Nonnegative matrix factorization for coherent set identification by direct low rank maximum likelihood estimation

Robert M. Polzin[1], Ilja Klebanov[1], Nikolas Nüsken[2], and Péter Koltai[3]

[1]Institute of Mathematics, Freie Universität Berlin, Germany
[2]Department of Mathematics, King's College London, UK
[3]Department of Mathematics, University of Bayreuth, Germany

Abstract

We analyze connections between two low rank modeling approaches from the last decade for treating dynamical data. The first one is the coherence problem (or coherent set approach), where groups of states are sought that evolve under the action of a stochastic matrix in a way maximally distinguishable from other groups. The second one is a low rank factorization approach for stochastic matrices, called Direct Bayesian Model Reduction (DBMR), which estimates the low rank factors directly from observed data. We show that DBMR results in a low rank model that is a projection of the full model, and exploit this insight to infer bounds on a quantitative measure of coherence within the reduced model. Both approaches can be formulated as optimization problems, and we also prove a bound between their respective objectives. On a broader scope, this work relates the two classical loss functions of nonnegative matrix factorization, namely the Frobenius norm and the generalized Kullback–Leibler divergence, and suggests new links between likelihood-based and projection-based estimation of probabilistic models.

Keywords: low rank modeling, coherent sets, maximum likelihood, nonnegative matrix factorization, clustering, Markov state model.

MSC classification: 65F55, 62M05, 37M10, 15A23, 60J22.

1. Introduction

1.1. Motivation and contributions

One of the fundamental concepts in statistics, data science and machine learning is that seemingly complicated data has an underlying simpler structure; and filtering out this

structure is the crucial task of *model reduction*. Apart from a simpler representation of the data that requires less storage, the main advantages include robustness when used for predictions as well as interpretability of the low-dimensional features. Since data is often given in the form of matrices $A \in \mathbb{R}^{m\times n}$, the above task typically boils down to matrix factorizations of the form $A \approx BC$, where $B \in \mathbb{R}^{m\times r}$ and $C \in \mathbb{R}^{r\times n}$ are matrices of lower dimensionality, $r \ll \min\{m, n\}$; see [UHZ+16] for a comprehensive overview. In many applications, data is inherently nonnegative, and incorporating the nonnegativity constraint on the factors B and C can add to the interpretability of the low rank approximation, which explains the success of *nonnegative* matrix factorization (NMF) over the past two decades [LS99, SG08, WZ12, LD14, Gil20]. Stochastic matrices are a frequent occurrence of nonnegative matrices in applications. Often they arise as (or from) data matrices, because the entities encoded by rows and columns of these matrices are in some probabilistic relationship.

One such application, which will be the guiding example of this paper, is determining coherent sets of a dynamical system [FSM10, FLS10], as made precise in section 2. In a nutshell, for a left stochastic "transition" matrix $P \in \mathbb{R}^{m\times n}$, we seek partitions of $\{1, 2, \ldots, n\}$ such that random transitions as described by P from different partition elements remain "maximally distinguishable" in the sense that the image random variables are decorrelated. We call these partition elements "coherent".

Approximating the transition matrix P by a product $P \approx \lambda\Gamma$ of two left stochastic matrices λ, Γ of lower dimension can be interpreted as a (soft) clustering of the corresponding states and coherent pairs are then identifiable, even more so if $\Gamma \in \{0, 1\}^{r\times n}$ has binary entries, which corresponds to (hard) clustering. Such a factorization is precisely what is provided by direct Bayesian model reduction (DBMR), a specific NMF algorithm proposed by [GH17, GONH18] for the identification of reduced models directly from the data, i.e. without approximating the "full" transition matrix P in the first place. One of the main motivations for this paper is the application of DBMR to the coherence problem described above. In comparison to the "classical" approach popularized by Froyland and others [FSM10, FLS10], which relies on a truncated singular value decomposition (SVD) of the transition matrix P (after a suitable rescaling, cf. Algorithm 1), the DBMR approach holds the following promises:

- As mentioned above, the "full" transition matrix P need not be approximated. While such an approximation often relies on a tremendous amount of samples, [GH17] argue that the (comparatively few) matrix entries of the low rank approximation require far less data.
- Whilst the classical approach requires an ambiguous post-processing step to identify coherent sets, often performed by k-means clustering [Den17], the DBMR output provides the coherent sets directly through the matrix Γ, while the matrix λ acts as a reduced transition matrix acting on these compound states. The left stochasticity of λ guarantees a form of structure preservation that does not hold for the classical approach, where the 'reduced transition matrix' is typically not a stochastic matrix. In fact, its columns may not sum to one and it can even have negative entries.

Conceptually speaking, truncated SVD provides an optimal low rank approximation with respect to the Frobenius norm as asserted by the Eckart–Young–Mirsky theorem

[HE15, Theorem 4.4.7], whilst DBMR corresponds to a (relaxed) maximum likelihood estimate of (λ, Γ) and hence minimizes the Kullback–Leibler (KL) divergence between the full and the low rank model, see Remark 6 in section 4.2.

In other words, both SVD and DBMR can be viewed as providing a "low-complexity" approximation to the transition matrix P,

$$A^* \in \underset{A \text{ "low-complexity"}}{\arg\min} d(A, P), \tag{1}$$

where d alludes to a "distance-like" quantity between matrices, and the notion of low-complexity has to be made precise (in our context this will amount to low rank and, in the case of DBMR, implicit sparsity constraints). Building on this parallel and following [LS00], [Gil20, Section 1.2], truncated SVD and DBMR can be succinctly formulated as solutions to **Problem 1** and **Problem 2** below, respectively. From the perspective of matrix factorization, the Frobenius norm $\|A - BC\|_F$ and the (generalized) Kullback–Leibler divergence $\mathrm{D}_{\mathrm{KL}}(A \parallel BC)$ (essentially applied to the vector consisting of matrix entries [LS00]) are two of the most fundamental distances minimized within NMF (and beyond) for the approximation $A \approx BC$ discussed above [WZ12]. For this reason, our comparison of the classical approach to the coherence problem with the one by DBMR should be seen on a broader scale—we derive connections between these two central objectives of matrix factorization and make the following two main contributions (with the terminology yet to be made precise):

(i) We prove that the DBMR output corresponds to the composition of the full model P and an orthogonal projection Π; that is, $P\Pi = \lambda\Gamma$. Based on this insight we deduce that the "degree of coherence" contained in the low rank model bounds the degree of coherence contained in the full model from below.

(ii) We derive an inequality involving the two measures of distance between the full and the low rank model mentioned above—the Frobenius norm (for the SVD approach) on the one hand, and the Kullback–Leibler divergence (for DBMR) on the other hand. To our knowledge, this is the first quantitative relationship between these two classical objectives of matrix factorization. To this end, we prove and utilize a novel *Pinsker-type inequality*, which could be of independent interest.

Next, we turn to the discussion of related work, and the remainder of the paper is structured as follows. The necessary notation and the coherence problem are introduced in section 2, while the computationally relevant form of the coherence problem arises from a relaxation detailed in section 3. In section 4 we summarize the low rank modeling approach referred to as DBMR. The material in this first part of the paper is encapsulated in **Problem 1** and **Problem 2** and those serve as the starting point for the analysis in the subsequent second part. The novel contributions are derived in sections 5 and 6, then they are illustrated by numerical examples in section 7, followed by a conclusion in section 8.

1.2. Related Work

Coherent sets and canonical variables. The term "coherent set" stems from fluid dynamics and dynamical systems [FSM10, FLS10], and the concept has been preceded by transport-related considerations around the term "coherent structures" (see the references in these papers). The abstract linear-algebraic problem, that the coherence problem boils down to in our setting, is equivalent to Canonical Correlation Analysis [Hot36] (see [KHMN19] for this observation) and it has also been transferred to other applications, e.g., nonequilibrium statistical physics [KCS16, KWNS18, WN20].

Orthogonal NMF and clustering. Coherent sets are a special form of clusters. Clustering itself is strongly related to NMF, in particular, through a modification called Orthogonal NMF, where the problem

$$\min_{B,C\geqslant 0} \|A - BC\|_F^2 \tag{2}$$

is augmented by the additional constraint $CC^\top = \mathrm{Id}_r$. It is shown in [DHS05] that the objectives of Orthogonal NMF and of r-means clustering (of the columns of A) are optimized by equivalent objects, rendering these two problems (however not their algorithmic solutions) equivalent. Based on this pioneering observation, there has been ever since activity on clustering and community detection via NMF [DLPP06, YL13, WKZ+18, LSZL20, OBA22]; see [LD14] for a survey. From a broader perspective, the generality of the formulation in (1) has been exploited to derive modifications of NMF, for instance replacing the distance-like quantity d [FBD09, FI11]; see [Gil20] for an overview. We would also like to point the reader to [SRS+08] which directly links NMF and probabilistic modeling with latent variables.

The orthogonality constraint of NMF implicitly appears in DBMR as well: The rows of the DBMR factor Γ turn out to be orthogonal, cf. Remark 4, hence, after rescaling of the factors λ and Γ, DBMR satisfies the constraints of orthogonal NMF.

Probabilistic models and estimation. With a probabilistic model in the background, the NMF *approximation* problem can be phrased as an *estimation* problem. This connection is used in Probabilistic Latent Semantic Analysis (PLSA) [Hof99, Hof01], where in the original motivation rows and columns of the data matrix correspond to words and documents, respectively. It has been shown in [DLP06] that PLSA is equivalent to NMF if the latter is formulated in terms of the (generalized) Kullback–Leibler divergence. As mentioned above as well as in Remark 6, the Kullback–Leibler divergence is associated with maximum likelihood estimation, hence both PLSA and DBMR essentially compute *most likely low rank models*. Indeed, Gerber and Horenko compare DBMR with PLSA in [GH17, SI sec. 5] and demonstrate superior scaling performance of DBMR for large problems.

Projection-based approximation. Considering our contributions (i) and (ii) mentioned above, we establish a new link between likelihood-based and projection-based approximation of probabilistic models. The latter class has also been extensively studied [DW05,

HS05, dLGDLR08, SS13], and provides bounds of the form where the eigenvalue error between original and projected model is bounded from above by the projection error of the associated eigenvectors. Sharper bounds hold if the model is reversible, essentially meaning that the probability matrix that one seeks to approximate is self-adjoint with respect to a suitable inner product. This applies to the classical approach to the coherence problem as well, since the singular value decomposition of A is equivalent to the eigenvalue decomposition of $A^\top A$, which is self-adjoint by construction.

2. Setup and Notation

Throughout this paper, we denote by $\mathbb{R}^r$ the Euclidean space of dimension $r \in \mathbb{N}$, equipped with the corresponding Euclidean norm $\|\cdot\|_2$ and inner product $\langle \cdot, \cdot \rangle_2$. For a vector $w \in \mathbb{R}^r_{>0}$ with positive entries and $x, y \in \mathbb{R}^r$, let

- $w^{-1} := (w_j^{-1})_{j=1,\dots,r}$ denote the componentwise inverse of w;
- $D_w := \operatorname{diag}(w) := (w_i \delta_{ij})_{i,j=1}^r \in \mathbb{R}^{r\times r}$ denote the corresponding diagonal matrix with entries w_i on its diagonal and δ_{ij} being the Kronecker delta;
- $\langle x, y \rangle_w := x^\top D_w y = \langle D_w^{1/2} x, D_w^{1/2} y \rangle_2$ denote the w-weighted inner product and $\|\cdot\|_w$ the associated norm. We call $x, y \in \mathbb{R}^r$ w-orthogonal if $\langle x, y \rangle_w = 0$.

For a matrix $A \in \mathbb{R}^{m\times n}$, $\|A\|_F := \left(\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2 \right)^{1/2}$ denotes the Frobenius norm of A and $\sigma_k(A)$ the k-th largest singular value. We abbreviate the j-th column of A by $A_{\bullet j}$. Throughout, a *left stochastic matrix* $A \in \mathbb{R}^{m\times n}$ will be a nonnegative matrix the columns of which sum to one. We will *not* require it to be square, slightly abusing standard terminology.

In the bulk of this work, we consider discrete state spaces modeled by finite sets of the form $[r] := \{1, \dots, r\}$, $r \in \mathbb{N}$. Probability measures on $[r]$ will be identified with probability vectors $p \in \mathbb{R}^r_{\geqslant 0}$, that is, with vectors having nonnegative entries that sum to one. For such probability vectors $u, v \in \mathbb{R}^m_{\geqslant 0}$ we define the Kullback–Leibler divergence between u and v by $\mathrm{D}_{\mathrm{KL}}(u \parallel v) = \sum_{i=1}^m u_i \log \frac{u_i}{v_i}$ if u is absolutely continuous with respect to v (interpreted as probability measures) and $\mathrm{D}_{\mathrm{KL}}(u \parallel v) = \infty$ otherwise. Here and in what follows, we use the conventions $\log 0 := -\infty$, $\frac{0}{0} := 0$, and set $c \log 0$ to $0, -\infty, +\infty$ for $c = 0$, $c > 0$, $c < 0$, respectively. Finally, we denote by $\mathbb{1}_E \in \mathbb{R}^r$ the indicator vector associated to a subset $E \subseteq [r]$: its i-th entry is 1 if $i \in E$ and 0 otherwise.

Following [FSM10], we introduce the concept of coherent sets induced by a stochastic (or deterministic) transition. In this context, we adapt a space-discrete setting: This can either be viewed as an approximation to a continuous-space dynamics, or else as a genuinely discrete system. Historically, the former case motivated the construction, and a straightforward connection between the space-continuous and discrete settings is briefly summarized in Appendix A.

Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space and consider two random variables $X : \Omega \to [n]$ and $Y : \Omega \to [m]$ with distributions $p \in \mathbb{R}^n_{\geqslant 0}$ and $q \in \mathbb{R}^m_{\geqslant 0}$, respectively, modeling the state of a random system at initial and final time. Here, $n, m \in \mathbb{N}$ denote the sizes

(cardinalities) of the respective discrete state spaces. Often, $X \sim p$ is considered to be an *input* and $Y \sim q$ to be an *output*,[1] which is why the elements of $[n]$ and $[m]$ will be called input and output categories/states, respectively. We assume that X and Y are coupled through the left stochastic transition matrix $P \in \mathbb{R}^{m\times n}$, $P_{ij} = \mathbb{P}[Y = i \mid X = j]$, and, hence, follow the joint distribution $\mathbb{P}[Y = i, X = j] = P_{ij}p_j$. Using matrix notation we write $(X, Y) \sim PD_p$ and note the relation $q = Pp$.

We next recall the *coherence problem* for the input-output pair (X, Y). On an intuitive level, we would like to obtain a coarse-grained understanding of the situation, for example allowing us to forecast Y given X, in a conceptually simple and computationally tractable, yet faithful way. To this end, we seek nontrivial[2] partitions $\mathcal{E} := (E_i)_{i=1}^r$ of $[n]$ and $\mathcal{F} := (F_i)_{i=1}^r$ of $[m]$ such that $X \in E_i$ implies $Y \in F_i$ with high probability; or, as we will say, (E_i, F_i) form a *coherent pair*. The number of subsets r is fixed for now and roughly corresponds to the complexity of the reduced model; typically we shall aim for $r \ll \min(m, n)$. Following [FSM10], we formulate the following two heuristic conditions for (E_i, F_i) to form a coherent pair:

1. $\mathbb{P}[Y \in F_i \mid X \in E_i] \approx 1$, and
2. $\mathbb{P}[X \in E_i] \approx \mathbb{P}[Y \in F_i]$.

The first condition demands that states from E_i transition predominantly to F_i. The second condition ensures that, in addition, *exclusively* the states from E_i transition to F_i, up to a small error. Taken together, these two conditions describe the scenario that the pair (E_i, F_i) evolves coherently, approximately unaffected by the dynamics on the complements $[n]\backslash E_i$ and $[m]\backslash F_i$.

An attempt to accordingly partition the system into a fixed number $r \in \mathbb{N}$ of coherent pairs is to consider the maximization problem

$$\max_{\substack{\mathcal{E},\ \mathcal{F} \\ \mathbb{P}[X\in E_k]\approx\mathbb{P}[Y\in F_k]}} \sum_{k=1}^{r} \mathbb{P}[Y \in F_k \mid X \in E_k], \tag{3}$$

carried out over all (nontrivial) partitions $\mathcal{E} = (E_k)_{k=1}^r$ of $[n]$ and $\mathcal{F} = (F_k)_{k=1}^r$ of $[m]$ that respect the second condition from above. Here, no partition elements are allowed to be empty sets. To make this a well-posed problem, the constraint $\mathbb{P}[X \in E_k] \approx \mathbb{P}[Y \in F_k]$ needs to be given a quantitative meaning. Note that simply requiring equality may easily render the set of admissible solutions empty. Irrespective of this choice, (3) tends to be a computationally hard combinatorial optimization problem; thus we will later discuss a numerically more approachable relaxation (see Problem 1 below).

Any partition $\mathcal{E} = (E_1, \dots, E_r)$ of $[n]$ can be encoded by an "assignment" $\gamma : [n] \to [r]$ or an "affiliation matrix" $\Gamma \in \{0, 1\}^{r\times n}$ via

$$\gamma(j) := k \text{ with } j \in E_k, \qquad \Gamma_{kj} := \delta_{k\gamma(j)} = \begin{cases} 1 & \text{if } j \in E_k, \\ 0 & \text{if } j \notin E_k. \end{cases} \tag{4}$$

[1] The present situation is sometimes called a *Bayesian relation model* [GH17]. It is a specific, "two-layer" instance of a *Bayesian network* or *decision network*, see, for instance, [Hec98].

[2] We say that the partition $(E_i)_{i=1}^r$ in nontrivial if none of the sets E_i are empty.

The partition $\mathcal{E}$ can then be characterized by $E_k = \gamma^{-1}(\{k\})$. To fix the terminology, we introduce the following notions:

Definition 1 (Affiliation matrix)**:** We call a left stochastic matrix with binary entries $\Gamma \in \{0,1\}^{r\times n}$, $r, n \in \mathbb{N}$, a *(hard) affiliation matrix*. We call the unique map $\gamma : [n] \to [r]$ satisfying $\Gamma_{kj} = \delta_{k\gamma(j)}$ the *assignment* corresponding to Γ.

