Estimating Latent-Variable Graphical Models using Moments and Likelihoods

Arun Tejasvi Chaganty CHAGANTY@CS.STANFORD.EDU
Percy Liang PLIANG@CS.STANFORD.EDU
Stanford University, Stanford, CA, USA

Abstract

Recent work on the method of moments enable consistent parameter estimation, but only for certain types of latent-variable models. On the other hand, pure likelihood objectives, though more universally applicable, are difficult to optimize. In this work, we show that using the method of moments in conjunction with composite likelihood yields consistent parameter estimates for a much broader class of discrete directed and undirected graphical models, including loopy graphs with high treewidth. Specifically, we use tensor factorization to reveal information about the hidden variables. This allows us to construct convex likelihoods which can be globally optimized to recover the parameters.

1. Introduction

Latent-variable graphical models provide compact representations of data and have been employed across many fields (Ghahramani & Beal, 1999; Jaakkola & Jordan, 1999; Blei et al., 2003; Quattoni et al., 2004; Haghighi & Klein, 2006). However, learning these models remains a difficult problem due to the non-convexity of the negative log-likelihood. Local methods such as expectation maximization (EM) are the norm, but are susceptible to local optima.

Recently, unsupervised learning techniques based on the spectral method of moments have offered a refreshing perspective on this learning problem (Mossel & Roch, 2005; Hsu et al., 2009; Bailly et al., 2010; Song et al., 2011; Anandkumar et al., 2011; 2012b;a; Hsu et al., 2012; Balle & Mohri, 2012). These methods exploit the linear algebraic properties of the model to factorize moments of the observed data distribution into parameters, providing strong theoretical guarantees. However, they apply to a limited set of models, and are thus not as broadly applicable as EM.

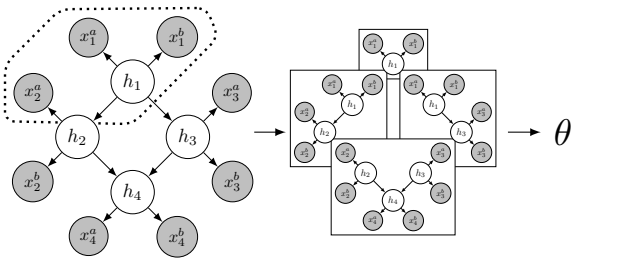


Figure 1. Overview of our approach: (i) we use tensor factorization to learn the *conditional moments* for each hidden variable; (ii) we optimize a composite likelihood to recover the hidden *marginals*; and (iii) we optimize another likelihood objective to the *model parameters*. Both likelihood objectives are convex.

In this paper, we show that a much broader class of discrete directed and undirected graphical models can be consistently estimated: specifically those in which *each* hidden variable has three conditionally independent observed variables ("views"). Our key idea is to leverage the method of moments, not to directly provide a consistent parameter estimate as in previous work, but as constraints on a likelihood-based objective. Notably, our method applies to latent undirected log-linear models with high treewidth.

The essence of our approach is illustrated in Figure 1, which contains three steps. First, we identify three views for each hidden variable h_i (for example, x_1^a, x_1^b and x_3^a are conditionally independent given h_1) and use the tensor factorization algorithm of Anandkumar et al. (2013) to estimate the *conditional moments* $\mathbb{P}(x_i^a \mid h_i)$ and $\mathbb{P}(x_i^b \mid h_i)$ for each i (Section 3). Second, we optimize a *composite marginal likelihood* to recover the marginals over subsets of hidden nodes (e.g., $\mathbb{P}(h_2, h_3, h_4)$). Normally, such a marginal likelihood objective would be non-convex, but given the conditional moments, we obtain a convex objective, which can be globally optimized using EM (see Sections 4 and 4.2). So far, our method has relied only on the conditional independence structure of the model and applies generically to both directed and undirected models.

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

The final step of turning hidden marginals into model parameters requires some specialization. In the directed case, this is simple normalization; in the undirected case, we need to solve another convex optimization problem (Section 5).

2. Setup

Let $\mathcal{G}$ be a discrete graphical model with observed variables $\mathbf{x} = (x_1, \ldots, x_L)$ and hidden variables $\mathbf{h} = (h_1, \ldots, h_M)$. We assume that the domains of the variables are $x_v \in [d]$ for all $v \in [L]$ and $h_i \in [k]$ for all $i \in [M]$, where $[n] = \{1, \ldots, n\}$. Let $\mathcal{X} \triangleq [d]^L$ and $\mathcal{H} \triangleq [k]^M$ be the joint domains of $\mathbf{x}$ and $\mathbf{h}$, respectively.

For undirected models $\mathcal{G}$, let $\mathcal{G}$ denote a set of cliques, where each clique $\mathcal{C} \subseteq \mathbf{x} \cup \mathbf{h}$ is a subset of nodes. The joint distribution is given by an exponential family: $p_\theta(\mathbf{x}, \mathbf{h}) \propto \prod_{\mathcal{C} \in \mathcal{G}} \exp(\theta^\top \phi_{\mathcal{C}}(\mathbf{x}_{\mathcal{C}}, \mathbf{h}_{\mathcal{C}}))$, where θ is the parameter vector, and $\phi_{\mathcal{C}}(\mathbf{x}_{\mathcal{C}}, \mathbf{h}_{\mathcal{C}})$ is the local feature vector which only depends on the observed ($\mathbf{x}_{\mathcal{C}}$) and hidden ($\mathbf{h}_{\mathcal{C}}$) variables in clique $\mathcal{C}$. Also define $\mathcal{N}(a) = \{b \neq a : \exists \mathcal{C} \supseteq \{a, b\}\}$ to be the neighbors of variable a.

For directed models $\mathcal{G}$, define $p_\theta(\mathbf{x}, \mathbf{h}) = \prod_{a \in \mathbf{x} \cup \mathbf{h}} p_\theta(a \mid \text{Pa}(a))$, where $\text{Pa}(a) \subseteq \mathbf{x} \cup \mathbf{h}$ are the parents of a variable a. The parameters θ are the conditional probability tables of each variable, and the cliques are $\mathcal{G} = \{\{a\} \cup \text{Pa}(a) : a \in \mathbf{x} \cup \mathbf{h}\}$.

Problem statement This paper focuses on the problem of parameter estimation: We are given n i.i.d. examples of the observed variables $\mathcal{D} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})$, where each $\mathbf{x}^{(i)} \sim p_{\theta^*}$ for some true parameters θ^*. Our goal is to produce a parameter estimate $\hat{\theta}$ that approximates θ^*.

The standard estimation procedure is maximum likelihood:

$$L_{\text{unsup}}(\theta) \triangleq \sum_{\mathbf{x} \in \mathcal{D}} \log p_\theta(\mathbf{x}) = \sum_{\mathbf{x} \in \mathcal{D}} \log \sum_{\mathbf{h} \in \mathcal{H}} p_\theta(\mathbf{x}, \mathbf{h}). \quad (1)$$

Maximum likelihood is statistically efficient, but in general computationally intractable because marginalizing over hidden variables $\mathbf{h}$ yields a non-convex objective. In practice, one uses local optimization procedures (e.g., EM or L-BFGS) on the marginal likelihood, but these can get stuck in local optima. We will later return to likelihoods, but let us first describe a method of moments approach for parameter estimation. To do this, let's introduce some notation.

Notation We use the notation $[\cdot]$ to indicate indexing; for example, $M[i]$ is the i-th row of a matrix M and $M[i, j]$ is the (i, j)-th element of M. For a tensor $T \in \mathbb{R}^{d \times \cdots \times d}$ and a vector $\mathbf{i} = (i_1, \ldots, i_\ell)$, define the projection $T[\mathbf{i}] = T[i_1, \ldots, i_\ell]$.

We use $\otimes$ to denote the tensor product: if $u \in \mathbb{R}^d, v \in \mathbb{R}^k$, then $u \otimes v \in \mathbb{R}^{d \times k}$. For an ℓ-th order tensor $T \in \mathbb{R}^{d \times \ldots \times d}$ and vectors $v_1, \cdots, v_\ell \in \mathbb{R}^d$, define the application:

$$T(v_1, \cdots, v_\ell) = \sum_{\mathbf{i}} T[\mathbf{i}] v_1[i_1] \cdots v_\ell[i_\ell].$$

Analogously, for matrices $M_1 \in \mathbb{R}^{d \times k}, \cdots, M_\ell \in \mathbb{R}^{d \times k}$:

$$T(M_1, \cdots, M_\ell)[\mathbf{j}] = \sum_{\mathbf{i}} T[\mathbf{i}] M_1[i_1, j_1] \cdots M_\ell[i_\ell, j_\ell].$$

We will use $\mathbb{P}(\cdot)$ to denote various moment tensors constructed from the true data distribution $p_{\theta^*}(\mathbf{x}, \mathbf{h})$:

$$M_i \triangleq \mathbb{P}(x_i), \quad M_{ij} \triangleq \mathbb{P}(x_i, x_j), \quad M_{ijk} \triangleq \mathbb{P}(x_i, x_j, x_k).$$

Here, M_i, M_{ij}, M_{ijk} are tensors of orders $1, 2, 3$ in $\mathbb{R}^d, \mathbb{R}^{d \times d}, \mathbb{R}^{d \times d \times d}$. Next, we define the *hidden marginals*:

$$Z_i \triangleq \mathbb{P}(h_i), \quad Z_{ij} \triangleq \mathbb{P}(h_i, h_j), \quad Z_{ijk} \triangleq \mathbb{P}(h_i, h_j, h_k).$$

These are tensors of orders $1, 2, 3$ in $\mathbb{R}^k, \mathbb{R}^{k \times k}, \mathbb{R}^{k \times k \times k}$. Finally, we define *conditional moments* $O^{(v|i)} \triangleq \mathbb{P}(x_v \mid h_i) \in \mathbb{R}^{d \times k}$ for each $v \in [L]$ and $i \in [M]$.

2.1. Assumptions

In this section, we state technical assumptions that hold for the rest of the paper, but that we feel are not central to our main ideas. The first one ensures that all realizations of each hidden variable are possible:

Assumption 1 (Non-degeneracy)**.** The marginal distribution of each hidden variable h_i has full support: $\mathbb{P}(h_i) \succ 0$.

