STCS 6701: Probabilistic Machine Learning
Homework 2

Due: Nov 14 at 11:59 pm ET

Instructions

- All homework should be typeset using LaTeX. Box your answers whenever appropriate.
- Standard late policy applies. Everyone has a total five late days throughout the semester. You are free to use them for whatever reason, no need to inform course staff.
- The homework should be turned in via Gradescope before the deadline (more details will be announced closer to the deadline).
- Turn in the code as well as the writeup.
- You can use any programming language you like.

Problem 1: Ideal Point Model

Motivation This problem is intended to give you an introduction to one of the most widely used latent variable models in political and social sciences. We consider a dataset of roll-call votes from the 113th U.S. Senate.

Your task is to uncover hidden ideological structure from these binary voting patterns. We will use a probit ideal-point model (a variant of Item Response Theory) that represents both senators and bills on a shared latent axis. This model is widely used to study polarization and party structure in real legislatures.

Question 1.1: The model

Data: $y_{ij} \in \{0, 1, \texttt{missing}\}$ = vote of senator i on bill j.
Latent Utility: We conceptualize the voting process as composed of the following elements

- Each senator has an ideological position on a hidden *left-right* axis.
- Each bill has a *location* on that axis: some are left-leaning, some are right-leaning, some are centrist.
- Some bills are more polarizing than others —a tax reform bill may split the chamber almost perfectly, while a ceremonial resolution passes nearly unanimously.
- A senator casts a "yea" if their **latent support** for a bill crosses some internal threshold, otherwise the senator casts a "nay" vote.

This hidden "support" for a bill is not observed directly, we only get to see yea/nay/didn't vote. So we imagine that behind every vote there is an unobserved continuous variable —a latent utility z_{ij}— that represents how strongly senator i supports bill j.

1. If $z_{ij} > 0$, the senator votes "yea."
2. If $z_{ij} < 0$, the senator votes "nay."

Now we want to connect z_{ij} to parameters that describe senators and bills.

1. **Senator position.** Suppose each senator has a hidden ideology θ_i on a *left-right* axis. If θ_i is large and positive, the senator is more conservative; if it is negative, more liberal.
2. **Bill location.** Suppose each bill has a threshold β_j, placing it on the same axis. A bill with $\beta_j = 0$ is centrist; a bill with $\beta_j = +2$ (i.e., some "large" arbitrary number) is very conservative; a bill with $\beta_j = -2$ (i.e., some "small" arbitrary number) is very liberal.
3. **Discrimination.** Not all bills are equally informative to a senator's ideology. Some bills divide senators sharply, others less so. To capture this we introduce a discrimination parameter α_j.

Therefore, latent utility takes the form

$$z_{ij} = \alpha_j(\theta_i - \beta_j) + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0, 1). \tag{1}$$
$$y_{ij} = 1(z_{ij} > 0) \tag{2}$$

a) **Explain in words:** If a senator's θ_i is far larger than a bill's β_j (and $\alpha_j > 0$), what does the model predict about the vote? What if θ_i is smaller?

b) **Role of** α_j: Compare two bills with the same β_j but different α_j: Which bill is more *polarizing*?

c) **Missing Votes:** How should we handle $y_{ij} = \texttt{missing}$ in this setup?

d) **Sketch**: Sketch the graphical model using plate notation.

e) **Sketch**: Draw a 1D lineshowing senators at positions θ_i, bills at positions β_j, and explain the threshold rule with a simple picture.

f) **Setting the Prior**: Suppose you choose a zero-centered prior for θ_i and β_j. How would you choose the prior variance(s) using the held-out data?

Question 1.2: Putting priors on the parameters

Right now, the latent parameters $\theta_i, \beta_j, \alpha_j$ are free-floating. To complete the model, we need to place prior distributions on these quantities.

Priors for $\theta_i, \beta_j, \alpha_j$. Throughout, assume the following priors

- $\theta_i \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$: senators' latent ideologies. (e.g., where a senator is in the political spectrum)
- $\beta_j \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2)$: bills' latent positions on the ideological axis. (e.g., where a bill is in the political spectrum).
- $\alpha_j \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2)$: how strongly a bill separates senators into yea/nay camps.

