Common SpatioTemporal Pattern Analysis

Ronald Phlypo, Nisrine Jrad, Bertrand Rivet, Marco Congedo

Common SpatioTemporal Pattern Analysis

Ronald Phlypo*, Nisrine Jrad, Bertrand Rivet, and Marco Congedo

Vision and Brain Signal processing (ViBS) Group, GIPSA Lab. Grenoble INP/UMR 5216 CNRS. BP 46, 961, Rue de la Houille Blanche, 38901 Saint Martin d'Hères, France.
firstname.lastname@gipsa-lab.grenoble-inp.fr

Abstract. In this work we present a method for the estimation of a rank-one pattern living in two heterogeneous spaces, when observed through a mixture in multiple observation sets. Using a well chosen representation for an observed set of second order tensors (matrices), a singular value decomposition of the set structure yields an accurate estimate under some widely acceptable conditions. The method performs a completely algebraic estimation in both heterogeneous spaces without the need for heuristic parameters. Contrary to existing methods, neither independence in one of the spaces, nor joint decorrelation in both of the heterogeneous spaces is required. In addition, because the method is not variance based in the input space, it has the critical advantage of being applicable with low signal-to-noise ratios. This makes this method an excellent candidate ,e.g., for the direct estimation of the spatio-temporal P300 pattern in passive exogenous brain computer interface paradigms. For these applications it is often sufficient to consider quasi-decorrelation in the temporal space only, while we do not want to impose a similar constraint in the spatial domain.

1 Introduction

Passive exogenous brain computer interface (BCI) paradigms, such as the P300 speller, require sound signal processing techniques to identify the typical cerebral responses measured as scalp potential differences. The difficulty of the signal processing is mainly due to the low signal-to-noise ratio of the electroencephalogram (EEG), i.e. the recording containing the temporally and spatially sampled potential field at the scalp. Indeed, the P300 voltage as measured at the scalp electrodes amounts to a few microvolts, while the ongoing spontaneous electroencephalogram is dominated by oscillations extending over several tens of microvolts. Fortunately, signal processing techniques may use the phase-lock between the stimulus onset and the P300 inflection (note that the P300 originates from a positive inflection around 300ms post-stimulus). A straightforward way to estimate the P300 waveform would thus be to mean out the unsynchronised ongoing cerebral activity over a set of observations (trials) aligned to the stimuli onsets.

* This work has been supported through the project grant ANR-09-BLAN-0330 (Gaze & EEG) of the ANR (National Research Agency), France.

However, plain averaging (taking the mean value over the trials) is biased by the appearance of high amplitude artifacts such as eye blinks present in some of the trials (outliers of the distribution over which the mean is taken), and this bias is inversely proportional with the number of trials that are available. One could use robust statistics to estimate the average waveform, or use the more commonly weighted averaging techniques [1]. Weighted averaging techniques optimise the ratio of the estimated signal energy with respect to that of the noise and ongoing background EEG conditional on the availability of the covariance structures of the signal part and the background/noise part. However, care should be taken in the estimation of the weights, since generally the covariance structures of the background EEG (together with that of the noise components) and that of the component of interest can only be roughly estimated, resulting in an estimation of inferior quality with respect to simple averaging [2,3].

It is worthy to note that the above methods make only use of the temporal diversity (the temporal distribution of the samples over a time window), but none does exploit the spatial diversity (the spatial distribution of the samples over the various electrodes). In [4], the careful use of principal component analysis with varimax rotation has been promoted. It has been noted that the variance itself might not be the optimal criterion to obtain a P300 estimate and that the correlation might be more appropriate. In addition, – and confirmed in [5] – the varimax rotation seems to be preferable, since here, compact temporal representations of the components are preferred over smeared out variances. The varimax rotation is also closely related to the independent component analysis (ICA) [6] (applied to event-related potentials in e.g. [7]). Despite the fact that the compactness of the representation uses temporal information, rather than mere spatial decorrelation based on the covariance structure, the methods still either need a proper pre-selection of that trials/electrodes that are not contaminated by artifacts [5] or need a posterior selection of the components [7]. This is because some artifacts, such as blinks, have a similar compact temporal representation.

