Dynamic Data Dimensionality Reduction: A Factorization Approach

Stefano Soatto Alessandro Chiuso Ying-Nian Wu

Department of Computer Science, UCLA, Los Angeles - CA 90095
Department of Electrical Engineering, Washington University, St.Louis - MO 63130
Dipartimento di Elettronica ed Informatica, Universitá di Padova, Italy

Keywords: canonical correlation, model reduction, balanced realization, subspace identification, dynamic scene analysis, minimum description length, image sequence compression, video coding, generative model, prediction error methods, ARMA regression, system identification, learning, E-M algorithm

Technical Report UCLA-CSD 010001, March 6, 2000

Abstract

In this report we present a technique for reducing the dimensionality of a stationary time series with an arbitrary covariance sequence. Contrary to popular belief, this problem can be solved in closed form in a way that is asymptotically efficient (i.e. optimal in the sense of maximum likelihood). Our technique can be seen as a special case of a more general theory of subspace identification borrowed from the field of dynamical systems. Although our scheme only captures second-order statistics, we suggest future developments that model higher-order moments.

1 Preliminaries

Let $\{y(t)\}_{t=1...\tau}$ be a discrete-time stochastic process with sample paths $y(t) \in \mathbb{R}^m$. We are interested in reducing the dimensionality of the "data" (i.e. a sample path of dimension $\mathcal{O}(m\tau)$) by extracting a *model* for the process $\{y\}$.

We restrict our attention to second-order statistics and assume that the process $\{y(t)\}$ generates a covariance sequence with rational spectrum. Later we will extend our approach to a more general class of models.

1.1 Model realization

It is well known that a positive definite covariance sequence with rational spectrum corresponds to an equivalence class of second order stationary processes [11]. It is then possible to choose as a representative of each class a Gauss-Markov model – that is the output of a linear dynamical system driven by white, zero-mean Gaussian noise – with the given covariance. In other words, we can assume that there exists a positive integer n, a process $\{x(t)\}$ (the "state") with initial condition $x_0 \in \mathbb{R}^n \sim \mathcal{N}(0, P)$ and a symmetric positive semi-definite matrix $\begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \geq 0$ such that $\{y(t)\}$ is the output of the following Gauss-Markov "ARMA" model[1]:

$$\begin{cases} x(t+1) = Ax(t) + v(t) & v(t) \sim \mathcal{N}(0, Q); \quad x(0) = x_0 \\ y(t) = Cx(t) + w(t) & w(t) \sim \mathcal{N}(0, R); \quad E[w(t)v^T(t)] = S \end{cases} \tag{1}$$

for some matrices $A \in \mathbb{R}^{n \times n}$ and $C \in \mathbb{R}^{m \times n}$. The goal of dimensionality reduction can be posed as the *identification* of the model above, that is the estimation of the model parameters A, C, Q, R, S from "output" measurements $y(1), \ldots, y(\tau)$.

[1] ARMA stands for auto-regressive moving average.

It is commonly believed that – even for the case of linear models driven by Gaussian noise such as (1) – this problem can only be solved iteratively, and is often formulated in the framework of expectation-maximization [14]. Instead, we will follow the philosophy of subspace identification methods as championed by Van Overschee and DeMoor [12], and show that a solution can be computed in closed form. This solution can be shown to be asymptotically efficient under suitable hypothesis.

1.2 Uniqueness and canonical model realizations

The first observation concerning the model (1) is that the choice of matrices A, C, Q, R, S is not unique, in the sense that there are infinitely many models that give rise to exactly the same measurement covariance sequence starting from suitable initial conditions. The first source of non-uniqueness has to do with the choice of basis for the state space: one can substitute A with TAT^{-1}, C with CT^{-1}, Q with TQT^T, S with TS, and choose the initial condition Tx_0, where $T \in \mathcal{GL}(n)$ is any invertible $n \times n$ matrix and obtain the same output covariance sequence.