As explained above, the distribution of the pair (X, Y) can be described in terms of the initial distribution p and the transition matrix P. In practical settings, these objects are typically approximated by their empirical (maximum likelihood) estimates based on finitely many samples, $\boldsymbol{D} = (X(u), Y(u))_{u=1}^{S}$, $S \in \mathbb{N}$, where the pairs $(X(u), Y(u))$ are assumed to be independent and identically distributed copies of (X, Y). The data $\boldsymbol{D}$ leads us to the *count matrix* $N \in \mathbb{N}_0^{m\times n}$ and the *empirical frequency estimators* $\hat{p} \in \mathbb{R}^n_{>0}$ and $\hat{P} \in \mathbb{R}^{m\times n}_{\geqslant 0}$ given by

$$N_{ij} := \#\{u \mid X(u) = j,\, Y(u) = i\}, \qquad \hat{p}_j = \frac{1}{S}\sum_{i=1}^{m} N_{ij}, \qquad \widehat{P}_{ij} = \frac{N_{ij}}{\sum_{i'=1}^{m} N_{i'j}}. \tag{5}$$

Here and in the following we assume the row and column sums of N to be strictly positive for each $i \in [m]$ and $j \in [n]$; otherwise, the associated input or output categories are removed and the sets $[n]$ and $[m]$ are restricted and relabeled accordingly. Note that $\hat{p}$ and $\hat{P}$ are, in fact, maximum likelihood estimates, cf. section 4 and equation (11) in particular. Note that the maximum likelihood estimate of q, $\hat{q} := (\sum_{j=1}^{n} N_{ij}/S)_{i\in[m]} \in \mathbb{R}^m_{>0}$ satisfies $\hat{q} = \hat{P}\hat{p}$, inheriting the above-mentioned relation $q = Pp$ of the exact quantities.

> As the estimation problem (i.e., the comparison of P with $\widehat{P}$, p with $\hat{p}$, and q with $\hat{q}$) is not the focus of the current work, we will not distinguish the exact quantities from their frequency estimators from now on, i.e. we assume them to coincide. In particular, for simplicity of notation, we will use P, p, q to denote the empirical estimators.

Further, throughout this manuscript, $\Lambda \in \mathbb{R}^{m\times n}$, $\lambda \in \mathbb{R}^{m\times r}$ and $\Gamma \in \mathbb{R}^{r\times n}$ denote left stochastic matrices such that $\Lambda = \lambda\Gamma$ and, for reasons to be clarified in section 3, we introduce transition matrices that are normalized with respect to the reference distributions p and q,

$$P' := D_q^{-1} P D_p \in \mathbb{R}^{m\times n}, \qquad \widetilde{P} := D_q^{-1/2} P D_p^{1/2} \in \mathbb{R}^{m\times n}, \qquad \widetilde{\Lambda} := D_q^{-1/2}\Lambda D_p^{1/2}. \tag{6}$$

Note that the normalized transition matrix P' transports *densities* with respect to the reference measure p at initial time to densities with respect to the reference measure q at final times. By $q = Pp$ it is immediate that $P'\mathbb{1}_{[n]} = \mathbb{1}_{[m]}$.

In this context, we also define our main measure of coherence within the pair (P, p), the intuition behind which will be explained in detail in section 3.

Definition 2 (Degree of coherence)**:** We define the *degree of r-coherence* $\mathcal{C}_r(p, P)$ in the pair (p, P) of input distribution and transition matrix as the sum $\sum_{i=1}^{r} \sigma_i(\widetilde{P})$ of the r

leading singular values of $\widetilde{P} := D_q^{-1/2} P D_p^{1/2}$ with $q = Pp$. We simply say that this is the *degree of coherence* in P and write $\mathcal{C}(P)$, if the integer r and the reference distribution p are clear from the context.

3. Classical approach to coherent sets

The following relaxation of the coherence problem (3) can be found in [FSM10] for a two-partition, and in [Den17, sec. 3.3] for an arbitrary number of coherent pairs. For details, the reader is referred to these works. Using (6) we obtain

$$\mathbb{P}\left[Y \in F \mid X \in E\right] = \frac{\sum_{i\in F}\sum_{j\in E} P_{ij}p_j}{\sum_{j\in E} p_j} = \frac{\langle \mathbb{1}_F, PD_p\mathbb{1}_E\rangle_2}{\langle \mathbb{1}_E, D_p\mathbb{1}_E\rangle_2} = \frac{\langle \mathbb{1}_F, P'\mathbb{1}_E\rangle_q}{\|\mathbb{1}_E\|_p^2}. \tag{7}$$

Note that, if $\mathcal{E} = (E_k)_{k=1}^r$ is a partition of $[n]$, then $\mathbb{1}_{E_k}, \mathbb{1}_{E_l}$ are p-orthogonal whenever $k \neq l$, i.e. $\langle \mathbb{1}_{E_k}, \mathbb{1}_{E_l}\rangle_p = 0$. To obtain a computationally feasible relaxation of the coherent set problem (3), we relax the condition that the vectors $\mathbb{1}_{E_k}$ should be indicator vectors, but keep their p-orthogonality. We thus replace $\mathbb{1}_{E_k}$ by vectors $e_k \in \mathbb{R}^n$ and $\mathbb{1}_{F_k}$ by vectors $f_k \in \mathbb{R}^m$ (how these new vectors can be related back to partition elements is explained below). Note that, although we do not require the system f_k to be q-orthogonal at this stage, this property will be a consequence of our analysis (see below). The constraint $\mathbb{P}[X \in E_k] \approx \mathbb{P}[Y \in F_k]$ can now be required with equality and translates to $\|e_k\|_p = \|f_k\|_q$, yielding the relaxed coherent-set maximization problem

$$\max_{\substack{e_1,\dots,e_r \\ f_1,\dots,f_r}} \sum_{k=1}^r \frac{\langle f_k, P'e_k\rangle_q}{\|e_k\|_p \|f_k\|_q}, \tag{8}$$

subject to $e_1, \dots, e_r$ being a p-orthogonal system in $\mathbb{R}^n$. Since (8) is invariant under (positive) scaling of e_k and f_k, we can further restrict the optimization to unit vectors. By noting that $f_k = P'e_k/\|P'e_k\|_q$ is a maximizer of the summands for fixed $e_1, \dots, e_r$, this further reduces to

$$\max_{(e_1,\dots,e_r)\ p\text{-orthonormal}} \sum_{k=1}^r \|P'e_k\|_q \quad \Longleftrightarrow \quad \max_{(\tilde{e}_1,\dots,\tilde{e}_r)\ \text{orthonormal}} \sum_{k=1}^r \|\widetilde{P}\tilde{e}_k\|_2, \tag{9}$$

after observing that any p-orthonormal system $(e_1, \dots, e_r)$ in $\mathbb{R}^n$ can be written as $e_k = D_p^{-1/2}\tilde{e}_k$, $k \in [r]$, for some orthonormal system $(\tilde{e}_1, \dots, \tilde{e}_r)$ in $\mathbb{R}^n$ and that $\|P'e_k\|_q = \|D_q^{1/2}P'D_p^{-1/2}\tilde{e}_k\|_2 = \|\widetilde{P}\tilde{e}_k\|_2$. By (the singular value version of) the Courant–Fischer theorem stated in Theorem 27, the right- and left-hand side of (9) are maximized by the r leading right singular vectors $\tilde{e}_k$ of $\widetilde{P}$ and by $e_k = D_p^{-1/2}\tilde{e}_k$, $k \in [r]$, respectively. The optimal value is equal to the sum of the leading r singular values.

Via the Eckart–Young–Mirsky theorem [HE15, Theorem 4.4.7], the task (9) is equivalent to finding the best rank-r approximation to $\widetilde{P}$ with respect to the Frobenius norm and also the spectral norm:

$$\arg\min_{\substack{A\in\mathbb{R}^{m\times n} \\ \operatorname{rank} A = r}} \left\|\widetilde{P} - A\right\|_F = \sum_{k=1}^r \sigma_k(\widetilde{P}) \tilde{f}_k \tilde{e}_k^\top. \tag{10}$$

In this light, the relaxed coherence problem is equivalent to a low rank approximation problem of the weighted transition matrix $\widetilde{P}$. To summarize the discussion so far, we formulate the relaxed coherence problem as follows:

Problem 1 (Relaxed coherence)**:** Given a count matrix $N \in \mathbb{N}^{m\times n}$ and a fixed rank $r \leqslant \min(m,n)$, find a rank-r matrix $A \in \mathbb{R}^{m\times n}$ that minimizes $\left\|\widetilde{P} - A\right\|_F$, where $\widetilde{P}$ is constructed from N via (5) and (6).

We emphasize that the right-singular vectors of a rank-r matrix $A \in \mathbb{R}^{m\times n}$ solving Problem 1 satisfy the (right) optimality condition in (9). Due to the left stochasticity of P the leading singular value of $\widetilde{P}$ can be shown[3] to be $\sigma_1 = 1$, with corresponding right singular vector $p^{1/2} := (\sqrt{p_1}, \ldots, \sqrt{p_n})^\top$. The maximizers e_k of the relaxed coherence problem (9) need not be approximate indicator vectors, but for well-pronounced coherent dynamics, their linear span (and likewise that of the f_k) is going to be close to the linear span of indicator vectors [KCS16], see also [DW05]. This observation suggests, by viewing the singular vectors e_k as *features* of the states $j \in [n]$, various approaches to extract a coherent r-partition from the singular vectors: A number of algorithms exist, differing in how post-processing steps are handled, and whether hard or soft clusters are sought. One can use k-means clustering [Den17, sec. 3.3], [BK17], PCCA+ [DW05, RW13], or SEBA [FRS19]. The final-time members of the coherent pairs are obtained in a similar manner from the vectors $f_k = D_q^{1/2}\tilde{f}_k$, where $\tilde{f}_k$ are the left singular vectors of $\widetilde{P}$, and are matched to the initial-time members such that the objective in (3) is maximal. We summarize this *classical approach* to coherent pairs, using k-means clustering in the postprocessing step, in Algorithm 1.

The relationship between (7) and (8) shows that the value of the latter is a measure for the coherence of an r-partition which motivates our usage of this value as the "degree of coherence" in Definition 2. Since the optimal value for (8) is the sum $\sum_{i=1}^r \sigma_i(\widetilde{P})$ over the r leading singular values of $\widetilde{P}$, it follows that the degree of coherence of an r-partition is bounded from above by r. In other words, tightness of the bound $\sum_{i=1}^r \sigma_i \leqslant r$ indicates coherence of the system at hand. In the case of complete coherence ($\sum_{i=1}^r \sigma_i = r$), the transition matrix P (and hence $\widetilde{P}$ as well) has the form where there are partitions $(E_k)_{k=1}^r$ of $[n]$ and $(F_k)_{k=1}^r$ of $[m]$ such that $P_{ij} > 0$ implies $i \in E_k$ and $j \in F_k$ for the same k.

The question of how to choose the number r of coherent sets can be answered by considering the singular spectrum of $\widetilde{P}$ [Fro13]: The aim is to have the leading r singular values close to one (and the corresponding singular vectors close to linear combinations of indicator vectors $\mathbb{1}_{E_k}$ corresponding to some partition $\mathcal{E} = (E_k)_{k=1}^r$), while the remaining singular values should, ideally, be substantially smaller than 1 (indicating no further coherence within the system). Consequently, the choice of r should be informed by the values and gaps in the spectrum.

[3] Since $q^{1/2} = \widetilde{P}p^{1/2}$ and p, q are probability vectors, i.e. $\|p\|_2 = \|q\|_2 = 1$, we have $\sigma_1 \geqslant 1$. To show $\sigma_1 \leqslant 1$, note that σ_1^2 is the leading eigenvalue of $\widetilde{P}^\top\widetilde{P}$ and hence also of the similar matrix $D_p^{-1/2}\widetilde{P}^\top\widetilde{P}D_p^{1/2} = P^\top D_q^{-1}PD_p$. It is a straightforward calculation that $\|P^\top D_q^{-1}PD_p\|_\infty = 1$, thus $\sigma_1^2 \leqslant 1$.

Algorithm 1 Classical approach to coherent pairs.

1: INPUT: Data subsumed into the count matrix N, number of coherent pairs $r \in \mathbb{N}$
2: Compute $P, p, \widetilde{P}$ via (5) and (6)
3: Compute SVD of $\widetilde{P}$: $\widetilde{P} = U\Sigma V^\top$ with orthogonal matrices $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ and rectangular matrix

$$\Sigma = \begin{pmatrix} \operatorname{diag}(\sigma_1, \ldots, \sigma_s) & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{m\times n}$$

with the singular values $\sigma_1 \geqslant \cdots \geqslant \sigma_s > 0$ of $\widetilde{P}$ on its diagonal
4: Truncate to the r leading singular values to obtain a low rank approximation $\widetilde{P}_{\mathrm{red}}$ of $\widetilde{P}$,

$$\widetilde{P}_{\mathrm{red}} = U \begin{pmatrix} \operatorname{diag}(\sigma_1, \ldots, \sigma_r) & 0 \\ 0 & 0 \end{pmatrix} V^\top,$$

and transform back to obtain a low rank approximation $P_{\mathrm{red}} := D_q^{-1/2} \widetilde{P}_{\mathrm{red}} D_p^{1/2}$ of P
5: Use the first r right singular vectors $\tilde{e}_1, \ldots, \tilde{e}_r$ of $\widetilde{P}$ (the first r columns of V) as features for a k-means-clustering of $[n]$ into r clusters [Den17, sec. 3.3], leading to the partition $\mathcal{E} = (E_k)_{k=1}^r$ of $[n]$; and obtain similarly the partition $\mathcal{F} = (F_k)_{k=1}^r$ of $[m]$ using the left singular vectors $\tilde{f}_1, \ldots, \tilde{f}_r$ of $\widetilde{P}$
6: "Match" the partitions $\mathcal{E}$ and $\mathcal{F}$ by reordering $\mathcal{F}$ such that the objective in (3) is maximized
7: OUTPUT: $\widetilde{P}_{\mathrm{red}}$, P_{red}, $\mathcal{E}$, $\mathcal{F}$

Remark 3: The observation, mentioned in section 1, that the algebraic form of the relaxed coherence problem is equivalent to Canonical Correlation Analysis (CCA), has been made in [KHMN19]. CCA is commonly described as a method that finds bases of the input and output space with maximal correlation under an assumed probabilistic relationship.

4. Likelihood-based estimation from data

4.1. Full versus low rank models

Solving the relaxed coherence problem as presented so far is a two-step procedure: First, p and P are estimated from observational data, and second, the dominant singular vectors are extracted from $\widetilde{P}$ (see Problem 1). Thus, it is natural to ask whether a low rank approximation of $\widetilde{P}$ can be obtained directly, merging estimation and projection.

For this purpose, recall the empirical estimators $\widehat{p}$ and $\widehat{P}$ for p and P from (5). It is classical and straightforward to show that $\pi = \widehat{p}$, $\Lambda = \widehat{P}$ maximize the likelihood of the data $\boldsymbol{D}$,

$$\mathbb{P}\left[\boldsymbol{D} \mid \pi, \Lambda\right] = \prod_{i=1}^{m} \prod_{j=1}^{n} \pi_j^{N_{ij}} \Lambda_{ij}^{N_{ij}} \,. \tag{11}$$

Reiterating the discussion from section 2, we overload the notation, dropping the hats in $(\pi^*, \Lambda^*) = (\widehat{p}, \widehat{P})$, and hence denoting the true objects (p, P) and their empirical (maximum likelihood) estimators by the same symbols. Traditionally, these maximum likelihood estimates are then used to approximate $\widetilde{P}$ and its r leading singular modes, from which a coherent r-partition can be extracted in various ways, see section 3. Note that this approach requires the approximation of $(m+1)n$ probability values, some of which might be very small, hence a large number of samples if often required.

However, if we are interested in a fixed number $r \ll \min(m, n)$ of singular modes, the effective information of interest is already represented by $\mathcal{O}(r(n+m))$ quantities, and hence it might be expected that a direct approach (circumventing the estimation of p and P) could provide accurate results based on a significantly reduced number of samples. More specifically, we will contrast the traditional procedure with a direct estimation of P by a low rank transition matrix $\Lambda = \lambda\Gamma$, where $\lambda \in \mathbb{R}^{m\times r}$ and $\Gamma \in \mathbb{R}^{r\times n}$ are left stochastic matrices still fulfilling $q = Pp$, in the maximum-likelihood framework of DBMR. Indeed, DBMR requires the computation of only $r(n+m)$ matrix entries and thus its output promises to be of low variance, even in the regime where only a few samples are available [GH17, Theorem; in particular Eq. 7].

In terms of interpretability, our alternative approach has another crucial advantage: DBMR maximizes a lower bound of the log-likelihood function, the optimum $\Lambda^* = \lambda^*\Gamma^*$ of which turns out to comprise a (hard) affiliation matrix Γ^* in the sense of Definition 1 (though starting with the assumption on Γ to be only left stochastic). As mentioned in section 2, there is a one-to-one correspondence between such affiliation matrices and partitions of $[n]$, providing a meaningful partition $\mathcal{E} := (E_k)_{k=1}^r$ of $[n]$ *without* any post-processing steps, while λ corresponds to a 'reduced transition matrix' on these compound states. A natural choice for the output partition $\mathcal{F} := (F_k)_{k=1}^r$ of $[m]$ is given by

$$F_k = \{i \in [m] \mid \lambda_{ik} = \max_{k' \in [r]} \lambda_{ik'}\}$$

(with arbitrary choice of category in case of non-uniqueness of the maximizer).

4.2. Direct Bayesian model reduction (DBMR)

In the following, we will discuss how the low rank model $\Lambda = \lambda\Gamma \approx P$, where $\lambda \in \mathbb{R}^{m\times r}$ and $\Gamma \in \mathbb{R}^{r\times n}$ are left stochastic matrices, can be estimated from the data in a maximum likelihood fashion similar to the derivation of (5) from (11). This approach, proposed by Gerber and Horenko [GH17], achieves both estimation and model reduction simultaneously, without the need to estimate the full model P in the first place. In other words, we assume that the output depends on the input through a *latent variable* $Z \in [r]$, illustrated by the graphical model $X \xrightarrow{\Gamma} Z \xrightarrow{\lambda} Y$, encapsulating the conditional independence assumption $\mathrm{Law}[Y \mid X, Z] = \mathrm{Law}[Y \mid Z]$. In this case, Γ and λ correspond to the transition matrices to and from the latent state, respectively:

$$\Gamma_{kj} = \mathbb{P}\big[Z = k \mid X = j\big], \qquad \lambda_{ik} = \mathbb{P}\big[Y = i \mid Z = k\big]. \tag{12}$$

Note that we can interpret $\Gamma_{kj} \in [0,1]$ as a (soft) *affiliation* of input category j to the latent state k. As we will see below, the DBMR solution in fact yields binary estimates $\Gamma_{kj} \in \{0,1\}$, interpreted as hard affiliations as in Definition 1.

