Beyond Predictive Accuracy: Methods for Interpretable and Robust Discovery of Latent Processes in Complex Systems

Nguyen Nguyen[1], Jiawei Li[3], Meng Lai[2], Ioannis Ch. Paschalidis[1,2] and Jonathan H. Huggins[3,2]

[1] *Division of Systems Engineering, Boston University, USA e-mail:* nguyenpn@bu.edu; yannisp@bu.edu

[2] *Faculty of Computing & Data Sciences, Boston University, USA*

[3] *Department of Mathematics & Statistics, Boston University, USA e-mail:* huggins@bu.edu

Abstract: The adoption of machine learning in critical domains such as healthcare and scientific discovery is hindered by a significant gap: while models excel at prediction, they often lack the interpretability and robustness required for high-stakes decision-making. This limitation stems from the difficulty of reliably identifying the unobserved, or latent, processes that govern complex systems, particularly when faced with imperfect data and unavoidable model misspecification. This work addresses this challenge by developing methodologies that shift the focus from predictive accuracy toward the robust and interpretable discovery of these latent structures. This work presents two primary contributions: (1) an asymptotically consistent spectral method of moments for Hierarchical Imitation Learning that provides a direct, reliable estimation of hidden decision-making policies, serving as both an asymptotically consistent standalone solution and a high-quality initialization that synergizes with Expectation-Maximization (EM) algorithms to prevent convergence to poor local optima, and (2) the Accumulated Cutoff Discrepancy Criterion (ACDC), a novel model selection framework that robustly identifies the true number of underlying processes by preventing overfitting to statistical noise and model artifacts. Collectively, these contributions advance a more robust and interpretable approach to machine learning, providing valuable tools for meaningful scientific discovery.

Keywords and phrases: Learning for control, Machine learning, Markov Decision Processes, Options, Imitation Learning, Method of Moments.

Chapter 1

Introduction

Many critical real-world applications in areas such as healthcare, autonomous systems, and scientific discovery are fundamentally driven by unobserved, or latent, processes. For example, processes might correspond to subpopulations that cannot be directly observed such as types of cells (Gorsky, Chan and Ma, 2020; Prabhakaran et al., 2016), behavioral genotypes (Stevens et al., 2019), or groups with canonical patterns of IQ development (Bauer, 2007). They could correspond to a variety of scientifically important objects such cell programs (Buettner et al., 2017; Kotliar et al., 2019; Risso et al., 2018), mutational processes in tumors (Kinker et al., 2020; Levitin et al., 2019; Seplyarskiy et al., 2021), or material types (Févotte and Dobigeon, 2015; Rajabi and Ghassemian, 2015). Processes can also be decision policies, such as in the case of *Hierarchical Reinforcement Learning* (HRL) or, in the presence of expert demonstration, *Hierarchical Imitation Learning* (HIL), where general sweeping decisions over large epochs consist of smaller specific decisions on finer (more granular) epochs (Barto and Mahadevan, 2003; Sutton, Precup and Singh, 1999).

While current machine learning techniques excels at prediction, a challenge remains in the consistent and interpretable identification of the underlying latent structures that govern observable phenomena. A lot of approaches, like HRL/HIL frameworks, often rely on local search like the *Expectation-Maximization* (EM) algorithms. These are prone to problems such as slow convergence, suboptimal local optima, and sensitivity to initialization. In tasks like learning from expert demonstrations, this can lead to policies that are inefficient or fail to capture the expert's true strategy.

Moreover, in practice, it is necessary to not only characterize the latent processes, but also determine how many such processes there are. *Akaike, Bayesian, and deviance information criteria* (AIC, BIC, DIC) (Akaike, 1974; Schwarz, 1978; Spiegelhalter et al., 2002) remains commonly used for such purpose due to their simplicity and ease of use. However, these criteria primarily optimize for predictive performance. Real-world difficulties such as sparse, incomplete, or noisy data, along with imperfect model knowledge, frequently lead to issues using these standard methods. Model misspecification is ubiquitous in complex systems, as simplifying assumptions are often necessary and the true generative process is typically unknown or too intricate to model perfectly. Consequently, as sample sizes increase, BIC and similar criteria tend to select increasingly complex models, adding spurious latent structures merely to better approximate the observed data, rather than identifying the genuinely distinct underlying processes (Cai, Campbell and Broderick, 2021). This overfitting hinders

interpretability, as the inferred structures may lack meaningful real-world correspondence, leading potentially to flawed scientific conclusions or ineffective interventions.

For machine learning to be effectively adopted in high-stakes fields like healthcare, models must, in addition to being accurate, be transparent, explainable, and robust against real-world data complications and model misspecification.

This work addresses the problem by presenting two primary contributions:

1. An asymptotically consistent spectral method of moments for HIL provides a direct, non-iterative estimation of hidden hierarchical decision-making policies within the Options Framework (Sutton, Precup and Singh, 1999). By leveraging low-order moments of observed state-action data and circumventing local search, it offers global convergence guarantees under mild conditions and requires only a single pass through the expert demonstration data. This approach can serve as a robust standalone method for policy recovery, or it can synergize effectively with existing EM algorithms by providing a high-quality initialization, thereby mitigating the risk of convergence to poor local optima.
2. The *Accumulated Cutoff Discrepancy Criterion* (ACDC), a novel, general-purpose model selection framework designed explicitly to identify the true number of latent processes, even when the assumed model family is misspecified. ACDC shifts the focus from predictive accuracy to explanatory power by incorporating a cutoff discrepancy threshold at the component level. This mechanism prevents the criterion from rewarding the addition of spurious components that only capture minor data deviations or noise, thereby directly combating the overfitting problem common to BIC/AIC in misspecified settings. It provides robust model selection consistency under intuitive assumptions for broad model classes.

Collectively, these contributions aim to provide valuable alternative and complementary methods for applying machine learning in high-stakes situations. Continue on, Chapter 2 will provide details to the spectral HIL method, including relevant background and related works. Chapter 3 will similarly provide the background and details to the ACDC, with particular focus on its application to the class of probabilistic matrix factorization (PMF) models. Finally, Chapter 4 will discuss the limitations of the proposed methods and the way forward.

Chapter 2

Spectral method of moments for Hierarchical Imitation Learning

2.1. Introduction

Hierarchical Reinforcement Learning (HRL) seeks to address the scalability and interpretability problem of Reinforcement Learning (RL) by introducing layers of abstraction over the decision process, enabling high-level decisions over larger time scales and low-level decisions over smaller time scales (Barto and Mahadevan, 2003; Sutton, Precup and Singh, 1999). The success of HRL relies on discovering suitable abstractions. In the literature, the problem of discovering suitable abstractions has been tackled both separately and in conjunction (in a single end-to-end process) with learning the optimal policy (Barto and Mahadevan, 2003). In specific instances where expert demonstrations are available, the process of discovering abstractions and learning optimal policies can be accelerated via *Hierarchical Imitation Learning* (HIL). Specifically, HIL involves computing a hierarchy of policies from expert demonstrations and is the extension of *Imitation Learning* (IL) to HRL. In this chapter, we develop a novel HIL approach for the HRL with options framework of Sutton, Precup and Singh (1999).

The HRL with options framework proposed by Sutton, Precup and Singh (1999) involves a two-tiered hierarchy of policies, with a high-level policy governing "options" or decision as to which of a finite set of low-level policies are used to select actions. A key challenge of HIL in this options framework is that in practice, only (low-level) states and actions are directly observed through expert demonstrations, not the (high-level) options. The options thus constitute hidden (or latent) variables, and so recent HIL works have drawn inspiration from Expectation-Maximization (EM) techniques for learning Hidden Markov models (HMMs) and other latent variable models (Daniel et al., 2016; Giammarino and Paschalidis, 2021; Zhang and Paschalidis, 2021). These EM techniques process state-action pairs from expert demonstrations with a Bayesian smoother to compute a surrogate function for the (log)likelihood, and subsequent maximization of this surrogate function over the policy space. Whilst local-convergence theoretical guarantees have recently been shown for such an EM approach in the context of HIL (Zhang and Paschalidis, 2021), the nature of EM techniques as local-search procedures means that they are prone to convergence to local (non-global) maxima, and slow convergence with associated high computational expense.

In HMMs and other specific classes of latent variable models, *methods of moments* have been developed to overcome convergence issues inherent with EM techniques (Anandkumar et al., 2014a; Hsu, Kakade and Zhang, 2012; Hsu and Kakade, 2013; Mattila, Rojas and Wahlberg, 2015; Mattila et al., 2017, 2020; Parikh et al., 2012). These moment methods are free of local convergence problems (Anandkumar et al., 2014a; Mattila et al., 2020), and often offer much faster practical convergence with less computational expense (Mattila, Rojas and Wahlberg, 2015; Mattila et al., 2017). Moment methods have therefore been used both by themselves and as initialization algorithms for EM techniques (cf. (Zhang et al., 2016)). Nevertheless, moment methods have not previously been investigated for HIL in the options framework.

This chapter describes a new method of moments for HIL in the HRL options framework of (Sutton, Precup and Singh, 1999). Inspired by the method of moments for HMMs developed in (Hsu, Kakade and Zhang, 2012), the proposed method of moments for HIL offers global convergence under mild regularity and non-degeneracy conditions, and has the practical advantage of only requiring a single pass through the expert demonstrations. It therefore serves as both a useful alternative and complementary technique to the previously developed but locally-convergent EM algorithms of (Daniel et al., 2016; Giammarino and Paschalidis, 2021; Zhang and Paschalidis, 2021).

Notation used in this chapter: Uppercase letters denote random variables, lowercase letters denote realizations. Uppercase bold letters denote matrices, lowercase bold letters denote vectors. Superscript on a quantity acts like a label in case there are many quantities with the same symbol. Subscript on a quantity denotes it being a subclass of the original quantity. The Kronecker product $\otimes$ is defined as

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & \dots & a_{1n}\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\boldsymbol{B} & \dots & a_{mn}\boldsymbol{B} \end{bmatrix},$$

where $\boldsymbol{A}$ is a $m \times n$ matrix, $\boldsymbol{B}$ is a $p \times q$ matrix, and $\boldsymbol{A} \otimes \boldsymbol{B}$ is a $mp \times nq$ matrix. The Hadamard (element-wise) product $\circ$ is defined as

$$\boldsymbol{A} \circ \boldsymbol{B} = \begin{bmatrix} a_{11}b_{11} & \dots & a_{1n}b_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1}b_{m1} & \dots & a_{mn}b_{mn} \end{bmatrix},$$

where $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{A} \circ \boldsymbol{B}$ are $m \times n$ matrices. Furthermore, $\mathbf{I}_m$ denotes an $m \times m$ identity matrix, $\mathbf{1}_{m\times n}$ denotes an $m \times n$ matrix with all of its entries equal to one, $\mathbf{0}_{m\times n}$ denotes an $m \times n$ matrix with all of its entries equal to zero, and $\mathbf{e}_j$ denotes the j^{th} unit vector. The Moore–Penrose inverse of a matrix $\boldsymbol{A}$ will be denoted $\boldsymbol{A}^{+}$ and its transpose by $\boldsymbol{A}^{\mathrm{T}}$.

2.2. Problem Formulation

In this section, we introduce the HRL with (Bauer, 2007)options framework (Barto and Mahadevan, 2003; Sutton, Precup and Singh, 1999) and formulate the associated HIL problem.

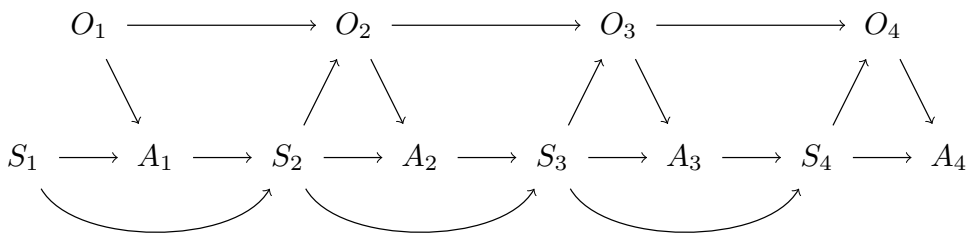


Fig 2.1: Bayesian network of the HRL option framework.

2.2.1. HRL with Options Framework

The HRL with options framework corresponds to the Bayesian network shown in Fig. 2.1 where O_t, S_t, and A_t denote the option, the state, and the action at time $t \geq 1$, respectively. The triple (O_t, S_t, A_t) forms a discrete-time Markov chain with O_t, S_t, and A_t defined over the finite spaces $\mathcal{O}$, $\mathcal{S}$, and $\mathcal{A}$, respectively. We denote the cardinality of these spaces as $|\mathcal{O}| = \omega$, $|\mathcal{S}| = \zeta$, and $|\mathcal{A}| = \alpha$.

The initial option and state pair (o_1, s_1) is sampled from an initial distribution $\pi_1(\cdot, \cdot)$. For $t \geq 1$, to advance one time step starting from the current pair (o_t, s_t), the action a_t is sampled from a low-level policy $\pi_{lo}(\cdot|s_t, o_t)$. Then, the resulting state s_{t+1} is sampled from an environment transition probability distribution $\Phi(\cdot|s_t, a_t)$. Finally, the next option o_{t+1} is sampled based on the new state and the previous option from the high-level policy $\pi_{hi}(\cdot|o_t, s_{t+1})$. The HRL with options framework is thus characterized by the policies π_{hi} and π_{lo}, and the transition distribution Φ.

Remark 1. *The framework we consider differs slightly from that in Sutton, Precup and Singh (1999) in that:*

1. *The termination random variable, along with its decision policy is omitted, with the option transition based solely on the high-level policy π_{hi}. This is due to the fact that the termination factor is only involved in the transition between options, without directly affecting any observables in any way. For the sake of simplicity, the termination policy is folded into the high-level policy as one single object.*
2. *The process starts with the pair (o_1, s_1) instead of (o_0, s_1). This difference is inconsequential as the resulting extra transition would be canceled out during the operations below.*

2.2.2. The HIL Problem

Suppose that an expert uses the HRL with options framework to generate a sequence of states and actions $\{(s_t, a_t)\}_{t=1}^T$. In the HIL problem, we seek to use this sequence to learn the expert's underlying low and high-level policies π_{lo} and π_{hi}. The associated options $\{o_t\}_{t=1}^T$ are not observed and constitute hidden (or latent) variables. The HIL problem is thus an instance of learning in the presence of latent variables which has motivated its solution via EM approaches in (Daniel et al., 2016; Giammarino and Paschalidis, 2021; Zhang and Paschalidis, 2021). Due to local convergence issues inherent in EM approaches, we shall take

a different approach and develop a method of moments for HIL inspired by the method of moments for HMMs developed in (Hsu, Kakade and Zhang, 2012).

To develop our method, we define the following matrices.

Definition 1. *For $s \in \mathcal{S}$, define $\mathbf{\Pi}_s^{lo} \in \mathbb{R}^{\omega \times \alpha}$ with*

$$\mathbf{\Pi}_s^{lo}[o, a] = \pi_{lo}(A_t = a | O_t = o, S_t = s)$$

as the matrix representation of π_{lo} under the state s.

Definition 2. *For $s \in \mathcal{S}$, define $\mathbf{\Pi}_s^{hi} \in \mathbb{R}^{\omega \times \omega}$ with*

$$\mathbf{\Pi}_s^{hi}[o, o'] = \pi_{hi}(O_{t+1} = o' | O_t = o, S_{t+1} = s)$$

as the matrix representation of π_{hi} under the state s.

Definition 3. *For $a \in \mathcal{A}$, define $\mathbf{\Phi}_a^A \in \mathbb{R}^{\zeta \times \zeta}$ with*

$$\mathbf{\Phi}_a^A[s, s'] = \Phi(S_{t+1} = s' | S_t = s, A_t = a)$$

as matrix representations of the transition dynamics.

Definition 4. *For $s' \in \mathcal{S}$, define $\mathbf{\Xi}_{s'} \in \mathbb{R}^{\zeta \times \omega}$ with*

$$\mathbf{\Xi}_{s'}[s, o] = P(S_t = s, O_t = o, S_{t+1} = s').$$

We also require the following mild regulatory assumptions.

Assumption 1 (Option-Action Identifiability)**.** *Under the same state, no two options contain the same policy for choosing an action, i.e., $\mathbf{\Pi}_s^{lo}$ has full row rank $\forall s \in \mathcal{S}$.*

Assumption 2 (Option-Option Identifiability)**.** *Under the same state, no two options give the same policy for choosing the next option, i.e., $\mathbf{\Pi}_s^{hi}$ has full rank $\forall s \in \mathcal{S}$.*

Assumption 3. *$\mathbf{\Xi}_s$ has full column rank $\forall s \in \mathcal{S}$.*

Assumption 4. *All actions have a non-zero chance of transitioning a state to all of its neighboring states and one state is another state's neighbor if there exists an action under which the probability of transitioning from the latter to the former is non-zero, i.e., for any $s, s' \in \mathcal{S}$, if there exists $a \in \mathcal{A}$ such that $\mathbf{\Phi}_a^A[s, s'] > 0$, then $\mathbf{\Phi}_{a'}^A[s, s'] > 0, \forall a' \in \mathcal{A}$.*

Assumption 5 (Stationary)**.** *The process (O_t, S_t) starts with the stationary distribution, that is $\boldsymbol{\pi}_s^1[o] = \boldsymbol{\pi}_s^\infty[o]$ where $\boldsymbol{\pi}_s^t \in \mathbb{R}^\omega$ with $\boldsymbol{\pi}_s^t[o] = P(O_t = o, s_t = s)$ for $s \in \mathcal{S}$.*

Remark 2. *Assumptions 1, 2, and 3 follow the same line of reasoning as Condition 1 of Hsu, Kakade and Zhang (2012); they remove malicious instances that can cause learning to confuse options that have the same transition/action probability. Assumption 4 is needed as the method of moments relies on the cancellation of certain terms across all actions, it can be interpreted as an action emission noise in the expert or a transition noise in the environment.*

2.3. Spectral method of moments

In this section, we develop our method of moments for HIL. We specifically identify observable moments of the states and actions, and show that they enable recovery of the low-policy π_{lo} via matrix diagonalization and the high-level policy π_{hi} via simple matrix algebra.

2.3.1. Moments in HIL

We note that under Assumption 5, the moments of the states, options, and actions are time-invariant. Thus, without loss of generality, we consider the moments $\boldsymbol{M}_a \in \mathbb{R}^{\zeta\zeta\times\zeta\alpha}$ for $a \in \mathcal{A}$ with

$$\begin{aligned}&\boldsymbol{M}_a\left[s_2\zeta + s_1, s_3\alpha + a_3\right]\\ &\quad= \; P(S_2 = s_2, S_1 = s_1, A_2 = a, S_3 = s_3, A_3 = a_3),\end{aligned}$$

and $\boldsymbol{K}_s \in \mathbb{R}^{\zeta\times\alpha}$ for $s \in \mathcal{S}$ with

$$\boldsymbol{K}_s[s_1, a_2] = P(S_1 = s_1, S_2 = s, A_2 = a_2).$$

Our goal now is to construct an expression of these observable moments that allows the recovery of the low-level policy via matrix diagonalization.

2.3.2. Diagonalizable Forms

We first examine the properties of $\boldsymbol{K}_s$ for $s \in \mathcal{S}$. Specifically, let $\boldsymbol{V}_s \in \mathbb{R}^{\alpha\times\omega}$ for $s \in \mathcal{S}$ be a matrix of right singular vectors corresponding to the ω largest singular values of $\boldsymbol{K}_s$. We then have the following lemma.

Lemma 1. *Define the block-diagonal matrices*

$$\boldsymbol{V} = \begin{bmatrix}\boldsymbol{V}_1 & & \\ & \ddots & \\ & & \boldsymbol{V}_\zeta\end{bmatrix} \text{ and } \boldsymbol{\Pi}^{lo} = \begin{bmatrix}\boldsymbol{\Pi}^{lo}_1 & & \\ & \ddots & \\ & & \boldsymbol{\Pi}^{lo}_\zeta\end{bmatrix}. \tag{2.1}$$

Then the product matrix

$$\boldsymbol{\Pi}^{lo}\boldsymbol{V} = \begin{bmatrix}\boldsymbol{\Pi}^{lo}_1\boldsymbol{V}_1 & & \\ & \ddots & \\ & & \boldsymbol{\Pi}^{lo}_\zeta\boldsymbol{V}_\zeta\end{bmatrix}$$

is invertible.

We next examine the properties of the moments $\boldsymbol{M}_a$. Before doing so, note that the moments $\boldsymbol{M}_a$ involve the transition dynamics $\boldsymbol{\Phi}^A_a$ as well as the underlying low- and high-level policies we are interested in. To remove the influence of the transition dynamics on $\boldsymbol{M}_a$, let us define the kernel matrix $\boldsymbol{\Psi} \in \mathbb{R}^{\zeta\times\zeta}$ with

$$\boldsymbol{\Psi}[s_2, s_3] = \begin{cases}\psi_{s_2 s_3}, & \text{if } \boldsymbol{\Phi}^A_a[s_2, s_3] > 0 \;\forall a \in \mathcal{A},\\ 0, & \text{otherwise,}\end{cases}$$

and normalizer matrices $\boldsymbol{N}_a \in \mathbb{R}^{\zeta\times\zeta}$ for $a \in \mathcal{A}$ with

$$\boldsymbol{N}_a[s_2, s_3] = \begin{cases}\dfrac{1}{\boldsymbol{\Phi}^A_a[s_2, s_3]}, & \text{if } \boldsymbol{\Phi}^A_a[s_2, s_3] > 0,\\ 0, & \text{otherwise,}\end{cases}$$

where $\psi_{s_2 s_3}$ are constants of choice such that $\boldsymbol{\Psi}$ is full rank (such constants will always exist under Assumption 4). We then may define the surrogate moments

$$\hat{\boldsymbol{M}}_a = (\boldsymbol{\Psi} \otimes \mathbf{1}_{\zeta\times\alpha}) \circ (\boldsymbol{N}_a \otimes \mathbf{1}_{\zeta\times\alpha}) \circ \boldsymbol{M}_a, \tag{2.2}$$

for $a \in \mathcal{A}$ and

$$\hat{\boldsymbol{M}} = \sum_{a\in\mathcal{A}} \hat{\boldsymbol{M}}_a \tag{2.3}$$

that do not depend on the transition dynamics where $\hat{\boldsymbol{M}}_a$ and $\hat{\boldsymbol{M}}$ have the same dimensions as $\boldsymbol{M}_a$.

These surrogate moments combined with Lemma 1 lead to the following theorem that establishes that the observable moments allow the recovery of the low-level policy via matrix diagonalization.

