} - \mathbf{E}^{\dagger}\mathbf{E}). \tag{40}$$

which confirms that the noise is orthogonal to the inputs. Furthermore, the state noise is the component of the future that is orthogonal to both the past and the inputs. Finally, the noise covariances are $\Sigma_{\gamma} = \mathrm{cov}(\Gamma) = \Gamma\Gamma^{\top}/T$, $\Sigma_{\psi} = \mathrm{cov}(\Psi) = \Psi\Psi^{\top}/T$.

3.3 Relation to subspace methods

Although our factorization is wholly novel, our approach shares two core tactics of subspace approaches: Comparison of past and future subspaces, and use of orthogonal projections to separate state from inputs. In most other regards our approach is distinct and produces different numerical results.

This is largely due to the $\mathbf{J}$-step: In the limit of infinite samples of finite-dimensional data, the residual

$$\lim_{T/d\to\infty} \|\underline{\mathbf{R}} - \overline{\mathbf{R}}\| = 0 \tag{41}$$

and therefore $\lim_{T/d\to\infty} \mathbf{J} = \mathbf{I}$. A wide variety of subspace system identification methods catalogued in [9] also depend on orthogonal

[3] Strictly speaking, the system is only approximately balanced for finite T, because $\mathbf{X}\mathbf{X}^{\top} \approx \underline{\mathbf{X}}\underline{\mathbf{X}}^{\top}$ has (very small) nonzero off-diagonal elements.

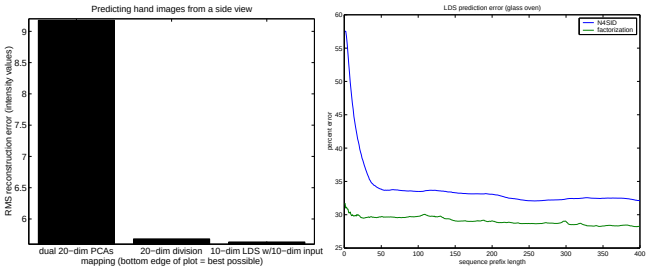

Figure 2: LEFT: Reconstruction error of a test target sequence from a test source sequence using 1) a least-squares mapping between eigen-codings of source and target images; 2) a rank-constrained division; and 3) an LDS estimated from a target image sequence and a very low-dimensional eigen-coding of the source sequence. The training (estimation) and test sequences are disjoint. The lower edge of the graph is the lower bound on how well test images can be reconstructed from linear combinations of training set images. RIGHT: Predicting the output of a glass oven from a history of its inputs.

representations of the past and future, but they make the assumption that equation 41 forms a reasonable approximation for finite-data, at least when $t \gg d$. This approximation can be quite good for long runs of low-dimensional data, but it can also be quite poor when data is limited, especially in our applications, where $t \ll d$. In our factorization, the matrix $\mathbf{J}$ recovers information that overlooked by subspace methods.

Because previous subspace methods have not reconciled $\overline{\mathbf{R}}$ and $\underline{\mathbf{R}}$, they typically yield two equally valid estimates for $\mathbf{C}$ that differ by an affine transform of $\overline{\mathbf{R}} - \underline{\mathbf{R}}$. Choosing either one results in a reconstruction residual and a bias in $\mathbf{A}$. Plugging the inequality

$$\|\mathbf{R}_\mathbf{J} - \mathbf{J}\underline{\mathbf{R}}\|_F \leq \|\mathbf{J}^\top\overline{\mathbf{R}} - \mathbf{J}\underline{\mathbf{R}}\|_F \leq \|\overline{\mathbf{R}} - \underline{\mathbf{R}}\|_F \qquad (42)$$

into equation 34 shows that our method will yield smaller state-to-frame (output) residuals $\|\{\underline{\mathbf{Z}}\mathbf{E}^\perp\} - \mathbf{C}'\underline{\mathbf{X}}\|_F$ than previous subspace methods. It is not necessarily the case that the state-to-state residual (equation 33) is also reduced: By using $\mathbf{J}$ to rotate $\overline{\mathbf{R}}$ and $\underline{\mathbf{R}}$ into alignment, $\mathbf{X}$ becomes a more faithful representation of the variations in the data, increasing the amount of variance that $\mathbf{A}$ must "explain." However, the inequality in equation 42 can be used to show that our method has lower frame-to-frame residuals $\|\{\overline{\mathbf{Z}}\mathbf{E}^\perp\} - \mathbf{C}\mathbf{A}\mathbf{C}'^\dagger\{\underline{\mathbf{Z}}\mathbf{E}^\perp\}\|_F$ (from equation 23) and previous-state-to-frame residuals $\|\{\overline{\mathbf{Z}}\mathbf{E}^\perp\} - \mathbf{C}\mathbf{A}\underline{\mathbf{X}}\|_F$, which are the important residuals to minimize for filtering and prediction applications.