Since (11) can be split into two optimization problems, one for π and one for Λ, estimating the factors λ and Γ from the observation data $\boldsymbol{D}$ via maximum likelihood estimation reduces to maximizing

$$\ell(\lambda, \Gamma) := \log \mathbb{P}[\boldsymbol{D} \mid \lambda\Gamma] = \sum_{i=1}^{m} \sum_{j=1}^{n} N_{ij} \log(\lambda\Gamma)_{ij}, \tag{13}$$

over all pairs (λ, Γ) of left stochastic matrices. Since the full model $P = \widehat{P}$ maximizes (11) without the low rank constraint, we obtain the natural bound $\ell(\lambda, \Gamma) \leqslant \ell(P, \mathrm{Id}_n)$. Since (13) has no closed-form maximizer, [GH17] suggest to relax the problem and maximize a lower bound of ℓ,

$$\hat{\ell}(\lambda, \Gamma) := \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{r} N_{ij} \Gamma_{kj} \log \lambda_{ik} \leqslant \ell(\lambda, \Gamma), \tag{14}$$

where we applied Jensen's inequality. This leads to the following formulation:

Problem 2 (DBMR)**:** Given a count matrix $N \in \mathbb{N}^{m \times n}$ and a fixed rank $r \leqslant \min(m, n)$, find left stochastic matrices $\lambda \in \mathbb{R}^{m \times r}$ and $\Gamma \in \mathbb{R}^{r \times n}$ that maximize $\hat{\ell}$ given by (14).

The DBMR algorithm 2 suggested by [GH17] maximizes $\hat{\ell}(\lambda, \Gamma)$ by an alternating optimization over λ and Γ with a computational cost that is linear in m and n: Maximizing $\hat{\ell}(\lambda, \Gamma)$ for fixed Γ yields a unique optimum

$$\lambda_{ik} = \frac{\sum_{j=1}^{n} \Gamma_{kj} N_{ij}}{\sum_{i'=1}^{m} \sum_{j'=1}^{n} \Gamma_{kj'} N_{i'j'}}. \tag{15}$$

On the other hand, for any fixed left stochastic matrix λ, maximizing $\hat{\ell}(\lambda, \Gamma)$ with respect to Γ decouples into n separate linear programs [GH17, Suppl. p. 19] solved by

$$\Gamma_{kj} = \begin{cases} 1, & k = \arg\max_{k'} \sum_{i=1}^{m} N_{ij} \log \lambda_{ik'} \\ 0, & \text{else.} \end{cases} \tag{16}$$

Possible non-uniqueness of the arg max is resolved such that there is only one nonzero entry in every column of Γ. This means that solutions to the DBMR problem necessarily describe *binary* (or hard) affiliations $\Gamma_{kj} \in \{0,1\}$ of the input states $j \in [n]$ to the latent states $k \in [r]$ as in Definition 1. The fact that the partial updates (15) and (16) are available in closed form motivates the alternating procedure described in Algorithm 2, introducing the iteration counter h.

We note that $\hat{\ell}(\lambda, \Gamma)$ is concave with respect to both λ and Γ, individually. Thus, $\hat{\ell}(\lambda^{(h)}, \Gamma^{(h)})$ increases in h. Since there are only finitely many values that Γ can take,

Algorithm 2 DBMR algorithm from [GH17].

1: INPUT: Data subsumed into the count matrix N, number of latent states $r \in \mathbb{N}$, and maximum iteration number $h_{\max}$
2: Set random stochastic matrix $\Gamma^{(0)} \in \{0,1\}^{r\times n}$ and $h = 0$
3: Set $\lambda^{(0)}$ by evaluating (15) for $\Gamma = \Gamma^{(0)}$
4: **while** $\hat{\ell}(\lambda^{(h)}, \Gamma^{(h)}) \neq \hat{\ell}(\lambda^{(h-1)}, \Gamma^{(h-1)})$ and $h < h_{\max}$ **do**
5: Set $\Gamma^{(h+1)}$ by evaluating (16) for $\lambda = \lambda^{(h)}$
6: Set $\lambda^{(h+1)}$ by evaluating (15) for $\Gamma = \Gamma^{(h+1)}$
7: $h \leftarrow h + 1$
8: **end while**
9: OUTPUT: $\lambda = \lambda^{(h)}$ and $\Gamma = \Gamma^{(h)}$

the algorithm converges, but possibly to a maximum of $\hat{\ell}$ that is only local (with respect to the updates (15) and (16)). A practical alternative stopping criterion is to stop if the relaxed likelihood $\hat{\ell}$ shows small improvements that are below a given threshold. Since the algorithm might only find a locally optimal solution, it is usually run several times with independent random initializations $\Gamma^{(0)}$, and the result with the highest relaxed likelihood value is taken.

Remark 4: Note that (16) implies that $\Gamma\Gamma^\top$ is a diagonal matrix: the rows of Γ are orthogonal, but typically not orthonormal. In Orthogonal NMF, as mentioned in section 1.2, the orthogonality requirement would translate to $\Gamma\Gamma^\top = \mathrm{Id}_K$. If Γ has full rank, this can be achieved by the replacement $\Gamma \to D\Gamma$, using a diagonal scaling D; cf. [DLPP06, below equation (10)]. In this case we also have $\lambda\Gamma = \big(\lambda D^{-1}\big)(D\Gamma)$, where the factors in the parentheses fulfill the requirements of Orthogonal NMF. DBMR hence yields a particular form of Orthogonal NMF.

DBMR as maximum likelihood estimate on a constraint set. For fixed $m, n \in \mathbb{N}$ and for $r \in \mathbb{N}$, $r \leqslant \min\{m, n\}$, denote

$$\mathcal{D}^r_\Lambda \coloneqq \left\{\Lambda = \lambda\Gamma \in \mathbb{R}^{m\times n} \mid \lambda \in \mathbb{R}^{m\times r}_{\geqslant 0},\ \Gamma \in \{0,1\}^{r\times n} \text{ both left stochastic}\right\}. \tag{17}$$

The set $\mathcal{D}^r_\Lambda$ comprises the set of transition matrices that we consider as low rank approximations (more precisely, as approximations of rank at most r) to the full-rank transition matrix P. The salient feature (in addition to the low rank constraint) is the sparsity assumption $\Gamma \in \{0,1\}^{r\times n}$ which, by the left stochasticity of Γ, leads to the interpretation of Γ as a (hard) affiliation matrix, see Definition 1.

If we restrict the maximum likelihood estimation from section 4.1 to the set $\mathcal{D}^r_\Lambda$,

$$\Lambda^* = \operatorname*{arg\,max}_{\Lambda\in\mathcal{D}^r_\Lambda} \left\{ \sum_{i=1}^{m}\sum_{j=1}^{n} N_{ij} \log \Lambda_{ij} \right\}, \tag{18}$$

then we obtain an alternative derivation of the DBMR algorithm:

Lemma 5: The maximizers of $\hat{\ell}$ in (14), setting $\Lambda^* = \lambda^*\Gamma^*$, coincide with the solutions to (18). In other words, the maximizers of ℓ and $\hat{\ell}$ coincide if we restrict the considerations to binary Γ.

Proof. According to [GH17], any maximizer $\Lambda^* = \lambda^*\Gamma^*$ of $\hat{\ell}$ satisfies $\Lambda^* \in \mathcal{D}^r_\Lambda$. Note that the inequality in (14) follows from Jensen's inequality, $\sum_{k=1}^r \Gamma_{kj} \log \lambda_{ik} \leqslant \log \sum_{k=1}^r \lambda_{ik}\Gamma_{kj}$, and is, in fact, an equality whenever Γ is a binary left stochastic matrix. Hence, ℓ and $\hat{\ell}$ coincide on $\mathcal{D}^r_\Lambda$ (when this set is viewed as the set of admissible pairs (λ, Γ) in (17)). □

It is well known that maximum likelihood estimation is inherently related to Kullback–Leibler minimization [Was04, Section 9.5]. The following remark establishes this connection in the specific context of DBMR:

Remark 6 (Connection to Kullback–Leibler divergences)**:** As mentioned in section 2, the joint distribution of the data (X, Y) is described by the probability values $P_{ij}p_j = \mathbb{P}[Y = i, X = j]$ (which are proportional to N_{ij}). We can likewise describe the joint distribution of the DBMR output by $\Lambda_{ij}p_j$, where $\Lambda \in \mathcal{D}^r_\Lambda$. The DBMR objective (18) can be rewritten in the form

$$\Lambda^* = \operatorname*{arg\,min}_{\Lambda \in \mathcal{D}^r_\Lambda} \mathrm{D}_{\mathrm{KL}}(\{P_{ij}p_j\} \parallel \{\Lambda_{ij}p_j\}), \tag{19}$$

where, with slight abuse of notation, we naturally extend the definition of D_{KL} from section 2 to matrices associated with joint distributions. In other words, DBMR attempts to match the true joint distribution in the Kullback–Leibler sense, while obeying the rank and sparsity constraints imposed by $\mathcal{D}^r_\Lambda$. Indeed,

$$\mathrm{D}_{\mathrm{KL}}(\{P_{ij}p_j\} \parallel \{\Lambda_{ij}p_j\}) = \sum_{i,j} \log\left(\frac{P_{ij}p_j}{\Lambda_{ij}p_j}\right) P_{ij}p_j \propto \sum_{i,j} N_{ij} \left(\log P_{ij} - \log \Lambda_{ij}\right). \tag{20}$$

Clearly, minimizing (20) is equivalent to maximizing (18), since $\sum_{i,j} N_{ij} \log P_{ij}$ does not depend on Λ.

Remark 7: In the case when the count matrix N is of low effective dimensionality, that is, when $\operatorname{rank} N = \widetilde{r} < r$, a natural question is whether the DBMR output $\Lambda^* \in \mathcal{D}^r_\Lambda$ lies in the smaller set $\mathcal{D}^{\widetilde{r}}_\Lambda$, i.e., whether DBMR automatically identifies the low rank structure of N. In general, this is not the case: Consider two linearly independent vectors $a, b \in \mathbb{R}^m$, $m \geqslant 3$, and $P = [a, \frac{a}{2} + \frac{b}{2}, b] \in \mathbb{R}^{m\times 3}$ (recall that N is simply a scaled version of P, implying here that $\operatorname{rank} N = \operatorname{rank} P = \widetilde{r} = 2$). For $r = 3$, the best approximation of P within $\mathcal{D}^r_\Lambda$ is clearly P itself, by choosing $\lambda^* = P$ and $\Gamma^* = \mathrm{Id}_3$. This solution is unique up to permutations of the columns of λ^* and rows of Γ^*, respectively, since, for any affiliation matrix $\Gamma \in \mathbb{R}^{3\times 3}$ with $\operatorname{rank}\Gamma < 3$ and any $\lambda \in \mathbb{R}^{m\times 3}$, the product $\lambda\Gamma$ has at least two identical columns and cannot coincide with P. Clearly, the above solution $\Lambda^* = \lambda^*\Gamma^*$ maximizes the likelihood in (13) and coincides with the DMBR output, at least for appropriate initializations (e.g. $\Gamma^{(0)} = \mathrm{Id}_3$).

Remark 8 (Choice of r)**:** Statistical approaches towards model selection may be applied to the problem of choosing an appropriate latent dimensionality r, see, for instance,

[Gil20, Section 5.2]. In the context of DBMR, Akaike's information criterion (AIC) [Aka74] as an estimator of the relative expectation of Kullback–Leibler distance based on maximized log-likelihood has been suggested by [GH17]. For the this setting, the AIC for the reduced model is defined by

$$\mathrm{AIC}(\lambda, \Gamma) := 2r(n+m) - 2\hat{\ell}(\lambda, \Gamma), \tag{21}$$

where $r(n+m)$ is the number of estimated parameters in λ and Γ and $\hat{\ell}$ is the DBMR objective defined by (14). The term $2r(n+m)$ in (21) is a bias-correction term depending on the number of latent states r. The optimal integer value of r can be obtained by performing the DBMR algorithm with different numbers of r (i.e., $r = 1, 2, 3, ...$) and then selecting the reduced model with the minimal AIC value in (21).

5. Relations between the full and the reduced model

5.1. DBMR as a projection

In order to analyze the approximation properties of the reduced model $\Lambda = \lambda\Gamma$ provided by DBMR, we will consider the (hard) affiliation matrix Γ fixed throughout this section (assume e.g. that it has already been computed). Then, in view of (15), DBMR constructs the low rank approximation Λ of P as follows: For each $k = 1, \ldots, r$,

- compute the column $\lambda_{\bullet k}$ as the weighted average of the columns of P associated with k: $\lambda_{\bullet k} \propto \sum_{j=1}^{n} \Gamma_{kj} p_j P_{\bullet j}$,
- replace the columns of P associated with k by this average: $\Lambda_{\bullet j} = \lambda_{\bullet k}$ if $\Gamma_{kj} = 1$.

The crucial observation in this section is that this process can be rewritten as a composition of P with a projection Π,

$$\Lambda = \lambda\Gamma = P\Pi, \tag{22}$$

as illustrated by the following example:

Example 9: Let $n = 5$ and $r = 2$ and assume that we have already computed

$$\Gamma = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

Then the observation described above concerning the λ-update-step (15) within DBMR can be illustrated as follows:

$$P = \begin{pmatrix} P_{\bullet 1} & P_{\bullet 2} & P_{\bullet 3} & P_{\bullet 4} & P_{\bullet 5} \end{pmatrix}$$

$\xrightarrow[\text{according to } \Gamma]{\text{group columns of } P}$

$$\lambda = \begin{pmatrix} \lambda_{\bullet 1} = \frac{p_1 P_{\bullet 1} + p_2 P_{\bullet 2}}{p_1 + p_2} & \lambda_{\bullet 2} = \frac{p_3 P_{\bullet 3} + p_4 P_{\bullet 4} + p_5 P_{\bullet 5}}{p_3 + p_4 + p_5} \end{pmatrix}$$

$\xrightarrow[\text{columns}]{\text{average}}$

$$\Lambda = \lambda\Gamma = \begin{pmatrix} \lambda_{\bullet 1} & \lambda_{\bullet 1} & \lambda_{\bullet 2} & \lambda_{\bullet 2} & \lambda_{\bullet 2} \end{pmatrix}.$$

$\xrightarrow{\text{augment}}$

Clearly, this procedure of obtaining $\Lambda = \lambda\Gamma$ can be written as a matrix product $\Lambda = P\Pi$ for

$$\Pi = \begin{pmatrix} \frac{p_1}{p_1+p_2} & \frac{p_1}{p_1+p_2} & 0 & 0 & 0 \\ \frac{p_2}{p_1+p_2} & \frac{p_2}{p_1+p_2} & 0 & 0 & 0 \\ 0 & 0 & \frac{p_3}{p_3+p_4+p_5} & \frac{p_3}{p_3+p_4+p_5} & \frac{p_3}{p_3+p_4+p_5} \\ 0 & 0 & \frac{p_4}{p_3+p_4+p_5} & \frac{p_4}{p_3+p_4+p_5} & \frac{p_4}{p_3+p_4+p_5} \\ 0 & 0 & \frac{p_5}{p_3+p_4+p_5} & \frac{p_5}{p_3+p_4+p_5} & \frac{p_5}{p_3+p_4+p_5} \end{pmatrix}.$$

This motivates the definition of Π in (23) below.

Remark 10: The procedure and example above explain why DBMR provides "good" low rank approximations Λ of P and why it identifies coherent pairs: P is a maximum likelihood estimate and Λ maximizes the same likelihood, but within the set $\mathcal{D}_\Lambda^r$ of low rank matrices, cf. (18). Hence, indirectly, Λ results to be as "close" to P as possible. In view of the discussion above, the best way to be "close" to P is to group its columns by *similarity* (and this clustering then defines Γ). This way, each column in P does not differ too much from the corresponding average column in Λ. As a consequence, in terms of coherence, the states within each group (i.e. partition element) evolve with similar probability vectors, hence "coherently".

After establishing equation (22) in Theorem 18 below, together with several properties of Π, we leverage this result to draw conclusions about the relationship between the full model P and the low rank model $\Lambda = \lambda\Gamma$, in particular, in the context of coherence. More precisely, we show that the degree of coherence (measured by the sum of leading singular values of the corresponding matrix as in section 3) associated to the DBMR approximation is bounded from above by the one of the full model (cf. Proposition 14), $\mathcal{C}(\Lambda) \leqslant \mathcal{C}(P)$.

In what follows, we work with the full and low rank transition matrices P and Λ as well as with their rescaled versions $\widetilde{P} = D_q^{-1/2} P D_p^{1/2} \in \mathbb{R}^{m\times n}$, $\widetilde{\Lambda} = D_q^{-1/2}\Lambda D_p^{1/2}$, see (6) and section 3. To identify the projection Π in (22), recall the assignment $\gamma : [n] \to [r]$ associated to a fixed (hard) affiliation matrix $\Gamma \in \mathbb{R}^{r\times n}$ from Definition 1. Motivated by Example 9, we define $\Pi \in \mathbb{R}^{n\times n}_{\geqslant 0}$ and $\widetilde{\Pi} = D_p^{-1/2}\Pi D_p^{1/2}$ by

$$\Pi_{ij} = \frac{p_i\, \delta_{\gamma(i)\gamma(j)}}{\sum_l p_l \delta_{\gamma(l)\gamma(i)}}, \qquad \widetilde{\Pi}_{ij} = \frac{\sqrt{p_i p_j}\, \delta_{\gamma(i)\gamma(j)}}{\sum_l p_l \delta_{\gamma(l)\gamma(i)}}, \tag{23}$$

noting that here and in subsequent sections, $\widetilde{\Pi}$ serves as an auxiliary object that facilitates computations. Indeed, it is straightforward to verify that $\widetilde{\Pi}$ is an orthogonal projection (cf. Lemma 24 in Appendix B). Consequently, Π is a p^{-1}-orthogonal projection, meaning that $\Pi^2 = \Pi$ and that Π is p^{-1}-symmetric:

$$\langle \Pi x, y\rangle_{p^{-1}} = \langle x, \Pi y\rangle_{p^{-1}}, \qquad \text{for all} \quad x, y \in \mathbb{R}^n. \tag{24}$$

The following result confirms the relation (22) and clarifies the structure of Π in terms of its eigenvector decomposition. An important role is played by the set $\mathrm{Ran}\gamma = \gamma([n]) \subset [r]$, which we refer to as the set of *active* latent states.