Theorem 1. *The product $\boldsymbol{V}^{\mathrm{T}}\hat{\boldsymbol{M}}^{+}\hat{\boldsymbol{M}}_a\boldsymbol{V}$ admits the factorization:*

$$\boldsymbol{V}^{\mathrm{T}}\hat{\boldsymbol{M}}^{+}\hat{\boldsymbol{M}}_a\boldsymbol{V} = \boldsymbol{B}^{-1}\boldsymbol{\Lambda}_a\boldsymbol{B}, \tag{2.4}$$

where

$$\boldsymbol{\Lambda}_a = \begin{bmatrix} \operatorname{diag}(\boldsymbol{\Pi}_1^{lo}\mathbf{e}_a) & & \\ & \ddots & \\ & & \operatorname{diag}(\boldsymbol{\Pi}_\zeta^{lo}\mathbf{e}_a) \end{bmatrix},$$

and

$$\boldsymbol{B} = (\boldsymbol{\Psi} \otimes \mathbf{I}_\omega)\boldsymbol{\Pi}^{hi}\boldsymbol{\Pi}^{lo}\boldsymbol{V}.$$

In order to compute the eigenbasis that jointly diagonalize (2.4) for all $a \in \mathcal{A}$, we find a vector $\boldsymbol{\eta} \in \mathbb{R}^\alpha$ such that the eigenvalues of

$$\sum_{a\in\mathcal{A}} \boldsymbol{\eta}_a \boldsymbol{V}^{\mathrm{T}}\hat{\boldsymbol{M}}^{+}\hat{\boldsymbol{M}}_a\boldsymbol{V} = \boldsymbol{B}\left(\sum_{a\in\mathcal{A}} \boldsymbol{\eta}_a\boldsymbol{\Lambda}_a\right)\boldsymbol{B}^{-1} \tag{2.5}$$

are well spread. In other words, we find $\boldsymbol{\eta}$ such that the values $\mathbf{e}_o^{\mathrm{T}}\boldsymbol{\Pi}_s^{lo}\boldsymbol{\eta}$ are distinct and non-zero for all $(o, s) \in \mathcal{O} \times \mathcal{S}$. As suggested in Hsu and Kakade (2013), this can be satisfied in most cases if $\boldsymbol{\eta}$ is sampled uniformly from the surface of a unit sphere in $\mathbb{R}^\alpha$.

The eigen-decomposition will yield an eigenbasis up to a permutation $\boldsymbol{\mathcal{P}} \in \mathbb{R}^{\omega\times\zeta}$ of the pair $(o, s) \in \mathcal{O} \times \mathcal{S}$. To put it differently, the diagonal matrix obtained from diagonalizing $\boldsymbol{V}^{\mathrm{T}}\hat{\boldsymbol{M}}^{+}\hat{\boldsymbol{M}}_a\boldsymbol{V}$ using this basis will be of the form $\boldsymbol{\mathcal{P}}\boldsymbol{\Lambda}_a\boldsymbol{\mathcal{P}}^{\mathrm{T}}$. With some further processing (see Appendix B), an order up to a permutation $\hat{\boldsymbol{\mathcal{P}}} \in \mathbb{R}^\omega$ of $o \in \mathcal{O}$ can be recovered, meaning the diagonal matrix obtained will be of the form $(\mathbf{I}_\zeta \otimes \hat{\boldsymbol{\mathcal{P}}})\boldsymbol{\Lambda}_a(\mathbf{I}_\zeta \otimes \hat{\boldsymbol{\mathcal{P}}}^{\mathrm{T}})$. This ordering corresponds to the relabeling of the options. Because this recovery process, while necessary, does not represent the main contribution of this work, it will be elaborated in the Appendix.

After obtaining the low-level policy matrices $\hat{\boldsymbol{\mathcal{P}}}\boldsymbol{\Pi}_s^{lo}$, the high-level policy matrices can be computed by the following theorem, up to the permutation $\hat{\boldsymbol{\mathcal{P}}}$ of the options.

Theorem 2.

$$\hat{\boldsymbol{\mathcal{P}}}\boldsymbol{\Pi}_{s'}^{hi}\hat{\boldsymbol{\mathcal{P}}}^{\mathrm{T}} = \sum_{s} \boldsymbol{w}_{s'}[s] \left(\hat{\boldsymbol{\mathcal{P}}}\boldsymbol{\Pi}_{s}^{lo}\boldsymbol{K}_{s}^{+}\hat{\boldsymbol{K}}_{ss'}\boldsymbol{\Pi}_{s'}^{lo^{+}}\hat{\boldsymbol{\mathcal{P}}}^{\mathrm{T}}\right), \tag{2.6}$$

where:

- $\hat{\boldsymbol{K}}_{ss'}$ *is a* $\zeta \times \alpha$ *submatrix of* $\hat{\boldsymbol{M}}$ *defined by*

$$\hat{\boldsymbol{K}}_{ss'}[s'', a] = \hat{\boldsymbol{M}}[s\zeta + s'', s'\alpha + a].$$

- $\boldsymbol{w}_{s'}$ *are length* ζ *weight vectors of choice subject to*

$$\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\Psi}\mathbf{e}_i^{\mathrm{T}} = 1, \ \forall i \in \mathcal{S}.$$

2.3.3. Proposed Method of Moments for HIL

Given the observed sequence $\{(s_t, a_t)\}_{t=1}^{T}$, our method of moments to learn the policies π_{lo} and π_{hi} is:

Step 1: Estimate $\boldsymbol{M}_a$, $\boldsymbol{K}_s$, and $\boldsymbol{\Phi}_a^A$ from data via:

$$\boldsymbol{M}_a\left[s_2\zeta + s_1, s_3\alpha + a_3\right] = \frac{\sum_{t=1}^{T-2} \mathbb{I}_{\{s_t=s_1, s_{t+1}=s_2, a_{t+1}=a, s_{t+2}=s_3, a_{t+2}=a_3\}}}{T-2},$$

$$\boldsymbol{K}_s\left[s_1, a_2\right] = \frac{\sum_{t=1}^{T-1} \mathbb{I}_{\{s_t=s_1, s_{t+1}=s, a_{t+1}=a_2\}}}{T-1},$$

$$\boldsymbol{\Phi}_a^A[s, s'] = \frac{\sum_{t=1}^{T-1} \mathbb{I}_{\{s_t=s, a_t=a, s_{t+1}=s'\}}}{\sum_{t=1}^{T-1} \mathbb{I}_{\{s_t=s, a_t=a\}}}.$$

Step 2: Compute the surrogate moments $\hat{\boldsymbol{M}}_a$ and $\hat{\boldsymbol{M}}$ according to Equations (2.2) and (2.3).

Step 3: Perform SVD on $\boldsymbol{K}_s$, and construct the matrix $\boldsymbol{V}$ according to Equation (2.1).

Step 4: Compute the joint eigenbasis $\boldsymbol{B}$ using Equation (2.5). Then, recover the order of its column using the algorithm discussed in the Appendix.

Step 5: Recover $\boldsymbol{\Pi}^{lo}$ using the diagonals that result from diagonalizations according to Equation (2.4).

Step 6: Compute $\boldsymbol{\Pi}^{hi}$ with Equation (2.6).

2.3.4. Performance discussion

The algorithm consists of two parts, data collection with complexity $\mathcal{O}(T)$, and data processing with complexity $\mathcal{O}(\zeta^4\alpha\omega)$, dominated by the cost of computing $\boldsymbol{V}^{\mathrm{T}}\hat{\boldsymbol{M}}^{+}\hat{\boldsymbol{M}}_a\boldsymbol{V}$ and its eigenbasis. This gives us the total time complexity of $\mathcal{O}(T + \zeta^4\alpha\omega)$.

Comparison to the EM methods presented in Zhang and Paschalidis (2021) and Giammarino and Paschalidis (2021), which has time complexity of $\mathcal{O}(T\omega^2)$ and $\mathcal{O}(\zeta\alpha\omega^3)$ per iteration respectively, can be difficult. This is due to the fact that they have different bottlenecks, along with the fact that the method of moments is parameter-less while EM methods need

initialization. However, a general rule is that the larger the number of samples is relative to the number of states and actions, the better the method of moments performs compared to EM.

Another thing to note is that the techniques mentioned can be synergistic, with the output of the method of moments being good initialization for EM methods to refine.

2.4. Experiment

In this section, we examine the proposed algorithm in numerical experiments. We will use a similar setup to Zhang and Paschalidis (2021) to test our model. Let there be a finite state machine with four states and the following parameters:

$$\mathbf{\Pi}_1^{hi} = \begin{bmatrix} 0.67 & 0.33 \\ 0.16 & 0.84 \end{bmatrix}, \quad \mathbf{\Pi}_2^{hi} = \begin{bmatrix} 0.88 & 0.12 \\ 0.16 & 0.84 \end{bmatrix},$$
$$\mathbf{\Pi}_3^{hi} = \begin{bmatrix} 0.84 & 0.16 \\ 0.12 & 0.88 \end{bmatrix}, \quad \mathbf{\Pi}_4^{hi} = \begin{bmatrix} 0.84 & 0.16 \\ 0.33 & 0.67 \end{bmatrix}.$$

$$\mathbf{\Pi}_1^{lo} = \begin{bmatrix} 0.6 & 0.4 \\ 0.1 & 0.9 \end{bmatrix}, \quad \mathbf{\Pi}_2^{lo} = \begin{bmatrix} 0.7 & 0.3 \\ 0.15 & 0.85 \end{bmatrix},$$
$$\mathbf{\Pi}_3^{lo} = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{bmatrix}, \quad \mathbf{\Pi}_4^{lo} = \begin{bmatrix} 0.9 & 0.1 \\ 0.35 & 0.65 \end{bmatrix}.$$

$$\mathbf{\Phi}_1^A = \begin{bmatrix} 0.7 & 0.1 & 0.1 & 0.1 \\ 0.4 & 0.4 & 0.1 & 0.1 \\ 0.3 & 0.3 & 0.3 & 0.1 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix},$$
$$\mathbf{\Phi}_2^A = \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.1 & 0.3 & 0.3 & 0.3 \\ 0.1 & 0.1 & 0.4 & 0.4 \\ 0.1 & 0.1 & 0.1 & 0.7 \end{bmatrix}.$$

The error will be measured by:

$$\text{ERROR} = \sqrt{\left\|\mathbf{\Pi}^{lo} - \overline{\mathbf{\Pi}}^{lo}\right\|_2^2 + \left\|\mathbf{\Pi}^{hi} - \overline{\mathbf{\Pi}}^{hi}\right\|_2^2},$$

where $\overline{\mathbf{\Pi}}^{lo}, \overline{\mathbf{\Pi}}^{hi}$ are the predicted values of $\mathbf{\Pi}^{lo}, \mathbf{\Pi}^{hi}$.

For intuition, we can think of the states as locations on a number line (i.e., states with larger index are further right), the actions are $\mathcal{A} = \{\text{move-left}, \text{move-right}\}$, and the options are

$$\mathcal{O} = \{\text{tend-to-move-left}, \text{tend-to-move-right}\}.$$

Looking at the numbers, we can see that the agent wants to alternately move from left to right and right to left.

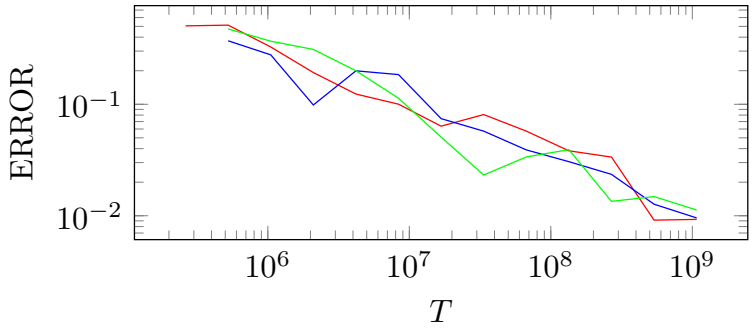


Fig 2.2: Log-log plot of the error versus the number of sample points for several realizations of the problem.

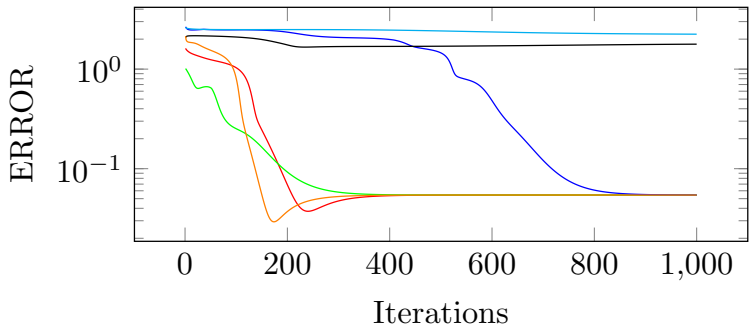


Fig 2.3: Iterations versus error of EM runs with various initializations, some of which do not converge.

In Fig. 2.2 we plot the error versus the number of samples for a few runs of our method in log scale. It can be seen that the error is polynomial relative to the number of samples.

For comparison purposes, Fig. 2.3 depicts a few EM runs with randomized initialization and a sample size of 3×10^5. It can be seen that initialization have a significant effect on EM's rate of convergence and whether or not it arrives at the correct optima. In contrast, the proposed method of moments does not require initialization.

2.5. Chapter summary

We developed a novel method of moments for Hierarchical Imitation Learning (HIL) that offers global convergence under mild regulatory conditions. Our method of moments for HIL is based on similar methods for HMMs and other latent variable models, and avoids the local convergence issues inherent in previous Expectation-Maximization (EM) approaches to HIL.

Chapter 3

Accumulated Cutoff Discrepancy Criterion

3.1. Introduction

Characterizing latent structures with meaningful real-world interpretations is a common task in scientific applications of statistical methods. This task often amounts to discovering unobserved physical "processes" that generate observable quantities. In practice, it is necessary to not only characterize each latent process but also determine *how many* such processes there are. This requires solving a model selection problem for a sequence of model families $\mathcal{M}^{(K)} = \{P_\theta : \theta \in \Theta_K\}$, $K = 1, 2, \ldots$, where K denotes the number of latent processes in the model. For example, K could be the number of components in a mixture model or the number of factors in a factor analysis model. Given some observed data $x_1, \ldots, x_N \overset{\text{iid}}{\sim} P_o$ that were generated by the output of K_o latent processes, the goal is to recover K_o and a model $\widetilde{P}_o \in \mathcal{M}^{(K_o)}$ such that $\widetilde{P}_o \approx P_o$. Likelihood-based model selection methods provide consistent estimation of K_o when $\mathcal{M}^{(K_o)}$ is *well-specified* (that is, $P_o \in \mathcal{M}^{(K_o)}$ but $P_o \notin \mathcal{M}^{(K)}$ if $K < K_o$). However, if $\mathcal{M}^{(K_o)}$ is *misspecified* (that is, $P_o \notin \mathcal{M}^{(K_o)}$), then these methods do not work as intended (Cai, Campbell and Broderick, 2021; Frühwirth-Schnatter, 2006; Guha, Ho and Nguyen, 2021; Miller and Dunson, 2019; Xue et al., 2024). The reason for this failure is that they optimize for some measure of *predictive performance* (e.g., some form of expected log loss). Therefore, as $N \to \infty$, these methods will select a sequence of models that converge to the distribution $P_\star \in \mathcal{M}^{(\infty)} = \bigcup_{K=1}^{\infty} \mathcal{M}^{(k)}$ that has minimal Kullback–Leibler divergence between P_o and all $P \in \mathcal{M}^{(\infty)}$. Since typically $P_\star \notin \mathcal{M}^{(K)}$ for any finite K, when using a predictive method for model selection, as the number of observations N increases, rather than obtaining better estimates, the opposite occurs: the distribution P_o is better estimated by adding spurious latent structures that compensate for the shortcomings of the parametric model. This problem is known as *overfitting* (Cai, Campbell and Broderick, 2021).

However, when trying discover real-world processes, the statistical problem is no longer predictive in nature, but rather *explanatory* (Shmueli, 2010). The aim is to infer meaningful constructs that capture the underlying causal structure – in this case the latent processes – even if doing so results in a fitted model with reduced predictive power. An alternative set of approaches are nonparametric or semi-parametric in nature. With weaker assumptions, they can often be more robust. However, because they do not rely on parametric assumptions,

they tend to be less sensitive and can underestimate the number of latent components. These nonparametric methods also lack the generality of the likelihood-based approaches. Hence, there is a need for robust model selection procedures that have the generality of likelihood-based methods like Akaike, Bayesian, and deviance information criteria (Akaike, 1974; Schwarz, 1978; Spiegelhalter et al., 2002) while retaining the robustness to parametric assumptions provided by the nonparametric methods. Notably, information criteria remain widely used due to their simplicity and the ease with which they can be incorporated into data analysis workflows – for example, when a user wants to use an existing (perhaps specialized) method to estimate the parameters of each model $\mathcal{M}^{(K)}$ $(K = 1, 2, \dots)$. We defer more detailed discussion of related work to Sections 3.2.6 and 3.3 once we have provided more context.

Summary of Contributions. In this chapter, we make the following contributions to the development of more reliable and robust methods for model selection for discovering real-world latent processes:

1. We define a formal notion of *robust model selection consistency* and argue that it captures key features any robust model selection method should satisfy.
2. We propose the *accumulated cutoff discrepancy criterion* (ACDC), as a simple, flexible approach to robust model selection. in which the observations are determined by combining unobserved outputs from K latent processes.
3. We show how to apply ACDC to the class of PMF models
4. We prove that ACDC provides robust model selection consistency for PMF models.
5. We illustrate the advantages of ACDC through two numerical experiments with simulated and real data.[1]

Note that while ACDC is broadly applicable, the primary focus of this chapter, and the core contribution of this thesis regarding ACDC, centers on its application within the domain of PMF models, including non-negative matrix factorization (NMF). We will be using mixture modeling as a simple scenario to facilitate an intuition to the component-level discrepancy, then move on to detail a more general way to construct this quantity.

3.2. Methodology

We first define robust model selection consistency, which naturally leads to a plug-in procedure, the accumulated cutoff discrepancy criterion (ACDC). We then describe how to apply ACDC for a broad class of latent variable models. In this section, we use mixture modeling as a running example to illustrate ideas. Let $\mathcal{F} = \{F_\phi \mid \phi \in \Phi\}$ denote the component distribution family and let $\eta \in \Delta_K$ denote the component weights, where $\Delta_K = \{\eta \in \mathbb{R}_+^K \mid \sum_{k=1}^K \eta_k = 1\}$ is the $(K-1)$-dimensional probability simplex. Denote the parameter set for the K-component mixture distributions by $\Theta^{(K)} = \Delta_K \times \Phi^K$, so the mixture model distribution family is

$$\mathcal{M}^{(K)} = \left\{ P_\theta = \textstyle\sum_{k=1}^K \eta_k F_{\phi_k} : \theta = (\eta, \phi_1, \dots, \phi_K) \in \Theta^{(K)} \right\}.$$

[1]Code to reproduce all experiments is available on GitHub: `https://github.com/TARPS-group/robust-model-selection-for-discovery`.

We can also write the generative process of the mixture model in terms of latent variables $z_n \in \{1, \dots, K\}$ that indicate which component observation n belongs to:

$$z_n \mid \theta \sim \mathsf{Categorical}(\eta) \qquad\qquad x_n \mid z_n = k, \theta \sim F_{\phi_k}.$$

Since we are interested in isolating the contribution of each component, it is this latent variable representation that will be most relevant. We discuss applications to other models in Sections 3.2.3 and 3.3.1.

3.2.1. Robust Model Selection Consistency

Generalizing the mixture model setting, consider a sequence of models $\mathcal{M}^{(1)}, \mathcal{M}^{(2)}, \dots, \mathcal{M}^{(K)}, \dots$, where K captures how many distinct latent components are generating the observed data. Assume that $\mathcal{M}^{(K)} = \{P_\theta \mid \theta \in \Theta^{(K)}\}$, where $\Theta^{(K)}$ is the parameter space and $P_\theta \in \mathcal{P}(\mathbb{X})$, the space of probability measures on the observation space $\mathbb{X}$. The objective is to identify the true number of processes K_o. Fix a *distribution-level discrepancy* $\mathcal{D}_{\text{dist}}$ on probability measures that will quantify the fit between P_θ and the data-generating distribution P_o. We do not assume a unique minimizing parameter since, at the very least, the component indices in latent variable models are non-identifiable. Let $\Theta_\star^{(K)}(P_o) := \arg\min_{\theta \in \Theta^{(K)}} \mathcal{D}_{\text{dist}}(P_o \mid P_\theta)$ denote the set of minimizing parameters. Alternatively, a practitioner might choose a parameter estimation procedure that is not model-based, in which case it might converge to a parameter value in some other set, which we also denote by $\Theta_\star^{(K)}(P_o)$.

The challenge in the misspecified setting is that for any $\theta_\star^{(K)} \in \Theta_\star^{(K)}(P_o)$, typically $\mathcal{D}_{\text{dist}}(P_o \mid P_{\theta_\star^{(K)}})$ is not minimized at $K = K_o$. In fact, in our settings of interest $\mathcal{M}^{(K)} \subsetneq \mathcal{M}^{(K+1)}$ but $P_o \notin \mathcal{M}^{(K)}$ for any K, so $\mathcal{D}_{\text{dist}}(P_o \mid P_{\theta_\star^{(K)}})$ is monotonically decreasing as K increases. To correctly recover K_o, the user must specify how much P_θ can deviate from P_o while remaining an acceptable approximation. Therefore, we introduce a second discrepancy which measures how well the *components* of P_o and P_θ match.

Since the components of P_o are unknown, the component contributions must be estimated based on model P_θ but using the distribution of data from P_o. Let the *component-level realized discrepancy* $\mathcal{D}_{\text{comp}}^{(K)}(\theta, k, P_o)$ quantify how close the inferred component $k \in \{1, \dots, K\}$ from P_o is to component k of the model $P_\theta^{(K)}$. To quantify the overall degree of component-level misspecification of P_o with true number of components K_o, define the *worst-case component-wise discrepancy*

$$\rho(P_o, K_o) := \sup_{\theta \in \Theta_\star^{(K_o)}(P_o)} \max_{k \in [K_o]} \mathcal{D}_{\text{comp}}^{(K_o)}(\theta, k, P_o).$$

For example, in the mixture model case we can construct a component-level realized discrepancy by inferring the component of P_o that would correspond to each mixture component. That is, if we assign each observation from P_o according to the conditional component probabilities $p(k \mid x, \theta) = \eta_k \frac{\mathrm{d}F_{\phi_k}}{\mathrm{d}P_\theta}(x)$, then the inferred kth component of P_o is

$$\widetilde{F}_{ok}^{(\theta)} = \frac{p(k \mid x, \theta)}{\int p(k \mid y, \theta) P_o(\mathrm{d}y)} P_o.$$

Given a choice of discrepancy measure $\mathcal{D}$ (e.g., $\mathcal{D}_{\text{dist}}$), we can set $\mathcal{D}^{(K)}_{\text{comp}}(\theta, k, P_o) = \mathcal{D}(\widetilde{F}^{(\theta)}_{ok} \mid F_{\phi_k})$.