The second source of non-uniqueness has to do with issues in spectral factorization that are beyond the scope of this paper [11]. Suffices to our purpose to say that one can transform the model (1) into a particular unique form – the so-called "innovation representation" – given by

$$\begin{cases} x(t+1) = Ax(t) + Ke(t) & x(0) = x_0 \sim \mathcal{N}(0, P) \\ y(t) = Cx(t) + e(t) & e(t) \sim \mathcal{N}(0, \Lambda); \end{cases} \tag{2}$$

where K is called "Kalman gain" and $e(t)$ the "innovation proces[2]". Therefore, without loss of generality, we restrict ourselves to models of the form (2).

We are therefore left with the first source of non-uniqueness: any given process has *not* a unique innovation model, but an *equivalence class* of models $\mathcal{R} \doteq \{[A] = TAT^{-1}, [C] = CT^{-1}, [K] = TK, \mid T \in \mathcal{GL}(n)\}$. In order to be able to identify a unique model of the type (2) from a sample path $y(t)$, it is therefore necessary to choose a representative of each equivalence class (i.e. a basis of the state-space): such a representative is called a *canonical model realization* (or simply canonical realization). It is canonical in the sense that it does not depend on the choice of the state space (because it has been fixed).

While there are many possible choices of canonical realizations (see for instance [10]), we are interested in one that is "tailored" to the data, in the sense of having a diagonal state covariance. Such a model realization is called *balanced* [3]. Since we are interested in data dimensionality reduction, we will make the following assumptions about the model (1):

$$m >> n; \ \text{rank}(C) = n \tag{3}$$

and choose the canonical realization that makes the columns of C orthogonal:

$$C^T C = \Sigma_n^{1/2} = \text{diag}\{\sigma_1^{1/2}, \ldots, \sigma_n^{1/2}\} \tag{4}$$

where, without loss of generality, $\sigma_i \geq \sigma_j$, for $i < j$ and σ_i are non negative. One can show that the numbers σ_i are an invariant of the process.

As we will see shortly, this assumption results in a unique innovation model. Such a model is characterized by a state space such that its covariance $\Sigma_n^{1/2} = E[x(t)x^T(t)]$ is diagonal, and its diagonal elements are uniquely defined by the data. We shall also make a simplifying assumption that $E[x(t+1)y^T(t)]$ has full row rank, that is the state is *constructible* in one step.

The real numbers σ_i, which one can estimate from the data $y(t)$, are called the "canonical correlation coefficients" [9], as they represent the canonical correlations between past (say $\mathcal{P}_t \doteq \overline{\text{span}}\{y(s), s < t\}$) and future (say $\mathcal{F}_t \doteq \overline{\text{span}}\{y(s), s \geq t\}$) of the process. In terms of data[3] the spaces $\mathcal{P}_t$ and $\mathcal{F}_t$ can be thought of as the row-span of the

[2]The innovation process can be interpreted as the one-step prediction error $e(t) \doteq y(t) - \hat{y}(t|t-1)$. It is the error one commits in predicting (in the Bayesian sense) the value of $y(t)$ given the values of $y(s)$, for $s < t$.

[3]The Hilbert spaces of zero-mean and finite variance random variables and the (row) space of semi-infinite sequences constructed from a realization (sample-path) are isometrically isomorphic with a suitable choice of inner product

doubly-infinite block-Hankel matrices:

$$\mathcal{F}_t \doteq \overline{\text{row span}} \begin{bmatrix} y(t) & y(t+1) & \dots & y(t+\tau-1) & \dots \\ y(t+1) & y(t+2) & \dots & y(t+\tau) & \dots \\ \vdots & \vdots & & \vdots & \\ y(T) & y(T+1) & \dots & y(T+\tau-1) & \dots \\ \vdots & \vdots & & \vdots & \end{bmatrix}$$

and

$$\mathcal{P}_t \doteq \overline{\text{row span}} \begin{bmatrix} y(t-1) & y(t) & \dots & y(t+\tau-2) & \dots \\ y(t-2) & y(t-1) & \dots & y(t+\tau-3) & \dots \\ \vdots & \vdots & & \vdots & \\ y(t-T) & y(1) & \dots & y(\tau-1) & \dots \\ \vdots & \vdots & & \vdots & \end{bmatrix}.$$

