Identifying Mixture Models

Leonard J. Schulman

Caltech

Simons Inst., March 22, 2022

Based on joint works with Chaitanya Swamy (Waterloo), Yuval Rabani (Hebrew U), Jian Li (Tsinghua), Spencer Gordon (Caltech) and Bijan Mazaheri (Caltech)

We are interested in Bayesian Networks $\mathcal{G} = (\mathcal{V} \sqcup \mathcal{U}, \mathcal{E})$ with **visible** vertices $\mathcal{V}$ and **hidden** or **latent** confounders $\mathcal{U}$.

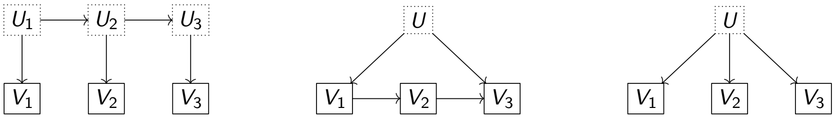


Figure 1: Hidden Markov Model; "Front Door"; MixProd

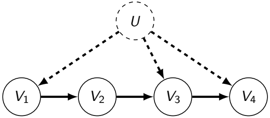
Write $\mathcal{P}$ for the restriction of the joint distribution to $\mathcal{V}$. This is what we can learn (up to sampling noise) from data. $\mathcal{P}$ is Markovian on the graph: factors as

$$\Pr(v_1, \ldots, v_n) = \prod_{i=1}^{n} \Pr(V_i = v_i \mid \mathbf{pa}(V_i))$$

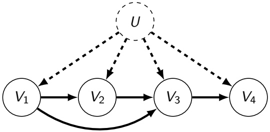
where $\mathbf{pa}(V_i)$ is the assignment to the parents of V_i. These conditionals are the **parameters** of the model.

Source Identification / Parameter learning

If we're given the distribution on all variables (including $\mathcal{U}$), we can easily identify all the parameters of the model. But we're actually only given $\mathcal{P}$ (or empirical $\widehat{\mathcal{P}}$). So what *can* we determine? In some cases [Pearl/ Tian/ Shpitser/ Huang/ Valtorta] can make remarkable deductions. E.g., in:



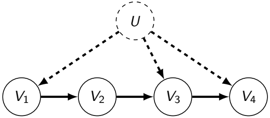
can deduce effect of an intervention at V_1 on V_3, despite confounder U.
But in most cases, there's little we can determine from $\mathcal{P}$. E.g., if single U can affect all visible variables:



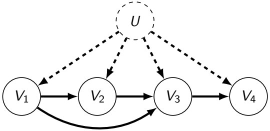
U can generate **any** distribution on $\mathcal{V}$.

Source Identification / Parameter learning

If we're given the distribution on all variables (including $\mathcal{U}$), we can easily identify all the parameters of the model. But we're actually only given $\mathcal{P}$ (or empirical $\widehat{\mathcal{P}}$). So what *can* we determine? In some cases [Pearl/ Tian/ Shpitser/ Huang/ Valtorta] can make remarkable deductions. E.g., in:



can deduce effect of an intervention at V_1 on V_3, despite confounder U.
But in most cases, there's little we can determine from $\mathcal{P}$. E.g., if single U can affect all visible variables:



U can generate **any** distribution on $\mathcal{V}$. But in order to do so, U needs to range over a large set. (Size 2^n for binary V_i's.)

Cardinality/dimension bounds on hidden variables

Cardinality or dimension bounds on hidden variables are a long-standing assumption e.g., for Hidden Markov Models. In our “non-parametric” context, natural assumption is cardinality.

$$k = \text{cardinality}(\text{range}(U))$$

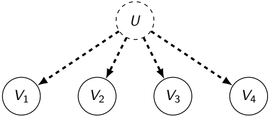


Figure 2: k-MixProd

Tower of increasingly general problems:

$$k\text{-MixIID} < k\text{-MixProd} < k\text{-MixBND}$$

In k-MixProd, the V_i are independent conditional on U.
In k-MixIID, they are moreover iid conditional on U.