4 Applications

Controlled image synthesis: We revisited the problem of view-mapping the hand, this time using an input/output LDS. Using the same test/train split as above, we estimated a 10-dimensional LDS, taking the training target sequence as outputs $\mathbf{Y}$ and a 10-dimensional eigen-coding of the training source sequence as inputs $\mathbf{E}$. The eigen-coded inputs do not contain enough information for a quality regression, so the LDS must propagate state through time to make up for the missing information. We ran the estimated LDS with the test source sequence as input, and found that its outputs reconstruct the test target sequence even better than the 20-dimensional rank-constrained division. See figure 2LEFT. This shows that the "learned" model of the hand's dynamics is successfully providing the information not in the inputs.

Industrial process prediction: To demonstrate the accuracy of our factorization in a lower-dimensional setting, we obtained some datasets used as benchmarks in the system identification community. The graph below shows prediction error for measurements taken from an analog industrial process (a glass oven). The input/output datasets, train/test splits, model orders, and error measure were taken from Overschee & deMoor [9], and models were estimated using our factorization and O&dM's "industrial-strength" **`subid`** implementation. The goal is to predict the next output given a history of inputs. Figure 2RIGHT shows that our factorization enjoys a clear advantage. Similar results were obtained from all datasets. It is not clear that this translates into improvements in process *control*, but the next section shows that it certainly does translate into improvements in process *simulation*.

Temporal texture synthesis: An LDS is a remarkably good model for image sequences in which intensity variations are due to local changes in the BRDF and/or small motions. This covers many natural phenomena including turbulently flowing water, rain, smoke, fire, wind-swayed vegetation. These are scenes for which motion-based video encoders such as MPEG typically offer little or no compression. The physics of these scenes—coupled harmonic oscillators that are lightly forced (typically by air currents) and friction-damped—is very well matched to the LDS model (e.g., see figure 3). Due to this goodness of fit, an LDS will offer good synthesis of novel sequences. One of the first demonstrations of the applicability of an LDS to temporal texture synthesis was given by Szummer [7]. The idea of whole-frame temporal texture synthesis was recently exploited to good effect by Soatto et al [6], who introduced a no-inputs first-order approximation to the Overschee method in [2].

We obtained Szummer's test sequences and took several new sequences of flowing rivers, swaying trees, steaming coffee, etc., with a hand-held camcorder. A 250-frame sequence of 320x240 images can be analyzed in roughly 40 seconds on a vintage 1998 Alpha CPU. We estimated an optimal dimensionality k for each sequence's eigen-coding from a lower bound on the mutual information between past and future (a paper on this is in preparation, however the results below do not depend on any value of k). The dimensionality r of the corresponding LDS was set to capture 95% of the mutual information. Each LDS was estimated using our factorization, a commercially available state-of-the-art implementation of the asymptotically optimal Overschee & de Moor (O&dM) subspace method [9], and the Doretto & Soatto (DS) factorization [2] using code available at their website. For each sequence, we found that all three methods yield different parameter estimates, with substantial differences in $\mathbf{A}$, $\mathbf{C}$, and the state covariance Γ, reflecting different inferences

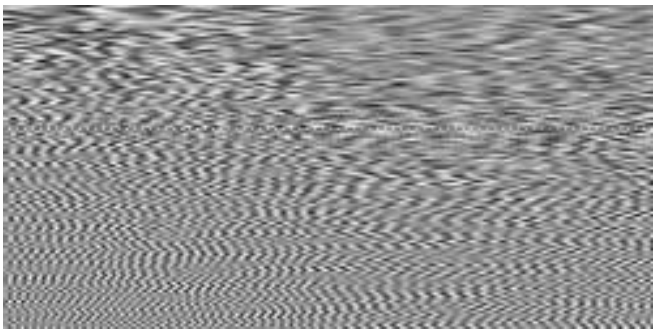

Figure 3: The coding of a river image sequence shows how it is decomposed into coupled oscillations of the basis images. Each image row depicts a column in $\underline{\mathbf{Q}}$. Time flows to the right; the low-frequency rows at top explain most of the variance in the sequence. Although each basis has a resonant frequency, the phase is affected by couplings between modes, visible as 2D textures. This particular sequence is clearly not wide-sense stationary—the deepening ripples on the right-hand side show how a gust of wind shifts the distribution of energy into the higher frequencies—yet the estimated LDS generates synthetic river sequences with high fidelity.