Theorem 11 (DBMR as a projection)**:** Let $\Gamma \in \{0,1\}^{r\times n}$ be a hard affiliation matrix (according to Definition 1) and λ be given by (15). Then Π as defined in (23) is a left stochastic and p^{-1}-orthogonal projection which satisfies $\lambda\Gamma = P\Pi$. Moreover, Π has the following properties:

(a) The rank of Π coincides with the number of active latent states, $\operatorname{rank}\Pi = \#\mathrm{Ran}\gamma$. In particular, $\operatorname{rank}\Pi \leqslant r$.

(b) The vectors $a^{(k)} \in \mathbb{R}^n$, associated to active latent states $k \in \mathrm{Ran}\gamma \subseteq [r]$ and defined by

$$a_i^{(k)} := p_i \delta_{\gamma(i)k}\,, \qquad i = 1,\dots,n, \tag{25}$$

are eigenvectors of Π with eigenvalue 1, i.e., $\Pi a^{(k)} = a^{(k)}$. They span the image space of Π, that is,

$$\mathrm{Span}\{a^k :\ k \in \operatorname{Ran}\gamma\} = \operatorname{Ran}\Pi. \tag{26}$$

The supports of the vectors $a^{(k)}$ are disjoint, that is, $a_i^{(k)}a_i^{(l)} = 0$, for all i and all $k \neq l$. In particular, these vectors are orthogonal as well as p-orthogonal.

Hence, the vector $p = \sum_{k\in\mathrm{Ran}\gamma} a^{(k)}$ is also an eigenvector of Π with eigenvalue 1.

(c) The DBMR output $\Lambda = \lambda\Gamma$ respects the final distribution in the sense that $\Lambda p = q$.

Proof. The proof is given in Appendix B. □

Remark 12: Let us comment on the previous result.

(a) Note that the assumptions in Theorem 11 on Γ and λ are satisfied at any (completed) iteration of the DBMR algorithm; convergence is not required.

(b) For a fixed (active) latent state $k \in [r]$, it is natural to consider the corresponding set $\gamma^{-1}(\{k\})$ comprised of the states in $[n]$ that are mapped to k. The disjoint union $\bigcup_k \gamma^{-1}(\{k\}) = [n]$ can then be viewed as a coarse-graining of the input set $[n]$ induced by the assignment γ. Theorem 11 shows that this interpretation persists at the level of the projection Π, and that its range encodes the same information. Indeed, the eigenvectors in (25) can be obtained as the restriction of p to $\gamma^{-1}(\{k\})$,

$$a_i^{(k)} = \begin{cases} p_i & \text{if } i \in \gamma^{-1}(\{k\}), \\ 0 & \text{otherwise.} \end{cases} \tag{27}$$

Corollary 13: Let $\Gamma \in \{0,1\}^{r\times n}$ be a hard affiliation matrix (according to Definition 1), λ be given by (15) and $\Lambda = \lambda\Gamma$. Further, let $\widetilde{P}$ and $\widetilde{\Lambda}$ be given by (6) and $\widetilde{\Pi}$ by (23). Then $\widetilde{P} - \widetilde{\Lambda}$ is orthogonal to any matrix of the form $A\widetilde{\Pi}$, $A \in \mathbb{R}^{m\times n}$ with respect to the Frobenius norm. In particular, since $\widetilde{\Lambda} = \widetilde{P}\widetilde{\Pi}$,

$$\|\widetilde{P} - \widetilde{\Lambda}\|_F^2 = \|\widetilde{P}\|_F^2 - \|\widetilde{\Lambda}\|_F^2. \tag{28}$$

Further, for a fixed (hard) affiliation matrix $\Gamma \in \{0,1\}^{r\times n}$ of rank r, the choice of λ in (15) results in a matrix $\widetilde{\Lambda}$ that is the best approximation of $\widetilde{P}$ in Frobenius norm:

$$\widetilde{\Lambda} = D_q^{-1/2}\lambda\Gamma D_p^{1/2} \in \operatorname*{arg\,max}_{\lambda'\in\mathbb{R}^{m\times r}} \|\widetilde{P} - D_q^{-1/2}\lambda'\Gamma D_p^{1/2}\|_F. \tag{29}$$

Proof. Note that for real matrices A and B of the same dimension, the Frobenius norm is induced by the inner product $\langle A, B\rangle_F := \operatorname{tr}(A^\top B)$, where $\operatorname{tr}(\cdot)$ denotes the trace. By (6) and (23) the identity $\Lambda = P\Pi$ from Theorem 11 implies $\widetilde{\Lambda} = \widetilde{P}\widetilde{\Pi}$. Therefore, using the properties $\widetilde{\Pi}^\top = \widetilde{\Pi}$ and $\widetilde{\Pi}^2 = \widetilde{\Pi}$ from Lemma 24 as well as invariance of the trace under cyclic permutation of factors, we obtain, for each $A \in \mathbb{R}^{m\times n}$

$$\langle \widetilde{P} - \widetilde{\Lambda}, A\widetilde{\Pi}\rangle_F = \operatorname{tr}\left(\left(\widetilde{P} - \widetilde{\Lambda}\right)^\top A\widetilde{\Pi}\right) = \operatorname{tr}\left(\widetilde{\Pi}\left(\mathrm{Id} - \widetilde{\Pi}\right)\widetilde{P}^\top A\right) = 0. \tag{30}$$

This readily implies (28).

For the last statement, let us define the diagonal matrices $D_N = \operatorname{diag}(\mathbb{1}^\top N)^{-1}$, $D_1 = \operatorname{diag}(\mathbb{1}^\top N\Gamma)^{-1}$, and $D_2 = \operatorname{diag}(\mathbb{1}^\top N\Gamma^\top\Gamma)^{-1}$. It is a straightforward calculation to see that $P = ND_N$, $D_1\Gamma = \Gamma D_2$, and $\Pi = D_N^{-1}\Gamma^\top\Gamma D_2$. Since Γ is assumed to have rank r, $\Gamma\Gamma^\top$ is invertible, yielding that for an arbitrary $\lambda' \in \mathbb{R}^{m\times r}$ we have

$$\begin{aligned}\lambda'\Gamma &= \lambda' D_1^{-1}D_1\Gamma = \lambda' D_1^{-1}\Gamma D_2 = \lambda' D_1^{-1}(\Gamma\Gamma^\top)^{-1}\Gamma\Gamma^\top\Gamma D_2\\ &= \lambda' D_1^{-1}(\Gamma\Gamma^\top)^{-1}\Gamma D_N\ D_N^{-1}\Gamma^\top\Gamma D_2 =: A'\Pi\end{aligned}$$

This means that $\lambda'\Gamma$ is of the form $A'\Pi$ for a suitable $A' \in \mathbb{R}^{m\times n}$, and (29) follows from (30). ∎

Hence, (15) is optimal in the sense of (29) in addition to being optimal in terms of the (relaxed) maximum likelihood from section 4.2.

Corollary 13 will be used in Corollary 20 below to relate the objectives degree of coherence $\mathcal{C}_r$ and relaxed likelihood $\hat{\ell}$.

5.2. Pointwise singular value bounds

Recall that $\sigma_i(M)$, $i \geqslant 1$, denotes the i-th singular value of a matrix M, in descending order. By Theorem 11(b), $\Pi p = p$ and hence $p^{1/2}$ is a right eigenvector of $\widetilde{\Pi}$ with eigenvalue 1:

$$\widetilde{\Pi}p^{1/2} = D_p^{-1/2}\Pi D_p^{1/2}p^{1/2} = D_p^{-1/2}\Pi p = D_p^{-1/2}p = p^{1/2}.$$

As discussed in section 3, we have that $\sigma_1(\widetilde{P}) = 1$ with corresponding right singular vector $p^{1/2}$. Thus, we also have that $\sigma_1(\widetilde{P}\widetilde{\Pi}) = 1$ with corresponding right singular vector $p^{1/2}$, since $\widetilde{\Pi}$ is an orthogonal projection and hence $\widetilde{P}\widetilde{\Pi}$ cannot have singular values larger than 1. Thus, $\widetilde{P}$ and $\widetilde{P}\widetilde{\Pi}$ share the same leading singular value with the same left and right singular vector pair. As for the comparison of the other singular values, the following holds true.

Proposition 14: With $\widetilde{P}$ defined in (6) and $\widetilde{\Pi}$ defined in (23), we have that

$$\sigma_i(\widetilde{P}\widetilde{\Pi}) \leqslant \sigma_i(\widetilde{P}), \quad i \leqslant \min\{m, n\}. \tag{31}$$

Proof. The claim follows from [GK78, (2.3) on pp. 27] by noting that $\|\widetilde{\Pi}\|_2 = 1$. For the readers' convenience, we give a proof based on the Courant–Fischer theorem: Let us fix $i \leqslant \min\{m, n\}$ such that $\sigma_i(\widetilde{P}\widetilde{\Pi}) > 0$ (otherwise, inequality (31) holds trivially), in particular, $i \leqslant \operatorname{rank}(\widetilde{\Pi})$. Let F_i denote the subspace spanned by the first i right singular vectors of $\widetilde{P}\widetilde{\Pi}$. Since the associated singular values are all nonzero, $\ker \widetilde{\Pi} \cap F_i = \{0\}$ and thereby

$$\dim F_i = \dim\{\widetilde{\Pi} f \mid f \in F_i\} =: \dim \widetilde{\Pi} F_i. \tag{32}$$

Since $\widetilde{\Pi}$ is an orthogonal projection, $\|\widetilde{\Pi} h\|_2 \leqslant \|h\|_2$ for each $h \in \mathbb{R}^n$, and Theorem 27 (Courant–Fischer) implies

$$\sigma_i(\widetilde{P}\widetilde{\Pi}) = \min_{\substack{h \in F_i \\ \|h\|_2 = 1}} \|\widetilde{P}\widetilde{\Pi} h\|_2 \leqslant \min_{\substack{\tilde{h} \in \widetilde{\Pi} F_i \\ \|\tilde{h}\|_2 = 1}} \|\widetilde{P}\tilde{h}\|_2 \leqslant \max_{W \in \mathfrak{W}_i} \min_{\substack{\tilde{h} \in W \\ \|\tilde{h}\|_2 = 1}} \|\widetilde{P}\tilde{h}\|_2 = \sigma_i(\widetilde{P}),$$

where $\mathfrak{W}_i$ denotes the set of i-dimensional subspaces of $\mathbb{R}^n$. This proves the claim. □

6. Relations between the Frobenius norm and the relaxed likelihood objectives

Recall that the degree of coherence $\mathcal{C}_r$ of the full matrix P and the reduced matrix Λ is defined via the singular values of the scaled transition matrices $\widetilde{P} = D_q^{-1/2} P D_p^{1/2}$ and $\widetilde{\Lambda} = D_q^{-1/2} \Lambda D_p^{1/2}$, respectively, and that $\sigma_i(\widetilde{\Lambda}) \leqslant \sigma_i(\widetilde{P})$ for each $i \leqslant \min\{m, n\}$ by Proposition 14.

Noting that the squared Frobenius norm $\|\cdot\|_F^2$ of a matrix equals the sum of its squared singular values, it is therefore natural to measure the discrepancy between full and reduced models by the corresponding difference in Frobenius norm. Theorem 18 below relates $\|\widetilde{P} - \widetilde{\Lambda}\|_F^2$ to the relaxed likelihood $\hat{\ell}$ from (14) that is maximized by DBMR, indicating that DBMR provides a quasi-optimal solution of the (relaxed) coherence problem (8).

Remark 15: In the context of Nonlinear Matrix Factorization, [DLP06, Equations (8)–(10)] show that, assuming small errors and linearizing the objective around the optimum, a maximum-likelihood estimation of nonnegative factor matrices can be connected to χ^2 statistics. This leads them to the minimization of the Frobenius-norm difference of an empirical frequency matrix and its factorized approximation as well as to a connection to the maximum likelihood setting. They do not elaborate this any further, eventually.

More generally, Theorem 18 establishes an a posteriori bound between the two main NMF objectives discussed in section 1, namely the Frobenius norm $\|A - BC\|_F$ and the (generalized[4]) Kullback–Leibler divergence $\mathrm{D}_{\mathrm{KL}}(A \parallel BC)$. For this purpose, we require

[4]The Kullback–Leibler divergence $\mathrm{D}_{\mathrm{KL}}(A \parallel BC)$ is generalized in the sense that A as well as BC represent *unnormalized* probability distributions.

Pinsker-like inequalities for the weighted ℓ^2 norm in Appendix C. These are based on the concept of *balancedness* of a vector $x \in \mathbb{R}^m$ that we introduce in the following. Roughly speaking, we call a vector $x \in \mathbb{R}^m$ balanced if $\|x\|_\infty/\|x\|_1 \ll 1$. Note that the inequality $\|x\|_\infty \leqslant \|x\|_1$ holds true in general and equality only holds for (multiples of) standard unit vectors. On the other hand, the above ratio is minimal if all entries of x have the same modulus, i.e. $x_i = \pm\|x\|_\infty$ for each $i = 1, \dots, m$. In other words, for x to be balanced, the "mass" of the vector (measured by $\|x\|_1$) should not be attributed to one or just a few entries, with the others being zero or close to zero, but should be distributed rather evenly among the entries. More generally, q-balancedness of x indicates that the ratio $|x_i|/q_i$ is close to constant in i, with $q \in \mathbb{R}^m_{>0}$ being a strictly positive probability vector:

Definition 16: For $m \in \mathbb{N}$ and a strictly positive probability vector $q \in \mathbb{R}^m_{>0}$, we define the *balancedness* and the *q-balancedness* of a vector $x \in \mathbb{R}^m$ by

$$\mathfrak{B}(x) := \frac{\|x\|_1}{m\|x\|_\infty} \in [\tfrac{1}{m}, 1], \qquad \mathfrak{B}_q(x) := \frac{\|x\|_1}{\max_i \frac{|x_i|}{q_i}} \in [\min_i q_i, 1], \qquad \text{if } x \neq 0,$$

and by $\mathfrak{B}(0) := \mathfrak{B}_q(0) := 1$.

Remark 17: Note that $\mathfrak{B}(x) = \mathfrak{B}_q(x)$ for the vector $q = (1/m, \dots, 1/m)$.

Theorem 18: Let $\Lambda \in \mathbb{R}^{m\times n}$, $\lambda \in \mathbb{R}^{m\times r}$ and $\Gamma \in \{0,1\}^{r\times n}$ be left stochastic matrices such that $\Lambda = \lambda\Gamma$ and let $\widetilde{P}$ and $\widetilde{\Lambda}$ be given by (6). Further, let $\alpha_j := \alpha(P_{\bullet j}, \Lambda_{\bullet j}) := \frac{2}{3}\max_i \frac{|P_{ij}-\Lambda_{ij}|}{P_{ij}} \in [0, \infty]$, $j = 1, \dots, n$,

$$\kappa_q^{(1)} := \tfrac{1}{2} \min_{j=1,\dots,n} \mathfrak{B}_q(P_{\bullet j} - \Lambda_{\bullet j}),$$
$$\kappa_q^{(2)} := \tfrac{1}{2} \min_{j=1,\dots,n} \mathfrak{B}_q(P_{\bullet j})(1 - \alpha_j),$$

$\kappa^{\mathrm{pr}} := \min_i q_i/2$ and $\kappa^{\mathrm{post}} := \max(\kappa_q^{(1)}, \kappa_q^{(2)})$. Then, $\kappa^{\mathrm{post}} \geqslant \kappa^{\mathrm{pr}}$ and, for $\kappa \in \{\kappa^{\mathrm{pr}}, \kappa^{\mathrm{post}}\}$,

$$\|\widetilde{P} - \widetilde{\Lambda}\|_F^2 \;\leqslant\; \kappa^{-1} \sum_{j=1}^n p_j \mathrm{D}_{\mathrm{KL}}(P_{\bullet j} \parallel \Lambda_{\bullet j}) \;=\; \frac{1}{\kappa S}\left(\hat{\ell}(P, \mathrm{Id}_n) - \hat{\ell}(\lambda, \Gamma)\right), \tag{33}$$

where $\hat{\ell}$ is the DBMR objective given by (14).

Proof. Note that $\kappa^{\mathrm{post}} \geqslant \kappa^{\mathrm{pr}}$ by the definition of $\kappa_q^{(1)}$ and $\mathfrak{B}_q$, so it suffices to to prove (33) for $\kappa = \kappa^{\mathrm{post}}$. Observe that $P_{ij}p_j = N_{ij}/S$ for any $i = 1, \dots, m$ and $j = 1, \dots, n$, and that $\ell(\lambda, \Gamma) = \hat{\ell}(\lambda, \Gamma)$, since Γ is an affiliation matrix, cf. (13)–(14). Hence, Proposition 26 implies (note that we do not need to verify the condition $\alpha_j < 1$ for each $j = 1, \dots, n$,

since, in this case, $\kappa_q^{(2)} < 0$ and $\kappa = \kappa_q^{(1)}$)

$$
\begin{aligned}
\|\widetilde{P} - \widetilde{\Lambda}\|_F^2 &= \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{1}{q_i}(P_{ij} - \Lambda_{ij})^2 p_j \\
&\leqslant \min\left(\sum_{j=1}^{n} p_j \sum_{i=1}^{m} \frac{(P_{ij} - \Lambda_{ij})^2}{q_i} \;,\; \frac{1}{\min_i q_i} \sum_{j=1}^{n} p_j \|P_{\bullet j} - \Lambda_{\bullet j}\|_2^2 \right) \\
&\leqslant \kappa^{-1} \sum_{j=1}^{n} p_j \mathrm{D}_{\mathrm{KL}}(P_{\bullet j} \parallel \Lambda_{\bullet j}) \\
&= \kappa^{-1} \sum_{j=1}^{n} p_j \sum_{i=1}^{m} P_{ij} \log \frac{P_{ij}}{\Lambda_{ij}} \\
&= (\kappa S)^{-1} \left(\sum_{j=1}^{n}\sum_{i=1}^{m} N_{ij} \log P_{ij} - \sum_{i=1}^{m}\sum_{j=1}^{n} N_{ij} \log \sum_{k=1}^{r} \lambda_{ik}\Gamma_{kj} \right) \\
&= (\kappa S)^{-1} \left(\sum_{j=1}^{n}\sum_{i=1}^{m} N_{ij} \sum_{k=1}^{n} \delta_{kj} \log P_{ik} - \sum_{i=1}^{m}\sum_{j=1}^{n} N_{ij} \sum_{k=1}^{r} \Gamma_{kj} \log \lambda_{ik} \right) \\
&= (\kappa S)^{-1} \Big(\hat{\ell}(P, \mathrm{Id}_n) - \hat{\ell}(\lambda, \Gamma) \Big).
\end{aligned}
$$

□

For the interpretation of this result, a few remarks are in order.