Cardinality/dimension bounds on hidden variables

Cardinality or dimension bounds on hidden variables are a long-standing assumption e.g., for Hidden Markov Models. In our "non-parametric" context, natural assumption is cardinality.

$$k = \text{cardinality}(\text{range}(U))$$

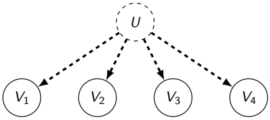


Figure 2: k-MixProd

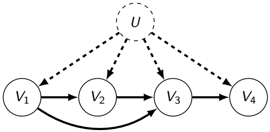


Figure 3: More general k-MixBND.

Tower of increasingly general problems:

$$k\text{-MixIID} < k\text{-MixProd} < k\text{-MixBND}$$

In k-MixProd, the V_i are independent conditional on U.
In k-MixIID, they are moreover iid conditional on U.

Cardinality/dimension bounds on hidden variables

Cardinality or dimension bounds on hidden variables are a long-standing assumption e.g., for Hidden Markov Models. In our "non-parametric" context, natural assumption is cardinality.

$$k = \text{cardinality}(\text{range}(U))$$

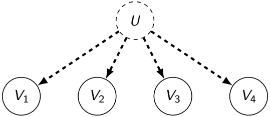


Figure 2: k-MixProd

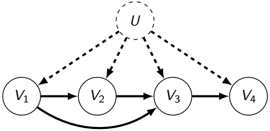


Figure 3: More general k-MixBND.

Tower of increasingly general problems:

$$\underbrace{k\text{-MixIID} < k\text{-MixProd}}_{\text{this lecture}} < k\text{-MixBND}$$

In k-MixProd, the V_i are independent conditional on U.

In k-MixIID, they are moreover iid conditional on U.

Talk outline

1. k-MixIID and the classical moment problem.
   Key concepts:
   1. Prony's algorithm.
   2. Hankel matrices.

   Theorems: Hankel condition number, sample complexity
   lower bound $\sim \exp(\Omega(k))$.
   (And upper bounds also for transportation distance reconstruction.)
2. k-MixProd.
   Key concepts:
   1. Method of synthetic bits.
   2. Hadamard Extensions.

   Theorems: Hadamard Extension condition number, sample complexity
   upper bound $\sim \exp O(k \log k)$.

1. k-MixIID and the classical moment problem

In k-MixIID: U is distributed on $\{1, \ldots, k\}$ according to an unknown prob. dist. π. Focus on case that V_i's are all binary.
For each $u \in \{1, \ldots, k\}$ there is an $0 \leq \mathbf{m}_u \leq 1$ s.t.

$$\Pr(V_i | U = u) = \mathbf{m}_u$$

so by conditional independence of the V_i, $V_R = \bigwedge_{i \in R} V_i$

$$\Pr(V_R = 1 | U = u) = \mathbf{m}_u^{|R|} .$$

So for the rv $Y = \#$ Heads

$$Y = |\{i : V_i = 1\}|$$

the moments of Y are linear combinations of the moments of the "k-spike" atomic probability distribution on $[0, 1]$

$$p = \sum_{u=1}^{k} \pi_u \delta_{\mathbf{m}_u} \qquad \text{(here } \delta_x \text{ is unit measure at } x\text{)}$$

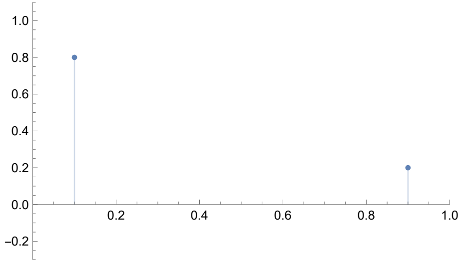


Figure 4: 2-spike dist. p with $\pi_1 = 0.8$ at $\mathbf{m}_1 = 0.1$, and $\pi_2 = 0.2$ at $\mathbf{m}_2 = 0.9$