To avoid being biased by the artifact components in the estimation of the P300 waveform, methods based on spatio-temporal models have been introduced, e.g. [8,9]. Whilst [8] imposes a regular waveform comprising three free parameters (a waveform equivalent to the gamma distribution function with latency, amplitude and form factor as free parameters), [9] derives an iteratively refined waveform obtained through iterations of ICA and a selection procedure reminiscent to an expectation-maximisation. The estimation of the latency in [8] results in an exhaustive search over all possible delays. Moreover, the energy function used in the estimation of the latency needs to be sufficiently smoothened to alleviate spurious minima (a consequence of the noise), which requires the introduction of an extra, heuristically chosen parameter. In [9], the algorithm also requires a heuristically chosen parameter setting, namely a lower threshold on the correlation between the current estimate of the template and the individual estimates. It should be mentioned that both proposed algorithms only have an implicit coupling between the spatial and the temporal subspace and estimate

them in an alternating fashion. As a consequence, estimation errors in each of the spaces carry over to the other space.

From the above, we may observe that the attention given in literature to this field of research has increased considerably during the last decades. Despite this interest in the topic, the authors are not aware of any attempt to explicitly estimate the spatio-temporal signal model without either modelling the signal of interest [8] or using a spatial estimation model with posterior calculation of the corresponding temporal waveforms [10,11] or algorithms alternating between spatial and temporal estimates [9,8]. In this contribution we present a method based on a joint spatio-temporal estimation of the P300. Incorporating the spatial distribution and the temporal waveform in a single, non-iterative algorithm without heuristic parameter selection, we aim at a more robust estimator. In addition, since no particular waveform will be imposed, nor a characteristic spatial distribution, the proposed method can be easily adopted for solving similar estimation problems in two coupled, heterogeneous spaces.

2 Methods

2.1 Notational Conventions

Upper case boldface ($\mathbf{A}$) and lower case boldface ($\mathbf{a}$) characters will respectively denote matrices and column vectors. Scalars and constants will be denoted by lower case light face (a) and upper case light face (A) characters, respectively. The i-th column of $\mathbf{A}$ is thus $\mathbf{a}_i$ and the jth entry of $\mathbf{a}$, a_j. An ensemble of matrices $\mathbf{A}^{(k)}$ will be denoted by the calligraphic upper case $\mathcal{A} = \{\mathbf{A}^{(k)} \,|\, k = 1, 2, \ldots, K\}$. The symbol $\otimes$ denotes the tensor product (Kronecker product) and the symbol $(\cdot)^T$ will stand for the matrix transposition operator.

2.2 Model

We suppose to have K samples of a process $\mathbf{X}$ defined on two coupled heterogeneous spaces. To ease the presentation, we will from hereon consider that these spaces are the spatial and the temporal dimensions typical of the P300, where the temporal dimension is taken relative to the stimulus onset. The K samples then simply represent the trials. The process realisations $\mathbf{X}^{(k)} \in \mathbb{R}^{M \times N}$ are matrices, with entries x_{mn} referred to their spatial index m, the arbitrarily chosen electrode index, and their temporal index n relative with respect to the stimulus onset[1]. Suppose the following generative model for $\mathbf{X}$:

$$\mathbf{X} = \begin{cases} \sigma_1 \boldsymbol{\upsilon}\boldsymbol{\nu}^T + \sum_{i=2}^{L} \sigma_i \boldsymbol{\alpha}_{i-1}\boldsymbol{\beta}_{i-1}^T & \text{with probability } p \\ \sum_{i=1}^{L} \sigma_i \boldsymbol{\mu}_i \boldsymbol{\zeta}_i^T & \text{with probability } 1-p \end{cases}, \tag{1}$$

[1] Although it is common to choose n monotonically increasing with respect to the relative time after onset, this is not necessary for the algorithm to function well. The only pre-requisite is that there is a bijective relation between the relative time after onset and the temporal index.