Remark 19:

(a) Note that (33) provides a (weaker) *a priori* bound for $\kappa = \kappa^{\mathrm{pr}}$ and a (sharper) *a posteriori* estimate for $\kappa = \kappa^{\mathrm{post}}$ due to its dependence on the solution Λ of the DBMR problem.

(b) In the proof of Theorem 18 we used only two of the four inequalities established in Proposition 26. Using all four inequalities, defining $\kappa := \max(\kappa^{(1)}, \kappa_q^{(1)}, \kappa^{(2)}, \kappa_q^{(2)})$ with

$$
\begin{aligned}
\kappa^{(1)} &:= \tfrac{m}{2} \min_{i=1,\dots,m} q_i \min_{j=1,\dots,n} \mathfrak{B}(P_{\bullet j} - \Lambda_{\bullet j}), \\
\kappa^{(2)} &:= \tfrac{m}{2} \min_{i=1,\dots,m} q_i \min_{j=1,\dots,n} \mathfrak{B}(P_{\bullet j})(1 - \alpha_j),
\end{aligned}
$$

one would obtain a seemingly sharper bound in (33). However,

$$
\begin{aligned}
\kappa_q^{(1)} &= \tfrac{1}{2} \min_{j=1,\dots,n} \frac{\|P_{\bullet j} - \Lambda_{\bullet j}\|_1}{\max_i \frac{|P_{ij} - \Lambda_{ij}|}{q_i}} \geqslant \frac{\min_i q_i}{2} \min_j \frac{\|P_{\bullet j} - \Lambda_{\bullet j}\|_1}{\|P_{\bullet j} - \Lambda_{\bullet j}\|_\infty} = \kappa^{(1)}, \\
\kappa_q^{(2)} &= \tfrac{1}{2} \min_j \frac{1 - \alpha_j}{\max_i \frac{P_{ij}}{q_i}} \qquad\qquad \geqslant \frac{\min_i q_i}{2} \min_j \frac{1 - \alpha_j}{\|P_{\bullet j}\|_\infty} \qquad = \kappa^{(2)},
\end{aligned}
$$

so this would not lead to an improvement over Theorem 18.

(c) Clearly, the higher the q-balancedness $\mathfrak{B}_q(P_{\bullet j})$ of $P_{\bullet j}$ in the formula of $\kappa_q^{(2)}$ is for each $j = 1, \dots, n$, the sharper the inequality (33) becomes. Note that this balancedness is large (for fixed j) if $P_{\bullet j} \approx q$. The dynamic interpretation of $P_{\bullet j} \approx q$ is that the state j is mapped to a distribution that is close to the final distribution q. If that is true for every j, then there is little coherence in the system, as $P \approx q\mathbb{1}_{[n]}^\top$, with singular values $\sigma_1 = 1$ and $\sigma_n \approx 0$, $n \geqslant 2$. In contrast, $\kappa_q^{(1)}$ is large if the q-balancedness of the *difference*, $\mathfrak{B}_q(P_{\bullet j} - \Lambda_{\bullet j})$, is large for each $j = 1, \dots, n$. On the one hand, this seems to be a less restrictive requirement than the previous one. On the other hand, it is harder to characterize a priori, as all we know about the columns of P and Λ is that they are probability vectors, hence their difference has zero mean.

(d) DBMR maximizes $\hat{\ell}(\lambda, \Gamma)$ over all pairs of stochastic matrices of given fixed dimensions, cf. Problem 2. Thus, within the bound given in Theorem 18, DBMR minimizes the Frobenius norm of the difference between the full model and the low rank model. By the Eckart–Young–Mirsky theorem [HE15, Theorem 4.4.7], the best rank-r approximation of a matrix with respect to the Frobenius norm is given by the composition of the leading r singular modes of the matrix, cf. (10). Theorem 18 thus states that the optimal DBMR solution is a quasi-optimal approximation of the leading r singular modes of $\widetilde{P}$, and hence to the coherence problem, as discussed in section 3. In particular, the bound in (33) is zero if $P = \Lambda$, as any reasonably tight bound of $\|\widetilde{P} - \widetilde{\Lambda}\|_F$ should be.

(e) The seeming dependence of (33) on S is deceptive, since $\hat{\ell}$ itself scales "linearly" with S. More precisely, the right-hand side of the bound converges almost surely as $S \to \infty$, if the data acquisition procedure is such that $\frac{1}{S}N$ converges almost surely for $S \to \infty$. This is obvious from (14). The i.i.d. sampling procedure assumed in section 2 satisfies this condition by the law of large numbers.

(f) We note that a related "balancedness" concept plays a role in a different a posteriori refinement of Pinsker's inequality [OW05, Theorem 2.1], relating the total variation distance and the Kullback–Leibler divergence.

An alternative interpretation to Theorem 18 arises by invoking Corollary 13.

Corollary 20: Under the assumptions of Theorem 18,

$$\mathcal{C}_r(\Lambda) \geqslant \frac{1}{\kappa S}\Big(\hat{\ell}(\lambda, \Gamma) - \hat{\ell}(P, \mathrm{Id}_n)\Big) + \big\|\widetilde{P}\big\|_F^2 . \tag{34}$$

Proof. Using Corollary 13 and noting that all singular values of $\widetilde{\Lambda}$ satisfy $\sigma_i(\widetilde{\Lambda}) \in [0, 1]$, we obtain

$$\mathcal{C}_r(\Lambda) = \sum_{i=1}^{r} \sigma_i(\widetilde{\Lambda}) \geqslant \sum_{i=1}^{r} \sigma_i^2(\widetilde{\Lambda}) = \big\|\widetilde{\Lambda}\big\|_F^2 = \|\widetilde{P}\|_F^2 - \|\widetilde{P} - \widetilde{\Lambda}\|_F^2.$$

The claim follows directly from Theorem 18. □

Therefore, relying on the a priori error bound in (33) with $\kappa = \kappa^{\mathrm{pr}}$, an increase of the DBMR objective $\hat{\ell}(\lambda, \Gamma)$ results in a sharper lower bound on the degree of coherence in $\Lambda = \lambda\Gamma$.

7. Numerical examples

We will consider two examples and compare the performance of Algorithm 2 implementing DBMR [GH17] which is available as open access, see section 9, with Algorithm 1 implementing the classical approach to coherence. In the first example we model a transition matrix with two perfectly coherent partition elements where one of these elements can again be subdivided into two strongly, but not perfectly coherent sets. The second example is a discrete version of a map with three large (and several small) coherent sets; see [FLS10, Example 1]. In each example we consider three different perturbations of the transition matrix: unperturbed ($\varepsilon = 0$), slightly perturbed ($\varepsilon = 2$ or 1) and strongly perturbed ($\varepsilon = 10$ or 4) in the following sense: Each data point $(x, y) \in \boldsymbol{D}$ is replaced by a uniform random point from the set

$$\big((x + \{-\varepsilon, \dots, \varepsilon\}) \bmod n\big) \times \big((y + \{-\varepsilon, \dots, \varepsilon\}) \bmod n\big). \tag{35}$$

Note that we assume the states to be ordered periodically, i.e., states 1 and n are adjacent. For DBMR we perform 100 independent runs with randomly generated initial affiliation matrices $\Gamma^{(0)}$ (i.e., the columns of this matrix are independent uniform random samples of the r canonical unit vectors) and the best result in terms of the DBMR objective (14) is taken. The following criteria for coherence are considered for the comparison between DBMR and the classical approach with $r = 3$ latent states:

(a) The second and third singular values σ_2, σ_3 (note that $\sigma_1 = 1$ by construction) of the low rank projected (and reweighted) transition matrices $\widetilde{\Lambda}$ (of DBMR) and of $\widetilde{P}_{\mathrm{red}}$ of the classical approach (note that, by construction, the latter coincide with the ones of the full rank transition matrix $\widetilde{P}$) are presented in Tables 1 and 2.

(b) The DBMR objective (14) is evaluated for the resulting affiliation matrices 'DBMR-Γ' and 'SVD-Γ' (which naturally correspond to partitions by (4)) of Algorithms 1 and 2, and compared to the 'default' affiliation matrix 'default-Γ' given by our construction of the example (see the descriptions below). For this purpose, the corresponding matrix λ is chosen to maximize $\hat{\ell}(\cdot, \Gamma)$ in (14) and, hence, is given by (15). In addition, we compare these values to the 'reference value' $\hat{\ell}(P, \mathrm{Id}_n)$ of the unreduced model. The corresponding values are presented in Tables 1 and 2.

(c) Finally, we compare the objectives of the different model reduction tools by considering the tightness of the bound in Theorem 18. Recall from Corollary 20 that the quantity $\|\widetilde{P} - \widetilde{\Lambda}\|_F^2$ is approximately monotonic in the degree of coherence $\mathcal{C}$ introduced in Definition 2. More precisely, the larger $\mathcal{C}(\Lambda)$, the smaller $\|\widetilde{P} - \widetilde{\Lambda}\|_F^2$. Tables 1 and 2 show both sides of the inequality (33) for Λ obtained in the best DBMR run. In addition, Figure 5 in Appendix E illustrates in how far the two objectives $\|\widetilde{P} - \widetilde{\Lambda}\|_F^2$ and $\hat{\ell}$ are in line by comparing their values for a large number of corresponding pairs (λ, Γ).

(d) A visual comparison is performed in Figures 1 and 2 by plotting the transition matrix P and its reduced versions P_{red} (SVD)[5] and $\Lambda = \lambda\Gamma$ (DBMR), where larger transition

[5]Note that P_{red} can have negative entries, so it need not be a transition matrix.

probabilities correspond to darker shades of gray. In addition, the partitions of the input states corresponding to the respective affiliation matrices (default-Γ, SVD-Γ and DBMR-Γ) are color-coded in yellow, green and red on the bottom line of the matrix images.

Example 21 (three coherent sets)**:** Our first example is an idealized dynamics having two perfectly coherent sets, one of which can further be subdivided into two less coherent sets. We take $n = m = 100$ input and output states and define the coherent sets $E_1 = \{1, \dots, 25\}$, $E_2 = \{26, \dots, 50\}$, and $E_3 = \{51, \dots, 100\}$ which partition both $[n]$ and $[m]$. The data set $\boldsymbol{D}$ consists of $S = 25000$ pairs (X_u, Y_u), $u = 1, \dots, S$, and is constructed such that

$$N_{ij} = \begin{cases} 8, & i,j \in E_1 \text{ or } i,j \in E_2, \\ 2, & i \in E_1, j \in E_2 \text{ or } i \in E_2, j \in E_1, \\ 5, & i,j \in E_3, \\ 0, & \text{otherwise.} \end{cases}$$

Hence, there are $\sum_{i=1}^{m} N_{ij} = 250$ transitions out of every state. As discussed above, we also consider two perturbed version of the above data given by (35) for $\varepsilon = 2, 10$.

The resulting coherence criteria (a), (b), (c) and (d) described above are summarized in Table 1 and visualized in Figure 1. Note that the rank of the unperturbed transition matrix P is $r = 3$, allowing the truncated SVD to match the exact transition matrix. Also, since P has only three different columns, the DBMR result with $r = 3$ latent states coincides with P (cf. Example 9). As expected, in both the full and the reduced models, coherence (measured by the singular values) as well as the values of $\hat{\ell}$, are decreasing with increasing perturbation strength. We observe that, for unperturbed and slightly perturbed data, the reduced models as well as the partitions visualized in Figure 1 are rather similar, as are the singular values σ_2, σ_3 and the values of $\hat{\ell}$ in Table 1. On the other hand, for strong perturbations ($\varepsilon = 10$), we report larger differences in all of the above criteria, suggesting that the two objectives $\mathcal{C}_3$ and $\hat{\ell}$ are not entirely aligned. Finally, we consider the tightness of the inequality (33) in Table 1 (c). Since the transition matrix P in the unperturbed case has rank three, DBMR is exact and both values are below machine precision. We observe for the perturbed cases that there is a factor of 10 between the left-hand side and right-hand side of the inequality, which could be indicative of the different nature of the objectives that are optimized in Problem 1 and Problem 2.

Example 22 (piecewise expanding interval map)**:** In our second example, the number of input and output states equals $n = m = 90$ with 90 transitions out of every state, totaling $S = 8100$ data points. Each input state j is paired, with equal frequencies, with three output states i such that $N_{ij} = 30$ for the corresponding pairs. We do not explicitly write down how these three output states are chosen, but instead refer the reader to Figure 2 (top left) as well as to [FLS10, Example 1], which this example was inspired by. The sets $E_1 = \{1, \dots, 30\}$, $E_2 = \{31, \dots, 60\}$, and $E_3 = \{61, \dots, 90\}$ are perfectly coherent (their output 'partners' being E_2, E_3, and E_1, respectively). There are also smaller perfectly coherent sets, for instance $\{61, 80, 81\}$, of which the output

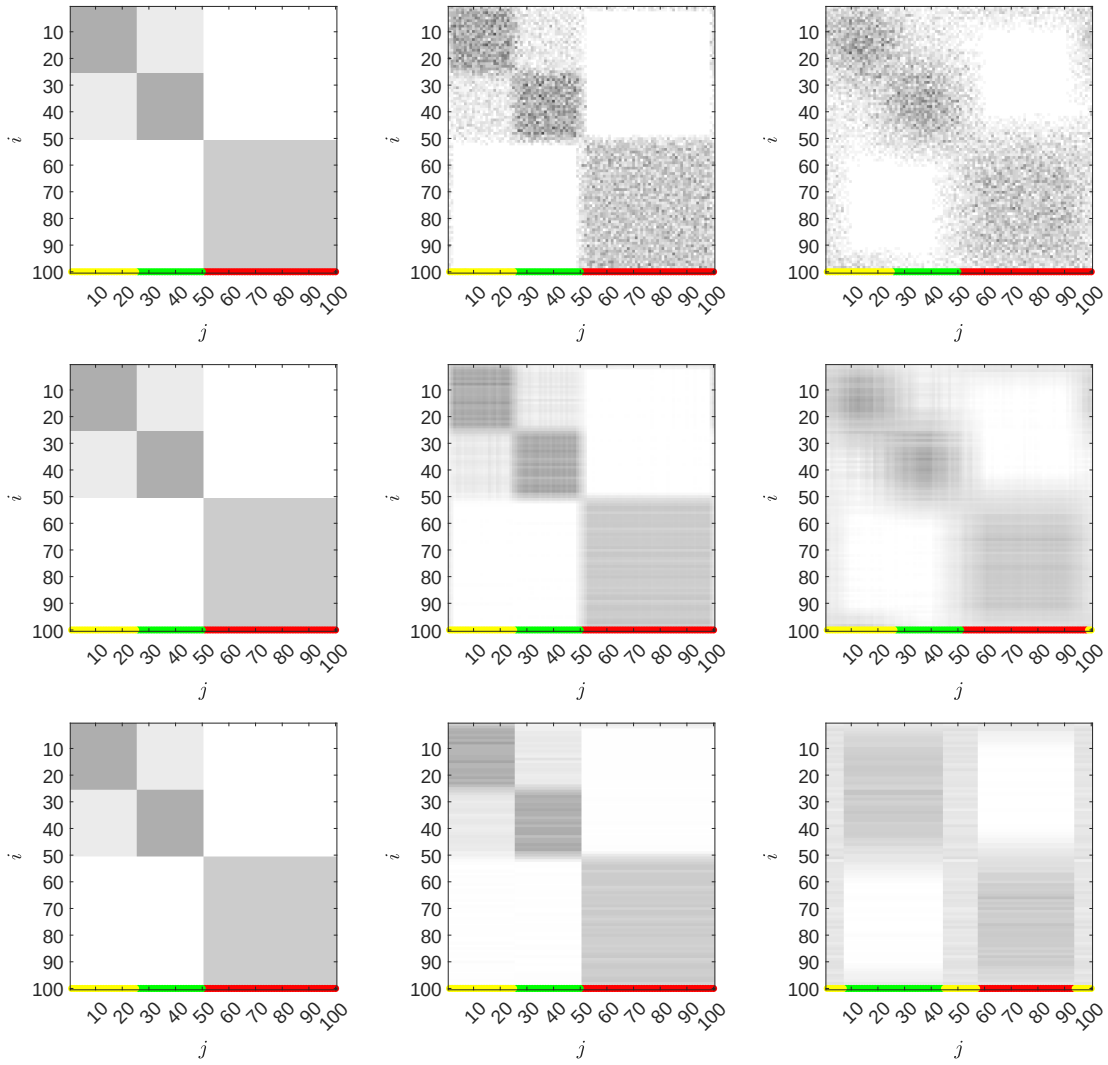

Figure 1: Coherent set identification for Example 21 with $r = 3$ clusters and 3 different sizes of perturbation: *left:* unperturbed; *center:* slightly perturbed, *right:* strongly perturbed. *Top:* Full transition matrix P. *Middle:* Reduced transition matrix P_{red} obtained within the classical Algorithm 1. *Bottom:* Reduced transition matrix $\Lambda = \lambda\Gamma$ of the DBMR Algorithm 2. The coloring on the bottom line of each plot corresponds to the clustering given by the corresponding affiliation matrix Γ (or partition $\mathcal{E}$), i.e. default-Γ (top), SVD-Γ (middle) and DBMR-Γ (bottom).