Let $X \sim p$ and let $\mu_j = E(X^j)$. Then

$$E(Y) = n \sum_u \pi_u \mathbf{m}_u = n\mu_1$$
$$E(Y^2) = n\mu_1 + n(n-1)\mu_2$$
$$E(Y^3) = n\mu_1 + 3n(n-1)\mu_2 + n(n-1)(n-2)\mu_3 \quad \ldots \text{etc.}$$

Triangular linear system with nonzero diagonal coefficients. So the moments of Y (0 through n), which we learn from $\mathcal{P}$, determine the moments μ_j of the k-spike dist. p.

The Moment Problem

Classical question: given μ_j ($j \geq 0$), are they the moments of a measure on $\mathbb{R}$?
Classical answer: yes iff for every $K \geq 1$, the **Hankel matrix**

$$H_K = \begin{pmatrix} \mu_0 & \mu_1 & \cdots & \mu_{K-1} \\ \mu_1 & \mu_2 & \cdots & \mu_K \\ \cdots & \cdots & \cdots & \cdots \\ \mu_{K-1} & \mu_{K+1} & \cdots & \mu_{2K-2} \end{pmatrix}$$

is nonnegative-definite.

Furthermore, the measure is **unique** provided the μ_j do not grow too quickly. For a distribution supported on $[0, 1]$ (Hausdorff moment problem), such as p, this is guaranteed.

For a k-spike distribution p, (1) How many moments are required to identify p, (2) How do we do so algorithmically?

Answers:

(1) $\mu_1, \ldots, \mu_{2k-1}$ suffice. (Easy to see necessary.)
(And can verify dist. is k-spike if we're also given μ_{2k}.)
Consequently sufficient to have $n = 2k - 1$ observable rv's.
(2) Algorithm of Prony (1795). Relies on the Hankel matrix which for k-spike dists is:

$$H_{k+1} = V_{k+1}^{\perp} \cdot \mathrm{diag}(\pi) \cdot V_{k+1} \tag{1}$$

where V_ℓ is the $k \times \ell$ Vandermonde matrix of the spike sites:

$$V = \begin{pmatrix} 1 & \mathbf{m}_1 & \mathbf{m}_1^2 & \ldots & \mathbf{m}_1^{\ell-1} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & \mathbf{m}_k & \mathbf{m}_k^2 & \ldots & \mathbf{m}_k^{\ell-1} \end{pmatrix}$$

$$\mathrm{diag}(\pi) = \begin{pmatrix} \pi_1 & 0 & 0 \\ 0 & \ldots & 0 \\ 0 & 0 & \pi_k \end{pmatrix}$$

This solves k-MixIID if you have perfect statistics, i.e., exact H_{k+1}.
"Living in Asymptotia"

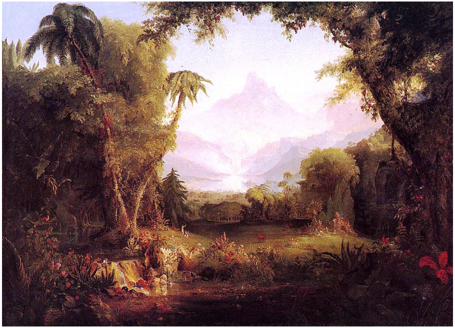


Figure 5: Thomas Cole, The Garden of Eden, 1828

This solves k-MixIID if you have perfect statistics, i.e., exact H_{k+1}.
“Living in Asymptotia”