Next, we assume the graphical model only has conditional independences given by the graph:

Assumption 2 (Faithful)**.** For any hidden variables $a, b, c \in \mathbf{h}$ such that an active trail[1] connects a and b conditioned on c, we have that a and b are dependent given c.

Finally, we assume the graphical model is in a canonical form in which all observed variables are leaves:

Assumption 3 (Canonical form)**.** For each observed variable x_v, there exists exactly one $\mathcal{C} \in \mathcal{G}$ such that $\mathcal{C} = \{x_v, h_i\}$ for some hidden node h_i.

The following lemma shows that this is not a real assumption (see the appendix for the proof):

Lemma 1 (Reduction to canonical form)**.** *Every graphical model can be transformed into canonical form. There is a one-to-one correspondence between the parameters of the transformed and original models.*

Finally, for clarity, we will derive our algorithms using exact moments of the true distribution p_{θ^*}. In practice, we would use moments estimated from data $\mathcal{D}$.

[1] See Koller & Friedman (2009) for a definition. We do not condition on observed variables.

3. Bottlenecks

We start by trying to reveal some information about the hidden variables that will be used by subsequent sections. Specifically, we review how the tensor factorization method of Anandkumar et al. (2013) can be used to recover the conditional moments $O^{(v|i)} \triangleq \mathbb{P}(x_v \mid h_i)$. The key notion is that of a bottleneck:

Definition 1 (Bottleneck)**.** A hidden variable h_i is said to be a *bottleneck* if (i) there exists three observed variables (views), $x_{v_1}, x_{v_2}, x_{v_3}$, that are conditionally independent given h_i (Figure 2(a)), and (ii) each $O^{(v|i)} \triangleq \mathbb{P}(x_v \mid h_i) \in \mathbb{R}^{d \times k}$ has full column rank k for each $v \in \{v_1, v_2, v_3\}$. We say that a subset of hidden variables $S \subseteq \mathbf{h}$ is bottlenecked if every $h \in S$ is a bottleneck. We say that a graphical model $\mathcal{G}$ is bottlenecked if all its hidden variables are bottlenecks.

For example, in Figure 1, x_1^a, x_1^b, x_2^a are views of the bottleneck h_1, and x_2^a, x_2^b, x_1^b are views of the bottleneck h_2. Therefore, the clique $\{h_1, h_2\}$ is bottlenecked. Note that views are allowed to overlap.

The full rank assumption on the conditional moments $O^{(v|i)} = \mathbb{P}(x_v \mid h_i)$ ensures that all states of h_i "behave differently." In particular, the conditional distribution of one state cannot be a mixture of that of other states.

Anandkumar et al. (2012a) provide an efficient tensor factorization algorithm for estimating $\mathbb{P}(x_v \mid h_i)$:

Theorem 1 (Tensor factorization)**.** *Let $h_i \in \mathbf{h}$ be a bottleneck with views $x_{v_1}, x_{v_2}, x_{v_3}$. Then there exists an algorithm* GETCONDITIONALS *that returns consistent estimates of $O^{(v|i)}$ for each $v \in \{v_1, v_2, v_3\}$ up to relabeling of the hidden variables.*

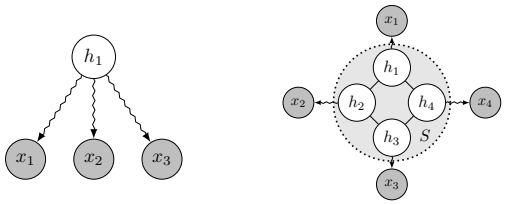
To simplify notation, consider the example in Figure 2(a) where $h_1 = 1, v_1 = 1, v_2 = 2, v_3 = 3$. The observed moments M_{12}, M_{23}, M_{13} and M_{123} can be factorized as follows:

$$M_{vv'} = \sum_h \pi^{(1)}[h] O^{(v|1)\top}[h] \otimes O^{(v'|1)\top}[h]$$

$$M_{123} = \sum_h \pi^{(1)}[h] O^{(1|1)\top}[h] \otimes O^{(2|1)\top}[h] \otimes O^{(3|1)\top}[h].$$

The GETCONDITIONALS algorithm first computes a whitening matrix $W \in \mathbb{R}^{d \times k}$ such that $W^\top M_{12} W = I_{k \times k}$, and uses W to transform M_{123} into a symmetric orthogonal tensor. Then a robust tensor power method is used to extract the eigenvectors of the whitened M_{123}; unwhitening yields the columns of $O^{(3|1)}$ (up to permutation). The other conditional moments can be recovered similarly.

The resulting estimate of $O^{(v|i)}$ based on n data points converges at a rate of $n^{-\frac{1}{2}}$ with a constant that depends polynomially on $\sigma_k(O^{(v|i)})^{-1}$, the inverse of the k-th largest singular value of $O^{(v|i)}$. Note that $\sigma_k(O^{(v|i)})$ can become quite small if h_i and x_v are connected via many intermediate hidden variables.[2]



(a) Bottleneck (b) Exclusive views

Figure 2. (a) A bottleneck h_1 has three conditionally independent views x_1, x_2, x_3. (b) A bidependent subset S has exclusive views $\{x_1, x_2, x_3, x_4\}$.

The tensor factorization method attacks the heart of the non-convexity in latent-variable models, providing some information about the hidden variables in the form of the conditional moments $O^{(v|i)} = \mathbb{P}(x_v \mid h_i)$. Note that GETCONDITIONALS only examines the conditional independence structure of the graphical model, not its parametrization.

If i is the single parent of v (e.g., $\mathbb{P}(x_1^a \mid h_1)$ in Figure 1), then this conditional moment is a parameter of the model, but this is in general not the case (e.g., $\mathbb{P}(x_2^a \mid h_1)$). Furthermore, there are other parameters (e.g., $\mathbb{P}(h_4 \mid h_2, h_3)$) which we do not have a handle on yet. In general, there is a gap between the conditional moments and the model parameters, which we will address in the next two sections.

4. Recovering hidden marginals

Having recovered conditional moments $O^{(v|i)} \triangleq \mathbb{P}(x_v \mid h_i)$, we now seek to compute the marginal distribution of sets of hidden variables $Z_S \triangleq \mathbb{P}(\mathbf{h}_S)$.

Example To gain some intuition, consider the directed grid model from Figure 1. We can express the observed marginals $M_{12} \triangleq \mathbb{P}(x_1^a, x_2^a) \in \mathbb{R}^{d \times d}$ as a linear function of the hidden marginals $Z_{12} \triangleq \mathbb{P}(h_1, h_2) \in \mathbb{R}^{k \times k}$, where the linear coefficients are based on the conditional moments $O^{(1|1)}, O^{(2|2)} \in \mathbb{R}^{d \times k}$:

$$M_{12} = O^{(1|1)} Z_{12} O^{(2|2)\top}.$$

We can then solve for Z_{12} by matrix inversion:

$$Z_{12} = O^{(1|1)\dagger} M_{12} O^{(2|2)\dagger\top}.$$

[2] To see this, suppose h_1 has a view x_v via a chain: $h_1 - h_2 \cdots - h_t - x_v$. In this example, if $\sigma_k(\mathbb{P}(h_{i+1} \mid h_i)) = a_k$ for each $i = 1, \cdots, t-1$, then $\sigma_k(O^{(v|1)}) = a_k^t \sigma_k(O^{(v|t)})$.

4.1. Exclusive views

For which subsets of hidden nodes can we recover the marginals? The following definition offers a characterization:

Definition 2 (Exclusive views)**.** Let $S \subseteq \mathbf{h}$ be a subset of hidden variables. We say $h_i \in S$ has an exclusive view x_v if the two conditions hold: (i) there exists some observed variable x_v which is conditionally independent of the others $S \backslash \{h_i\}$ given h_i (Figure 2(b)), and (ii) the conditional moment matrix $O^{(v|i)} \triangleq \mathbb{P}(x_v \mid h_i)$ has full column rank k and can be recovered. We say that S has the *exclusive views property* if every $h_i \in S$ has an exclusive view.

Estimating hidden marginals We now show that if a subset of hidden variables S has the exclusive views property, then we can recover the marginal distribution $\mathbb{P}(\mathbf{h}_S)$. Consider any $S = \{h_{i_1}, \dots, h_{i_m}\}$ with the exclusive views property. Let x_{v_j} be an exclusive view for h_{i_j} in S and define $\mathcal{V} = \{x_{v_1}, \dots, x_{v_m}\}$. By the exclusive views property, the marginal over the observed variables $\mathbb{P}(\mathbf{x}_\mathcal{V})$ factorizes according to the marginal over the hidden variables $\mathbb{P}(\mathbf{h}_S)$ times the conditional moments:

$$\begin{aligned} M_\mathcal{V} &\triangleq \mathbb{P}(\mathbf{x}_\mathcal{V}) \\ &= \sum_{\mathbf{h}_S} \mathbb{P}(\mathbf{h}_S)\, \mathbb{P}(x_{v_1} \mid h_{i_1}) \cdots \mathbb{P}(x_{v_m} \mid h_{i_m}) \\ &= Z_S(O^{(v_1|i_1)}, \dots, O^{(v_m|i_m)}) \\ &= Z_S(\mathbf{O}), \end{aligned}$$

where $\mathbf{O} = O^{(v_1|i_1)} \otimes \cdots \otimes O^{(v_m|i_m)}$ is the tensor product of all the conditional moments. Vectorizing, we have that $Z_S \in \mathbb{R}^{k^m}$, $M_\mathcal{V} \in \mathbb{R}^{d^m}$, and $\mathbf{O} \in \mathbb{R}^{d^m \times k^m}$. Since each $O^{(v|i)}$ has full column rank k, the tensor product $\mathbf{O}$ has full column rank k^m. Succinctly, $M_\mathcal{V}$ (which can be estimated directly from data) is a linear function of Z_S (what we seek to recover). We can solve for the hidden marginals Z_S simply by multiplying $M_\mathcal{V}$ by the pseudoinverse of $\mathbf{O}$:

$$Z_S = M_\mathcal{V}(O^{(v_1|i_1)^\dagger}, \cdots, O^{(v_m|i_m)^\dagger}).$$

Algorithm 1 summarizes the procedure, GETMARGINALS. Given Z_S, the conditional probability tables for S can easily be obtained via renormalization.