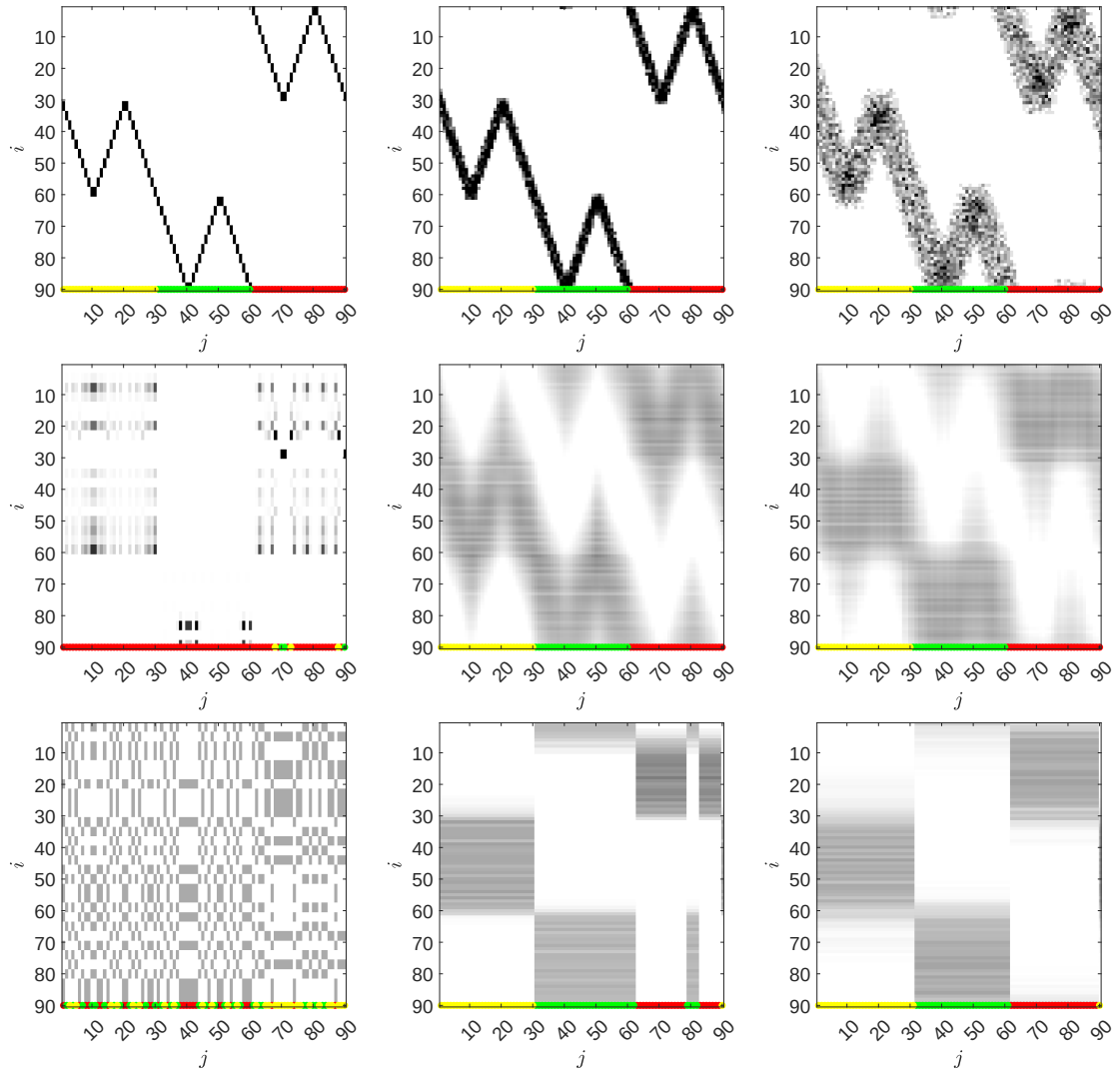

Figure 2: Coherent set identification for Example 22 with $r = 3$ clusters and 3 different sizes of perturbation: *left:* unperturbed; *center:* slightly perturbed, *right:* strongly perturbed. *Top:* Full transition matrix P. *Middle:* Reduced transition matrix P_{red} obtained within the classical Algorithm 1. *Bottom:* Reduced transition matrix $\Lambda = \lambda\Gamma$ of the DBMR Algorithm 2. The coloring on the bottom line of each plot corresponds to the clustering given by the corresponding affiliation matrix Γ (or partition $\mathcal{E}$), i.e. default-Γ (top), SVD-Γ (middle) and DBMR-Γ (bottom).

	Perturbation		$\varepsilon = 0$		$\varepsilon = 2$		$\varepsilon = 10$	
(a)	$\sigma_2(\widetilde{P})$	$\sigma_3(\widetilde{P})$	1.000	0.600	0.939	0.545	0.725	0.362
	$\sigma_2(\widetilde{\Lambda})$	$\sigma_3(\widetilde{\Lambda})$	1.000	0.600	0.918	0.528	0.702	0.071
(b)	$\hat{\ell}(\lambda, \Gamma)$ for default-Γ		$-0.954 \cdot 10^5$		$-0.997 \cdot 10^5$		$-1.077 \cdot 10^5$	
	$\hat{\ell}(\lambda, \Gamma)$ for SVD-Γ		$-0.954 \cdot 10^5$		$-0.997 \cdot 10^5$		$-1.076 \cdot 10^5$	
	$\hat{\ell}(\lambda, \Gamma)$ for DBMR-Γ		$-0.954 \cdot 10^5$		$-0.997 \cdot 10^5$		$-1.072 \cdot 10^5$	
	reference value $\hat{\ell}(P, \mathrm{Id}_n)$		$-0.954 \cdot 10^5$		$-0.951 \cdot 10^5$		$-1.012 \cdot 10^5$	
(c)	$\|\widetilde{P} - \widetilde{\Lambda}\|_F^2$		$5.2083 \cdot 10^{-31}$		0.4123		0.5443	
	$\kappa^{-1} \sum_{j=1}^n p_j \mathrm{D}_{\mathrm{KL}}(P_{\bullet j} \parallel \Lambda_{\bullet j})$		$8.2991 \cdot 10^{-17}$		4.3351		3.9866	
	κ		$0.1562^{(2)}$		$0.0424^{(1)}$		$0.0621^{(1)}$	

Table 1: Coherence criteria (a), (b) and (c) discussed above for the comparison of the classical approach to coherence (Algorithm 1) and DBMR (Algorithm 2). In the last row, the superscript 1 or 2 indicates whether $\kappa = \kappa_q^{(1)}$ or $\kappa = \kappa_q^{(2)}$.

'partner' is $\{1, 2, 3\}$. There are, in fact, 30 such 3-element coherent sets, and arbitrary unions of them are also perfectly coherent. Note that the smaller a coherent set, the more its coherence will be affected by the perturbations. For the small perturbation we use $\varepsilon = 1$ and for the large one we use $\varepsilon = 4$, cf. Figure 2 (top middle and right).

The coherence criteria (a), (b), (c) and (d) described above are reported in Table 2 and visualized in Figure 2. In the unperturbed case, Algorithms 1 and 2 both identified perfectly coherent sets that we comment on in Remark 23 below. In the slightly perturbed case, the $\hat{\ell}$-value of DBMR-Γ was worse than the ones of default-Γ and SVD-Γ. This shows that, even with a large number of 100 independent runs, DBMR was incapable of identifying the global optimum of $\hat{\ell}$, which we attribute to the large number of small coherent sets (in the unperturbed case), presumably resulting in a large number of local optima. We did not observe this issue in the strongly perturbed case, where both Algorithms 1 and 2 identified the default partition. This suggests that the perturbation of $\varepsilon = 4$ was sufficient to 'smoothen out' many of the local optima. We also point out that, compared to Example 21, the inequality (33) is sharper — the deviating factor is between 2 and 8 rather than 10.

Remark 23: In Example 22 with no perturbation, visualized in Figure 2 (left), there is a large number of small perfectly coherent sets. Hence, each partitioning of these sets into three groups will again produce perfectly coherent sets. In that sense, both the classical Algorithm 1 and the DBMR Algorithm 2 identify perfectly coherent sets (we verified that the sum of the leading three singular values of $\widetilde{\Lambda} = D_q^{-1/2} \lambda \Gamma D_p^{1/2}$, i.e. the

	Perturbation		$\varepsilon = 0$		$\varepsilon = 1$		$\varepsilon = 4$	
(a)	$\sigma_2(\widetilde{P})$	$\sigma_3(\widetilde{P})$	1.0000	1.0000	0.9849	0.9846	0.8961	0.8948
	$\sigma_2(\widetilde{\Lambda})$	$\sigma_3(\widetilde{\Lambda})$	1.0000	1.0000	0.9548	0.9234	0.8518	0.8400
(b)	$\hat{\ell}(\lambda, \Gamma)$ for default-Γ		$-0.2755 \cdot 10^5$		$-0.2828 \cdot 10^5$		$-0.3002 \cdot 10^5$	
	$\hat{\ell}(\lambda, \Gamma)$ for SVD-Γ		$-0.3212 \cdot 10^5$		$-0.2828 \cdot 10^5$		$-0.3002 \cdot 10^5$	
	$\hat{\ell}(\lambda, \Gamma)$ for DBMR-Γ		$-0.2755 \cdot 10^5$		$-0.2861 \cdot 10^5$		$-0.3002 \cdot 10^5$	
	reference value $\hat{\ell}(P, \mathrm{Id}_n)$		$-0.0890 \cdot 10^5$		$-0.1817 \cdot 10^5$		$-0.2531 \cdot 10^5$	
(c)	$\|\widetilde{P} - \widetilde{\Lambda}\|_F^2$		27.000		7.5525		2.0543	
	$\kappa^{-1} \sum_{j=1}^n p_j \mathrm{D}_{\mathrm{KL}}(P_{\bullet j} \parallel \Lambda_{\bullet j})$		69.0776		40.3973		15.6709	
	κ		$0.0333^{(1)}$		$0.0319^{(1)}$		$0.0370^{(1)}$	

Table 2: Coherence criteria (a), (b) and (c) discussed above for the comparison of the classical approach to coherence (Algorithm 1) and DBMR (Algorithm 2). In the last row, the superscript 1 or 2 indicates whether $\kappa = \kappa_q^{(1)}$ or $\kappa = \kappa_q^{(2)}$.

degree of coherence $\mathcal{C}_3(p, \lambda\Gamma)$, has the maximal possible value of $r = 3$, (λ, Γ) being the DBMR output), showing that this DBMR result is not inferior to the classical one with respect to our measure of coherence. Furthermore, DBMR performs a partitioning into groups of equal size (30 states each), cf. Figure 2 (bottom left), while the group sizes resulting from the classical approach (namely 84/3/3) strongly differ, cf. Figure 2 (middle left). As argued by [FSM10] in the context of $r = 2$ coherent sets, equal group size is a preferable property in terms of coherence. In fact, [FSM10, Section III.A] *imposes* the two coherent sets to have approximately the same mass. In that sense, the DBMR result is preferable to the one of the classical approach. Note that this preferable property of coherent sets having large size is not reflected by our measures of coherence, namely the objective in (3) and the degree of coherence in Definition 2.

Classical approach versus DBMR for coherent set identification. In this manuscript, we have compared analytically as well as empirically the performance of two approaches to identify coherent sets of dynamical systems — the classical approach (Algorithm 1) and DBMR (Algorithm 2). Let us shortly summarize the advantages and disadvantages of using DBMR for this task:

On the one hand, DBMR performs worse in terms of coherence *given our measure of coherence* (see Definition 2). This is almost a tautology since the classical approach by construction maximizes the degree of coherence $\mathcal{C}_r$, whilst DBMR optimizes a different objective. Furthermore, DBMR comes with the risk of running into local maxima of $\hat{\ell}$, cf. Example 22 with slight perturbation.

On the other hand, DBMR seems to promote large coherent sets (an attractive property in many settings of applied interest), whilst the classical approach might identify coherent sets that are very small, cf. Remark 23 and Example 22 with no perturbation. In this light, the DBMR objective might be preferred, but future work developing a systematic comparison between optimization objectives for coherent sets is needed. In this direction, we conjecture that the entropic characterization of DBMR (see Remark 6) can be shown to provide a theoretical foundation for the observed size-sensitive properties of DBMR.

The 'reduced transition matrix' λ (operating on the compound input states grouped by Γ) of DBMR is a left stochastic matrix. Therefore, DBMR is structure preserving in this sense, while the 'reduced transition matrix' $\widetilde{P}_{\mathrm{red}}$ of the classical approach can have negative entries and its columns typically do not sum to one. A further advantage of DBMR might be that it does not require the approximation of the entire mn entries of the "full" transition matrix P; instead it estimates the $r(n+m)$ entries (essentially, even only $rm+n$ entries, since Γ has only one nonzero entry per column) of the factors λ and Γ directly from the data. [GH17] argue that the (comparatively few) matrix entries of the low rank approximation require far less data. However, our experiments have neither verified nor refuted this intuition.

8. Conclusion and Outlook

In this paper, we have suggested and analyzed the application of Direct Bayesian Model Reduction (DBMR; Algorithm 2, [GH17]) for the identification of coherent sets and compared it with the classical approach based on truncated singular value decomposition (Algorithm 1). Both approaches perform a certain factorization of a matrix $A \approx BC$ into low rank matrices B, C, but maximize two different objectives, namely the 'degree of coherence' as the sum of the leading singular values, corresponding to the minimization of the Frobenius norm $\|A - BC\|_F$, and the relaxed likelihood $\hat{\ell}$ from (14), connected to maximum likelihood estimation and minimization of the (generalized) Kullback–Leibler divergence $\mathrm{D}_{\mathrm{KL}}(A \parallel BC)$. Therefore, on a broader scale, our contributions also establish connections between these two central minimization problems for matrix factorization.

The above-mentioned comparison is based on two central results, Theorems 11 and 18. The first shows that the DBMR output $\Lambda = \lambda\Gamma$ can be written as a composition of the full model P with an orthogonal projection Π, $\Lambda = P\Pi$. While this is insightful in its own right, it also gives us the necessary tools to derive bounds on the degree of coherence of the reduced model Λ in Proposition 14. The second theorem establishes a connection between the Frobenius norm distance and the Kullback–Leibler divergence mentioned above, which, to the best of our knowledge, is the first relationship of this kind. For this purpose, we have derived certain Pinsker-type inequalities for the (weighted) ℓ^2 norm in Appendix C, which might be of independent interest.

In our numerical experiments, DBMR was able to identify meaningful coherent sets. It is well known that DBMR can get stuck in local maxima of its objective function, which we also observed (Example 22 with slight perturbation $\varepsilon = 1$) even though we used

a large number 100 of independent runs of DBMR. The singular values, and thereby the degree of coherence of the corresponding reduced models was slightly inferior to the classical approach, which is hardly surprising since the classical approach optimizes precisely this objective. However, the additional computations in Appendix E, and in particular Figure 5, show that the two objectives are mostly aligned, backing up our theoretical findings from Theorem 18.

An important advantage of DBMR over the classical approach is that its low rank model $\Lambda = \lambda\Gamma$ is a product of a left stochastic matrix λ and an affiliation matrix Γ, which has a clear probabilistic interpretation of a reduced transition matrix λ that operates on compound states clustered by Γ. This structure preservation is missed by the classical approach, where the 'reduced transition matrix' $\widetilde{P}_{\mathrm{red}}$ can have negative entries and the columns typically do not sum to one.

The connection between coherence—understood widely as subgroups of states subject to similar evolution—and matrix factorization remains an active field of research and various future directions are imaginable. It is arguable whether the sum of the leading singular values of $\widetilde{P}$ is a good measure for the 'degree of coherence', see the discussion towards the end of section 7. Establishing and optimizing other objectives, such as the DBMR objective $\hat{\ell}$ from (14), and analyzing connections between these objectives would deepen our theoretical understanding of both matrix factorization and the study of coherent structures. The coherence problem (3) has a symmetry in the sense that matching partitions of *both* input and output space are sought. In contrast, DBMR in its current form gives merely partitions of the input space. Future research could hence address the development of efficient 'symmetrized' versions of DBMR.

9. Code availability

MATLAB code for the Bayesian-Model-Reduction-Toolkit [GH17] is available at

https://github.com/SusanneGerber/Bayesian-Model-Reduction-Toolkit,

and the adaptation for this paper is available at

https://github.com/RobertMaltePolzin/DBMR_Coherence.

10. Acknowledgment

The authors thank Robin Chemnitz, Mattes Mollenhauer, Rupert Klein, Peter Névir, and Niklas Wulkow for discussions and helpful suggestions. This research has been partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the grant CRC 1114 'Scaling Cascades in Complex Systems', Project Number 235221301, projects A01 "Coupling a multiscale stochastic precipitation model to large scale atmospheric flow dynamics", A02 "Multiscale data and asymptotic model assimilation for atmospheric flows" and A08 "Characterization and prediction of quasi-stationary atmospheric states". The research of IK was funded by the DFG under

Germany's Excellence Strategy (EXC-2046/1, project 390685689) through the project EF1-10 of the Berlin Mathematics Research Center MATH+.

A. The coherence problem: From continuous to discrete space

The coherence problem that we consider here has one of its roots in fluid dynamics. There, the (Lagrangian) evolution of passive tracers advected by the flow field is described by a flow $\phi^{t,\tau} : \Omega \to \Omega$ on some (mostly two or three-dimensional) spatial domain Ω. The flow is nonautonomous, and $\phi^{t,\tau}$ denotes the dynamical evolution from time t to $t+\tau$. Of particular interest are non-trivial subsets $A \subset \Omega$ that "evolve coherently" under the flow on some time interval $[t, t+\tau]$, meaning that sets $\phi^{t,s}(A)$, $0 \leqslant s \leqslant \tau$, only experience minimal filamentation. In other words, the flow does not "disperse" the set A.

The setting can be simplified by considering the flow at only discrete time instances. For our purposes, only two time instances are enough, say t and $t+\tau$. The mapping $T := \phi^{t,\tau}$ does not need to leave any set in some state space invariant, hence the states at initial and final time can belong to different sets. Formally, the dynamics thus boils down to a mapping $T : \Omega_1 \to \Omega_2$. We assume that Ω_1, Ω_2 are measurable spaces, T is a measurable map, and suppress the underlying sigma algebras.

The coherence problem for the map T can now be vaguely stated as the task to find nontrivial subsets $E \subset \Omega_1$ and $F \subset \Omega_2$ such that $T(E) \approx F$ and E, F are relatively 'simple' in terms of their geometry and balancedness. The latter can be made precise by requiring that the relation $T(E) \approx F$ persists under slight (random) perturbations. We refer to [FSM10] for further details. The sets E, F are then called a *finite time coherent pair*. There are precise functional-analytic formulations of this problem available in [Fro13, Fro15]. Instead of taking this route, we can discretize the dynamics first and state the coherence problem in the discrete setting directly. This was done in [FSM10] to arrive at a problem that is numerically accessible via matrix analysis. The same setting arises in situations where a precise observation of the state of the system is not possible and only quantized (discrete) observations are performed.

To get to the discrete setting, we subdivide the subsets Ω_1 and Ω_2 into collections of mutually disjoint partition elements $\{B_1, ..., B_n\}$ and $\{C_1, ..., C_m\}$, respectively. We assume that the initial state X is an Ω_1-valued random variable with law μ; thus μ is a probability measure supported on Ω_1. Let ν denote the pushforward of μ by T, i.e., the law of $Y := T(X)$. We then define the *transition matrix* $P \in \mathbb{R}^{m \times n}$ by

$$P_{ij} = \frac{\mu(B_j \cap T^{-1}(C_i))}{\nu(B_j)} = \mathbb{P}\left[Y \in C_i \mid X \in B_j\right]. \tag{36}$$

Note that P is left stochastic, i.e., $P_{ij} \geqslant 0$ for all i, j, and $\sum_{i=1}^{m} P_{ij} = 1$ for all j. We further define the discrete distributions at initial and final times by $p \in \mathbb{R}^n$ and $q \in \mathbb{R}^m$, respectively, with

$$p_j = \mu(B_j), \quad j = 1, \dots, n, \qquad q_i = \nu(C_i), \quad i = 1, \dots, m.$$

It follows that the discrete initial distribution is mapped to the discrete final one by the transition matrix,

$$q = Pp. \tag{37}$$

We also assume that $p > 0$ and $q > 0$, componentwise. If not, the associated partition elements are removed from the sets B_j and C_i, respectively (and the sets Ω_1 and Ω_2 are restricted accordingly).