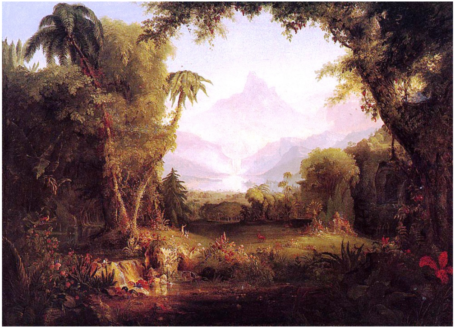


Figure 5: Thomas Cole, The Garden of Eden, 1828

Figure 6: Gustave Doré, Adam and Eve Driven out of Eden, 1865

Sample size

Prony's alg. notoriously unstable as a function of empirical dist. $\widehat{\mathcal{P}}$. Is this a property of the algorithm or of the problem? When spikes collide, model parameters not identifiable, so accuracy of $\widehat{\mathcal{P}}$ (hence sample size) must depend on: **separation parameter**

$$\zeta = \min_{i \neq j} |\mathbf{m}_i - \mathbf{m}_j| .$$

Theorem 1 (Rabani S Swamy '14)

For any $n \in O(k)$, $\|\widehat{\mathcal{P}} - \mathcal{P}\|_\infty \leq \zeta^{O(k)}$ (therefore sample size $\geq (1/\zeta)^{\Omega(k)}$) is necessary even to determine parameters within $\pm 1/k$.
(Neglecting dependence on mixture weights.)

That paper also gave sample size upper bound of $(1/\zeta)^{O(k^2)}$. Since improved [Li Rabani **S** Swamy '15], [Kim, Koehler, Moitra, Mossel, Ramnarayan '19], [Gordon Mazaheri Rabani **S** '20] to $(1/\zeta)^{O(k)}$; also give reconstruction in Weierstrass-1 (transportation) distance. Key is an upper bound on condition number of Hankel H_k.

2. k-MixProd

Recall k-MixProd, a much more general problem than k-MixIID.

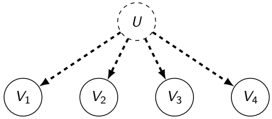


Figure 7: k-MixProd

Parameters: prior π on hidden variable U, and an $n \times k$ matrix

$$\mathbf{m}_{iu} = \Pr(V_i = 1 | U = u)$$

Prior work focused on **learning** rather than **identifying** the model.
"Learning" = reconstruct any model $(\pi, \mathbf{m})$ creating statistics close to the observed statistics.
"Identifying" = learning in regions of parameter space $(\pi, \mathbf{m})$ where there is a stable invertibility guarantee:

$$\forall \varepsilon \exists \delta \text{ s.t. if } \text{dist}((\pi, \mathbf{m}), (\pi', \mathbf{m}')) > \delta \text{ then } |\mu(\pi, \mathbf{m}) - \mu(\pi', \mathbf{m}')| > \varepsilon.$$

Identification gives stronger output guarantees than *Learning*, under stronger assumptions.

Identification as a goal goes back at least to [Koopmans, Reiersol 1950], [Koopmans 1950], [Teicher 1963], [Blischke 1964], [Yakowitz, Spragins 1968]

Since more is assumed, runtime might be better.

For our motivations, identification is the right problem, since it tells you how the system will function if you **intervene** (set some of the random variables).

Literature on k-MixProd

Mixture models began with [Newcomb 1886], [Pearson 1894]. See [Everitt, Hand 1981], [Titterington et al. 1985], [Lindsay 1995], [McLachlan et al. 2019]. Abundant literature for discrete variables thanks to disparate motivations, e.g., astronomy, population genetics, bioinformatics, image recognition, text classification; see [Pritchard et al. 2000], [Ji et al. '05], [Juan, Vidal '02, '04]. Iterative methods (EM) often used [Juan et al. '04], [Li et al. '16], [Palmer et al. '16], [Carrerira-Perpiñán, Renals '00], [Najafi et al. '20] ...

Algorithms with provable guarantees, some for Gaussians: $k = 2$: [Kearns et al. '94], [Freund, Mansour '99], [Dasgupta '99], [Cryan, Goldberg, Goldberg '02]. General k: [Feldman, O'Donnell, Servedio '08], [Chaudhuri, Rao '08], [Moitra, Valiant '10], [Arora et al. '12], [Anandkumar et al. '12ab], [Rabani et al. '14], [Hardt, Price '15], [Li et al. '15], [Kim et al. '19], [Chen, Moitra '19], [Wu, Yang '20], [Rabani et al. '20] ...