In practice one only considers the finite past and future, i.e. takes only a finite number of block rows in these matrices (under our hypothesis we can take just one) and, as only finite data are available, infinite sequences are approximated by finite ones (finite number of columns). The canonical correlations σ_i can be related to the mutual information between past and future by the well-known formula

$$I(\mathcal{P}_t, \mathcal{F}_t) = \sum_{i=1}^{n} log\left(\frac{1}{1-\sigma_i^2}\right) \tag{5}$$

which is independent of t by stationariety [1]. This interpretation can be used in the reduction context as one wants to reduce the dimension of the state n while keeping the maximum amount of information. If $x(t)$ is chosen in this basis, according to (5), the reduction just consist in dropping its last components.

The problem we set out to solve can then be formulated as follows: *given* measurements of a sample path of the process: $y(1), \dots, y(\tau)$; $\tau >> n$, estimate $\hat{A}, \hat{C}, \hat{K}, \hat{\Lambda}$, a canonical realization of the process $\{y(t)\}$. Ideally, we would want the maximum likelihood solution from the finite sample:

$$\hat{A}(\tau), \hat{C}(\tau), \hat{K}(\tau), \hat{\Lambda}(\tau) = \arg \min_{A,C,K,\Lambda} p(y(1), \dots, y(\tau) | A, C, K, \Lambda). \tag{6}$$

We will first derive a closed-form suboptimal solution and then show that it asymptotically maximizes the likelihood.

2 Closed-form solution for the linear Gaussian case

The idea behind dynamic dimensionality reduction is to use the dynamic properties of the processes involved to determine a low rank approximation. To understand this fact, let us for a moment assume that we only consider the output equation. Let $Y_1^{\tau-1} \doteq [y(1), \dots, y(\tau-1)] \in \mathbb{R}^{m \times \tau-1}$ with $\tau > n$, and similarly for $X_1^{\tau-1}$ and $W_1^{\tau-1}$, and notice that $Y_1^{\tau-1} = CX_1^{\tau-1} + W_1^{\tau-1}$; $C \in \mathbb{R}^{m \times n}$; $C^T C = \Sigma_n^{1/2}$ by our assumptions (3) and (4). Consider the problem of finding the best estimate of C in the sense of Frobenius: $\hat{C}(\tau), \hat{X}(\tau) = \arg\min_{C, X_1^{\tau-1}} \|Y_1^{\tau-1} - CX_1^{\tau-1}\|_F$. Let $Y_1^{\tau-1} = U\Sigma V^T$; $U \in \mathbb{R}^{m \times n}$; $U^T U = I$; $V \in \mathbb{R}^{\tau \times n}$, $V^T V = I$ be the singular value decomposition (SVD) [8] with Σ diagonal. Let U_n, V_n be the matrices formed with the first n columns of U and V respectively. Let us also denote with $\Sigma_n = \text{diag}\{\sigma_1, \dots, \sigma_n\}$. It follows immediately from the fixed rank approximation property of the SVD [8] that $\hat{C}(\tau) = U_n$; $\hat{X}(\tau) = \Sigma_n V_n^T$. However, this solution is "static" in the sense that it does not take into account the fact that the rows of X have a very particular structure (determined by equation 2). In particular, the state $x(t)$ (i.e. a column of X) represents a very special low-rank approximation of the output $y(t)$: it is the one that makes the "past" and the "future" conditionally independent. Therefore, in the context of dimensionality reduction of finite data, it is natural to choose the state to be the n-dimensional subspace of the past measurements which retains the maximum amount information to predict future measurements at any given time according to formula (5). In the rest of this section we shall construct such an approximation.

The way we construct this approximation can be summarized as follows. Let L_f be the lower triangular Cholesky factor [8] of $\frac{1}{\tau-1}Y_2^\tau(Y_2^\tau)^T$ and L_p that of $\frac{1}{\tau-1}Y_1^{\tau-1}(Y_1^{\tau-1})^T$. We want to find the "best", in the sense of weighted Frobenius norm, i.e. $\|L_f^{-1}\left(Y_2^\tau - \Phi Y_1^{\tau-1}\right)\|_F$, n-dimensional subspace of the (finite) past (say $Y_1^{\tau-1}$) to predict the (finite) future (say Y_2^τ). The reason why the norm is weighted by the inverse cholesky factor is that we want to compute relative error and not absolute one. The answer to this question [8] is found by computing the first n principal directions in $Y_1^{\tau-1}$ with respect to Y_2^τ and choosing the state as the space spanned by these components. Principal directions are computed using the Singular Value Decomposition as follows.

