Likelihood Based Inference in Fully and Partially Observed Exponential Family Graphical Models with Intractable Normalizing Constants

Anindya Bhadra
www.stat.purdue.edu/∼bhadra

Purdue University

Overview

- (Undirected) probabilistic graphical models encode a Markov random field, denoting conditional independence relationships.

- May include latent variables, used in generative models such as Boltzmann machines (also called energy-based models).

- Apart from very specific cases (such as multivariate Gaussian), these models have an intractable normalizing constant that affect any likelihood-based inference (MLE or Bayesian).

- Goal: To provide a tractable approach for likelihood-based inference in these models.

- *Joint work with Yujie Chen and Antik Chakraborty (Purdue).*

Fully observed exponential family graphical models

- Let p denote the number of variables.

- We consider models of the form: $p_\theta(x) = \exp(-E_\theta(x))/z(\theta)$. More explicitly:

$$p_\theta(x_1, \dots, x_p) = \frac{1}{z(\theta)} \exp\left\{ \sum_{j \in V} \theta_j T_j(x_j) + \sum_{(j,k) \in E} \theta_{jk} T_{jk}(x_j, x_k) + \sum_{j \in V} C(x_j) \right\}.$$

- Restriction to exponential family offers crucial advantages, and $E_\theta(x)$ and $\nabla_\theta E_\theta(x)$ have simple forms that are easy to evaluate.

- Examples:
  - Ising: $C(x_j) = 0$, $T_j(x_j) = x_j$ and $T_{jk}(x_j, x_k) = x_j x_k$. Sample space: $\{0, 1\}^p$ and $\Theta = \mathbb{R}^{p \times p}$.
  - Poisson graphical model (Besag, 1974): $C(x_j) = \log x_j!$, $T_j(x_j) = x_j$ and $T_{jk}(x_j, x_k) = x_j x_k$, sample space: $\{0, 1, \dots, \}$, $\Theta \in \mathbb{R}^{p \times p}$ with non-positive off-diagonal elements.

Typical inference methodology in fully observed cases: pseudolikelihood and MCMC for doubly intractable models

- The node-conditional distribution of $X_j \mid X_{-j}$ is:

$$p_\theta(x_j \mid x_{-j}) = \frac{1}{z(\theta; x_{-j})} \exp\left\{ \theta_j T_j(x_j) + C(x_j) + 2 \sum_{k \in N(j)} \theta_{jk} T_{jk}(x_j, x_k) \right\},$$

  a univariate exp. family distribution, with known $z(\cdot)$.

- Besag (1974, JRSSB) proposed using $\prod_{j=1}^{p} p(X_j \mid X_{-j})$, the pseudolikelihood, which is tractable, instead of likelihood.

- Doubly intractable MCMC: Exchange algorithm (Murray et al., 2006), contrastive divergence (Hinton, 2002) etc. One idea: Run auxiliary chain to generate samples from $p_\theta(\cdot)$. Then MC approximate using:

$$\nabla_\theta \log z(\theta) = \mathbb{E}_{Y \sim p_\theta}\{-\nabla_\theta(E_\theta(Y))\}$$

- Score matching: Hyvärinen (2005, JMLR).

Partially observed exponential family graphical models

- Models of the form: $p_\theta(\mathbf{v}, \mathbf{h}) = z(\theta)^{-1} \exp(-E_\theta(\mathbf{v}, \mathbf{h}))$ and $z(\theta) = \sum_{\mathbf{v},\mathbf{h}} \exp(-E_\theta(\mathbf{v}, \mathbf{h}))$.

- Specifically,

$$\log p_\theta(\mathbf{v}, \mathbf{h}) = \sum_j \theta_{jj} v_j + \sum_k \theta_{kk} h_k + \sum_{j \neq j'} \theta_{jj'} v_j v_{j'} + \sum_{k \neq k'} \theta_{kk'} h_k h_{k'} + \sum_{j,k} \theta_{jk} v_j h_k - \log z(\theta),$$

  for $\theta \in \mathbb{R}^{(p+m)\times(p+m)}$.

- Note: $p_\theta(\mathbf{v}) = \sum_{\mathbf{h}} p_\theta(\mathbf{v}, \mathbf{h})$ is not in exponential family in general! It is the "product of experts" model of Hinton (2002).