The provable "learning" algorithms use grid-and-search in parameter space. Due to grid search, this is very expensive: to learn a model which reproduces statistics within variation distance ε, [FOS'08] runtime is $(nk/\varepsilon)^{O(k^3)}$, [CM'19] $k^{O(k^3)}(n/\varepsilon)^{O(k^2)}$.

Identifying a k-MixProd model

We study k-MixProd under a ζ-separation assumption:

$$\forall i \forall u \neq u' : \ | \mathbf{m}_{iu} - \mathbf{m}_{iu'} | > \zeta > 0 \tag{2}$$

Comments:

(a) Separation for $\zeta = 0$ was shown by [Tahmasebi Motahari Maddah-Ali '18] to imply that the mapping $(\pi, \mathbf{m}) \to \mu$ is injective. (Provided $\pi > 0$, and up to the obvious symmetry of permuting columns.) Algebraic result, no algorithm.

(b) It is clear that some kind of separation guarantee is necessary: e.g., two identical columns make the model unidentifiable.
The separation assumption (2) is a little stronger than necessary. We provide a sufficient weaker assumption in [Gordon **S** '22]. However it is not algorithmic.
Full characterization and efficient algorithm beyond ζ-separation remain open problems.

We give an algorithm (different approach from [TMM'18] entirely) for ζ-separated k-MixProd. Near-optimal in sample complexity.

Theorem 2 (Gordon Mazaheri Rabani S, manuscript)

For a k-MixProd model on $n \geq 3k-3$ bits, we can identify a model $(\hat{\pi}, \hat{\mathbf{m}})$ with all parameters within $\pm\varepsilon$ of true $(\pi, \mathbf{m})$, in runtime and sample complexity

$$(1/\zeta)^{O(k \log k)} \varepsilon^{-2} n \log n.$$

What is the key challenge? In k-MixIID the observables $V_1, \ldots, V_n$ were iid conditional on the hidden variable U. So our **multilinear moments** (3)

$$E(V_1 V_2) = \sum_u \Pr(u) \Pr(V_1 = 1|u) \Pr(V_2 = 1|u) = \sum \pi_u \mathbf{m}_{1u} \mathbf{m}_{2u} \quad (3)$$

$$= \sum_u \pi_u \mathbf{m}_{1u}^2 \quad (4)$$

were actually higher moments (4) of the single k-spike distribution p.
But now each V_i has a unique dependence on U. What good does it do to combine information between different V_i?

Hadamard extension of $\mathbf{m}$

Given two row vectors $\mathbf{m}_1$ and $\mathbf{m}_2$ in $\mathbb{R}^k$, their **Hadamard product** is

$$\mathbf{m}_1 \odot \mathbf{m}_2 \in \mathbb{R}^k$$
$$(\mathbf{m}_1 \odot \mathbf{m}_2)_u = \mathbf{m}_{1u}\, \mathbf{m}_{2u}$$

For $n \times k$ matrix $\mathbf{m}$, its Hadamard Extension is the $2^n \times k$ matrix with rows indexed by $S \subseteq [n]$, multilinear version of Vandermonde:

$$\mathbb{H}(\mathbf{m})_S = \bigodot_{i \in S} \mathbf{m}_i$$

or explicitly: $$\mathbb{H}(\mathbf{m})_{S,u} = \prod_{i \in S} \mathbf{m}_{iu}$$

(appearing first, not with this name, in [Chen Moitra'19].) E.g.,

$$\mathbf{m} = \begin{pmatrix} 1/2 & 1/3 & 1/5 \\ 1/7 & 1/11 & 1/13 \end{pmatrix} \quad \Rightarrow \quad \mathbb{H}(\mathbf{m}) = \begin{pmatrix} 1 & 1 & 1 \\ 1/2 & 1/3 & 1/5 \\ 1/7 & 1/11 & 1/13 \\ 1/14 & 1/33 & 1/65 \end{pmatrix}$$