which says that with probability p we observe the spatio-temporal structure $\mathbf{u}\mathbf{v}^T$ up to some angular noise on $\mathbf{u}$ and $\mathbf{v}$, yielding $\boldsymbol{\upsilon}\boldsymbol{\nu}^T$. Eq. (1) is completely defined by further fixing

$$\begin{array}{ll} \boldsymbol{\upsilon} \sim P_{VMF}(\mathbf{u}, \kappa) & \boldsymbol{\nu} \sim P_{VMF}(\mathbf{v}, \kappa) \\ \boldsymbol{\alpha}_i \sim P_{VMF}([\mathbf{U}^\perp]_i, \kappa_2) & \boldsymbol{\beta}_i \sim P_{VMF}([\mathbf{V}^\perp]_i, \kappa_2) \\ \boldsymbol{\mu}_i \sim P_{VMF}(\mathbf{q}_i, \kappa_2) & \boldsymbol{\zeta}_i \sim P_{VMF}(\mathbf{r}_i, \kappa_2) \end{array} \tag{2}$$

and for σ_i any (non-degenerate) distribution on the (positive) real numbers may be chosen. For our convenience we have chosen the uniform distribution on $(0, 1)$ in the simulations. $\mathbf{U}^\perp \in \mathbb{R}^{M\times(L-1)}$ and $\mathbf{V}^\perp \in \mathbb{R}^{N\times(L-1)}$ are random vectors constraint to form an orthogonal basis for the orthogonal complement of $\mathbf{u}$ and $\mathbf{v}$, respectively; $\mathbf{q}_i$ and $\mathbf{r}_i$ are L vectors from an orthogonal basis for $\mathbb{R}^M$ and $\mathbb{R}^N$, respectively. $P_{VMF}(\boldsymbol{\mu}, \kappa)$ stands for the Von Mises-Fischer probability distribution function with mean $\boldsymbol{\mu} \in \mathbb{S}^{N-1}$ and form parameter κ (in particular, $\kappa = 0$ means a uniform distribution on the hypersphere $\mathbb{S}^{N-1}$ and $\kappa \to +\infty$ means an improper Dirac distribution at $\boldsymbol{\mu}$). Remark that we have only chosen two different values for all κ's (κ and κ_2) to simplify the representation (and the subsequent simulations), however, κ might differ for each of the vectors in the model.

For the specific case of the P300 the above means we suppose that the spatio-temporal pattern – up to some directional noise – is present with probability p in our trials. We will detail this further in the next paragraph.

2.3 Representation of a Matrix

The samples, our trials, simply form a collection of observed matrices $\mathcal{X} = \{\mathbf{X}^{(k)}\}_{k=1}^K$. These matrices can be represented by a multitude of equivalent forms, the most straightforward being

$$\mathbf{X}^{(k)} = \begin{pmatrix} x_{11}^{(k)} & x_{12}^{(k)} & \dots & x_{1N}^{(k)} \\ x_{21}^{(k)} & x_{22}^{(k)} & \dots & x_{2N}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{M1}^{(k)} & x_{M2}^{(k)} & \dots & x_{MN}^{(k)} \end{pmatrix} \in \mathbb{R}^{M\times N} \quad .$$

In fact, for matrices, being second order tensors defined on two (possibly heterogeneous) spaces (space and time for the EEG), we have that each element corresponds to a spatial position (the electrode index m) and a relative time (temporal index n). If we take the respective canonical bases $\mathcal{E}^s = \{\mathbf{e}_m^s\}_{m=1}^M$ and $\mathcal{E}^t = \{\mathbf{e}_n^t\}_{n=1}^N$ for each of these spaces, the coefficients of $\mathbf{X}^{(k)}$ with respect to the (tensor product) basis

$$\left(\mathbf{e}_1^s \otimes \mathbf{e}_1^t, \mathbf{e}_2^s \otimes \mathbf{e}_1^t, \dots, \mathbf{e}_M^s \otimes \mathbf{e}_1^t, \mathbf{e}_1^s \otimes \mathbf{e}_2^t, \mathbf{e}_2^s \otimes \mathbf{e}_2^t, \dots, \mathbf{e}_M^s \otimes \mathbf{e}_N^t\right)$$

are given as $\left(x_{11}^{(k)}, x_{21}^{(k)}, \dots, x_{M1}^{(k)}, x_{12}^{(k)}, x_{22}^{(k)}, \dots, x_{MN}^{(k)}\right)^T \in \mathbb{R}^{M\cdot N}$.

