Invited Paper

Non-normality in Neural Networks

Ankit Kumar[a, b, c] and Kristofer Bouchard[b, c, d, e]

[a]Department of Physics, University of California, Berkeley, CA 94720
[b]Redwood Center for Theoretical Neuroscience, University of California, Berkeley, CA, 94720
[c]Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720
[d]Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720
[e]Hellen Wills Neuroscience Institute, University of California, Berkeley, CA, 94720

ABSTRACT

A central goal of neuroscience is to link synaptic connectivity of neural circuits to produced dynamics and computations. Anatomical and functional connectivity within neural systems is asymmetric, which upon linearization gives rise to non-normal dynamics. Particular linear combinations of neurons that are involved in circuit function are canonically identified in systems neuroscience via PCA, which seeks subspaces which maximize variance. We have recently proposed Dynamical Components Analysis (DCA), which seeks subspaces of activity in which the mutual information between past and future activity (i.e., 'the dynamic memory') is largest. Here, we show that the presence of non-normality leads to a divergence between these subspaces and consequently, the importance of single neurons that are identified by each method. Applied to *in-vivo* electrophysiology recordings from diverse brain areas, subspaces of past-future mutual information are better able to predict animal and human behavior than subspaces of high variance. Finally, we discuss possible consequences of non-normality for the training and function of *in-silico* recurrent neural networks.

Keywords: Dimensionality Reduction, Neuroscience, Recurrent Neural Networks

1. INTRODUCTION

The brain performs sophisticated computations through the complex, highly recurrent dynamics of networks of neurons.[1] Unravelling the principles by which the structure of these biological neural networks gives rise to dynamics, and in turn to function, is central not only to neuroscience, but also to fields such as artificial intelligence and neuromorphic computing that seek to emulate these principles. Perhaps the most salient structural constraint on biological neural networks is known as Dale's Law, which requires that a given neuron projects either excitatory or inhibitory synapses, but not both. Recent studies of the patterns of synaptic connectivity in cortex have also revealed that asymmetric connectivity motifs are commonplace.[2] The consequence of these basic anatomical facts is that the synaptic connectivity matrix is non-normal. Dynamics driven by non-normal operators exhibit rich, non-trivial transient behavior. These transients have diverse consequences for phenomena in fluid dynamics, optical physics, and ecology.[3–5] However, despite their presence being a necessary consequence of neural anatomy, the implications of non-normal dynamics for the function of biological and biologically inspired neural networks remain poorly understood.[6] Here, we explore the consequences of non-normality on the transmission of information through subspaces of collective dynamics in biological neural networks.

A robust result that holds across diverse brain areas is that the distinct computations and behaviors are produced through collective activity unfolding within distinct subspaces with dimensionality much less than the ambient dimensionality (i.e. the number of neurons in a given circuit or brain region). Rigorous analysis of brain function therefore requires characterizing the normative properties of dynamics within subspaces that support particular computations. Methodologically, the simplest approach to extracting subspaces of collective

Further author information: (Send correspondence to K.B.)
K.B.: E-mail: KEBouchard@lbl.gov

AI and Optical Data Sciences III, edited by Bahram Jalali, Ken-ichi Kitayama, Proc. of SPIE Vol. 12019, 120190A
 · 0277-786X · doi: 10.1117/12.2613472

population activity from multivariate neural recordings are linear dimensionality reduction methods.[7] The workhorse of this family of methods in systems neuroscience is principal components analysis (PCA), which extracts subspaces of highest variance in firing rates. Yet, in order to give rise to complex behaviors, neural circuits must generate activity with complex temporal dependencies, not necessarily activity with high variance. The capacity of a dynamical system to retain information within its recurrent activity can be made precise using the predictive information (PI), which measures the mutual information shared between the past and future of a time series.[8] We previously introduced Dynamical Components Analysis (DCA)[9] to identify linear subspaces with highest PI. In this work, we show that the degree of non-normality tunes the divergence between PCA and DCA components. In electrophysiology recordings form macaque motor cortex, rat hippocampus, and human speech areas, we demonstrate that behavioral variables can be better predicted from DCA components than PCA components, indicating that predictive information is a better normative measure of subspace dynamics relevant to circuit function.