A complete list of observable statistics of our model is $\Pr(V_R)$, where $V_R = \bigwedge_{i \in R} V_i$, ranging over all $R \subseteq [n]$. These probabilities are given by the vector

$$\begin{pmatrix} \Pr(V_R) \end{pmatrix} = \mathbb{H}(\mathbf{m}) \begin{pmatrix} \pi_1 \\ \dots \\ \pi_k \end{pmatrix}$$

Of course, we know only the vector on the LHS, not $\mathbb{H}(\mathbf{m})$ or π.

A complete list of observable statistics of our model is $\Pr(V_R)$, where $V_R = \bigwedge_{i \in R} V_i$, ranging over all $R \subseteq [n]$. These probabilities are given by the vector

$$\begin{pmatrix} \\ \Pr(V_R) \\ \\ \end{pmatrix} = \mathbb{H}(\mathbf{m}) \begin{pmatrix} \pi_1 \\ \dots \\ \pi_k \end{pmatrix}$$

Of course, we know only the vector on the LHS, not $\mathbb{H}(\mathbf{m})$ or π.

It turns out that we will be able to use $\mathbb{H}(\mathbf{m})$ in our algorithm, *without* knowing it.

In order to be able to use $\mathbb{H}(\mathbf{m})$ at finite sample size, though, we *also* need to understand something about its numerical stability (not just rank). Discuss this first; later the algorithm.

Condition number of Hadamard extensions

Lemma 3

Let A be any set of $k-1$ ζ-separated rows of $\mathbf{m}$. Write $\mathbf{m}|_A = \mathbf{m}$ restricted to the rows $i \in A$. Then the k'th-largest singular value of $\mathbb{H}(\mathbf{m}|_A)$ satisfies:

$$\sigma_k(\mathbb{H}(\mathbf{m}|_A)) \geq \zeta^{O(k)}.$$

Effectively a far generalization of the eigenvalue lower bound for Hankel matrices; here Vandermonde $\hookrightarrow$ Hadamard Extension. Clearest using

Lemma 4 (Feldman O'Donnell Servedio '08)

Let M be an $r \times k$ matrix, $r \geq k$. Then $\exists$ a set J of k rows s.t. $\sigma_k(M|_J) \geq \frac{\sigma_k(M)}{\sqrt{k(r-k)+1}}$.

Corollary 5

$\mathbb{H}(\mathbf{m}|_A)$ has a $k \times k$ submatrix $\mathcal{A}$ with $\sigma_k(\mathcal{A}) \geq \zeta^{O(k)}$.

An alternative way of arranging the values $\Pr(V_R)$ is in the $2^n \times 2^n$ matrix

$$C = \mathbb{H}(\mathbf{m})\,\mathrm{diag}(\pi)\mathbb{H}(\mathbf{m})^\top$$

If $R \cap R' = \emptyset$ then $C_{R,R'} = \Pr(V_{R\cup R'})$. So we can observe some, **not all,** entries of this matrix. E.g., first column corresp. to $\emptyset$ so fully observable:

$$(\mathbb{H}(\mathbf{m})\,\mathrm{diag}(\pi)\mathbb{1})_R = \Pr(V_R).$$

If for $A, B \subseteq [n]$, $A \cap B = \emptyset$ then we can observe the entire smaller matrix

$$\mathbb{H}(\mathbf{m}\,|_B)\,\mathrm{diag}(\pi)\mathbb{H}(\mathbf{m}\,|_A)^\top. \quad \text{multilinear gen'l of Hankel matrix}$$

In particular if $A, B \subseteq [n]$, $|A| = |B| = k-1$, $A \cap B = \emptyset$, then by Cor. 5, $\mathbb{H}(\mathbf{m}\,|_A)$ has a $k \times k$ submatrix $\mathcal{A}$, and $\mathbb{H}(\mathbf{m}\,|_B)$ has a $k \times k$ submatrix $\mathcal{B}$, such that we have good conditioning of the $k \times k$ matrix

$$C_{\mathcal{B}\mathcal{A}} = \mathcal{B}\,\mathrm{diag}(\pi)\mathcal{A}^\top$$

Lemma 6

$\sigma_k(C_{\mathcal{B}\mathcal{A}}) \geq \zeta^{O(k)}$.