Identifiability. The model likelihood depends only on the product $\alpha_j(\theta_i - \beta_j)$. This leads to certain *symmetries* in the parameters:

1. **Translation:** Show that if we add a constant c to all senator positions and all bill positions,
$$(\theta_i - \beta_j) = (\theta_i + c) - (\beta_j + c),$$
the likelihood is unchanged. What does this mean about the absolute location of the ideological axis?
2. **Scaling:** Show that if we multiply all senator and bill positions by $k > 0$ and divide all discriminations by k,
$$\alpha_j(\theta_i - \beta_j) = \tfrac{\alpha_j}{k}(k\theta_i - k\beta_j),$$
the likelihood is unchanged. What does this mean about the scale of the ideological axis?
3. **Interpretation:** Do these symmetries affect how you interpret the parameters? Are there any other symmetries in the parameters?

Question 1.3: CAVI

In this section you will derive CAVI updates for this model.

a) **Joint distribution:** Using the priors on the previous question write down the joint distribution $p(y, z, \alpha, \theta, \beta)$.

b) **Latent utilities z_{ij}.** Recall that $z_{ij} \mid \theta_i, \beta_j, \alpha_j \sim \mathcal{N}(\alpha_j(\theta_i - \beta_j), 1)$. Show that conditioning on y_{ij} leads to a truncated Normal:

$$z_{ij} \mid y_{ij}, \theta, \beta, \alpha \sim \begin{cases} \mathcal{N}(\mu_{ij}, 1) \text{ truncated to } (0, \infty), & y_{ij} = 1, \\ \mathcal{N}(\mu_{ij}, 1) \text{ truncated to } (-\infty, 0], & y_{ij} = 0, \end{cases}$$

where $\mu_{ij} = \alpha_j(\theta_i - \beta_j)$. If $y_{ij} = -1$ (missing), explain why no z_{ij} is drawn.

c) **Senator positions θ_i.** Derive the conditional distribution of θ_i given z, β, α under your chosen prior from 2.1. Show it is Gaussian, and write down its mean and variance.

d) **Bill locations β_j.** Derive the conditional distribution of β_j given z, θ, α under your chosen prior from 2.1. Show it is Gaussian, and write down its mean and variance.

e) **Bill discriminations α_j.** Assume the prior you proposed in 2.1 for α_j. Derive the conditional distribution of α_j given z, θ, β. If you chose a Normal prior, it will be Normal; if you chose a truncated Normal prior, it will be truncated Normal (see Useful Formulas for information on the Truncated Normal).

f) **Marginalizing out z.** Write down $p(y, \alpha, \theta, \beta)$. What is $p(y|\alpha, \theta, \beta)$?

g) **Variational family.** Assume a factorization

$$q(\theta, \beta, \alpha, z) = \Big(\prod_{i=1}^{n} q(\theta_i)\Big)\Big(\prod_{j=1}^{d} q(\beta_j)\, q(\alpha_j)\Big)\Big(\prod_{i=1}^{n}\prod_{j=1}^{p} q(z_{ij})\Big),$$

where $q(\theta_i)$ and $q(\beta_j)$ are Normal, $q(\alpha_j)$ is either Normal or truncated Normal (depending on your chosen prior), and $q(z_{ij})$ is a truncated Normal as in 1.2(b). Write the full family explicitly and state which moments of each factor you will need for updates.

h) **Coordinate updates.** Using the identity

$$\log q^*(v) \propto \mathbb{E}_{-v}[\log p(y, z, \theta, \beta, \alpha)],$$

derive expressions for the optimal factors up to Normal/truncated Normal forms. Specifically:

(i) $q(z_{ij})$: update the mean parameter $\bar{\mu}_{ij}$ and give a formula for $\mathbb{E}_q[z_{ij}]$ using standard truncated-Normal moments (see Useful Formulas).

(ii) $q(\theta_i)$ and $q(\beta_j)$: write the precision and mean in terms of expectations $\mathbb{E}_q[\alpha_j]$, $\mathbb{E}_q[\alpha_j^2]$, and $\mathbb{E}_q[z_{ij}]$.

(iii) $q(\alpha_j)$: treat $\{z_{ij}\}_{i=1}^{n}$ as responses in a linear regression on $(\theta_i - \beta_j)$. Write the resulting mean and variance, and note how the update changes.