Mathematical Notation: We use boldface type $\mathbf{x}$ to indicate vector valued variables. $\lambda(A)$ and $\sigma(A)$ refer to eigenvalues and singular values, respectively. Angled brackets $\langle\cdot\rangle$ denote the expectation with respect to a probability measure, and $||\cdot||_F, ||\cdot||_{\text{op}}$ denote the Frobenius and operator norms, respectively.

2. HIGH PI VS. HIGH VARIANCE SUBSPACES IN NON-NORMAL LINEAR SYSTEMS

Let $\mathbf{x}_t$ be the multivariate firing rates of a population of N neurons indexed by time t. Let $\{\Lambda_0, \Lambda_1, \Lambda_2, ...\Lambda_T\}$ be the corresponding (assumed stationary) autocovariance sequence $\Lambda_\tau = \langle \mathbf{x}_j \mathbf{x}_{j+\tau}^\top \rangle$. The predictive information (PI) measures the mutual information contained between the past stochastic trajectories of a time series over timescale T with the future trajectories over timescale T. As $T \to \infty$, the PI may be identified with the subextensive growth of the entropy of the time series with T.[8] Assuming the firing rates are well modeled by Gaussian processes, the predictive information measures the extent to which the future of the time series is linearly predictable from the past (i.e. "self-predictive"), and can be expressed, up to constants, in terms of the autocovariance sequence:

$$\text{PI(T)} = I(\mathbf{x}_{-T}, \mathbf{x}_{-T+1}..., \mathbf{x}_{-1} | \mathbf{x}_0, ..., \mathbf{x}_{T-2}..., \mathbf{x}_{T-1}) \approx \log\det \Sigma_{2T} - 2\log\det \Sigma_T \qquad (1)$$
$$[\Sigma_T]_{ij} = \Lambda_{|i-j|}, \{i, j = 1, ..., T\}$$

where Σ_T is the spatiotemporal covariance matrix over timescale T. In what follows, we will also be interested in state space parameterizations of the dynamics. That is, we model the firing rates as arising from linear, Markovian dynamics, and dynamics within latent subspaces, $\mathbf{y}_t$ as arising from linear readouts of the firing rates:

$$\mathbf{y}_t = C\mathbf{x}_t \quad \mathbf{x}_{t+1} = A\mathbf{x}_t + B\mathbf{w}_t \quad \mathbf{w}_t \sim \mathcal{N}(0, I_{N\times N}), A \in \mathbb{R}^{N\times N}, B \in \mathbb{R}^{N\times N}, C \in \mathbb{R}^{d\times N} \qquad (2)$$

A is stable if $|\lambda_{max}(A)| < 1$. In this case, further assuming $BB^\top > 0$, Λ_0, the state covariance is positive semidefinite and arises from the solution of the Lyapunov equation: $\Lambda_0 - A\Lambda_0 A^\top = BB^\top, \Lambda_0 = \sum_{k=0}^\infty A^k BB^\top \left(A^\top\right)^k$. A is non-normal if $AA^\top \neq A^\top A$. For non-normal A, it is possible for $\lambda_{\max}(A_S) := \lambda_{\max}\left(\frac{A+A^\top}{2}\right) > 1$ even if A is a stable matrix. In this case, the system dynamics exhibit transient amplification, whereby $||A^k||$ initially increases with k, before asymptotically decaying.[10] The rank d PCA subspace corresponds to the space spanned by eigenvectors corresponding to the largest d eigenvalues of Λ_0. For non-normal systems, then, PCA identifies subspaces exhibiting the largest total amplification of inputs.

By contrast, DCA seeks out subspaces within (i.e. readouts of) multivariate time series that contain maximum (Gaussian) predictive information. In terms of the autocovariance sequence, the rank-d DCA subspace solves $\text{argmax}_C PI(T, C)$, where $PI(T, C)$ is obtained from eqs. 1-2 by replacing $\Lambda_{|i-j|} \in \mathbb{R}^{N\times N}$ with the projected autocorrelation matrix $C\Lambda_{|i-j|}C^\top \in \mathbb{R}^{d\times d}$. When the state space parameterization (eq. 2) is used to model the data generation process, the predictive information can be expressed in terms of the state variances associated