Reducing k-MixProd to k-MixIID: method of synthetic bits

Fix disjoint $A, B \subseteq [n]$ and well-conditioned $C_{\mathcal{B}\mathcal{A}}$ as above. Let $\mathbf{m}_1$ be any row *outside of* $A \cup B$.

Strategy: we use the rows of $\mathcal{B}$ to **synthesize** a row equivalent to $\mathbf{m}_1$; we use $\mathcal{A}$ to determine the weights of this synthesis.
Recall that

$$E(V_1) = \mathbf{m}_1 \operatorname{diag}(\pi) \mathbb{1}$$

We wish we had a variable V_1' that was iid to V_1 conditional on U; if so we'd be able to observe

$$E(V_1 V_1') = (\mathbf{m}_1 \odot \mathbf{m}_1) \operatorname{diag}(\pi) \mathbb{1}$$

We don't have such a V_1' but the next-best thing is to construct $\mathbf{m}_1 \odot \mathbf{m}_1$.
Concretely (marking in violet quantities we can compute):

(1) Let

$$v_1 := \mathbf{m}_1 \operatorname{diag}(\pi)\mathcal{A}^\top$$

(We can observe v_1 because row 1 is not in A.)
In particular if $S = \emptyset$ is among the sets used in $\mathcal{A}$, then v_1 has an entry

$$(v_1)_\emptyset = \mathbf{m}_1 \operatorname{diag}(\pi)\mathbb{1} = E(V_1).$$

(2) Let

$$u_1 := v_1 C_{\mathcal{B}\mathcal{A}}^{-1}$$

u_1 is a set of weights that synthesize a copy of $\mathbf{m}_1$ out of $\mathcal{B}$:

$$\begin{aligned} u_1\mathcal{B} &= [\mathbf{m}_1 \operatorname{diag}(\pi)\mathcal{A}^\top] C_{\mathcal{B}\mathcal{A}}^{-1}\mathcal{B} \\ &= \mathbf{m}_1 \operatorname{diag}(\pi)\mathcal{A}^\top (\mathcal{A}^\top)^{-1} \operatorname{diag}(\pi)^{-1}\mathcal{B}^{-1}\mathcal{B} \\ &= \mathbf{m}_1 \end{aligned}$$

(3) Since row 1 is not in B, we can replace every $R \in \mathcal{B}$ by $R \cup \{1\}$. Form the $k \times k$ matrix $\bar{\mathcal{B}}$ with these "upshifted" rows, then let

$$v_2 := u_1 C_{\bar{\mathcal{B}}\mathcal{A}}$$

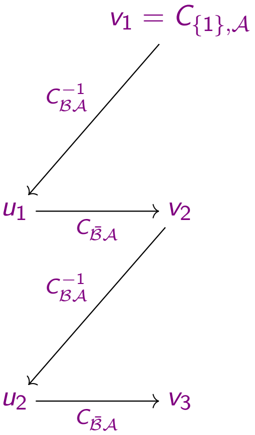
This gets us a second moment! v_2 has an entry

$$\begin{aligned}(v_2)_\emptyset &= (u_1 C_{\bar{\mathcal{B}}\mathcal{A}})_\emptyset \\ &= u_1 \bar{\mathcal{B}}\,\mathrm{diag}(\pi)\mathbb{1} \\ &= (\mathbf{m}_1 \odot (u_1 \mathcal{B}))\,\mathrm{diag}(\pi)\mathbb{1} \qquad \text{Hadamard prod. distributes} \\ &= (\mathbf{m}_1 \odot \mathbf{m}_1)\,\mathrm{diag}(\pi)\mathbb{1} \\ &= E(V_1 V_1')\end{aligned}$$