Definition 5 (Robust model selection consistency)**.** *Fix a function $\kappa : \mathbb{R}_+ \times \mathbb{N} \to \mathbb{R}_+$. A model selection procedure $\widehat{K}(x_{1:N}, \rho) \in \mathbb{N}$ is* κ-robustly consistent for $\Theta_\star$ and $\mathcal{D}_{\text{comp}}$ *if, for any data-generating distribution P_o and true component number K_o that satisfies*

$$\inf_{\theta \in \Theta^{(K)}_\star(P_o)} \mathcal{D}_{\text{dist}}(P_o \mid P_\theta) \geq \kappa(\rho(P_o, K_o), K) \quad \text{for all } K \in \{1, \dots, K_o - 1\}, \tag{3.1}$$

it holds that, for $x_1, x_2, \dots \overset{iid}{\sim} P_o$,

$$\mathbb{P}\big\{\widehat{K}(x_{1:N}, \rho(P_o, K_o)) = K_o\big\} \overset{N\to\infty}{\longrightarrow} 1.$$

To interpret Definition 5 and the role of the function κ, it is helpful to compare robust model selection consistency to classical model selection consistency. Figure 3.1 provides a cartoon illustration of the differences. Classical model consistency typically requires that (1) $P_o \in \mathcal{M}^{(K_o)}$ and (2) $P_o \notin \mathcal{M}^{(K)}$ for $K < K_o$. Robust consistency weakens the first condition by allowing for a worst-case discrepancy $\rho(P_o, K_o) \neq 0$, rather than needing $\rho(P_o, K_o) = 0$. However, robust consistency strengthens the second by requiring a gap between P_o and all models for $K < K_o$. The size of this gap specified in Eq. (3.1) in terms of κ. Hence, we call κ the *gap function*. It would be natural to ask that $\kappa(\rho, K) = \rho$, although this may not always be possible.

3.2.2. A Plug-in Procedure

Inspired by Definition 5, we propose a simple plug-in procedure for model selection,. Assuming $\rho_o = \rho(P_o, K_o)$ were known, we would want to find the smallest value of K such that for all $k \in \{1, \dots, K\}$, we have $\mathcal{D}^{(K)}_{\text{comp}}(\theta_\star, k, P_o) \leq \rho_o$ for $\theta_\star \in \Theta^{(K_o)}_\star(P_o)$. However, since P_o and $\Theta^{(K_o)}_\star(P_o)$ are unavailable, we propose to instead use the empirical distribution $\widehat{P}_o = N^{-1}\sum_{n=1}^N \delta_{x_n}$ (here δ_x denotes the Dirac measure at x) and a point estimator $\widehat{\theta}^{(K)}$. Hence, we obtain the plug-in estimator $\widehat{\mathcal{D}}^{(K,k)}_{\text{comp}} = \mathcal{D}^{(K)}_{\text{comp}}(\widehat{\theta}^{(K)}, k, \widehat{P}_o)$. Since values of $\widehat{\mathcal{D}}^{(K,k)}_{\text{comp}} < \rho_o$ are not important from a model selection perspective, we truncate the estimator by replacing it with $\max(0, \widehat{\mathcal{D}}^{(K,k)}_{\text{comp}} - \rho)$, where ρ is a best estimate of ρ_o. We can view taking this maximum as serving a similar role to how the coarsened posterior conditions on the discrepancy having a known upper bound.

Rather than taking the maximum over the component-wise discrepancy estimates, for better robustness to noisy estimates we use a sum, which results in a robust model selection loss

$$\mathcal{R}^\rho(x_{1:n}, K) = \sum_{k=1}^K \max(0, \widehat{\mathcal{D}}^{(K,k)}_{\text{comp}} - \rho), \tag{3.2}$$

where for notational simplicity we have left the dependence of $\widehat{\mathcal{D}}^{(K,k)}_{\text{comp}}$ on $x_{1:n}$ (as a function of $\widehat{P}_o$ and $\widehat{\theta}^{(K)}$) implicit. Since the loss is the sum (i.e., accumulation) of discrepancies that

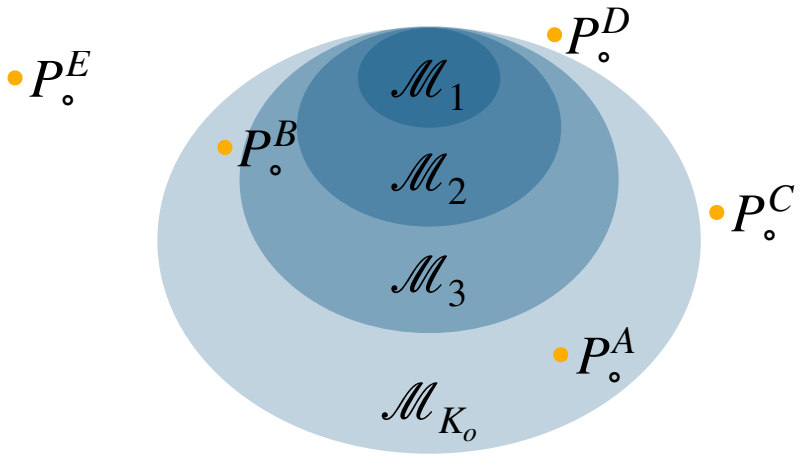


Fig 3.1: Cartoon illustration of the differences between traditional and robust model selection consistency where the true model corresponds to $K_o = 4$, with the nested models indicated by the gray ovals. We contrast five possible data-generating distributions, $P_o^A, \ldots, P_o^E$, indicated by the gold points. Since $P_o^A, P_o^B \in \mathcal{M}^{(K_o)}$ but are not in $\mathcal{M}^{(K)}$ for $K < K_o = 4$, for these models K_o can be consistently estimated. However, since $P_o^C, P_o^D, P_o^E \notin \mathcal{M}^{(K_o)}$, for these distributions K_o cannot be estimated consistently in the traditional sense. On the other hand P_o^A, P_o^B, P_o^C, and P_o^D are close to $\mathcal{M}^{(K_o)}$, so K_o could potentially be robustly and consistently estimated in these four cases. However, P_o^B and P_o^D are also close to $\mathcal{M}^{(3)}$, so robustly consistent estimation of K_o is feasible only for P_o^A and P_o^C. Since P_o^E is far from $\mathcal{M}^{(K_o)}$, K_o would not be consistently estimable – either in the traditional or robust sense.

have been truncated (i.e., cut off) at ρ, we call Eq. (3.2) the *accumulated cutoff discrepancy criterion* (ACDC). The corresponding robust model estimator is

$$\widehat{K}^{\rho}(x_{1:n}) = \min\{\arg\min_{K} \mathcal{R}^{\rho}(x_{1:n}, K)\}.$$

Since $\arg\min_K$ may return a set of values if the loss is equal to zero for more than one value of K, it is necessary to include an additional min operation to select the smallest value from the set. We provide a number of methods for determining ρ in Section 3.2.5.

3.2.3. Modeling Framework

Guaranteeing robust model consistency requires specific choices of model and component-level discrepancy. In this paper, we consider a flexible modeling framework in which the observed data are the result of K distinct latent sources (Fig. 3.2). This framework will lead to a natural choice of component-level discrepancy, of which the mixture model discrepancy proposed Section 3.2.1 is a particular case.

We allow each observation $x_n \in \mathcal{X}$ to have sample-specific covariates $w_n \in \mathcal{W}$. Assume that x_n depends on process-level contributions $y_{n1}, \ldots, y_{nK} \in \mathcal{Y}$ via the deterministic function $g^{(K)} \colon \mathcal{Y}^{\otimes K} \to \mathcal{X}$:

$$x_n = g^{(K)}(y_{n1}, \ldots, y_{nK}).$$

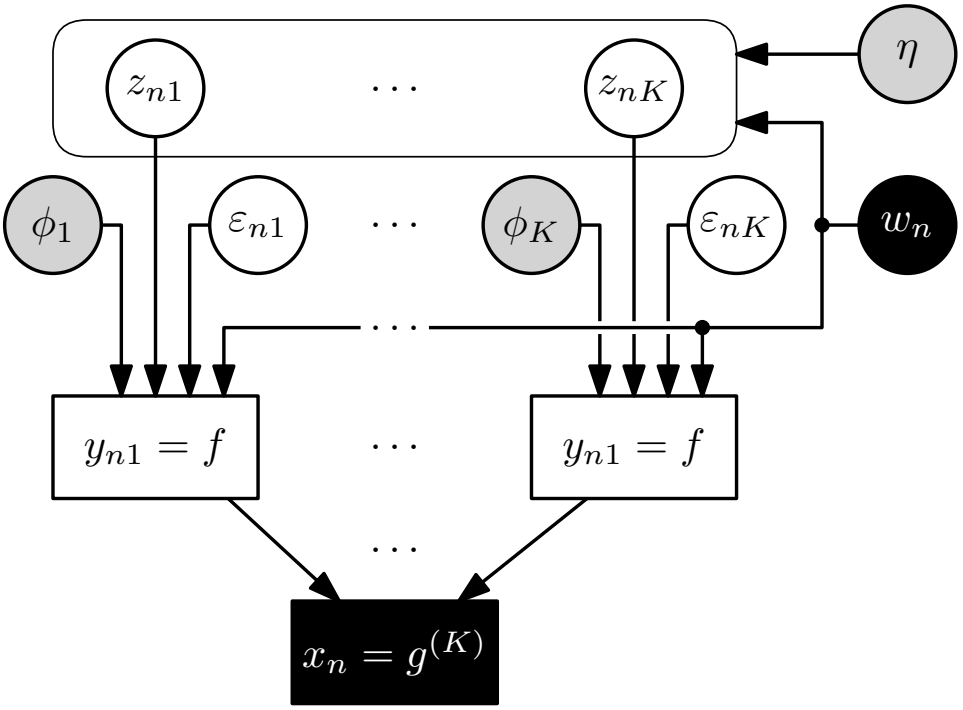


Fig 3.2: Graphical representation of the general form for model $\mathcal{M}^{(K)}$ for a single observation x_n. Circles denote random variables while squares denote deterministic variables. A gray background indicates global parameters while a black background indicates an observed quantity.

For example, we could have $g^{(K)}(y_1, \ldots, y_K) = \sum_{k=1}^{K} y_k$ or $g^{(K)}(y_1, \ldots, y_K) = \max_k y_k$. The process-level contributions are in turn determined by an observation-specific latent variable $z_n = (z_{n1}, \ldots, z_{nK}) \in \mathcal{Z}^{(K)} \subseteq \mathcal{Z}^{\otimes K}$ and the process-specific parameters $\phi_1, \ldots, \phi_K \in \Phi$. We assume both $\mathcal{Y}$ and $\mathcal{Z}$ contain a *null value* $\emptyset$, which indicates no contribution. Specifically, we assume $g^{(K)}$ has the following *no contribution property*: for all $y_1, \ldots, y_{K-1} \in \mathcal{Y}$,

$$g^{(K)}(y_1, \ldots, y_{K-1}, \emptyset) = g^{(K-1)}(y_1, \ldots, y_{K-1}).$$

Given a deterministic function $f \colon \mathcal{W} \times \mathcal{Z} \times \Phi \times \mathbb{R} \to \mathcal{Y}$ and independent noise random variables $\varepsilon_{nk} \overset{\text{iid}}{\sim} G$, the component-level contributions are given by

$$y_{nk} = \begin{cases} \emptyset, & \text{if } z_{nk} = \emptyset, \\ f(w_n, z_{nk}, \phi_k, \varepsilon_{nk}), & \text{otherwise.} \end{cases}$$

If there are no covariates, we drop the dependence on w_n and write $f(z_{nk}, \phi_k, \varepsilon_{nk})$ instead. Typically $\mathcal{Z} \subseteq \mathbb{R}$ and z_{nk} represents the activity level of the kth process for the nth observation. In such cases, usually $\emptyset = 0$. For a global parameter $\eta \in \mathcal{E}^{(K)}$, we assume the observation-specific latent variables are independent but may depend on the sample-specific covariates:

$$z_n \mid \eta, w_n \overset{\text{ind}}{\sim} H^{(K)}_{\eta, w_n}.$$

If there are no covariates, we drop the dependence on w_n and write $H^{(K)}_{\eta}$ instead.

Our framework captures many common model types:

Example 1 (Mixture Modeling). *We can recover a general mixture model by taking* $H_\eta^{(K)} = \mathsf{Categorical}(\eta)$*, so* $z_{nk} \in \{0,1\}$ *and* $\sum_{k=1}^K z_{nk} = 1$*. Given a mixture component distribution family* $\mathcal{F} = \{F_\phi \mid \phi \in \Phi\}$*, define* f *such that, for* $\varepsilon \sim G$*, it holds that* $f(0,\phi,\varepsilon) = 0$ *and* $f(1,\phi,\varepsilon) \sim F_\phi$*. Finally, take* $g^{(K)}$ *to be the summation operator. Hence, if* $z_{nk} = 1$*, then* $x_n = y_{nk} \sim F_{\phi_k}$*.*

Example 2 (Probabilistic Matrix Factorization). *For probabilistic matrix factorization (PMF),* $x_n \in \mathbb{R}^D$*. Let* $\mathcal{Z} \subseteq \mathbb{R}$ *and* $\Phi \subseteq \mathbb{R}^D$*. We assume that* $\mathcal{F} = \{F_\mu \mid \mu \in \mathbb{R}^D\}$ *is a location family of distributions satisfying* $\int x F_\mu(\mathrm{d}x) = \mu$ *for all* $\mu \in \mathbb{R}^D$*. Let* f *satisfy* $f(z,\phi,\varepsilon) \sim F_{z\phi}$ *if* $\varepsilon \sim G$*. For example, in nonnegative matrix factorization,* $F_\mu = \mathsf{Poiss}(\mu)$ *while in classical factor analysis* $F_\mu = \mathcal{N}(\mu, \sigma^2)$*, where* σ^2 *can also be learned. Finally, take* $g^{(K)}$ *to be the summation operator.*

Our framework is similarly applicable to supervised variants of probabilistic matrix factorization models, functional clustering problems, and a variety of other latent variable models (Blei and Lafferty, 2007; Carvalho et al., 2008; Chiou and Li, 2007; Cunningham and Yu, 2014; Dunson, 2000; West, 2003).

3.2.4. Constructing the Component-level Discrepancy

To define the component-level discrepancy, it is tempting to directly apply the approach we took for mixture models and quantify the difference between the distributions of $y_{n1}, \ldots, y_{nK}$ when $x_n \sim P_o$ and the modeled distributions of $y_{n1}, \ldots, y_{nK}$. However, the distributions of $y_{1k}, y_{2k}, \ldots, y_{nk}, \ldots$ may be different due to the sample-specific dependence on w_n and z_{nk}. To address this issue, we instead consider the discrepancy between the conditional distribution of the noise variables ε_{nk} given $x_{1:N}$ and the assumed noise distribution G. However, we must exclude ε_{nk} if $y_{nk}^{(K)} = \emptyset$ because then it is no longer part of the graphical model (see Fig. 3.2). Dropping the dependence on n, the inferred distribution of the kth noise variable is

$$\widetilde{G}_k^{(\theta)} = \int \mathcal{L}(\varepsilon_k \mid y_k \neq \emptyset, x, w, \theta) P_o(\mathrm{d}x, \mathrm{d}w),$$

where $\mathcal{L}(\varepsilon_k \mid \cdots)$ denotes a conditional law of ε_k. Hence, define the discrepancy for the kth component by $\mathcal{D}_{\mathrm{comp}}^{(K)}(\theta, k, P_o) = \mathcal{D}(\widetilde{G}_k^{(\theta)} \mid G)$.

To define an empirical version of $\mathcal{D}_{\mathrm{comp}}^{(K)}(\theta, k, P_o)$, let $\widehat{G}_{nk}^{(K)} = \mathcal{L}(\varepsilon_{nk} \mid x_n, w_n, \widehat{\theta}^{(K)})$, define the usage indicators $u_{nk}^{(K)} = \mathbb{1}(y_{nk}^{(K)} \neq \emptyset)$, and denote the number of observations for component k as $N_k^{(K)} = \sum_{n=1}^N u_{nk}^{(K)}$. Then the empirical distribution of the noise variables for component k is

$$\widehat{G}_k^{(K)} = \frac{1}{N_k^{(K)}} \sum_{n=1}^N u_{nk}^{(K)} \widehat{G}_{nk}^{(K)}.$$

Hence, an estimate of discrepancy for the kth component is given by

$$\widehat{\mathcal{D}}_{\mathrm{comp}}^{(K,k)} = \mathcal{D}(\widehat{G}_k^{(K)} \mid G). \tag{3.3}$$

In some scenarios, it is necessary to replace the divergence with an estimator because $\mathcal{D}(\widehat{G}_k^{(K)} \mid G)$ is undefined (e.g., the KL divergence) or not efficiently computable (e.g., the Wasserstein distance).

3.2.5. Choosing ρ

The value of ρ is problem dependent, as it quantifies the maximum amount of model misspecification of each process. We propose two complementary approaches to selecting ρ that take advantage of the fact that the robust loss is a piecewise linear function of ρ. Therefore, given a fitted model for each candidate K, we can easily compute the loss for all values of ρ.

Using domain knowledge. The first approach aims to leverage domain knowledge. Specifically, it is frequently the case that some related datasets are available with "ground-truth" labels either through manual labeling or via *in silico* mixing of data where group labels are directly observed (see, e.g., de Souto et al., 2008). In such cases, an empirically optimal ρ value for one or more such datasets with ground-truth labels can be determined by maximizing a problem-appropriate accuracy metric such as F-measure. Because ρ quantifies the divergence between the true process distributions and the model estimates, we expect the values found using this approach will generalize to new datasets that are reasonably similar. We illustrate this approach in **??**. Alternatively, if real data with ground truth K_o is unavailable, plausible simulated data could be used to calibrate ρ instead (Xue et al., 2024).

A generally applicable approach. For applications where there are no related datasets with ground-truth labels available, we propose a second approach. After estimating the model parameters for each fixed K and computing all process-wise divergences, we plot the loss as a function of ρ for each $K \in \{K_{\min}, \ldots, K_{\max}\}$. For readability, introduce a small positive constant $\lambda \ll 1$ and plot $\mathcal{R}^{\rho}(x_{1:N}, K) + \lambda K$ so it is possible to distinguish the lines when the loss is exactly zero. The optimal model is determined by identifying the number of processes which is best over a substantial range of ρ values, with ρ as small as possible. The idea behind this selection rule is to identify the first instance of stability, indicating that allowing just a small amount of additional misspecification (by increasing ρ) doesn't notably improve the loss. However, subsequent stable regions that appear afterward might introduce too much tolerance, potentially resulting in model underfitting. This approach is similar in spirit to the one introduced for heuristically selecting the α parameter for the coarsened posterior (Miller and Dunson, 2019).

Automation. Building on the same heuristic intuition, we can automate the selection of ρ by defining a minimum width $\Delta_{\min}$ for which an interval can be recognized as the stability region. Specifically, we keep track of the smallest ρ value, ρ_{start}, for which a K-specific penalized loss becomes minimal among all other K-specific losses. We then identify the next ρ value, ρ_{end}, at which this same loss curve is no longer the minimum. The difference between $\Delta = \rho_{\text{end}} - \rho_{\text{start}}$ defines an interval of stability. If $\Delta \geq \Delta_{\min}$ (i.e., Δ has the predefined minimum width), it is recognized as a stability interval, the corresponding value of K is chosen to be $\widehat{K}$. Otherwise, ρ_{end} becomes ρ_{start} and repeat the procedure to compute the new ρ_{end} and Δ, check if $\Delta \geq \Delta_{\min}$, and so forth. The value of $\Delta_{\min}$ should be set based on preliminary manual experiments to estimate a suitable stability region width for automated selection in larger batches. This approach allows users to adjust the interval threshold to balance the tradeoff between avoiding underfitting ($\Delta_{\min}$ sufficiently small) and ensuring appropriately conservative and stable model selection ($\Delta_{\min}$ sufficiently large).

3.2.6. Related Work

There is limited work on general-purpose approaches to robust model selection with the goal of ensuring interpretability. Recent work on robustifying likelihoods to small model perturbations offer one promising strategy (Chakraborty, Bhattacharya and Pati, 2023; Dewaskar et al., 2023; Wu et al., 2024) . However, these methods aim to replace existing parameter estimation methods rather than augment them – which is a key goal of the present work. Perhaps most closely related to our work is Miller and Dunson (2019), which proposes an elegant robust Bayesian model selection procedure that employs a technique they call *coarsening*. Unlike the standard posterior, which assumes the data were generated from the assumed model (that is, it conditions on $x = x_{\text{obs}}$), the *coarsened posterior* conditions on the "true model data" being close to the observed data (that is, it conditions on the estimated Kullback–Leibler divergence between x and x_{obs} being less than some threshold γ) – hence, sacrificing predictive power for greater robustness. However, using the coarsened posterior approach has a potentially high computational cost because it requires running Markov chain Monte Carlo dozens of times to heuristically determine a suitable robustness threshold (Miller and Dunson, 2019; Xue et al., 2024). In addition, while coarsening offers good robustness in many scenarios, it does not have have any formal correctness guarantees and can fail in simple situations .

3.3. Application to PMF

We now turn to illustrating the use of ACDC in some representative applications while also providing theoretical support for our approach by showing that ACDC is robustly consistent. For readability, theorems are stated informally in the main text. Formal statements of assumptions and results are deferred to Appendix C.

For the experiments in this section, we use the KL divergence as the discrepancy measure $\mathcal{D}$ for the calculation of the estimated component-level discrepancy $\widehat{\mathcal{D}}_{\text{comp}}^{(K,k)}$ defined in Eq. (3.3). We compare to BIC since, like ACDC, it only requires a point estimate for each value of K. Alternative criteria like AIC and DIC would give similar results. However, since BIC has a larger penalty than AIC (which DIC generalizes), it will be more conservative and hence tend to choose smaller values of K. See, for example, Cai, Campbell and Broderick (2021); Miller and Dunson (2019); Xue et al. (2024) for numerical examples showing that Bayesian model selection does not resolve the overfitting problem.

3.3.1. Probabilistic Matrix Factorization

ACDC is robustly consistent for PMF.

Theorem 3. *ACDC using $\widehat{\mathcal{D}}_{\text{comp}}^{(K,k)}$ defined in Eq. (3.3) is κ-robustly consistent for probabilistic matrix factorization if the underlying component discrepancy $\mathcal{D}$ is the KL divergence, $\mathcal{D}_{\text{dist}} = d_{\text{BL}}$, and $\kappa(\rho, K) = \sqrt{K\rho/2}$.*

In our numerical experiments, we take $G = \mathsf{Unif}([0,1]^D)$, the uniform distribution on the D-dimensional hypercube. This choice is without loss of generality when using KL divergence as the discrepancy measure since it is invariant to diffeomorphisms of the noise variables ε_{nk}. Moreover, it leads to a universal choice of $f(z, \phi, \cdot) = F_{z\phi}^{-1}$, the inverse CDF. However, in

specific scenarios other choices for G and f might be preferred due to considerations such as ease of implementation or stability of KL divergence estimation. For example, in the Gaussian case, we could take $G = \mathcal{N}(0, I)$ and $f(z, \phi, \varepsilon) = z\phi + \sigma\varepsilon$.

In addition to BIC, we compare ACDC to *parallel analysis* (PA) (Buja and Eyuboglu, 1992; Horn, 1965), which is a commonly used method for model selection for probabilistic matrix factorization. To account for permutation invariance of the parameters, BIC is computed as

$$BIC\left(x_{1:N}, z^{(K)}_{1:K,1:N}, \phi^{(K)}_{1:K}\right) = K\log(N) - 2\log\left(p\left(x_{1:N} \;\middle|\; z^{(K)}_{1:K,1:N}, \phi^{(K)}_{1:K}\right)\right) + 2\log(K!).$$

Parallel analysis is carried out by generating the scree plot (ordered PCA eigenvalues) of the data against that of randomly generated matrices of the same size. These random matrices are generated by independently permuting each row of the data matrix. The results of Dobriban (2020) show that, in general, PA can be conservative and miss factors with a low signal-to-noise ratio.