with causal and acausal Kalman filters of the firing rates. These causal and acausal Kalman filters exhibit state space dynamics of the form:

$$\begin{aligned} \hat{\mathbf{x}}_{t+1} &= (A - KC)\hat{\mathbf{x}}_t + K\mathbf{y}_t \\ \tilde{\mathbf{x}}_t &= (A^\top - \bar{K}C)\hat{\mathbf{x}}_{t+1} + \bar{K}\mathbf{y}_{t+1} \end{aligned} \tag{3}$$

where $\hat{\mathbf{x}}_t$ and $\tilde{\mathbf{x}}$ denote causal and acausal minimum mean square estimates (MMSE) of $\mathbf{x}_t$ given observations $\mathbf{y}_t = C\mathbf{x}_t$, and $K, \bar{K}$ are the associated Kalman gains associated from solution of dual Riccati equations. We denote the state covariance of the causal (acausal) Kalman filter with zero initial conditions after T observations as P_T (Q_T). Then, the PI may be expressed as:[11]

$$PI(T, C) = -\frac{1}{2}\log\det I - P_T Q_T \tag{4}$$

For a discrete time stochastic process whose spectral density ϕ satisfies $\frac{1}{2\pi}\int_{-\pi}^{\pi}\log\det\phi(e^{i\theta})d\theta > -\infty$, $||P_T Q_T||_{\text{op}} < 1$, guaranteeing eq. 4 is also finite.[12] Processes that satisfy this condition are known as being purely nondeterministic in time series analysis, which roughly coincides with the process having finite memory of its own past.

We now present our first main result:

Theorem 1. For symmetric, stable A and $B = I$, the subspaces found by PCA and DCA coincide, and are given by the subspaces spanned by the eigenvectors corresponding to the largest eigenvalues of A.

Proof See Appendix.

Thus for normal dynamical systems, direction of state space with maximum variance and with maximum memory of their past coincide. For non-normal systems, this equivalence no longer holds. We measure the divergence in angle between the leading PCA and DCA direction as we increase the nonnormality vis a vis the asymmetry of entries in randomly generated A matrices (**Fig. 1C**, red). We also measure the spearman rank correlation between the loadings onto PCA and DCA subspaces (**Fig. 1C**, black). If $C \in \mathbb{R}^{dN}$ is the matrix that projects onto the subspace of interest, then the loadings are defined as the norm of the rows of this matrix, and measure the contribution of ambient degrees of freedom (i.e., neurons) onto latent degrees of freedom. As the nonnormality is increased, the subspace angle between PCA and DCA increases, and the correlation between loadings decreases.

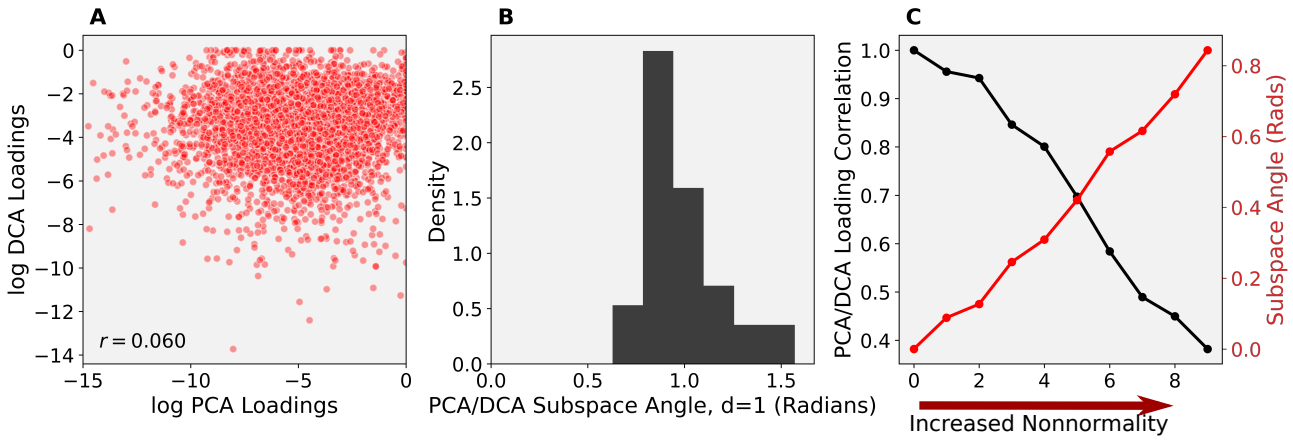

Figure 1: Spearman r between loadings onto (black) and subspace angles between (red) PCA/DCA subspaces averaged across many random synthetic systems as the asymmetry between entries in the matrix A, and therefore the nonnormality, is increased.