From the theory of matrix algebra, each matrix can be represented on a maximum of $L = \min(M, N)$ basis vectors (in the product space). A possible reduced representation can be found through the singular value decomposition of $\mathbf{X}^{(k)}$ (where we have dropped the superscripts in what follows to augment readability)

$$\mathbf{X} = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Psi}^T$$

from which we obtain that the coordinates of $\mathbf{X}$ in its specific basis

$$(\boldsymbol{\phi}_1 \otimes \boldsymbol{\psi}_1, \boldsymbol{\phi}_2 \otimes \boldsymbol{\psi}_2, \ldots \boldsymbol{\phi}_L \otimes \boldsymbol{\psi}_L, \boldsymbol{\phi}_2 \otimes \boldsymbol{\psi}_1 \ldots \boldsymbol{\phi}_M \otimes \boldsymbol{\psi}_N)$$

are given by $\left(\lambda_1, \lambda_2, \ldots, \lambda_L, \underbrace{0, 0, \ldots 0}_{M \cdot N - L}\right)$. In other words, the matrix $\mathbf{X}^{(k)}$ lies in a L-dimensional subspace of $\mathbb{R}^{M \cdot N}$. This representation will form the basis of the spatio-temporal method that will be developed in the rest of this contribution.

2.4 Some Useful Properties of the Tensor Product Basis

The kronecker product is nowadays a widely used operation in the manipulation of multi-way arrays and tensors and it is known to have some attractive properties (see e.g. [12]), especially for our application.

Property 1 (Transposability of Orthogonality) *If ϕ_m and $\phi_{m'}$ are two orthogonal vectors in one of the two spaces, then the vectors $\phi_m \otimes \psi_n$ and $\phi_{m'} \otimes \psi_n$ are orthogonal vectors in the product space.*

It follows from the distributivity and the associativity of the tensor product that $\langle \phi_m \otimes \psi_n, \phi_{m'} \otimes \psi_{n'} \rangle = \langle \phi_m, \phi_{m'} \rangle \langle \psi_n, \psi_{n'} \rangle$, or, the correlation in the product space is the product of the correlations in the respective spaces. In other words, to have a high correlation in product space, the correlation in both spaces should be high.

Practically, we have that the P300 template should be a spatial as well as a temporal representative and it clearly does not suffice to be a representative in only a single one of these spaces. In addition, for activity other than the evoked potential to be captured in a spatio-temporal representation, it should correlate spatially as well as temporally over the different trials. In other words, background activity may be spatially correlated as long as it is temporally uncorrelated and vice versa.

2.5 Angles Between Subspaces

From the representation of $\mathbf{X}^{(k)}$ discussed in section 2.3, we may retain the basis for the subspace on which $\mathbf{X}^{(k)}$ is defined, rather than its coefficients. For two matrices $\mathbf{X}^{(k)}$ and $\mathbf{X}^{(k')}$ we could then try to find the best common representative. Putting $\boldsymbol{\Xi}^{(k)} = \left(\boldsymbol{\xi}_1^{(k)}, \boldsymbol{\xi}_2^{(k)}, \ldots \boldsymbol{\xi}_L^{(k)}\right)$ and $\boldsymbol{\Xi}^{(k')} = \left(\boldsymbol{\xi}_1^{(k')}, \boldsymbol{\xi}_2^{(k')}, \ldots \boldsymbol{\xi}_L^{(k')}\right)$

as some orthonormal subspace bases for $\mathbf{X}^{(k)}$ and $\mathbf{X}^{(k')}$ respectively, we could define the angle between subspaces, analogously as in [13], as

$$\cos(\theta_i) = \max_{\boldsymbol{\omega}_i, \boldsymbol{\rho}_i} \boldsymbol{\omega}_i^T {\boldsymbol{\Xi}^{(k)}}^T \boldsymbol{\Xi}^{(k')} \boldsymbol{\rho}_i, \text{ subject to } \begin{matrix} \langle \boldsymbol{\omega}_i, \boldsymbol{\omega}_{i'} \rangle = 0 \\ \langle \boldsymbol{\rho}_i, \boldsymbol{\rho}_{i'} \rangle = 0 \end{matrix}, \forall i' < i \ .$$

The angle between the subspaces $\boldsymbol{\Xi}^{(k)}$ and $\boldsymbol{\Xi}^{(k')}$ is defined to be θ_L. It can be demonstrated that $0 \leq \cos(\theta_i) \leq 1$ and the index i for which $1 = \cos(\theta_i) > \cos(\theta_i + 1)$ determines the dimension of the common subspace.