Table 1: RMS frame-to-frame errors for 3 LDS factorizations.

sequence	O&dM [9]	DS[2]	here
blown trees	fails	140.164	**93.140**
blown trees, truncated	83.193	72.512	**71.366**
river	fails	118.877	**54.118**
river, truncated	127.823	25.219	**24.228**
fire	fails	403.757	**188.552**
fire, truncated	468.330	175.702	**170.446**
hand	fails	162.384	**15.408**
hand, truncated	22.235	13.864	**12.163**

about the hidden state $\mathbf{X}$. As predicted by equation 42, our factorization consistently yielded the smallest residuals. The table below compares the frame-to-frame residuals (vis-a-vis equation 13) for some sequences randomly taken from our collection. All algorithms were given the same data and settings for k and r. (In order to provide conditions under which the O&dM method would work, we also ran "truncated" trials with halved values of k (and thus r), essentially removing high-frequency components of the data.)

For all sequences and all choices of r and k, our factorization yields the lowest error. A paired F-test analysis of variance indicates that these claims are statistically significant at $p < 10^{-4}$ levels (probability of error) for the "truncated" trials and $p < 10^{-6}$ levels for the fully automatic trials. Note that all algorithms do well modeling the low-frequency components of the temporal texture; our method does substantially better when high-frequency components must be modeled as well. In summary our methods reliably produce more accurate data reconstructions per model parameter. Measurements of reconstruction error on held-out subsequences show a similar advantage at the $p < 0.08$ level.

Following the example of [6], a random walk in our LDS state spaces yields high-quality synthetic videos, producing sequences of several times the length of the original video with high realism but containing none of the original frames[4].

5 Summary

We have considered here the problem of predicting an image from a previous or a related image using linear subspaces. We gave novel factorizations for a static mapping, which is a rank-constrained matrix division, and for a mapping in which the system has hidden state, which is system identification of an LDS, a.k.a. Kalman filter. All results are closed-form and inherit the SVD's optimal reconstruction properties when the matrices used to effect the mapping have reduced dimension. To demonstrate the applicability of these solutions to video synthesis, we explored the use of rank-constrained division to do view-mappings and the use of an LDS to synthesize realistic temporal video texture in response to low-dimensional control signals or via random walks. As predicted analytically and confirmed numerically, the novel factorization exhibits more accurate reconstructions of estimation data and predictions of held-out test-data.

References

[1] B. de Moor, P. van Overschee, and W. Favoreel. *Applied and Computational Control, Signals and Circuits*, chapter Numerical algorithms for subspace state space system identification – An Overview', pages 247–311. Birkhauser Books, 1999.

[2] S. Doretto and S. Soatto. Dynamic data factorization. Technical Report TR2001-0001, UCLA Computer Science, http://www.vision.cs.ucla.edu/papers/TR2000-0001.pdf, 2001.

[3] G. Golub and A. van Loan. *Matrix Computations*. Johns Hopkins U. Press, 1996.

[4] F. D. la Torres and M. Black. Dynamic coupled component analysis. In *Proc., Computer Vision and Pattern Recognition*, volume II, pages 643–649, 2001.

[5] B. Moore. Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transactions on Automatic Control*, 26(1):17–32, 1981.

[6] S. Soatto, G. Doretto, and Y. Wu. Dynamic textures. In *Intl. Conf. on Computer Vision*, 2001.

[7] M. Szummer and R. W. Picard. Temporal texture modeling. In *IEEE International Conference on Image Processing*, 1996.

[8] P. van Overschee and B. de Moor. Subspace algorithms for the stochastic identification problem. *Automatica*, 29(3):649–660, March 1993.

[9] P. van Overschee and B. de Moor. *Subspace identification for linear systems*. Kluwer Academic, 1996.

[4] In contrast, video synthesized from models estimated with the code from [2] suffered from state explosions or extinctions within a few hundred frames, suggesting bias in the estimated system matrices and/or covariance matrices.