Problem 2: Gaussian Matrix Factorization (MFVI from conditionals + features)

§Model

Let $X \in \mathbb{R}^{n\times m}$ be a partially observed ratings matrix with observed index set $\Omega \subseteq \{1,\ldots,n\} \times \{1,\ldots,m\}$. Fix latent dimension K. For users i and items j we have factors $\theta_i, \beta_j \in \mathbb{R}^K$.

$$\theta_i \sim \mathcal{N}(0, \eta_\theta^2 I_K), \qquad \beta_j \sim \mathcal{N}(0, \eta_B^2 I_K), \qquad x_{ij} \mid \theta_i, \beta_j \sim \mathcal{N}(\theta_i^\top \beta_j, \sigma^2),\ (i,j) \in \Omega.$$

Define $\Omega_i = \{j : (i,j) \in \Omega\}$ and $\Omega^j = \{i : (i,j) \in \Omega\}$. Throughout, $\sigma^2, \eta_\theta^2, \eta_B^2$ are known.

(Given) Complete conditionals. You may use the following (conjugate) complete conditionals in your derivations.

$$p(\theta_i \mid \{\beta_j\}, X, \sigma^2, \eta_\theta^2) = \mathcal{N}\big(\mu_{\theta i},\ \Sigma_{\theta i}\big), \quad \Sigma_{\theta i}^{-1} = \eta_\theta^{-2} I_K + \sigma^{-2} \sum_{j\in\Omega_i} \beta_j \beta_j^\top, \quad \mu_{\theta i} = \Sigma_{\theta i}\, \sigma^{-2} \sum_{j\in\Omega_i} \beta_j\, x_{ij}, \tag{3}$$

$$p(\beta_j \mid \{\theta_i\}, X, \sigma^2, \eta_B^2) = \mathcal{N}\big(\mu_{\beta j},\ \Sigma_{\beta j}\big), \quad \Sigma_{\beta j}^{-1} = \eta_B^{-2} I_K + \sigma^{-2} \sum_{i\in\Omega^j} \theta_i \theta_i^\top, \quad \mu_{\beta j} = \Sigma_{\beta j}\, \sigma^{-2} \sum_{i\in\Omega^j} \theta_i\, x_{ij}. \tag{4}$$

Question 2.1: Mean-field VI (CAVI from the conditionals)

We approximate the posterior with a factorized family

$$q(\Theta, B) = \prod_{i=1}^n q(\theta_i) \prod_{j=1}^m q(\beta_j), \qquad q(\theta_i) = \mathcal{N}(m_{\theta i}, V_{\theta i}), \quad q(\beta_j) = \mathcal{N}(m_{\beta j}, V_{\beta j}).$$

a) **ELBO pieces.** Write the ELBO $\mathcal{L}(q)$ and list the expectations it comprises.

b) **CAVI updates.** Using the identity $\log q^\star(x_v) \propto \mathbb{E}_{-v}[\log p(x_v \mid x_{-v})]$ and the *given* complete conditionals 3 and 4, derive the optimal CAVI updates by replacing unknowns with their q-expectations. State the updates for $V_{\theta_i}^{-1}, V_{\beta_j}^{-1}, m_{\theta_i}$, and m_{β_j}.

c) **Algorithm sketch.** Give pseudocode for one CAVI sweep:

$$\{q(\theta_i)\}_{i=1}^n \ \to\ \{q(\beta_j)\}_{j=1}^m,$$

including which expectations are recomputed and a convergence criterion (e.g., ELBO monotone ascent or small parameter change).

d) **Flagging uncertain recommendations** Using q derive the approximate posterior predictive variance $\mathrm{Var}(x_{ij} \mid \text{data})$ on a holdout set (i.e.,$(i,j) \notin \Omega$) and explain how you would use this variance to flag uncertain recommendations.

e) **Setting hyperparameters** Explain how you could set η_θ^2, η_B^2, and σ^2 using heldout data.

Question 2.3: Adding item features (e.g., Genre) and updating CAVI

Let $g_j \in \mathbb{R}^p$ be a (possibly multi-hot) feature vector for item j (e.g., genres).

(additive side term). Augment the likelihood with an additive linear effect of features:

$$x_{ij} \mid \theta_i, \beta_j, \gamma \sim \mathcal{N}(\theta_i^\top \beta_j + g_j^\top \gamma,\ \sigma^2), \qquad \gamma \sim \mathcal{N}(0, \eta_\gamma^2 I_p).$$

(Given) Conditional for γ. You may use

$$
\begin{aligned}
p(\gamma \mid \Theta, B, X, G) &= \mathcal{N}(\mu_\gamma, \Sigma_\gamma) \\
\Sigma_\gamma^{-1} &= \eta_\gamma^{-2} I_p + \sigma^{-2} \sum_{(i,j)\in\Omega} g_j g_j^\top \\
\mu_\gamma &= \Sigma_\gamma \, \sigma^{-2} \sum_{(i,j)\in\Omega} g_j \big(x_{ij} - \theta_i^\top \beta_j\big).
\end{aligned}
$$

a) **Mean-field extension.** Extend the variational family to include $q(\gamma) = \mathcal{N}(m_\gamma, V_\gamma)$. Derive the CAVI update for $q(\gamma)$ by replacing $\theta_i^\top \beta_j$ with $\mathbb{E}_q[\theta_i]^\top \mathbb{E}_q[\beta_j]$.