The transition matrix P together with the (initial) distribution p characterize a one-step random transition that jumps from some element B_j of the initial partition to some element C_i of the final partition. This way, if one were to consider a sequence of partitions with diameter converging to zero, the associated sequence of transition matrices P would constitute a *small random perturbation* [Kif86] of T, see [Fro98]. Thus, formulating the coherence problem for this discrete dynamics (see main text) will automatically deliver coherent sets that are robust with respect to small random perturbations of T.

Finally, we note that approximating dynamical properties of T through the discretization (36) is attributed to Ulam [Ula60].

B. Proof of Theorem 11

In this section, we present the proof of Theorem 11. Here, $\gamma\colon [n] \to [r]$ denotes the assignment corresponding to Γ (cf. Definition 1).

Lemma 24: The matrix $\widetilde{\Pi} \in \mathbb{R}^{n\times n}$ given by (23) is symmetric, $\widetilde{\Pi}^\top = \widetilde{\Pi}$, and a projection, $\widetilde{\Pi}^2 = \widetilde{\Pi}$. Consequently, Π is a p^{-1}-orthogonal projection (for p^{-1}-symmetry, see (24)).

Proof. Symmetry of $\widetilde{\Pi}$ follows directly from its definition in (23) and the fact that $\widetilde{\Pi}_{ij} \neq 0$ if and only if $\gamma(i) = \gamma(j)$. In order to show that $\widetilde{\Pi}$ is a projection we compute

$$(\widetilde{\Pi}^2)_{ij} = \sum_k \frac{\sqrt{p_i p_k}\,\delta_{\gamma(i)\gamma(k)}}{\sum_l p_l \delta_{\gamma(l)\gamma(k)}} \frac{\sqrt{p_k p_j}\,\delta_{\gamma(k)\gamma(j)}}{\sum_{l'} p_{l'} \delta_{\gamma(l')\gamma(j)}} = \frac{\sqrt{p_i p_j}\,\delta_{\gamma(i)\gamma(j)}}{\sum_l p_l \delta_{\gamma(l)\gamma(i)}} \sum_k \frac{p_k\,\delta_{\gamma(k)\gamma(j)}}{\sum_{l'} p_{l'} \delta_{\gamma(l')\gamma(j)}} = \widetilde{\Pi}_{ij},$$

where we use the fact that nonzero terms in the first sum require $\gamma(i) = \gamma(j) = \gamma(k)$. It follows that Π is a p^{-1}-orthogonal projection:

$$\begin{aligned} \Pi^2 &= (D_p^{1/2}\widetilde{\Pi} D_p^{-1/2})^2 = D_p^{1/2}\widetilde{\Pi}^2 D_p^{-1/2} = \Pi, \\ \langle u, \Pi v\rangle_{p^{-1}} &= \langle u, D_p^{-1/2}\widetilde{\Pi} D_p^{-1/2} v\rangle_2 = \langle D_p^{1/2}\widetilde{\Pi} D_p^{-1/2} u, D_p^{-1} v\rangle_2 = \langle \Pi u, v\rangle_{p^{-1}}, \qquad u, v \in \mathbb{R}^n. \end{aligned}$$

□

Proof of Theorem 11. Π is left stochastic by definition and a p^{-1}-orthogonal projection by Lemma 24. Since $N_{ij} = SP_{ij}p_j$ and $\sum_{i=1}^m N_{ij} = Sp_j$ by (5) as well as $\Gamma_{kj} = \delta_{k\gamma(j)}$ by Definition 1, equation (15) implies

$$\lambda_{ik} = \frac{\sum_{j=1}^n \Gamma_{kj} N_{ij}}{\sum_{i'=1}^m \sum_{j'=1}^n \Gamma_{kj'} N_{i'j'}} = \frac{\sum_{j=1}^n \delta_{k\gamma(j)} P_{ij} p_j}{\sum_{j'=1}^m \delta_{k\gamma(j')} p_{j'}}.$$

Hence,

$$(\lambda\Gamma)_{il} = \sum_{k=1}^{r} \left(\frac{\sum_{j=1}^{n} \delta_{k\gamma(j)} P_{ij} p_j}{\sum_{j'=1}^{m} \delta_{k\gamma(j')} p_{j'}} \delta_{k\gamma(l)} \right) = \sum_{j=1}^{n} P_{ij} \frac{p_j \, \delta_{\gamma(j)\gamma(l)}}{\sum_{j'} p_{j'} \, \delta_{\gamma(l)\gamma(j')}} = (P\Pi)_{il},$$

proving $\lambda\Gamma = P\Pi$. The eigenvector properties in (b) follow from

$$\left(\Pi a^{(k)}\right)_i = \sum_{j=1}^{n} \frac{p_i \delta_{\gamma(i)\gamma(j)}}{\sum_l p_l \delta_{\gamma(l)\gamma(i)}} p_j \delta_{\gamma(j)k} = p_i \delta_{\gamma(i)k} \sum_{j=1}^{n} \frac{p_j \, \delta_{\gamma(i)\gamma(j)}}{\sum_l p_l \delta_{\gamma(l)\gamma(i)}} = a_i^{(k)},$$

$$(\Pi p)_i = \sum_{j=1}^{n} \frac{p_i \delta_{\gamma(i)\gamma(j)}}{\sum_l p_l \delta_{\gamma(l)\gamma(i)}} p_j = p_i \sum_{j=1}^{n} \frac{p_j \, \delta_{\gamma(i)\gamma(j)}}{\sum_l p_l \delta_{\gamma(l)\gamma(i)}} = p_i,$$

where we use the fact that nonzero terms in the first sums require $\gamma(i) = \gamma(j)$. Since Π is a projection, any of its eigenvalues can either be 0 or 1. For (26), it is hence sufficient to show that if $\langle b, a^{(k)} \rangle_{p^{-1}} = 0$ for all $k \in \operatorname{Ran} \gamma$, then necessarily $\Pi b = 0$. This follows by noting that

$$\langle b, a^{(k)} \rangle_{p^{-1}} = \sum_{i=1}^{n} b_i \delta_{\gamma(i)k}, \tag{38}$$

so that

$$\sum_{j=1}^{n} \Pi_{ij} b_j = \frac{p_i \sum_{j=1}^{n} \delta_{\gamma(i)\gamma(j)} b_j}{\sum_{l=1}^{n} p_l \delta_{\gamma(l)\gamma(i)}} = 0$$

if (38) is satisfied for all $k \in \operatorname{Ran} \gamma$.

The disjointness of the supports of the vectors $a^{(k)}$ follows from their definition (and Γ being a hard affiliation matrix). Item is now a direct consequence of (b). Item (c) follows directly from (b) and $\Lambda = P\Pi$, $\Lambda p = P\Pi p = Pp = q$. □

C. Pinsker's inequalities for the (weighted) ℓ^2 norm

The classical formulation of Pinsker's inequality [Tsy09, Lemma 2.5] bounds the squared ℓ^1 norm (or, equivalently, the squared total variation norm) of the difference of two probability vectors $u, v \in \mathbb{R}^m$ by the Kullback–Leibler divergence,

$$\|u - v\|_1^2 \leqslant 2 \, \mathrm{D}_{\mathrm{KL}}(u \parallel v). \tag{39}$$

In this section, we derive a similar result for the (possibly weighted) ℓ^2 norm in place of the ℓ^1 norm. While this can easily be achieved by applying the inequality $\|x\|_2 \leqslant \|x\|_1$, $x \in \mathbb{R}^m$, our aim is to obtain bounds that are as sharp as possible. For this purpose, we use the concepts of balancedness and q-weighted balancedness from Definition 16 and state four versions of Pinsker's inequality in Proposition 26 below which are particularly sharp in cases where either

(a) the difference $u - v$ has high balancedness, or

(b) we require a bound of the q-weighted ℓ^2 norm and $u - v$ has high q-balancedness, or

(c) the vector u has high balancedness and the difference $u - v$ is small, or

(d) we require a bound of the q-weighted ℓ^2 norm, u has high q-balancedness and the difference $u - v$ is small,

respectively.

Lemma 25: For any $x \in \mathbb{R}$ with $x > -1$,

$$\log(1 + x) \leqslant x - \frac{x^2}{2} + \frac{x^3}{3}.$$

Proof. The function $f(x) = x - x^2/2 + x^3/3 - \log(1 + x)$ satisfies $f(0) = 0$ and, since $(1 + x)f'(x) = x^3$,

$$f'(x) \begin{cases} \geqslant 0 & \text{if } x \geqslant 0, \\ \leqslant 0 & \text{if } -1 < x < 0, \end{cases}$$

proving the claim by the fundamental theorem of calculus. □

The following result is utilized in Theorem 18.

Proposition 26: Let $u, v, q \in \mathbb{R}^m$, $m \in \mathbb{N}$, be probability vectors such that $q > 0$ (componentwise) and $\alpha(u, v) := \frac{2}{3} \max_i \frac{|u_i - v_i|}{u_i} \in [0, \infty]$.[6] Then

(a) $\|u - v\|_2^2 \leqslant \dfrac{2\, \mathrm{D}_{\mathrm{KL}}(u \parallel v)}{m\, \mathfrak{B}(u - v)}$,

(b) $\displaystyle\sum_{i=1}^{m} \frac{|u_i - v_i|^2}{q_i} \leqslant \frac{2\, \mathrm{D}_{\mathrm{KL}}(u \parallel v)}{\mathfrak{B}_q(u - v)}$,

(c) $\|u - v\|_2^2 \leqslant \dfrac{2\, \mathrm{D}_{\mathrm{KL}}(u \parallel v)}{m\, \mathfrak{B}(u)(1 - \alpha(u, v))}$, if $\alpha(u, v) < 1$.

(d) $\displaystyle\sum_{i=1}^{m} \frac{|u_i - v_i|^2}{q_i} \leqslant \frac{2\, \mathrm{D}_{\mathrm{KL}}(u \parallel v)}{\mathfrak{B}_q(u)(1 - \alpha(u, v))}$, if $\alpha(u, v) < 1$.

Proof. Let $\eta := v - u$. If $\eta = 0$, there is nothing to show. Otherwise, Hölder's inequality and Pinsker's inequality (39) yield

$$\|\eta\|_2^2 \leqslant \|\eta\|_\infty \|\eta\|_1 = \frac{\|\eta\|_\infty}{\|\eta\|_1} \|\eta\|_1^2 \leqslant \frac{2\, \mathrm{D}_{\mathrm{KL}}(u \parallel v)}{m\, \mathfrak{B}(u - v)},$$

$$\sum_{i=1}^{m} \frac{|\eta_i|^2}{q_i} \leqslant \max_i \frac{|\eta_i|}{q_i} \|\eta\|_1 = \frac{\max_i \frac{|\eta_i|}{q_i}}{\|\eta\|_1} \|\eta\|_1^2 \leqslant \frac{2\, \mathrm{D}_{\mathrm{KL}}(u \parallel v)}{\mathfrak{B}_q(u - v)},$$

proving (a) and (b). In order to show (c) and (d), first note that, if $v_i = 0$ and $u_i \neq 0$ for some i, then $\mathrm{D}_{\mathrm{KL}}(u \parallel v) = \infty$ and there is also nothing to show. On the other hand,

[6] Recall the convention $\frac{0}{0} = 0$.

if $u_i = 0$ and $v_i \neq 0$ for some i, then the condition $\alpha(u,v) < 1$ is violated. Finally, if $u_i = 0$ and $v_i = 0$ for some i, then we can reduce the dimension m by one and work with $[m]\backslash\{i\}$ without changing any of the quantities involved. Hence, we can assume $u_i \neq 0 \neq v_i$ for each $i \in [m]$. Now, since $\sum_i \eta_i = 0$ and assuming that $\alpha(u,v) < 1$, Lemma 25 implies

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(u \parallel v) &= -\sum_{i=1}^{m} u_i \log\left(\frac{u_i + \eta_i}{u_i}\right) \\
&\geqslant \sum_{i=1}^{m} u_i \left(-\frac{\eta_i}{u_i} + \frac{\eta_i^2}{2u_i^2} - \frac{\eta_i^3}{3u_i^3}\right) \\
&= \sum_{i=1}^{m} \frac{\eta_i^2}{2u_i}\left(1 - \frac{2\eta_i}{3u_i}\right) \\
&\geqslant \frac{1-\alpha(u,v)}{2\|u\|_\infty} \|u - v\|_2^2,
\end{aligned}$$

proving (c). The proof of (d) goes similarly with a slight modification of the last step:

$$\mathrm{D}_{\mathrm{KL}}(u \parallel v) \geqslant \sum_{i=1}^{m} \frac{\eta_i^2}{q_i} \frac{q_i}{2u_i}\left(1 - \frac{2\eta_i}{3u_i}\right) \geqslant \frac{1-\alpha(u,v)}{2\max_i \frac{u_i}{q_i}} \sum_{i=1}^{m} \frac{|u_i - v_i|^2}{q_i}.$$

□

D. Courant–Fischer Theorem for Singular Values

The common formulation of the Courant–Fischer (or min-max) theorem is stated for the eigenvalues of quadratic matrices. However, in section 3 we require a version for singular values of an arbitrary matrix, which is a well-known consequence. Let us state and prove the precise version that we are going to use:

Theorem 27: Let $n_1, n_2 \in \mathbb{N}$ and $M \in \mathbb{R}^{n_1 \times n_2}$ be an arbitrary matrix with $\operatorname{rank}(M) := s \leqslant \min(n_1, n_2)$ and ordered positive singular values $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_s > 0$. Further, let $M = U\Sigma V^\top$ be a singular value decomposition of M with $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_s)$ and with $U \in \mathbb{R}^{n_1 \times s}$, $V \in \mathbb{R}^{n_2 \times s}$ having orthonormal columns. Then, denoting by $\mathfrak{W}_k$ the set of k-dimensional subspaces of $\mathbb{R}^{n_2}$,

$$\max_{W \in \mathfrak{W}_k} \min_{x \in W,\, \|x\|_2 = 1} \|Mx\|_2 = \sigma_k, \qquad k \in [s], \tag{40}$$

$$\max_{(e_1,\ldots,e_r)\ \text{orthonormal}} \sum_{k=1}^{r} \|Me_k\|_2 = \sum_{k=1}^{r} \sigma_k, \qquad r \in [s], \tag{41}$$

where (40) is maximized by $W_k^\star = \operatorname{span}(V_{\bullet 1}, \ldots, V_{\bullet k})$. (with the inner minimization problem solved by $x = V_{\bullet k}$), while (41) is maximized by the right singular vectors $e_1 = V_{\bullet 1}, \ldots, e_r = V_{\bullet r}$.

Proof. First note that, for each $k \in [s]$, $\sigma_k = \sqrt{\lambda_k}$, where $\lambda_1 \geqslant \cdots \geqslant \lambda_s > 0$ are the positive eigenvalues of $M^\top M$ and that

$$\|Mx\|_2^2 = \langle Mx, Mx\rangle_2 = \langle x, M^\top Mx\rangle_2 \leqslant \|x\|_2 \|M^\top Mx\|_2,$$

with equality if and only if x and $M^\top Mx$ are collinear, i.e. whenever x is an eigenvector of $M^\top M$. Hence, the inequality "$\leqslant$" in (40) follows directly from the classical Courant–Fischer theorem [HJ13, Theorem 4.2.6]. To see the equality for $W = W_k^\star$, consider any $x = \sum_{j=1}^k \alpha_j V_{\bullet j} \in W_k^\star$, $\alpha \in \mathbb{R}^k$, and observe that

$$\|Mx\|_2^2 = \|\sum_{j=1}^k \alpha_j \sigma_j U_{\bullet j}\|_2^2 = \sum_{j=1}^k (\alpha_j \sigma_j)^2 \geqslant \sigma_k^2 \sum_{j=1}^k \alpha_j^2 = \sigma_k^2 \|x\|_2^2,$$

with equality for $x = V_{\bullet k}$.

To see (41), note that a recursive application of (40) implies the inequality "$\leqslant$" in (41), while the choice $e_k = V_{\bullet k}$ satisfies equality:

$$\sum_{k=1}^r \|MV_{\bullet k}\|_2 = \sum_{k=1}^r \|\sigma_k U_{\bullet k}\|_2 = \sum_{k=1}^r \sigma_k.$$

□

E. Collective analysis of DBMR runs

To support the the investigations in section 7, here we provide a brief analysis involving multiple DBMR runs. Due to the non-concavity of the objective $\hat{\ell}$, DBMR can settle in different local maxima of $\hat{\ell}$, depending on the initialization $\Gamma^{(0)}$.

Results of 100 runs of DBMR are presented in Figures 3 (Example 21) and 4 (Example 22). The top row of images compare the first 5 singular values of $\widetilde{\Lambda}$ for every converged pair (λ, Γ) (blue crosses) with the singular values of the full model $\widetilde{P}$ (red dots). Since $\operatorname{rank} \widetilde{\Lambda} = 3$, we do not show its 4th and 5th singular values, which are always zero. The bottom row of images present, for every converged pair (λ, Γ), the DBMR objective $\hat{\ell}(\lambda, \Gamma)$ versus the degree of coherence $\mathcal{C}_3(\Lambda)$. In the bottom panels a colorbar indicates the number of solutions in the histogram bins. In each of the bottom rows, one panel is showing the results as a scatter plot instead of a histogram, for better visual experience.