Let us denote $Y_2^\tau \doteq [y(2), \ldots, y(\tau)] \in \mathbb{R}^{m\times\tau-1}$ with $\tau > n$, and similarly for X_2^τ and W_2^τ, and notice that

$$Y_2^\tau = CX_2^\tau + W_2^\tau; \quad C \in \mathbb{R}^{m\times n}; \tag{7}$$

Let us denote with $\hat{Y}_2^\tau$ the orthogonal projection[4] of the rows of Y_2^τ onto the row-span of the matrix $Y_1^{\tau-1}$ and similarly by $\hat{Y}_3^{\tau+1}$ the orthogonal projection of the rows of $Y_3^{\tau+1}$ onto the row-span of $Y_1^{\tau-1}$ and Y_2^τ. Let us compute the singular value decomposition (SVD) [8]

$$\frac{1}{\tau-1}L_f^{-1}Y_2^\tau\left(Y_1^{\tau-1}\right)^T L_p^{-T} = U\Sigma V^T; \quad U \in \mathbb{R}^{m\times m};\ U^TU = I;\ V \in \mathbb{R}^{m\times m},\ V^TV = I \tag{8}$$

with $\Sigma = \mathrm{diag}\{\sigma_1, \ldots, \sigma_n, \ldots \sigma_m\}$. Ideally (in the absence of noise and when the data are generated according to a Gauss-Markov model) $\sigma_{n+1} = \sigma_{n+2} = \ldots = \sigma_m = 0$. In practice, however, this does not hold. One can therefore choose n by choosing a threshold on the value of σ_i. Let us denote with U_n, V_n the matrices formed with the firs n columns of U and V respectively. Let us also denote with $\Sigma_n = \mathrm{diag}\{\sigma_1, \ldots, \sigma_n\}$. Consider the problem of finding the best estimate of C and the best estimate of X of the form $\hat{X}_1^{\tau-1} = MY_1^{\tau-1}$ for some linear operator $M \in \mathbb{R}^{(n\times m)}$ acting on the "past", in the sense of Frobenius: $\hat{C}(\tau), \hat{M} = \arg\min_{C,M}\|L_f^{-1}\left(Y_2^\tau - CMY_1^{\tau-1}\right)\|_F$. It is shown in [8] that this is solved by choosing $\hat{X}_1^{\tau-1}$ as the space spanned by the first n principal directions in $Y_1^{\tau-1}$ with respect to Y_2^τ. It is also shown in [8] that these principal directions are given by the formula:

$$\Sigma_n^{1/2}V_n^TL_p^{-1}Y_1^{\tau-1} = \Sigma_n^{-1/2}U_n^TL_f^{-1}\hat{Y}_2^\tau$$

Therefore, we have

$$\hat{C}(\tau) = L_fU_n\Sigma_n^{1/2};\ \hat{X}_1^{\tau-1} = \Sigma_n^{1/2}V_n^TL_p^{-1}Y_1^{\tau-1} = \Sigma_n^{-1/2}U_n^TL_f^{-1}\hat{Y}_2^\tau. \tag{9}$$

$\hat{A}$ can be determined uniquely, again in the sense of Frobenius, by solving the following linear problem: $\hat{A}(\tau) = \arg\min_A\|\hat{X}_2^\tau - A\hat{X}_1^{\tau-1}\|_F$ which is trivially done in closed form using $\hat{X}_2^\tau(\tau) \doteq \Sigma_n^{-1/2}U_n^TL_f^{-1}\hat{Y}_3^{\tau+1}$ as

$$\hat{A}(\tau) = \frac{1}{\tau-1}\hat{X}_2^\tau\left(\hat{X}_1^{\tau-1}\right)^T\Sigma_n^{-1}. \tag{10}$$

Notice that $\hat{C}(\tau)$ is uniquely determined up to a change of sign of the components of C and x. Also note that

$$E[\hat{x}(t)\hat{x}^T(t)] \equiv \lim_{\tau\to\infty}\frac{1}{\tau}\sum_{k=1}^{\tau}\hat{x}(t+k)\hat{x}^T(t+k) = \Sigma_n^{1/2}V_n^TV_n\Sigma_n^{1/2} = \Sigma_n \tag{11}$$

which is diagonal. Thus the resulting model is *stochastically balanced*[5].