b) **How do θ, β updates change?** Show that the *precisions* $V_{\theta i}^{-1}$ and $V_{\beta j}^{-1}$ are unchanged, and only the *means* are residualized by the feature term:

$$
m_{\theta i} = V_{\theta i} \, \sigma^{-2} \sum_{j\in\Omega_i} m_{\beta j} \big(x_{ij} - g_j^\top m_\gamma\big), \qquad m_{\beta j} = V_{\beta j} \, \sigma^{-2} \sum_{i\in\Omega^j} m_{\theta i} \big(x_{ij} - g_j^\top m_\gamma\big).
$$

Explain why this residualization is the only change.

c) **User features** Suppose we are also given user features $f_i \in \mathbb{R}^p$. Suggest a way of extending the model to account for this feature. Which CAVI updates would you expect to change and why?

Question 3: Implementation & Report (choose *one*)

Pick exactly **one** of the following and implement it end-to-end:[1]

- **Option A (Ideal Point Model; Probit IRT, 1D):** Implement a CAVI (using §1.2) on the 113th Senate dataset (votes.csv, senators.txt).
- **Option B (Gaussian Matrix Factorization):** Implement CAVI updates for the model in §2 on a standard explicit-feedback dataset (e.g. MovieLens 100K). Use the coordinate-wise updates you derived and evaluate out-of-sample.
- **Option C (Project-aligned Probabilistic Latent Factor Model):** If your final project involves a *probabilistic latent factor model*—interpreted broadly to include both linear matrix factorization and nonlinear variants such as variational autoencoders (VAEs) or other probabilistic latent-variable models—implement a minimal working version on a real dataset of your choice. Clearly describe the dataset and the generative process (likelihood, priors, and inference or optimization method). Be sure to **introduce the model and its probabilistic assumptions**, and **interpret the latent variables** (e.g., what structure or semantics they capture). Include both quantitative evaluation (e.g., held-out log-likelihood or predictive metrics) and qualitative visualizations that illuminate the learned latent structure.

What to implement (minimal checklist)

Common to both options

a) **Posterior predictive.**
   - *Option A:* For held-out (i, j), compute $\hat{p}_{ij} = \Phi\big(\hat{\alpha}_j(\hat{\theta}_i - \hat{\beta}_j)\big)$.
   - *Option B:* For held-out (i, j), compute $\widehat{x}_{ij} = \hat{\theta}_i^\top \hat{\beta}_j$.

b) **Metrics (pick at least two).**
   - *Binary (Option A):* average log-posterior-predictive on the held-out set + one metric of your choosing.
   - *Ratings (Option B):* Average log-posterior-predictive on a held-out set + one metric of your choosing.

c) **Convergence diagnostics.**

Option A specific (Ideal Point)

d) **Substantive plots.**
   - **Ideology axis:** posterior means of θ_i with intervals, colored by party; label ∼5 outliers.
   - **Bill landscape:** scatter of $(\hat{\beta}_j, \hat{\alpha}_j)$; annotate ∼5 highest $\hat{\alpha}_j$ bills.

e) **Implementation details.** Plot the ELBO vs iteration, number of iterations to convergence etc,.

Option B specific (Matrix Factorization)

f) **Latent dimension sweep.** Run for $K \in \{2, 5, 10, 20\}$ (or a comparable set) and report RMSE/MAE vs. K.

g) **Performance with vs without genre as a feature**.

[1] For MAP and coordinatewise MAP implementations, feel free to use automatic differentiation frameworks like torch, pyro, jax, or tensorflow.

h) **Uncertainty (first-order).** Estimate $\mathrm{Var}(x_{ij} \mid \text{data})$, and discuss how you could use this to flag uncertain recommendations (estimate this analytically or by drawing samples from the predictive).

i) **Visualization (for $K = 2$).** Scatter plots of $\{\hat{\theta}_i\}$ and $\{\hat{\beta}_j\}$; comment on clusters (genres/users).

j) **Convergence checks** Plot the ELBO vs iteration.