While other robust model selection approaches for matrix factorization exist, they all have limitations that lead us to not include them in our empirical comparison. The method of Liu (2019) is limited to Gaussian nonnegative matrix factorization (NMF). Pelizzola, Laursen and Hobolth (2023) aim to address the problem of robustness for the case of Poisson NMF using two different approaches: a negative binomial instead of a Poisson likelihood to improve the model's ability to handle overdispersed data, and a testing routine inspired by cross validation. While this approach provides reasonable results, using the negative binomial only targets a very specific type of data–model mismatch, and the cross validation approach does not have any correctness guarantees. Bai and Ng (2002) propose an information criterion-based approach for factor analysis and provide an asymptotic consistency result for the case where the input dimension tends towards infinity. However, their main result applies only to the Gaussian NMF model with principle component analysis (PCA) as the estimation method. Particularly in NMF applications, another common approach is to evaluate the stability of the NMF solution across multiple runs. The *cophenetic correlation coefficient* (Brunet et al., 2004) is one such example. A similar stability-based principle is adopted by the widely-used toolset SigProfilerExtractor (Islam et al., 2022), which uses a consensus bootstrap approach to ensure results are consistent and reproducible. However, these approaches are computationally costly due to its reliance on repeated NMF executions. Furthermore, the results of Xue et al. (2024) demonstrate empirically that consensus the bootstrap methodology – much like PA – tend to be conservative, underestimating the true number of factors.

3.3.2. Mutational Process Discovery

Exposure to, and presence of, carcinogenic processes such as UV radiation, tobacco smoke, defective DNA repair mechanisms, and naturally occurring biochemical reactions, generate characteristic patterns of somatic mutations known as *mutational signatures* (Alexandrov et al., 2013; Nik-Zainal et al., 2012). Mutational signature-based analyses have contributed to novel insights in cancer research (e.g., Alexandrov et al., 2013, 2020; Li et al., 2020; Nik-Zainal et al., 2012) and are leading to emerging translations in clinical settings (e.g., Chakravarty and Solit, 2021). The most widely used approach to signature discovery is to fit a Poisson non-negative matrix factorization (NMF) model. For the nth tumor sample, the data consist

of a count vector $x_n \in \mathbb{N}^D$, where D is the number of mutation types being considered (e.g., there are $D = 96$ single-base substitution types with a trinucleotide context). The number of mutations of type d in sample n due to mutational process k is given by $y_{nkd}^{(K)} \sim \mathsf{Poiss}(\phi_{kd}^{(K)} z_{nk}^{(K)})$ and hence the total number of mutations of type d in sample n is $x_{nd} = \sum_{k=1}^{K} y_{nkd}^{(K)}$.

Because it is nearly impossible to obtain ground-truth signatures for real data, we use simulated breast cancer data based on the COSMIC v2 catalog and the pan-cancer analysis of whole genomes (PCAWG), following the procedure of Xue et al. (2024). First, we applied nonnegative least squares regression to the count matrix and COSMIC signatures, resulting in best fit exposure vectors. We then selected the signatures with significant loadings contributions and used them as the ground-truth signatures ϕ_o, with the inferred per-sample loadings serving as the ground truth exposures, $z_{o1}, \ldots, z_{oN}$. Finally, we generated four synthetic datasets: one well specified, and three others each with a different form of model misspecification. The forms of misspecification we use are from Xue et al. (2024):

- **Perturbation:** for each x_n, the signatures ϕ_o are stochastically perturbed before being used to simulate the observed counts.
- **Contamination:** for each x_n, in addition to the ground truth signatures ϕ_o, a randomly generated signature with small exposure is included in the sampling process.
- **Overdispersion:** the data is sampled from a negative binomial distribution instead of a Poisson distribution.

For each value of K, we compute the MLE of the signature and loadings parameters using the multiplicative update algorithm (Lee and Seung, 2000). As is standard practice, we quantify the signature recovery error using the cosine difference $D_{\cos}(\phi, \phi_\star) = 1 - \langle \widetilde{\phi}, \widetilde{\phi}_\star \rangle$, where for any vector v, we define $\widetilde{v} = v / \|v\|_2$. For the exposures, we quantify the error using the relative average difference $D_{\text{rad}}(z, z_\star) = |\overline{z} - \overline{z}_\star| / \overline{z}_\star$, where for a vector $v \in \mathbb{R}^N$, $\overline{v} = N^{-1} \sum_{n=1}^{N} v_n$. To evaluate the quality of an estimate as a whole, we perform bipartite matching against the ground truth by minimizing the metric

$$D^{(K)}(\sigma) = \sum_{k=1}^{K} \left[D_{\cos}\left(\phi_k^{(K)}, \phi_{o\sigma(k)}\right) + 0.1 \tanh D_{\text{rad}}\left(z_k^{(K)}, z_{o\sigma(k)}\right) \right],$$

where $\sigma \colon [K_o] \to [K]$ denotes an injective matching function. We bound and down-weight the D_{rad} using the $0.1 \tanh$ transform so that signature reconstruction accuracy is the main determinant for matching and exposure accuracy acts a tiebreaker when the signature accuracies are ambiguous. Given the optimal matching $\sigma_\star = \arg\min_\sigma D^{(K)}(\sigma)$, the accuracy scores are defined as the worst-case cosine and relative average errors, given by $L_\phi^{(K)} = \max\{D_{\cos}(\phi_k^{(K)}, \phi_{o\sigma_\star(k)}) : k = 1, \ldots, K\}$, and $L_z^{(K)} = \max\{D_{\text{rad}}(\overline{z}_k^{(K)}, \overline{z}_{o\sigma_\star(k)}) : k = 1, \ldots, K\}$.

Figures 3.3 and E.1 show that, across all four datasets, BIC selects $\widehat{K} = K_{\max}$ – even when the data is well specified. On the other hand, PA consistently underestimates K_o, estimating $\hat{K} = 2$. Finally, ACDC selects $\widehat{K} = 7$ or 8, which correspond to the some of the largest values of K for which the parameter estimates still have reasonably small error and meaningful decomposition. These results suggest ACDC is a promising alternative to existing approaches for selecting the number of mutational signatures, which all suffer from some combination of high computational cost and lack of statistical rigor (Alexandrov et al., 2020; Xue et al., 2024).

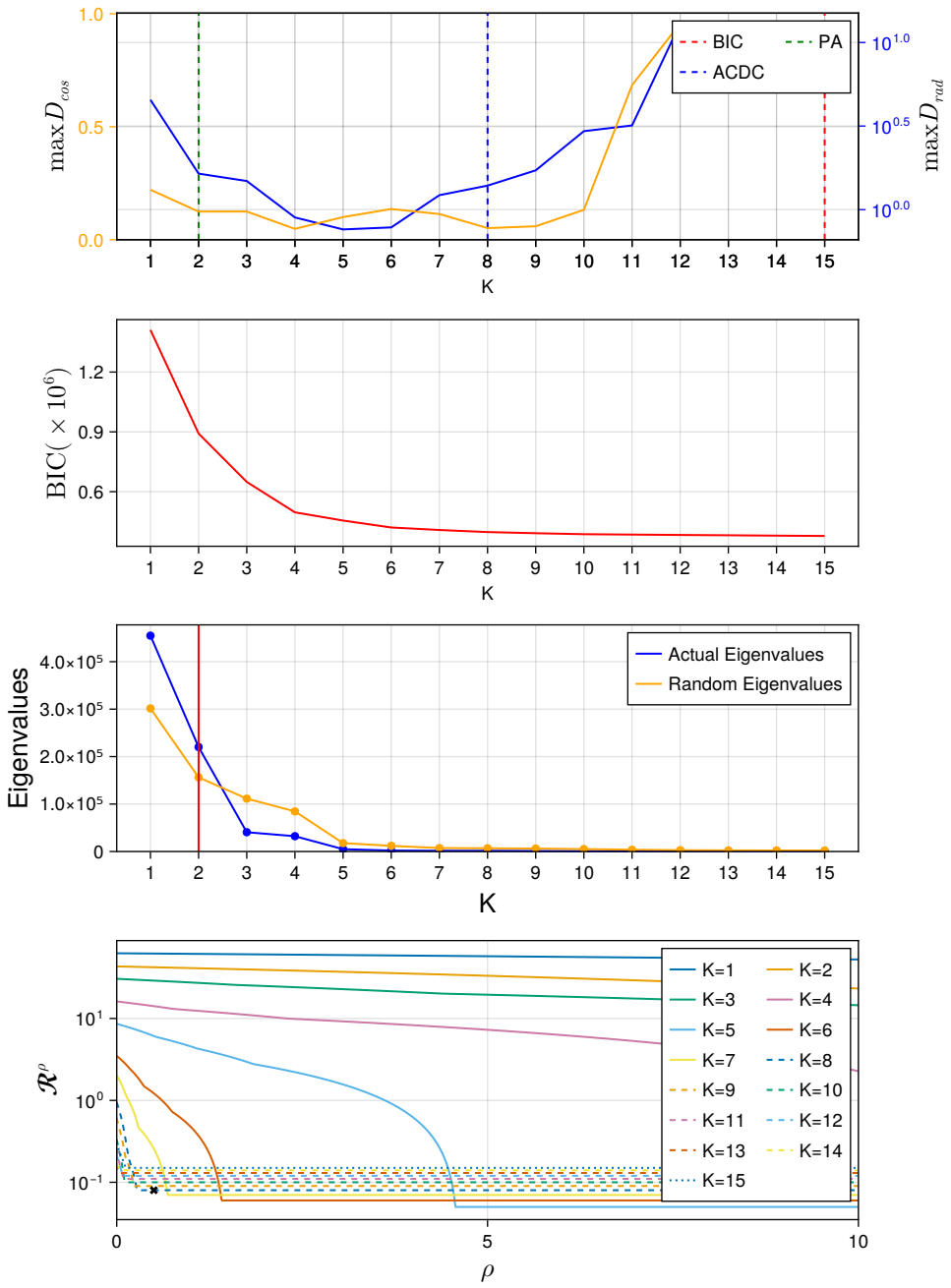


Fig 3.3: Estimation quality of mutational signature discovery for perturbed synthetic breast cancer data (Section 3.3.2). **(top)** Errors of signature and loadings estimates. **(middle top)** K versus value of BIC. BIC selects $\widehat{K} = K_{\max}$. **(middle bottom)** Scree plot generated from the dataset, indicating that $K = 2$ is the optimal choice of K. **(bottom)** Structurally aware loss for $K \in \{1, \ldots, K\}$, the wide region with the smallest ρ is marked by the cross mark, indicating $K = 7$ or $K = 8$ is the most appropriate choice of K.

3.3.3. Materials Discovery using Hyperspectral Imaging

Hyperspectral remote sensing data is used in applications such as environmental monitoring and city planning (Brook and Dor, 2016; Ji et al., 2016; Lin and Zhang, 2017). Hyperspectral data is collected as an image in which each pixel specifies the intensity of light at each observed wavelength. However, due to low spatial resolution, each pixel can be a mixture of materials, each reflecting different amounts of light at each wavelength. *Hyperspectral unmixing* refers to the unsupervised extraction of spectral signatures corresponding to materials (called *end-members*) and the abundances of these materials from each pixel. While it is reasonable to assume the intensities are observed with Gaussian noise, the contributions from each material obviously cannot be negative. However, incorporating this non-negativity constraint into the model is non-trivial (e.g., if using off-the-shelf parameter estimation methods), and so the constraint is often ignored. Thus, to illustrate the benefits of our approach in enabling principled model selection in a setting with clear misspecification, we will use a Gaussian factor analysis model.

We apply the model to a 307×307 hyperspectral image of an urban area, with each pixel representing a plot 2m $\times$ 2m in size.[2] After discarding certain wavelengths due to dense water vapor and atmospheric effects, each pixel consists of 162 channels with wavelengths ranging from 400nm to 2500nm. There are three versions of ground truth, containing 4, 5, and 6 end-members respectively. The 6 end-member version includes an additional material named "metal", which, through manual inspection, was found to contribute to only a small part of the image. As a result, none of the NMF algorithms we tested accurately recovered this end-member. Therefore, we use the ground truth with 5 end-members for this experiment (visualized in Fig. E.2(a))

We obtain point estimates of the factorization using a modified implementation of the coordinate descent method (Cichocki and Phan, 2009). We judge the quality of the estimates by quantifying how close the inferred material abundances are to the ground truth using the soft adjusted Rand index (sARI; Flynt, Dean and Nugent, 2019). Figure 3.4 shows that ACDC selects $\widehat{K} = K_o = 5$, which also maximizes the sARI. BIC overfits, selecting $\widehat{K} = 7$, while PA underfits, selecting $\widehat{K} = 3$. Fig. E.2(b)–(e) provides a qualitative understanding of the result. The small gap between $K = 3$ and $K = 4$ can be explained through the addition of the "grass" material that is very similar to the already included "tree" material. The large gap between $K = 4$ and $K = 5$ can be explained by the addition of the very different "asphalt" material, Finally, $K = 6$ only adds a "residual" material compared to $K = 5$, which can be interpreted as a different shade of "grass" and is clearly an artifact of overfitting.

3.4. Chapter summary

In this chapter, we have developed a general theoretical and methodological framework for ensuring mechanistic interpretability when selecting the number of latent components in a latent variable model. As two applications, we showed that our ACDC method is robustly consistent and empirically effective for mixture models and probabilistic matrix factorization. These results open the door to a number of directions for future work.

[2] https://rslab.ut.ac.ir/data

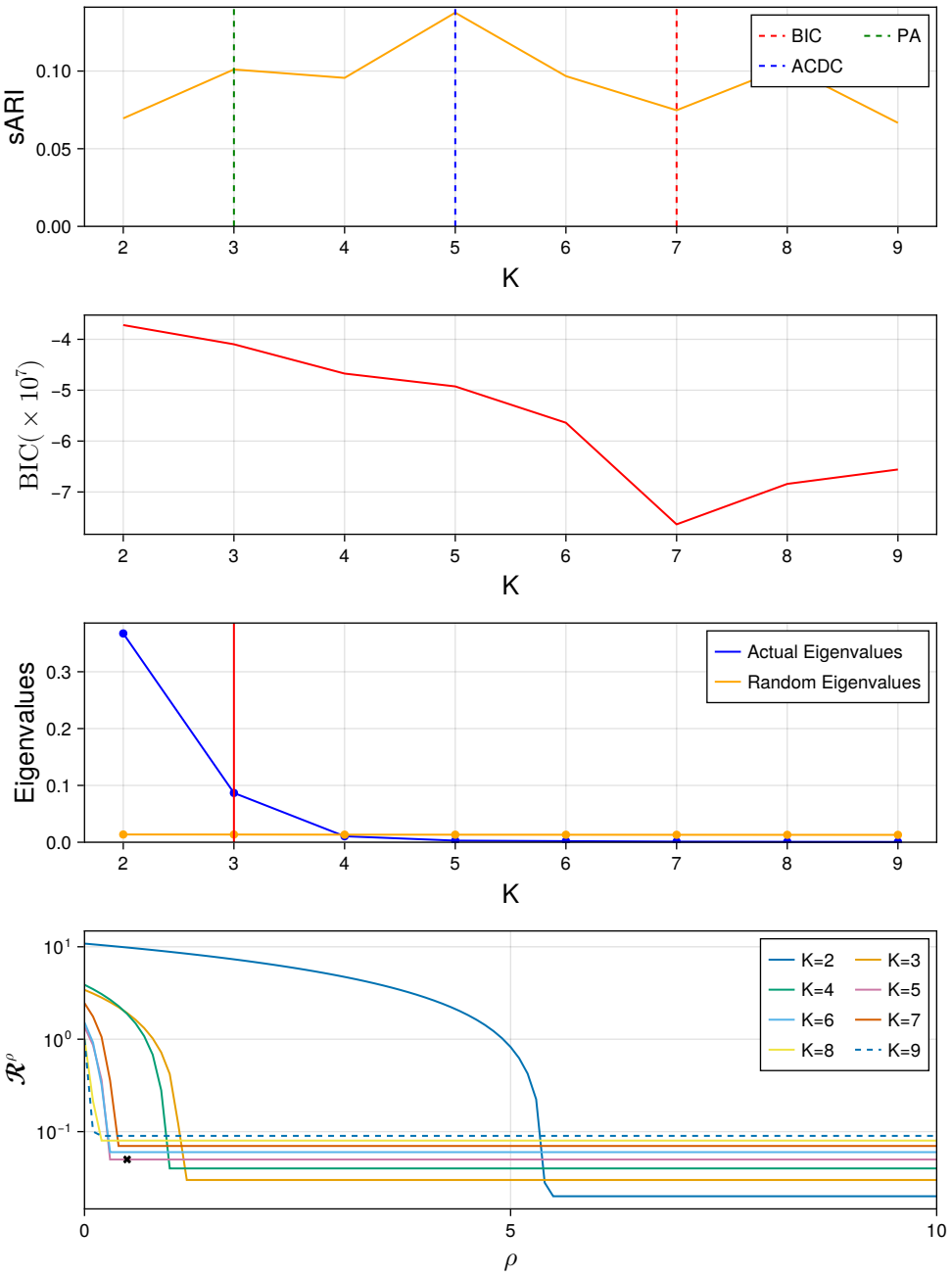


Fig 3.4: Model selection of end-members for hyperspectral urban dataset (Section 3.3.3). **(top)** Quality of solution measured by sARI, showing $K = 5$ is the solution with material distribution that best matches the ground truth. **(middle top)** Quality of solution ranked using BIC, indicating $K = 7$ is the best fitting solution. **(middle bottom)** Scree plot generated from the dataset, indicating that $K = 3$ is the optimal choice of K. **(bottom)** Structurally aware loss for $K = 2, \ldots, 9$, the wide region with the smallest ρ is marked by the cross mark, indicating $K = 5$ is the most appropriate choice of K.

Chapter 4

Discussion and Future Work

4.1. Summary of Contributions

This thesis develops methodologies that shift the focus of machine learning from predictive accuracy toward the robust and interpretable discovery of latent structures. Two primary contributions are presented to address this challenge:

1. A Spectral Method of Moments for Hierarchical Imitation Learning (HIL): This work introduces a novel, non-iterative algorithm for learning hierarchical policies from expert demonstrations where high-level options are latent. By leveraging low-order observable moments, this method provides global and asymptotically consistent parameter estimation, directly overcoming the limitations of local optima and initialization sensitivity inherent in standard EM approaches.
2. The Accumulated Cutoff Discrepancy Criterion (ACDC): This work validates and applies ACDC, a general model selection framework, for robustly identifying the true number of latent processes (K_o) under model misspecification. The primary focus of this thesis was on the application and validation of ACDC within the domain of PMF/NMF. By introducing a "cutoff discrepancy" (ρ), ACDC avoids overfitting to statistical noise and model artifacts, a common failure mode for traditional criteria like BIC/AIC that are optimized for prediction.

Collectively, these contributions provide tools and options for addressing the full process of robust discovery, from determining "how many" processes exist (ACDC) to defining "what they are" (Spectral HIL).

4.2. Limitations and Future Works

While the proposed methods offer significant advantages, they also present clear limitations and corresponding avenues for future research. These are discussed separately for each contribution.

For the method of moments, its global convergence guarantees rely on mild regulatory conditions, such as full rank and identifiability. Future work should include a thorough investigation of the method's performance and sensitivity when these assumptions are nearly or partially violated, which may occur in noisy, real-world data. Additionally, we can also

examine its extension to situations in which the options form a semi-Markov (rather than a Markov) process.

As for ACDC, in applications where related labeled datasets are not available, we are only able to provide a heuristic method for calibrating the degree of misspecification, as quantified by ρ. Ideally, we would like to have a more rigorous criteria. However, given the nonparametric nature of misspecification, we suspect that a fully general solution does not exist; for example, the coarsened posterior similarly requires a heuristic calibration step. One alternative to explore in the future is the simulation-based calibration method from Xue et al. (2024), which was developed for the coarsened posterior but could easily be used with ACDC as well. However, generally speaking, a user must have *some* prior knowledge about the degree of model misspecification – and believe that the misspecification will be reasonably small as measured by the chosen discrepancy $\mathcal{D}$. Moreover, if the degree of misspecification is very large, we should not expect any robust model selection procedure to work well (cf. P_o^E in Fig. 3.1).

It would be valuable to further develop our robust consistency theory; for example similar results could be proved for other common model classes. It would also be useful to characterize how quickly $\mathbb{P}(\widehat{K} = K_o)$ converges to 1 and to quantify the variability of $\widehat{K}$.

A high-impact direction is the application of ACDC to HMM state selection. Pohle et al. (2017) identify the failure of BIC/AIC for HMMs as a "notorious problem," demonstrating that these criteria consistently overestimate K to "mop up" any unmodeled structure (e.g., outliers, temporal variation, or non-Markovian dynamics). As they only offer a "pragmatic" solution based on heuristics and expert knowledge, ACDC presents an opportunity to develop a formal, statistical criterion that is robust to these misspecifications in HMMs. Another interesting direction for future work is to apply ACDC to other model classes such as supervised factor analysis and extend it to apply to models outside of the framework we developed in Section 3.2.3 – for example, (nonlinear) variational autoencoders (Kingma and Welling, 2014, 2019) and semiparametric matrix factorization models (Anandkumar et al., 2014b; Rohe and Zeng, 2023). This aligns with the emerging field of Causal Representation Learning (CRL), which seeks to learn interpretable causal factors from high-dimensional data (Moran and Aragam, 2025). ACDC could provide a crucial tool for robustly selecting the true number of latent causal factors in a VAE, a problem that Moran and Aragam (2025) identify as a key open statistical question.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19** 716-723.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. and Stratton, M. R. (2013). Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports* **3** 246.

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., Islam, S. M. A., Lopez-Bigas, N., Klimczak, L. J., McPherson, J. R., Morganella, S., Sabarinathan, R., Wheeler, D. A., Mustonen, V., PCAWG Mutational Signatures Working Group, Getz, G., Rozen, S. G., Stratton, M. R. and PCAWG Consortium (2020). The repertoire of mutational signatures in human cancer. *Nature* **578** 94–101.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M. and Telgarsky, M. (2014a). Tensor decompositions for learning latent variable models. *Journal of machine learning research* **15** 2773–2832.

Anandkumar, A., Ge, R., Hsu, D. and Kakade, S. M. (2014b). Tensor Decompositions for Learning Latent Variable Models. **15** 2773 – 2883.

Anderson, T. W. (1963). Asymptotic Theory for Principal Component Analysis. *The Annals of Mathematical Statistics* **34** 122–148.

Anderson, T. W. and Amemiya, Y. (1988). The Asymptotic Normal Distribution of Estimators in Factor Analysis under General Conditions. *The Annals of Statistics* **16** 759–771.

Bai, J. and Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica* **70** 191–221.

Barto, A. G. and Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete event dynamic systems* **13** 41–77.

Bauer, D. J. (2007). Observations on the Use of Growth Mixture Models in Psychological Research. *Multivariate Behavioral Research* **42** 757-786.

Bharti, A., Naslidnyk, M., Key, O., Kaski, S. and Briol, F.-X. (2023). Optimally-weighted Estimators of the Maximum Mean Discrepancy for Likelihood-Free Inference. In *Proceedings of the 40th International Conference on Machine Learning* (A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato and J. Scarlett, eds.). *Proceedings of Machine Learning Research* **202** 2289–2312. PMLR.