Let $\hat{e}(t) \doteq y(t) - \hat{C}(\tau)\hat{x}(t)$ be the estimated innovation which is, strictly speaking, the one-step ahead prediction error based on just one measurement $y(t-1)$, i.e. it is the transient innovation. Its covariance can be obtained by

$$\hat{\Lambda}(\tau) = \frac{1}{\tau-1}\hat{E}_1^{\tau-1}\left(\hat{E}_1^{\tau-1}\right)^T$$

where $\hat{E}_1^{\tau-1}$ is defined as $\hat{E}_1^{\tau-1} = [\hat{e}(1), \ldots, \hat{e}(\tau-1)]$. The input-to-state matrix K can be estimated, following (2) as follows. Compute $\hat{K}(\tau) = \arg\min_K\|\hat{X}_2^\tau - K\hat{E}_1^{\tau-1}\|_F$ which yields

$$\hat{K}(\tau) = \frac{1}{\tau-1}X_2^\tau\left(E_1^{\tau-1}\right)^T\hat{\Lambda}(\tau)^{-1}. \tag{12}$$

[4]The orthogonal projection is computed via $\hat{Y}_2^\tau \doteq Y_2^\tau\left(Y_1^{\tau-1}\right)^T\left[Y_1^{\tau-1}\left(Y_1^{\tau-1}\right)^T\right]^{-1}Y_1^{\tau-1}$.

[5]Strictly speaking this model is *finite interval*-balanced as in this case σ_i are the canonical correlation coefficients between finite past and finite future, namely $Y_1^{\tau-1}$ and Y_2^τ.

Note that we have used the finite past and future ($Y_1^{\tau-1}$ and Y_2^{τ}) to obtain our estimate. As we have pointed out, the difference $y(t) - \hat{C}(\tau)\hat{x}(t)$ is the transient innovation i.e. it is the prediction error based on the finite past. It coincides with the stationary one if and only if the system has finite memory m (in our case ha memory one), which happens if and only if $(A - KC)^m$ is nilpotent (in our case $A - KC = 0$).

It follows easily that its variance is bigger than the variance of the true innovation. For the same reason also the matrix $\hat{K}(\tau)$ is not the stationary K (which is called the "Kalman" gain). There is a procedure, which could in principle give us consistent estimates Λ and K even using the finite past and future under some assumptions. This is based on solving some Riccati equation, but would require the introduction of a certain "backward " model for the process y, which is beyond our scopes. This fact, however, seems to have been partially overlooked also in the system identification community, and we will not comment on it further beyond warning the reader that there are indeed procedures which gives consistent and asymptotically efficient estimates [6].

2.1 Choice of model order

In the algorithm above we have assumed that the order of the model n was given. In practice, this needs to be inferred from the data. Following [3], we propose determining the model order empirically from the singular values $\sigma_1, \sigma_2, \ldots$, by choosing n as the cutoff where the value of σ drops below a threshold. A threshold can also be imposed on the difference between adjacent singular values. The problem of order estimation is currently a research topic in the field of subspace identification and can not be considered as a fully solved problem. There are, however, numerous procedures based either on Information Theoretic criteria or Hypothesis testing ([4, 13] and references therein) which can be used to choose a threshold.

2.2 Asymptotic properties

While the solution given above is clearly suboptimal, it is possible to show that a slight modification of the algorithm reported above is asymptotically efficient. This proof of this fact has not yet been published and is due to D. Bauer [5]. The proof is valid for a particulr type of algorithms, which, in our setup, would substantially increase the computational complexity

3 Extensions to nonlinear, non-Gaussian models

While the algorithm proposed above depends critically on the linear structure of the problem and only captures the second-order statistics of the process $\{y(t)\}$, some simple modifications allow extending it to non-linear models and/or to non-Gaussian processes.