Theorem 2 (Hidden marginals from exclusive views)**.** *If $S \subseteq \mathbf{x}$ is a subset of hidden variables with the exclusive views property, then Algorithm 1 recovers the marginals $Z_S = \mathbb{P}(\mathbf{h}_S)$ up to a global relabeling of the hidden variables determined by the labeling from* GETCONDITIONALS.

Relationship to bottlenecks The bottleneck property allows recovery of conditional moments, and the exclusive views property allows recovery of hidden marginals. But we will now show that the latter property is in fact implied by the former property for special sets of hidden variables, which we call bidependent sets (in analogy with biconnected components), in which conditioning on one variable does not break the set apart:

Algorithm 1 GETMARGINALS (pseudoinverse)

Input: Hidden subset $S = \{h_{i_1}, \dots, h_{i_m}\}$ with exclusive views $\mathcal{V} = \{x_{v_1}, \dots, x_{v_m}\}$ and conditional moments $O^{(v_j|i_j)} = \mathbb{P}(x_{v_j} \mid h_{i_j})$.

Output: Marginals $Z_S = \mathbb{P}(\mathbf{h}_S)$.

Return $Z_S \leftarrow M_\mathcal{V}(O^{(v_1|i_1)^\dagger}, \dots, O^{(v_m|i_m)^\dagger})$.

Definition 3 (Bidependent set)**.** We say that a subset of nodes S is *bidependent* if conditioned on any $a \in S$, there is an active trail between any other two nodes $b, c \in S$.

Note that all cliques are bidependent, but bidependent sets can have more conditional independences (e.g., $\{h_1, h_2, h_3\}$ in Figure 2(b)). This will be important in Section 5.1.

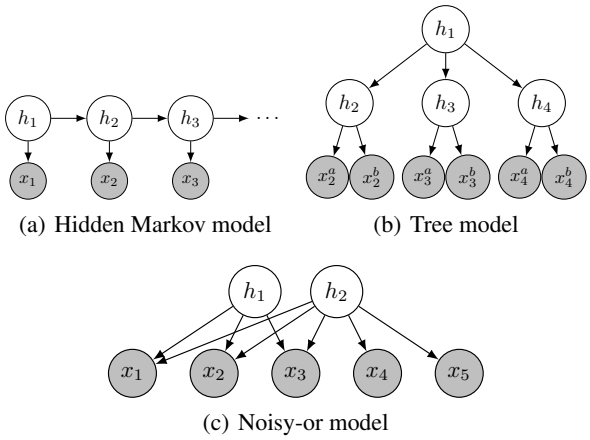
Bidependent sets are significant because they guarantee exclusive views if they are bottlenecked:

Lemma 2 (Bottlenecked implies exclusive views)**.** *Let $S \subseteq \mathbf{h}$ be a bidependent subset of hidden variables. If S is bottlenecked, then S has the exclusive views property.*

Proof. Let S be a bidependent subset and fix any $h_0 \in S$. Since h_0 is a bottleneck, it has three conditionally independent views, say x_1, x_2, x_3 without loss of generality. For condition (i), we will show that at least one of the views is conditionally independent of $S \backslash \{h_0\}$ given h_0. For the sake of contradiction, suppose that each observed variable x_i is conditionally dependent on some $h_i \in S \backslash \{h_0\}$ given h_0, for $i \in \{1, 2, 3\}$. Then conditioned on h_0, there is an active trail between h_1 and h_2 because S is biconnected. This means there is also an active trail $x_1 - h_1 - h_2 - x_2$ conditioned on h_0. Since the graphical model is faithful by assumption, we have $x_1 \not\perp x_2 \mid h_0$, contradicting the fact that x_1 and x_2 are conditionally independent given h_0. To show condition (ii), assume, without loss of generality, that x_1 is an exclusive view. Then we can recover $O^{(1|0)} = \mathbb{P}(x_1 \mid h_0)$ via GETCONDITIONALS. □

Remarks. Note that having only two independent views for each $h_i \in S$ is sufficient for condition (i) of the exclusive views property, while three is needed for condition (ii). The bottleneck property (Definition 1) can also be relaxed if some cliques share parameters (see examples below).

Our method extends naturally to the case in which the observed variables are real-valued ($x_v \in \mathbb{R}^d$), as long as the



(a) Hidden Markov model (b) Tree model

(c) Noisy-or model

Figure 3. (a) and (b): graphical models that satisfy the exclusive views property; (c) a graphical model that does not.

hidden variables remain discrete. In this setting, the conditional moments $O^{(v|i)} \triangleq \mathbb{E}(x_v \mid h_i) \in \mathbb{R}^{d \times k}$ would no longer be distributions but general rank k matrices.

Example: hidden Markov model. In the HMM (Figure 3(a)), h_2 is a bottleneck, so we can recover $O \triangleq \mathbb{P}(x_2 \mid h_2)$. While the first hidden variable h_1 is not a bottleneck, it still has an exclusive view x_1 with respect to the clique $\{h_1, h_2\}$, assuming parameter sharing across emissions ($\mathbb{P}(x_1 \mid h_1) = O$).

Example: latent tree model. In the latent tree model (Figure 3(b)), h_1 is not directly connected to an observed variable, but it is still a bottleneck, with views x_2^a, x_3^a, x_4^a, for example. The clique $\{h_1, h_2\}$ has exclusive views $\{x_2^a, x_3^a\}$.

Non-example In Figure 3(c), h_1 does not have exclusive views. Without parameter sharing, the techniques in this paper are insufficient. In the special case where the graphical model represents a binary-valued noisy-or network, we can use the algorithm of Halpern & Sontag (2013), which first learns h_2 and subtracts off its influence, thereby making h_1 a bottleneck.

4.2. Composite likelihood

So far, we have provided a method of moments estimator which used (i) tensor decomposition to recover conditional moments and (ii) matrix pseudoinversion to recover the hidden marginals. We will now improve statistical efficiency by replacing (ii) with a convex likelihood-based objective.

Of course, optimizing the original marginal likelihood (Equation 1) is subject to local optima. However, we make two changes to circumvent non-convexity: The first is that we already have the conditional moments from tensor decomposition, so effectively a subset of the parameters are fixed. However, this alone is not enough, for the full likelihood is still non-convex. The second change is that we will optimize a *composite likelihood objective* (Lindsay, 1988) rather than the full likelihood.

Consider a subset of hidden nodes $S = \{h_{i_1}, \ldots, h_{i_m}\}$, with exclusive views $\mathcal{V} = \{x_{v_1}, \ldots, x_{v_m}\}$. The expected composite log-likelihood over $\mathbf{x}_\mathcal{V}$ given parameters $Z_S \triangleq \mathbb{P}(\mathbf{h}_S)$ with respect to the true distribution $\mathcal{M}_\mathcal{V}$ can be written as follows:

$$\begin{aligned}\mathcal{L}_{\text{cl}}(Z_S) &\triangleq \mathbb{E}[\log \mathbb{P}(\mathbf{x}_\mathcal{V})] \\ &= \mathbb{E}[\log \sum_{\mathbf{h}_S} \mathbb{P}(\mathbf{h}_S)\, \mathbb{P}(\mathbf{x}_\mathcal{V} \mid \mathbf{h}_S)] \\ &= \mathbb{E}[\log Z_S(O^{(v_1|i_1)}[x_{v_1}], \cdots, O^{(v_m|i_m)}[x_{v_m}])] \\ &= \mathbb{E}[\log Z_S(\mathbf{O}[\mathbf{x}_\mathcal{V}])]. \end{aligned} \tag{2}$$

The final expression is an expectation over the log of a linear function of Z_S, which is concave in Z_S. Unlike maximum likelihood in fully-observed settings, we do not have a closed-form solution, so we use EM to optimize it. However, since the function is concave, EM is guaranteed to converge to the *global* maximum. Algorithm 2 summarizes our algorithm.

Algorithm 2 GETMARGINALS (composite likelihood)

Input: Hidden subset $S = \{h_{i_1}, \ldots, h_{i_m}\}$ with exclusive views $\mathcal{V} = \{x_{v_1}, \ldots, x_{v_m}\}$ and conditional moments $O^{(v_j|i_j)} = \mathbb{P}(x_{v_j} \mid h_{i_j})$.

Output: Marginals $Z_S = \mathbb{P}(\mathbf{h}_S)$.

Return $Z_S = \arg\max_{Z_S \in \Delta_{k^m - 1}} \mathbb{E}[\log Z_S(\mathbf{O}[\mathbf{x}_\mathcal{V}])]$.

4.3. Statistical efficiency

We have proposed two methods for estimating the hidden marginals Z_S given the conditional moments $\mathbf{O}$, one based on computing a simple pseudoinverse, and the other based on composite likelihood. Let $\hat{Z}_S^{\text{pi}}$ denote the pseudoinverse estimator and $\hat{Z}_S^{\text{cl}}$ denote the composite likelihood estimator.[3]

The Cramér-Rao lower bound tells us that maximum likelihood yields the most statistically efficient composite estimator for Z_S given access to only samples of $\mathbf{x}_\mathcal{V}$.[4] Let us go one step further and quantify the *relative efficiency*

[3] For simplicity, assume that $\mathbf{O}$ is known. In practice, $\mathbf{O}$ would be estimated via tensor factorization.

[4] Of course, we could improve statistical efficiency by maximizing the likelihood of all of $\mathbf{x}$, but this would lead to a non-convex optimization problem.

of the pseudoinverse estimator compared to the composite likelihood estimator.

