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: $y_{ij} \in \{0, 1, \text{missing}\}$ = 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

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

$$y_{ij} = \nVdash(z_{ij} > 0) \tag{2}$$

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?

We are given

$$\begin{aligned} z_{ij} &= \alpha_j(\theta_i - \beta_j) + \epsilon_{ij}, & \epsilon_{ij} &\sim \mathcal{N}(0,1), \\ y_{ij} &= \nVdash(z_{ij} > 0). \end{aligned}$$

Define the mean of the latent utility

$$\mu_{ij} \triangleq \alpha_j(\theta_i - \beta_j).$$

Then we can rewrite $z_{ij} = \mu_{ij} + \epsilon_{ij}$, so conditional on $(\theta_i, \alpha_j, \beta_j)$ we have

$$z_{ij} \mid \theta_i, \alpha_j, \beta_j \sim \mathcal{N}(\mu_{ij}, 1).$$

Before proceeding with Question 1.1(a), let us derive the probability that the model produces a "yea":

$$\begin{aligned} \Pr(y_{ij} = 1 \mid \theta_i, \alpha_j, \beta_j) &= \Pr(z_{ij} > 0 \mid \theta_i, \alpha_j, \beta_j) \\ &= \Pr(\mu_{ij} + \epsilon_{ij} > 0) \\ &= \Pr(\epsilon_{ij} > -\mu_{ij}) \\ &= 1 - \Phi(-\mu_{ij}) = \Phi(\mu_{ij}) = \Phi\big(\alpha_j(\theta_i - \beta_j)\big), \end{aligned}$$

where $\Phi(\cdot)$ is the standard normal CDF and we used the symmetry $\Phi(-x) = 1 - \Phi(x)$. Similarly,

$$\Pr(y_{ij} = 0 \mid \theta_i, \alpha_j, \beta_j) = 1 - \Phi\big(\alpha_j(\theta_i - \beta_j)\big).$$

Case 1: θ_i is far larger than β_j. Assume $\theta_i \gg \beta_j$ and $\alpha_j > 0$. Then $\theta_i - \beta_j \gg 0$ and hence $\mu_{ij} = \alpha_j(\theta_i - \beta_j) \gg 0$. Because $\Phi(\mu_{ij})$ is very close to 1 when μ_{ij} is large and positive, the model predicts $y_{ij} = 1$ with probability near one: the latent utility is almost surely positive, so the senator is very likely to vote "yea."

Case 2: θ_i is far smaller than β_j. Assume $\theta_i \ll \beta_j$ and $\alpha_j > 0$. Then $\theta_i - \beta_j \ll 0$, implying $\mu_{ij} \ll 0$. Now $\Phi(\mu_{ij})$ is close to 0, so $\Pr(y_{ij} = 1 \mid \theta_i, \alpha_j, \beta_j)$ is near zero. In this situation the latent utility is very likely negative, so the senator will almost surely vote "nay."

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

For a fixed senator i and bill j, the latent utility satisfies

$$\begin{aligned} z_{ij} &= \alpha_j(\theta_i - \beta_j) + \epsilon_{ij}, & \epsilon_{ij} &\sim \mathcal{N}(0,1), \\ y_{ij} &= \nVdash(z_{ij} > 0). \end{aligned}$$

From the preliminary derivation we already have

$$\Pr(y_{ij} = 1 \mid \theta_i, \alpha_j, \beta_j) = \Phi\big(\alpha_j(\theta_i - \beta_j)\big).$$

Define $p_j(\theta_i) \triangleq \Pr(y_{ij} = 1 \mid \theta_i, \alpha_j, \beta_j)$ and introduce $u = \alpha_j(\theta_i - \beta_j)$ so that $p_j(\theta_i) = \Phi(u)$. Differentiating with respect to θ_i gives

$$\frac{\partial p_j(\theta_i)}{\partial \theta_i} = \phi(u)\,\alpha_j = \alpha_j\,\phi\big(\alpha_j(\theta_i - \beta_j)\big),$$

where $\phi(\cdot)$ is the standard normal density. Thus α_j controls how sharply the vote probability $p_j(\theta_i)$ rises or falls along the ideological axis.

Now compare two bills j and k with the same location β but different discriminations $\alpha_j > \alpha_k > 0$. Their vote probabilities for senator i are

$$p_j(\theta_i) = \Phi\big(\alpha_j(\theta_i - \beta)\big), \qquad p_k(\theta_i) = \Phi\big(\alpha_k(\theta_i - \beta)\big).$$

Case 1: θ_i near the bill location β. When $\theta_i = \beta$ exactly we obtain $p_j(\beta) = p_k(\beta) = \Phi(0) = 0.5$. For θ_i close (but not equal) to β, the arguments $\alpha_j(\theta_i - \beta)$ and $\alpha_k(\theta_i - \beta)$ remain close to zero, so both probabilities stay near 0.5. However, the derivatives at $\theta_i = \beta$ are

$$\frac{\partial p_j(\theta_i)}{\partial \theta_i}\bigg|_{\theta_i=\beta} = \alpha_j \phi(0), \qquad \frac{\partial p_k(\theta_i)}{\partial \theta_i}\bigg|_{\theta_i=\beta} = \alpha_k \phi(0),$$

and since $\alpha_j > \alpha_k$, the curve $p_j(\theta_i)$ changes more rapidly with ideology near the cutpoint. In other words, a small ideological shift around β causes a larger swing in the probability of voting "yea" for bill j than for bill k; bill j is therefore more polarizing.

Case 2: θ_i far from β. If $|\theta_i - \beta| \to \infty$, then $\alpha_j(\theta_i - \beta)$ and $\alpha_k(\theta_i - \beta)$ tend to $\pm\infty$ with the same sign, so

$$p_j(\theta_i) \to \nVdash(\theta_i > \beta), \qquad p_k(\theta_i) \to \nVdash(\theta_i > \beta),$$

and both probabilities saturate at 0 or 1. In this regime the derivatives shrink to zero because the Gaussian density $\phi\big(\alpha_\ell(\theta_i - \beta)\big)$ vanishes as its argument diverges. The main difference between j and k therefore lies in how quickly they transition between the extremes: larger α_j yields a steeper sigmoid, so p_j leaves the ambiguous middle region more abruptly.

Overall, holding $\beta_j = \beta_k = \beta$ and $\alpha_j > \alpha_k > 0$, bill j is more polarizing than bill k because its discrimination parameter makes the probability curve $p_j(\theta_i)$ change more steeply (and thus more abruptly) from near 0 to near 1 as a senator's ideology moves across β.

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

We start from the model, for each (i, j):

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

From the earlier derivation, we obtained an explicit expression for the voting probability:

$$\Pr(y_{ij} = 1 \mid \theta_i, \alpha_j, \beta_j) = \Pr(z_{ij} > 0 \mid \theta_i, \alpha_j, \beta_j) = \Phi\big(\alpha_j(\theta_i - \beta_j)\big),$$

where $\Phi(\cdot)$ is the standard normal CDF. Therefore,

$$\Pr(y_{ij} = 0 \mid \theta_i, \alpha_j, \beta_j) = 1 - \Pr(y_{ij} = 1 \mid \theta_i, \alpha_j, \beta_j) = 1 - \Phi\big(\alpha_j(\theta_i - \beta_j)\big).$$

Bernoulli representation. Define $p_{ij} := \Pr(y_{ij} = 1 \mid \theta_i, \alpha_j, \beta_j) = \Phi\big(\alpha_j(\theta_i - \beta_j)\big)$. Then, for a single observed vote $y_{ij} \in \{0, 1\}$,

$$\Pr(y_{ij} \mid \theta_i, \alpha_j, \beta_j) = p_{ij}^{y_{ij}} (1 - p_{ij})^{1 - y_{ij}}.$$

Substituting $p_{ij} = \Phi\big(\alpha_j(\theta_i - \beta_j)\big)$ gives

$$\Pr(y_{ij} \mid \theta_i, \alpha_j, \beta_j) = [\Phi\big(\alpha_j(\theta_i - \beta_j)\big)]^{y_{ij}} [1 - \Phi\big(\alpha_j(\theta_i - \beta_j)\big)]^{1 - y_{ij}}.$$

Joint likelihood. Define the index set of non-missing votes $O := \{(i,j) : y_{ij} \in \{0,1\}\}$. Assuming conditional independence of votes given $\{\theta_i\}$ and $\{\alpha_j, \beta_j\}$,

$$p(\{y_{ij}\}_{(i,j)\in O} \mid \{\theta_i\}, \{\alpha_j, \beta_j\}) = \prod_{(i,j)\in O} \Pr(y_{ij} \mid \theta_i, \alpha_j, \beta_j),$$

which by the Bernoulli form becomes

$$p(\{y_{ij}\}_{(i,j)\in O} \mid \{\theta_i\}, \{\alpha_j, \beta_j\}) = \prod_{(i,j)\in O} [\Phi\big(\alpha_j(\theta_i - \beta_j)\big)]^{y_{ij}} [1 - \Phi\big(\alpha_j(\theta_i - \beta_j)\big)]^{1-y_{ij}}.$$

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?

1. Specify the priors with unknown variances. Choose zero-centered Gaussian priors for the latent senator and bill locations:

$$\begin{aligned} \theta_i &\sim \mathcal{N}(0, \sigma_\theta^2), \quad i = 1, \ldots, N_{\text{sen}}, \\ \beta_j &\sim \mathcal{N}(0, \sigma_\beta^2), \quad j = 1, \ldots, N_{\text{bill}}. \end{aligned}$$

Treat $\sigma_\theta^2, \sigma_\beta^2$ as hyperparameters that we will choose using the held-out votes.

2. Split the data: training vs held-out. Let O be the set of observed (non-missing) votes as before, and split it into

- a training set O_{train},
- a held-out set O_{val},

with $O_{\text{train}} \cap O_{\text{val}} = \emptyset$ and $O_{\text{train}} \cup O_{\text{val}} = O$.

We will fit the model on O_{train} for each candidate choice of $(\sigma_\theta^2, \sigma_\beta^2)$, and evaluate predictive performance on O_{val}.

3. For fixed variances, fit θ, β on the training set. Recall the Bernoulli form:

$$\begin{aligned} p_{ij}(\theta_i, \beta_j) &:= \Pr(y_{ij} = 1 \mid \theta_i, \alpha_j, \beta_j) = \Phi\big(\alpha_j(\theta_i - \beta_j)\big), \\ \Pr(y_{ij} \mid \theta_i, \alpha_j, \beta_j) &= p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}}. \end{aligned}$$

For a fixed pair $(\sigma_\theta^2, \sigma_\beta^2)$, the training log-posterior (up to an additive constant) is

$$\begin{aligned} \ell_{\text{train}}(\theta, \beta \mid \sigma_\theta^2, \sigma_\beta^2) = &\sum_{(i,j)\in O_{\text{train}}} [y_{ij} \log p_{ij}(\theta_i, \beta_j) + (1 - y_{ij}) \log(1 - p_{ij}(\theta_i, \beta_j))] \\ &- \frac{1}{2\sigma_\theta^2} \sum_i \theta_i^2 - \frac{1}{2\sigma_\beta^2} \sum_j \beta_j^2. \end{aligned}$$

For that choice of $(\sigma_\theta^2, \sigma_\beta^2)$, compute the MAP estimates

$$\hat{\theta}(\sigma_\theta^2, \sigma_\beta^2), \quad \hat{\beta}(\sigma_\theta^2, \sigma_\beta^2)$$

by maximizing ℓ_{train}.

4. Compute held-out predictive log-likelihood. Using $\hat{\theta}, \hat{\beta}$, compute predicted probabilities on held-out pairs:

$$\hat{p}_{ij} := \Phi\big(\alpha_j(\hat{\theta}_i - \hat{\beta}_j)\big), \quad (i,j) \in O_{\text{val}}.$$

The held-out log-likelihood for this choice of variances is

$$L_{\text{val}}(\sigma_\theta^2, \sigma_\beta^2) := \sum_{(i,j)\in O_{\text{val}}} \left[y_{ij} \log \hat{p}_{ij} + (1 - y_{ij}) \log(1 - \hat{p}_{ij})\right].$$

5. Choose the prior variances by maximizing held-out performance. Finally, treat $\sigma_\theta^2, \sigma_\beta^2$ as tuning parameters and choose them to maximize the held-out log-likelihood:

$$(\sigma_\theta^{2*}, \sigma_\beta^{2*}) = \arg \max_{(\sigma_\theta^2, \sigma_\beta^2)} L_{\text{val}}(\sigma_\theta^2, \sigma_\beta^2).$$

In practice one would:

- pick a grid of candidate values for σ_θ^2 and σ_β^2,
- for each pair, fit MAP $\hat{\theta}, \hat{\beta}$ on O_{train},
- compute L_{val},
- select the pair that yields the largest L_{val}.

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?

We already have, from 1.1, for each (i, j),

$$\Pr(y_{ij} = 1 \mid \theta_i, \alpha_j, \beta_j) = \Phi\big(\alpha_j(\theta_i - \beta_j)\big),$$

and

$$\Pr(y_{ij} = 0 \mid \theta_i, \alpha_j, \beta_j) = 1 - \Phi\big(\alpha_j(\theta_i - \beta_j)\big).$$

So the likelihood depends on the parameters only through the term $\alpha_j(\theta_i - \beta_j)$.

Show translation invariance. Consider a constant $c \in \mathbb{R}$ and define transformed parameters

$$\theta_i' = \theta_i + c, \qquad \beta_j' = \beta_j + c, \qquad \alpha_j' = \alpha_j.$$

Compute the transformed difference:

$$\theta_i' - \beta_j' = (\theta_i + c) - (\beta_j + c) = \theta_i - \beta_j.$$

Then the probit argument under the transformed parameters is

$$\alpha_j'(\theta_i' - \beta_j') = \alpha_j(\theta_i - \beta_j),$$

so the "yea" probability is unchanged:

$$\Pr(y_{ij} = 1 \mid \theta_i', \alpha_j', \beta_j') = \Phi\big(\alpha_j'(\theta_i' - \beta_j')\big) = \Phi\big(\alpha_j(\theta_i - \beta_j)\big) = \Pr(y_{ij} = 1 \mid \theta_i, \alpha_j, \beta_j).$$

The same holds for $\Pr(y_{ij} = 0 \mid \cdot)$, so for each (i, j),

$$\Pr(y_{ij} \mid \theta_i', \alpha_j', \beta_j') = \Pr(y_{ij} \mid \theta_i, \alpha_j, \beta_j).$$

Because the joint likelihood is the product over (i, j) of these terms, the entire likelihood is unchanged by the transformation

$$\theta_i \mapsto \theta_i + c, \qquad \beta_j \mapsto \beta_j + c, \qquad \alpha_j \text{ unchanged.}$$

This shows that the absolute location (origin) of the ideological axis is not identified by the data: only relative positions $\theta_i - \beta_j$ matter. In other words, the model can only recover ideologies and bill locations up to an additive constant; choosing where "0" is on the axis is a matter of convention (e.g., setting $\mu_\theta = 0$, or fixing one $\beta_j = 0$).

2. **Scaling:** Show that if we multiply all senator and bill positions by $k > 0$ and divide all discriminations by k,
$$\alpha_j(\theta_i - \beta_j) = \frac{\alpha_j}{k}(k\theta_i - k\beta_j),$$
the likelihood is unchanged. What does this mean about the scale of the ideological axis?

We already know from 1.1 that, for each (i, j),

$$\Pr(y_{ij} = 1 \mid \theta_i, \alpha_j, \beta_j) = \Phi\big(\alpha_j(\theta_i - \beta_j)\big),$$

$$\Pr(y_{ij} = 0 \mid \theta_i, \alpha_j, \beta_j) = 1 - \Phi\big(\alpha_j(\theta_i - \beta_j)\big).$$

So again the likelihood depends on the parameters only through the product

$$\alpha_j(\theta_i - \beta_j).$$

Show scaling invariance. Take a constant $k > 0$ and define the transformed parameters

$$\theta'_i = k\,\theta_i, \qquad \beta'_j = k\,\beta_j, \qquad \alpha'_j = \frac{\alpha_j}{k}.$$

First compute the transformed difference:

$$\theta'_i - \beta'_j = k\theta_i - k\beta_j = k(\theta_i - \beta_j).$$

Now plug into the probit argument with the transformed discrimination:

$$\alpha'_j(\theta'_i - \beta'_j) = \frac{\alpha_j}{k}\big[k(\theta_i - \beta_j)\big] = \alpha_j(\theta_i - \beta_j).$$

Therefore, for every (i, j),

$$\Pr(y_{ij} = 1 \mid \theta'_i, \alpha'_j, \beta'_j) = \Phi\big(\alpha'_j(\theta'_i - \beta'_j)\big) = \Phi\big(\alpha_j(\theta_i - \beta_j)\big) = \Pr(y_{ij} = 1 \mid \theta_i, \alpha_j, \beta_j),$$

and similarly
$$\Pr(y_{ij} = 0 \mid \theta'_i, \alpha'_j, \beta'_j) = \Pr(y_{ij} = 0 \mid \theta_i, \alpha_j, \beta_j).$$

Thus for each observation,

$$\Pr(y_{ij} \mid \theta'_i, \alpha'_j, \beta'_j) = \Pr(y_{ij} \mid \theta_i, \alpha_j, \beta_j),$$

and the product over all (i, j) (the joint likelihood) is unchanged by the transformation

$$\theta_i \mapsto k\theta_i, \qquad \beta_j \mapsto k\beta_j, \qquad \alpha_j \mapsto \alpha_j/k.$$

This shows that the scale (units) of the ideological axis is not identified by the data: if we stretch or compress all θ_i and β_j by a positive factor k and compensate by shrinking α_j by $1/k$, the likelihood is exactly the same. The model only identifies ideological positions up to a multiplicative constant; choosing the "unit" of ideology (e.g. fixing $\sigma^2_\theta = 1$ or constraining the variance of θ_i) is a matter of convention needed to fix the scale.