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics* **1** 17 – 35.

Brook, A. and Dor, E. B. (2016). Quantitative Detection of Settled Dust Over Green Canopy Using Sparse Unmixing of Airborne Hyperspectral Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **9** 884–897.

Brunet, J.-P., Tamayo, P., Golub, T. R. and Mesirov, J. P. (2004). Metagenes and Molecular Pattern Discovery Using Matrix Factorization. *Proceedings of the National Academy of Sciences* **101** 4164–4169.

Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C. and Stegle, O. (2017). F-scLVM: Scalable and Versatile Factor Analysis for Single-Cell RNA-seq. *Genome Biology* **18** 212.

Buja, A. and Eyuboglu, N. (1992). Remarks on Parallel Analysis. *Multivariate Behavioral Research* **27** 509–540.

Cai, D., Campbell, T. and Broderick, T. (2021). Finite mixture models do not reliably learn the number of components. In *Proceedings of the 38th International Conference on Machine Learning* (M. Meila and T. Zhang, eds.). *Proceedings of Machine Learning Research* **139** 1158–1169. PMLR.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q. and and, M. W. (2008). High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. *Journal of the American Statistical Association* **103** 1438–1456.

Chakraborty, A., Bhattacharya, A. and Pati, D. (2023). Robust probabilistic inference via a constrained transport metric. *arXiv*.

Chakravarty, D. and Solit, D. B. (2021). Clinical cancer genomic profiling. *Nature Reviews. Genetics* **22** 483–501.

Chiou, J. and Li, P. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69** 679–699.

Cichocki, A. and Phan, A.-H. (2009). Fast Local Algorithms for Large Scale Nonnegative Matrix and Tensor Factorizations. *IEICE Transactions* **92-A** 708–721.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley-Interscience, Hoboken, NJ.

Cunningham, J. P. and Yu, B. M. (2014). Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience* **17** 1500–1509.

Daniel, C., Van Hoof, H., Peters, J. and Neumann, G. (2016). Probabilistic inference for determining options in reinforcement learning. *Machine Learning* **104** 337–357.

de Souto, M. C. P., Costa, I. G., de Araujo, D. S. A., Ludermir, T. B. and Schliep, A. (2008). Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics* **9** 497 - 497.

Devarajan, K. (2019). Non-Negative Matrix Factorization Based on Generalized Dual Divergence.

Dewaskar, M., Tosh, C., Knoblauch, J. and Dunson, D. B. (2023). Robustifying likelihoods by optimistically re-weighting data. *arXiv*.

Dobriban, E. (2020). Permutation Methods for Factor Analysis and PCA. *The Annals of Statistics* **48**.

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62** 355–366.

Févotte, C. and Dobigeon, N. (2015). Nonlinear Hyperspectral Unmixing with Robust

Nonnegative Matrix Factorization. *IEEE Transactions on Image Processing* **24** 4810–4819.

Flynt, A., Dean, N. and Nugent, R. (2019). sARI: A Soft Agreement Measure for Class Partitions Incorporating Assignment Probabilities. *Advances in Data Analysis and Classification* **13** 303–323.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models. Springer Series in Statistics*. Springer New York.

Fu, X., Huang, K. and Sidiropoulos, N. D. (2018). On Identifiability of Nonnegative Matrix Factorization. *IEEE Signal Processing Letters* **25** 328–332.

Giammarino, V. and Paschalidis, I. C. (2021). Online Baum-Welch algorithm for Hierarchical Imitation Learning. In *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE.

Gibbs, A. L. and Su, F. E. (2002). On Choosing and Bounding Probability Metrics. *International Statistical Review* **70** 419–435.

Gorsky, S., Chan, C. and Ma, L. (2020). Coarsened mixtures of hierarchical skew normal kernels for flow cytometry analyses. *arXiv*.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. and Smola, A. (2012). A Kernel Two-Sample Test. *Journal of Machine Learning Research* **13** 723-773.

Guha, A., Ho, N. and Nguyen, X. (2021). On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli* **27** 2159 – 2188.

Horn, J. L. (1965). A Rationale and Test for the Number of Factors in Factor Analysis. *Psychometrika* **30** 179–185.

Hsu, D., Kakade, S. M. and Zhang, T. (2012). A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences* **78** 1460–1480.

Hsu, D. and Kakade, S. M. (2013). Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science* 11–20.

Islam, S. M. A., Díaz-Gay, M., Wu, Y., Barnes, M., Vangara, R., Bergstrom, E. N., He, Y., Vella, M., Wang, J., Teague, J. W., Clapham, P., Moody, S., Senkin, S., Li, Y. R., Riva, L., Zhang, T., Gruber, A. J., Steele, C. D., Otlu, B., Khandekar, A., Abbasi, A., Humphreys, L., Syulyukina, N., Brady, S. W., Alexandrov, B. S., Pillay, N., Zhang, J., Adams, D. J., Martincorena, I., Wedge, D. C., Landi, M. T., Brennan, P., Stratton, M. R., Rozen, S. G. and Alexandrov, L. B. (2022). Uncovering Novel Mutational Signatures by de Novo Extraction with SigProfilerExtractor. *Cell Genomics* **2**.

Ji, C., Jia, Y., Li, X. and Wang, J. (2016). Research on Linear and Nonlinear Spectral Mixture Models for Estimating Vegetation Fractional Cover of Nitraria Bushes. *National Remote Sensing Bulletin* **20** 1402–1412.

Kingma, D. P. and Welling, M. (2014). Auto-Encoding Variational Bayes. *International Conference on Learning Representations*.

Kingma, D. P. and Welling, M. (2019). An Introduction to Variational Autoencoders. *Foundations and Trends in Machine Learning* **12** 307–392.

Kinker, G. S., Greenwald, A. C., Tal, R., Orlova, Z., Cuoco, M. S., McFarland, J. M., Warren, A., Rodman, C., Roth, J. A., Bender, S. A., Kumar, B., Rocco, J. W., Fernandes, P. A., Mader, C. C., Keren-Shaul, H., Plotnikov, A., Barr, H., Tsherniak, A., Rozenblatt-Rosen, O., Krizhanovsky, V.,

Puram, S. V., Regev, A. and Tirosh, I. (2020). Pan-Cancer Single Cell RNA-seq Uncovers Recurring Programs of Cellular Heterogeneity. *Nature genetics* **52** 1208–1218.

Kotliar, D., Veres, A., Nagy, M. A., Tabrizi, S., Hodis, E., Melton, D. A. and Sabeti, P. C. (2019). Identifying Gene Expression Programs of Cell-Type Identity and Cellular Activity with Single-Cell RNA-Seq. *eLife* **8** e43803.

Lee, D. and Seung, H. S. (2000). Algorithms for Non-negative Matrix Factorization. In *Advances in Neural Information Processing Systems* (T. Leen, T. Dietterich and V. Tresp, eds.) **13**. MIT Press.

Levitin, H. M., Yuan, J., Cheng, Y. L., Ruiz, F. J., Bush, E. C., Bruce, J. N., Canoll, P., Iavarone, A., Lasorella, A., Blei, D. M. and Sims, P. A. (2019). De Novo Gene Signature Identification from Single-cell RNA-seq with Hierarchical Poisson Factorization. *Molecular Systems Biology* **15** e8557.

Li, Y., Roberts, N. D., Wala, J. A., Shapira, O., Schumacher, S. E., Kumar, K., Khurana, E., Waszak, S., Korbel, J. O., Haber, J. E., Imielinski, M., PCAWG Structural Variation Working Group, Weischenfeldt, J., Beroukhim, R., Campbell, P. J. and PCAWG Consortium (2020). Patterns of somatic structural variation in human cancer genomes. *Nature* **578** 112–121.

Lin, H. and Zhang, X. (2017). Retrieving the Hydrous Minerals on Mars by Sparse Unmixing and the Hapke Model Using MRO/CRISM Data. *Icarus* **288** 160–171.

Liu, Z. (2019). Model Selection for Nonnegative Matrix Factorization by Support Union Recovery. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 3407–3411.

Mattila, R., Rojas, C. R. and Wahlberg, B. (2015). Evaluation of Spectral Learning for the Identification of Hidden M arkov Models. *IFAC-PapersOnLine* **48** 897-902. 17th IFAC Symposium on System Identification SYSID 2015.

Mattila, R., Rojas, C. R., Krishnamurthy, V. and Wahlberg, B. (2017). Identification of hidden Markov models using spectral learning with likelihood maximization. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)* 5859-5864.

Mattila, R., Rojas, C., Moulines, E., Krishnamurthy, V. and Wahlberg, B. (2020). Fast and Consistent Learning of Hidden Markov Models by Incorporating Non-Consecutive Correlations. In *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. Singh, eds.). *Proceedings of Machine Learning Research* **119** 6785–6796. PMLR.

Miller, J. W. and Dunson, D. B. (2019). Robust Bayesian Inference via Coarsening. *Journal of the American Statistical Association* **114** 1113-1125.

[author] Moran, Gemma E.G. E. Aragam, BryonB. (2025). Towards Interpretable Deep Generative Models via Causal Representation Learning.

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L. A., Menzies, A., Martin, S., Leung, K., Chen, L., Leroy, C., Ramakrishna, M., Rance, R., Lau, K. W., Mudie, L. J., Varela, I., McBride, D. J., Bignell, G. R., Cooke, S. L., Shlien, A., Gamble, J., Whitmore, I., Maddison, M., Tarpey, P. S., Davies, H. R., Papaemmanuil, E., Stephens, P. J., McLaren, S., Butler, A. P., Teague, J. W., Jönsson, G., Garber, J. E., Silver, D., Miron, P., Fatima, A., Boyault, S., Langerød, A., Tutt, A., Martens, J. W. M., Aparicio, S. A.

J. R., Borg, Å., Salomon, A. V., Thomas, G., Børresen-Dale, A.-L., Richardson, A. L., Neuberger, M. S., Futreal, P. A., Campbell, P. J., Stratton, M. R. and Breast Cancer Working Group of the International Cancer Genome Consortium (2012). Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell* **149** 979–993.

Parikh, A. P., Song, L., Ishteva, M., Teodoru, G. and Xing, E. P. (2012). A spectral algorithm for latent junction trees. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence* 675–684.

Pelizzola, M., Laursen, R. and Hobolth, A. (2023). Model Selection and Robust Inference of Mutational Signatures Using Negative Binomial Non-Negative Matrix Factorization. *BMC Bioinformatics* **24** 187.

[author] Pohle, JenniferJ., Langrock, RolandR., family=Beest, prefix=van useprefix=falsep. u. given=Floris Schmidt, Niels MartinN. M. (2017). Selecting the Number of States in Hidden Markov Models - Pitfalls, Practical Challenges and Pragmatic Solutions.

Prabhakaran, S., Azizi, E., Carr, A. and Pe'er, D. (2016). Dirichlet Process Mixture Model for Correcting Technical Variation in Single-Cell Gene Expression Data. In *Proceedings of The 33rd International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.). *Proceedings of Machine Learning Research* **48** 1070–1079. PMLR, New York, New York, USA.

Rajabi, R. and Ghassemian, H. (2015). Spectral Unmixing of Hyperspectral Imagery Using Multilayer NMF. *IEEE Geoscience and Remote Sensing Letters* **12** 38–42.

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. and Vert, J.-P. (2018). A General and Flexible Method for Signal Extraction from Single-Cell RNA-seq Data. *Nature Communications* **9** 284.

Rohe, K. and Zeng, M. (2023). Vintage factor analysis with Varimax performs statistical inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **85** 1037–1060.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics* **6** 461–464.

Seplyarskiy, V. B., Soldatov, R. A., Koch, E., McGinty, R. J., Goldmann, J. M., Hernandez, R. D., Barnes, K., Correa, A., Burchard, E. G., Ellinor, P. T., McGarvey, S. T., Mitchell, B. D., Vasan, R. S., Redline, S., Silverman, E., Weiss, S. T., Arnett, D. K., Blangero, J., Boerwinkle, E., He, J., Montgomery, C., Rao, D. C., Rotter, J. I., Taylor, K. D., Brody, J. A., Chen, Y.-D. I., de Las Fuentes, L., Hwu, C.-M., Rich, S. S., Manichaikul, A. W., Mychaleckyj, J. C., Palmer, N. D., Smith, J. A., Kardia, S. L. R., Peyser, P. A., Bielak, L. F., O'Connor, T. D., Emery, L. S., Gilissen, C., Wong, W. S. W., Kharchenko, P. V. and Sunyaev, S. (2021). Population Sequencing Data Reveal a Compendium of Mutational Processes in Human Germline. *Science (New York, N.Y.)* **373** 1030–1035.

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science* **25** 289 – 310.

Simon-Gabriel, C.-J. and Schölkopf, B. (2018). Kernel Distribution Embeddings - Universal Kernels, Characteristic Kernels and Kernel Metrics on Distributions. *Journal of Machine Learning Research* **19** 1 – 29.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002).

Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64** 583-639.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B. and Lanckriet, G. R. G. (2010). Hilbert Space Embeddings and Metrics on Probability Measures. *Journal of Machine Learning Research* **11** 1517 – 1561.

Stevens, E., Dixon, D. R., Novack, M. N., Granpeesheh, D., Smith, T. and Linstead, E. (2019). Identification and analysis of behavioral phenotypes in autism spectrum disorder via unsupervised machine learning. *International Journal of Medical Informatics* **129** 29-36.

Sutton, R. S., Precup, D. and Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence* **112** 181-211.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics. Springer Series in Statistics*. Springer.

Villani, C. (2009). *Optimal transport: old and new* **338**. Springer.

West, M. (2003). Bayesian Factor Regression Models in the "Large p, Small n" Paradigm . *Bayesian Statistics* **7**.

Wu, B., Weinstein, E. N., Salehi, S., Wang, Y. and Blei, D. M. (2024). Adaptive Nonparametric Perturbations of Parametric Bayesian Models. *arXiv*.

Xue, C., Miller, J. W., Carter, S. L. and Huggins, J. H. (2024). Robust discovery of mutational signatures using power posteriors. *bioRxiv* 2024.10.23.619958.

Zhang, Z. and Paschalidis, I. (2021). Provable hierarchical imitation learning via em. In *International Conference on Artificial Intelligence and Statistics* 883–891. PMLR.

Zhang, Y., Chen, X., Zhou, D. and Jordan, M. I. (2016). Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *The Journal of Machine Learning Research* **17** 3537–3580.

Zhao, R. and Tan, V. Y. F. (2017). A Unified Convergence Analysis of the Multiplicative Update Algorithm for Regularized Nonnegative Matrix Factorization.

Supplementary Materials

Appendix A: Proofs for Method of moments for HIL

A.1. Lemma 1

Statement: let $\boldsymbol{V}_s \in \mathbb{R}^{\alpha \times \omega}$ for $s \in \mathcal{S}$ be a matrix of right singular vectors corresponding to the ω largest singular values of $\boldsymbol{K}_s$. We then have the following lemma. Define the block-diagonal matrices

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V}_1 & & \\ & \ddots & \\ & & \boldsymbol{V}_\zeta \end{bmatrix} \text{ and } \boldsymbol{\Pi}^{lo} = \begin{bmatrix} \boldsymbol{\Pi}_1^{lo} & & \\ & \ddots & \\ & & \boldsymbol{\Pi}_\zeta^{lo} \end{bmatrix}.$$

Then the product matrix

$$\boldsymbol{\Pi}^{lo}\boldsymbol{V} = \begin{bmatrix} \boldsymbol{\Pi}_1^{lo}\boldsymbol{V}_1 & & \\ & \ddots & \\ & & \boldsymbol{\Pi}_\zeta^{lo}\boldsymbol{V}_\zeta \end{bmatrix}$$

is invertible.

Proof. We have

$$\begin{aligned} \boldsymbol{K}_s[s_1, a_2] &= \sum_{o_1}\sum_{o_2} P(O_1 = o_1, S_1 = s_1, S_2 = s) \\ &\quad \times \pi_{hi}(O_2 = o_2 | O_1 = o_1, S_2 = s) \\ &\quad \times \pi_{lo}(A_2 = a_2 | O_2 = o_2, S_2 = s) \\ &= \left\{\boldsymbol{\Xi}_s \boldsymbol{\Pi}_s^{hi} \boldsymbol{\Pi}_s^{lo}\right\}[s_1, a_2]. \end{aligned} \tag{A.1}$$

This implies $\text{rowspan}(\boldsymbol{K}_s) \subseteq \text{rowspan}(\boldsymbol{\Pi}_s^{lo})$. In addition, because $\boldsymbol{\Xi}_s$ is full column rank and $\boldsymbol{\Pi}_s^{hi}$ is full rank (Assumptions 2 and 3),

$$\boldsymbol{\Pi}_s^{lo} = (\boldsymbol{\Xi}_s \boldsymbol{\Pi}_s^{hi})^+ \boldsymbol{K}_s,$$

which implies $\text{rowspan}(\boldsymbol{\Pi}_s^{lo}) \subseteq \text{rowspan}(\boldsymbol{K}_s)$. Thus,

$$\text{rowspan}(\boldsymbol{V}_s^{\mathrm{T}}) = \text{rowspan}(\boldsymbol{K}_s) = \text{rowspan}(\boldsymbol{\Pi}_s^{lo}).$$

Therefore, $\boldsymbol{\Pi}_s^{lo}\boldsymbol{V}_s$ is invertible. Since s is chosen arbitrarily, this applies for all s. Because $\boldsymbol{\Pi}_s^{lo}\boldsymbol{V}_s$ is invertible for all s, it follows that $\boldsymbol{\Pi}^{lo}\boldsymbol{V}$ is also invertible. □

A.2. Theorem 1

Statement: The product $\boldsymbol{V}^{\mathrm{T}}\hat{\boldsymbol{M}}^{+}\hat{\boldsymbol{M}}_a\boldsymbol{V}$ admits the factorization:

$$\boldsymbol{V}^{\mathrm{T}}\hat{\boldsymbol{M}}^{+}\hat{\boldsymbol{M}}_a\boldsymbol{V} = \boldsymbol{B}^{-1}\boldsymbol{\Lambda}_a\boldsymbol{B}, \tag{A.2}$$

where

$$\boldsymbol{\Lambda}_a = \begin{bmatrix} \mathrm{diag}(\boldsymbol{\Pi}_1^{lo}\mathbf{e}_a) & & \\ & \ddots & \\ & & \mathrm{diag}(\boldsymbol{\Pi}_\zeta^{lo}\mathbf{e}_a) \end{bmatrix},$$

and

$$\boldsymbol{B} = (\boldsymbol{\Psi} \otimes \mathbf{I}_\omega)\boldsymbol{\Pi}^{hi}\boldsymbol{\Pi}^{lo}\boldsymbol{V}.$$

Proof. Let

$$\boldsymbol{\Xi} = \begin{bmatrix} \boldsymbol{\Xi}_1 & & \\ & \ddots & \\ & & \boldsymbol{\Xi}_\zeta \end{bmatrix} \text{ and } \boldsymbol{\Pi}^{hi} = \begin{bmatrix} \boldsymbol{\Pi}_1^{hi} & & \\ & \ddots & \\ & & \boldsymbol{\Pi}_\zeta^{hi} \end{bmatrix}.$$

Then,

$$\begin{aligned} \boldsymbol{M}_a\left[s_2\zeta + s_1, s_3\alpha + a_3\right] = & \sum_{s_2',o_1}\sum_{s_2'',o_2}\sum_{s_2''',o_2'}\sum_{s_3,o_2'}\sum_{s_3',o_2''} \\ & P(O_1 = o_1, S_1 = s_1, S_2 = s_2 = s_2') \\ & \times \pi_{hi}(O_2 = o_2, S_2 = s_2''|O_1 = o_1, S_2 = s_2') \\ & \times \pi_{lo}(A_2 = a, S_2 = s_2''', O_2 = o_2'|O_2 = o_2, S_2 = s_2'') \\ & \times P(S_3 = s_3, O_2 = o_2''|A_2 = a, S_2 = s_2''', O_2 = o_2') \\ & \times \pi_{hi}(O_3 = o_3, S_3 = s_3'|O_2 = o_2'', S_3 = s_3) \\ & \times \pi_{lo}(A_3 = a_3, S_3 = s_3''|O_3 = o_3, S_3 = s_3') \\ = & \left\{\boldsymbol{\Xi}\boldsymbol{\Pi}^{hi}\boldsymbol{\Lambda}_a(\boldsymbol{\Phi}_a^A \otimes \mathbf{I}_\omega)\boldsymbol{\Pi}^{hi}\boldsymbol{\Pi}^{lo}\right\}\left[s_2\zeta + s_1, s_3\alpha + a_3\right]. \end{aligned}$$

Consider the $\zeta \times \alpha$ submatrix

$$\begin{aligned} \boldsymbol{U}_{s_2s_3}^a[s_1, a_3] &= \boldsymbol{M}_a[s_2\zeta + s_1, s_3\alpha + a_3] \\ \Leftrightarrow \boldsymbol{U}_{s_2s_3}^a &= \boldsymbol{\Xi}_{s_2}\boldsymbol{\Pi}_{s_2}^{hi}\mathrm{diag}(\boldsymbol{\Pi}_{s_2}^{lo}\mathbf{e}_a)\boldsymbol{\Phi}_a^A[s_2, s_3]\boldsymbol{\Pi}_{s_3}^{hi}\boldsymbol{\Pi}_{s_3}^{lo}. \end{aligned}$$

Notice that $\boldsymbol{\Phi}_a^A[s_2, s_3]$ is a real number that can be estimated using observable data. We define

$$\begin{aligned} \hat{\boldsymbol{U}}_{s_2s_3}^a &= \frac{1}{\boldsymbol{\Phi}_a^A[s_2, s_3]}\boldsymbol{U}_{s_2s_3}^a \\ &= \boldsymbol{\Xi}_{s_2}\boldsymbol{\Pi}_{s_2}^{hi}\mathrm{diag}(\boldsymbol{\Pi}_{s_2}^{lo}\mathbf{e}_a)\boldsymbol{\Pi}_{s_3}^{hi}\boldsymbol{\Pi}_{s_3}^{lo} \end{aligned} \tag{A.3}$$

for all $s_2, s_3 \in \mathcal{S}$ such that $\mathbf{\Phi}_a^A[s_2, s_3] > 0, \forall a \in \mathcal{A}$, and $\hat{\boldsymbol{U}}^a_{s_2 s_3} = \mathbf{0}_{\zeta \times \alpha}$ otherwise.