Abusing notation slightly, think of $M_{\mathcal{V}}$ as just a flat multinomial over d^m outcomes and Z_S as a multinomial over k^m outcomes, where the two are related by $\mathbf{O} \in \mathbb{R}^{d^m \times k^m}$. We will not need to access the internal tensor structure of $M_{\mathcal{V}}$ and Z_S, so to simplify the notation, let $m = 1$ and define $\mu = M_{\mathcal{V}} \in \mathbb{R}^d$, $z = Z_S \in \mathbb{R}^k$, and $O = \mathbf{O} \in \mathbb{R}^{d \times k}$. The hidden marginals z and observed marginals μ are related via $\mu = Oz$.

Note that z and μ are constrained to lie on simplexes Δ_{k-1} and Δ_{d-1}, respectively. To avoid constraints, we reparameterize z and μ using $\widetilde{z} \in \mathbb{R}^{k-1}$ and $\widetilde{\mu} \in \mathbb{R}^{d-1}$:

$$\mu = \begin{bmatrix} \widetilde{\mu} \\ 1 - \mathbf{1}^\top \widetilde{\mu} \end{bmatrix} \qquad z = \begin{bmatrix} \widetilde{z} \\ 1 - \mathbf{1}^\top \widetilde{z} \end{bmatrix}.$$

In this representation, $\widetilde{\mu}$ and $\widetilde{z}$ are related as follows,

$$\begin{bmatrix} \widetilde{\mu} \\ 1 - \mathbf{1}^\top \widetilde{\mu} \end{bmatrix} = \begin{bmatrix} O_{\neg d, \neg k} & O_{\neg d, k} \\ O_{d, \neg k} & O_{d,k} \end{bmatrix} \begin{bmatrix} \widetilde{z} \\ 1 - \mathbf{1}^\top \widetilde{z} \end{bmatrix}$$
$$\widetilde{\mu} = \underbrace{(O_{\neg d, \neg k} - O_{\neg d, k} \mathbf{1}^\top)}_{\triangleq \widetilde{O}} \widetilde{z} + O_{\neg d, k}.$$

The pseudoinverse estimator is defined as $\widehat{\widetilde{z}}^{\text{pi}} = \widetilde{O}^\dagger(\widehat{\widetilde{\mu}} - O_{\neg d,k})$, and the composite likelihood estimator is given by $\widehat{\widetilde{z}}^{\text{cl}} = \arg\max_{\widetilde{z}} \hat{\mathbb{E}}[\ell(x; \widetilde{z})]$, where $\ell(x; \widetilde{z}) = \log(\mu[x])$ is the log-likelihood function.

First, we compute the asymptotic variances of the two estimators.

Lemma 3 (Asymptotic variances)**.** *The asymptotic variances of the pseudoinverse estimator $\widehat{\widetilde{z}}^{\text{pi}}$ and composite likelihood estimator $\widehat{\widetilde{z}}^{\text{cl}}$ are:*

$$\Sigma^{\text{pi}} = \widetilde{O}^\dagger(\widetilde{D} - \widetilde{\mu}\widetilde{\mu}^\top)\widetilde{O}^{\dagger\top},$$
$$\Sigma^{\text{cl}} = \left(\widetilde{O}^\top(\widetilde{D}^{-1} + \widetilde{d}^{-1}\mathbf{1}\mathbf{1}^\top)\widetilde{O}\right)^{-1},$$

where $\widetilde{D} \triangleq \operatorname{diag}(\widetilde{\mu})$ and $\widetilde{d} \triangleq 1 - \mathbf{1}^\top \widetilde{\mu}$.

Next, let us compare the relative efficiencies of the two estimators: $e^{\text{pi}} \triangleq \frac{1}{k-1} \operatorname{tr}(\Sigma^{\text{cl}}(\Sigma^{\text{pi}})^{-1})$. From the Cramér-Rao bound (van der Vaart, 1998), we know that $\Sigma^{\text{cl}} \preceq \Sigma^{\text{pi}}$. This implies that the relative efficiency, e^{pi}, lies between 0 and 1, and when $e^{\text{pi}} = 1$, the pseudoinverse estimator is said to be (asymptotically) efficient. To gain intuition, let us explore two special cases:

Lemma 4 (Relative efficiency when $\widetilde{O}$ is invertible)**.** *When $\widetilde{O}$ is invertible, the asymptotic variances of the pseudoinverse and composite likelihood estimators are equal, $\Sigma^{\text{cl}} = \Sigma^{\text{pi}}$, and the relative efficiency is 1.*

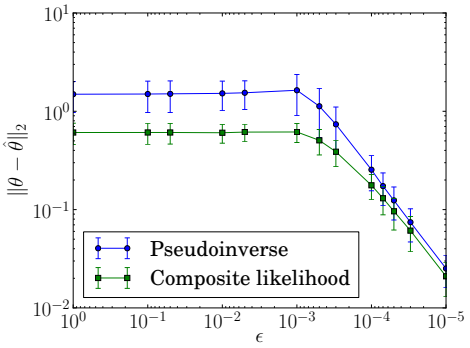


Figure 4. Comparison of parameter estimation error ($\|\hat{\theta} - \theta\|_2$) versus error in moments (ϵ) for a hidden Markov model with $k = 2$ hidden and $d = 5$ observed values. Empirical moments $\widehat{M}_{123}$ were generated by adding Gaussian noise, $\mathcal{N}(0, \epsilon I)$, to expected moments M_{123}. Results are averaged over 400 trials.

Lemma 5 (Relative efficiency with uniform observed marginals)**.** *Let the observed marginals μ be uniform: $\mu = \frac{1}{d}\mathbf{1}$. The efficiency of the pseudoinverse estimator is:*

$$e^{\text{pi}} = 1 - \frac{1}{k-1} \frac{\|\mathbf{1}_U\|^2}{1 + \|\mathbf{1}_U\|^2} \left(1 - \frac{1}{d - \|\mathbf{1}_U\|^2}\right), \quad (3)$$

where $\mathbf{1}_U \triangleq \widetilde{O}\widetilde{O}^\dagger \mathbf{1}$, the projection of $\mathbf{1}$ onto the column space of $\widetilde{O}$. Note that $0 \le \|\mathbf{1}_U\|_2^2 \le k - 1$.

When $\|\mathbf{1}_U\|_2 = 0$, the pseudoinverse estimator is efficient: $e^{\text{pi}} = 1$. When $\|\mathbf{1}_U\|_2 > 0$ and $d > k$, the pseudoinverse estimator is strictly inefficient. In particular, if $\|\mathbf{1}_U\|_2^2 = k - 1$, and we get:

$$e^{\text{pi}} = 1 - \frac{1}{k}\left(1 - \frac{1}{1 + d - k}\right). \quad (4)$$

Based on Equation 3 and Equation 4, we see that the pseudoinverse gets progressively worse compared to the composite likelihood as the gap between k and d increases for the special case wherein the observed moments are uniformly distributed. For instance, when $k = 2$ and $d \to \infty$, the efficency of the pseudolikelihood estimator is half that of the composite likelihood estimator. Empirically, we observe that the composite likelihood estimator also leads to more accurate estimates in general non-asymptotic regimes (see Figure 4).

5. Recovering parameters

We have thus far shown how to recover the conditional moments $O^{(v|i)} = \mathbb{P}(x_v \mid h_i)$ for each exclusive view x_v of each hidden variable h_i, as well as the hidden marginals $Z_S = \mathbb{P}(\mathbf{h}_S)$ for each bidependent subset of hidden variables S. Now all that remains to be done is to recover the parameters.

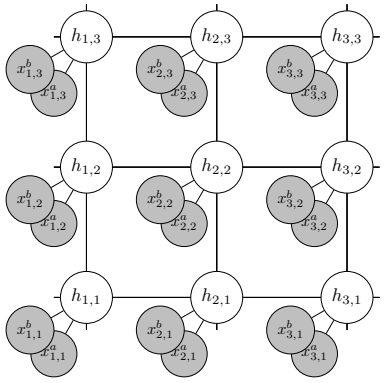


Figure 5. Example: undirected grid model where each hidden variable has two conditionally independent observations. This model has high treewidth, but we can estimate it efficiently using pseudolikelihood.

Since our graphical model is in canonical form (Assumption 3), all cliques $\mathcal{C} \in \mathcal{G}$ either consist of hidden variables $\mathbf{h}_\mathcal{C}$ or are of the form $\{x_v, h_i\}$. The key observation is that the clique marginals are actually sufficient statistics of the model p_θ. How we turn these clique marginals $\{\mathbb{P}(\mathbf{x}_\mathcal{C}, \mathbf{h}_\mathcal{C})\}_{\mathcal{C}\in\mathcal{G}}$ into parameters θ depends on the exact model parametrization.

For directed models, the parameters are simply the local conditional tables $p_\theta(a \mid \text{Pa}(a))$ for each clique $\mathcal{C} = \{a\} \cup \text{Pa}(a)$. These conditional distributions can be obtained by simply normalizing $Z_\mathcal{C}$ for each assignment of $\text{Pa}(a)$.

For undirected log-linear models, the canonical parameters θ cannot be obtained locally, but we can construct a global convex optimization problem to solve for θ. Suppose we were able to observe $\mathbf{h}$. Then we could optimize the *supervised* likelihood, which is concave:

$$\begin{aligned} L_{\text{sup}}(\theta) &\triangleq \mathbb{E}_{(\mathbf{x},\mathbf{h})\sim p_{\theta^*}}[\log p_\theta(\mathbf{x}, \mathbf{h})] \\ &= \theta^\top \left(\sum_{\mathcal{C}\in\mathcal{G}} \mathbb{E}[\phi(\mathbf{x}_\mathcal{C}, \mathbf{h}_\mathcal{C})] \right) - A(\theta). \end{aligned} \quad (5)$$

Of course we don't have supervised data, but we do have the marginals $\mathbb{P}(\mathbf{x}_\mathcal{C}, \mathbf{h}_\mathcal{C})$, from which we can easily compute the expected features:

$$\mu_\mathcal{C} \triangleq \mathbb{E}[\phi(\mathbf{x}_\mathcal{C}, \mathbf{h}_\mathcal{C})] = \sum_{\mathbf{x}_\mathcal{C}, \mathbf{h}_\mathcal{C}} \mathbb{P}(\mathbf{x}_\mathcal{C}, \mathbf{h}_\mathcal{C}) \phi(\mathbf{x}_\mathcal{C}, \mathbf{h}_\mathcal{C}). \quad (6)$$

Therefore, we can optimize the supervised likelihood objective without actually having any supervised data! In the finite data regime, the method of moments yields the estimate $\hat{\mu}^{\text{mom}}_\mathcal{C}$ which approximates the true $\mu_\mathcal{C}$. In supervised learning, we obtain a different estimate $\hat{\mu}^{\text{sup}}_\mathcal{C}$ of $\mu_\mathcal{C}$ based on an empirical average over data points. In the limit of infinite data, both estimators converge to $\mu_\mathcal{C}$.