3. **Interpretation:** Do these symmetries affect how you interpret the parameters? Are there any other symmetries in the parameters?

   The translation and scaling symmetries mean that the absolute origin and units of the ideological axis are not identified by the data. One can shift all θ_i and β_j by a constant, or multiply them all by a positive constant and rescale α_j accordingly, without changing any vote probabilities. As a result, θ_i and β_j are only interpretable up to an affine transformation (shift and rescale); what is identified are relative positions, orderings, and distances, not the raw numerical values themselves.

   If the sign of α_j is not constrained, there is one more symmetry: a global sign flip

   $$\theta_i \mapsto -\theta_i, \qquad \beta_j \mapsto -\beta_j, \qquad \alpha_j \mapsto -\alpha_j$$

   also leaves the likelihood unchanged. This means the orientation of the ideological axis ("which side is left/right") is arbitrary until one imposes a convention (for example, fixing some known liberal to have $\theta_i > 0$ or requiring most $\alpha_j > 0$).

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 $p(y, z, \alpha, \theta, \beta)$.

Model pieces. From before, for each pair (i, j):

$$\begin{aligned} z_{ij} &= \alpha_j(\theta_i - \beta_j) + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim \mathcal{N}(0, 1), \\ y_{ij} &= \nVdash(z_{ij} > 0). \end{aligned}$$

Priors:

Senators' ideologies:

$$\theta_i \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2), \quad i = 1, \ldots, N.$$

Bills' locations:

$$\beta_j \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2), \quad j = 1, \ldots, J.$$

Bills' discriminations:

$$\alpha_j \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2), \quad j = 1, \ldots, J.$$

For CAVI, it is convenient to make explicit the conditional distributions:

Latent utilities:

$$z_{ij} \mid \theta_i, \alpha_j, \beta_j \sim \mathcal{N}\big(\alpha_j(\theta_i - \beta_j), 1\big).$$

Observed votes are deterministic given z_{ij}:

$$p(y_{ij} \mid z_{ij}) = \nVdash\{y_{ij} = 1, z_{ij} > 0\} + \nVdash\{y_{ij} = 0, z_{ij} \leq 0\}.$$

Factorization of the joint. Collect notation:

$$y = \{y_{ij}\}, \quad z = \{z_{ij}\}, \quad \theta = \{\theta_i\}, \quad \beta = \{\beta_j\}, \quad \alpha = \{\alpha_j\}.$$

Using conditional independence (given θ, β, α, all z_{ij} are independent across i, j; given z_{ij}, the y_{ij} are independent), the joint distribution factors as

$$p(y, z, \alpha, \theta, \beta) = \left[\prod_{i=1}^{N} p(\theta_i)\right] \left[\prod_{j=1}^{J} p(\beta_j)\right] \left[\prod_{j=1}^{J} p(\alpha_j)\right] \left[\prod_{i=1}^{N}\prod_{j=1}^{J} p(z_{ij} \mid \theta_i, \alpha_j, \beta_j)\right] \left[\prod_{i=1}^{N}\prod_{j=1}^{J} p(y_{ij} \mid z_{ij})\right].$$

Now plug in the specific normal and indicator forms:

Priors:

$$p(\theta_i) = \mathcal{N}(\theta_i \mid \mu_\theta, \sigma_\theta^2), \quad p(\beta_j) = \mathcal{N}(\beta_j \mid \mu_\beta, \sigma_\beta^2), \quad p(\alpha_j) = \mathcal{N}(\alpha_j \mid \mu_\alpha, \sigma_\alpha^2),$$

Latent utilities:

$$p(z_{ij} \mid \theta_i, \alpha_j, \beta_j) = \mathcal{N}\big(z_{ij} \mid \alpha_j(\theta_i - \beta_j), 1\big),$$

Vote indicators:

$$p(y_{ij} \mid z_{ij}) = \nVdash\{y_{ij} = 1,\, z_{ij} > 0\} + \nVdash\{y_{ij} = 0,\, z_{ij} \le 0\}.$$

So the full joint is

$$\begin{aligned} p(y, z, \alpha, \theta, \beta) = &\left[\prod_{i=1}^{N} \mathcal{N}(\theta_i \mid \mu_\theta, \sigma_\theta^2)\right] \left[\prod_{j=1}^{J} \mathcal{N}(\beta_j \mid \mu_\beta, \sigma_\beta^2)\right] \left[\prod_{j=1}^{J} \mathcal{N}(\alpha_j \mid \mu_\alpha, \sigma_\alpha^2)\right] \\ &\times \left[\prod_{i=1}^{N}\prod_{j=1}^{J} \mathcal{N}\big(z_{ij} \mid \alpha_j(\theta_i - \beta_j),\, 1\big)\right] \left[\prod_{i=1}^{N}\prod_{j=1}^{J} p(y_{ij} \mid z_{ij})\right]. \end{aligned}$$

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.

We start from the latent utility model, for each (i, j):

Latent utility:

$$z_{ij} \mid \theta_i, \beta_j, \alpha_j \sim \mathcal{N}(\mu_{ij}, 1),$$

where

$$\mu_{ij} := \alpha_j(\theta_i - \beta_j).$$

Observed vote:

$$y_{ij} = \nVdash(z_{ij} > 0).$$

Equivalently, the conditional distribution of y_{ij} given z_{ij} is

$$p(y_{ij} \mid z_{ij}) = \nVdash\{y_{ij} = 1,\, z_{ij} > 0\} + \nVdash\{y_{ij} = 0,\, z_{ij} \le 0\}.$$

Conditional distribution of z_{ij} given y_{ij}. We want $p(z_{ij} \mid y_{ij}, \theta, \beta, \alpha)$. By Bayes' rule (up to a proportionality constant):

$$p(z_{ij} \mid y_{ij}, \theta, \beta, \alpha) \;\propto\; p(y_{ij} \mid z_{ij})\, p(z_{ij} \mid \theta_i, \beta_j, \alpha_j).$$

But we know

$$p(z_{ij} \mid \theta_i, \beta_j, \alpha_j) = \mathcal{N}(z_{ij} \mid \mu_{ij}, 1).$$

Now consider two cases.

Case 1: $y_{ij} = 1$. If $y_{ij} = 1$, then by the definition of y_{ij},

$$p(y_{ij} = 1 \mid z_{ij}) = \nVdash\{z_{ij} > 0\}.$$

So the conditional density is

$$p(z_{ij} \mid y_{ij} = 1, \theta, \beta, \alpha) \;\propto\; \nVdash\{z_{ij} > 0\}\, \mathcal{N}(z_{ij} \mid \mu_{ij}, 1).$$

This is exactly a Normal $\mathcal{N}(\mu_{ij}, 1)$ density restricted to the domain $(0, \infty)$ and renormalized. Therefore:

$$z_{ij} \mid y_{ij} = 1, \theta, \beta, \alpha \sim \mathcal{N}(\mu_{ij}, 1) \text{ truncated to } (0, \infty).$$

Case 2: $y_{ij} = 0$. If $y_{ij} = 0$, then

$$p(y_{ij} = 0 \mid z_{ij}) = \nVdash\{z_{ij} \leq 0\}.$$

So

$$p(z_{ij} \mid y_{ij} = 0, \theta, \beta, \alpha) \ \propto \ \nVdash\{z_{ij} \leq 0\}\, \mathcal{N}(z_{ij} \mid \mu_{ij}, 1),$$

which is a Normal $\mathcal{N}(\mu_{ij}, 1)$ restricted to $(-\infty, 0]$. Hence:

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

Putting both cases together:

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

with $\mu_{ij} = \alpha_j(\theta_i - \beta_j)$.

Missing case: $y_{ij} = -1$. Recall the augmented model for a single (i, j):

$$\begin{gathered} z_{ij} \mid \theta_i, \beta_j, \alpha_j \sim \mathcal{N}(\mu_{ij}, 1), \qquad \mu_{ij} := \alpha_j(\theta_i - \beta_j), \\ y_{ij} = \nVdash(z_{ij} > 0), \quad y_{ij} \in \{0, 1\}. \end{gathered}$$

For observed votes $y_{ij} \in \{0, 1\}$, we showed

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

Now consider the missing case $y_{ij} = -1$. Define the index set of non-missing votes

$$O := \{(i, j) : y_{ij} \in \{0, 1\}\}.$$

Start from the full augmented joint over all pairs (i, j):

$$\begin{aligned} p\big(\{y_{ij}, z_{ij}\}_{i,j}, \theta, \beta, \alpha\big) = &\left[\prod_{(i,j) \in O} p(y_{ij} \mid z_{ij})\, p(z_{ij} \mid \theta_i, \beta_j, \alpha_j) \right] \\ &\times \left[\prod_{(i,j) \notin O} p(z_{ij} \mid \theta_i, \beta_j, \alpha_j) \right] p(\theta)\, p(\beta)\, p(\alpha). \end{aligned}$$

To obtain the joint distribution involving only observed votes and their latent utilities, integrate out the $\{z_{ij}\}_{(i,j)\notin O}$:

$$\begin{aligned} p\big(\{y_{ij}, z_{ij}\}_{(i,j)\in O}, \theta, \beta, \alpha\big) &= \int \Bigg[\prod_{(i,j)\in O} p(y_{ij} \mid z_{ij})\, p(z_{ij} \mid \theta_i, \beta_j, \alpha_j) \Bigg] \\ &\quad \times \Bigg[\prod_{(i,j)\notin O} p(z_{ij} \mid \theta_i, \beta_j, \alpha_j) \Bigg] p(\theta)\, p(\beta)\, p(\alpha) \prod_{(i,j)\notin O} dz_{ij} \\ &= \Bigg[\prod_{(i,j)\in O} p(y_{ij} \mid z_{ij})\, p(z_{ij} \mid \theta_i, \beta_j, \alpha_j) \Bigg] p(\theta)\, p(\beta)\, p(\alpha), \end{aligned}$$

because each factor $\int p(z_{ij} \mid \theta_i, \beta_j, \alpha_j)\, dz_{ij} = 1$ for $(i,j) \notin O$.

Thus the joint distribution over observed votes, their latent utilities, and the parameters is

$$p\big(\{y_{ij}, z_{ij}\}_{(i,j)\in O}, \theta, \beta, \alpha\big) = \Bigg[\prod_{(i,j)\in O} p(y_{ij} \mid z_{ij})\, p(z_{ij} \mid \theta_i, \beta_j, \alpha_j) \Bigg] p(\theta)\, p(\beta)\, p(\alpha).$$

Pairs (i,j) with $y_{ij} = -1$ are not in O, so they do not appear in this product and there is no likelihood term involving y_{ij}. Introducing a corresponding z_{ij} that appears only in its prior conditional $p(z_{ij} \mid \theta_i, \beta_j, \alpha_j)$ would integrate out to 1 and would not change the posterior. Therefore, when $y_{ij} = -1$ (missing), we simply omit the latent utility and 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.

Step 1: Write the conditional for a single senator i.

Prior (from 2.1):

$$\theta_i \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2), \quad \text{independently across } i.$$

Latent utilities (for all bills j with a non-missing vote from senator i):

$$z_{ij} \mid \theta_i, \beta_j, \alpha_j \sim \mathcal{N}(\mu_{ij}, 1), \qquad \mu_{ij} := \alpha_j(\theta_i - \beta_j).$$

Define the set of bills for which senator i has a non-missing vote:

$$O_i := \{j : (i,j) \in O\}.$$

Conditional on z, β, α, the full conditional for θ_i is

$$p(\theta_i \mid z, \beta, \alpha) \;\propto\; p(\theta_i) \prod_{j\in O_i} p(z_{ij} \mid \theta_i, \beta_j, \alpha_j).$$

Plug in the Gaussian forms:

Prior:

$$p(\theta_i) \propto \exp\left\{ -\frac{1}{2\sigma_\theta^2} (\theta_i - \mu_\theta)^2 \right\}.$$

Likelihood terms (for each $j \in O_i$):

$$p(z_{ij} \mid \theta_i, \beta_j, \alpha_j) \propto \exp\Big\{-\frac{1}{2}(z_{ij} - \alpha_j(\theta_i - \beta_j))^2\Big\}.$$

Thus

$$p(\theta_i \mid z, \beta, \alpha) \ \propto \ \exp\left\{-\frac{1}{2\sigma_\theta^2}(\theta_i - \mu_\theta)^2 - \frac{1}{2}\sum_{j \in O_i}(z_{ij} - \alpha_j(\theta_i - \beta_j))^2\right\}.$$

We now show this kernel is Gaussian in θ_i by expanding the exponent and completing the square.

Step 2: Expand the quadratic in θ_i.

First expand the prior term:

$$-\frac{1}{2\sigma_\theta^2}(\theta_i - \mu_\theta)^2 = -\frac{1}{2\sigma_\theta^2}(\theta_i^2 - 2\mu_\theta\theta_i + \mu_\theta^2).$$

Next expand each likelihood term. Rewrite the mean:

$$z_{ij} - \alpha_j(\theta_i - \beta_j) = z_{ij} - \alpha_j\theta_i + \alpha_j\beta_j = \alpha_j\theta_i - (z_{ij} + \alpha_j\beta_j)\,(-1),$$

so

$$(z_{ij} - \alpha_j(\theta_i - \beta_j))^2 = (\alpha_j\theta_i - (z_{ij} + \alpha_j\beta_j))^2 = \alpha_j^2\theta_i^2 - 2\alpha_j\theta_i(z_{ij} + \alpha_j\beta_j) + (z_{ij} + \alpha_j\beta_j)^2.$$

Sum over $j \in O_i$:

$$\sum_{j \in O_i}(z_{ij} - \alpha_j(\theta_i - \beta_j))^2 = \left(\sum_{j \in O_i}\alpha_j^2\right)\theta_i^2 - 2\theta_i\sum_{j \in O_i}\alpha_j(z_{ij} + \alpha_j\beta_j) + \sum_{j \in O_i}(z_{ij} + \alpha_j\beta_j)^2.$$

Now the exponent in $p(\theta_i \mid \cdot)$ is

$$\begin{aligned}\log p(\theta_i \mid z, \beta, \alpha) &= \text{const} - \frac{1}{2\sigma_\theta^2}(\theta_i^2 - 2\mu_\theta\theta_i + \mu_\theta^2) - \frac{1}{2}\sum_{j \in O_i}(z_{ij} - \alpha_j(\theta_i - \beta_j))^2 \\ &= \text{const} - \frac{1}{2}\left[\left(\frac{1}{\sigma_\theta^2} + \sum_{j \in O_i}\alpha_j^2\right)\theta_i^2 - 2\theta_i\left(\frac{\mu_\theta}{\sigma_\theta^2} + \sum_{j \in O_i}\alpha_j(z_{ij} + \alpha_j\beta_j)\right)\right],\end{aligned}$$

where "const" collects all terms independent of θ_i.

Define the posterior precision and linear term:

$$A_i := \frac{1}{\sigma_\theta^2} + \sum_{j \in O_i}\alpha_j^2, \qquad B_i := \frac{\mu_\theta}{\sigma_\theta^2} + \sum_{j \in O_i}\alpha_j(z_{ij} + \alpha_j\beta_j).$$

Then

$$\log p(\theta_i \mid z, \beta, \alpha) = \text{const} - \frac{1}{2}(A_i\theta_i^2 - 2B_i\theta_i).$$

Step 3: Complete the square and identify the Gaussian.

Write the quadratic in completed-square form:

$$A_i\theta_i^2 - 2B_i\theta_i = A_i\left(\theta_i^2 - 2\frac{B_i}{A_i}\theta_i\right) = A_i\left(\theta_i - \frac{B_i}{A_i}\right)^2 - A_i\left(\frac{B_i}{A_i}\right)^2.$$

So

$$\log p(\theta_i \mid z, \beta, \alpha) = \text{const}' - \frac{A_i}{2}\left(\theta_i - \frac{B_i}{A_i}\right)^2,$$

where const′ is another constant not depending on θ_i.

This is exactly the kernel of a Normal distribution with

variance

$$\text{Var}(\theta_i \mid z, \beta, \alpha) = A_i^{-1} = \left(\frac{1}{\sigma_\theta^2} + \sum_{j \in O_i} \alpha_j^2\right)^{-1},$$

mean

$$\mathbb{E}[\theta_i \mid z, \beta, \alpha] = \frac{B_i}{A_i} = \left(\frac{1}{\sigma_\theta^2} + \sum_{j \in O_i} \alpha_j^2\right)^{-1}\left(\frac{\mu_\theta}{\sigma_\theta^2} + \sum_{j \in O_i} \alpha_j(z_{ij} + \alpha_j\beta_j)\right).$$

For each senator i, the conditional distribution of θ_i given z, β, α is Gaussian:

$$\theta_i \mid z, \beta, \alpha \sim \mathcal{N}(m_{\theta_i}, v_{\theta_i}),$$

with

$$v_{\theta_i} = \left(\frac{1}{\sigma_\theta^2} + \sum_{j \in O_i} \alpha_j^2\right)^{-1},$$

$$m_{\theta_i} = v_{\theta_i}\left(\frac{\mu_\theta}{\sigma_\theta^2} + \sum_{j \in O_i} \alpha_j(z_{ij} + \alpha_j\beta_j)\right),$$

where $O_i = \{j : (i, j) \in O\}$ indexes the bills on which senator i cast a non-missing vote.

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.

We proceed exactly as for the θ_i case, but now treating β_j as the unknown and (θ, α, z) as given.

Step 1: Prior and conditional likelihood for a single bill j.

From 2.1, the prior on the bill location is

$$\beta_j \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2),$$

independently across j.

For all senators i with a non-missing vote on bill j, the latent utilities satisfy

$$z_{ij} \mid \theta_i, \beta_j, \alpha_j \sim \mathcal{N}(\mu_{ij}, 1), \qquad \mu_{ij} := \alpha_j(\theta_i - \beta_j).$$

Let

$$O_j := \{i : (i,j) \in O\}$$

be the set of senators who (non-missingly) voted on bill j.