3. RESULTS IN NEURAL DATA

We next applied PCA and DCA to electrophysiological recordings from rat hippocampus (HPC), macaque motor cortex (M1), and human ventral sensorimotor cortex (vSMC), an area associated with speech production. In the former two datasets, we obtain firing rates by binning recorded spikes and applying a boxcox tranformation to the binned counts. The latter dataset is comprised of preprocessed voltages recorded from the surface via electrocorticography (ECoG).[13]

We measure the functional relevance of a subspace by linearly predicting behavioral variables (animal position for HPC, reach velocity for M1, and syllable spoken for vSMC). As compared to PCA subspaces at the same dimensionality, behavior can be better decoded from DCA subspaces across brain areas (**Fig. 2**). The DCA decoding vs. dimension curve exhibits a more prominent elbow, indicating that the method better uncovers low dimensionality. DCA/PCA subspaces angle are consistently large (average of 0.98 radians, **Fig. 1A**) and single unit loadings are nearly uncorrelated (spearman $r = 0.06$, **Fig. 1B**). These results indicate that subspaces and single units that are variance-maximizing strongly diverge from those with maximum predictive information, a hallmark of nonnormality. Importantly, the latter contain dynamics more relevant to behavior.

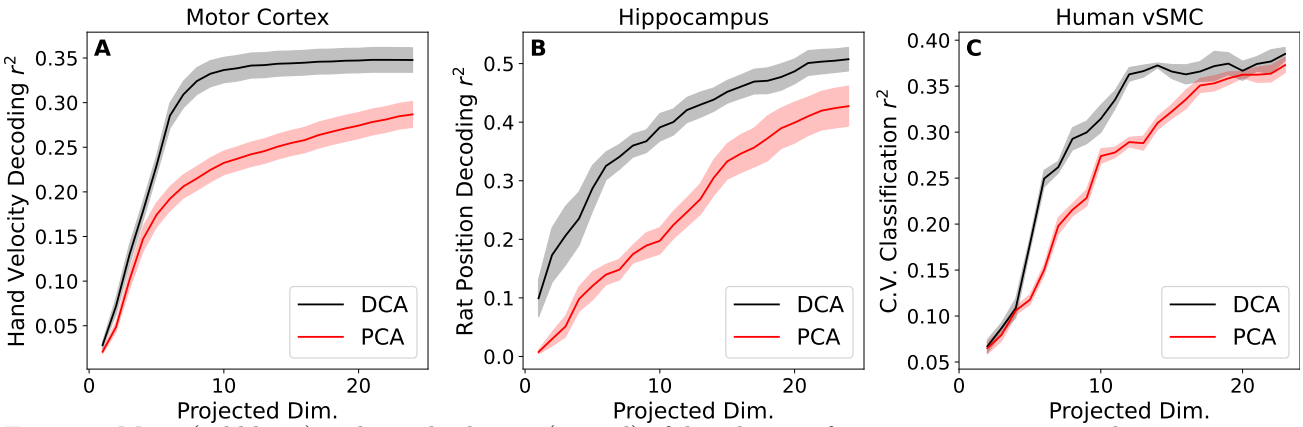

Figure 2: Mean (sold lines) and standard error (spread) of decoding performance vs. projection dimension across recording sessions (28 for M1, 8 for HPC, 1 for vSMC) and 5 folds of the data.