Algorithm 3 GETPARAMETERS

Input: Conditional moments $O^{(v|i)} = \mathbb{P}(x_v \mid h_i)$ and hidden marginals $Z_S = \mathbb{P}(\mathbf{h}_S)$.
Output: Parameters θ.
  if $\mathcal{G}$ is directed **then**
    Normalize $\mathbb{P}(a, \text{Pa}(a))$ for $a \in \mathbf{x} \cup \mathbf{h}$.
  else if $\mathcal{G}$ is undirected with low treewidth **then**
    Compute features $\mu_\mathcal{C}$ for $\mathcal{C} \in \mathcal{G}$ (Equation 6).
    Optimize full likelihood (Equation 5).
  else if $\mathcal{G}$ is undirected with high treewidth **then**
    Compute features $\mu_{\{a\}\cup\mathcal{N}(a)}$ for $a \in \mathbf{h}$ (Equation 8).
    Optimize pseudolikelihood (Equation 7).
  end if

Remark If we have exclusive views for only a subset of the cliques, we can still obtain the expected features $\mu_\mathcal{C}$ for those cliques and use posterior regularization (Graça et al., 2008), measurements (Liang et al., 2009), or generalized expectation criteria (Mann & McCallum, 2008) to encourage $\mathbb{E}_{p_\theta}[\phi(\mathbf{x}_\mathcal{C}, \mathbf{h}_\mathcal{C})]$ to match $\mu_\mathcal{C}$. The resulting objective functions would be non-convex, but we expect local optima to be less of an issue.

5.1. Pseudolikelihood

While we now have a complete algorithm for estimating directed and undirected models, optimizing the full likelihood (Equation 5) can still be computationally intractable for undirected models with high treewidth due to the intractability of the log-partition function $A(\theta)$. One can employ various variational approximations of $A(\theta)$ (Wainwright & Jordan, 2008), but these generally lead to inconsistent estimates of θ. We thus turn to an older idea of pseudolikelihood (Besag, 1975). The pseudolikelihood objective is a sum over the log-probability of each variable a given its neighbors $\mathcal{N}(a)$:

$$L_{\text{pseudo}}(\theta) \triangleq \mathbb{E}_{(\mathbf{x},\mathbf{h})\sim p_{\theta^*}} \left[\sum_{a\in\mathbf{x}\cup\mathbf{h}} \log p_\theta(a \mid \mathcal{N}(a)) \right]. \quad (7)$$

In the fully-supervised setting, it is well-known that pseudolikelihood provides consistent estimates which are computationally efficient but less statistically efficiency.[5]

Let $\phi_{a,\mathcal{N}(a)}(a, \mathcal{N}(a)) = \sum_{\mathcal{C}\ni a} \phi_\mathcal{C}(\mathbf{x}_\mathcal{C}, \mathbf{h}_\mathcal{C})$ denote the sum over cliques $\mathcal{C}$ that contain a; note that $\phi_{a,\mathcal{N}(a)}$ only depends on a and its neighbors $\mathcal{N}(a)$. We can write each conditional log-likelihood from Equation 7 as:

$$p_\theta(a \mid \mathcal{N}(a)) = \exp(\theta^\top \phi_{a,\mathcal{N}(a)}(a, \mathcal{N}(a)) - A_a(\theta; \mathcal{N}(a))),$$

where the conditional log-partition function

[5] Coincidentally, this is the same high-level motivation for using method of moments in the first place.

$A_a(\theta; \mathcal{N}(a)) = \log \sum_{\alpha \in [k]} \exp(\theta^\top \phi_{a,\mathcal{N}(a)}(\alpha, \mathcal{N}(a)))$ involves marginalizing only over the single variable a.

If we knew the marginals for each neighborhood,

$$\mu_{a,\mathcal{N}(a)} \triangleq \mathbb{E}[\phi_{a,\mathcal{N}(a)}(a, \mathcal{N}(a))], \tag{8}$$

then we would be able to optimize the pseudolikelihood objective again without having access to any labeled data. Unfortunately, $\{a\} \cup \mathcal{N}(a)$ does not always have exclusive views. For example, consider $a = h_1$ and $\mathcal{N}(a) = \{h_2, h_3, h_4\}$ in Figure 3(b).

However, we can decompose $\{a\} \cup \mathcal{N}(a)$ as follows: conditioning on a partitions $\mathcal{N}(a)$ into independent subsets; let $\mathcal{B}(a)$ be the collection of these subsets, which we will call *sub-neighborhoods*. For example, $\mathcal{B}(h_1) = \{\{h_2\}, \{h_3\}, \{h_4\}\}$ in Figure 3(b) and $\mathcal{B}(h_{2,2}) = \{\{h_{1,2}, h_{2,3}, h_{3,2}, h_{2,1}\}\}$ contains a single sub-neighborhood in Figure 5.

A key observation is that for each sub-neighborhood $B \in \mathcal{B}(a)$, each $\{a\} \cup B$ is bidependent: conditioning on a does not introduce new independencies within B by construction of $\mathcal{B}(a)$, and conditioning on any $b \in B$ does not either since every other $b' \in B \backslash \{b\}$ is connected to a. Assuming $\mathcal{G}$ is bottlenecked, by Lemma 2 we have that $\{a\} \cup B$ has exclusive views. Hence, we can recover $\mathbb{P}(a, B)$ for each a and $B \in \mathcal{B}(a)$. Based on conditional independence of the sub-neighborhoods B given a, we have that $\mathbb{P}(a, \mathcal{N}(a)) = \mathbb{P}(a) \prod_{B \in \mathcal{B}(a)} \mathbb{P}(B \mid a)$. This allows us to compute the expected features $\mu_{a,\mathcal{N}(a)}$ and use them in the optimization of the pseudolikelihood objective.

Note that our pseudolikelihood-based approach does depend exponentially on the size of the sub-neighborhoods, which could be exceed the largest clique size. Therefore, each node essentially should have low degree or locally exhibit a lot of conditional independence. On the positive side, we can handle graphical models with high treewidth; neither sample nor computational complexity necessarily depends on the treewidth. For example, an $n \times n$ grid model has a treewidth of n, but the degree is at most 4.

6. Discussion

For latent-variable models, there has been tension between local optimization of likelihood, which is broadly applicable but offers no global theoretical guarantees, and the spectral method of moments, which provides consistent estimators but are limited to models with special structure. The purpose of this work is to show that the two methods can be used synergistically to produce consistent estimates for a broader class of directed and undirected models.

Our approach provides consistent estimates for a family of models in which each hidden variable is a *bottleneck*—that is, it has three conditionally independent observations. This bottleneck property of Anandkumar et al. (2013) has been exploited in many other contexts, including latent Dirichlet allocation (Anandkumar et al., 2012b), mixture of spherical Gaussians (Hsu & Kakade, 2013), probabilistic grammars (Hsu et al., 2012), noisy-or Bayesian networks (Halpern & Sontag, 2013), mixture of linear regressions (Chaganty & Liang, 2013), and others. Each of these methods can be viewed as "preprocessing" the given model into a form that exposes the bottleneck or tensor factorization structure. The model parameters correspond directly to the solution of the factorization.

In contrast, the bottlenecks in our graphical models are given by assumption, but the conditional distribution of the observations given the bottleneck can be quite complex. Our work can therefore be viewed as "postprocessing", where the conditional moments recovered from tensor factorization are used to further obtain the hidden marginals and eventually the parameters. Along the way, we developed the notion of exclusive views and bidependent sets, which characterize conditions under which the conditional moments can reveal the dependency structure between hidden variables. We also made use of custom likelihood functions which were constructed to be easy to optimize.

Another prominent line of work in the method of moments community has focused on recovering *observable operator representations* (Jaeger, 2000; Hsu et al., 2009; Bailly et al., 2010; Balle & Mohri, 2012). These methods allow prediction of new observations, but do not recover the actual parameters of the model, making them difficult to use in conjunction with likelihood-based models. Song et al. (2011) proposed an algorithm to learn observable operator representations for latent tree graphical models, like the one in Figure 3(b), assuming the graph is bottlenecked. Their approach is similar to our first step of learning conditional moments, but they only consider trees. Parikh et al. (2012) extended this approach to general graphical models which are bottlenecked using a latent junction tree representation. Consequently, the size of the observable representations is exponential in the treewidth. In contrast, our algorithm only constructs moments of the order of size of the cliques (and sub-neighborhoods for pseudolikelihood), which can be much smaller.

An interesting direction is to examine the necessity of the bottleneck property. Certainly, three views is in general needed to ensure identifiability (Kruskal, 1977), but requiring *each* hidden variable to be a bottleneck is stronger than what we would like. We hope that by judiciously leveraging likelihood-based methods in conjunction with the method of moments, we can generate new hybrid techniques for estimating even richer classes of latent-variable models.

References

Anandkumar, A., Chaudhuri, K., Hsu, D., Kakade, S. M., Song, L., and Zhang, T. Spectral methods for learning multivariate latent tree structure. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

Anandkumar, A., Hsu, D., and Kakade, S. M. A method of moments for mixture models and hidden Markov models. In *Conference on Learning Theory (COLT)*, 2012a.

Anandkumar, A., Liu, Y., Hsu, D., Foster, D. P., and Kakade, S. M. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 917–925, 2012b.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. Tensor decompositions for learning latent variable models. Technical report, ArXiv, 2013.