By the Definitions (2.2, A.3)

$$\begin{aligned}\hat{\boldsymbol{M}}_a &= (\mathbf{\Psi} \otimes \mathbf{1}_{\zeta \times \alpha}) \circ \begin{bmatrix} \hat{\boldsymbol{U}}^a_{11} & \dots & \hat{\boldsymbol{U}}^a_{1\zeta} \\ \vdots & \ddots & \vdots \\ \hat{\boldsymbol{U}}^a_{\zeta 1} & \dots & \hat{\boldsymbol{U}}^a_{\zeta\zeta} \end{bmatrix} \\ &= \mathbf{\Xi}\mathbf{\Pi}^{hi}\mathbf{\Lambda}_a(\mathbf{\Psi} \otimes \mathbf{I}_\omega)\mathbf{\Pi}^{hi}\mathbf{\Pi}^{lo}.\end{aligned}$$

By the Definition (2.3)

$$\begin{aligned}\hat{\boldsymbol{M}} &= \sum_{a \in \mathcal{A}} \hat{\boldsymbol{M}}_a \\ &= \mathbf{\Xi}\mathbf{\Pi}^{hi}\left(\sum_{a \in \mathcal{A}} \mathbf{\Lambda}_a\right)(\mathbf{\Psi} \otimes \mathbf{I}_\omega)\mathbf{\Pi}^{hi}\mathbf{\Pi}^{lo} \\ &= \mathbf{\Xi}\mathbf{\Pi}^{hi}(\mathbf{\Psi} \otimes \mathbf{I}_\omega)\mathbf{\Pi}^{hi}\mathbf{\Pi}^{lo}.\end{aligned}$$

Finally, we can write Equation (A.2) as

$$\boldsymbol{V}^{\mathrm{T}}\hat{\boldsymbol{M}}^{+}\hat{\boldsymbol{M}}_a\boldsymbol{V} = \left(\mathbf{\Pi}^{hi}\mathbf{\Pi}^{lo}\boldsymbol{V}\right)^{-1}\left(\mathbf{\Psi}^{-1} \otimes \mathbf{I}_\omega\right)\mathbf{\Lambda}_a(\mathbf{\Psi} \otimes \mathbf{I}_\omega)\mathbf{\Pi}^{hi}\mathbf{\Pi}^{lo}\boldsymbol{V}, \tag{A.4}$$

and the proof is complete. □

A.3. Theorem 2

Statement:

$$\hat{\boldsymbol{\mathcal{P}}}\mathbf{\Pi}^{hi}_{s'}\hat{\boldsymbol{\mathcal{P}}}^{\mathrm{T}} = \sum_s \boldsymbol{w}_{s'}[s]\left(\hat{\boldsymbol{\mathcal{P}}}\mathbf{\Pi}^{lo}_s\boldsymbol{K}_s^{+}\hat{\boldsymbol{K}}_{ss'}\mathbf{\Pi}^{lo+}_{s'}\hat{\boldsymbol{\mathcal{P}}}^{\mathrm{T}}\right),$$

where:

- $\hat{\boldsymbol{K}}_{ss'}$ is a $\zeta \times \alpha$ submatrix of $\hat{\boldsymbol{M}}$ defined by

$$\hat{\boldsymbol{K}}_{ss'}[s'', a] = \hat{\boldsymbol{M}}[s\zeta + s'', s'\alpha + a]. \tag{A.5}$$

- $\boldsymbol{w}_{s'}$ are length ζ weight vectors of choice subject to

$$\boldsymbol{w}_i^{\mathrm{T}}\mathbf{\Psi}\mathbf{e}_i^{\mathrm{T}} = 1, \ \forall i \in \mathcal{S}.$$

Proof. According to Definition (A.5) and Equation (A.1), $\boldsymbol{K}_s$ and $\hat{\boldsymbol{K}}_{ss'}$ can be written as following:

$$\boldsymbol{K}_s = \mathbf{\Xi}_s\mathbf{\Pi}^{hi}_s\mathbf{\Pi}^{lo}_s,$$

$$\hat{\boldsymbol{K}}_{ss'} = \mathbf{\Psi}[s, s']\left(\mathbf{\Xi}_s\mathbf{\Pi}^{hi}_s\mathbf{\Pi}^{hi}_{s'}\mathbf{\Pi}^{lo}_{s'}\right).$$

Therefore,

$$\begin{aligned}\boldsymbol{\Pi}_s^{lo}\boldsymbol{K}_s{}^{+}\hat{\boldsymbol{K}}_{ss'}\boldsymbol{\Pi}_{s'}^{lo+} &= \boldsymbol{\Psi}[s,s']\left(\boldsymbol{\Pi}_s^{lo}\boldsymbol{\Pi}_s^{lo+}\boldsymbol{\Pi}_s^{hi-1}\boldsymbol{\Xi}_s{}^{+}\boldsymbol{\Xi}_s\boldsymbol{\Pi}_s^{hi}\boldsymbol{\Pi}_{s'}^{hi}\boldsymbol{\Pi}_{s'}^{lo}\boldsymbol{\Pi}_{s'}^{lo+}\right)\\ &= \boldsymbol{\Psi}[s,s']\boldsymbol{\Pi}_{s'}^{hi}.\end{aligned}$$

With that, we contract the left hand side of Equation (2.6)

$$\begin{aligned}&\sum_s \boldsymbol{w}_{s'}[s]\left(\hat{\boldsymbol{\mathcal{P}}}\boldsymbol{\Pi}_s^{lo}\boldsymbol{K}_s^{+}\hat{\boldsymbol{K}}_{ss'}\boldsymbol{\Pi}_{s'}^{lo+}\hat{\boldsymbol{\mathcal{P}}}^{\mathrm{T}}\right)\\ =&\sum_s \boldsymbol{w}_{s'}[s]\boldsymbol{\Psi}[s,s']\left(\hat{\boldsymbol{\mathcal{P}}}\boldsymbol{\Pi}_{s'}^{hi}\hat{\boldsymbol{\mathcal{P}}}^{\mathrm{T}}\right)\\ =&\hat{\boldsymbol{\mathcal{P}}}\boldsymbol{\Pi}_{s'}^{hi}\hat{\boldsymbol{\mathcal{P}}}^{\mathrm{T}}.\end{aligned}$$

The last equality holds because we chose $\boldsymbol{w}_{s'}$ such that $\sum_s \boldsymbol{w}_{s'}[s]\boldsymbol{\Psi}[s,s'] = 1$. Analysis of the choice of $\boldsymbol{w}_{s'}$ will be reserved for future work. □

Appendix B: Order recovery process for method of moments

From Equation (A.4), we know the eigenbasis $\boldsymbol{B} \in \mathbb{R}^{\zeta\omega\times\zeta\omega}$ is of the form

$$\boldsymbol{B} = \left(\boldsymbol{\Pi}^{hi}\boldsymbol{\Pi}^{lo}\boldsymbol{V}\right)^{-1}\left(\boldsymbol{\Psi}^{-1}\otimes\mathbf{I}_\omega\right). \tag{B.1}$$

The eigen-decomposition (2.4) will introduce an unknown scaling and permutation to the columns of the basis:

$$\hat{\boldsymbol{B}} = \boldsymbol{B}\mathrm{diag}(\boldsymbol{c})\boldsymbol{\mathcal{P}}, \tag{B.2}$$

where $\boldsymbol{c}$ is a vector that corresponds to the scaling of each column, and $\boldsymbol{\mathcal{P}}$ is a permutation operator on the columns.

For convenience, define the following shorthands:

$$\boldsymbol{X}_s = \left(\boldsymbol{\Pi}_s^{hi}\boldsymbol{\Pi}_s^{lo}\boldsymbol{V}_s\right)^{-1},$$

$$\boldsymbol{X} = \left(\boldsymbol{\Pi}^{hi}\boldsymbol{\Pi}^{lo}\boldsymbol{V}\right)^{-1} = \mathrm{diag}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_\zeta),$$

and $\boldsymbol{\Gamma} = \boldsymbol{\Psi}^{-1}$ with elements γ_{ij}. Rewrite (B.1) as $\boldsymbol{B} = \boldsymbol{X}\left(\boldsymbol{\Gamma}\otimes\mathbf{I}_\omega\right)$. Recalling Equation (B.2), we have that the structure of the sub-matrices $\hat{\boldsymbol{J}}_s \in \mathbb{R}^{\omega\times\zeta\omega}$ of the basis $\hat{\boldsymbol{B}}$ is $\hat{\boldsymbol{J}}_s = \begin{bmatrix}\gamma_{s1}\boldsymbol{X}_s & \ldots & \gamma_{s\zeta}\boldsymbol{X}s\end{bmatrix}\mathrm{diag}(\boldsymbol{c})\boldsymbol{\mathcal{P}}$. It is easily seen that each of these sub-matrices contains ω sets of ζ linearly dependent vectors, and that the grouping of these linearly dependent columns are identical due to them sharing the same permutation $\boldsymbol{\mathcal{P}}$. We separate and represent these sets as $\zeta\omega\times\zeta$ matrices $\hat{\boldsymbol{Q}}_o$ given by

$$\hat{\boldsymbol{Q}}_o = \begin{bmatrix}\gamma_{11}\boldsymbol{X}_1\mathbf{e}_o & \ldots & \gamma_{1\zeta}\boldsymbol{X}_1\mathbf{e}_o\\ \vdots & \ddots & \vdots\\ \gamma_{\zeta 1}\boldsymbol{X}_\zeta\mathbf{e}_o & \ldots & \gamma_{\zeta\zeta}\boldsymbol{X}_\zeta\mathbf{e}_o\end{bmatrix}\mathrm{diag}(\boldsymbol{c}_o)\boldsymbol{\mathcal{P}}_o, \tag{B.3}$$

where $\boldsymbol{c}_o$ and $\boldsymbol{\mathcal{P}}_o$ are unknown scaling factor and permutation corresponding to the group o.

We define $\boldsymbol{d}_o = \begin{bmatrix}\boldsymbol{X}_1^{\mathrm{T}}\mathbf{e}_o^{\mathrm{T}} & \dots & \boldsymbol{X}_\zeta^{\mathrm{T}}\mathbf{e}_o^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{\zeta\omega}$. Then, (B.3) can be rewritten as $\hat{\boldsymbol{Q}}_o = \mathrm{diag}(\boldsymbol{d}_o)(\boldsymbol{\Gamma}\otimes\mathbf{1}_\omega)\mathrm{diag}(\boldsymbol{c}_o)\boldsymbol{\mathcal{P}}_o$. In order to recover the original ordering of the columns of $\hat{\boldsymbol{Q}}_o$, we need to somehow match them with the columns of $\boldsymbol{\Gamma}\otimes\mathbf{1}_\omega$, which is a known quantity.

Before proceeding, let us introduce the element-wise inverse operator $\oslash$ such that $\boldsymbol{A}^{\oslash}$ corresponds to the matrix formed by inverting each element of the matrix $\boldsymbol{A}$, and so that $(\boldsymbol{A}\boldsymbol{\mathcal{P}})^{\oslash} = \boldsymbol{A}^{\oslash}\boldsymbol{\mathcal{P}}$, and $[\mathrm{diag}(\boldsymbol{u})\boldsymbol{A}\mathrm{diag}(\boldsymbol{v})]^{\oslash} = \mathrm{diag}(\boldsymbol{u})^{-1}\boldsymbol{A}^{\oslash}\mathrm{diag}(\boldsymbol{v})^{-1}$. Let's assume that $\boldsymbol{\Gamma}\otimes\mathbf{1}_\omega$ has no zero entries. If there are zero entries, we can further partition $\hat{\boldsymbol{Q}}_o$ and $\boldsymbol{\Gamma}\otimes\mathbf{1}_\omega$ into column groups that has the same rows with zero entries, remove those rows, and use the following procedure on each of the groups before combining the result.

We have the following reduction:

$$\begin{aligned}\hat{\boldsymbol{Q}}_o^{\oslash}\hat{\boldsymbol{Q}}_o^{\mathrm{T}} &= \mathrm{diag}(\boldsymbol{d}_o)^{-1}(\boldsymbol{\Gamma}^{\oslash}\boldsymbol{\Gamma}^{\mathrm{T}}\otimes\mathbf{1}_\omega\mathbf{1}_\omega^{\mathrm{T}})\mathrm{diag}(\boldsymbol{d}_o)\\ &= (\boldsymbol{d}_o^{\oslash}\boldsymbol{d}_o^{\mathrm{T}})\circ(\boldsymbol{\Gamma}^{\oslash}\boldsymbol{\Gamma}^{\mathrm{T}}\otimes\mathbf{1}_\omega\mathbf{1}_\omega^{\mathrm{T}}).\end{aligned}$$

Therefore $\hat{\boldsymbol{Q}}_o^{\oslash}\hat{\boldsymbol{Q}}_o^{\mathrm{T}}\circ(\boldsymbol{\Gamma}^{\oslash}\boldsymbol{\Gamma}^{\mathrm{T}}\otimes\mathbf{1}_\omega\mathbf{1}_\omega^{\mathrm{T}})^{\oslash} = \boldsymbol{d}_o^{\oslash}\boldsymbol{d}_o^{\mathrm{T}}$. It is easily verifiable that $\mathrm{diag}(\boldsymbol{d}_o^{\oslash}\boldsymbol{d}_o^{\mathrm{T}}\mathbf{e}_i)\mathrm{diag}(\boldsymbol{d}_o) = \boldsymbol{d}_o[i]\mathbf{I}$.

We have $\mathrm{diag}(\boldsymbol{d}_o^{\oslash}\boldsymbol{d}_o^{\mathrm{T}}\mathbf{1}_{\zeta\omega})\hat{\boldsymbol{Q}}_o = (\mathbf{1}_{\zeta\omega}^{\mathrm{T}}\boldsymbol{d}_o)(\boldsymbol{\Gamma}\otimes\mathbf{1}_\omega)\mathrm{diag}(\boldsymbol{c}_o)\boldsymbol{\mathcal{P}}_o$. Notice that the columns of this matrix are just multiples of the columns of $(\boldsymbol{\Gamma}\otimes\mathbf{1}_\omega)$. As such a matching can be computed by checking linear dependence between the columns of the two matrices.

Appendix C: Robust Consistency of ACDC

We now show that, under reasonable assumptions, ACDC is robustly consistent for probabilistic matrix factorization. Due to the complex dependence structures in probabilistic matrix factorization, we limit our result when $\mathcal{D}$ is the KL divergence, as stated informally in Theorem 3.

C.1. Theory for Probabilistic Matrix Factorization

We will develop a general theory, then discuss how Theorem 3 follows as a corollary. First, we consider the requirements for the distribution-level discrepancy $\mathcal{D}_{\mathrm{dist}}(\cdot,\cdot)$ (assumed to be a metric), the discrepancy $\mathcal{D}(\cdot \mid \cdot)$ used to construct the component-level discrepancy, and the discrepancy estimator $\widehat{\mathcal{D}}(\cdot \mid \cdot)$ – noting that sometimes it will be possible to take $\widehat{\mathcal{D}}(\cdot \mid \cdot) = \mathcal{D}(\cdot \mid \cdot)$.

Assumption C.1. *For $y_{Nn} \in \mathcal{X}$ $(N = 1, 2, \dots; n = 1, \dots, N)$, define the empirical distribution $\widehat{P}_N = N^{-1}\sum_{n=1}^{N}\delta_{y_{N,n}}$ and assume $\widehat{P}_N \to P$ in distribution. The distribution-level and component-level discrepancies satisfy the following conditions:*

(a) The distribution-level discrepancy metric detects empirical convergence: $\mathcal{D}_{\mathrm{dist}}(\widehat{P}_N, P) \to 0$ *as* $N \to \infty$.

(b) The distribution-level discrepancy metric is jointly convex in its arguments: *for all* $w \in (0,1)$ *and distributions* P, P', Q, Q',

$$\mathcal{D}_{\mathrm{dist}}(wP + (1-w)P', wQ + (1-w)Q') \le w\,\mathcal{D}_{\mathrm{dist}}(P,Q) + (1-w)\,\mathcal{D}_{\mathrm{dist}}(P',Q').$$

(c) The discrepancy estimator is consistent: *For any distributions P, Q, if $\mathcal{D}(P \mid Q) < \infty$ and $\widehat{P}_N \to P$ in distribution, then $\widehat{\mathcal{D}}(\widehat{P}_N \mid Q) \to \mathcal{D}(P \mid Q)$ as $N \to \infty$.*
(d) Smoothness of the discrepancy estimator: *The map $\phi \mapsto \widehat{\mathcal{D}}(\widehat{P}_N \mid F_\phi)$ is continuous.*
(e) The discrepancy bounds the metric: *There exists a continuous, non-decreasing function $\widetilde{\kappa} : \mathbb{R} \to \mathbb{R}$ such that $\mathcal{D}_{\text{dist}}(P, Q) \leq \widetilde{\kappa}(\mathcal{D}(P \mid Q))$ for all distributions P, Q.*
(f) The distribution-level discrepancy metric between components is finite: *For all $\phi, \phi' \in \Theta$, it holds that $\mathcal{D}_{\text{dist}}(F_\phi, F_{\phi'}) < \infty$.*

Remark 3 (Discussion of Assumption C.1)**.** *A wide variety of metrics satisfy Assumption C.1(a), including the bounded Lipschitz metric, the Kolmogorov metric, maximum mean discrepancies with sufficiently regular bounded kernels, and the Wasserstein metric with a bounded cost function (Simon-Gabriel and Schölkopf, 2018; Sriperumbudur et al., 2010; van der Vaart and Wellner, 1996; Villani, 2009). Assumption C.1(b) also holds for a range of metrics. For example, it is easy to show that all integral probability metrics – which includes the bounded Lipschitz metric, maximum mean discrepancy, and 1-Wasserstein distance – are jointly convex. Assumption C.1(c) is a natural requirement that the limiting divergence can be estimated consistently. Such estimators are well-studied for many common discrepancies. Assumption C.1(d) will typically hold as long as the map $\phi \mapsto F_\phi$ is well-behaved. For example, for the Kullback–Leibler divergence estimators and standard maximum mean discrepancy estimators (Bharti et al., 2023; Gretton et al., 2012), when F_ϕ admits a density f_ϕ, it suffices for the map $\phi \mapsto f_\phi(x)$ to be continuous for P_o-almost every x. Assumption C.1(e) is not overly restrictive. Assumption C.1(f) trivially holds for bounded metrics such as the bounded Lipschitz metric and integral probability measures with uniformly bounded test functions.*

Our second assumption requires the parameter estimation procedure to be sufficiently well-behaved, in the sense that, for each fixed number of components $K \leq K_o$, it should consistently estimate an asymptotic parameter $\theta_\star^{(K)}$. We do not make any explicit assumptions that such parameters are, in any sense, "optimal."

For a given K, denote the empirical distribution of estimated coefficients by $\widehat{H}^{(K,N)} = N^{-1} \sum_{n=1}^{N} \delta_{\widehat{z}_n^{(K)}}$.

Assumption C.2. *For each $K \in \{1, \ldots, K_o\}$, there exists $\phi_\star^{(K)} \in \Phi^{(K)}$ and distribution $H_\star^{(K)}$ such that $\widehat{\phi}^{(K,N)} \xrightarrow{p} \phi_\star^{(K)}$ and $\widehat{H}^{(K,N)} \xrightarrow{d} H_\star^{(K)}$ as $N \to \infty$, after possible reordering of components.*

This should holds for most applications under mild conditions (Anderson, 1963; Anderson and Amemiya, 1988; Devarajan, 2019; Fu, Huang and Sidiropoulos, 2018; Zhao and Tan, 2017).

Let $\frac{\mathrm{d}y}{\mathrm{d}\varepsilon}(\cdot, \phi, z)$ be the Jacobian matrix of the mapping $\varepsilon \mapsto f(z, \phi, \varepsilon)$, and $\mathcal{E}(y, \phi, z) = \{\varepsilon \mid f(z, \phi, \varepsilon) = y)\}$. Let

$$Q(y_{nk} \mid \phi_k, z_{nk}, \widetilde{G}_k) = \int_{\mathcal{E}(y_{nk}, \phi_k, z_{nk})} \widetilde{G}_k(\varepsilon) \det \left(\frac{\mathrm{d}y}{\mathrm{d}\varepsilon}(\varepsilon, \phi_k, z_{nk}) \right)^{-1}$$

denote the conditional distribution of $y_{nk}^{(K)}$ given $z_{nk}^{(K)}$, $\phi_k^{(K)}$ and noise distribution $\widetilde{G}$, and let

$$Q^{(K)}\left(y_n \mid \phi, z_n, \widetilde{G}_{1:K}\right) = \prod_{k=1}^{K} Q(y_{nk} \mid \phi_k, z_{nk}, \widetilde{G}_k).$$

We can now define the conditional distributions for the limiting model as

$$Q_{\star k}^{(K)}(y_{nk} \mid z_{nk}) = Q\left(y_{nk} \mid \phi_{\star k}^{(K)}, z_{nk}, G\right) \quad \text{and} \quad Q_{\star}^{(K)}(y_n \mid z_n) = \prod_{k=1}^{K} Q_{\star k}^{(K)}(y_{nk} \mid z_{nk}),$$

and the empirical conditional distributions as

$$\widehat{Q}_k^{(K,N)}(y_n \mid z_n) = Q\left(y_n \mid \widehat{\phi}_k^{(K,N)}, z, \widehat{G}_k^{(K,N)}\right) \quad \text{and} \quad \widehat{Q}^{(K,N)}(y_n \mid z_n) = \prod_{k=1}^{K} \widehat{Q}_k^{(K,N)}(y_k \mid z_k).$$

Assumption C.3. *The data-generating distribution and model satisfy the following conditions:*

(a) For all $K \in \{1, \ldots, K_o\}$, $k \in \{1, \ldots, K\}$, and $z \in \mathcal{Z}$, the distributions $Q_{\star k}^{(K)}(\cdot \mid z)$ has the same support on $\mathcal{Y}$.

(b) For all $K \in \{1, \ldots, K_o\}$, $k \in \{1, \ldots, K\}$, $z \in \mathcal{Z}$, and distribution Q' on $\mathcal{Y}$,

$$\lim_{\phi \to \phi_{k\star}^{(K)}} \widehat{\mathcal{D}}(Q', Q(\cdot \mid \phi, z, G)) = \widehat{\mathcal{D}}(Q', Q_{\star k}^{(K)}(\cdot \mid z)).$$

(c) There exists a function $C : \mathcal{Y}^K \to \mathbb{R}$ such that for all $y_{1:K} \in \mathcal{Y}^K$, if $z_{1:K} \sim H_{\star}^{(K)}$, then

$$\operatorname{Var}\left(Q_{\star}^{(K)}\left(y_{1:K} \mid z_{1:K}\right)\right) \leq C(y_{1:K}) \cdot \mathbb{E}\left[Q_{\star}^{(K)}\left(y_{1:K} \mid z_{1:K}\right)\right],$$

and

$$\int_{\mathcal{X}_D^{\otimes K}} C(y_{1:K})\, dy_{1:K} < \infty.$$

Assumption C.3 ensures that the data generating distribution can be sufficiently approximated via empirical sampling. The first two parts are mild regularity conditions while Assumption C.3(c) is applicable to the PMF models used in Section 3.3 (see Appendix D.2 for details).

The following theorem and **??** immediately imply Theorem 3.

Theorem C.1. *For probabilistic matrix factorization, if Assumptions C.1 to C.3 hold with $\mathcal{D}(\cdot \mid \cdot) = \mathrm{KL}(\cdot \mid \cdot)$, then ACDC is κ-robustly consistent for $\kappa(\rho, K) = \widetilde{\kappa}(K\rho)$.*

See Appendix D for a proof.