We observe that in the unperturbed case DBMR recovers $\sigma_2 = 1$ perfectly for all runs, but $\sigma_3 = 0.6$ is recovered only in 60% of the runs, and in the remaining runs it converges to a (degenerate) model with effectively $r = 2$ latent states; i.e., $\sigma_3(\widetilde{\Lambda}) = 0$. The bottom left panel of Figure 3 shows that the degenerate results are suboptimal, in the sense that the corresponding iterations get stuck in a suboptimal local maximum. As the perturbation in the data is increased, σ_2 is close to optimal and we start to get results between the previous two extreme cases of $\sigma_3(\widetilde{\Lambda})$, and the associated degrees of coherence $\mathcal{C}_3(\widetilde{\Lambda})$ spread out from the previous two values somewhat. For large perturbation, this

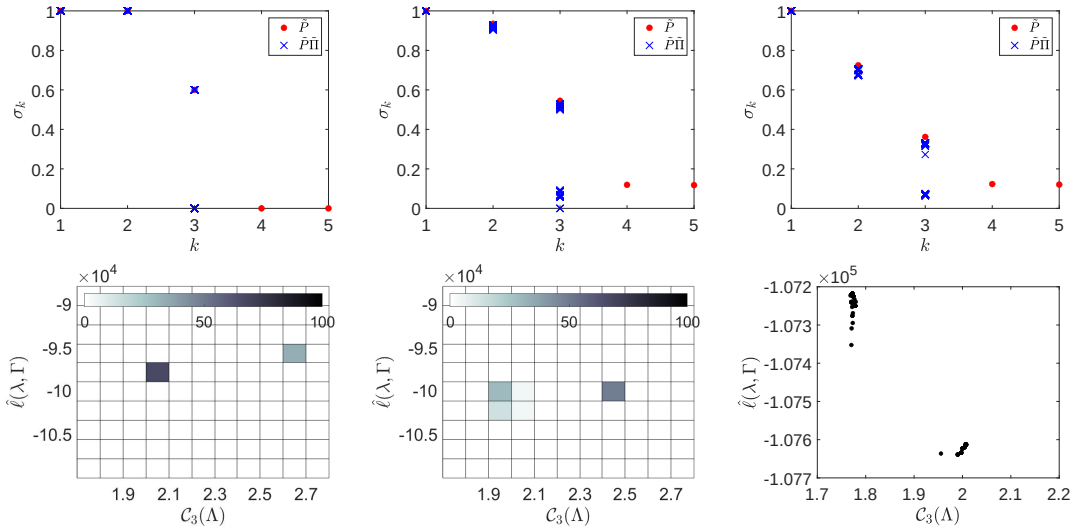

Figure 3: *Top:* First 5 singular values of $\widetilde{P}$ and $\widetilde{\Lambda} = \widetilde{P}\widetilde{\Pi}$ for the transition matrix of Example 21. *Bottom:* Likelihood bound $\hat{\ell}(\lambda, \Gamma)$ in dependence of the degree of coherence $\mathcal{C}_3(\Lambda)$. Results are shown for 100 runs of DBMR with $r = 3$ latent states. *Left:* unperturbed; *center:* slightly perturbed *right:* strongly perturbed.

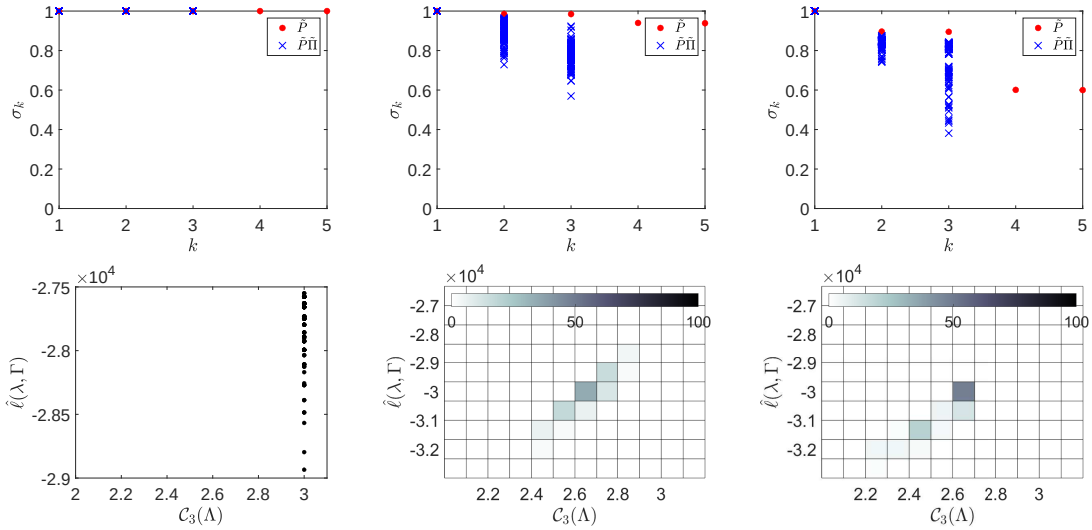

Figure 4: *Top:* First 5 singular values of $\widetilde{P}$ and $\widetilde{\Lambda} = \widetilde{P}\widetilde{\Pi}$ for the transition matrix of Example 22. *Bottom:* Likelihood bound $\hat{\ell}(\lambda, \Gamma)$ in dependence of the degree of coherence $\mathcal{C}_3(\Lambda)$. Results are shown for 100 runs of DBMR with $r = 3$ latent states. *Left:* unperturbed; *center:* slightly perturbed *right:* strongly perturbed.

process continues, and we see clustering of the objectives $(\mathcal{C}, \hat{\ell})$ around $(1.75, -1.072 \cdot 10^5)$ and $(2, -1.076 \cdot 10^5)$. That is, the objectives work against one another. We thus note that while increasing the perturbation improved the success rate of the DBMR runs finding a global maximum, it also turned the harmonious objectives into mildly conflicting ones.

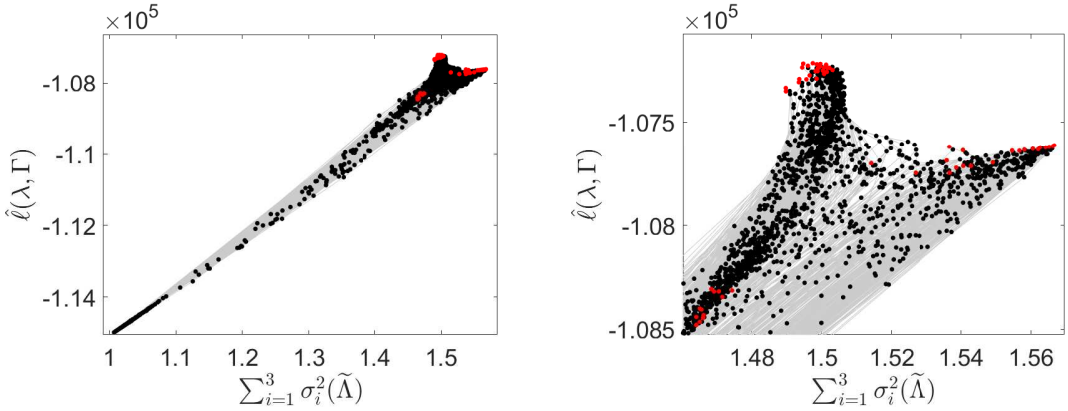

Figure 5: The two objectives $\hat{\ell}(\lambda, \Gamma)$ and $\|\widetilde{\Lambda}\|_F^2$ for all iterates of 1000 randomly initialized DBMR runs in the strongly perturbed case of Example 21. Grey lines connect data points for successive DBMR iterates, red dots indicate the endpoints of the 1000 runs (local optima of the DBMR objective (14)). The right panel is a close-up of the region with the highest values of both objectives, corresponding to the top right corner of the left panel.

Figure 4 (Example 22) shows that for the unperturbed case all DBMR runs converge to coherence-optimal partitions. However, for the perturbed cases, many local optima trap the runs. As discussed in Example 22, we attribute these local minima to the many coherent sets of different sizes present in this system. In this example we observe no conflict between the optimization criteria $\mathcal{C}$ and $\hat{\ell}$.

To get a more comprehensive picture for Example 21 with large perturbation, where the conflict between the optimization criteria arises, we consider 1000 new DBMR runs. Every run is initialized, as before, with an affiliation matrix $\Gamma^{(0)}$ of which every column is a uniform i.i.d. sample of one of the three standard unit vectors. This time, instead of considering only the converged DBMR solutions, we keep all iterates of every run, that is, whole "DBMR trajectories": 6674 pairs of (λ, Γ) in total. For all these, we depict $\hat{\ell}(\lambda, \Gamma)$ and $\sum_{i=1}^3 \sigma_i(\widetilde{\Lambda})^2 = \|\widetilde{\Lambda}\|_F^2$, where the latter equality is due to $\operatorname{rank}(\widetilde{\Lambda}) \leqslant r = 3$. By Corollary 20 this is an equivalent objective to $\|\widetilde{P} - \widetilde{\Lambda}\|_F^2$, since $\|\widetilde{P}\|_F^2$ depends only on the data and not on DBMR iterates. The results are shown in the left-hand panel of Figure 5, with a close-up on the region with the conflicting optima on the right. In the left-hand panel all initial points of DBMR runs satisfy $\|\widetilde{\Lambda}\|_F^2 \leqslant 1.1$. We observe that, although there is some "spread" during the DBMR iterations and in particular in the local DBMR optima, the correlation between the objectives is quite high for this set of matrices. We also observe a third cluster of local optima around $\hat{\ell}(\lambda, \Gamma) \approx -1.084 \cdot 10^5$ which was not found by the previous 100 runs, cf. Figure 3 (bottom right panel). It seems

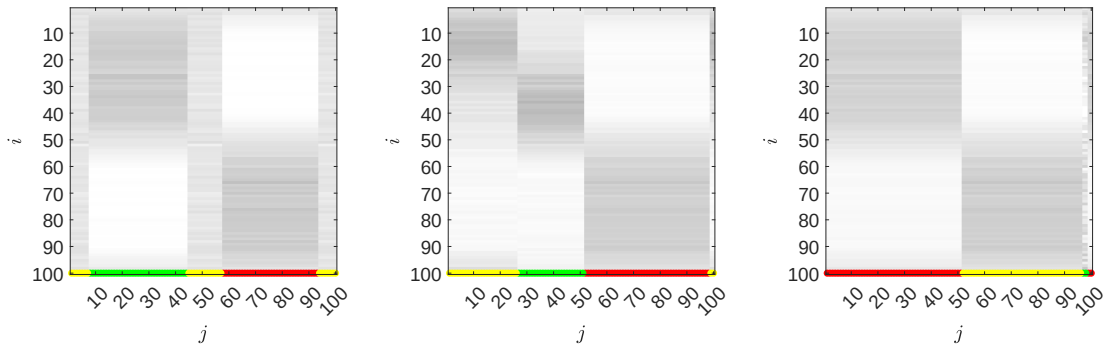

Figure 6: Three DBMR output transition matrices $\Lambda = \lambda\Gamma$ for optimal (left: $\hat{\ell}(\lambda, \Gamma) \approx -1.072 \cdot 10^5$) and suboptimal (center: $\hat{\ell}(\lambda, \Gamma) \approx -1.076 \cdot 10^5$, right: $\hat{\ell}(\lambda, \Gamma) \approx -1.084 \cdot 10^5$) local maxima in Example 21 with large perturbation, each corresponding to one of the three red clusters in Figure 5. The coloring of the bottom line indicates the partition given by the affiliation matrix Γ.

to correspond to degenerate local minima essentially belonging to a coherent 2-partition, as can be seen by the corresponding DBMR transition matrix Λ, depicted in Figure 6 (right). This figure illustrates three DBMR transition matrices Λ, each corresponding to one DBMR optimum from the three red clusters in Figure 5. Its left panel is identical to the bottom right panel Figure 1 with the highest DBMR objective value, while the center panel with objective value around $-1.076 \cdot 10^5$ corresponds to a clustering that is closer to the coherence-optimal partition (Figure 1 middle right).

References

[Aka74] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Automatic Control*, AC-19:716–723, 1974. System identification and time-series analysis.

[BK17] R. Banisch and P. Koltai. Understanding the geometry of transport: diffusion maps for Lagrangian trajectory data unravel coherent sets. *Chaos*, 27(3):035804, 16, 2017.

[Den17] A. Denner. *Coherent structures and transfer operators*. PhD thesis, Technische Universität München, 2017.

[DHS05] C. Ding, X. He, and H. D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM international conference on data mining*, pages 606–610. SIAM, 2005.

[dLGDLR08] D. M. de Lachapelle, D. Gfeller, and P. De Los Rios. Shrinking matrices while preserving their eigenpairs with application to the spectral coarse

graining of graphs. *Submitted to SIAM Journal on Matrix Analysis and Applications*, 2008.

[DLP06] C. Ding, T. Li, and W. Peng. Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence chi-square statistic, and a hybrid method. In *AAAI*, volume 42, pages 137–43, 2006.

[DLPP06] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 126–135, 2006.

[DW05] P. Deuflhard and M. Weber. Robust Perron cluster analysis in conformation dynamics. *Linear Algebra Appl.*, 398:161–184, 2005.

[FBD09] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural computation*, 21(3):793–830, 2009.

[FI11] C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, 23(9):2421–2456, 2011.

[FLS10] G. Froyland, S. Lloyd, and N. Santitissadeekorn. Coherent sets for nonautonomous dynamical systems. *Phys. D*, 239(16):1527–1541, 2010.

[Fro98] G. Froyland. Approximating physical invariant measures of mixing dynamical systems in higher dimensions. *Nonlinear Anal.*, 32(7):831–860, 1998.

[Fro13] G. Froyland. An analytic framework for identifying finite-time coherent sets in time-dependent dynamical systems. *Phys. D*, 250:1–19, 2013.

[Fro15] G. Froyland. Dynamic isoperimetry and the geometry of Lagrangian coherent structures. *Nonlinearity*, 28(10):3587–3622, 2015.

[FRS19] G. Froyland, C. P. Rock, and K. Sakellariou. Sparse eigenbasis approximation: Multiple feature extraction across spatiotemporal scales with application to coherent set identification. *Commun. Nonlinear Sci. Numer. Simul.*, 77:81–107, 2019.

[FSM10] G. Froyland, N. Santitissadeekorn, and A. Monahan. Transport in time-dependent dynamical systems: finite-time coherent sets. *Chaos*, 20(4):043116, 10, 2010.

[GH17] S. Gerber and I. Horenko. Toward a direct and scalable identification of reduced models for categorical processes. *Proceedings of the National Academy of Sciences*, 114(19):4863–4868, 2017.

[Gil20] N. Gillis. *Nonnegative Matrix Factorization*. Society for Industrial and Applied Mathematics, 2020.

[GK78] I. Gohberg and M. G. Kreĭn. *Introduction to the theory of linear nonselfadjoint operators*, volume 18. American Mathematical Soc., 1978.

[GONH18] S. Gerber, S. Olsson, F. Noé, and I. Horenko. A scalable approach to the computation of invariant measures for high-dimensional Markovian systems. *Scientific reports*, 8(1):1796, 2018.

[HE15] T. Hsing and R. Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons, 2015.

[Hec98] D. Heckerman. *A tutorial on learning with Bayesian networks*. Springer, 1998.

[HJ13] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, 2 edition, 2013.

[Hof99] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

[Hof01] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, 42(1-2):177–196, 2001.

[Hot36] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3-4):321–377, 12 1936.

[HS05] W. Huisinga and B. Schmidt. Metastability and dominant eigenvalues of transfer operators. *Lecture Notes in Computational Science and Engineering*, 49:167, 2005.

[KCS16] P. Koltai, G. Ciccotti, and C. Schütte. On metastability and Markov state models for non-stationary molecular dynamics. *The Journal of Chemical Physics*, 145(17):174103, 2016.

[KHMN19] S. Klus, B. E. Husic, M. Mollenhauer, and F. Noé. Kernel methods for detecting coherent structures in dynamical data. *Chaos*, 29(12):123112, 15, 2019.

[Kif86] Y. Kifer. General random perturbations of hyperbolic and expanding transformations. *Journal D'Analyse Mathématique*, 47:111–150, 1986.

[KWNS18] P. Koltai, H. Wu, F. Noé, and C. Schütte. Optimal data-driven estimation of generalized Markov state models for non-equilibrium dynamics. *Computation*, 6(1), 2018.

[LD14] T. Li and C. Ding. Nonnegative matrix factorizations for clustering: A survey. In C. C. Aggarwal and C. K. Reddy, editors, *Data Clustering: Algorithms and Applications*, Data Mining and Knowledge Discovery Series, chapter 7, pages 149–176. Chapman and Hall/CRC, 1. edition, 2014.

[LS99] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[LS00] D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2000.

[LSZL20] H. Lu, X. Sang, Q. Zhao, and J. Lu. Community detection algorithm based on nonnegative matrix factorization and pairwise constraints. *Physica A: Statistical Mechanics and its Applications*, 545:123491, 2020.

[OBA22] M. Ortiz-Bouza and S. Aviyente. Community detection in multiplex networks based on orthogonal nonnegative matrix tri-factorization. *arXiv preprint arXiv:2205.00626*, 2022.

[OW05] E. Ordentlich and M. J. Weinberger. A distribution dependent refinement of Pinsker's inequality. *IEEE Transactions on Information Theory*, 51(5):1836–1840, 2005.

[RW13] S. Röblitz and M. Weber. Fuzzy spectral clustering by PCCA+: Application to Markov state models and data classification. *Advances in Data Analysis and Classification*, 7(2):147–179, 2013.

[SG08] A. P. Singh and G. J. Gordon. A unified view of matrix factorization models. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part II 19*, pages 358–373. Springer Berlin Heidelberg, 2008.

[SRS+08] M. Shashanka, B. Raj, P. Smaragdis, et al. Probabilistic latent variable models as nonnegative factorizations. *Computational intelligence and neuroscience*, 2008, 2008.

[SS13] C. Schütte and M. Sarich. *Metastability and Markov State Models in Molecular Dynamics*, volume 24. American Mathematical Soc., 2013.

[Tsy09] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[UHZ+16] M. Udell, C. Horn, R. Zadeh, S. Boyd, et al. Generalized low rank models. *Foundations and Trends® in Machine Learning*, 9(1):1–118, 2016.

[Ula60] S. M. Ulam. *A collection of mathematical problems*, volume 8. Interscience Publishers, 1960.

[Was04] L. Wasserman. *All of statistics: a concise course in statistical inference*, volume 26. Springer, 2004.

[WKZ+18] W. Wu, S. Kwong, Y. Zhou, Y. Jia, and W. Gao. Nonnegative matrix factorization with mixed hypergraph regularization for community detection. *Information Sciences*, 435:263–281, 2018.

[WN20] H. Wu and F. Noé. Variational approach for learning Markov processes from time series data. *Journal of Nonlinear Science*, 30(1):23–66, 2020.

[WZ12] Y.-X. Wang and Y.-J. Zhang. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on knowledge and data engineering*, 25(6):1336–1353, 2012.

[YL13] J. Yang and J. Leskovec. Overlapping community detection at scale: A nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 587–596, 2013.