Bailly, R., Habrard, A., and Denis, F. A spectral approach for probabilistic grammatical inference on trees. In *Algorithmic Learning Theory*, pp. 74–88. Springer, 2010.

Balle, B. and Mohri, M. Spectral learning of general weighted automata via constrained matrix completion. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 2159–2167, 2012.

Besag, J. The analysis of non-lattice data. *The Statistician*, 24:179–195, 1975.

Blei, D., Ng, A., and Jordan, M. I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Chaganty, A. and Liang, P. Spectral experts for estimating mixtures of linear regressions. In *International Conference on Machine Learning (ICML)*, 2013.

Ghahramani, Z. and Beal, M. J. Variational inference for Bayesian mixtures of factor analysers. In *Advances in Neural Information Processing Systems*, 1999.

Graça, J., Ganchev, K., and Taskar, B. Expectation maximization and posterior constraints. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

Haghighi, A. and Klein, D. Prototype-driven learning for sequence models. In *North American Association for Computational Linguistics (NAACL)*, 2006.

Halpern, Y. and Sontag, D. Unsupervised learning of noisy-or Bayesian networks. In *Uncertainty in Artificial Intelligence (UAI)*, pp. 272–281, 2013.

Hsu, D. and Kakade, S. M. Learning mixtures of spherical Gaussians: Moment methods and spectral decompositions. In *Innovations in Theoretical Computer Science (ITCS)*, 2013.

Hsu, D., Kakade, S. M., and Zhang, T. A spectral algorithm for learning hidden Markov models. In *Conference on Learning Theory (COLT)*, 2009.

Hsu, D., Kakade, S. M., and Liang, P. Identifiability and unmixing of latent parse trees. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

Jaakkola, T. S and Jordan, M. I. Variational probabilistic inference and the QMR-DT network. *Journal of Artificial Intelligence Research*, 10:291–322, 1999.

Jaeger, H. Observable operator models for discrete stochastic time series. *Neural Computation*, 2000.

Koller, D. and Friedman, N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Kruskal, J. B. Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and Applications*, 18:95–138, 1977.

Liang, P., Jordan, M. I., and Klein, D. Learning from measurements in exponential families. In *International Conference on Machine Learning (ICML)*, 2009.

Lindsay, B. Composite likelihood methods. *Contemporary Mathematics*, 80:221–239, 1988.

Mann, G. and McCallum, A. Generalized expectation criteria for semi-supervised learning of conditional random fields. In *Human Language Technology and Association for Computational Linguistics (HLT/ACL)*, pp. 870–878, 2008.

Mossel, E. and Roch, S. Learning nonsingular phylogenies and hidden markov models. In *Theory of computing*, pp. 366–375. ACM, 2005.

Parikh, A., Song, L., Ishteva, M., Teodoru, G., and Xing, E. A spectral algorithm for latent junction trees. In *Uncertainty in Artificial Intelligence (UAI)*, 2012.

Quattoni, A., Collins, M., and Darrell, T. Conditional random fields for object recognition. In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

Song, Le, Xing, E. P, and Parikh, A. P. A spectral algorithm for latent tree graphical models. In *International Conference on Machine Learning (ICML)*, 2011.

van der Vaart, A. W. *Asymptotic statistics*. Cambridge University Press, 1998.

Wainwright, M. and Jordan, M. I. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–307, 2008.

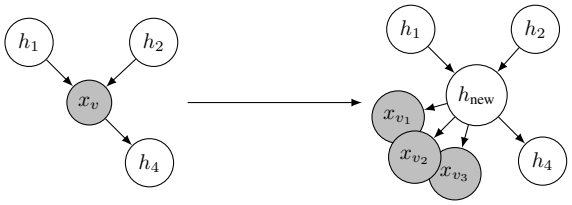


Figure 6. Reduction to canonical form.

A. Proofs

Due to space limitations, we have omitted some proofs from the main body of the paper. The proofs are provided below.

A.1. Lemma 1

Proof. Let x_v be an observed variable which is contained in more than one clique or in cliques of size larger than 2. We apply the following simple transformation (see Figure 6 for directed models): first, replace x_v with a a new hidden variable h_{new}; for directed models, this means that the parents and children of x_v become the parents and children of h_{new}. Second, create three fresh observed variables $x_{v_1}, x_{v_2}, x_{v_3}$, connecting them to h_{new}, and making all new nodes to deterministically take on identical values. We add three copies so that h_{new} is guaranteed to be a bottleneck. By construction, there is a one-to-one mapping between the joint distributions of the old and new graphical models, and thus the parameters as well. We repeatedly apply this procedure until the graphical model is in canonical form. □

A.2. Lemma 3

In Section 4.2, we compared the asymptotic variance Σ_S^{cl} of the composite likelihood estimator with that of the pseudoinverse estimator, Σ_S^{pi}, for a subset of hidden variables S. Now we will derive these asymptotic variances in detail.

Recall, that in Section 4.2 we simplified notation by taking $m = 1$ and flattening the moments $M_{\mathcal{V}}$ and hidden marginals Z_S into vectors $\mu \in \mathbb{R}^d$ and $z \in \mathbb{R}^k$ respectively. The conditional moments, O, is a now matrix $O \in \mathbb{R}^{d \times k}$ and the hidden marginals z and observed marginals μ are related via $\mu = Oz$.

Lemma (Asymptotic variances)**.** *The asymptotic variances of the pseudoinverse estimator $\widehat{\widetilde{z}}^{\text{pi}}$ and composite likelihood estimator $\widehat{\widetilde{z}}^{\text{cl}}$ are:*

$$\Sigma^{\text{pi}} = \widetilde{O}^{\dagger}(\widetilde{D} - \widetilde{\mu}\widetilde{\mu}^{\top})\widetilde{O}^{\dagger\top},$$
$$\Sigma^{\text{cl}} = \left(\widetilde{O}^{\top}(\widetilde{D}^{-1} + \widetilde{d}^{-1}\mathbf{1}\mathbf{1}^{\top})\widetilde{O}\right)^{-1},$$

where $\widetilde{D} \triangleq \operatorname{diag}(\widetilde{\mu})$ and $\widetilde{d} \triangleq 1 - \mathbf{1}^{\top}\widetilde{\mu}$.

Proof for Lemma 3. First, let us look at the asymptotic variance of the pseudoinverse estimator $\hat{z}^{\text{pi}} = \widetilde{O}^{\dagger}(\widehat{\widetilde{\mu}} - O_{\neg d,k})$. Note that $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$, where each x_i is an independent draw from the multinomial distribution μ. Hence the variance of $\hat{\mu}$ is $(D - \mu\mu^{\top})$ where $D \triangleq \operatorname{diag}(\mu)$. Recall that $\widehat{\widetilde{\mu}}$ is just the first $d-1$ entries of $\hat{\mu}$, so the variance of $\widehat{\widetilde{\mu}}$ is $(\widetilde{D} - \widetilde{\mu}\widetilde{\mu}^{\top})$ where $\widetilde{D} \triangleq \operatorname{diag}(\widetilde{\mu})$. Since $\widetilde{z}$ is just a linear transformation of $\widetilde{\mu}$, the asymptotic variance of $\widehat{\widetilde{z}}^{\text{pi}}$ is:

$$\begin{aligned}\Sigma^{\text{pi}} &= \widetilde{O}^{\dagger}\operatorname{Var}(\widehat{\widetilde{\mu}})\widetilde{O}^{\dagger\top} \\ &= \widetilde{O}^{\dagger}(\widetilde{D} - \widetilde{\mu}\widetilde{\mu}^{\top})\widetilde{O}^{\dagger\top}.\end{aligned}$$

Now, let us look at the variance of the composite likelihood estimator. Using the delta-method (van der Vaart, 1998) we have that the asymptotic variance of $\widehat{\widetilde{z}}^{\text{cl}} = \arg\max_{\widetilde{z}} \hat{\mathbb{E}}[\ell(x;\widetilde{z})]$ is,

$$\Sigma^{\text{cl}} = \mathbb{E}[\nabla^2\ell(x;\widetilde{z}^*)]^{-1}\operatorname{Var}[\nabla\ell(x;\widetilde{z}^*)]\,\mathbb{E}[\nabla^2\ell(x;\widetilde{z}^*)]^{-1},$$

where $\ell(x;\widetilde{z})$ is the log-likelihood of the observations x given parameters $\widetilde{z}$. We can write $\ell(x;\widetilde{z})$ in terms of $\widetilde{z}$ and $\widetilde{O}$ as,

$$\begin{aligned}\ell(x;\widetilde{z}) &= \log(\mu[x]) \\ &= \log\left(e_x^{\top}\begin{bmatrix}\widetilde{O} \\ -\mathbf{1}^{\top}\widetilde{O}\end{bmatrix}\widetilde{z} + e_x^{\top}\begin{bmatrix}O_{\neg d,k} \\ 1 - \mathbf{1}^{\top}O_{\neg d,k}\end{bmatrix}\right),\end{aligned}$$

where e_x is an indicator vector on x.

Taking the first derivative,

$$\begin{aligned}\nabla\ell(x;\widetilde{z}) &= \frac{1}{\mu[x]}\begin{bmatrix}\widetilde{O} \\ -\mathbf{1}^{\top}\widetilde{O}\end{bmatrix}^{\top} e_x \\ &= \begin{bmatrix}\widetilde{O} \\ -\mathbf{1}^{\top}\widetilde{O}\end{bmatrix}^{\top} D^{-1}e_x, \qquad (9)\end{aligned}$$

where $D \triangleq \operatorname{diag}(\mu)$.