Appendix D: Robust Consistency for Probabilistic Matrix Factorization

D.1. Proof of Theorem C.1

Notation. Let

$$Q^{(K)}\left(y_{1:K} \mid \phi_{1:K}, H, G_{1:K}\right) = \mathbb{E}_{z_{1:K}\sim H}\left[Q^{(K)}\left(y_{1:K} \mid \phi_{1:K}, z_{1:K}, G_{1:K}\right)\right]$$

denote the joint distribution of $y_{n,1:K}$ when $z^{(K)}_{n,1:K}$ is marginalized out, and let

$$P^{(K)}(x \mid \phi_{1:K}, H, G_{1:K}) = \int_{\sigma(x)} Q^{(K)}\left(y_{1:K} \mid \phi_{1:K}, H, G_{1:K}\right) \mathrm{d}y_{1:K}$$

be the marginal probability of drawing x_n, where $\sigma(x) = \{y_{1:K} \mid \sum_k y_k = x\}$. Similarly, we define the model distributions

$$Q^{(K)}_\star(y_{1:K}) = \mathbb{E}_{z_{1:K}\sim H^{(K)}_\star}\left[Q^{(K)}_\star\left(y_{1:K} \mid z_{1:K}\right)\right] \quad \text{and} \quad P^{(K)}_\star(x) = \int_{\sigma(x)} Q^{(K)}_\star(y_{1:K})\mathrm{d}y_{1:K},$$

the empirical distributions

$$\widehat{Q}^{(K,N)}(y_{1:K}) = \mathbb{E}_{z_{1:K}\sim \widehat{H}^{(K,N)}}\left[\widehat{Q}^{(K,N)}\left(y_{1:K} \mid z_{1:K}\right)\right] \quad \text{and} \quad \widehat{P}^{(K,N)}(x) = \int_{\sigma(x)} \widehat{Q}^{(K,N)}(y_{1:K})\mathrm{d}y_{1:K},$$

and the bridging distributions

$$\check{Q}^{(K,N)}(y_{1:K}) = \mathbb{E}_{z_{1:K}\sim \widehat{H}^{(K,N)}}\left[Q^{(K)}_\star\left(y_{1:K} \mid z_{1:K}\right)\right] \quad \text{and} \quad \check{P}^{(K,N)}(x) = \int_{\sigma(x)} \check{Q}^{(K,N)}(y_{1:K})\mathrm{d}y_{1:K},$$

Let

$$\mathcal{R}^\rho\left(\widehat{G}^{(K,N)}_{1:K}\right) = N\sum_{k=1}^{K} \max\left(0, \widehat{\mathcal{D}}\left(\widehat{G}^{(K,N)}_k, G\right) - \rho\right).$$

Approach. We show that (1) if $K = K_o$, then $\mathcal{R}^\rho(\widehat{G}^{(K,N)}_{1:K}) \to 0$ in probability, and (2) if $K < K_o$, then $\mathcal{R}^\rho(\widehat{G}^{(K,N)}_{1:K}) \to \infty$ in probability. The conclusion follows immediately from these two results.

Proof of part (1). If $K = K_o$, then it follows from Assumption C.1(c,d) and Assumption C.2 that $\widehat{\mathcal{D}}\left(\widehat{G}^{(K_o,N)}_k, G\right) \to \mathcal{D}\left(G^{(K_o)}_\star, G\right)$ in probability. Hence, it follows that there exists $\varepsilon > 0$ such that

$$\widehat{\mathcal{D}}\left(\widehat{G}^{(K_o,N)}_k, G\right) < \rho - \varepsilon + o_P(1).$$

Using this inequality, we have

$$\mathcal{R}^\rho\left(\widehat{G}^{(K,N)}_{1:K}\right) \le N\sum_{k=1}^{K} \max\left(0, -\varepsilon + o_P(1)\right).$$

Hence, we can conclude that $\lim_{N\to\infty} \mathbb{P}\left\{\mathcal{R}^\rho\left(\widehat{G}^{(K,N)}_{1:K}\right) = 0\right\} = 1$.

Proof of part (2). Consider the case of $K < K_o$. We have

$$\begin{aligned}
\mathcal{D}_{\text{dist}}(P_o, P_\star^{(K)}) &\leq \mathcal{D}_{\text{dist}}\left(P_o, \widehat{P}^{(K,N)}\right) + \mathcal{D}_{\text{dist}}\left(\widehat{P}^{(K,N)}, P_\star^{(K)}\right) \\
&\leq \mathcal{D}_{\text{dist}}\left(P_o, \widehat{P}^{(K,N)}\right) + \mathcal{D}_{\text{dist}}\left(\widehat{P}^{(K,N)}, \check{P}^{(K,N)}\right) + \mathcal{D}_{\text{dist}}\left(\check{P}^{(K,N)}, P_\star^{(K)}\right) \\
&= o_P(1) + \mathcal{D}_{\text{dist}}\left(\widehat{P}^{(K,N)}, \check{P}^{(K,N)}\right) + \mathcal{D}_{\text{dist}}\left(\check{P}^{(K,N)}, P_\star^{(K)}\right) \qquad \text{(D.1)} \\
&\leq \widetilde{\kappa}\left(\widehat{\mathcal{D}}\left(\widehat{P}^{(K,N)}, \check{P}^{(K,N)}\right) + o_P(1)\right) + \widetilde{\kappa}\left(\mathcal{D}\left(\check{P}^{(K,N)}, P_\star^{(K)}\right)\right) + o_P(1) \\
&\leq \widetilde{\kappa}\left(\widehat{\mathcal{D}}\left(\widehat{Q}^{(K,N)}, \check{Q}^{(K,N)}\right)\right) + \widetilde{\kappa}\left(\mathcal{D}\left(\check{Q}^{(K,N)}, Q_\star^{(K)}\right)\right) + o_P(1). \qquad \text{(D.2)}
\end{aligned}$$

where Eq. (D.1) follows from Assumption C.1(a). A bound on $\widehat{\mathcal{D}}\left(\widehat{Q}^{(K,N)}, \check{Q}^{(K,N)}\right)$ is given by

$$\begin{aligned}
&\widehat{\mathcal{D}}\left(\widehat{Q}^{(K,N)}, \check{Q}^{(K,N)}\right) \\
&= \widehat{\mathcal{D}}\left(\mathbb{E}_{z_{1:K}\sim\widehat{H}^{(K,N)}}\left[\prod_k \widehat{Q}_k^{(K,N)}(\cdot \mid z_k)\right], \mathbb{E}_{z_{1:K}\sim\widehat{H}^{(K,N)}}\left[\prod_k Q_{\star k}^{(K)}(\cdot \mid z_k)\right]\right) \\
&\leq \mathbb{E}_{z_{1:K}\sim\widehat{H}^{(K,N)}}\left[\widehat{\mathcal{D}}\left(\prod_k \widehat{Q}_k^{(K,N)}(\cdot \mid z_k), \prod_k Q_{\star k}^{(K)}(\cdot \mid z_k)\right)\right] + o_P(1) \qquad \text{(D.3)} \\
&= \sum_k \mathbb{E}_{z_{1:K}\sim\widehat{H}^{(K,N)}}\left[\widehat{\mathcal{D}}\left(\widehat{Q}_k^{(K,N)}(\cdot \mid z_k), Q_{\star k}^{(K)}(\cdot \mid z_k)\right)\right] + o_P(1) \\
&= \sum_k \mathbb{E}_{z_{1:K}\sim\widehat{H}^{(K,N)}}\left[\widehat{\mathcal{D}}\left(\widehat{Q}^{(K,N)}(\cdot \mid z_k), Q\left(\cdot \mid \phi_k^{(K,N)}, z_k, G\right)\right) + o_P(1)\right] + o_P(1) \qquad \text{(D.4)} \\
&\leq \sum_k \widehat{\mathcal{D}}\left(\widehat{G}_k^{(K,N)}, G\right) + o_P(1). \qquad \text{(D.5)}
\end{aligned}$$

where Eq. (D.3) follows from the fact that the KL divergence is convex with respect to both of its arguments, Eq. (D.4) follows from Assumption C.3(b), and Eq. (D.5) follows from the fact that KL divergence is invariant under diffeomorphism. With this we bound

$$\begin{aligned}
\widetilde{\kappa}\left(\mathcal{D}\left(\widehat{Q}^{(K,N)}, \check{Q}^{(K,N)}\right)\right) &\leq \widetilde{\kappa}\left(\sum_k \mathcal{D}\left(\widehat{G}_k^{(K,N)}, G\right) + o_P(1)\right) \\
&\leq \widetilde{\kappa}\left(\sum_k \mathcal{D}\left(\widehat{G}_k^{(K,N)}, G\right)\right) + o_P(1). \qquad \text{(D.6)}
\end{aligned}$$

Next we will bound $\mathcal{D}\left(\check{Q}^{(K,N)}, Q_\star^{(K)}\right)$. For the chi-squared distance $D_{\chi^2}(P, Q) = \int \frac{(P(x)-Q(x))^2}{Q(x)}\, dx$

(Cover and Thomas, 2006), we have that

$$
\begin{aligned}
&\mathbb{E}_{h_{1:N}}\left[D_{\chi^2}\left(\check{Q}^{(K,N)}, Q_\star^{(K)}\right)\right] \\
&= \mathbb{E}_{h_{1:N}}\left[\int_{\mathcal{X}_D^{\otimes K}} \frac{(\check{Q}^{(K,N)} - Q_\star^{(K)})^2}{Q_\star^{(K)}}\, dy_{1:K}\right] \\
&= \int_{\mathcal{X}_D^{\otimes K}} \mathbb{E}_{h_{1:N}}\left[\frac{(\check{Q}^{(K,N)} - Q_\star^{(K)})^2}{Q_\star^{(K)}}\right] dy_{1:K} \quad \text{[by bounded covergence]} \\
&= \int_{\mathcal{X}_D^{\otimes K}} \mathbb{E}_{h_{1:N}}\left[\frac{(\check{Q}^{(K,N)})^2}{Q_\star^{(K)}} - 2\check{Q}^{(K,N)} + Q_\star^{(K)}\right] dy_{1:K} \\
&= \int_{\mathcal{X}_D^{\otimes K}} \left[\mathbb{E}_{h_{1:N}}\left(\frac{(\check{Q}^{(K,N)})^2}{Q_\star^{(K)}}\right) - 2Q_\star^{(K)} + Q_\star^{(K)}\right] dy_{1:K} \quad [\text{ since } \mathbb{E}[\check{Q}^{(K,N)}] = Q_\star^{(K)}] \\
&= \int_{\mathcal{X}_D^{\otimes K}} \frac{\mathbb{E}_{h_{1:N}}\left((\check{Q}^{(K,N)})^2\right)}{Q_\star^{(K)}} - Q_\star^{(K)}\, dy_{1:K} \\
&= \int_{\mathcal{X}_D^{\otimes K}} \frac{\mathrm{Var}_{h_{1:N}}(\check{Q}^{(K,N)}) + \left(\mathbb{E}_{h_{1:N}}\check{Q}^{(K,N)}\right)^2}{Q_\star^{(K)}} - Q_\star^{(K)}\, dy_{1:K} \\
&= \int_{\mathcal{X}_D^{\otimes K}} \frac{\mathrm{Var}_{h_{1:N}}(\check{Q}^{(K,N)})}{Q_\star^{(K)}} dy_{1:K} \\
&= \frac{1}{N}\int_{\mathcal{X}_D^{\otimes K}} \frac{\mathrm{Var}_{z_{1:K}\sim H_\star^{(K)}}\left(Q_\star^{(K)}(y_{1:K} \mid z_{1:K})\right)}{Q_\star^{(K)}} dy_{1:K} \quad \text{[by variance of sample mean: } \frac{1}{N}\mathrm{Var}(X)] \\
&\leq \frac{1}{N}\int_{\mathcal{X}_D^{\otimes K}} \frac{C(y_{1:K}) \cdot \mathbb{E}_{z_{1:K}\sim H_\star^{(K)}}\left[Q_\star^{(K)}(y_{1:K} \mid z_{1:K})\right]}{Q_\star^{(K)}(y_{1:K})} dy_{1:K} \\
&\leq \frac{1}{N}\int_{\mathcal{X}_D^{\otimes K}} C(y_{1:K}) dy_{1:K}.
\end{aligned}
$$

Since $\mathrm{KL} \leq \log(1 + D_{\chi^2}) \leq D_{\chi^2}$ (Gibbs and Su, 2002), it follows that $\chi^2 \to 0 \Rightarrow \mathrm{KL} \to 0$. Together with Assumption C.3, it follows that

$$\lim_{N\to\infty} \mathbb{E}_{h_{1:N}}\left[\mathcal{D}_{KL}\left(\check{Q}^{(K,N)}, Q_\star^{(K)}\right)\right] = 0.$$

and therefore

$$\widetilde{\kappa}\left(\mathcal{D}\left(\check{Q}^{(K,N)}, Q_\star^{(K)}\right)\right) \leq \widetilde{\kappa}\left(o_P(1)\right) \leq o_P(1). \tag{D.7}$$

Using Eqs. (D.6) and (D.7), we rewrite Eq. (D.2) as

$$\mathcal{D}_{\mathrm{dist}}(P_o, P_\star^{(K)}) \leq \widetilde{\kappa}\left(\sum_k \widehat{\mathcal{D}}\left(\widehat{G}_k^{(K,N)}, G\right)\right) + o_P(1).$$

Therefore, under the conditions for κ-robust consistency, we are guaranteed that

$$\widetilde{\kappa}(K\rho) \leq \widetilde{\kappa}\left(\sum_k \widehat{\mathcal{D}}\left(\widehat{G}_k^{(K,N)}, G\right)\right) + o_P(1).$$

Because $\widetilde{\kappa}$ is monotonic,

$$K\rho \leq \sum_k \widehat{\mathcal{D}}\left(\widehat{G}_k^{(K,N)}, G\right) + o_P(1).$$

This holds if and only if there exists $\ell \in \{1, \ldots, K\}$ such that $\mathcal{D}\left(G_{\star\ell}^{(K)}, G\right) \geq \rho$. Hence, for some $\varepsilon > 0$, $\mathcal{D}\left(G_{\star\ell}^{(K)}, G\right) = \rho + \varepsilon$. As a result

$$\begin{aligned}
\mathcal{R}^\rho\left(\widehat{G}_{1:K}^{(K,N)}\right) &\geq N \max\left(0, \widehat{\mathcal{D}}\left(\widehat{G}_\ell^{(K,N)}, G\right) - \rho\right) \\
&= N \max\left(0, \mathcal{D}\left(G_{\star\ell}^{(K)}, G\right) - \rho + o_P(1)\right) \qquad \text{(D.8)} \\
&= N \max\left(0, \varepsilon + o_P(1)\right) \\
&\to \infty
\end{aligned}$$

where Eq. (D.8) follows from Assumption C.1(c).

D.2. Verifying Assumption C.3 for Applications

We show that Assumption C.3 holds for both the PMF models used for the experiments in Section 3.3. It is sufficient to verify the assumption for a single element y_{nk} since the variance and integrability conditions can often be checked component-wise. Hence, we drop the dependence on n and k in our notation.

Poisson PMF

Consider the Poisson model with $y \sim \mathsf{Poiss}(\lambda)$ and $\lambda = Wh$. For convenience, we assume $h \sim \mathsf{Gamma}(\alpha, \beta)$. To compute the first and second moments of

$$P(y \mid h) = \frac{(Wh)^y e^{-Wh}}{y!},$$

we will use the identity $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\,\mathrm{d}t$ and integration by substitution. For the first moment we have

$$\begin{aligned}
\mathbb{E}_h\big[P(y \mid h)\big] &= \int_0^\infty \frac{(Wh)^y e^{-Wh}}{y!} \frac{\beta^\alpha}{\Gamma(\alpha)} h^{\alpha-1} e^{-\beta h}\,\mathrm{d}h \\
&= \frac{W^y \beta^\alpha}{y!\,\Gamma(\alpha)} \int_0^\infty h^{y+\alpha-1} e^{-(W+\beta)h}\,\mathrm{d}h \\
&= \frac{W^y \beta^\alpha\,\Gamma(y+\alpha)}{y!\,\Gamma(\alpha)(W+\beta)^{y+\alpha}},
\end{aligned}$$

while the second moment is

$$\begin{aligned}\mathbb{E}_h\big[P^2(y \mid h)\big] &= \int_0^\infty \frac{(Wh)^{2y}e^{-2Wh}}{(y!)^2}\,\frac{\beta^\alpha}{\Gamma(\alpha)}h^{\alpha-1}e^{-\beta h}\,\mathrm{d}h \\ &= \frac{W^{2y}\beta^\alpha}{\Gamma(\alpha)(y!)^2}\int_0^\infty h^{2y+\alpha-1}e^{-(2W+\beta)h}\,\mathrm{d}h \\ &= \frac{W^{2y}\beta^\alpha\,\Gamma(2y+\alpha)}{\Gamma(\alpha)(y!)^2(2W+\beta)^{2y+\alpha}}.\end{aligned}$$

Taking the ratio of the second to the first moment, define

$$C(y) = \frac{\mathbb{E}_h[P^2(y \mid h)]}{\mathbb{E}_h[P(y \mid h)]} = \frac{W^y\,\Gamma(2y+\alpha)}{y!\,\Gamma(y+\alpha)}\left(\frac{W+\beta}{2W+\beta}\right)^{y+\alpha}\cdot\left(\frac{1}{2W+\beta}\right)^y,$$

which is continuous and finite for all $y \in \mathbb{N}$. Now, using Stirling's approximation $\Gamma(z) \sim \sqrt{2\pi}\,z^{z-1/2}e^{-z}$, we have

$$\frac{\Gamma(2y+\alpha)}{y!\,\Gamma(y+\alpha)} \sim \frac{(2y)^{2y+\alpha-1/2}e^{-2y}}{\sqrt{2\pi}\,y^{y+\alpha-1/2}e^{-y}\cdot\sqrt{2\pi}\,y^{y+1/2}e^{-y}} \sim \frac{2^{2y+\alpha}}{\sqrt{\pi y}}.$$

Substituting into $C(y)$ yields

$$\begin{aligned}C(y) &\sim \frac{W^y}{\sqrt{\pi y}}\left(\frac{W+\beta}{2W+\beta}\right)^\alpha\left(\frac{W+\beta}{2W+\beta}\right)^y\left(\frac{1}{2W+\beta}\right)^y 2^{2y} \\ &\sim \frac{1}{\sqrt{\pi y}}\left(\frac{W+\beta}{2W+\beta}\right)^\alpha\left(\frac{2^2W(W+\beta)}{(2W+\beta)^2}\right)^y \\ &\sim \frac{1}{\sqrt{\pi y}}\left(\frac{4W^2+4W\beta}{4W^2+\beta^2+4W\beta}\right)^y.\end{aligned}$$

Because $0 < \frac{4W^2+4W\beta}{4W^2+\beta^2+4W\beta} < 1$, the ratio $C(y)$ decays exponentially as $y \to \infty$, hence $\sum_{y=0}^\infty C(y) < \infty$.

Gaussian PMF

Consider the Gaussian setting $y \sim \mathcal{N}(\phi h, \sigma^2)$ and, following common practice, we let $h \sim \mathcal{N}(\mu, \tau^2)$. To compute the moments of $p(y \mid h)$, we integrate over the latent variable h by combining the terms in the exponential, completing the square, and using the Gaussian integral identity $\int_{-\infty}^\infty e^{-(ax^2+bx+c)}\,\mathrm{d}x = \sqrt{\frac{\pi}{a}}e^{\frac{b^2}{4a}-c}$. The first moment is

$$\begin{aligned}\mathbb{E}_h[p(y \mid h)] &= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-\phi h)^2}{2\sigma^2}\right)\cdot\frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{(h-\mu)^2}{2\tau^2}\right)\,dh \\ &\propto \exp\left\{-\frac{(y-\phi\mu)^2}{2(\phi^2\tau^2+\sigma^2)}\right\},\end{aligned}$$

while the second moment is

$$\begin{aligned}\mathbb{E}_h[p(y \mid h)^2] &= \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\phi h)^2}{2\sigma^2}\right)\right)^2 \cdot \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(h-\mu)^2}{2\tau^2}\right) dh \\ &\propto \exp\left\{-\frac{(y-\phi\mu)^2}{2(\phi^2\tau^2 + \frac{1}{2}\sigma^2)}\right\}.\end{aligned}$$

Therefore, up to a multiplicative constant, the ratio of moments is

$$\begin{aligned}C(y) &= \frac{\mathbb{E}_h[P^2(y \mid h)]}{\mathbb{E}_h[P(y \mid h)]} \\ &\propto \frac{\exp\left\{-\frac{(y-\phi\mu)^2}{2(\phi^2\tau^2+\frac{1}{2}\sigma^2)}\right\}}{\exp\left\{-\frac{(y-\phi\mu)^2}{2(\phi^2\tau^2+\sigma^2)}\right\}} \\ &= \exp\left\{-(y-\phi\mu)^2\left(\frac{1}{2(\phi^2\tau^2+\frac{1}{2}\sigma^2)} - \frac{1}{2(\phi^2\tau^2+\sigma^2)}\right)\right\}.\end{aligned}$$

Since $\frac{1}{2(\phi^2\tau^2+\frac{1}{2}\sigma^2)} > \frac{1}{2(\phi^2\tau^2+\sigma^2)}$, we get $C(y) \propto \exp\{-a(y-\phi\mu)^2\}$ for some $a > 0$. Therefore, $C(y)$ decays exponentially as $y \to \infty$ and hence $\int C(y)\mathrm{d}y < \infty$.

Appendix E: Conditional Sampling for PMF Models

E.1. Poisson NMF

Recall the standard Poisson NMF model:

$$\begin{aligned}y_{nk}^{(K)} &\sim \mathsf{Poiss}\left(\phi_k^{(K)} z_{nk}^{(K)}\right) && \text{for } n = 1,\dots,N, k = 1,\dots,K \\ x_n &= \sum_{k=1}^{K} y_{nk}^{(K)} && \text{for } n = 1,\dots,N,\end{aligned}$$

Applying Bayes' rule, we can sample $\varepsilon_{nk} \mid x_n$ for any given n, k using the following procedure, with each dimension d sampled independently:

$$\begin{aligned}y_{n,1:K,[d]}^{(K)} \mid x_{n,[d]} &\sim \mathsf{Multi}\left(x_{n,[d]}; \widehat{p}_{n,1:K,d}\right), \\ \varepsilon_{n,k,[d]} \mid y_{n,k,[d]}^{(K)} &\sim \mathsf{Unif}\left(\mathcal{F}_{\mathsf{Poiss}}\left(y_{n,k,[d]}^{(K)} - 1; p_{n,k,d}\right),\ \mathcal{F}_{\mathsf{Poiss}}\left(y_{n,k,[d]}^{(K)}; p_{n,k,d}\right)\right),\end{aligned}$$

where

$$p_{n,k,d} = \phi_{k,[d]}^{(K)} z_{nk}^{(K)}, \qquad \widehat{p}_{n,k,d} = \frac{p_{n,k,d}}{\sum_{k'=1}^{K} p_{n,k',d}},$$

and $\mathcal{F}_{\mathsf{Poiss}}(\cdot;\lambda)$ is the cdf of $\mathsf{Poiss}(\lambda)$.