We then sought to characterize the features of single units constituting PCA vs. DCA subspaces. A canonical measure of the importance of single neurons to circuit function is the extent to which their firing rates can be predicted from external, behavioral variables (i.e., an encoding model), and their weight in a linear regression onto behavioral variables from the full population activity (i.e., a decoding model). We sought to determine the extent to which participation in a PCA or DCA subspace can be predicted from simple marginal statistics. To this end, we calculated spearman rank correlations of PCA/DCA loadings with single unit variance, single unit PI, weight in a linear decoder, and encoding r^2 (**Fig. 3**). We find that high loadings in PCA subspaces can be almost entirely predicted from single unit predictive information in M1 ((**Fig. 3A**) and single unit variance in Hippocampus (**Fig. 3B**). In M1, PCA loadings also correlate more strongly with single unit encoding and decoding statistics. By contrast, DCA loadings (black/grey bars) show no correlation with single unit variance or PI, and little correlation with decoding and encoding statistics. Subspaces with long dynamical memories, as measured by DCA, therefore reflect non-trivial collective dynamics that cannot be recapitulated from single unit features.

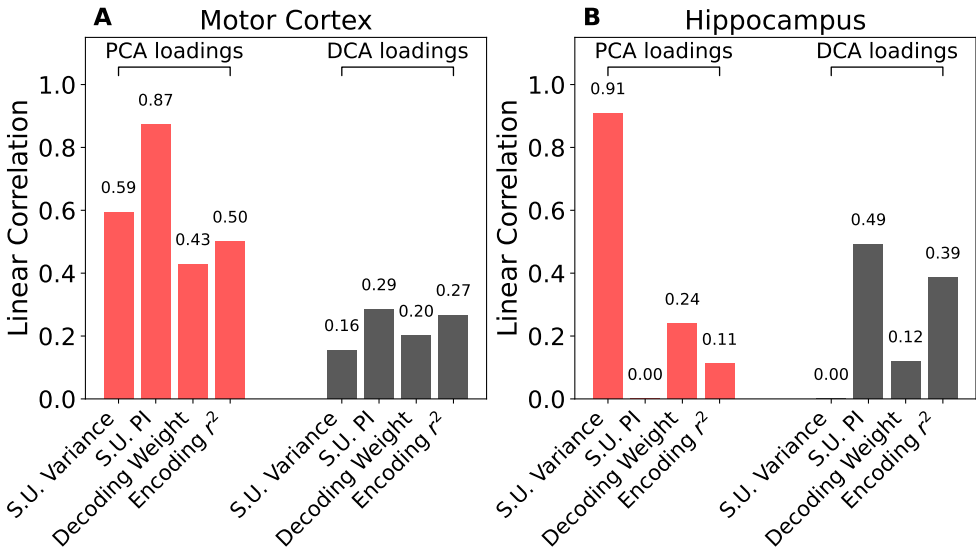

Figure 3: Average spearman r between PCA loadings (red) and DCA loadings (grey) and single unit variance, single unit PI (red), weight in a population decoder of animal behavior, and r^2 of a linear encoder of single unit neural activity of behavioral variables.

4. CONCLUSION AND OUTLOOK

In this work, we explored the interplay between non-normality, variance, and predictive information in dynamical systems. Predictive information quantifies the persistence of knowledge of the past state of a dynamical system into the future, and can therefore be interpreted as a dynamical memory. The persistence of information in memory, however, is not the only computational function required of biological neural networks. Prior work in the neuroscience literature has identified transient amplification arising from non-normality to be a possible mechanism underlying amplification of relevant patterns within visual stimuli.[6] Crucially, in a non-normal system, the subspaces in which dynamics exhibit maximum variance (i.e. amplification) and long term memory are distinct, as revealed by the divergence between PCA and DCA directions. A normative function of anatomical constraints that give rise to nonnormality may therefore be to enable distinct types of computations within distinct subspaces even at the level of linear dynamics. In our future work, we will explore this hypothesis across diverse brain areas and behavioral/task conditions.

Our identification of the importance of long dynamical memories in neural circuit function parallels the importance of enforcing orthogonal weights in the successful training of artificial recurrent neural networks. Training of RNNs is typically done through backpropagation through time, which is well known to suffer from exploding or vanishing gradients. To a linear approximation, orthogonal dynamics matrices enable gradients to neither vanish nor diverge, but remain constant with respect to network states arbitrarily far in the past. This favorable property has been exploited in several works to achieve state of the art results on sequence prediction tasks.[14] More recent work has also shown that more general non-normal weight matrix architectures can outperform purely orthogonal dynamics in tasks[15] that involve computations beyond memorization of past inputs. These results similarly suggest a functional role for the divergence between directions of long memory traces and directions of maximum amplification of inputs in artificial recurrent neural networks.