It is easily verified that the expectation of the first derivative is indeed $\mathbf{0}$:

$$\begin{aligned}\mathbb{E}[\nabla\ell(x;\widetilde{z})] &= \begin{bmatrix}\widetilde{O} \\ -\mathbf{1}^{\top}\widetilde{O}\end{bmatrix}^{\top} D^{-1}\,\mathbb{E}[e_x] \\ &= \begin{bmatrix}\widetilde{O} \\ -\mathbf{1}^{\top}\widetilde{O}\end{bmatrix}^{\top} D^{-1}\mu \\ &= \begin{bmatrix}\widetilde{O} \\ -\mathbf{1}^{\top}\widetilde{O}\end{bmatrix}^{\top}\mathbf{1} \\ &= \widetilde{O}^{\top}\mathbf{1} - \widetilde{O}^{\top}\mathbf{1} \\ &= \mathbf{0}.\end{aligned}$$

Taking the second derivative,

$$\nabla^2 \ell(x;\widetilde{z}) = \frac{1}{\mu[x]^2}\begin{bmatrix}\widetilde{O}\\ -\mathbf{1}^\top\widetilde{O}\end{bmatrix}^\top e_x e_x^\top \begin{bmatrix}\widetilde{O}\\ -\mathbf{1}^\top\widetilde{O}\end{bmatrix} = \begin{bmatrix}\widetilde{O}\\ -\mathbf{1}^\top\widetilde{O}\end{bmatrix}^\top D^{-1} e_x e_x^\top D^{-1} \begin{bmatrix}\widetilde{O}\\ -\mathbf{1}^\top\widetilde{O}\end{bmatrix}. \quad (10)$$

From Equation 9 and Equation 10, we get

$$\begin{aligned}
\mathbb{E}[\nabla^2\ell(x;\widetilde{z}^*)] &= -\begin{bmatrix}\widetilde{O}\\ -\mathbf{1}^\top\widetilde{O}\end{bmatrix}^\top D^{-1}\,\mathbb{E}[e_x e_x^\top]D^{-1}\begin{bmatrix}\widetilde{O}\\ -\mathbf{1}^\top\widetilde{O}\end{bmatrix}\\
\mathrm{Var}[\nabla\ell(x;\widetilde{z}^*)] &= \begin{bmatrix}\widetilde{O}\\ -\mathbf{1}^\top\widetilde{O}\end{bmatrix}^\top D^{-1}\,\mathbb{E}[e_x e_x^\top]D^{-1}\begin{bmatrix}\widetilde{O}\\ -\mathbf{1}^\top\widetilde{O}\end{bmatrix}\\
&= \begin{bmatrix}\widetilde{O}\\ -\mathbf{1}^\top\widetilde{O}\end{bmatrix}^\top D^{-1}DD^{-1}\begin{bmatrix}\widetilde{O}\\ -\mathbf{1}^\top\widetilde{O}\end{bmatrix}\\
&= \begin{bmatrix}\widetilde{O}\\ -\mathbf{1}^\top\widetilde{O}\end{bmatrix}^\top \begin{bmatrix}\widetilde{D}^{-1} & \mathbf{0}\\ \mathbf{0}^\top & \widetilde{d}^{-1}\end{bmatrix}\begin{bmatrix}\widetilde{O}\\ -\mathbf{1}^\top\widetilde{O}\end{bmatrix}\\
&= \widetilde{O}^\top\widetilde{D}^{-1}\widetilde{O} + \widetilde{d}^{-1}\widetilde{O}^\top\mathbf{1}\mathbf{1}^\top\widetilde{O},
\end{aligned}$$

where $\widetilde{D} = \mathrm{diag}(\widetilde{\mu})$ and $\widetilde{d} = 1 - \mathbf{1}^\top\widetilde{\mu}$ are the diagonal elements of D. As expected, $\mathbb{E}[\nabla^2\ell(x)] = -\mathrm{Var}[\nabla\ell(x)]$ because $\hat{z}$ is a maximum likelihood estimator.

Finally, the asymptotic variance of Σ^{cl} is,

$$\begin{aligned}
\Sigma^{\mathrm{cl}} &= \mathbb{E}[\nabla^2\ell(x;\widetilde{z}^*)]^{-1}\,\mathrm{Var}[\nabla\ell(x;\widetilde{z}^*)]\,\mathbb{E}[\nabla^2\ell(x;\widetilde{z}^*)]^{-1}\\
&= \mathrm{Var}[\nabla\ell(x;\widetilde{z}^*)]^{-1}\\
&= \left(\widetilde{O}^\top\widetilde{D}^{-1}\widetilde{O} + \widetilde{d}^{-1}\widetilde{O}^\top\mathbf{1}\mathbf{1}^\top\widetilde{O}\right)^{-1}.
\end{aligned}$$

Given our assumptions, $\mathbf{1} \succ \mu \succ \mathbf{0}$. Consequently, $\widetilde{D}$ is invertible and the asymptotic variance is finite. □

A.3. Comparing the pseudoinverse and composite likelihood estimators

In Lemma 3, we derived concrete expressions for the asymptotic variances of the pseudoinverse and composite likelihood estimators, Σ^{pi} and Σ^{cl} respectively. In this section, we will use the asymptotic variances to compare the two estimators for two special cases.

Recall that the relative efficiency of the pseudoinverse estimator with respect to the composite likelihood estimator is $e^{\mathrm{pi}} = \frac{1}{\widetilde{k}}\,\mathrm{tr}(\Sigma^{\mathrm{cl}}(\Sigma^{\mathrm{pi}})^{-1})$, where $\widetilde{k} = k - 1$. The Cramér-Rao lower bound tells us that $\Sigma^{\mathrm{cl}} \preceq \Sigma^{\mathrm{pi}}$: thus the relative efficiency e^{pi} lies between 0 and 1. When $e^{\mathrm{pi}} = 1$, the pseudoinverse estimator is said to be efficient.

We will make repeated use of the Sherman-Morrison formula to simplify matrix inverses:

$$(A + \alpha uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{\alpha^{-1} + v^\top A^{-1}u},$$

where A is an invertible matrix, u, v are vectors and α is a scalar constant. Unless otherwise specified, we $\|u\|$ to denote the Euclidean norm of a vector u.

First, let us consider the case where $\widetilde{O}$:

Lemma 6 (Relative efficiency when $\widetilde{O}$ is invertible)**.** *When $\widetilde{O}$ is invertible, the asymptotic variances of the pseudoinverse and composite likelihood estimators are equal, $\Sigma^{\mathrm{cl}} = \Sigma^{\mathrm{pi}}$, and the relative efficiency is* 1.

Proof. Given that $\widetilde{O}$ is invertible we can simplify the expression of the asymptotic variance of the composite likelihood estimator, Σ^{cl}, as follows:

$$\begin{aligned}
\Sigma^{\mathrm{cl}} &= \left(\widetilde{O}^\top(\widetilde{D}^{-1} + \widetilde{d}^{-1}\mathbf{1}\mathbf{1}^\top)\widetilde{O}\right)^{-1}\\
&= \widetilde{O}^{-1}\left(\widetilde{D}^{-1} - \widetilde{d}^{-1}\mathbf{1}\mathbf{1}^\top\right)^{-1}\widetilde{O}^{-\top}\\
&= \widetilde{O}^{-1}\left(\widetilde{D} - \frac{\widetilde{D}\mathbf{1}\mathbf{1}^\top\widetilde{D}}{\widetilde{d} + \mathbf{1}^\top\widetilde{D}\mathbf{1}}\right)\widetilde{O}^{-\top}.
\end{aligned}$$

Note that $\widetilde{D}\mathbf{1} = \widetilde{\mu}$ and $\widetilde{d} = 1 - \mathbf{1}^\top\widetilde{\mu}$. This gives us,

$$\begin{aligned}
\Sigma^{\mathrm{cl}} &= \widetilde{O}^{-1}\left(\widetilde{D} - \frac{\widetilde{\mu}\widetilde{\mu}^\top}{1 - \mathbf{1}^\top\widetilde{\mu} + \mathbf{1}^\top\widetilde{\mu}}\right)\widetilde{O}^{-\top}\\
&= \widetilde{O}^{-1}(\widetilde{D} - \widetilde{\mu}\widetilde{\mu}^\top)\widetilde{O}^{-\top}\\
&= \Sigma^{\mathrm{pi}}.
\end{aligned}$$

□

Next, we consider the case where the observed moments μ is the uniform distribution.

Lemma 7 (Relative efficiency with uniform observed moments)**.** *Let the observed marginals μ be uniform: $\mu = \frac{1}{d}\mathbf{1}$. The efficiency of the pseudoinverse estimator is,*

$$e^{\mathrm{pi}} = 1 - \frac{1}{k-1}\frac{\|\mathbf{1}_U\|_2^2}{1 + \|\mathbf{1}_U\|_2^2}\left(1 - \frac{1}{d - \|\mathbf{1}_U\|_2^2}\right), \quad (11)$$

where $\mathbf{1}_U \triangleq \widetilde{O}\widetilde{O}^\dagger\mathbf{1}$, the projection of $\mathbf{1}$ onto the column space of $\widetilde{O}$. Note that $0 \le \|\mathbf{1}_U\|_2^2 \le k - 1$.