What to turn in

Please submit a short report (**2–4 pages of main text**; this is a *soft limit*, but avoid going significantly over 4 pages). Your writeup should read like a concise research note describing what you set out to do, how you did it, and what you found. Your report must include all the elements listed in the *What to implement* section above.

The structure below is intended as a set of *guidelines* to help organize your report, rather than a rigid template.

i. **Introduction.** State the problem you are addressing and briefly motivate why the modeling approach is appropriate.

ii. **Model.** Write down the full joint distribution for your generative model, specifying all distributions and parameterizations (e.g., Gamma shape–rate). Include a graphical model if relevant.

iii. **Inference.** Summarize your inference approach and main implementation choices in the main text (details may go in an Appendix).

iv. **Data & setup.** Describe your dataset or image, preprocessing steps, choice of K, priors, and any design decisions. Compare alternative choices where appropriate.

v. **Results.** Present key outcomes: number of clusters in factors, posterior summaries of component parameters, and representative visualizations of the latent factors, quantitative checks of model fits etc.

Question 4: Final Project

Write an “aspirational abstract” for your final project. Note you are not committed to deliver everything you mention on the abstract. Rather, preparing the abstract is a chance to think concretely and envision a successful final project.

Formulas & Identities You May Find Useful

Notation. $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard Normal pdf and cdf. For a Normal truncated to (a, b) we use the shorthands $\alpha = (a-\mu)/\sigma$, $\beta = (b-\mu)/\sigma$, and $Z = \Phi(\beta) - \Phi(\alpha)$.

A. Probit link and data augmentation

If $z \sim \mathcal{N}(\mu, 1)$ and $y = \mathbb{I}\{z > 0\}$, then

$$\mathbb{P}(y = 1) = \Phi(\mu), \qquad \log p(y \mid \mu) = y \log \Phi(\mu) + (1-y) \log\big(1 - \Phi(\mu)\big).$$

B. Truncated Normal moments (general and one-sided)

Let $Y \sim \mathcal{N}(\mu, \sigma^2)$ truncated to (a, b) with the α, β, Z above. Then

$$\mathbb{E}[Y] = \mu + \sigma \cdot \lambda, \quad \text{where} \;\; \lambda = \frac{\phi(\alpha) - \phi(\beta)}{Z},$$

$$\mathrm{Var}(Y) = \sigma^2 \left[1 + \frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{Z} - \lambda^2\right], \qquad \mathbb{E}[Y^2] = \mathrm{Var}(Y) + \big(\mathbb{E}[Y]\big)^2.$$

One-sided cases:

$$Y \sim \mathcal{N}(\mu, \sigma^2) \text{ truncated to } (0, \infty): \;\; \alpha = \tfrac{-\mu}{\sigma}, \;\; Z = 1 - \Phi(\alpha), \;\; \lambda = \tfrac{\phi(\alpha)}{Z}.$$

$$Y \sim \mathcal{N}(\mu, \sigma^2) \text{ truncated to } (-\infty, 0]: \;\; \alpha = \tfrac{0-\mu}{\sigma}, \;\; Z = \Phi(\alpha), \;\; \lambda = \tfrac{-\phi(\alpha)}{Z}.$$

Half-Normal (special case). If $\mu = 0$ and truncation is $(0, \infty)$, Y is Half-Normal: $\mathbb{E}[Y] = \sigma\sqrt{2/\pi}$, $\mathrm{Var}(Y) = \sigma^2\big(1 - 2/\pi\big)$. *Note:* Using $|X|$ with $X \sim \mathcal{N}(0, \sigma^2)$ samples this case correctly. For $\mu \neq 0$, $|X|$ *does not* produce the correct truncated Normal—use inverse-CDF, rejection sampling, or an off-the-shelf routine.

C. Sampling from a truncated Normal

Sampling is typically done with either rejection sampling using the inverse cdf or using exponential rejection schemes. However, **in practice, you may just want to use an off-the-shelf sampler like** `scipy.stats.truncnorm`.

D. Gaussian identities for CAVI

If $x \sim \mathcal{N}(m, V)$ then

$$\mathbb{E}[x] = m, \qquad \mathbb{E}[xx^\top] = V + mm^\top.$$

If $x \sim \mathcal{N}(m_x, V_x)$ and $y \sim \mathcal{N}(m_y, V_y)$ are independent, then

$$\mathbb{E}[x^\top y] = m_x^\top m_y, \quad \mathrm{Var}(x^\top y) = m_x^\top V_y m_x + m_y^\top V_x m_y + \mathrm{tr}(V_x V_y).$$