REFERENCES

[1] Vyas, S., Golub, M. D., Sussillo, D., and Shenoy, K. V., "Computation Through Neural Population Dynamics," *Annual Review of Neuroscience* **43**, 249–275 (July 2020). Publisher: Annual Reviews.

[2] Schröter, M., Paulsen, O., and Bullmore, E. T., "Micro-connectomics: probing the organization of neuronal networks at the cellular scale," *Nature Reviews Neuroscience* **18**, 131–146 (Mar. 2017).

[3] Schmid, P. J., "Nonmodal Stability Theory," *Annual Review of Fluid Mechanics* **39**, 129–162 (Jan. 2007). Publisher: Annual Reviews.

[4] Asllani, M. and Carletti, T., "Topological resilience in non-normal networked systems," *Physical Review E* **97**, 042302 (Apr. 2018). Publisher: American Physical Society.
[5] El-Ganainy, R., Khajavikhan, M., Christodoulides, D. N., and Ozdemir, S. K., "The dawn of non-Hermitian optics," *Communications Physics* **2**, 37 (Mar. 2019).
[6] Murphy, B. K. and Miller, K. D., "Balanced amplification: a new mechanism of selective amplification of neural activity patterns," *Neuron* **61**(4), 635–648 (2009). Publisher: Elsevier.
[7] Cunningham, J. P. and Yu, B. M., "Dimensionality reduction for large-scale neural recordings," *Nature neuroscience* **17**, 1500–1509 (Nov. 2014). Edition: 2014/08/24.
[8] Bialek, W., Nemenman, I., and Tishby, N., "Predictability, Complexity, and Learning," *Neural Computation* **13**, 2409–2463 (Nov. 2001).
[9] Clark, D., Livezey, J., and Bouchard, K., "Unsupervised Discovery of Temporal Structure in Noisy Data with Dynamical Components Analysis," in [*Advances in Neural Information Processing Systems*], Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F. d., Fox, E., and Garnett, R., eds., **32**, Curran Associates, Inc. (2019).
[10] Trefethen, L. N. and Embree, M., [*Spectra and pseudospectra*], Princeton university press (2020).
[11] Desai, U. B. and Pal, D., "A realization approach to stochastic model reduction and balanced stochastic realizations," in [*1982 21st IEEE Conference on Decision and Control*], 1105–1112, IEEE (1982).
[12] Bingham, N. H., "Szegö's theorem and its probabilistic descendants," *Probability Surveys* **9**, 287–324 (2012). Publisher: Institute of Mathematical Statistics and Bernoulli Society.
[13] Bouchard, K. E., Mesgarani, N., Johnson, K., and Chang, E. F., "Functional organization of human sensorimotor cortex for speech articulation," *Nature* **495**, 327–332 (Mar. 2013).
[14] Arjovsky, M., Shah, A., and Bengio, Y., "Unitary evolution recurrent neural networks," in [*International Conference on Machine Learning*], 1120–1128, PMLR (2016).
[15] Kerg, G., Goyette, K., Touzel, M. P., Gidel, G., Vorontsov, E., Bengio, Y., and Lajoie, G., "Non-normal Recurrent Neural Network (nnRNN): learning long time dependencies while improving expressivity with transient dynamics," *arXiv preprint arXiv:1905.12080* (2019).
[16] Kailath, T., Sayed, A. H., and Hassibi, B., [*Linear estimation*], no. BOOK, Prentice Hall (2000).

5. APPENDIX

Proof of Theorem 1 We recall that the transfer function associated with the model 2 for $B = I$ is $W(z) = C(Iz - A)^{-1}$. Order the eigenvalues of A as $\lambda_1 > \lambda_2 > ... > \lambda_N$. Let v_j be the eigenvector corresponding to λ_j. For simplicity, we assume that the eigenvalues of A are distinct. We will first show that $C^* = \sum_{j=1}^{d} v_j v_j^\top$, where $v_j v_j^\top$ is the projection onto the j^{th} eigenspace of A is the optimal solution to the PCA. Writing $A^k = \sum_{i=1}^{N} \lambda_i^k v_i v_i^\top$, it is straightforward to show that the state covariance can be expressed as:

$$\Lambda_0 = \sum_{i=1}^{N} \frac{1}{1-\lambda_i^2} v_i v_i^\top \tag{5}$$

PCA can formally be expressed as the optimization problem $\text{argmax}_C \ \text{Tr}\left(C\Lambda_0 C^\top\right)$ for $C \in \mathbb{R}^{d\times N}$. From the representation above and Ky Fan's principle (ref), it follows immediately that C^* is the optimal PCA subspace.