3.1 Choice of basis and independent components

The estimate of C derived in Section 2 can be interpreted in a functional sense as determining a choice of (orthonormal) basis elements for the space of measurements. This basis captures the principal directions of the correlation structure of the data. However, one may want to capture higher-order statistical structure in the data, for instance by requiring that the state vectors are not only orthogonal (in the sense of correlation), but also *statistically independent*. This could be done in the context of independent component analysis (ICA) by seeking for

$$\hat{C} = \arg\min KL(p(y(1), \ldots, y(\tau)) | p(Cx(1), \ldots Cx(\tau))) \tag{13}$$

subject to

$$p(x(1), \ldots, x(\tau)) = p(x(1)) \cdots p(x(\tau)) \tag{14}$$

where KL indicates Kullback-Leibler's divergence. Although this seems appealing from the point of view of dimensionality reduction of a static set, $x(t)$ are not independent for they have to satisfy (1). Nevertheless, $\hat{C}$ could be chosen according to the independence criterion so as to extract maximally independent components [2], and then $\hat{X}$ and $\hat{A}$ could be inferred from the estimated C according to a least-squares or maximum likelihood criterion.

3.2 Features and kernels

The model (1) assumes the existence of a state process with realization $x(t)$ of which the data $y(t)$ are a linear combination; the algorithm in Section 2 uses the geometry of the measurement space (specifically its inner product $\langle y(i), y(j) \rangle$) in order to recover the matrix C.

More in general, however, one could postulate the existence of a state of which the data are a *nonlinear* combination. If h is an invertible map, then we could have $y(t) = h(Cx(t) + w(t))$. Equivalently, if $\phi \doteq h^{-1}$, one could postulate the existence of a function ϕ acting on the data in such a way that its image is a linear combination of states $x(t)$:

$$\phi(y(t)) = Cx(t) + w(t). \tag{15}$$

Such images $\phi(t) \doteq \phi(y(t))$ are called *features*. Applying the algorithm in Section 2 to the transformed data is equivalent to changing the geometry of the measurement space by choosing an inner product in feature space:

$$\langle\langle y(i), y(j) \rangle\rangle \doteq \langle \phi(y(i)), \phi(y(j)) \rangle \tag{16}$$

in a way that is reminiscent of kernel methods [15]. The feature map ϕ could either be assigned "ad-hoc", or it could be part of the identification procedure. Ad-hoc choices may include wavelet transforms, Gabor filters etc.

4 Applications

4.1 Dynamic image modeling, coding and synthesis

We have tested the power of the algorithm above in modeling visual scenes that exhibit certain stationariety properties, so-called "dynamic textures" [7]. A small sample (50-100 images) of scenes containing water, smoke, foliage etc. has been used to identify a model, which can then be used for coding with orders of magnitude compression factors, or for synthesis of novel views.

In figure 1 we show a few images of the original sample (top) as well as simulated ones (bottom), and the value of the correlation coefficients. The order of the model has been truncated to 50. Complete video sequences can be downloaded from `http://vision.ucla.edu`. A different sequence depicting water undergoing vortical motion is shown in figure 2, together with the synthetic sequence and the correlation coefficients.

References

[1] H. Akaike. Canonical correlation analysis of time series and the use of an information criterion. *System identification: advances and case studies*, R. Mehra and D. Lainiotis (Eds.):27–96, 1976.

[2] S. Amari and F. Cardoso. Blind source separation– semiparametric statistical approach. *IEEE Trans. Signal Processing*, 45(11):2692–2700, 1997.

[3] K Arun and S. Y. Kung. Balanced approximation of stochastic systems. *SIAM Journal of Matrix Analysis and Applications*, 11(1):42–68, 1990.

[4] D. Bauer. *Some asymptotic theory for the estimation of linear systems using maximul likelihood methods or subspace algorithms.* PhD thesis, Technical University, Wien, 1998.