However, the angle between subspaces does not easily extend to multiple matrices if we do not restrict ourselves to pairwise comparisons. But we are looking for a common subspace to all observations (i.e. the intersection of all subspaces). Unfrotunately, it suffices that there exists a single observation that does not comprise the subspace of interest to have the zero vector as the resulting intersection. In other words, a single trial not containing a P300 response to the stimulus would compromise the analysis.

2.6 Best Representative Subspace

To avoid the above drawback, we use the following trick:

1. Compose $\mathbf{T} = \left(\boldsymbol{\Xi}^{(1)}, \boldsymbol{\Xi}^{(2)}, \ldots, \boldsymbol{\Xi}^{(k)}, \ldots \boldsymbol{\Xi}^{(K)} \right)$
2. Calculate the major singular value $\lambda_1(\mathbf{T})$ of $\mathbf{T}$ and its corresponding left and right eigenvectors $\boldsymbol{\rho}$ and $\boldsymbol{\omega}$
3. Calculate the best rank-one approximation to the *vectorized matrix* $\boldsymbol{\rho}$
4. Calculate $\tilde{\omega}_k = \sum_{i=(k-1)L+1}^{kL} \omega_i^2$

The best rank one approximation to $\boldsymbol{\rho}$ may be written as $\lambda_1(\boldsymbol{\rho})\, \hat{\mathbf{u}} \otimes \hat{\mathbf{v}}$. Concerning the vector $\tilde{\boldsymbol{\omega}}$, remark that its norm $\|\tilde{\boldsymbol{\omega}}\|_1 = 1$ and that the weights $\tilde{\omega}_\ell$ thus form a partition of 1. In fact, $\lambda_1(\mathbf{T})$ may be seen as an approximation to p in Eq. (1) and $\tilde{\omega}_\ell$ the relative probability that the ℓ-th observation adheres to the upper equation of the model in Eq. (1).

3 Results

In this section we will display the results obtained on synthetic data only, due to a lack of space. All data in the simulations have been generated in accordance to the model in Eq. (1). We have taken $\mathbf{u} \in \mathbb{R}^3$ and $\mathbf{v} \in \mathbb{R}^5$, $K = 25$ and have run 100 Monte Carlo realisations.

In the presentation of the results we have taken the mean over the different values of κ_2 since we have seen that its influence on the end results is less significant than that of p or κ. The results of this study are given in Figure 1(a) in terms of $\sqrt{\rho_{\mathbf{u}} \rho_{\mathbf{v}}} = \sqrt{\langle \hat{\mathbf{u}}, \mathbf{u} \rangle \langle \hat{\mathbf{v}}, \mathbf{v} \rangle}$, i.e. the geometric mean of the correlations, which is simply the square root of the inner product taken in the product space.

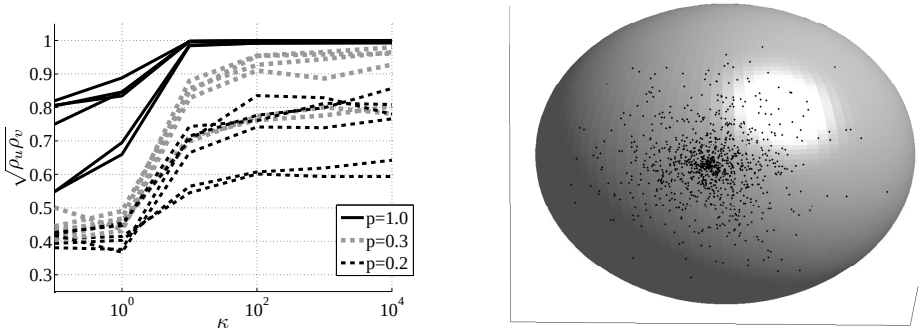

Fig. 1. (a) The mean performance over 100 Monte Carlo runs of the algorithm when varying κ and p in the model of Eq. (1). $\kappa_2 = 10^{-1}, 10^0 \dots 10^4$ with lower values in κ_2 resulting in a line closer to the x-axis. (b) 1000 samples from the Von Mises-Fisher distribution on the sphere $\mathbb{S}^2$ for $\kappa = 10$ (the knee in Figure (a)).

4 Discussion

From the Figures 1(a)&1(b), we observe that the performance degrades with an increase in the corruption probability $(1 - p)$. Also, we clearly have a similar behaviour of the correlation as a function of κ. We observed also that for fixed p and κ, augmenting the number of observations results in a better estimation performance (results not shown), which supports the assumption that the used statistics are consistent.