(4) Synthesize again! Weight rows of B to create u_2 s.t. $u_2 B = \mathbf{m}_1 \odot \mathbf{m}_1$.

$$u_2 := v_2 C_{\mathcal{B}\mathcal{A}}^{-1}$$

and keep going!



v_1 includes $E(V_1)$

v_2 includes $E(V_1 V_1')$

v_3 includes $E(V_1 V_1' V_1'')$

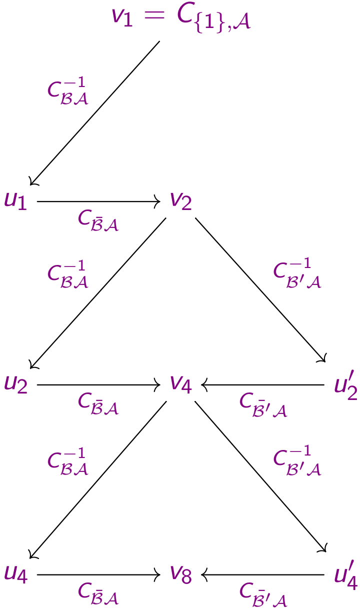
After $2k-1$ levels, can apply k-MaxIID algorithm.

Operator norm of $C_{\mathcal{B}\mathcal{A}}^{-1}$ is bounded by $(1/\zeta)^{O(k)}$ so after these $2k-1$ levels, errors blow up by $\leq (1/\zeta)^{k^2}$. Improve this by:

Synthetic bits method with repeated squaring

Needs $n = 3k - 3$ instead of $n = 2k - 1$. Use disjoint sets A, B, B' each with $k - 1$ ζ-separated rows.

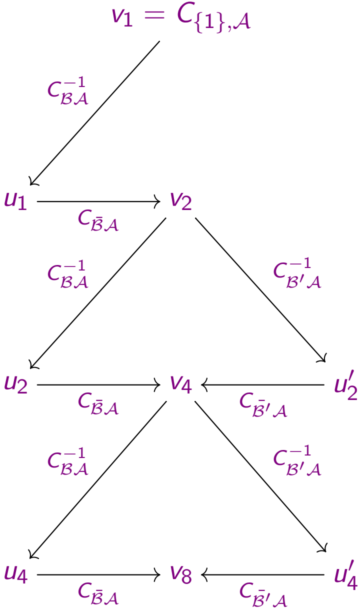
After these $\lg k$ levels, errors blow up by $\leq (1/\zeta)^{k \lg k}$.
$\Rightarrow$ sample size matches (almost) the $(1/\zeta)^k$ lower bound.



Synthetic bits method with repeated squaring

Needs $n = 3k - 3$ instead of $n = 2k - 1$. Use disjoint sets A, B, B' each with $k - 1$ ζ-separated rows.

After these $\lg k$ levels, errors blow up by $\leq (1/\zeta)^{k \lg k}$.
$\Rightarrow$ sample size matches (almost) the $(1/\zeta)^k$ lower bound. Proves Thm 2 (k-MixProd analysis).



Onwards:

1. "Learn" k-MixProd in Weierstrass (transportation) distance in time similar to identification? (Do have such results for k-MixIID.) I.e., $\sim \exp(k \lg k)$ rather than $\exp(k^3)$?

2. Parametric models.

3. Multiple cardinality- or dimension-bounded confounders.

Onwards:

1. "Learn" k-MixProd in Weierstrass (transportation) distance in time similar to identification? (Do have such results for k-MixIID.) I.e., $\sim \exp(k \lg k)$ rather than $\exp(k^3)$?
2. Parametric models.
3. Multiple cardinality- or dimension-bounded confounders.