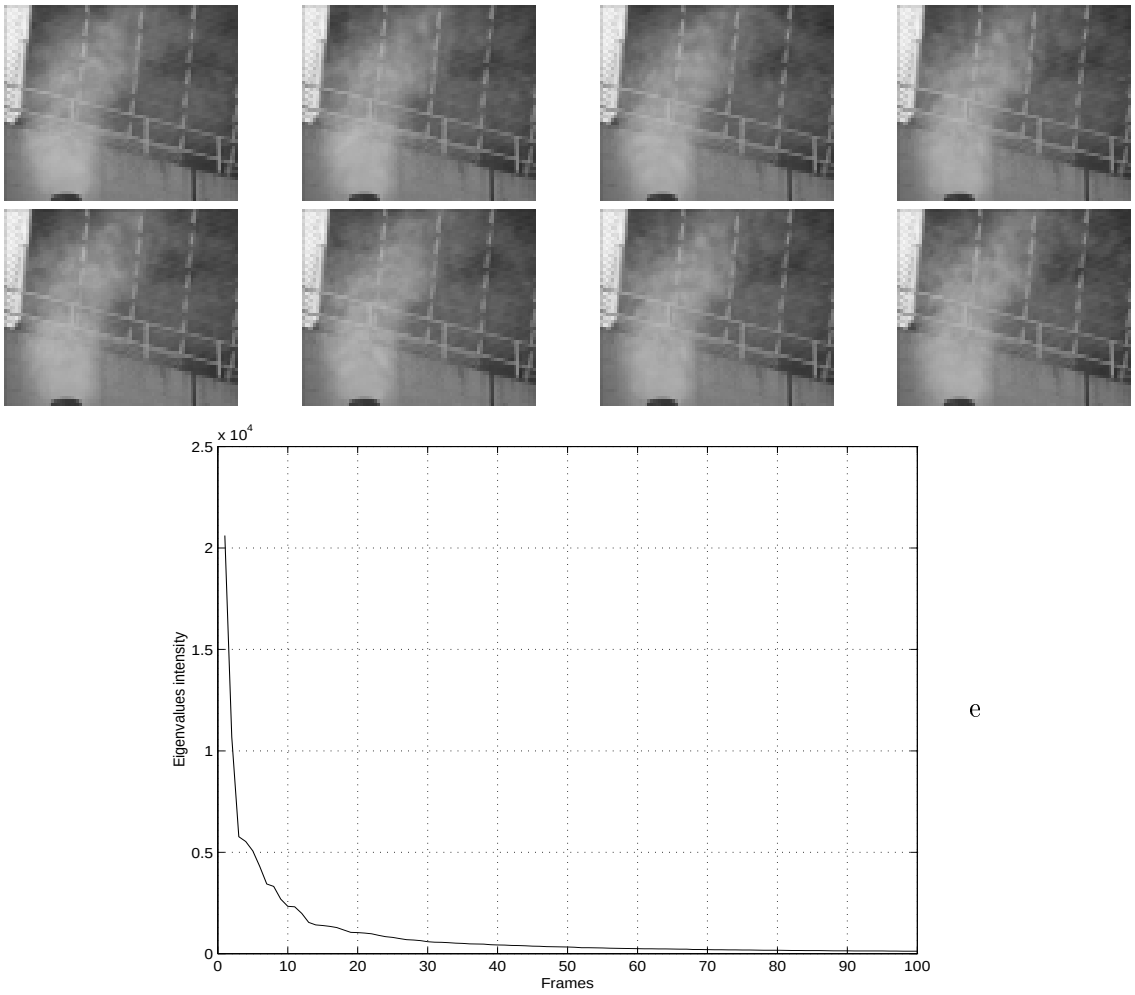
[5] D. Bauer. Asymptotic efficiency of the ccs subspace method in the case of no exogeneous inputs. *(submitted)*, 2000.

[6] A. Chiuso. *Gometric methods for subspace identification.* PhD thesis, Università di Padova, Feb 2000.

[7] G. Doretto, P. Pundir, Y. Wu, and S. Soatto. Dynamic textures. In *Technical Report UCLA CSD-200032*, page (submitted), 2000.