Since in practice the P300 does not correlate spatio-temporally with the background activity and noise, the proposed methodology is promising in the estimation of the P300 and related waveforms. Preliminary results on real data (not shown here) confirm this assumption. The above assumption contrasts with those of the principal component analysis of the observed matrices or that of the best rank-one approximation to the three way array $\mathcal{X}$ composed by stacking matrices $\mathbf{X}^{(k)}$, since the latter algorithms are variance rather than occurrence based. It is in this perspective that our method also resembles (kernel-based) clustering methods. However, in the latter, each of the vectors in $\mathbf{T}$ is attributed a cluster index, while the authors are not aware of any attempt to jointly estimate the best joint rank-1 matrix representation for a *clustering* of the matrices.

Note that the spatio-temporal pattern that results from the method does not need to occur in all observations, even not in one single observation. From Figure 1(a), we observe that acceptable performance is already achieved for $\kappa \gtrsim 10$ and $p > 0.4$. Actually, it suffices that the chosen spatio-temporal pattern is the closest to the majority of the subspaces spanned by the observations. This is reminiscent to the largest singular value ($0 \leq \cos(\theta_1) \leq 1$) that may be found

in the algorithm calculating angles between subspaces [13], an angle whose cosine does not necessarily equal one.

5 Conclusion

The proposed method seems to be promising for estimating a pattern in a product space of two heterogeneous spaces as is the case in the spatio-temporal P300 estimate. The method results in a direct spatio-temporal decomposition, rather than a spatial decomposition with posterior temporal estimation as often witnessed in literature. Moreover, instead of imposing independence, decorrelation, orthogonality or sparsity in one or both of the heterogeneous spaces, the method only relies on the re-occurrence of a spatio-temporal pattern in a subset of the observations, a pattern that may be subjected to noise (partial re-occurence). In addition, the algorithm has an algebraic solution without heuristic parameters to choose.

References

1. Hoke, M., Ross, B., Wickesberg, R., Lütkenhöner, B.: Weighted averaging theory and application to electric response audiometry. Electroencephalography and Clinical Neurophysiology **57**(5) (1984) 484–489
2. Davila, C.E., Mobin, M.S.: Weighted averaging of evoked potentials. IEEE Transactions on Biomedical Engineering **39**(4) (1992) 338–345
3. Lütkenhöner, B., Hoke, M., Pantev, C.: Possibilities and limitations of weighted averaging. Biological Cybernetics **52**(6) (1985) 40–416
4. Chapman, R.M., McCrary, J.W.: EP component identification and measurement by principal component analysis. Brain and Cognition **27** (1995) 288–310
5. Kayser, J., Tenke, C.E.: Optimizing PCA methodology for ERP component identification and measurement: theoretical rationale and empirical evaluation. Clinical Neurophysiology **114** (2003) 2307–2325
6. Comon, P.: Independent component analysis, a new concept? Signal Processing **36** (1994) 287–314
7. Makeig, S., Westerfield, M., Jung, T.P., Covington, J., J, T., T, S.: Independent components of the late positive response complex in a visual spatial attention task. Journal of Neuroscience **19** (1999) 2665–2680
8. Li, R., Keil, A., Principe, J.C.: Single-trial P300 estimation with a spatiotemporal filtering method. Journal of Neuroscience Methods **177** (2009) 488–496
9. Iyer, D., Zouridakis, G.: Single-trial evoked potential estimation: Comparison between independent component analysis and wavelet denoising. Clinical Neurophysiology **118** (2007) 495–504
10. Wang, Y., Berg, P., Scherg, M.: Common spatial subspace decomposition applied to analysis of brain responses under multiple task conditions: a simulation study. Clinical Neurophysiology **110** (1999) 604–614
11. Krusienski, D.J.: A method for visualizing independent spatio-temporal patterns of brain activity. EURASIP Journal on Advances in Signal Processing **2009** (2009)
12. Laub, A.J.: Matrix Analysis for Scientists and Engineers. Society for Industrial and Applied Mathematics (2005)

13. Golub, G.H., Van Loan, C.F.: Matrix Computations. 3rd edn. The Johns Hopkins University Press (1996)