When $\|\mathbf{1}_U\|_2 = 0$, the pseudoinverse estimator is efficient: $e^{\mathrm{pi}} = 1$. When $\|\mathbf{1}_U\|_2 > 0$ and $d > k$, the pseudoinverse estimator is strictly inefficient. In particular, if $\|\mathbf{1}_U\|_2^2 = k - 1$, and we get:

$$e^{\mathrm{pi}} = 1 - \frac{1}{k}\left(1 - \frac{1}{1 + d - k}\right). \quad (12)$$

Proof. Next, let us consider the case where the moments are the uniform distribution, where $\mu = \frac{1}{d}\mathbf{1}$ and $\widetilde{D} = \frac{1}{d}I$. The expressions for Σ^{cl} can be simplified as follows,

$$\begin{aligned}
\Sigma^{\text{cl}} &= \left(\widetilde{O}^\top(dI + d\mathbf{1}\mathbf{1}^\top)\widetilde{O}\right)^{-1} \\
&= \frac{1}{d}\left(\widetilde{O}^\top\widetilde{O} + \widetilde{O}^\top\mathbf{1}\mathbf{1}^\top\widetilde{O}\right)^{-1} \\
&= \frac{1}{d}\left((\widetilde{O}^\top\widetilde{O})^{-1} - \frac{(\widetilde{O}^\top\widetilde{O})^{-1}\widetilde{O}^\top\mathbf{1}\mathbf{1}^\top\widetilde{O}(\widetilde{O}^\top\widetilde{O})^{-1}}{1 + \mathbf{1}^\top\widetilde{O}(\widetilde{O}^\top\widetilde{O})^{-1}\widetilde{O}^\top\mathbf{1}}\right) \\
&= \frac{1}{d}\left(\widetilde{O}^\dagger\widetilde{O}^{\dagger\top} - \frac{(\widetilde{O}^\dagger\widetilde{O}^{\dagger\top}\widetilde{O}^\top)\mathbf{1}\mathbf{1}^\top(\widetilde{O}\widetilde{O}^\dagger\widetilde{O}^{\dagger\top})}{1 + (\mathbf{1}^\top\widetilde{O}\widetilde{O}^\dagger)(\widetilde{O}^{\dagger\top}\widetilde{O}^\top\mathbf{1})}\right),
\end{aligned}$$

where we have used the property $(\widetilde{O}^\top\widetilde{O})^{-1} = \widetilde{O}^\dagger\widetilde{O}^{\dagger\top}$ in the last step. Next, we use the pseudoinverse property, $\widetilde{O}\widetilde{O}^\dagger\widetilde{O}^{\dagger\top} = \widetilde{O}^{\dagger\top}$,

$$\begin{aligned}
\Sigma^{\text{cl}} &= \frac{1}{d}\left(\widetilde{O}^\dagger\widetilde{O}^{\dagger\top} - \frac{\widetilde{O}^\dagger\mathbf{1}\mathbf{1}^\top\widetilde{O}^{\dagger\top}}{1 + \|\widetilde{O}\widetilde{O}^\dagger\mathbf{1}\|^2}\right) \\
&= \frac{1}{d}\left(\widetilde{O}^\dagger\widetilde{O}^{\dagger\top} - \frac{\widetilde{O}^\dagger\mathbf{1}\mathbf{1}^\top\widetilde{O}^{\dagger\top}}{1 + \|\mathbf{1}_U\|^2}\right),
\end{aligned}$$

where $\mathbf{1}_U \triangleq \widetilde{O}\widetilde{O}^\dagger\mathbf{1} = \widetilde{O}^{\dagger\top}\widetilde{O}^\top\mathbf{1}$ is the projection of $\mathbf{1}$ onto the column space of $\widetilde{O}$.

Next, we can simplify the expression for $(\Sigma^{\text{pi}})^{-1}$,

$$\begin{aligned}
\Sigma^{\text{pi}} &= \widetilde{O}^\dagger\left(\frac{I}{d} - \frac{\mathbf{1}\mathbf{1}^\top}{d^2}\right)\widetilde{O}^{\dagger\top} \\
(\Sigma^{\text{pi}})^{-1} &= \left(\frac{1}{d}\widetilde{O}^\dagger\widetilde{O}^{\dagger\top} - \frac{1}{d^2}\widetilde{O}^\dagger\mathbf{1}\mathbf{1}^\top\widetilde{O}^{\dagger\top}\right)^{-1} \\
&= d\Bigg((\widetilde{O}^\dagger\widetilde{O}^{\dagger\top})^{-1} \\
&\qquad + \frac{(\widetilde{O}^\dagger\widetilde{O}^{\dagger\top})^{-1}\widetilde{O}^\dagger\mathbf{1}\mathbf{1}^\top\widetilde{O}^{\dagger\top}(\widetilde{O}^\dagger\widetilde{O}^{\dagger\top})^{-1}}{d - \mathbf{1}^\top\widetilde{O}^{\dagger\top}(\widetilde{O}^\dagger\widetilde{O}^{\dagger\top})^{-1}\widetilde{O}^\dagger\mathbf{1}}\Bigg).
\end{aligned}$$

Using the properties $(\widetilde{O}^\dagger\widetilde{O}^{\dagger\top})^{-1} = \widetilde{O}^\top\widetilde{O}$ and $\widetilde{O}^\top\widetilde{O}\widetilde{O}^\dagger = \widetilde{O}^\top$, we get,

$$\begin{aligned}
(\Sigma^{\text{pi}})^{-1} &= d\left(\widetilde{O}^\top\widetilde{O} + \frac{\widetilde{O}^\top\widetilde{O}\widetilde{O}^\dagger\mathbf{1}\mathbf{1}^\top\widetilde{O}^{\dagger\top}\widetilde{O}^\top\widetilde{O}}{d - \mathbf{1}^\top\widetilde{O}^{\dagger\top}\widetilde{O}^\top\widetilde{O}\widetilde{O}^\dagger\mathbf{1}}\right) \\
&= d\left(\widetilde{O}^\top\widetilde{O} + \frac{\widetilde{O}^\top\mathbf{1}\mathbf{1}^\top\widetilde{O}}{d - \|\widetilde{O}^\dagger\widetilde{O}\mathbf{1}\|^2}\right) \\
&= d\left(\widetilde{O}^\top\widetilde{O} + \frac{\widetilde{O}^\top\mathbf{1}\mathbf{1}^\top\widetilde{O}}{d - \|\mathbf{1}_U\|^2}\right).
\end{aligned}$$

Now, we are ready to study the relative efficiency.

$$\begin{aligned}
e^{\text{pi}} &= \frac{1}{\widetilde{k}}\operatorname{tr}(\Sigma^{\text{cl}}(\Sigma^{\text{pi}})^{-1}) \\
&= \frac{1}{\widetilde{k}}\operatorname{tr}\left(\frac{1}{d}\left(\widetilde{O}^\dagger\widetilde{O}^{\dagger\top} - \frac{\widetilde{O}^\dagger\mathbf{1}\mathbf{1}^\top\widetilde{O}^{\dagger\top}}{1 + \|\mathbf{1}_U\|^2}\right)\right. \\
&\qquad \left. d\left(\widetilde{O}^\top\widetilde{O} + \frac{\widetilde{O}^\top\mathbf{1}\mathbf{1}^\top\widetilde{O}}{d - \|\mathbf{1}_U\|^2}\right)\right) \\
&= \frac{1}{\widetilde{k}}\operatorname{tr}(I) + \frac{1}{\widetilde{k}}\operatorname{tr}\left(\frac{\widetilde{O}^\dagger\widetilde{O}^{\dagger\top}\widetilde{O}^\top\mathbf{1}\mathbf{1}^\top\widetilde{O}}{d - \|\mathbf{1}_U\|^2}\right) \\
&\quad - \frac{1}{\widetilde{k}}\operatorname{tr}\left(\frac{\widetilde{O}^\dagger\mathbf{1}\mathbf{1}^\top\widetilde{O}^{\dagger\top}\widetilde{O}^\top\widetilde{O}}{1 + \|\mathbf{1}_U\|^2}\right) \\
&\quad - \frac{1}{\widetilde{k}}\operatorname{tr}\left(\frac{\widetilde{O}^\dagger\mathbf{1}\mathbf{1}^\top\widetilde{O}^{\dagger\top}\widetilde{O}^\top\mathbf{1}\mathbf{1}^\top\widetilde{O}}{(d - \|\mathbf{1}_U\|^2)(1 + \|\mathbf{1}_U\|^2)}\right)
\end{aligned}$$

Next we apply the property that the trace is invariant under cyclic permutations,

$$\begin{aligned}
e^{\text{pi}} &= 1 + \frac{1}{\widetilde{k}}\frac{\|\widetilde{O}^{\dagger\top}\widetilde{O}^\top\mathbf{1}\|^2}{d - \|\mathbf{1}_U\|^2} - \frac{1}{\widetilde{k}}\frac{\|\widetilde{O}\widetilde{O}^\dagger\mathbf{1}\|^2}{1 + \|\mathbf{1}_U\|^2} \\
&\quad - \frac{1}{\widetilde{k}}\frac{(\mathbf{1}^\top\widetilde{O}^{\dagger\top}\widetilde{O}^\top\mathbf{1})^2}{(d - \|\mathbf{1}_U\|^2)(1 + \|\mathbf{1}_U\|^2)}.
\end{aligned}$$

Note that $\widetilde{O}\widetilde{O}^\dagger$ is a symmetric projection matrix and thus, $\widetilde{O}\widetilde{O}^\dagger = (\widetilde{O}\widetilde{O}^\dagger)^\top$ and $\widetilde{O}\widetilde{O}^\dagger = (\widetilde{O}\widetilde{O}^\dagger)(\widetilde{O}\widetilde{O}^\dagger)$. Then,

$$\begin{aligned}
e^{\text{pi}} &= 1 + \frac{1}{\widetilde{k}}\frac{\|\mathbf{1}_U\|^2}{d - \|\mathbf{1}_U\|^2} - \frac{1}{\widetilde{k}}\frac{\|\mathbf{1}_U\|^2}{1 + \|\mathbf{1}_U\|^2} \\
&\quad - \frac{1}{\widetilde{k}}\frac{\|\mathbf{1}_U\|^4}{(1 + \|\mathbf{1}_U\|^2)(d - \|\mathbf{1}_U\|^2)} \\
&= 1 - \frac{\|\mathbf{1}_U\|^2}{\widetilde{k}(1 + \|\mathbf{1}_U\|^2)}\left(1 - \frac{1}{d - \|\mathbf{1}_U\|^2}\right).
\end{aligned}$$

Note that $\mathbf{1}_U$ is the projection of $\mathbf{1}$ on to a k-dimensional subspace, thus, $0 \le \|\mathbf{1}_U\|^2 \le k$. When $\mathbf{1}_U = \mathbf{0}$, the relative efficiency e^{pi} is 1: the pseudoinverse estimator is efficient. When $\|\mathbf{1}_U\| > 0$ and $d > k$, the pseudoinverse estimator is strictly inefficient.

Consider the case when $\|\mathbf{1}_U\|^2 = \widetilde{k}$. Then, the relative efficiency is,

$$\begin{aligned}
e^{\text{pi}} &= 1 - \frac{1}{\widetilde{k} + 1}\left(1 - \frac{1}{d - \widetilde{k}}\right) \\
&= 1 - \frac{1}{k}\left(1 - \frac{1}{1 + d - k}\right).
\end{aligned}$$

□