Next, we show that C^* is optimal for DCA. We require the state variance associated with the forward and reverse time Kalman filters. We first show that for symmetric A and $B = I$, these two Kalman filters coincide. For this, we require the following Lemma (Theorem 5.4.2 in[16]):

Lemma 1. Let eq. 2 be a forward time Markovian realization of y_t, with $\Lambda_0 > 0$. Then, a reverse time realization of the process is given by:

$$\mathbf{y}_t = C\mathbf{x}_t \quad \mathbf{x}_t^r = A^r \mathbf{x}_{t+1}^r + \mathbf{w}_t^r$$
$$A^r = \Lambda_0 A^\top \Lambda_0^{-1}, \mathbf{w}_t^r \sim \mathcal{N}(0, Q), Q = \Lambda_0 - \Lambda_0 A \Lambda_0^{-1} A \Lambda_0$$

For symmetric A and $B = I$, A and Λ_0 commute, so that $A^r = A$. Furthermore, using the representation eq. 5, one has that $Q = I$, and so the forward and reverse time realizations coincide. This leads to the following Lemma:

Lemma 2. For symmetric, stable A and $B = I$, the forward and reverse time Kalman filter dynamics coincide, i.e. $P_T = Q_T$.

Using Lemma 2, the PI can be seen to be equal to $-\frac{1}{2}\sum_{i=1}^{n}\log\left(1-\sigma_i(T)^2\right)$ where $\sigma_i(T)$ are the eigenvalues of P_T. Since A is symmetric, we may diagonalize it by a unitary transformation. This decouples the MIMO transfer function $W(z)$ into a set of single input, single output transfer functions that each describe the dynamical system $x_{t+1}^{(i)} = \lambda_i x_t^{(i)} + w_t$, $w_t \sim \mathcal{N}(0,1)$, where λ_i is the i^{th} eigenvalue of A. The problem of the selection of the optimal readout of rank d reduces to the selection of d of these SISO systems to pass through a Kalman filter. The state variances of the corresponding Kalman filters are given by iterations of a univariate Riccati equation:[16]

$$\sigma_i^2(T) = \lambda_i^2\sigma_i^2(T-1) - \left(\lambda_i\sigma_i^2(T-1) - \frac{\lambda_i}{1-\lambda_i^2}\right)^2\left(\sigma_i^2(T-1) - \frac{1}{1-\lambda_i^2}\right)^{-1} \qquad \sigma_i(0) = 0$$

Implicitly, each $\sigma_i(T)$ also depends on the choice of λ_i. We can show inductively that $\sigma_i(T) > \sigma_j(T)$ whenever $\lambda_i > \lambda_j$. First, we note that $\sigma_i(1)^2 = \frac{\lambda_i^2}{1-\lambda_i^2}$, which is an increasing function of λ_i for $0 < \lambda_i < 1$. Next, assuming that $\sigma_i(T-1) > \sigma_j(T-1)$ for $\lambda_i > \lambda_j$, we evaluate $\frac{d\sigma_i^2(T)}{d\lambda_i}$:

$$\frac{d\sigma_i^2(T)}{d\lambda_i} = \frac{2\lambda_i}{(1-\lambda_i^2)^2}$$

which is also an increasing function of λ_i for $0 < \lambda_i < 1$. Thus, each Riccati iteration preserves the ordering of the σ_i with respect to λ_i. We note that $\lim_{T\to\infty}\sigma_i^2(T) \leq \frac{1}{1-\lambda_i^2}$, as the steady state Kalman state variance is bounded from above by the steady state variance of the underlying dynamical system. In the eigenbasis of A, then, the optimal rank-d C matrix is comprised of the canonical basis vectors coinciding with the top d largest eigenvalues of A, yielding the desired result.