E.2. Normal Factor Analysis Model

Recall the usual Gaussian factor analysis model:

$$\begin{aligned}
\Sigma_k^{(K)} &= \left[\sigma_{k,1}^2, \ldots, \sigma_{k,D}^2\right]^{\mathrm{T}} && \text{for } k = 1, \ldots, K, \\
y_{nk}^{(K)} &\overset{\text{e.w}}{\sim} \mathcal{N}\left(\phi_k^{(K)} z_{nk}^{(K)}, \Sigma_k^{(K)}\right) && \text{for } n = 1, \ldots, N, k = 1, \ldots, K, \\
x_n &= \sum_{k=1}^{K} y_{nk} && \text{for } n = 1, \ldots, N.
\end{aligned}$$

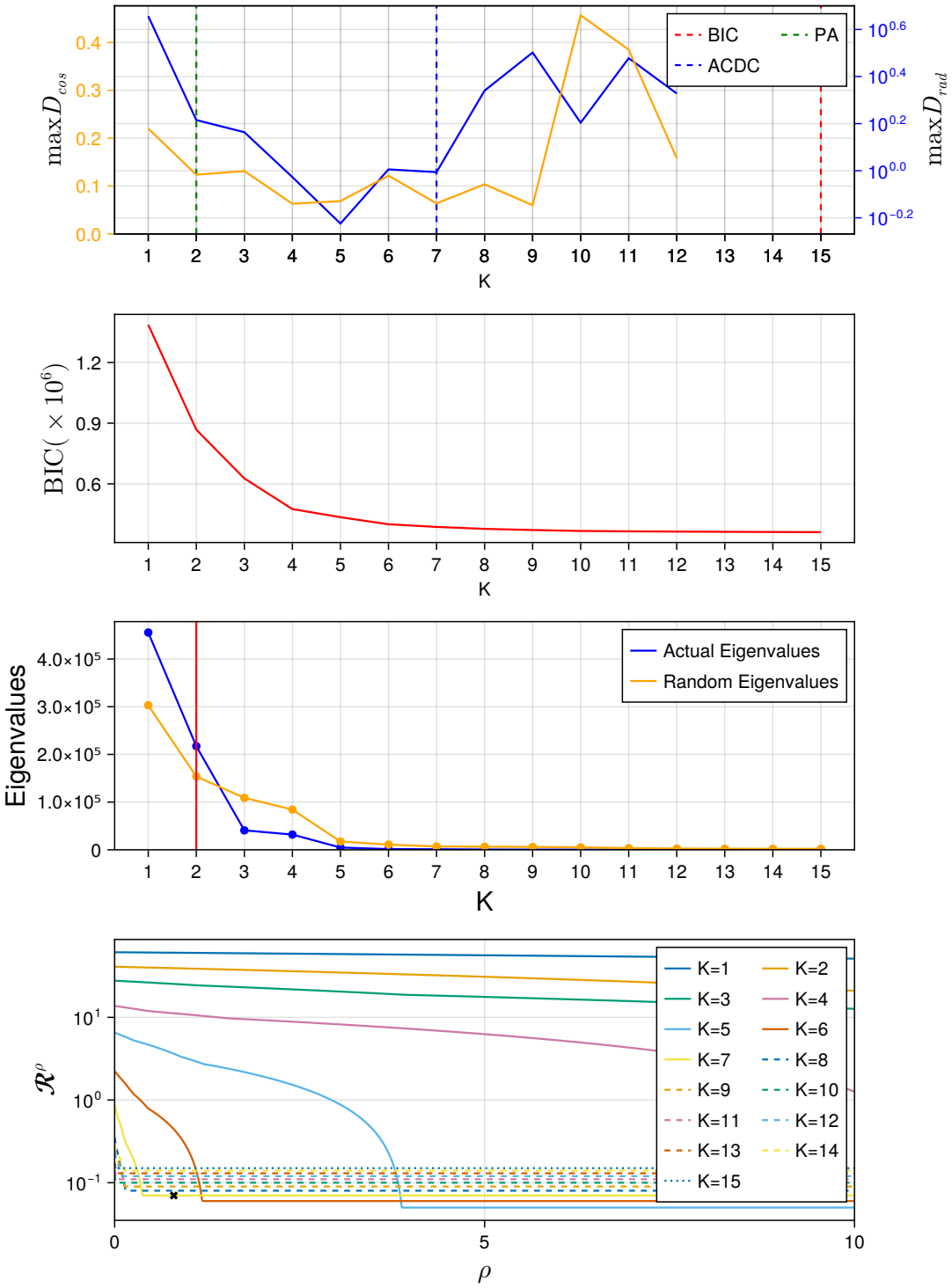
Again, applying Bayes' rule, we can sample $\varepsilon_{nk} \mid x_n$ for any given n, k using the following formulation, with each dimension d sampled independently:

$$\begin{aligned}
y_{n,k,[d]}^{(K)} \mid x_{n,[d]}, y_{n,1:k-1,[d]}^{(K)}, \sigma_{1:K,d} &\sim \begin{cases} \mathcal{N}(\widetilde{\mu}_{n,d,k}, \widetilde{\sigma}_{k,d}^2) & \text{if } k \neq K, \\ \delta\left(\overline{x}_{n,k,d}\right) & \text{if } k = K, \end{cases} \\
\varepsilon_{n,k,[d]} &= \mathcal{F}_{\mathcal{N}}\left(y_{n,k,[d]}^{(K)}; \mu_{n,k,d}, \sigma_{k,d}^2\right),
\end{aligned}$$

where

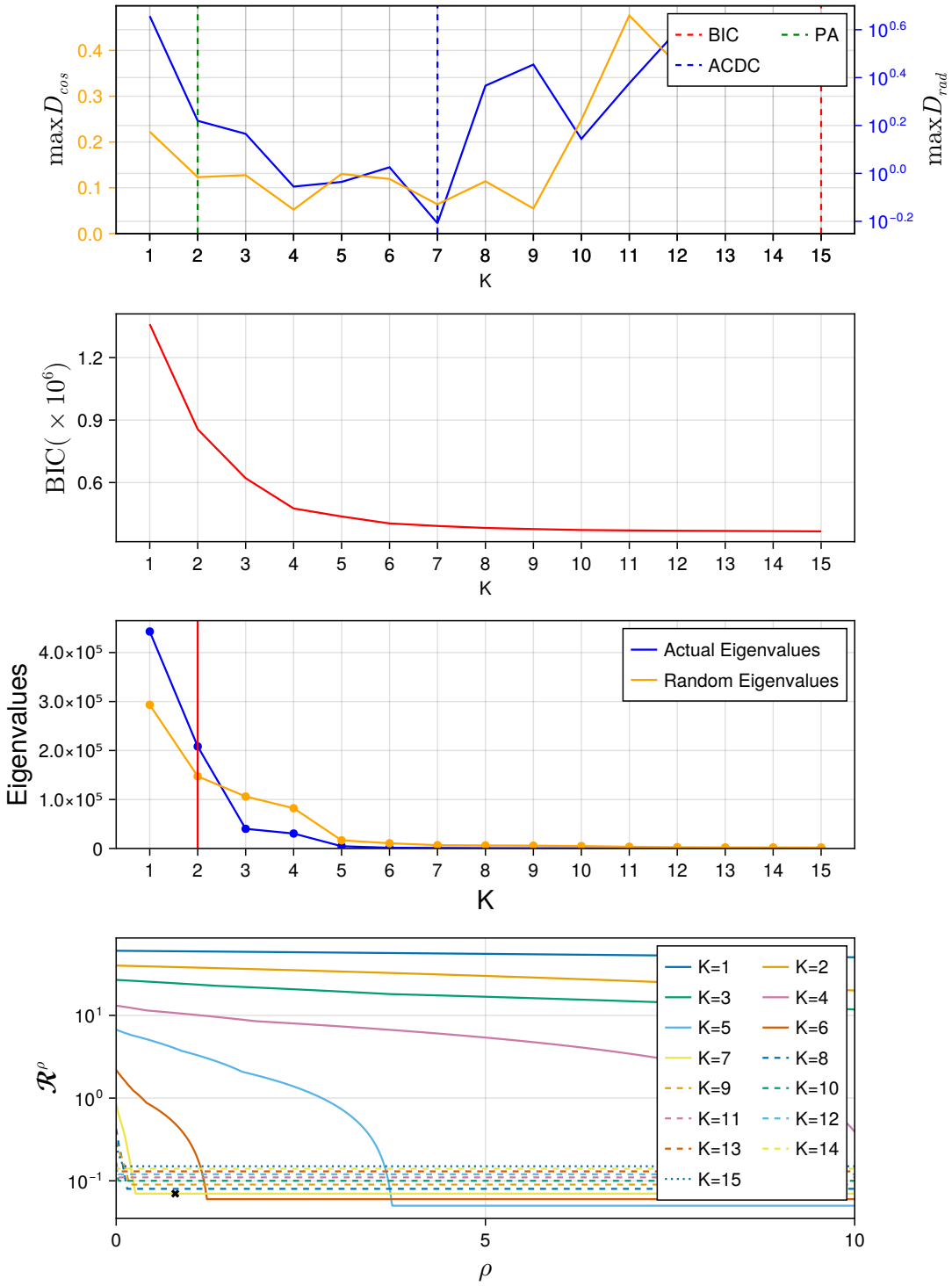
$$\begin{aligned}
\overline{x}_{n,k,d} &= x_{n,[d]} - \sum_{k'=1}^{k-1} y_{n,k',[d]}^{(K)}, & \mu_{n,k,d} &= \phi_{k,[d]}^{(K)} z_{nk}^{(K)}, \\
\overline{\mu}_{n,k,d} &= \sum_{k'=k+1}^{K} \mu_{n,k',d}, & \overline{\sigma}_{k,d}^2 &= \sum_{k'=k+1}^{K} \sigma_{k',d}^2, \\
\widetilde{\mu}_{n,k,d} &= \frac{\sigma_{k,d}^{-2}\mu_{n,k,d} - \overline{\sigma}_{k,d}^{-2}(\overline{\mu}_{n,k,d} - \overline{x}_{n,k,d})}{\sigma_{k,d}^{-2} + \overline{\sigma}_{k,d}^{-2}}, & \widetilde{\sigma}_{k,d}^2 &= \frac{\sigma_{k,d}^2 \overline{\sigma}_{k,d}^2}{\sigma_{k,d}^2 + \overline{\sigma}_{k,d}^2},
\end{aligned}$$

and $\mathcal{F}_{\mathcal{N}}(\cdot; \mu, \sigma^2)$ is the cdf of $\mathcal{N}(\mu, \sigma^2)$.



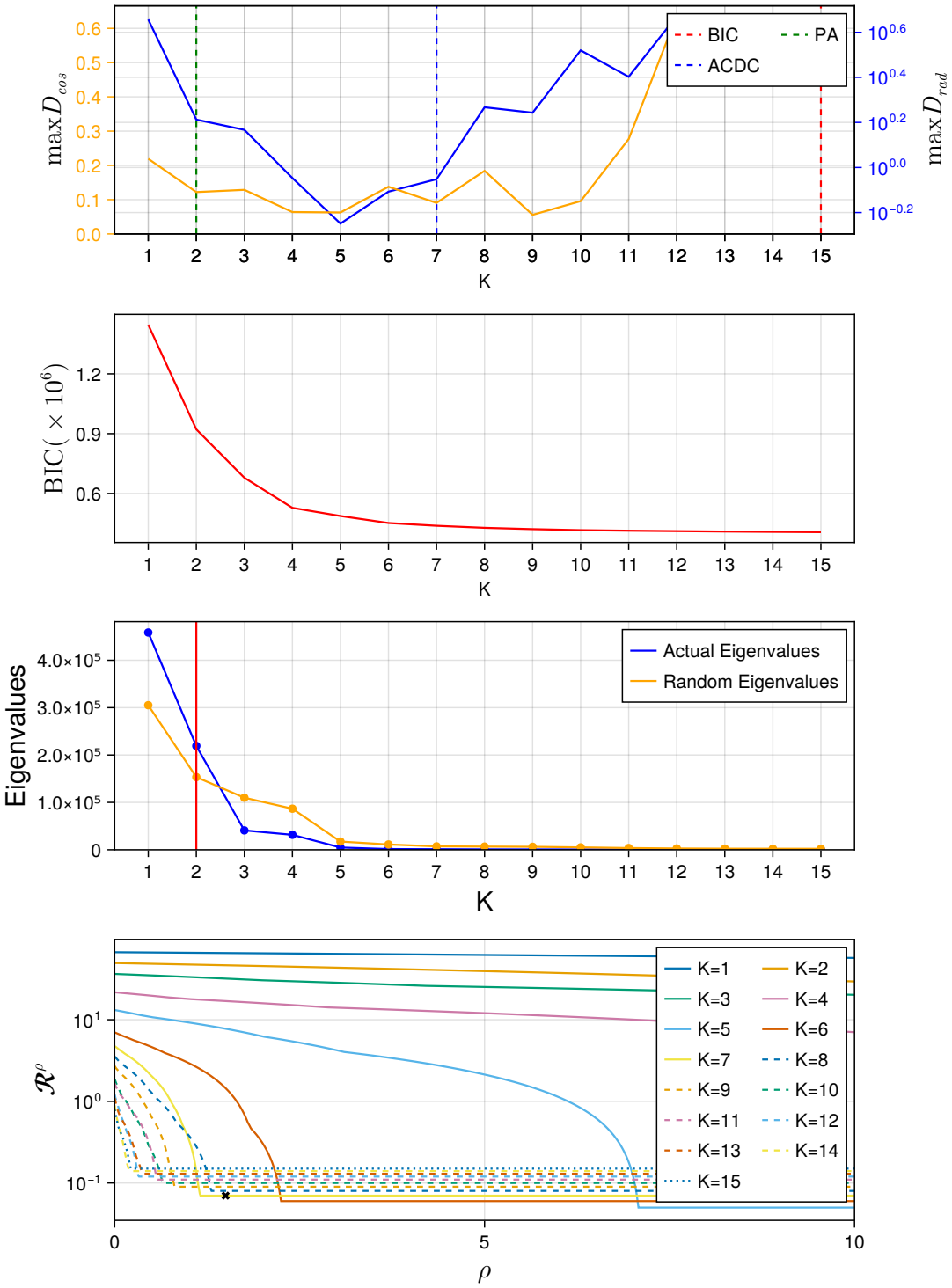
(a) Well specified

Fig E.1: Estimation quality for various scheme of data generation. See Fig. 3.3 caption for explanation.



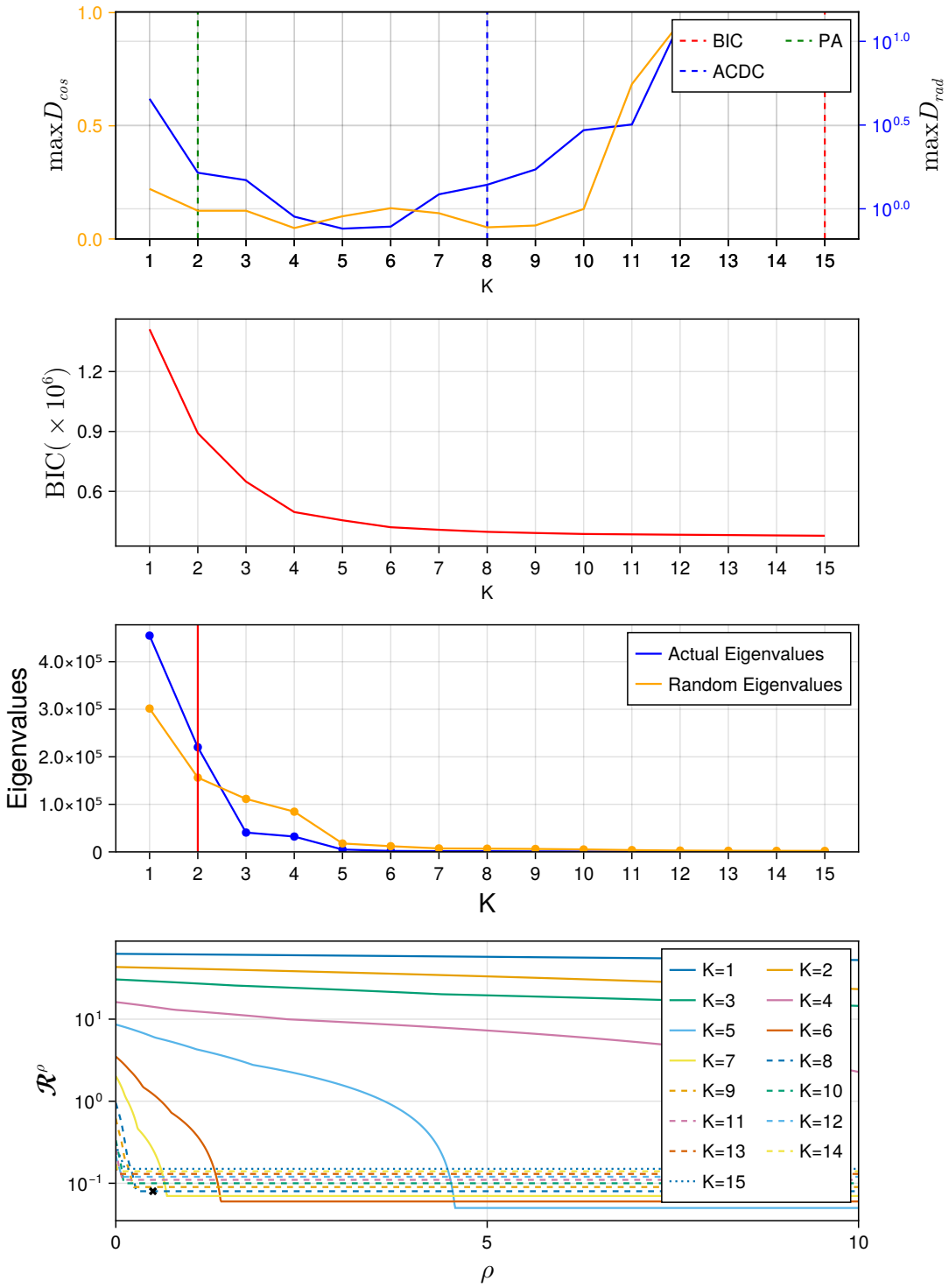
(b) Contaminated

Fig E.1: Estimation quality for various scheme of data generation. See Fig. 3.3 caption for explanation.



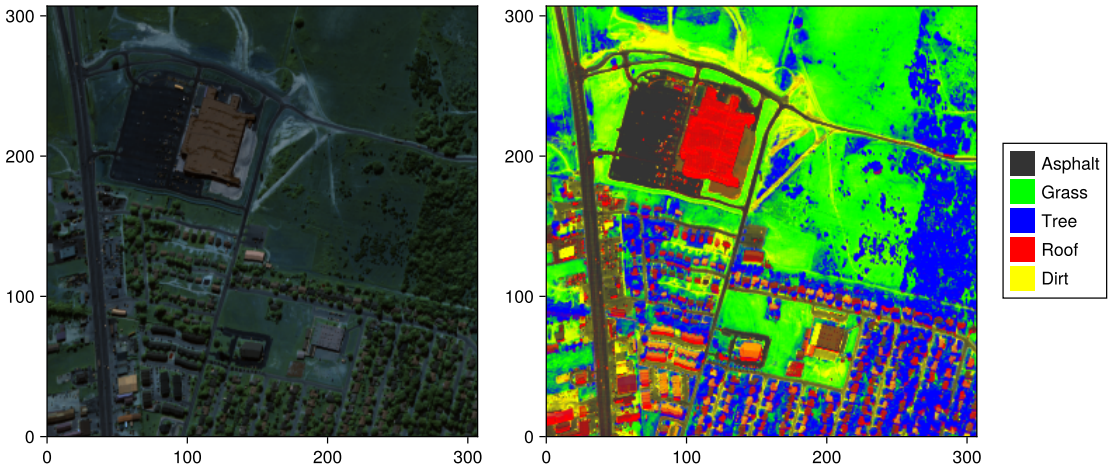
(c) Overdispersed

Fig E.1: Estimation quality for various scheme of data generation. See Fig. 3.3 caption for explanation.

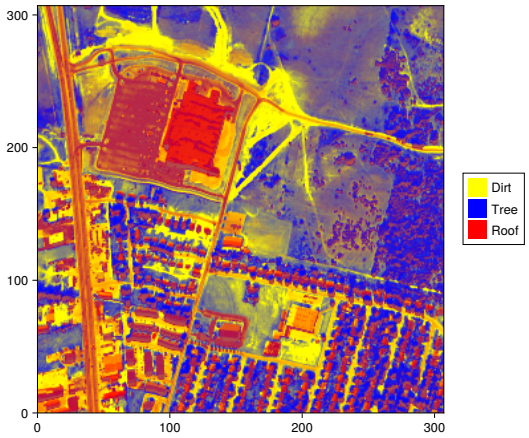


(d) Perturbed

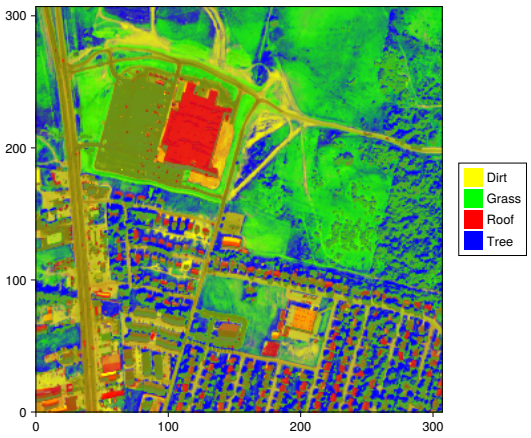
Fig E.1: Estimation quality for various scheme of data generation. See Fig. 3.3 caption for explanation.



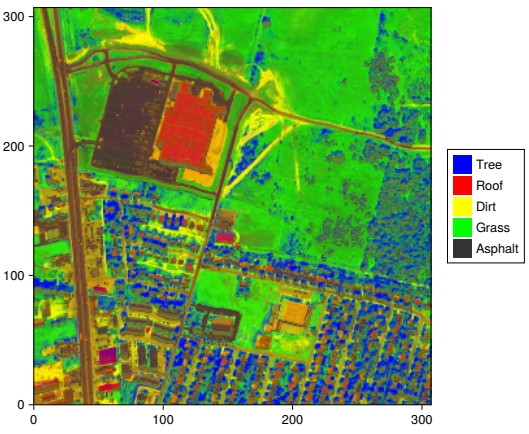
(a) Unlabeled (left) and labeled (right) versions of the urban dataset



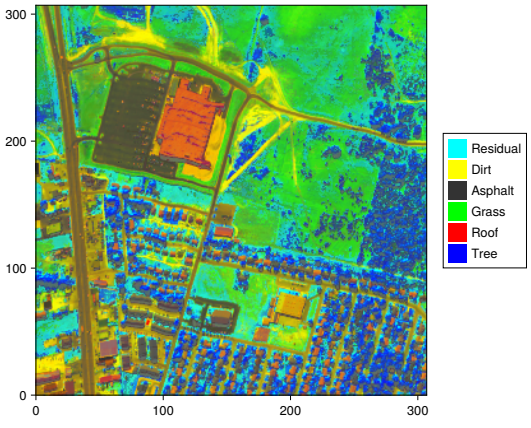
(b) $K = 3$



(c) $K = 4$



(d) $K = 5$



(e) $K = 6$

Fig E.2: Visualization of the hyperspectral urban datasets. **(a)** Data and ground truth labels. **(b, c, d, e)** Inferred end-member abundance maps for $K = 3, 4, 5, 6$.