Given z, θ, α, the full conditional for β_j is

$$p(\beta_j \mid z, \theta, \alpha) \;\propto\; p(\beta_j) \prod_{i \in O_j} p(z_{ij} \mid \theta_i, \beta_j, \alpha_j).$$

Plug in the Gaussian pieces:

Prior:

$$p(\beta_j) \propto \exp\left\{ -\frac{1}{2\sigma_\beta^2}(\beta_j - \mu_\beta)^2 \right\}.$$

For each $i \in O_j$,

$$p(z_{ij} \mid \theta_i, \beta_j, \alpha_j) \propto \exp\left\{ -\frac{1}{2}(z_{ij} - \alpha_j(\theta_i - \beta_j))^2 \right\}.$$

Thus

$$p(\beta_j \mid z, \theta, \alpha) \;\propto\; \exp\left\{ -\frac{1}{2\sigma_\beta^2}(\beta_j - \mu_\beta)^2 - \frac{1}{2}\sum_{i \in O_j}(z_{ij} - \alpha_j(\theta_i - \beta_j))^2 \right\}.$$

We now expand the exponent as a quadratic function of β_j.

Step 2: Expand the quadratic in β_j.

First expand the prior term:

$$-\frac{1}{2\sigma_\beta^2}(\beta_j - \mu_\beta)^2 = -\frac{1}{2\sigma_\beta^2}(\beta_j^2 - 2\mu_\beta\beta_j + \mu_\beta^2).$$

For the likelihood terms, write

$$z_{ij} - \alpha_j(\theta_i - \beta_j) = z_{ij} - \alpha_j\theta_i + \alpha_j\beta_j = \alpha_j\beta_j + c_{ij}, \quad c_{ij} := z_{ij} - \alpha_j\theta_i.$$

Then

$$(z_{ij} - \alpha_j(\theta_i - \beta_j))^2 = (\alpha_j\beta_j + c_{ij})^2 = \alpha_j^2\beta_j^2 + 2\alpha_j\beta_j c_{ij} + c_{ij}^2.$$

Summing over $i \in O_j$:

$$\sum_{i \in O_j}(z_{ij} - \alpha_j(\theta_i - \beta_j))^2 = \left(\sum_{i \in O_j}\alpha_j^2\right)\beta_j^2 + 2\beta_j\sum_{i \in O_j}\alpha_j c_{ij} + \sum_{i \in O_j} c_{ij}^2.$$

Because α_j does not depend on i,

$$\sum_{i \in O_j}\alpha_j^2 = |O_j|\,\alpha_j^2.$$

Now collect the terms that depend on β_j in the exponent of $p(\beta_j \mid \cdot)$:

$$\begin{aligned}\log p(\beta_j \mid z, \theta, \alpha) &= \text{const} - \frac{1}{2\sigma_\beta^2}(\beta_j^2 - 2\mu_\beta \beta_j + \mu_\beta^2) - \frac{1}{2}\sum_{i \in O_j}(\alpha_j^2 \beta_j^2 + 2\alpha_j \beta_j c_{ij} + c_{ij}^2) \\ &= \text{const} - \frac{1}{2}\left[\left(\frac{1}{\sigma_\beta^2} + |O_j|\alpha_j^2\right)\beta_j^2 - 2\beta_j\left(\frac{\mu_\beta}{\sigma_\beta^2} - \alpha_j \sum_{i \in O_j} c_{ij}\right)\right],\end{aligned}$$

where "const" collects all terms that do not depend on β_j.

Define

$$A_j := \frac{1}{\sigma_\beta^2} + |O_j|\alpha_j^2, \qquad B_j := \frac{\mu_\beta}{\sigma_\beta^2} - \alpha_j \sum_{i \in O_j} c_{ij} = \frac{\mu_\beta}{\sigma_\beta^2} - \alpha_j \sum_{i \in O_j}(z_{ij} - \alpha_j \theta_i).$$

Then

$$\log p(\beta_j \mid z, \theta, \alpha) = \text{const} - \frac{1}{2}(A_j \beta_j^2 - 2B_j \beta_j).$$

Step 3: Complete the square and identify the Gaussian.

Complete the square in β_j:

$$A_j \beta_j^2 - 2B_j \beta_j = A_j\left(\beta_j - \frac{B_j}{A_j}\right)^2 - A_j\left(\frac{B_j}{A_j}\right)^2.$$

So

$$\log p(\beta_j \mid z, \theta, \alpha) = \text{const}' - \frac{A_j}{2}\left(\beta_j - \frac{B_j}{A_j}\right)^2,$$

which is the kernel of a Normal distribution with

variance

$$\operatorname{Var}(\beta_j \mid z, \theta, \alpha) = A_j^{-1} = \left(\frac{1}{\sigma_\beta^2} + |O_j|\alpha_j^2\right)^{-1},$$

mean

$$\mathbb{E}[\beta_j \mid z, \theta, \alpha] = \frac{B_j}{A_j} = \left(\frac{1}{\sigma_\beta^2} + |O_j|\alpha_j^2\right)^{-1}\left(\frac{\mu_\beta}{\sigma_\beta^2} - \alpha_j \sum_{i \in O_j}(z_{ij} - \alpha_j \theta_i)\right).$$

Final form. For each bill j, the conditional distribution of β_j given z, θ, α is Gaussian:

$$\beta_j \mid z, \theta, \alpha \sim \mathcal{N}(m_{\beta_j}, v_{\beta_j}),$$

with

$$v_{\beta_j} = \left(\frac{1}{\sigma_\beta^2} + |O_j|\alpha_j^2\right)^{-1},$$

$$m_{\beta_j} = v_{\beta_j}\left(\frac{\mu_\beta}{\sigma_\beta^2} - \alpha_j \sum_{i \in O_j}(z_{ij} - \alpha_j \theta_i)\right),$$

where $O_j = \{i : (i, j) \in O\}$ is the set of senators who cast a non-missing vote on bill j.

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).

We proceed just like for θ_i and β_j, now treating α_j as the unknown with (z, θ, β) given. Suppose the prior is

$$\alpha_j \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2).$$

Step 1: Write the joint in exponential form.

For a given bill j, the complete conditional is proportional to the prior on α_j times the likelihood of the observed latent utilities z_{ij} for all senators $i \in O_j$ who voted on bill j:

$$p(\alpha_j \mid z, \theta, \beta) \propto p(\alpha_j) \prod_{i \in O_j} p(z_{ij} \mid \theta_i, \beta_j, \alpha_j).$$

The latent utility model is

$$z_{ij} \mid \theta_i, \beta_j, \alpha_j \sim \mathcal{N}(\alpha_j(\theta_i - \beta_j), 1).$$

Let

$$x_{ij} := \theta_i - \beta_j,$$

so

$$z_{ij} \mid \alpha_j \sim \mathcal{N}(\alpha_j x_{ij}, 1).$$

Then

$$\begin{aligned}
\log p(\alpha_j \mid z, \theta, \beta) &= \text{const} - \frac{1}{2\sigma_\alpha^2}(\alpha_j - \mu_\alpha)^2 - \frac{1}{2} \sum_{i \in O_j} (z_{ij} - \alpha_j x_{ij})^2 \\
&= \text{const} - \frac{1}{2\sigma_\alpha^2}(\alpha_j^2 - 2\mu_\alpha \alpha_j) - \frac{1}{2} \sum_{i \in O_j} (z_{ij}^2 - 2 z_{ij} \alpha_j x_{ij} + \alpha_j^2 x_{ij}^2).
\end{aligned}$$

Step 2: Collect quadratic terms in α_j.

The coefficient of α_j^2 in the exponent is

$$-\frac{1}{2} \left[\frac{1}{\sigma_\alpha^2} + \sum_{i \in O_j} x_{ij}^2 \right].$$

The coefficient of α_j is

$$\frac{\mu_\alpha}{\sigma_\alpha^2} + \sum_{i \in O_j} z_{ij} x_{ij}.$$

Define

$$A_j := \frac{1}{\sigma_\alpha^2} + \sum_{i \in O_j} (\theta_i - \beta_j)^2, \qquad B_j := \frac{\mu_\alpha}{\sigma_\alpha^2} + \sum_{i \in O_j} z_{ij}(\theta_i - \beta_j).$$

Then

$$\log p(\alpha_j \mid z, \theta, \beta) = \text{const} - \frac{1}{2}(A_j \alpha_j^2 - 2 B_j \alpha_j).$$

Step 3: Complete the square.

Complete the square in α_j:

$$A_j\alpha_j^2 - 2B_j\alpha_j = A_j\left(\alpha_j - \frac{B_j}{A_j}\right)^2 - A_j\left(\frac{B_j}{A_j}\right)^2.$$

So

$$\log p(\alpha_j \mid z, \theta, \beta) = \text{const}' - \frac{A_j}{2}\left(\alpha_j - \frac{B_j}{A_j}\right)^2,$$

which is the kernel of a Normal distribution.

Final form (Normal prior). For each bill j, the conditional distribution is Gaussian:

$$\alpha_j \mid z, \theta, \beta \sim \mathcal{N}(m_{\alpha_j}, v_{\alpha_j}),$$

with

$$v_{\alpha_j} = \left(\frac{1}{\sigma_\alpha^2} + \sum_{i\in O_j}(\theta_i - \beta_j)^2\right)^{-1},$$

$$m_{\alpha_j} = v_{\alpha_j}\left(\frac{\mu_\alpha}{\sigma_\alpha^2} + \sum_{i\in O_j} z_{ij}(\theta_i - \beta_j)\right),$$

where $O_j = \{i : (i,j) \in O\}$ is the set of senators who cast a non-missing vote on bill j.

Truncated Normal prior. If instead the prior is

$$\alpha_j \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2) \text{ truncated to } (0, \infty),$$

the conditional is the same Normal $\mathcal{N}(m_{\alpha_j}, v_{\alpha_j})$ truncated to $(0, \infty)$.

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

Step 1: Start from the joint with z.

Restricting to non-missing entries $(i,j) \in O$,

$$p(y, z, \alpha, \theta, \beta) = \left[\prod_{(i,j)\in O} p(y_{ij} \mid z_{ij})\, p(z_{ij} \mid \theta_i, \beta_j, \alpha_j)\right] p(\theta)\, p(\beta)\, p(\alpha),$$

where

$$z_{ij} \mid \theta_i, \beta_j, \alpha_j \sim \mathcal{N}\big(\mu_{ij}, 1\big), \quad \mu_{ij} := \alpha_j(\theta_i - \beta_j),$$

and

$$y_{ij} = \mathbf{1}(z_{ij} > 0), \quad y_{ij} \in \{0, 1\}.$$

The conditional for y_{ij} given z_{ij} is

$$p(y_{ij} \mid z_{ij}) = \begin{cases} 1, & y_{ij} = 1,\ z_{ij} > 0, \\ 1, & y_{ij} = 0,\ z_{ij} \le 0, \\ 0, & \text{otherwise.} \end{cases}$$

Equivalently,

$$p(y_{ij} \mid z_{ij}) = [\mathbf{1}(z_{ij} > 0)]^{y_{ij}} [\mathbf{1}(z_{ij} \le 0)]^{1-y_{ij}}.$$

Step 2: Marginalize out z.

We want

$$p(y, \alpha, \theta, \beta) = \int p(y, z, \alpha, \theta, \beta)\, dz.$$

Plug in the factorized joint and separate the integral:

$$\begin{aligned} p(y, \alpha, \theta, \beta) &= \int \left[\prod_{(i,j)\in O} p(y_{ij} \mid z_{ij})\, p(z_{ij} \mid \theta_i, \beta_j, \alpha_j) \right] p(\theta)\, p(\beta)\, p(\alpha)\, dz \\ &= p(\theta)\, p(\beta)\, p(\alpha) \prod_{(i,j)\in O} \int p(y_{ij} \mid z_{ij})\, p(z_{ij} \mid \theta_i, \beta_j, \alpha_j)\, dz_{ij}. \end{aligned}$$

For each fixed (i, j),

$$\int p(y_{ij} \mid z_{ij})\, p(z_{ij} \mid \theta_i, \beta_j, \alpha_j)\, dz_{ij}.$$

Step 3: Evaluate the inner integral.

Recall

$$z_{ij} \mid \theta_i, \beta_j, \alpha_j \sim \mathcal{N}(\mu_{ij}, 1), \quad \mu_{ij} = \alpha_j(\theta_i - \beta_j).$$

So the density is

$$p(z_{ij} \mid \theta_i, \beta_j, \alpha_j) = \phi(z_{ij} - \mu_{ij}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z_{ij} - \mu_{ij})^2}{2} \right).$$

We treat the two cases $y_{ij} = 1$ and $y_{ij} = 0$ separately.

Case A: $y_{ij} = 1$.

Then

$$p(y_{ij} = 1 \mid \theta_i, \alpha_j, \beta_j) = \int p(y_{ij} = 1 \mid z_{ij})\, p(z_{ij} \mid \theta_i, \beta_j, \alpha_j)\, dz_{ij}.$$

But $p(y_{ij} = 1 \mid z_{ij}) = \mathbf{1}(z_{ij} > 0)$, so

$$\begin{aligned} p(y_{ij} = 1 \mid \theta_i, \alpha_j, \beta_j) &= \int_0^\infty \phi(z_{ij} - \mu_{ij})\, dz_{ij} \\ &= \Pr(z_{ij} > 0 \mid \theta_i, \alpha_j, \beta_j). \end{aligned}$$

Standardize: let $U = z_{ij} - \mu_{ij} \sim \mathcal{N}(0, 1)$. Then

$$\Pr(z_{ij} > 0) = \Pr(U > -\mu_{ij}) = 1 - \Phi(-\mu_{ij}) = \Phi(\mu_{ij}),$$

using symmetry of the standard normal CDF. Hence

$$p(y_{ij} = 1 \mid \theta_i, \alpha_j, \beta_j) = \Phi(\mu_{ij}) = \Phi(\alpha_j(\theta_i - \beta_j)),$$

which matches what we had in 1.1.

Case B: $y_{ij} = 0$.

Similarly,

$$p(y_{ij} = 0 \mid \theta_i, \alpha_j, \beta_j) = \int p(y_{ij} = 0 \mid z_{ij})\, p(z_{ij} \mid \theta_i, \beta_j, \alpha_j)\, dz_{ij}.$$

Now $p(y_{ij} = 0 \mid z_{ij}) = \mathbf{1}(z_{ij} \leq 0)$, so

$$\begin{aligned} p(y_{ij} = 0 \mid \theta_i, \alpha_j, \beta_j) &= \int_{-\infty}^{0} \phi(z_{ij} - \mu_{ij})\, dz_{ij} \\ &= \Pr(z_{ij} \leq 0 \mid \theta_i, \alpha_j, \beta_j). \end{aligned}$$

Again standardize $U = z_{ij} - \mu_{ij}$:

$$\Pr(z_{ij} \leq 0) = \Pr(U \leq -\mu_{ij}) = \Phi(-\mu_{ij}) = 1 - \Phi(\mu_{ij}),$$

so

$$p(y_{ij} = 0 \mid \theta_i, \alpha_j, \beta_j) = 1 - \Phi(\alpha_j(\theta_i - \beta_j)).$$

We can combine both cases into the Bernoulli-pmf form

$$p(y_{ij} \mid \theta_i, \alpha_j, \beta_j) = [\Phi(\alpha_j(\theta_i - \beta_j))]^{y_{ij}}\, [1 - \Phi(\alpha_j(\theta_i - \beta_j))]^{1 - y_{ij}}.$$

Thus we have explicitly shown that

$$\int p(y_{ij} \mid z_{ij})\, p(z_{ij} \mid \theta_i, \beta_j, \alpha_j)\, dz_{ij} = p(y_{ij} \mid \theta_i, \alpha_j, \beta_j).$$

Step 4: Plug back into the marginal.

Return to

$$p(y, \alpha, \theta, \beta) = p(\theta)\, p(\beta)\, p(\alpha) \prod_{(i,j) \in O} \int p(y_{ij} \mid z_{ij})\, p(z_{ij} \mid \theta_i, \beta_j, \alpha_j)\, dz_{ij}.$$

By the calculation above, each integral is $p(y_{ij} \mid \theta_i, \alpha_j, \beta_j)$, so

$$p(y, \alpha, \theta, \beta) = p(\theta)\, p(\beta)\, p(\alpha) \prod_{(i,j) \in O} p(y_{ij} \mid \theta_i, \alpha_j, \beta_j),$$

i.e.

$$p(y, \alpha, \theta, \beta) = p(\theta)\, p(\beta)\, p(\alpha) \prod_{(i,j) \in O} [\Phi(\alpha_j(\theta_i - \beta_j))]^{y_{ij}} [1 - \Phi(\alpha_j(\theta_i - \beta_j))]^{1 - y_{ij}}.$$

Step 5: Make $p(y \mid \alpha, \theta, \beta)$ explicit.

By the definition of conditional density,

$$p(y \mid \alpha, \theta, \beta) = \frac{p(y, \alpha, \theta, \beta)}{p(\alpha, \theta, \beta)} = \frac{p(y, \alpha, \theta, \beta)}{p(\theta)\, p(\beta)\, p(\alpha)}.$$

Using the expression we just derived for $p(y, \alpha, \theta, \beta)$,

$$p(y \mid \alpha, \theta, \beta) = \prod_{(i,j) \in O} p(y_{ij} \mid \theta_i, \alpha_j, \beta_j) = \prod_{(i,j) \in O} [\Phi(\alpha_j(\theta_i - \beta_j))]^{y_{ij}} [1 - \Phi(\alpha_j(\theta_i - \beta_j))]^{1 - y_{ij}}.$$

Then, equivalently,

$$p(y, \alpha, \theta, \beta) = p(\theta)\, p(\beta)\, p(\alpha)\, p(y \mid \alpha, \theta, \beta).$$

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

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

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.

Assume the mean-field factorization

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

with each factor in a parametric family as follows.

Senator positions θ_i**.** For each $i = 1, \dots, n$,

$$q(\theta_i) = \mathcal{N}\big(m_{\theta_i},\, s^2_{\theta_i}\big),$$

with free variational parameters $m_{\theta_i} \in \mathbb{R}$ and $s^2_{\theta_i} > 0$.

Bill locations β_j**.** For each $j = 1, \dots, d$,

$$q(\beta_j) = \mathcal{N}\big(m_{\beta_j},\, s^2_{\beta_j}\big),$$

with variational parameters $m_{\beta_j} \in \mathbb{R}$, $s^2_{\beta_j} > 0$.

Bill discriminations α_j**.** For each $j = 1, \dots, d$,

- if the prior on α_j is Normal,
$$q(\alpha_j) = \mathcal{N}\big(m_{\alpha_j},\, s^2_{\alpha_j}\big),$$
- if the prior on α_j is truncated Normal (e.g. $\alpha_j > 0$),
$$q(\alpha_j) = \mathcal{N}\big(m_{\alpha_j},\, s^2_{\alpha_j}\big) \text{ truncated to } (0, \infty),$$

with variational parameters $m_{\alpha_j} \in \mathbb{R}$, $s^2_{\alpha_j} > 0$ in either case.

Latent utilities z_{ij}**.** For each observed vote $(i, j) \in O$ with $y_{ij} \in \{0, 1\}$,

$$q(z_{ij}) = \begin{cases} \mathcal{N}\big(m_{z_{ij}},\, s^2_{z_{ij}}\big) \text{ truncated to } (0, \infty), & y_{ij} = 1, \\ \mathcal{N}\big(m_{z_{ij}},\, s^2_{z_{ij}}\big) \text{ truncated to } (-\infty, 0], & y_{ij} = 0, \end{cases}$$

with variational parameters $m_{z_{ij}} \in \mathbb{R}$, $s^2_{z_{ij}} > 0$.

For missing votes $y_{ij} = -1$, no $q(z_{ij})$ factor is introduced (those pairs are excluded from the product).

Putting this together, the full family is

$$q(\theta, \beta, \alpha, z) = \left[\prod_{i=1}^{n} \mathcal{N}\big(\theta_i \mid m_{\theta_i}, s^2_{\theta_i}\big)\right] \left[\prod_{j=1}^{d} \mathcal{N}\big(\beta_j \mid m_{\beta_j}, s^2_{\beta_j}\big)\, q(\alpha_j)\right]$$
$$\times \prod_{(i,j) \in O} q(z_{ij}),$$

with $q(\alpha_j)$ and $q(z_{ij})$ as specified above.

Moments needed for CAVI updates.

From these factors, the coordinate updates will require the following expectations:

- **For each senator** i:
  - $\mathbb{E}_q[\theta_i] = m_{\theta_i}$
  - $\mathbb{E}_q[\theta_i^2] = s_{\theta_i}^2 + m_{\theta_i}^2$
- **For each bill location** j:
  - $\mathbb{E}_q[\beta_j] = m_{\beta_j}$
  - $\mathbb{E}_q[\beta_j^2] = s_{\beta_j}^2 + m_{\beta_j}^2$
- **For each bill discrimination** j:
  - $\mathbb{E}_q[\alpha_j]$
  - $\mathbb{E}_q[\alpha_j^2]$

  (If $q(\alpha_j)$ is Normal, these are m_{α_j} and $s_{\alpha_j}^2 + m_{\alpha_j}^2$; if truncated Normal, they are given by the standard truncated Normal moment formulas.)
- **For each latent utility** z_{ij} **with** $(i,j) \in O$:
  - $\mathbb{E}_q[z_{ij}]$
  - $\mathbb{E}_q[z_{ij}^2]$

  (Again, computed using the truncated Normal formulas for the appropriate truncation region.)

Because of the mean-field factorization,

$$q(\theta, \beta, \alpha, z) = q(\theta)\, q(\beta)\, q(\alpha)\, q(z),$$

all mixed expectations factor into products of first moments, for example:

- $\mathbb{E}_q[\alpha_j\, \theta_i] = \mathbb{E}_q[\alpha_j]\, \mathbb{E}_q[\theta_i]$,
- $\mathbb{E}_q[\alpha_j\, \beta_j] = \mathbb{E}_q[\alpha_j]\, \mathbb{E}_q[\beta_j]$,
- $\mathbb{E}_q[\alpha_j\, (\theta_i - \beta_j)] = \mathbb{E}_q[\alpha_j]\, (\mathbb{E}_q[\theta_i] - \mathbb{E}_q[\beta_j])$,

so all coordinate updates can be written in terms of the one- and two-moment quantities listed above.

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.

Throughout, let

- $O = \{(i,j) : y_{ij} \in \{0,1\}\}$ be the set of observed votes,
- $O_i = \{j : (i,j) \in O\}$ votes observed for senator i,
- $O_j = \{i : (i,j) \in O\}$ votes observed on bill j.

We also keep the priors

$$\theta_i \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2), \quad \beta_j \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2), \quad \alpha_j \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2).$$

The augmented model has, for $(i,j) \in O$,

$$z_{ij} \mid \theta_i, \beta_j, \alpha_j \sim \mathcal{N}(\mu_{ij}, 1), \quad \mu_{ij} := \alpha_j(\theta_i - \beta_j),$$

$$y_{ij} = \mathbf{1}(z_{ij} > 0), \quad y_{ij} \in \{0,1\}.$$

The variational family factorizes as

$$q(\theta, \beta, \alpha, z) = \left(\prod_i q(\theta_i)\right)\left(\prod_j q(\beta_j)\, q(\alpha_j)\right)\left(\prod_{(i,j)\in O} q(z_{ij})\right).$$

We use the standard CAVI identity

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

(i) Update for $q(z_{ij})$.

We isolate terms of the joint that involve a given z_{ij}:

$$\log p(y, z, \theta, \beta, \alpha) = \log p(y_{ij} \mid z_{ij}) + \log p(z_{ij} \mid \theta_i, \beta_j, \alpha_j) + (\text{terms not involving } z_{ij}).$$

So

$$\log q^*(z_{ij}) \propto \mathbb{E}_{-z_{ij}}[\log p(y_{ij} \mid z_{ij})] + \mathbb{E}_{-z_{ij}}[\log p(z_{ij} \mid \theta_i, \beta_j, \alpha_j)].$$

The likelihood term enforces the truncation:

- If $y_{ij} = 1$, support is $z_{ij} > 0$,
- If $y_{ij} = 0$, support is $z_{ij} \le 0$.

Within that support, $\log p(y_{ij} \mid z_{ij})$ is constant in z_{ij}.

The Gaussian prior conditional:

$$p(z_{ij} \mid \theta_i, \beta_j, \alpha_j) = \mathcal{N}(z_{ij} \mid \mu_{ij}, 1), \quad \mu_{ij} = \alpha_j(\theta_i - \beta_j).$$

Ignoring constants in z_{ij}, we have inside the support

$$\log q^*(z_{ij}) \propto \mathbb{E}_{-z_{ij}}\left[-\frac{1}{2}(z_{ij} - \mu_{ij})^2\right] = -\frac{1}{2}\left(z_{ij}^2 - 2 z_{ij}\, \mathbb{E}_q[\mu_{ij}]\right) + \text{const}.$$

Thus, up to truncation, the optimal factor is a Normal with

$$\bar{\mu}_{ij} := \mathbb{E}_q[\mu_{ij}] = \mathbb{E}_q[\alpha_j(\theta_i - \beta_j)].$$

Under the mean-field factorization,

$$\mathbb{E}_q[\alpha_j(\theta_i - \beta_j)] = \mathbb{E}_q[\alpha_j]\,(\mathbb{E}_q[\theta_i] - \mathbb{E}_q[\beta_j]).$$

So

$$q(z_{ij}) \propto \mathcal{N}(z_{ij} \mid \bar{\mu}_{ij}, 1) \times \begin{cases} \mathbf{1}(z_{ij} > 0), & y_{ij} = 1, \\ \mathbf{1}(z_{ij} \leq 0), & y_{ij} = 0. \end{cases}$$

That is,

- If $y_{ij} = 1$: $q(z_{ij})$ is $\mathcal{N}(\bar{\mu}_{ij}, 1)$ truncated to $(0, \infty)$.
- If $y_{ij} = 0$: $q(z_{ij})$ is $\mathcal{N}(\bar{\mu}_{ij}, 1)$ truncated to $(-\infty, 0]$.

Moment $\mathbb{E}_q[z_{ij}]$.

Let $\phi(\cdot)$ and $\Phi(\cdot)$ be the standard Normal pdf and cdf.

- If $y_{ij} = 1$ (truncate to $(0, \infty)$):

$$\mathbb{E}_q[z_{ij}] = \bar{\mu}_{ij} + \frac{\phi(\bar{\mu}_{ij})}{\Phi(\bar{\mu}_{ij})}.$$

- If $y_{ij} = 0$ (truncate to $(-\infty, 0]$):

$$\mathbb{E}_q[z_{ij}] = \bar{\mu}_{ij} - \frac{\phi(\bar{\mu}_{ij})}{1 - \Phi(\bar{\mu}_{ij})}.$$

These are the standard first moments for a Normal$(\bar{\mu}_{ij}, 1)$ truncated above or below at 0, respectively.

(ii) Updates for $q(\theta_i)$ and $q(\beta_j)$.

$q(\theta_i)$.

We collect all terms involving θ_i: the prior and the likelihood pieces for z_{ij} with $j \in O_i$.

$$\log p(\theta_i \mid \cdot) \propto \log p(\theta_i) + \sum_{j \in O_i} \log p(z_{ij} \mid \theta_i, \beta_j, \alpha_j).$$

Using

$$p(\theta_i) = \mathcal{N}(\theta_i \mid \mu_\theta, \sigma_\theta^2), \quad p(z_{ij} \mid \theta_i, \beta_j, \alpha_j) = \mathcal{N}(z_{ij} \mid \alpha_j(\theta_i - \beta_j), 1),$$

we can write

$$\log q^*(\theta_i) \propto -\frac{1}{2\sigma_\theta^2}(\theta_i - \mu_\theta)^2 - \frac{1}{2}\sum_{j \in O_i}(z_{ij} - \alpha_j(\theta_i - \beta_j))^2.$$

Expand the quadratic in θ_i, then take expectation over $\alpha_j, \beta_j, z_{ij}$ (all independent of θ_i under q). The resulting form is

$$\log q^*(\theta_i) = -\frac{1}{2}\,\lambda_{\theta_i}\,\theta_i^2 + \eta_{\theta_i}\,\theta_i + \text{const},$$

with

$$\lambda_{\theta_i} = \frac{1}{\sigma_\theta^2} + \sum_{j \in O_i} \mathbb{E}_q[\alpha_j^2],$$

$$\eta_{\theta_i} = \frac{\mu_\theta}{\sigma_\theta^2} + \sum_{j \in O_i} (\mathbb{E}_q[\alpha_j]\,\mathbb{E}_q[z_{ij}] + \mathbb{E}_q[\alpha_j^2]\,\mathbb{E}_q[\beta_j]).$$

Thus $q(\theta_i)$ is Normal,

$$q(\theta_i) = \mathcal{N}(m_{\theta_i}, v_{\theta_i}),$$

with

$$v_{\theta_i} = \lambda_{\theta_i}^{-1}, \quad m_{\theta_i} = v_{\theta_i}\,\eta_{\theta_i}.$$

So the update for $q(\theta_i)$ uses the moments $\mathbb{E}_q[\alpha_j]$, $\mathbb{E}_q[\alpha_j^2]$, $\mathbb{E}_q[z_{ij}]$, and $\mathbb{E}_q[\beta_j]$.

$q(\beta_j)$.

Similarly, for each bill j, collect terms involving β_j: its prior and the likelihood pieces for z_{ij} with $i \in O_j$:

$$\log p(\beta_j \mid \cdot) \propto \log p(\beta_j) + \sum_{i \in O_j} \log p(z_{ij} \mid \theta_i, \beta_j, \alpha_j).$$

With

$$p(\beta_j) = \mathcal{N}(\beta_j \mid \mu_\beta, \sigma_\beta^2),$$

and the same Normal likelihood, expanding in β_j and taking expectations gives

$$\log q^*(\beta_j) = -\frac{1}{2}\,\lambda_{\beta_j}\,\beta_j^2 + \eta_{\beta_j}\,\beta_j + \text{const},$$

where

$$\lambda_{\beta_j} = \frac{1}{\sigma_\beta^2} + \sum_{i \in O_j} \mathbb{E}_q[\alpha_j^2],$$

$$\eta_{\beta_j} = \frac{\mu_\beta}{\sigma_\beta^2} + \sum_{i \in O_j} (\mathbb{E}_q[\alpha_j^2]\,\mathbb{E}_q[\theta_i] - \mathbb{E}_q[\alpha_j]\,\mathbb{E}_q[z_{ij}]).$$

Thus

$$q(\beta_j) = \mathcal{N}(m_{\beta_j}, v_{\beta_j}),$$

with

$$v_{\beta_j} = \lambda_{\beta_j}^{-1}, \quad m_{\beta_j} = v_{\beta_j}\,\eta_{\beta_j}.$$

Again, the update depends on $\mathbb{E}_q[\alpha_j]$, $\mathbb{E}_q[\alpha_j^2]$, $\mathbb{E}_q[z_{ij}]$, $\mathbb{E}_q[\theta_i]$.

(iii) Update for $q(\alpha_j)$.

For a fixed bill j, the terms involving α_j are its prior and the likelihood contributions $p(z_{ij} \mid \theta_i, \beta_j, \alpha_j)$ for $i \in O_j$:

$$\log p(\alpha_j \mid \cdot) \propto \log p(\alpha_j) + \sum_{i \in O_j} \log p(z_{ij} \mid \theta_i, \beta_j, \alpha_j).$$

Using the Normal prior

$$p(\alpha_j) = \mathcal{N}(\alpha_j \mid \mu_\alpha, \sigma_\alpha^2),$$

and writing the likelihood as a linear regression:

$$z_{ij} = \alpha_j x_{ij} + \varepsilon_{ij}, \quad x_{ij} := \theta_i - \beta_j, \quad \varepsilon_{ij} \sim \mathcal{N}(0, 1),$$

the log-likelihood term is

$$\sum_{i \in O_j} \log \mathcal{N}(z_{ij} \mid \alpha_j x_{ij}, 1) \propto -\frac{1}{2} \sum_{i \in O_j} (z_{ij} - \alpha_j x_{ij})^2.$$

Expanding in α_j and taking expectation over $z_{ij}, \theta_i, \beta_j$ (but not α_j) gives

$$\log q^*(\alpha_j) = -\frac{1}{2}\, \lambda_{\alpha_j}\, \alpha_j^2 + \eta_{\alpha_j}\, \alpha_j + \text{const},$$

where

$$\lambda_{\alpha_j} = \frac{1}{\sigma_\alpha^2} + \sum_{i \in O_j} \mathbb{E}_q[x_{ij}^2], \quad x_{ij} = \theta_i - \beta_j,$$

$$\eta_{\alpha_j} = \frac{\mu_\alpha}{\sigma_\alpha^2} + \sum_{i \in O_j} \mathbb{E}_q[z_{ij} x_{ij}].$$

Under the mean-field factorization, z_{ij}, θ_i, and β_j are independent under q, so

$$\mathbb{E}_q[x_{ij}] = \mathbb{E}_q[\theta_i] - \mathbb{E}_q[\beta_j],$$

$$\mathbb{E}_q[x_{ij}^2] = \mathbb{E}_q[\theta_i^2] - 2\, \mathbb{E}_q[\theta_i]\, \mathbb{E}_q[\beta_j] + \mathbb{E}_q[\beta_j^2],$$

$$\mathbb{E}_q[z_{ij} x_{ij}] = \mathbb{E}_q[z_{ij}]\, (\mathbb{E}_q[\theta_i] - \mathbb{E}_q[\beta_j]).$$

So $q(\alpha_j)$ is Gaussian in the unconstrained case:

$$q(\alpha_j) = \mathcal{N}(m_{\alpha_j}, v_{\alpha_j}),$$

with

$$v_{\alpha_j} = \lambda_{\alpha_j}^{-1}, \quad m_{\alpha_j} = v_{\alpha_j}\, \eta_{\alpha_j}.$$

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 $\Omega \subseteq \{1,\dots,n\} \times \{1,\dots,m\}$. Fix latent dimension K. For users i and items j we have factors $\theta_i, \beta_j \in \mathbb{R}^K$.

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

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.

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

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

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

We approximate the posterior with a factorized family

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

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

We use the mean-field family

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

The joint model:

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

So the ELBO is

$$\mathcal{L}(q) = \mathbb{E}_q[\log p(X, \Theta, B)] - \mathbb{E}_q[\log q(\Theta, B)].$$

Break it into pieces.

Likelihood term.

$$\log p(X \mid \Theta, B) = \sum_{(i,j)\in\Omega} \log \mathcal{N}(x_{ij} \mid \theta_i^\top \beta_j, \sigma^2).$$

Up to an additive constant in q,

$$\log \mathcal{N}(x_{ij} \mid \theta_i^\top \beta_j, \sigma^2) = -\frac{1}{2\sigma^2}(x_{ij} - \theta_i^\top \beta_j)^2 + \text{const}.$$

So

$$\mathbb{E}_q[\log p(X \mid \Theta, B)] = -\frac{1}{2\sigma^2} \sum_{(i,j)\in\Omega} \mathbb{E}_q[(x_{ij} - \theta_i^\top \beta_j)^2] + \text{const},$$

with

$$\mathbb{E}_q[(x_{ij} - \theta_i^\top \beta_j)^2] = x_{ij}^2 - 2x_{ij}\, \mathbb{E}_q[\theta_i^\top \beta_j] + \mathbb{E}_q[(\theta_i^\top \beta_j)^2].$$

Prior terms.

$$\log p(\Theta) = \sum_{i=1}^{n} \log \mathcal{N}(\theta_i \mid 0, \eta_\theta^2 I_K),$$

$$\log p(B) = \sum_{j=1}^{m} \log \mathcal{N}(\beta_j \mid 0, \eta_B^2 I_K).$$

Up to constants,

$$\mathbb{E}_q[\log p(\Theta)] = -\frac{1}{2\eta_\theta^2} \sum_{i=1}^{n} \mathbb{E}_q[\theta_i^\top \theta_i] + \text{const},$$

$$\mathbb{E}_q[\log p(B)] = -\frac{1}{2\eta_B^2} \sum_{j=1}^{m} \mathbb{E}_q[\beta_j^\top \beta_j] + \text{const}.$$

Entropy (variational) terms.

$$\mathbb{E}_q[\log q(\Theta, B)] = \sum_{i=1}^{n} \mathbb{E}_q[\log q(\theta_i)] + \sum_{j=1}^{m} \mathbb{E}_q[\log q(\beta_j)].$$

For each Gaussian factor,

$$\mathbb{E}_q[\log q(\theta_i)] = -\frac{1}{2}\big(\log |V_{\theta_i}| + K(1 + \log 2\pi)\big),$$

$$\mathbb{E}_q[\log q(\beta_j)] = -\frac{1}{2}\big(\log |V_{\beta_j}| + K(1 + \log 2\pi)\big).$$

Putting it together (up to constants independent of q):

$$\begin{aligned}
\mathcal{L}(q) = &-\frac{1}{2\sigma^2} \sum_{(i,j)\in\Omega} \big(x_{ij}^2 - 2x_{ij}\, \mathbb{E}_q[\theta_i^\top \beta_j] + \mathbb{E}_q[(\theta_i^\top \beta_j)^2]\big) \\
&- \frac{1}{2\eta_\theta^2} \sum_i \mathbb{E}_q[\theta_i^\top \theta_i] - \frac{1}{2\eta_B^2} \sum_j \mathbb{E}_q[\beta_j^\top \beta_j] \\
&- \sum_i \mathbb{E}_q[\log q(\theta_i)] - \sum_j \mathbb{E}_q[\log q(\beta_j)] + \text{const}.
\end{aligned}$$

Expectations needed under q.

Under the mean-field Gaussians,

First and second moments of user factors:

$$\mathbb{E}_q[\theta_i] = m_{\theta_i}, \quad \mathbb{E}_q[\theta_i \theta_i^\top] = V_{\theta_i} + m_{\theta_i} m_{\theta_i}^\top.$$

First and second moments of item factors:

$$\mathbb{E}_q[\beta_j] = m_{\beta_j}, \quad \mathbb{E}_q[\beta_j \beta_j^\top] = V_{\beta_j} + m_{\beta_j} m_{\beta_j}^\top.$$

Cross terms for likelihood:

$$\mathbb{E}_q[\theta_i^\top \beta_j] = m_{\theta_i}^\top m_{\beta_j},$$

$$\mathbb{E}_q[(\theta_i^\top \beta_j)^2] = \mathbb{E}_q[\beta_j^\top (\theta_i \theta_i^\top) \beta_j] = \operatorname{tr}\big((V_{\theta_i} + m_{\theta_i} m_{\theta_i}^\top)(V_{\beta_j} + m_{\beta_j} m_{\beta_j}^\top)\big).$$

Entropy pieces for each Gaussian factor:

$\mathbb{E}_q[\log q(\theta_i)]$, $\mathbb{E}_q[\log q(\beta_j)]$ as above in terms of $V_{\theta_i}, V_{\beta_j}$.

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_{θ_i}, and m_{β_j}.

Let the mean-field family be

$$q(\Theta, B) = \prod_{i=1}^{n} q(\theta_i) \prod_{j=1}^{m} q(\beta_j)$$

with

$$q(\theta_i) = \mathcal{N}(m_{\theta_i}, V_{\theta_i}), \quad q(\beta_j) = \mathcal{N}(m_{\beta_j}, V_{\beta_j}).$$

Define the index sets

$$\Omega_i := \{j : (i,j) \in \Omega\}, \quad \Omega_j := \{i : (i,j) \in \Omega\}.$$

The given complete conditionals are

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

$$\mu_{\theta_i} = \Sigma_{\theta_i} \, \sigma^{-2} \sum_{j \in \Omega_i} \beta_j x_{ij},$$

and

$$p(\beta_j \mid \{\theta_i\}, X) = \mathcal{N}(\mu_{\beta_j}, \Sigma_{\beta_j}), \quad \Sigma_{\beta_j}^{-1} = \eta_B^{-2} I_K + \sigma^{-2} \sum_{i \in \Omega_j} \theta_i \theta_i^\top,$$

$$\mu_{\beta_j} = \Sigma_{\beta_j} \, \sigma^{-2} \sum_{i \in \Omega_j} \theta_i x_{ij}.$$

Using

$$\log q^\star(x_v) \propto \mathbb{E}_{-v}[\log p(x_v \mid x_{-v})],$$

the optimal variational factor has the same Gaussian form, with each occurrence of β_j, $\beta_j \beta_j^\top$ and θ_i, $\theta_i \theta_i^\top$ replaced by their expectations under q.

Because

$$\mathbb{E}_q[\beta_j] = m_{\beta_j}, \quad \mathbb{E}_q[\beta_j \beta_j^\top] = V_{\beta_j} + m_{\beta_j} m_{\beta_j}^\top,$$

$$\mathbb{E}_q[\theta_i] = m_{\theta_i}, \quad \mathbb{E}_q[\theta_i \theta_i^\top] = V_{\theta_i} + m_{\theta_i} m_{\theta_i}^\top,$$

the CAVI updates are:

User factors θ_i:

Precision:

$$V_{\theta_i}^{-1} = \eta_\theta^{-2} I_K + \sigma^{-2} \sum_{j \in \Omega_i} \mathbb{E}_q[\beta_j \beta_j^\top] = \eta_\theta^{-2} I_K + \sigma^{-2} \sum_{j \in \Omega_i} (V_{\beta_j} + m_{\beta_j} m_{\beta_j}^\top).$$

Mean:

$$m_{\theta_i} = V_{\theta_i}\, \sigma^{-2} \sum_{j \in \Omega_i} \mathbb{E}_q[\beta_j]\, x_{ij} = V_{\theta_i}\, \sigma^{-2} \sum_{j \in \Omega_i} m_{\beta_j}\, x_{ij}.$$

Item factors β_j:

Precision:

$$V_{\beta_j}^{-1} = \eta_B^{-2} I_K + \sigma^{-2} \sum_{i \in \Omega_j} \mathbb{E}_q[\theta_i \theta_i^\top] = \eta_B^{-2} I_K + \sigma^{-2} \sum_{i \in \Omega_j} (V_{\theta_i} + m_{\theta_i} m_{\theta_i}^\top).$$

Mean:

$$m_{\beta_j} = V_{\beta_j}\, \sigma^{-2} \sum_{i \in \Omega_j} \mathbb{E}_q[\theta_i]\, x_{ij} = V_{\beta_j}\, \sigma^{-2} \sum_{i \in \Omega_j} m_{\theta_i}\, x_{ij}.$$

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).

INPUT:

- X – ratings matrix (with missing entries)
- Ω – set of observed indices (i, j)
- Ω_i – for each i, $\Omega_i = \{j : (i, j) \in \Omega\}$
- Ω_j – for each j, $\Omega_j = \{i : (i, j) \in \Omega\}$
- σ^2 – noise variance
- η_θ^2, η_B^2 – prior variances
- K – latent dimension
- `tol` – convergence tolerance (e.g., 10^{-4})
- `max_iters` – maximum CAVI iterations

INITIALIZE:

- For $i = 1$ to n:
  - $m_{\theta_i} \leftarrow$ random K-vector
  - $V_{\theta_i} \leftarrow I_K$ ($K \times K$ matrix)
- For $j = 1$ to m:
  - $m_{\beta_j} \leftarrow$ random K-vector

```
        – $V_{\beta_j} \leftarrow I_K$
REPEAT for $t = 1, 2, \ldots,$ max_iters:
    // Store old means for convergence check
    $m_\theta^{\text{old}} \leftarrow \{m_{\theta_i}\}$ for all $i$
    $m_\beta^{\text{old}} \leftarrow \{m_{\beta_j}\}$ for all $j$
    1. Compute expectations under $q(\beta_j)$ for current iteration
    For $j = 1$ to $m$:
      • $\mathbb{E}_\beta[j] \leftarrow m_{\beta_j}$    // $\mathbb{E}_q[\beta_j]$
      • $\mathbb{E}_{\beta\text{-outer}}[j] \leftarrow V_{\beta_j} + m_{\beta_j} m_{\beta_j}^\top$    // $\mathbb{E}_q[\beta_j \beta_j^\top]$
    2. Update $q(\theta_i)$ block
    For $i = 1$ to $n$:
      • $\text{Prec}_{\theta_i} \leftarrow \eta_\theta^{-2} I_K$    // precision (inverse covariance) of $\theta_i$
      • $b_{\theta_i} \leftarrow \mathbf{0}_K$    // linear term for $\theta_i$ mean
      • For each $j \in \Omega_i$:
          – $\text{Prec}_{\theta_i} \leftarrow \text{Prec}_{\theta_i} + \sigma^{-2} \mathbb{E}_{\beta\text{-outer}}[j]$
          – $b_{\theta_i} \leftarrow b_{\theta_i} + \sigma^{-2} \mathbb{E}_\beta[j] \cdot X[i, j]$
      • $V_{\theta_i} \leftarrow \text{Prec}_{\theta_i}^{-1}$
      • $m_{\theta_i} \leftarrow V_{\theta_i} \cdot b_{\theta_i}$
    3. Compute expectations under updated $q(\theta_i)$
    For $i = 1$ to $n$:
      • $\mathbb{E}_\theta[i] \leftarrow m_{\theta_i}$    // $\mathbb{E}_q[\theta_i]$
      • $\mathbb{E}_{\theta\text{-outer}}[i] \leftarrow V_{\theta_i} + m_{\theta_i} m_{\theta_i}^\top$    // $\mathbb{E}_q[\theta_i \theta_i^\top]$
    4. Update $q(\beta_j)$ block
    For $j = 1$ to $m$:
      • $\text{Prec}_{\beta_j} \leftarrow \eta_B^{-2} I_K$    // precision (inverse covariance) of $\beta_j$
      • $b_{\beta_j} \leftarrow \mathbf{0}_K$    // linear term for $\beta_j$ mean
      • For each $i \in \Omega_j$:
          – $\text{Prec}_{\beta_j} \leftarrow \text{Prec}_{\beta_j} + \sigma^{-2} \mathbb{E}_{\theta\text{-outer}}[i]$
          – $b_{\beta_j} \leftarrow b_{\beta_j} + \sigma^{-2} \mathbb{E}_\theta[i] \cdot X[i, j]$
      • $V_{\beta_j} \leftarrow \text{Prec}_{\beta_j}^{-1}$
      • $m_{\beta_j} \leftarrow V_{\beta_j} \cdot b_{\beta_j}$
    5. Convergence criterion
      • $\Delta_\theta \leftarrow \max_i \|m_{\theta_i} - m_{\theta_i}^{\text{old}}\|_2$
      • $\Delta_\beta \leftarrow \max_j \|m_{\beta_j} - m_{\beta_j}^{\text{old}}\|_2$
```

- $\Delta \leftarrow \max(\Delta_\theta, \Delta_\beta)$
- If $\Delta <$ `tol`: **break**

OUTPUT:

$\{m_{\theta_i}, V_{\theta_i}\}$ for $i = 1, \ldots, n$

$\{m_{\beta_j}, V_{\beta_j}\}$ for $j = 1, \ldots, m$

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.

We consider a single held-out user-item pair $(i, j) \notin \Omega$.

The generative model for the rating is

$$x_{ij} \mid \theta_i, \beta_j \sim \mathcal{N}(\theta_i^\top \beta_j, \sigma^2).$$

Under the variational posterior,

$$q(\theta_i) = \mathcal{N}(m_{\theta_i}, V_{\theta_i}), \qquad q(\beta_j) = \mathcal{N}(m_{\beta_j}, V_{\beta_j}),$$

and, by the mean-field assumption,

$$q(\theta_i, \beta_j) = q(\theta_i)\, q(\beta_j),$$

so θ_i and β_j are independent under q.

The posterior predictive for x_{ij} is approximated by integrating out θ_i, β_j under q:

$$p(x_{ij} \mid \text{data}) \approx \iint p(x_{ij} \mid \theta_i, \beta_j)\, q(\theta_i)\, q(\beta_j)\, d\theta_i\, d\beta_j.$$

We now derive $\mathrm{Var}(x_{ij} \mid \text{data})$ under this approximation.

Step 1: Law of total variance.

Define the latent mean function

$$\mu_{ij}(\theta_i, \beta_j) := \theta_i^\top \beta_j.$$

The law of total variance gives

$$\mathrm{Var}(x_{ij} \mid \text{data}) \approx \mathbb{E}_q\big[\mathrm{Var}(x_{ij} \mid \theta_i, \beta_j)\big] + \mathrm{Var}_q\big(\mathbb{E}[x_{ij} \mid \theta_i, \beta_j]\big),$$

where the expectation and variance are with respect to $q(\theta_i, \beta_j)$.

From the Gaussian likelihood,

$$\mathrm{Var}(x_{ij} \mid \theta_i, \beta_j) = \sigma^2, \qquad \mathbb{E}[x_{ij} \mid \theta_i, \beta_j] = \mu_{ij}(\theta_i, \beta_j) = \theta_i^\top \beta_j.$$

Therefore,

$$\mathbb{E}_q\big[\mathrm{Var}(x_{ij} \mid \theta_i, \beta_j)\big] = \mathbb{E}_q[\sigma^2] = \sigma^2,$$

and

$$\mathrm{Var}_q\big(\mathbb{E}[x_{ij} \mid \theta_i, \beta_j]\big) = \mathrm{Var}_q(\theta_i^\top \beta_j).$$

So

$$\mathrm{Var}(x_{ij} \mid \mathrm{data}) \approx \sigma^2 + \mathrm{Var}_q(\theta_i^\top \beta_j).$$

The problem is now reduced to computing $\mathrm{Var}_q(\theta_i^\top \beta_j)$.

Step 2: Moments of $\theta_i^\top \beta_j$ under q.

Under q,

$$\theta_i \sim \mathcal{N}(m_{\theta_i}, V_{\theta_i}), \qquad \beta_j \sim \mathcal{N}(m_{\beta_j}, V_{\beta_j}),$$

with $\theta_i \perp \beta_j$ (independent).

We will compute:

- $\mathbb{E}_q[\theta_i^\top \beta_j]$,
- $\mathbb{E}_q[(\theta_i^\top \beta_j)^2]$,
- $\mathrm{Var}_q(\theta_i^\top \beta_j) = \mathbb{E}_q[(\theta_i^\top \beta_j)^2] - \big(\mathbb{E}_q[\theta_i^\top \beta_j]\big)^2$.

Step 2.1: First moment $\mathbb{E}_q[\theta_i^\top \beta_j]$.

Using linearity of expectation and independence:

$$\mathbb{E}_q[\theta_i^\top \beta_j] = \mathbb{E}_q[\theta_i]^\top \mathbb{E}_q[\beta_j] = m_{\theta_i}^\top m_{\beta_j}.$$

Step 2.2: Second moment $\mathbb{E}_q[(\theta_i^\top \beta_j)^2]$.

Start from the scalar square:

$$(\theta_i^\top \beta_j)^2 = (\beta_j^\top \theta_i)(\theta_i^\top \beta_j) = \beta_j^\top (\theta_i \theta_i^\top) \beta_j.$$

Use the identity for quadratic forms in terms of trace:

$$x^\top A x = \mathrm{tr}(A x x^\top).$$

Here, take $x = \beta_j$, $A = \theta_i \theta_i^\top$. Then

$$\beta_j^\top (\theta_i \theta_i^\top) \beta_j = \mathrm{tr}(\theta_i \theta_i^\top \, \beta_j \beta_j^\top).$$

Thus

$$(\theta_i^\top \beta_j)^2 = \mathrm{tr}(\theta_i \theta_i^\top \, \beta_j \beta_j^\top).$$

Now take expectation with respect to q:

$$\mathbb{E}_q[(\theta_i^\top \beta_j)^2] = \mathbb{E}_q\big[\,\mathrm{tr}(\theta_i \theta_i^\top \, \beta_j \beta_j^\top)\big].$$

Use linearity of trace and expectation:

$$\mathbb{E}_q\big[\,\mathrm{tr}(\theta_i \theta_i^\top \, \beta_j \beta_j^\top)\big] = \mathrm{tr}\,\big(\mathbb{E}_q[\theta_i \theta_i^\top \, \beta_j \beta_j^\top]\big).$$

Under the mean-field factorization, θ_i and β_j are independent, so

$$\mathbb{E}_q[\theta_i \theta_i^\top \, \beta_j \beta_j^\top] = \mathbb{E}_q[\theta_i \theta_i^\top]\, \mathbb{E}_q[\beta_j \beta_j^\top].$$

Define the second-moment matrices

$$S_{\theta_i} := \mathbb{E}_q[\theta_i \theta_i^\top], \qquad S_{\beta_j} := \mathbb{E}_q[\beta_j \beta_j^\top].$$

Then

$$\mathbb{E}_q[(\theta_i^\top \beta_j)^2] = \operatorname{tr}\left(S_{\theta_i} S_{\beta_j}\right).$$

For Gaussian factors,

$$S_{\theta_i} = V_{\theta_i} + m_{\theta_i} m_{\theta_i}^\top, \qquad S_{\beta_j} = V_{\beta_j} + m_{\beta_j} m_{\beta_j}^\top.$$

Step 2.3: Variance $\mathrm{Var}_q(\theta_i^\top \beta_j)$.

By definition,

$$\mathrm{Var}_q(\theta_i^\top \beta_j) = \mathbb{E}_q[(\theta_i^\top \beta_j)^2] - \left(\mathbb{E}_q[\theta_i^\top \beta_j]\right)^2.$$

We already have:

- $\mathbb{E}_q[(\theta_i^\top \beta_j)^2] = \operatorname{tr}(S_{\theta_i} S_{\beta_j})$,
- $\mathbb{E}_q[\theta_i^\top \beta_j] = m_{\theta_i}^\top m_{\beta_j}$.

Thus

$$\mathrm{Var}_q(\theta_i^\top \beta_j) = \operatorname{tr}(S_{\theta_i} S_{\beta_j}) - (m_{\theta_i}^\top m_{\beta_j})^2.$$

Using $S_{\theta_i} = V_{\theta_i} + m_{\theta_i} m_{\theta_i}^\top$ and $S_{\beta_j} = V_{\beta_j} + m_{\beta_j} m_{\beta_j}^\top$, expanding and simplifying yields the equivalent expression

$$\mathrm{Var}_q(\theta_i^\top \beta_j) = \operatorname{tr}(V_{\theta_i} V_{\beta_j}) + m_{\beta_j}^\top V_{\theta_i} m_{\beta_j} + m_{\theta_i}^\top V_{\beta_j} m_{\theta_i}.$$

Step 3: Posterior predictive variance.

Putting everything together:

$$\mathrm{Var}(x_{ij} \mid \text{data}) \approx \sigma^2 + \mathrm{Var}_q(\theta_i^\top \beta_j).$$

So the two equivalent formulas are:

Compact form (via second-moment matrices):

$$\boxed{\mathrm{Var}(x_{ij} \mid \text{data}) \approx \sigma^2 + \operatorname{tr}\left(S_{\theta_i} S_{\beta_j}\right) - (m_{\theta_i}^\top m_{\beta_j})^2}$$

with

$$S_{\theta_i} = V_{\theta_i} + m_{\theta_i} m_{\theta_i}^\top, \qquad S_{\beta_j} = V_{\beta_j} + m_{\beta_j} m_{\beta_j}^\top.$$

Expanded form (in terms of V and m only):

$$\boxed{\mathrm{Var}(x_{ij} \mid \text{data}) \approx \sigma^2 + \operatorname{tr}(V_{\theta_i} V_{\beta_j}) + m_{\beta_j}^\top V_{\theta_i} m_{\beta_j} + m_{\theta_i}^\top V_{\beta_j} m_{\theta_i}}$$

Here $m_{\theta_i}, V_{\theta_i}, m_{\beta_j}, V_{\beta_j}$ are the current variational means and covariances from the MFVI algorithm.

Using this variance to flag uncertain recommendations.

For a held-out pair $(i, j) \notin \Omega$:

- *Predictive mean:*
$$\hat{x}_{ij} := \mathbb{E}_q[x_{ij} \mid \text{data}] \approx m_{\theta_i}^\top m_{\beta_j}.$$
- *Predictive variance:*
$$\widehat{\text{Var}}(x_{ij} \mid \text{data}) \approx \sigma^2 + \text{Var}_q(\theta_i^\top \beta_j),$$
using either of the formulas above.

To flag uncertain recommendations, one can:

(a) Compute $\widehat{\text{Var}}(x_{ij} \mid \text{data})$ for all candidate (i, j) in the recommendation pool.

(b) Choose a threshold τ (e.g., the 80th or 90th percentile of the variance values).

(c) Mark all pairs with
$$\widehat{\text{Var}}(x_{ij} \mid \text{data}) > \tau$$
as high-uncertainty.

(d) Use this flag to:
   - avoid recommending high-uncertainty items,
   - or surface them only as "exploration" queries,
   - or ask the user for more feedback on those items.

Pairs with lower variance are treated as more reliable, high-confidence recommendations.

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

Let $\Omega \subset \{1, \dots, n\} \times \{1, \dots, m\}$ be the set of observed entries.

Split Ω into a training and validation set:
$$\Omega_{\text{train}} \cup \Omega_{\text{val}} = \Omega, \quad \Omega_{\text{train}} \cap \Omega_{\text{val}} = \varnothing.$$

During fitting, only entries in Ω_{train} are used.

Step 1: Fit MFVI for a fixed hyperparameter triple.

Fix a candidate triple $(\eta_\theta^2, \eta_B^2, \sigma^2)$.

Run CAVI on Ω_{train} until convergence, obtaining variational factors
$$q(\theta_i) = \mathcal{N}(m_{\theta_i}, V_{\theta_i}), \qquad q(\beta_j) = \mathcal{N}(m_{\beta_j}, V_{\beta_j})$$
for all i, j that appear in Ω_{train}.

Step 2: Posterior predictive on the validation set.

For each held-out pair $(i, j) \in \Omega_{\text{val}}$, approximate the posterior predictive distribution of x_{ij}.

Predictive mean:
$$\hat{\mu}_{ij} := \mathbb{E}_q[x_{ij} \mid \text{data}] \approx m_{\theta_i}^\top m_{\beta_j}.$$

Define
$$S_{\theta_i} := V_{\theta_i} + m_{\theta_i} m_{\theta_i}^\top, \qquad S_{\beta_j} := V_{\beta_j} + m_{\beta_j} m_{\beta_j}^\top,$$
so that
$$\mathbb{E}_q[\theta_i^\top \beta_j] = m_{\theta_i}^\top m_{\beta_j},$$
$$\mathbb{E}_q[(\theta_i^\top \beta_j)^2] = \text{tr}(S_{\theta_i} S_{\beta_j}),$$

and hence

$$\mathrm{Var}_q(\theta_i^\top \beta_j) = \mathrm{tr}(S_{\theta_i} S_{\beta_j}) - (m_{\theta_i}^\top m_{\beta_j})^2.$$

Using the law of total variance,

$$\mathrm{Var}(x_{ij} \mid \text{data}) \approx \sigma^2 + \mathrm{Var}_q(\theta_i^\top \beta_j) =: \hat{v}_{ij}.$$

Thus the approximate posterior predictive for each held-out entry is

$$p(x_{ij} \mid \text{data}) \approx \mathcal{N}\big(x_{ij};\ \hat{\mu}_{ij},\ \hat{v}_{ij}\big).$$

Step 3: Validation objective.

Define the held-out predictive log-likelihood for this triple $(\eta_\theta^2, \eta_B^2, \sigma^2)$ as

$$\mathcal{L}_{\text{val}}(\eta_\theta^2, \eta_B^2, \sigma^2) = \sum_{(i,j)\in\Omega_{\text{val}}} \log \mathcal{N}\big(x_{ij};\ \hat{\mu}_{ij},\ \hat{v}_{ij}\big).$$

Step 4: Hyperparameter selection and refitting.

Search over a grid (or finite set) of candidate values for $(\eta_\theta^2, \eta_B^2, \sigma^2)$.

For each candidate, compute $\mathcal{L}_{\text{val}}$ as above.

Select

$$(\eta_\theta^2, \eta_B^2, \sigma^2)^\star = \arg\max_{(\eta_\theta^2, \eta_B^2, \sigma^2)} \mathcal{L}_{\text{val}}(\eta_\theta^2, \eta_B^2, \sigma^2).$$

Finally, with $(\eta_\theta^2, \eta_B^2, \sigma^2)^\star$ fixed, refit the variational MF model on all observed entries Ω.

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:

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

(Given) Conditional for γ. You may use

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

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 $\mathbb{E}_q[\theta_i]^\top \mathbb{E}_q[\beta_j]$.

Start from the given complete conditional for γ:

$$p(\gamma \mid \Theta, B, X, G) = \mathcal{N}(\mu_\gamma, \Sigma_\gamma),$$

with

$$\Sigma_\gamma^{-1} = \eta_\gamma^{-2} I_p + \sigma^{-2} \sum_{(i,j)\in\Omega} g_j g_j^\top,$$

$$\mu_\gamma = \Sigma_\gamma \sigma^{-2} \sum_{(i,j)\in\Omega} g_j (x_{ij} - \theta_i^\top \beta_j).$$

Write its log-density in quadratic form. For some constant $C(\Theta, B)$ that does not depend on γ,

$$\log p(\gamma \mid \Theta, B, X, G) = -\frac{1}{2}\gamma^\top \Sigma_\gamma^{-1} \gamma + \gamma^\top \Sigma_\gamma^{-1} \mu_\gamma + C(\Theta, B).$$

Substitute the expressions for Σ_γ^{-1} and μ_γ:

$$\Sigma_\gamma^{-1} = A \quad \text{where} \quad A := \eta_\gamma^{-2} I_p + \sigma^{-2} \sum_{(i,j)\in\Omega} g_j g_j^\top,$$

$$\Sigma_\gamma^{-1}\mu_\gamma = A\mu_\gamma = \sigma^{-2} \sum_{(i,j)\in\Omega} g_j (x_{ij} - \theta_i^\top \beta_j).$$

So the conditional can be written as

$$\log p(\gamma \mid \Theta, B, X, G) = -\frac{1}{2}\gamma^\top A \gamma + \gamma^\top b(\Theta, B) + C(\Theta, B),$$

where

$$b(\Theta, B) := \sigma^{-2} \sum_{(i,j)\in\Omega} g_j (x_{ij} - \theta_i^\top \beta_j).$$

Note that:

- A depends only on $G, \sigma^2, \eta_\gamma^2$, not on Θ, B.
- All dependence on Θ, B is in the linear term $b(\Theta, B)$ and the constant $C(\Theta, B)$.

Apply the CAVI identity.

The mean-field family is extended as

$$q(\Theta, B, \gamma) = \left(\prod_i q(\theta_i)\right)\left(\prod_j q(\beta_j)\right) q(\gamma),$$

with

$$q(\theta_i) = \mathcal{N}(m_{\theta_i}, V_{\theta_i}), \quad q(\beta_j) = \mathcal{N}(m_{\beta_j}, V_{\beta_j}), \quad q(\gamma) = \mathcal{N}(m_\gamma, V_\gamma).$$

The CAVI identity for the optimal factor is

$$\log q^\star(\gamma) \propto \mathbb{E}_{q(\Theta,B)}[\log p(\gamma \mid \Theta, B, X, G)].$$

Insert the quadratic form:

$$\log q^\star(\gamma) \propto \mathbb{E}_{q(\Theta,B)}\left[-\frac{1}{2}\gamma^\top A\gamma + \gamma^\top b(\Theta, B) + C(\Theta, B)\right].$$

The expectation over Θ, B does not touch γ, so:

- The quadratic term is unchanged, because A is constant:

$$\mathbb{E}_{q(\Theta,B)}\left[-\frac{1}{2}\gamma^\top A\gamma\right] = -\frac{1}{2}\gamma^\top A\gamma.$$

- The linear term becomes $\gamma^\top \mathbb{E}_q[b(\Theta, B)]$.
- The constant term becomes another constant (irrelevant for normalization of $q(\gamma)$).

Thus

$$\log q^\star(\gamma) \propto -\frac{1}{2}\gamma^\top A\gamma + \gamma^\top\, \mathbb{E}_q[b(\Theta, B)] + \text{const}.$$

This is again the log-density of a multivariate Normal in γ with precision matrix A and mean $A^{-1}\mathbb{E}_q[b(\Theta, B)]$.

Computing $\mathbb{E}_q[b(\Theta, B)]$.

Recall

$$b(\Theta, B) = \sigma^{-2} \sum_{(i,j)\in\Omega} g_j(x_{ij} - \theta_i^\top \beta_j).$$

Take the expectation under the mean-field $q(\Theta, B)$:

$$\mathbb{E}_q[b(\Theta, B)] = \sigma^{-2} \sum_{(i,j)\in\Omega} g_j(x_{ij} - \mathbb{E}_q[\theta_i^\top \beta_j]).$$

Under mean-field,

$$q(\Theta, B) = \left(\prod_i q(\theta_i)\right)\left(\prod_j q(\beta_j)\right) \Rightarrow \theta_i \perp \beta_j \text{ under } q,$$

so

$$\mathbb{E}_q[\theta_i^\top \beta_j] = \mathbb{E}_q[\theta_i]^\top \mathbb{E}_q[\beta_j] = m_{\theta_i}^\top m_{\beta_j}.$$

Therefore

$$\mathbb{E}_q[b(\Theta, B)] = \sigma^{-2} \sum_{(i,j)\in\Omega} g_j(x_{ij} - m_{\theta_i}^\top m_{\beta_j}).$$

Final CAVI update for $q(\gamma)$.

From the quadratic form,

$$q^\star(\gamma) = \mathcal{N}(m_\gamma, V_\gamma),$$

with

Precision:

$$V_\gamma^{-1} = A = \eta_\gamma^{-2} I_p + \sigma^{-2} \sum_{(i,j)\in\Omega} g_j g_j^\top,$$

Mean:

$$m_\gamma = V_\gamma\, \mathbb{E}_q[b(\Theta, B)] = V_\gamma\, \sigma^{-2} \sum_{(i,j)\in\Omega} g_j(x_{ij} - m_{\theta_i}^\top m_{\beta_j}).$$

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:

$$m_{\theta_i} = V_{\theta_i}\sigma^{-2}\sum_{j\in\Omega_i} m_{\beta_j}(x_{ij} - g_j^\top m_\gamma), \quad m_{\beta_j} = V_{\beta_j}\sigma^{-2}\sum_{i\in\Omega^j} m_{\theta_i}(x_{ij} - g_j^\top m_\gamma).$$

Explain why this residualization is the only change.

Start from the feature-augmented likelihood for a single observed pair $(i, j) \in \Omega$:

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

Define the residualized response

$$r_{ij} := x_{ij} - g_j^\top \gamma.$$

Then

$$r_{ij} \mid \theta_i, \beta_j, \gamma \sim \mathcal{N}\big(\theta_i^\top \beta_j,\ \sigma^2\big),$$

which has the same form as the original model without features, but with x_{ij} replaced by r_{ij}.

1. Conditional for θ_i: precision unchanged, mean residualized.

Collect all terms in the joint log-density that depend on θ_i:

$$\log p(\theta_i \mid \{\beta_j\}, \gamma, X, G) = \log p(\theta_i) + \sum_{j\in\Omega_i} \log p(x_{ij} \mid \theta_i, \beta_j, \gamma) + \text{const},$$

where $\Omega_i = \{j : (i, j) \in \Omega\}$.

Using the Gaussian prior $\theta_i \sim \mathcal{N}(0, \eta_\theta^2 I_K)$ and the residual form,

$$\log p(\theta_i) = -\tfrac{1}{2}\eta_\theta^{-2}\,\theta_i^\top \theta_i + \text{const},$$

$$\log p(x_{ij} \mid \theta_i, \beta_j, \gamma) = -\tfrac{1}{2\sigma^2}\big(r_{ij} - \theta_i^\top \beta_j\big)^2 + \text{const}.$$

Expand the likelihood term in θ_i:

$$-\tfrac{1}{2\sigma^2}\big(r_{ij} - \theta_i^\top \beta_j\big)^2 = -\tfrac{1}{2\sigma^2}\big((\theta_i^\top \beta_j)^2 - 2r_{ij}\,\theta_i^\top \beta_j + r_{ij}^2\big).$$

The contribution of all $j \in \Omega_i$ to the log-density in θ_i is:

- *Quadratic term:*

$$-\tfrac{1}{2}\theta_i^\top \left(\eta_\theta^{-2} I_K + \sigma^{-2}\sum_{j\in\Omega_i} \beta_j\beta_j^\top\right)\theta_i.$$

- *Linear term:*

$$\theta_i^\top \left(\sigma^{-2}\sum_{j\in\Omega_i} \beta_j r_{ij}\right) = \theta_i^\top \left(\sigma^{-2}\sum_{j\in\Omega_i} \beta_j(x_{ij} - g_j^\top \gamma)\right).$$

Thus the conditional remains Gaussian,

$$p(\theta_i \mid \{\beta_j\}, \gamma, X, G) = \mathcal{N}\big(\mu_{\theta_i}^{\text{cond}}, \Sigma_{\theta_i}^{\text{cond}}\big),$$

with

- *Precision:*
$$(\Sigma^{\text{cond}}_{\theta_i})^{-1} = \eta_\theta^{-2} I_K + \sigma^{-2} \sum_{j \in \Omega_i} \beta_j \beta_j^\top,$$
exactly the same as in the model without features.
- *Mean:*
$$\mu^{\text{cond}}_{\theta_i} = \Sigma^{\text{cond}}_{\theta_i} \, \sigma^{-2} \sum_{j \in \Omega_i} \beta_j (x_{ij} - g_j^\top \gamma).$$

In mean-field VI, γ and β_j are replaced by their expectations:

$$\mathbb{E}_q[\gamma] = m_\gamma, \quad \mathbb{E}_q[\beta_j] = m_{\beta_j}.$$

Hence the CAVI update for $q(\theta_i) = \mathcal{N}(m_{\theta_i}, V_{\theta_i})$ is:

- *Precision:*
$$V_{\theta_i}^{-1} = \eta_\theta^{-2} I_K + \sigma^{-2} \sum_{j \in \Omega_i} \mathbb{E}_q[\beta_j \beta_j^\top] \quad \text{(same structure as before, no } g_j \text{ term)},$$
- *Mean:*
$$m_{\theta_i} = V_{\theta_i} \, \sigma^{-2} \sum_{j \in \Omega_i} m_{\beta_j} \, (x_{ij} - g_j^\top m_\gamma),$$

which is exactly the original mean update with the residualized response $x_{ij} - g_j^\top m_\gamma$ in place of x_{ij}.

2. Conditional for β_j: precision unchanged, mean residualized.

The same argument applies for β_j. Collect the terms involving β_j:

$$\log p(\beta_j \mid \{\theta_i\}, \gamma, X, G) = \log p(\beta_j) + \sum_{i \in \Omega^j} \log p(x_{ij} \mid \theta_i, \beta_j, \gamma) + \text{const},$$

where $\Omega^j = \{i : (i, j) \in \Omega\}$.

Using the residuals $r_{ij} = x_{ij} - g_j^\top \gamma$ and the prior $\beta_j \sim \mathcal{N}(0, \eta_B^2 I_K)$, the same expansion yields:

- *Precision:*
$$(\Sigma^{\text{cond}}_{\beta_j})^{-1} = \eta_B^{-2} I_K + \sigma^{-2} \sum_{i \in \Omega^j} \theta_i \theta_i^\top,$$
again unchanged from the no-feature model.
- *Mean:*
$$\mu^{\text{cond}}_{\beta_j} = \Sigma^{\text{cond}}_{\beta_j} \, \sigma^{-2} \sum_{i \in \Omega^j} \theta_i (x_{ij} - g_j^\top \gamma).$$

Taking expectations under q gives the CAVI update for $q(\beta_j) = \mathcal{N}(m_{\beta_j}, V_{\beta_j})$:

- *Precision:*
$$V_{\beta_j}^{-1} = \eta_B^{-2} I_K + \sigma^{-2} \sum_{i \in \Omega^j} \mathbb{E}_q[\theta_i \theta_i^\top],$$
- *Mean:*
$$m_{\beta_j} = V_{\beta_j} \, \sigma^{-2} \sum_{i \in \Omega^j} m_{\theta_i} \, (x_{ij} - g_j^\top m_\gamma).$$

This matches the stated residualized form.

3. Why residualization is the only change.

The key point is how the additive feature term enters the likelihood.

The log-likelihood contribution for a single (i, j) is

$$-\frac{1}{2\sigma^2}\big(x_{ij} - \theta_i^\top \beta_j - g_j^\top \gamma\big)^2.$$

Expanding in θ_i or β_j:

- The quadratic term in θ_i or β_j comes from $(\theta_i^\top \beta_j)^2$ and leads to the precision matrices. This term does not involve $g_j^\top \gamma$, so the Hessian with respect to θ_i or β_j is unchanged. Consequently, the precisions $V_{\theta_i}^{-1}$ and $V_{\beta_j}^{-1}$ are identical to the no-feature case.
- The linear term in θ_i (or β_j) involves the product $\beta_j(x_{ij} - g_j^\top \gamma)$ (or $\theta_i(x_{ij} - g_j^\top \gamma)$). Here, the feature term appears only by shifting the effective response from x_{ij} to the residual $x_{ij} - g_j^\top \gamma$.
- Under mean-field VI, γ is replaced by its mean m_γ, so every occurrence of x_{ij} in the mean updates is replaced by $x_{ij} - g_j^\top m_\gamma$.

Therefore:

- The curvature of the conditional log-densities in θ_i and β_j is unchanged.
- Only the linear term (and thus the conditional mean) is modified, through a simple residualization of the observed rating by the feature contribution.

This is why the only change in the CAVI updates is

$$x_{ij} \quad \longrightarrow \quad x_{ij} - g_j^\top m_\gamma$$

inside the mean formulas, while the precision matrices remain the same.

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?

A natural extension that incorporates user features $f_i \in \mathbb{R}^p$ is to add a user-side linear term to the mean of the Gaussian observation model. Introduce a coefficient vector $\delta \in \mathbb{R}^p$ and define

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

The variational family is then extended with a Gaussian factor

$$q(\delta) = \mathcal{N}(m_\delta, V_\delta).$$

It is convenient to rewrite the likelihood in terms of residuals. For a single pair (i, j),

$$x_{ij} - g_j^\top \gamma - f_i^\top \delta \mid \theta_i, \beta_j, \gamma, \delta \sim \mathcal{N}(\theta_i^\top \beta_j,\ \sigma^2),$$

so the likelihood has the same form as the original matrix factorization model, but with

$$x_{ij} \text{ replaced by } r_{ij} := x_{ij} - g_j^\top \gamma - f_i^\top \delta.$$

Under the mean-field approximation, γ and δ are replaced by their variational means m_γ and m_δ, giving the effective residual

$$r_{ij} \approx x_{ij} - g_j^\top m_\gamma - f_i^\top m_\delta.$$

Effect on the CAVI updates.

The structure of the CAVI updates for θ_i, β_j, and γ can be described in terms of this residualization.

Updates for $q(\theta_i)$ and $q(\beta_j)$.

The Gaussian complete conditionals for θ_i and β_j have precisions (inverse variances) determined by the quadratic terms in θ_i and β_j:

$$V_{\theta_i}^{-1} = \eta_\theta^{-2} I_K + \sigma^{-2} \sum_{j \in \Omega_i} \beta_j \beta_j^\top, \quad V_{\beta_j}^{-1} = \eta_B^{-2} I_K + \sigma^{-2} \sum_{i \in \Omega^j} \theta_i \theta_i^\top.$$

These terms do not involve x_{ij}, g_j, or f_i, so under variational expectations the precisions remain unchanged:

$$V_{\theta_i}^{-1} = \eta_\theta^{-2} I_K + \sigma^{-2} \sum_{j \in \Omega_i} \mathbb{E}_q[\beta_j \beta_j^\top], \quad V_{\beta_j}^{-1} = \eta_B^{-2} I_K + \sigma^{-2} \sum_{i \in \Omega^j} \mathbb{E}_q[\theta_i \theta_i^\top].$$

The linear terms in the complete conditionals, however, are modified by the residuals. Under the extended model, the CAVI means become

$$m_{\theta_i} = V_{\theta_i}\, \sigma^{-2} \sum_{j \in \Omega_i} m_{\beta_j}\, (x_{ij} - g_j^\top m_\gamma - f_i^\top m_\delta),$$

$$m_{\beta_j} = V_{\beta_j}\, \sigma^{-2} \sum_{i \in \Omega^j} m_{\theta_i}\, (x_{ij} - g_j^\top m_\gamma - f_i^\top m_\delta),$$

so the only change relative to the no-user-feature case is an additional residualization by the user feature term $f_i^\top m_\delta$.

Update for $q(\gamma)$.

The conditional distribution for γ now involves residuals that subtract both the latent factor contribution and the user feature contribution:

$$x_{ij} - \theta_i^\top \beta_j - f_i^\top \delta.$$

Taking expectations under q, the term $(x_{ij} - \theta_i^\top \beta_j)$ in the given conditional is replaced by

$$x_{ij} - \mathbb{E}_q[\theta_i]^\top \mathbb{E}_q[\beta_j] - f_i^\top m_\delta = x_{ij} - m_{\theta_i}^\top m_{\beta_j} - f_i^\top m_\delta.$$

The precision V_γ^{-1} retains the same form as in the item-feature-only model, since it depends only on $\sum_{(i,j) \in \Omega} g_j g_j^\top$, which is unchanged.

New update for $q(\delta)$.

The additional parameter δ has a Gaussian complete conditional analogous to that of γ, but built from user features f_i rather than item features g_j. Its precision involves

$$V_\delta^{-1} = \eta_\delta^{-2} I_p + \sigma^{-2} \sum_{(i,j) \in \Omega} f_i f_i^\top,$$

and its mean is proportional to a sum of residuals of the form

$$x_{ij} - \theta_i^\top \beta_j - g_j^\top \gamma,$$

which, under the variational approximation, become

$$x_{ij} - m_{\theta_i}^\top m_{\beta_j} - g_j^\top m_\gamma.$$

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

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,.

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

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.**

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 3: Project-aligned Probabilistic Latent Factor Model

(i) Introduction

In this section I implement a semi-supervised, probabilistic latent factor model to analyze a large panel of New York City property records. The primary goal is twofold: (1) to learn a low-dimensional latent representation of parcels that captures meaningful structure across assessed values, land characteristics, and building attributes; and (2) to use this representation to model the (log) sale price, including when prices are missing or partially observed. This setup matches the "project-aligned" option because the same model architecture underlies my course project on robust price imputation and latent structure in urban property markets.

The dataset is high-dimensional, noisy, and heavily skewed in both covariates and outcomes. A flexible generative model with latent factors is therefore natural: it can exploit shared structure across covariates, explicitly model missingness, and provide both point predictions and uncertainty estimates for (log) sale prices. I use a semi-supervised variant of a Missing-data Importance-Weighted Autoencoder (MIWAE) with a Student-t mixture prior over latent factors, trained end-to-end on approximately 1.2×10^5 NYC parcels.

(ii) Model

Let $i = 1, \ldots, n$ index properties. For each property we observe a covariate vector $x_i \in \mathbb{R}^D$ (parcel and building characteristics) and, for a subset, a scalar outcome $y_i = \log(\text{sale\_price}_i)$. The model introduces a latent factor $z_i \in \mathbb{R}^K$ with $K = 3$.

Prior on latent factors. I place a K-component Student-t mixture prior on z_i:

$$c_i \sim \text{Categorical}(\pi), \qquad z_i \mid c_i = k \sim t_\nu\big(\mu_k, \Sigma_k\big),$$

where $\pi \in \Delta^{K-1}$ are mixture weights, $\nu > 0$ is the degrees-of-freedom parameter, and (μ_k, Σ_k) are component means and covariances. In the implementation, the prior is parameterized in terms of unconstrained logits and Cholesky factors, but conceptually it is a mixture of heavy-tailed components that can capture multi-modal, robust structure in latent space.

Likelihood for covariates. Conditional on z_i, the observed covariates x_i are generated by a diagonal-covariance Gaussian decoder:

$$x_i \mid z_i \sim \mathcal{N}\big(\mu_x(z_i), \text{diag}(\sigma_x^2(z_i))\big),$$

where $\mu_x(\cdot)$ and $\log \sigma_x^2(\cdot)$ are outputs of a neural decoder network. In practice, each feature dimension is scaled to have approximately zero mean and unit variance before training, so the decoder operates in standardized units.

Likelihood for log sale price. For the (log) sale price y_i, I add a supervised Gaussian head on top of z_i:

$$y_i \mid z_i \sim \mathcal{N}\big(\mu_y(z_i), \sigma_y^2(z_i)\big),$$

with $\mu_y(\cdot)$ and $\log \sigma_y^2(\cdot)$ given by a small "price head" network that takes z_i as input. This defines a conditional generative model for y_i given z_i and allows the same latent factors to jointly explain both covariates and prices.

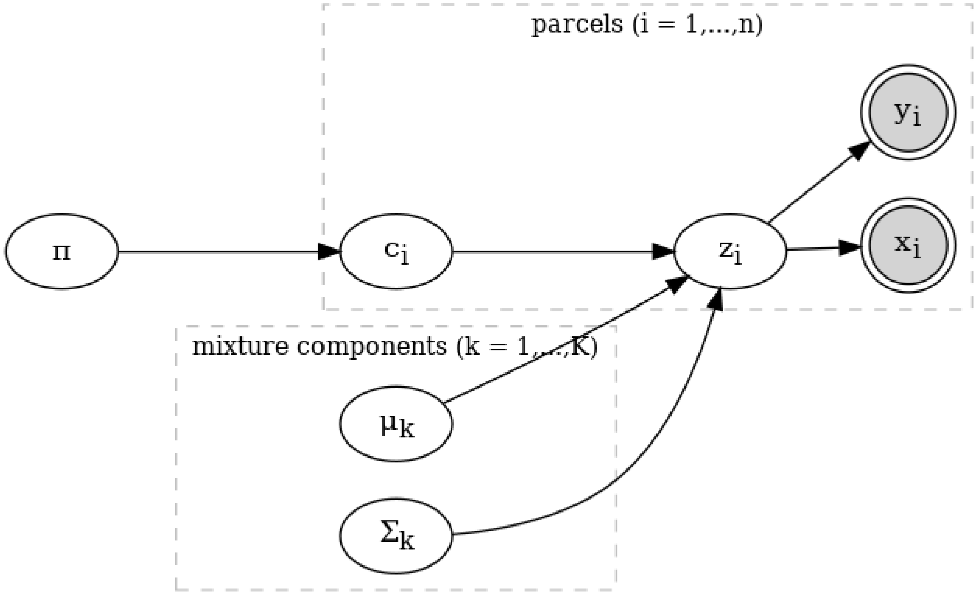


Figure 1: Graphical model of the semi-supervised MIWAE with a Student-t mixture prior.

Joint distribution. The full joint distribution of observed and latent variables for one property is

$$p(x_i, y_i, z_i, c_i) = p(c_i)\, p(z_i \mid c_i)\, p(x_i \mid z_i)\, p(y_i \mid z_i)^{\mathbb{I}\{y_i \text{ observed}\}},$$

and across properties I assume conditional independence given the global parameters:

$$p(X, Y, Z, C) = \prod_{i=1}^{n} p(x_i, y_i, z_i, c_i),$$

where Y includes only observed y_i in the product. Missing prices simply remove the corresponding likelihood factor while still contributing to learning via x_i.

A schematic graphical model for a single property is shown in Figure 1.

(iii) Inference

Direct maximum-likelihood estimation is intractable because the posterior $p(z_i, c_i \mid x_i, y_i)$ is non-Gaussian and mixture-structured. I therefore use amortized variational inference with an importance-weighted objective (MIWAE):

Variational family. For each property I introduce an encoder $q_\phi(z_i \mid x_i, y_i)$, parameterized as a diagonal-covariance Gaussian

$$q_\phi(z_i \mid x_i, y_i) = \mathcal{N}\big(m_\phi(x_i, y_i), \operatorname{diag}(v_\phi(x_i, y_i))\big),$$

with $m_\phi(\cdot)$ and $\log v_\phi(\cdot)$ implemented by a neural network. The discrete mixture indicator c_i is marginalized out analytically through the prior; I do not introduce a separate variational distribution over c_i.

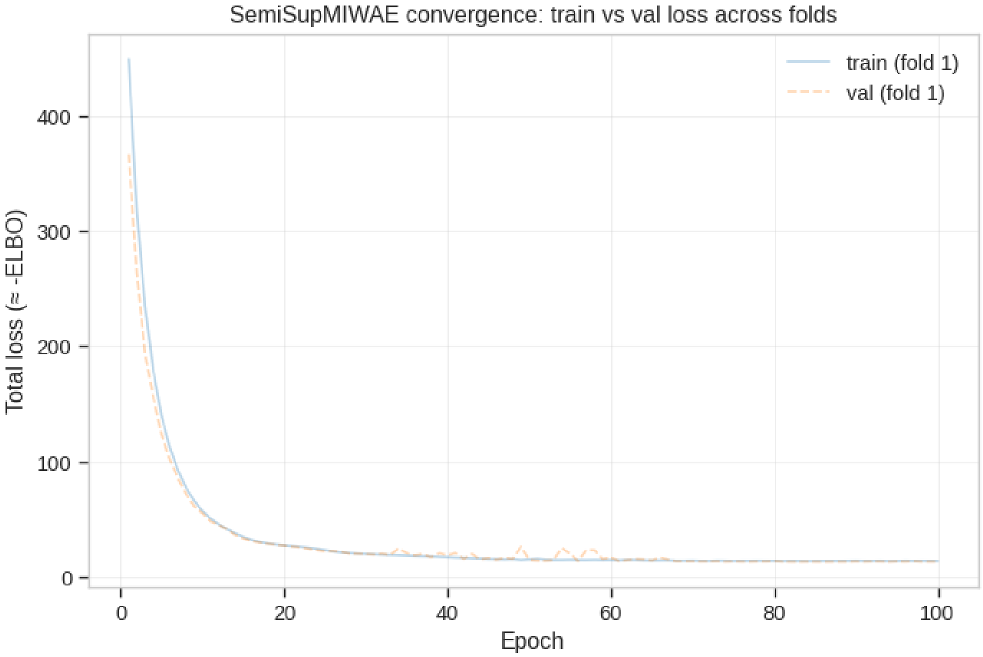


Figure 2: Convergence diagnostics for the SemiSupMIWAE model: training and validation loss versus epoch.

MIWAE objective. For partially observed (x_i, y_i) pairs, the training objective aggregates K_{IW} importance samples from q_ϕ and maximizes a Monte Carlo estimate of the log marginal likelihood:

$$\mathcal{L}_{\mathrm{MIWAE}}(\theta, \phi) = \sum_{i=1}^{n} \mathbb{E}_{z_i^{(1:K_{\mathrm{IW}})} \sim q_\phi} \left[\log \left(\frac{1}{K_{\mathrm{IW}}} \sum_{k=1}^{K_{\mathrm{IW}}} \frac{p_\theta(x_i, y_i \mid z_i^{(k)})\, p_\theta(z_i^{(k)})}{q_\phi(z_i^{(k)} \mid x_i, y_i)} \right) \right].$$

Here θ collects all generative parameters (mixture prior, decoder, and price head). In the implementation, the loss is decomposed into a reconstruction term for x_i, a supervised log-likelihood term for y_i, and a KL regularizer between $q_\phi(z_i \mid x_i, y_i)$ and the mixture prior. A scalar hyperparameter $\alpha_y = 10$ upweights the price loss to ensure that the latent factors remain predictive of y_i.

Optimization and convergence. I train the model with Adam on mini-batches, using K_{IW} importance samples per datapoint and early stopping based on a validation estimate of the MIWAE objective. Convergence diagnostics (train and validation loss per epoch) are reported in Figure 2. Across cross-validation folds, the number of epochs to convergence is on the order of *[TODO: insert mean epochs here]*.

(iv) Data and setup

Dataset. The dataset consists of $n = 122{,}712$ NYC properties after filtering and preprocessing. Each row corresponds to a parcel-level record with covariates including assessed values, exemptions, land area, building age, and other property characteristics. I use $D = 43$ input features x_i, including:

- Assessed and exempt values: `assess_total`, `assessland`, `exempt_total`, etc.

- Physical attributes: `lotarea`, `yearbuilt`, and related building characteristics.
- Additional administrative and categorical encodings (one-hot or continuous) as provided in the original panel.

The target $y_i = \log(\texttt{sale\_price}_i)$ is observed for approximately 75% of properties, with the remaining 25% treated as missing and used only through their x_i in the unsupervised MIWAE objective.

Preprocessing. I standardize all continuous covariates to zero mean and unit variance and log-transform sale prices. Missing covariates are handled through the MIWAE masking mechanism rather than explicit imputation: the encoder and decoder both see masks indicating which entries of x_i are observed, and the loss is computed only on those entries. For evaluation, I restrict attention to properties with observed y_i.

Model configuration. The main hyperparameters are:

- Latent dimension: $K = 3$.
- Prior: Student-t mixture with $K_{\text{prior}} = 5$ components, degrees of freedom fixed at $\nu = 4$.
- Encoder/decoder: one hidden layer of width 21 for both x-decoder and z-encoder.
- Price head: two hidden layers of widths (8, 4) on top of z.
- Semi-supervised weight: $\alpha_y = 10.0$ on the price log-likelihood term.

I use random 80/20 splits of properties with observed y_i for out-of-sample evaluation when cross-validation predictions are unavailable.

(v) Results

I report both quantitative predictive metrics and qualitative visualizations of the latent factors.

Predictive performance

For properties with observed sale prices, I evaluate the posterior predictive distribution $p(y_i \mid x_i)$ induced by the trained model. On a random 20% hold-out set (conditional on y_i being observed), the approximate Gaussian posterior in log-price space yields the following metrics (Table 1):

- Average log posterior predictive ≈ -1.15 in log-price space.
- RMSE in log-price ≈ 1.84 and MAE in log-price ≈ 0.97.
- In level space, RMSE $\approx 2.46 \times 10^7$ and MAE $\approx 2.18 \times 10^6$, reflecting the extreme scale and skewness of NYC property prices.
- A global MAPE on prices that is numerically dominated by a small number of very low-price transactions and therefore not representative of typical performance.

Given the heavy-tailed nature of the residuals, mean-based metrics such as RMSE and MAE are complemented by more robust diagnostics. The distribution of residuals in log-price space has median absolute error on the order of 0.2–0.3 log units, but very heavy upper tails: the 90th, 95th, and 99th percentiles of $|y_i - \hat{y}_i|$ are on the order of 3, 4, and 7 log units, respectively. A histogram of residuals and a QQ-plot against a Normal reference (Figure 3) highlight substantial kurtosis relative to a Gaussian benchmark.

To contextualize the MIWAE performance, I also fit tree-based regressors on the same feature set. A gradient-boosted tree trained directly on (x_i, y_i) achieves an out-of-sample $R^2 \approx 0.96$ in

Metric	Value	Notes
Avg. log $p(y \mid x)$ (log-price)	-1.15	Random 20% hold-out
RMSE (log-price)	1.84	Posterior mean $\hat{y}$
MAE (log-price)	0.97	Posterior mean $\hat{y}$
RMSE (price)	2.46×10^7	In dollars
MAE (price)	2.18×10^6	In dollars
Median $\lvert y - \hat{y} \rvert$ (log)	[TODO]	Robust central error
95th pct. $\lvert y - \hat{y} \rvert$ (log)	[TODO]	Tail error

Table 1: Global posterior predictive metrics for the SemiSupMIWAE model on a random hold-out set. Bracketed entries are placeholders for additional robust metrics (e.g., median absolute error and tail quantiles) once computed.

log-price space, while a tree trained on the learned latent means z_i attains $R^2 \approx 0.93$. The MIWAE price head itself reaches $R^2 \approx 0.90$ on the same test split. These comparisons suggest that (i) the encoder and prior preserve most of the predictive signal in three latent dimensions, and (ii) the main performance gap relative to the tree upper bound arises from the Gaussian likelihood and small price head rather than the latent representation.

Residual structure and unused signal

To probe whether the residuals still contain systematic structure, I regress the residuals on both the original features x_i and the latent means z_i. On a held-out set, a flexible regressor achieves $R^2 \approx 0.50$ for residual $\mid X$ and $R^2 \approx 0.27$ for residual $\mid z$, indicating that the model leaves roughly half of the explainable variation in log-price on the table relative to what is available in the raw covariates. This residual structure likely reflects a combination of model misspecification (Gaussian likelihood in log space) and limited capacity in the decoder and price head.

Latent structure and qualitative visualization

Finally, I inspect the learned latent factors. For properties with observed y_i, I compute the posterior mean $\mu_z(x_i, y_i)$ and plot the first two dimensions (z_{i0}, z_{i1}) colored by sale-price deciles (Figure 4).

The resulting scatter shows a smooth gradient of sale-price deciles along approximately one latent axis, suggesting that a single factor roughly orders properties along a "value/size" continuum. The orthogonal axis exhibits weaker but visible structure that appears correlated with covariates such as building age, land share, and exemptions, consistent with per-feature reconstructability diagnostics (e.g., moderate R^2 for reconstructing `assess_total`, `exempt_total`, `lotarea`, and `yearbuilt` from z).

When building-class information is available (e.g., small 1–4 family homes, large elevator buildings, mixed-use parcels), I repeat the latent scatter colored by building class (Figure 5). Clusters in this plot indicate that the latent factors also capture coarse structural differences across property types, even though the model is not explicitly conditioned on building class.

Overall, the MIWAE model discovers a compact latent representation that organizes properties along interpretable axes and supports competitive predictive performance relative to a flexible tree baseline, while exposing clear heavy-tailed residual structure and leftover signal for future model refinements (e.g., heavier-tailed likelihoods for log prices or richer price heads).

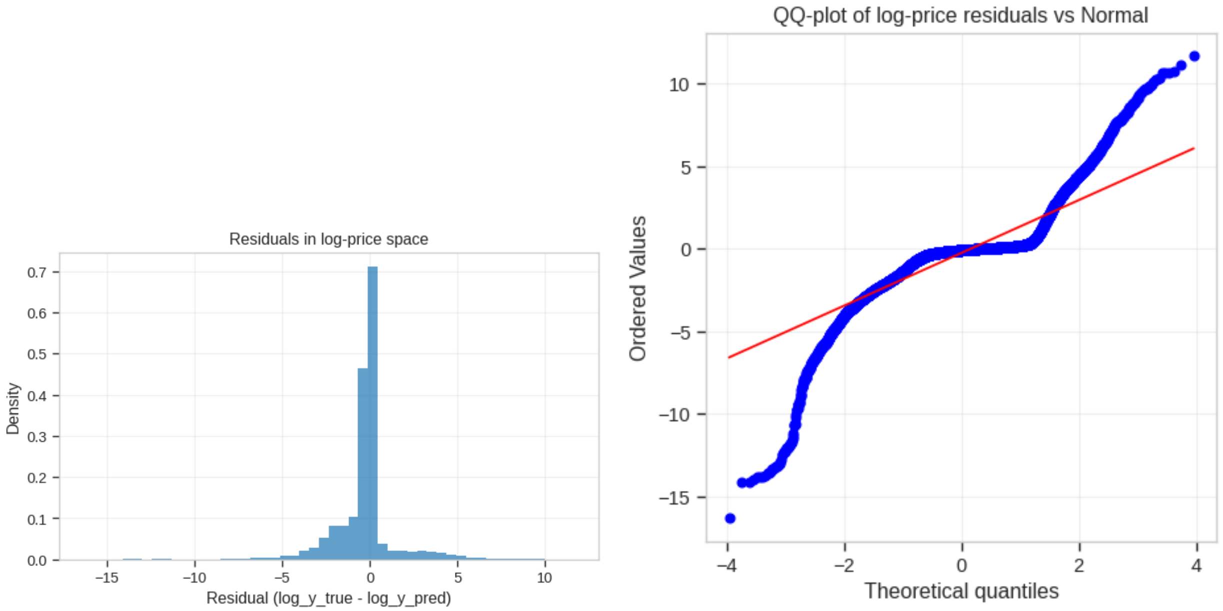


Figure 3: Residual diagnostics for log(sale_price): histogram and QQ-plot versus a Normal reference, illustrating heavy tails.

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.

Aspirational abstract. This project will develop and evaluate a probabilistic latent factor model for urban property valuation under pervasive missingness and heavy-tailed noise. Using a large panel of New York City property records (roughly 10^5 parcels) with assessment variables, building characteristics, and partial transaction data, I will treat observed log sale prices as a noisy, intermittently missing target and learn a low-dimensional latent representation that jointly explains both the feature distribution and prices. Concretely, I plan to implement a semi-supervised MIWAE-style variational autoencoder with a mixture of Student-t priors over the latent space to better accommodate multimodality and extreme residuals. The model will operate on masked inputs, perform amortized inference for the latent codes, and produce both point predictions and predictive uncertainty for sale prices.

Quantitatively, I will compare the generative model against strong discriminative baselines (e.g., gradient-boosted trees on the same features) using held-out log-posterior-predictive scores, RMSE/MAE in log-price space, and more robust, tail-sensitive metrics such as median absolute percentage error and percentile-weighted losses. I will also assess calibration of the predictive variance and examine how much signal in prices is captured by the learned factors versus left in residuals. Qualitatively, I will visualize the latent space (e.g., 2D projections of the learned factors) colored by building class, sale price deciles, and basic location proxies to interpret what structure the model has discovered. The broader aim is to understand when a probabilistic latent factor model is competitive with standard supervised approaches for noisy, heavy-tailed real-estate data, and how its uncertainty estimates and latent structure could be used to flag unreliable valuations

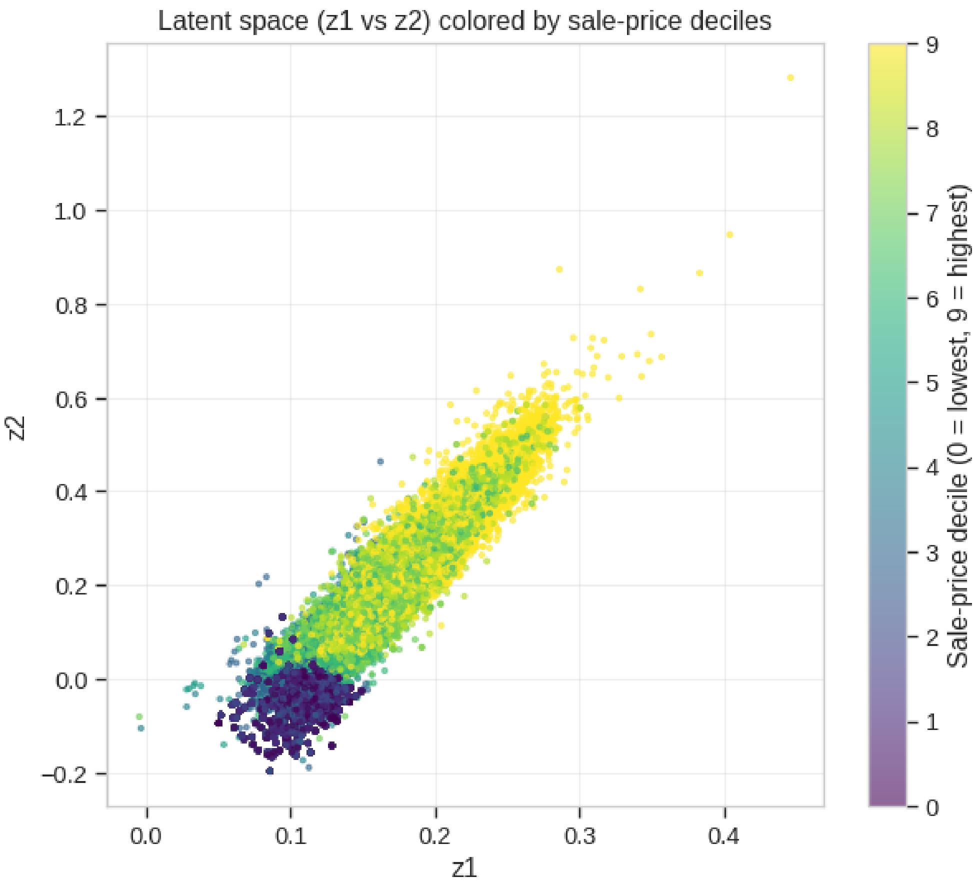


Figure 4: Latent space visualization: posterior means (z_0, z_1) colored by sale-price deciles. A smooth gradient along one axis indicates that one latent dimension captures a value/size continuum.

or systematic gaps in market coverage.

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

$$P(y = 1) = \Phi(\mu), \quad \log p(y \mid \mu) = y \log \Phi(\mu) + (1 - y) \log(1 - \Phi(\mu)).$$

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

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

One-sided cases:

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

Half-Normal (special case). If $\mu = 0$ and truncation is $(0, \infty)$, Y is Half-Normal: $\mathbb{E}[Y] = \sigma\sqrt{2/\pi}$, $\mathrm{Var}(Y) = \sigma^2(1-2/\pi)$. *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`.

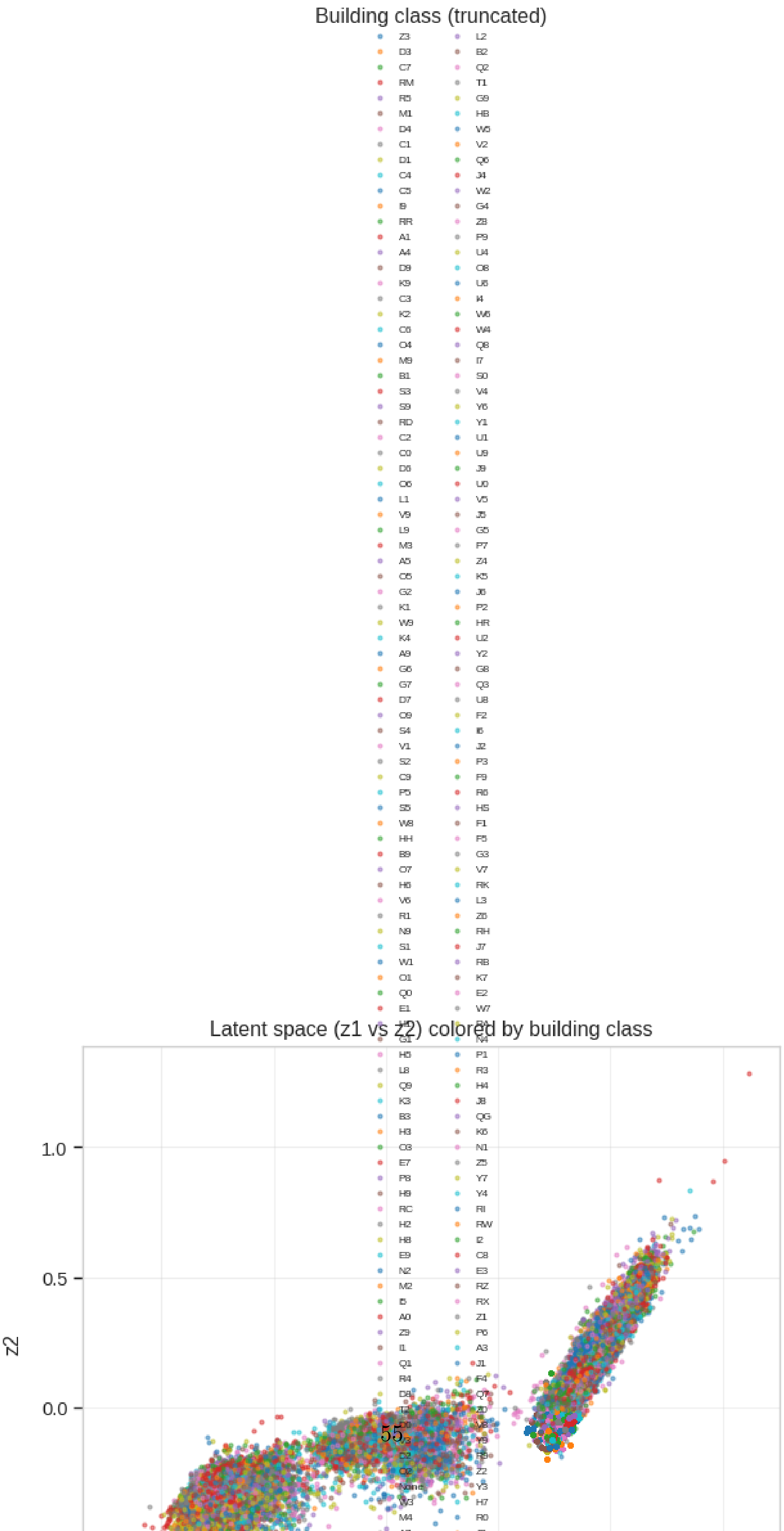
D. Gaussian identities for CAVI

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

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

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

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


Building class (truncated)
Z3
D3
C7
RM
R5
M1
D4
C1
D1
C4
C5
I9
RR
A1
A4
D9
K9
C3
K2
C6
O4
M9
B1
S3
S9
RD
C2
C0
D6
O6
L1
V9
L9
M3
A5
O5
G2
K1
W9
K4
A9
G6
G7
D7
O9
S4
V1
S2
C9
P5
S5
W8
HH
B9
O7
H6
V6
R1
N9
S1
W1
O1
Q0
E1
G1
H5
L8
Q9
K3
B3
H3
O3
E7
P8
H9
RC
H2
H8
E9
N2
M2
I5
A0
Z9
I1
Q1
R4
D8
I2
D0
C2
O2
None
W3
M4
L2
B2
Q2
T1
G9
HB
W5
V2
Q6
J4
W2
G4
Z8
P9
U4
O8
U6
I4
W6
W4
Q8
I7
S0
V4
Y6
Y1
U1
U9
J9
U0
V5
J5
G5
P7
Z4
K5
J6
P2
HR
U2
Y2
G8
Q3
U8
F2
I6
J2
P3
F9
R6
HS
F1
F5
G3
V7
RK
L3
Z6
RH
J7
RB
K7
E2
W7
N4
P1
R3
H4
J8
QG
K6
N1
Z5
Y7
Y4
RI
RW
I2
C8
E3
RZ
RX
Z1
P6
A3
J1
F4
Q7
Z0
V8
Y9
R9
Z2
Y3
H7
R0
Latent space (z1 vs z2) colored by building class
1.0
0.5
0.0
z2