Examples in partially observed cases: Boltzmann machines

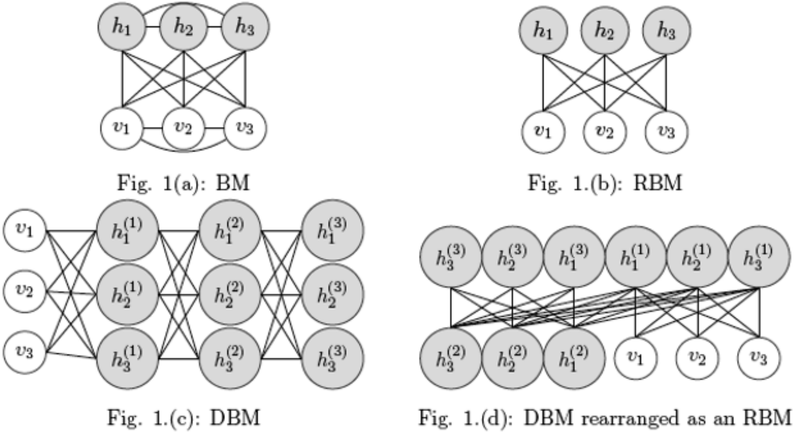


Figure 1: From left to right: BM, RBM, DBM with three visible nodes. DBM has three layers of hidden variables. The notation is: hidden nodes ($h_j \in \{0, 1\}$, shaded in gray), visible nodes ($v_k \in \{0, 1\}$, transparent). Deep hidden nodes in layer l are denoted by $h^{(l)}$.

Some remarks on Boltzmann machines

- One of the earliest forms of *generative models* that allows learning a latent structure for $p_\theta(\mathbf{h} \mid \mathbf{v})$.
- Can capture interactions not belonging in the exponential family. In fact, with enough hidden nodes, is a universal approximator of any distribution on $\{0, 1\}^p$ (Montufar and Ay, 2011, Neural Comp.).
- Currently not very popular due to difficulties in training.
- We will try to understand the cause of this difficulty, and what can be done about it.

Typical inference methodology in partially observed cases: contrastive divergence

- Due to Hinton (2002, Neural Comp.).
- Consider gradient based learning. We have:

$$\frac{\partial \ell(\theta)}{\partial \theta} = \frac{\partial \log \sum_{\mathbf{h}} p_\theta(\mathbf{v}, \mathbf{h})}{\partial \theta} = \sum_{v,h} \frac{\partial E_\theta(\mathbf{v}, \mathbf{h})}{\partial \theta} p_\theta(\mathbf{v}, \mathbf{h}) - \sum_{h} \frac{\partial E_\theta(\mathbf{v}, \mathbf{h})}{\partial \theta} p_\theta(\mathbf{h} \mid \mathbf{v}),$$

- Recall, in RBM: $E_\theta(\mathbf{v}, \mathbf{h}) = -\sum_{j,k} \theta_{jk} v_j h_k$. Thus, Hinton writes:

$$\nabla_{\theta^{(t)}} \log p_\theta(\mathbf{h}, \mathbf{v}) = \langle \mathbf{h}\mathbf{v}^T \rangle_{\mathrm{data}} - \langle \mathbf{h}\mathbf{v}^T \rangle_{\mathrm{model}},$$

  - Subscript "data" denote an expectation with respect to $p_{\theta^{(t)}}(\mathbf{h} \mid \mathbf{v})$ at the observed $\mathbf{v}$, which is analytic.
  - Subscript "model" denote an expectation with respect to $p_{\theta^{(t)}}(\mathbf{h}, \mathbf{v})$; which is typically not available in closed form.

Hinton's solution: the contrastive divergence (CD) algorithm

- At every iteration $\theta^{(t)}$, run a K-step Gibbs sampler sampling from $p_{\theta^{(t)}}(\mathbf{h} \mid \mathbf{v})$ and $p_{\theta^{(t)}}(\mathbf{v} \mid \mathbf{h})$.
- For RBM, these are simply batch draws from independent Bernoullis, because $\mathbf{h} \mid \mathbf{v}$ are conditionally independent, and so are $\mathbf{v} \mid \mathbf{h}$.
- If we allow *within layer* connections in $\mathbf{h}$ or $\mathbf{v}$, can't batch sample the Bernoullis anymore; explains why RBM is used and BM avoided!
- With $K \to \infty$ converge to $p_{\theta^{(t)}}(\mathbf{h}, \mathbf{v})$. Then can use Monte Carlo average to compute $\langle \mathbf{h}\mathbf{v}^T \rangle_{\text{model}}$.
- Hinton suggests using $K = 1$, because large K is computationally prohibitive! Persistent CD seems to work even better.

Towards full likelihood

- The success of CD has led to renewed interests in RBM-based architectures (e.g., DBM).
- Yet, it is know that CD-based solutions differ from maximum likelihood solutions (Sutskever and Tieleman, 2010).
- Classical results in statistics (Fisher, 1922; Rao, 1945) suggest asymptotic efficiency of likelihood-based solutions, so they are worth investigating.

Geyer (1991) estimate of $z(\theta)$

- Let $p_\theta(x) = q_\theta(x)/z(\theta)$, with $q_\theta(x)$ and $\nabla_\theta q_\theta(x)$ easy to evaluate.

- Suppose n i.i.d. data are observed. For $\theta, \phi \in \Theta$, we have:

$$\ell(\theta) - \ell(\phi) = \log \frac{p_\theta(\mathbf{X})}{p_\phi(\mathbf{X})} = \sum_{i=1}^{n} \log \frac{q_\theta(X_{i\bullet})}{q_\phi(X_{i\bullet})} - n \log \frac{z(\theta)}{z(\phi)}.$$

- Geyer (1991) proposed the importance sampling estimate:

$$\frac{z(\theta)}{z(\phi)} = \frac{1}{z(\phi)} \int q_\theta(x)dx = \frac{1}{z(\phi)} \int \frac{q_\theta(x)}{q_\phi(x)} q_\phi(x)dx = \mathbb{E}_{Y \sim p_\phi}\left[\frac{q_\theta(Y)}{q_\phi(Y)}\right],$$

motivating the Monte Carlo estimate $\frac{1}{N}\sum_{i=1}^{N} \frac{q_\theta(Y_{i\bullet})}{q_\phi(Y_{i\bullet})}$,

Our Monte Carlo estimate of $z(\theta)$ with $\phi = \mathrm{diag}(\theta)$

- Geyer's method is for generic θ, ϕ.

- Unbiasedness is guaranteed by construction, but the variance can become unbounded (see Geyer and Thompson, 1992 for examples)

- We choose the trial density $p_\phi(\cdot)$ with:

$$\boxed{\phi = \mathrm{diag}(\theta)}.$$

- Two key benefits:
  - $z(\phi) = \prod_{j=1}^{p} z(\theta_{jj})$ is known in closed form (product of univariate exponential family normalizing constants)
  - A sample $Y \sim p_\phi$ can be obtained by sampling $Y_j \sim p_{\theta_{jj}}$ independently and setting $Y = (Y_1, \ldots, Y_p)$.

Key propositions: bounded variance and exponential concentration of sample mean

- Monte Carlo estimate of $\nabla_\theta z(\theta)$ is similarly available as:

$$\frac{\nabla_\theta z(\theta)}{z(\phi)} = \mathbb{E}_{Y \sim p_\phi}\left[\frac{\nabla_\theta q_\theta(Y)}{q_\phi(Y)}\right].$$

- Main result 1: Monte Carlo estimates of $z(\theta)$ and $\nabla_\theta z(\theta)$ have bounded variances under mild conditions (see Prop. 3.2 of the paper).

- Main result 2: When the sample space is *bounded* (e.g., Ising, BMs), there is exponential concentration of the sample mean around the true mean (see Prop. 3.3 of the paper).

Maximum likelihood inference: fully observed case

- We use the Geyer estimates of $z(\theta)$ and $\nabla_\theta z(\theta)$.
- Projected gradient descent for MLE looks like:

$$\begin{aligned}\theta^{(t+1)} &= \mathcal{P}_\Theta\left(\theta^{(t)} + \gamma \nabla_\theta \ell(\theta^{(t)})\right) \\ &= \mathcal{P}_\Theta\left(\theta^{(t)} + \gamma \frac{\nabla_\theta\, q_{\theta^{(t)}}(\mathbf{X})}{q_{\theta^{(t)}}(\mathbf{X})} - \gamma \frac{\nabla_\theta z(\theta^{(t)})}{z(\theta^{(t)})}\right).\end{aligned}$$

- The projection $\mathcal{P}_\Theta$ is needed to ensure we stay in the valid parameter space (e.g., non-positive off-diagonals for Possion model).
- In high dimensions, we add an ℓ_1 penalty.

Bayesian inference: fully observed case

- In high dimensions, a blind random walk M–H encounters problems.
- Better to follow the gradient to propose a move.
- We use Hamiltonian Monte Carlo, again using the Geyer estimates of $z(\theta)$ and $\nabla_\theta z(\theta)$.
- We used scale mixtures of Laplace priors for the elements of θ.

Likelihood inference: partially observed case (RBMs)

- Consider BMs. In this case the data are $(\mathbf{h}, \mathbf{v})$ and the complete data model is Ising.

- Possible to use EM:

$$\theta^{(t+1)} = \theta^{(t)} + \gamma \left\{ \mathbb{P}(\mathbf{h} = \mathbf{1} \mid \mathbf{v}, \theta = \theta^{(t)})\mathbf{v}^T - \nabla_\theta \log z(\theta^{(t)}) \right\}$$

$$= \theta^{(t)} + \gamma \left\{ \mathbb{P}(\mathbf{h} = \mathbf{1} \mid \mathbf{v}, \theta = \theta^{(t)})\mathbf{v}^T - \frac{\mathbb{E}_{(\mathbf{h},\mathbf{v}) \sim p_\phi} \left[\frac{\nabla_\theta e^{-E_\theta(\mathbf{v},\mathbf{h})}}{e^{-E_\phi(\mathbf{v},\mathbf{h})}} \right]}{\mathbb{E}_{(\mathbf{h},\mathbf{v}) \sim p_\phi} \left[\frac{e^{-E_\theta(\mathbf{v},\mathbf{h})}}{e^{-E_\phi(\mathbf{v},\mathbf{h})}} \right]} \right\}.$$

- Bayesian inference is similar using HMC.

Inference for full BMs

- Recall again why RBMs are preferred over (full) BMs. Batch Gibbs sampling of $\mathbf{h} \mid \mathbf{v}$ and $\mathbf{v} \mid \mathbf{h}$ possible in RBM.

- In our case, the sampling model is the diagonal model ($\phi = \mathrm{diag}(\theta)$). It does not matter if there are within layer connections (within $\mathbf{h}$ or within $\mathbf{v}$) or not!

- The complete model $p_\theta(\mathbf{h}, \mathbf{v})$ and the conditional model $p_\theta(\mathbf{h} \mid \mathbf{v})$ are both Ising (exponential family), even though the marginal model $p_\theta(\mathbf{v})$ is not.

- Consequently, we can handle a full BM with this approach. We can also estimate the marginal likelihood for BMs by Chib's method:

$$p_\theta(\mathbf{v}) = p_\theta(\mathbf{h}, \mathbf{v}) / p_\theta(\mathbf{h} \mid \mathbf{v}).$$

Some theoretical results on likelihood based inference

- Under some assumptions, we are able to establish consistency of the ℓ_1 penalized estimator and posterior consistency of the Bayes estimator.
- There is support for the proposed gradient-based learning using the estimated gradient rather than true gradient.
- See paper for details.

Results: Ising for movie ratings data

- In the Movielens data (https://grouplens.org/datasets/movielens/), $162,000$ users rated $62,000$ movies on a scale 0–5, in increments of 0.5.

- We use a subset of $n = 303$ users who all rated the same $p = 50$ movies. We set:
$$X_{ij} = \mathbf{1}(\text{rating}_{ij} \geq 4.5),$$
denoting whether the ith user liked the jth movie.

- If $X_{ij} = X_{ik} = 1$, it means user i likes both movies j and k. Positive and negative values of $\{\theta_{jk}\}$ can now be interpreted as preferences.

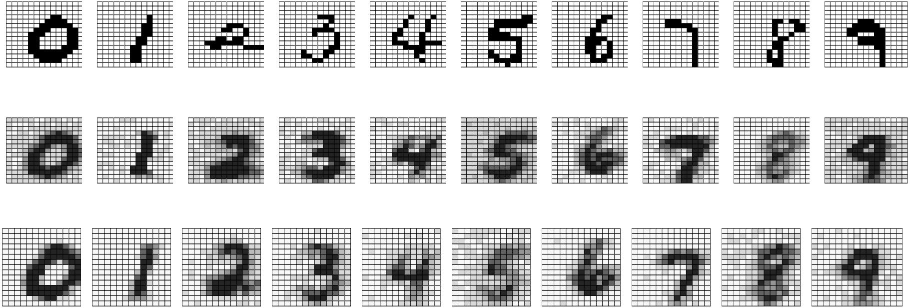
Results: Ising for movie ratings data

Positive Edge	$\hat{\theta}_{jk}$	Negative Edge	$\hat{\theta}_{jk}$
Lord of the Rings (2003) - Lord of the Rings (2001)	0.77	Inception (2010) - Batman (1989)	-0.95
Lord of the Rings (2002) - Lord of the Rings (2001)	0.72	Terminator (1984) - Dances with Wolves (1990)	-0.80
Lord of the Rings (2003) - Lord of the Rings (2002)	0.61	True Lies (1994) - Star Wars: Episode IV (1977)	-0.53
Star Wars: Episode V (1980) - Star Wars: Episode IV (1977)	0.56	Memento (2000) - Fugitive (1993)	-0.49
Terminator (1984) - Terminator 2: Judgment Day (1991)	0.55	Inception (2010) - Terminator (1984)	-0.48
Star Wars: Episode VI (1983) - Star Wars: Episode IV (1977)	0.53	Dances with Wolves (1990) - Twelve Monkeys (1995)	-0.44
Star Wars: Episode VI (1983) - Star Wars: Episode V (1980)	0.48	Independence Day (1996) - Batman (1989)	-0.44
Raiders of the Lost Ark (1981) - Star Wars: Episode V (1980)	0.40	Godfather (1972) - Independence Day (1996)	-0.42
Godfather (1972) - Schindler's List (1993)	0.25	Inception (2010)-Independence Day (1996)	-0.41
Raiders of the Lost Ark (1981) - Star Wars: Episode IV (1977)	0.24	Braveheart (1995) - Toy Story (1995)	-0.39

Table: Top 10 positive and negative interactions in the Movie Ratings Network.

- Clear Lord of the Rings and Star Wars clusters. Positive edges dominated by Spielberg–Lucas et al.
- We thought the negative edges were interesting. Batman, 1989 (Director: Tim Burton) is different in style than Christopher Nolan's Batman franchise (Director of Inception).
- No causal conclusions!

Results: BM and RBM for MNIST



- Row 1: Observed digits.
- Row 2: RBM-MLE reconstruction.
- Row 3: BM-MLE reconstruction.

Additional results

- We also analyzed count RNA-seq data in breast cancer using Poisson graphical models.
- BM seems to give good results with a fewer number of hidden nodes than RBMs.
- See paper for details.

Conclusions and future works

- Likelihood inference is difficult in intractable models because one needs to simulate auxiliary data from $p_\theta(\cdot)$ after every update of θ.
- Exchange algorithm (Murray et al., 2006), double MH (Liang, 2010), CD (Hinton, 2002) all suffer from this drawback.
- Sampling auxiliary data from $p_\phi(\cdot)$, the "diagonal model," offers crucial advantages.

Conclusions and future works

- Our estimates of $z(\theta)$ and $\nabla_\theta z(\theta)$ are unbiased.
- But, ratio of unbiased estimates is not in general unbiased (although it is consistent under mild conditions). We use this ratio estimator for $\nabla_\theta \log z(\theta) = \nabla_\theta z(\theta)/z(\theta)$.
- Thus, our HMC is an "approximate" MCMC (in the sense of Alquier et al., 2016). The estimate is not finite sample unbiased, and hence, not a pseudo-marginal approach (like exchange algorithm).
- We are working on a valid pseudo marginal scheme as well.

References

- Chen, Y., **Bhadra, A.** and Chakraborty, A. (2024+). Likelihood Based Inference in Fully and Partially Observed Exponential Family Graphical Models with Intractable Normalizing Constants. *(submitted).* [arXiv:2404.17763v1]
  - Code: `https://github.com/chenyujie1104/ExponentialGM`