[8] G. Golub and C. Van Loan. *Matrix computations.* Johns Hopkins University Press, 2 edition, 1989.

[9] H. Hotelling. Relations between two sets of variables. *Biometrica*, 28:321–377, 1936.



e

Figure 1:

[10] T. Kailath. *Linear Systems*. Prentice Hall, 1980.

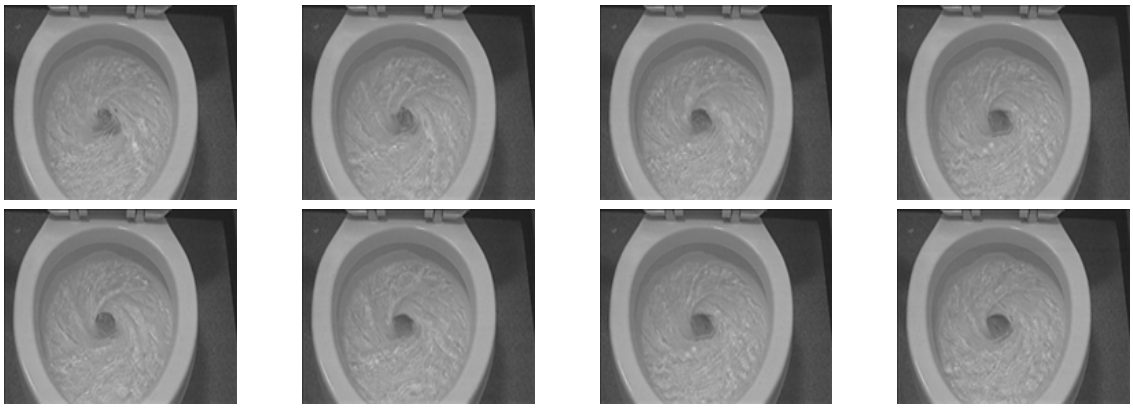
[11] L. Ljung. *System Identification: theory for the user*. Prentice Hall, 1987.

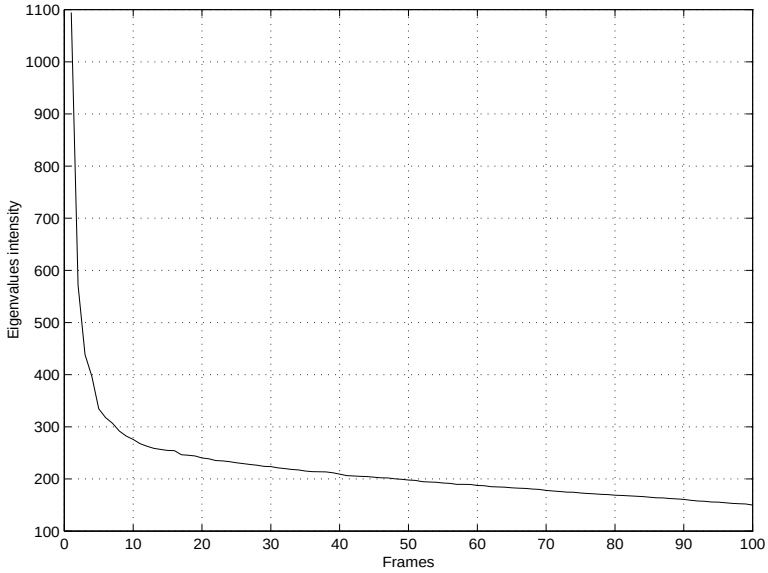
[12] P. Van Overschee and B. De Moor. Subspace algorithms for the stochastic identification problem. *Automatica*, 29:649–660, 1993.

[13] K. Peternell. *Identification of linear time-invariant systems via subspace and realization-based methods*. PhD thesis, Technical University, Wien, 1995.

[14] S. Roweis and Z. Ghahramani. A unifying review of linear gaussian models. *Neural Computation*, 11(2), 1999.

[15] B. Schölkopf and A. Smola. Kernel pca: pattern reconstruction via approximate pre-images. In *8th International Conference on Articial Neural Networks*, pages 147 – 152, 1998.





e

Figure 2: