STAT 479: Homework 5

Due: 11:59PM Mar 18, 2025 by Canvas

1. **True-False Conceptual Questions** (20 points)

   Write TRUE or FALSE for each of these statements.

   (a) The Bayesian Information Criterion (BIC) includes a penalty term for model complexity, discouraging overly complex models even if they fit the data well.

   (b) The Chow-Liu algorithm can learn Bayesian networks with arbitrary structure, including cycles.

   (c) In a causal graphical model, observing a common effect (collider) of two variables can create a spurious association between them, even if they were originally independent.

   (d) In causal discovery, controlling for a confounder ensures that any observed association between two variables must be causal.

   (e) The Expectation-Maximization (EM) algorithm improves the model parameters at each iteration by directly maximizing the likelihood of the observed data.

   (f) The likelihood function computed during EM is guaranteed to increase or remain constant at each iteration.

   (g) Variational inference approximates the posterior distribution by converting inference into an optimization problem that minimizes the Kullback-Leibler (KL) divergence.

   (h) The Evidence Lower Bound (ELBO) is a key objective in variational inference, providing a tractable lower bound on the log-likelihood.

   (i) The mean-field approximation in variational inference assumes that latent variables are independent, simplifying computation but potentially reducing accuracy.

   (j) Unlike Markov Chain Monte Carlo (MCMC), variational inference always produces an unbiased estimate of the posterior.

   Answer:

   (a)

   (b)

   (c)

(d)

(e)

(f)

(g)

(h)

(i)

(j)

2. **Deriving EM Updates for a Gaussian Mixture Model (GMM)** (20 points)

A Gaussian Mixture Model assumes that data points are generated from a mixture of K Gaussian distributions, each with a mean μ_k, covariance Σ_k, and a mixing weight π_k. As introduced in class, the Expectation-Maximization (EM) algorithm iteratively estimates these parameters.

(a) **Setting Up the Model**

We model the data as being drawn from K Gaussian components:

$$p(x|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

where π_k are the mixing weights summing to 1, and $\mathcal{N}(x|\mu_k, \Sigma_k)$ is a Gaussian density function. Which of the following best describes the latent variables in the GMM framework?

A. The mixing weights π_k that determine the prior probability of each Gaussian component.

B. The covariance matrices Σ_k, which control the shape of each Gaussian distribution.

C. The component assignments z_n, which indicate which Gaussian component generated each data point.

D. The observed data points x_n, which follow a mixture of Gaussians.

(b) **Expectation Step (E-Step)**

In the E-step, we compute the posterior responsibility γ_{nk}, which represents the probability that data point x_n was generated by component k:

$$\gamma_{nk} = p(z_n = k|x_n, \theta^{(t)}) = \frac{\pi_k^{(t)} \mathcal{N}(x_n|\mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} \pi_j^{(t)} \mathcal{N}(x_n|\mu_j^{(t)}, \Sigma_j^{(t)})}$$

What is the main role of γ_{nk} in the EM algorithm?

A. It represents the maximum likelihood estimate of the Gaussian parameters.

B. It updates the mixing weights to reflect the proportion of data points assigned to each cluster.

C. It acts as a "soft" assignment of each data point to the Gaussian components.

D. It maximizes the log-likelihood function directly.

(c) **Maximization Step (M-Step)**

In the M-step, we update the parameters of the Gaussians by maximizing the expected complete-data log-likelihood. The updates are:

$$\pi_k^{(t+1)} = \frac{1}{N} \sum_{n=1}^{N} \gamma_{nk}$$

$$\mu_k^{(t+1)} = \frac{\sum_{n=1}^{N} \gamma_{nk} x_n}{\sum_{n=1}^{N} \gamma_{nk}}$$

$$\Sigma_k^{(t+1)} = \frac{\sum_{n=1}^{N} \gamma_{nk} (x_n - \mu_k^{(t+1)})(x_n - \mu_k^{(t+1)})^T}{\sum_{n=1}^{N} \gamma_{nk}}$$

What does the M-step accomplish?

A. It reassigns each data point to a single Gaussian component.

B. It updates the model parameters to maximize the likelihood given the current soft assignments.

C. It computes the posterior probability of each data point belonging to a Gaussian component.

D. It eliminates one Gaussian component per iteration to simplify the model.

(d) **Convergence and Likelihood Maximization**

The EM algorithm repeats the E-step and M-step iteratively until convergence, typically when the log-likelihood:

$$\log p(X|\theta) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)$$

stabilizes.

Which of the following statements is **true** regarding the convergence properties of EM?

A. EM always finds the global maximum of the likelihood function.

B. EM maximizes a lower bound on the likelihood at each iteration, ensuring non-decreasing likelihood.
C. EM can decrease the likelihood in some iterations.
D. EM requires computing second-order derivatives to estimate parameter updates.

Answer:

(a)

(b)

(c)

(d)

3. **Causal Discovery** (20 points)

Causal discovery aims to infer causal relationships from observational data, often using DAGs to represent causality. Consider the following DAG:

$$X \to Z \leftarrow Y, \quad X \to W \to Y$$

Answer the following questions about this causal structure.

(a) Based on d-separation, are X and Y independent given no observed variables?
(b) Suppose we condition on Z. Are X and Y independent then?
(c) If you want to estimate the causal effect of X on Y, should you adjust for W? Why or why not?
(d) Suppose an additional variable U is added, where $U \to X$ and $U \to Y$. Explain how U acts as a confounder and how you would adjust for it to estimate the causal effect of X on Y.

Answer:

(a)

(b)

(c)

(d)

4. **I-Equivalence and Structure Discovery** (20 points)

Consider a Bayesian network structure learning problem where we aim to discover the best graphical representation of a given dataset. The **I-equivalence class** of a Bayesian network refers to the set of all graph structures that encode the same set of conditional independence relationships.

(a) **Defining I-Equivalence:** Two Bayesian networks are I-equivalent if they represent the same conditional independence relationships. Given the following two structures, determine whether they are I-equivalent and explain why.

$$G_1 : A \rightarrow B \rightarrow C, \quad G_2 : A \leftarrow B \rightarrow C$$

(b) **Learning DAGs vs. Learning I-Equivalence Classes:** Explain why learning the true DAG structure from observational data is harder than learning its I-equivalence class. What additional information would we need to uniquely determine the true DAG?

(c) **Graph Reversibility in I-Equivalence:** Suppose we have two DAGs that belong to the same I-equivalence class. How can we determine whether we can reverse an edge direction in one DAG while still preserving the same conditional independence structure? Provide an example.

(d) **Implications for Causal Discovery:** If two Bayesian networks are I-equivalent, does that mean they imply the same causal relationships? Why or why not?

Answer:

(a)

(b)

(c)

(d)

5. **Understanding Lower Bounds in Probabilistic Models** (20 points)

When training probabilistic models, we often approximate the data likelihood using a lower bound because directly computing the likelihood is intractable. We can define a sequence of progressively tighter bounds as follows:

$$\mathcal{L}_k(\mathbf{x}) = \mathbb{E}_{\mathbf{z}^{(1)},\ldots,\mathbf{z}^{(k)} \sim q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{1}{k} \sum_{i=1}^{k} \frac{p_\theta(\mathbf{x}, \mathbf{z}^{(i)})}{q_\phi(\mathbf{z}^{(i)} \mid \mathbf{x})} \right].$$

(a) **Jensen's Inequality Refresher:** Jensen's inequality states that for a concave function $f(x)$,

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)].$$

Show how Jensen's inequality implies that

$$\log p(\mathbf{x}) \geq \mathcal{L}_k(\mathbf{x})$$

for any $k \in \mathbb{N}$. Remember, $\log(\cdot)$ is a concave function.

(b) **Why More Samples Help:** Show that $\mathcal{L}_k(\mathbf{x})$ gets tighter as k increases: $\mathcal{L}_{k+1}(\mathbf{x}) > \mathcal{L}_k(\mathbf{x})$.

You can use the following lemma without proof:

Lemma: Let $I_k \subset [k+1] := \{1, \ldots, k+1\}$ with $|I_k| = k$ be a uniformly distributed subset of distinct indices from $[k+1]$. Then for any sequence of numbers $a_1, \ldots, a_{k+1}$,

$$\mathbb{E}_{I_k}\left[\frac{\sum_{i \in I_k} a_i}{k}\right] = \frac{\sum_{i=1}^{k+1} a_i}{k+1} \tag{1}$$

The above two results show that

$$\log p(\mathbf{x}) \geq \mathcal{L}_{k+1}(\mathbf{x}) \geq \mathcal{L}_k(\mathbf{x}).$$

However, the above inequalities do not guarantee $\mathcal{L}_k(\mathbf{x}) \to \log p(\mathbf{x})$ when $k \to \infty$. (The proof is left as an exercise to the reader. Or you can come to my office hours.)

Answer:

(a)

(b)

6. **OPTIONAL Markov Chain Monte Carlo Programming** (0 points)

This question is an OPTIONAL programming exercise. It will not impact your grade.

Nowadays, statistical modeling of sport data has become an important part of sports analytics and is often a critical reference for the managers in their decision-making process. In this part, we will work on a real world example in professional sports. Specifically, we are going to use the data from the 2013-2014 Premier League, the top-flight English professional league for men's football clubs, and build a predictive model on the number of goals scored in a single game by the two opponents. Bayesian hierarchical model is a good candidate for this kind of modeling task. We model each team's strength (both attacking and defending) as latent variables. Then in each game, the goals scored by the home team is a random variable conditioned on the attacking strength of the home team and the defending strength of the away team. Similarly, the goals scored by the away team is a random variable conditioned on the attack strength of the away team and the defense strength of the home team. Therefore, the distribution of the scoreline of a specific game is dependent on the relative strength between the home team A and the away team B, which also depends on the relative strength between those teams with their other opponents.

Table 1: 2013-2014 Premier League Teams

Index	0	1	2	3	4
Team	Arsenal	Aston Villa	Cardiff City	Chelsea	Crystal Palace
Index	5	6	7	8	9
Team	Everton	Fulham	Hull City	Liverpool	Manchester City
Index	10	11	12	13	14
Team	Manchester United	Newcastle United	Norwich City	Southampton	Stoke City
Index	15	16	17	18	19
Team	Sunderland	Swansea City	Tottenham Hotspurs	West Bromwich Albion	West Ham United

Here we consider using the same model as described by Baio and Blangiardo (2010). The Premier League has 20 teams, and we index them as in Table 1. Each team would play 38 matches every season (playing each of the other 19 teams home and away), which totals 380 games in the entire season. For the g-th game, assume that the index of home team is $h(g)$ and the index of the away team is $a(g)$. The observed number of goals (y_{g0}, y_{g1}) of home and away team is modeled as independent Poisson random variables:

$$y_{gj}|\theta_{gj} \sim \text{Poisson}(\theta_{gj}), \quad j = 0, 1 \tag{2}$$

where $\theta = (\theta_{g0}, \theta_{g1})$ represents the scoring intensity in the g-th game for the team playing at home ($j = 0$) and away ($j = 1$), respectively. We put a log-linear model for the θs:

$$\log \theta_{g0} = home + att_{h(g)} - def_{a(g)} \tag{3}$$

$$\log \theta_{g1} = att_{a(g)} - def_{h(g)} \tag{4}$$

Note that team strength is broken into attacking and defending strength. And home represents home-team advantage, and in this model is assumed to be constant across teams. The prior on the home is a normal distribution:

$$home \sim \mathcal{N}(0, \tau_0^{-1}) \tag{5}$$

where we set the precision $\tau_0 = 0.0001$.

The team-specific attacking and defending effects are modeled as:

$$att_t \sim \mathcal{N}(\mu_{att}, \tau_{att}^{-1}) \tag{6}$$

$$def_t \sim \mathcal{N}(\mu_{def}, \tau_{def}^{-1}) \tag{7}$$

We use conjugate priors as the hyper-priors on the attack and defense means and precisions:

$$\mu_{att} \sim \mathcal{N}(0, \tau_1^{-1}) \tag{8}$$

$$\mu_{def} \sim \mathcal{N}(0, \tau_1^{-1}) \tag{9}$$

$$\tau_{att} \sim \text{Gamma}(\alpha, \beta) \tag{10}$$

$$\tau_{def} \sim \text{Gamma}(\alpha, \beta) \tag{11}$$

where the precision $\tau_1 = 0.0001$, and we set parameters $\alpha = \beta = 0.1$.

This hierarchical Bayesian model can be represented using a directed acyclic graph as shown in Figure 1.

The goals of each game are $\mathbf{y} = \{y_{gj} | g = 0, 1, ..., 379, j = 0, 1\}$ are the observed variables, and parameters $\boldsymbol{\theta} = \{home, att_0, def_0, ..., att_{19}, def_{19}\}$ and hyper-parameters $\boldsymbol{\eta} = (\mu_{att}, \mu_{def}, \tau_{att}, \tau_{def})$ are unobserved variables that we need to make inference on. To ensure identifiability, we enforce a corner constraint on the parameters (pinning one team's parameters to 0,0). Here we use the first team as reference and assign its attacking and defending strength to be 0:

$$att_0 = def_0 = 0 \tag{12}$$

In this question, we want to estimate the posterior mean of the attacking and defending strength for each team, i.e. $\mathbb{E}_{p(\boldsymbol{\theta},\boldsymbol{\eta}|\mathbf{y})}[att_i]$, $\mathbb{E}_{p(\boldsymbol{\theta},\boldsymbol{\eta}|\mathbf{y})}[def_i]$, and $\mathbb{E}_{p(\boldsymbol{\theta},\boldsymbol{\eta}|\mathbf{y})}[home]$.

(a) Find the joint likelihood $p(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\eta})$.

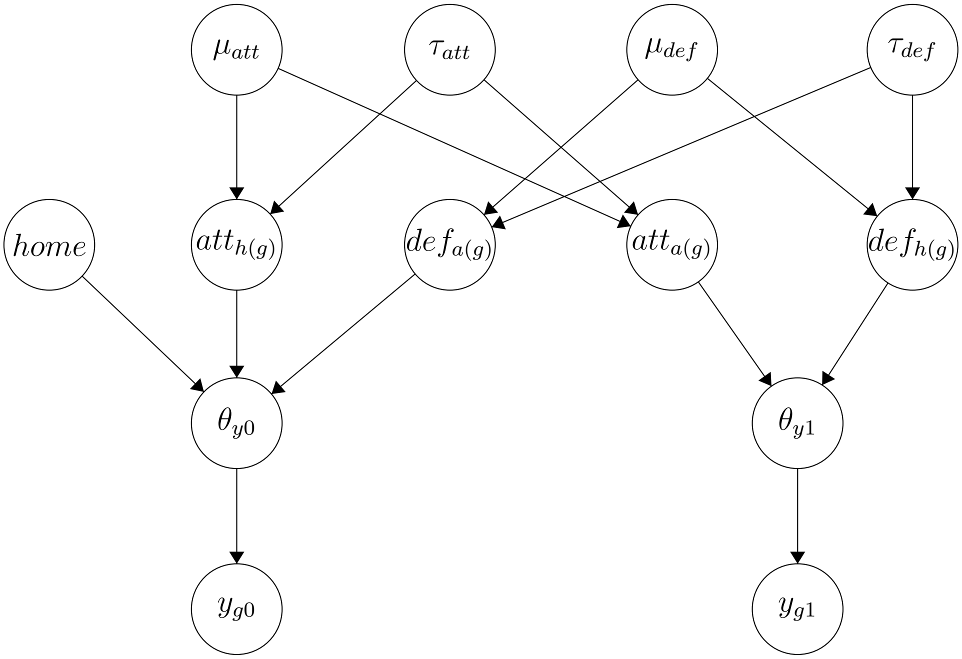

Figure 1: The DAG representation of the hierarchical Bayesian model. Figure adapted from Baio & Blangiardo.

(b) Write down the Metropolis-Hastings algorithm for sampling from posterior $p(\boldsymbol{\theta}, \boldsymbol{\eta}|\mathbf{y})$, and derive the acceptance function for a proposal distribution of your choice (e.g. isotropic Gaussian).

(c) Implement the Metropolis-Hastings algorithm to inference the posterior distribution. The data can be found from `https://lengerichlab.github.io/pgm-spring-2025/assets/hw/hw5/premier_league_2013_2014.dat`, which contains a 380×4 matrix. The first column is the number of goals y_{g0} scored by the home team, the second column is the number of goals y_{g1} scored by the away team, the third column is the index for the home team $h(g)$, and the fourth column is the index for the away team $a(g)$.

- Use an isotropic Gaussian proposal distribution $\mathcal{N}(0, \sigma^2 I)$ and use 0.1 as the starting point.
- Run the MCMC chain for 5000 steps to burn in and then collect 5000 samples with t steps in between (i.e., run M-H for $5000t$ steps and collect only each t-th sample). This is called thinning, which reduces the autocorrelation of the MCMC samples introduced by the Markovian process. The parameter sets are $\sigma = 0.005, 0.05, 0.5$, and $t = 1, 5, 20, 50$.
- Plot the trace plot of the burn in phase and the MCMC samples for the latent variable *home* using proposal distributions with different σ and t.
- Estimate the rejection ratio for each parameter setting, report your results

in a table.

- Comment on the results. Which parameter setting worked the best for the algorithm?
- Use the results from the optimal parameter setting:
    1. plot the posterior histogram of variable *home* from the MCMC samples.
    2. plot the estimated attacking strength $\mathbb{E}_{p(\boldsymbol{\theta},\boldsymbol{\eta}|\mathbf{y})}[att_i]$ against the estimated defending strength $\mathbb{E}_{p(\boldsymbol{\theta},\boldsymbol{\eta}|\mathbf{y})}[def_i]$ for each the team in one scatter plot. Please make sure to identify the team index of each point on your scatter plot using the index to team mapping in Table 1.