Exploring a Non-Parametric Bayesian Perspective on Compositional Contextual Inference in Motor Learning

Daniel Kornai

January 16, 2025

1 Formulating The Problem

Visual stimulus will impact motor learning if the learner believes that visual and motor signals share a common latent cause. This common latent cause [sometimes called *context*] may be used to index visuo-motor experiences in a way that permits compact and generalisable representations.

In a series of trials of a motor learning experiment, we observe a D dimensional force vector $\mathbf{f}$, and a T dimensional binary visual cue $\mathbf{x}$ a total of N times. If we believe these observations were caused by a latent variable that simultaneously controls both types of sensory experience, then the $N \times D$ matrix $\mathbf{F}$ and the $N \times T$ matrix $\mathbf{X}$ can be factored into a pair of basis sets [$\mathbf{A}$ and $\mathbf{Y}$ for forces and visuals respectively] and a shared binary feature ownership matrix $\mathbf{Z}$. That is, the i-th row of $\mathbf{A}$ and $\mathbf{Y}$ together form a prototypical visuo-motor sensory context, which we believe is contributing to our observations in the n-th trial if $\mathbf{Z}_{n,i} = 1$. Additional expressive richness [and compressive capability] is conferred on such a factorisation model if multiple contexts can be *composed* with each other.

In this setup we are essentially searching for the solutions to the matrix equations $\mathbf{X} \approx \mathbf{ZY}$ and $\mathbf{F} \approx \mathbf{ZA}$, which nicely corresponds to finding the posterior distributions of $\mathbf{Z}$, $\mathbf{Y}$, and $\mathbf{A}$ given our data $\mathbf{X}$, $\mathbf{F}$, and contraints given by the priors on our unknowns. From Bayes rule,

$$P(\mathbf{Z}, \mathbf{Y}, \mathbf{A}|\mathbf{X}, \mathbf{F}) \propto P(\mathbf{Z})P(\mathbf{X}|\mathbf{Z}, \mathbf{Y})P(\mathbf{Y})p(\mathbf{F}|\mathbf{Z}, \mathbf{A})p(\mathbf{A}) \tag{1}$$

where we assumed the independence structure shown in the graphical model of Fig. 1, which we will detail later. The formulation of this problem as Bayesian inference expresses our desire to find representations that generates observations with high probability [as captured by the likelihoods], and finding a representations that are in general probable, as expressed by the priors.
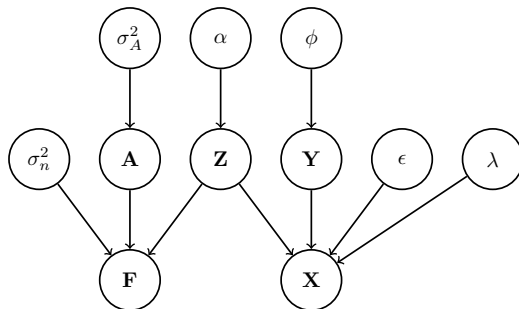

Figure 1: Graphical model illustrating dependencies among variables in the generative model.

Our prior expectation is that rows of $\mathbf{A}$ are drawn from $\mathcal{N}(0, \sigma_A^2 \mathbf{I})$. Forces are assumed to be observed under a linear Gaussian model where the hyperparameter σ_n^2 controls the magnitude of observation noise.

$$f_{n,:} = \mathbf{x}_{n,:}\mathbf{A} + \epsilon \sim \mathcal{N}(0, \sigma_n^2, \mathbf{I})$$

Each element of $\mathbf{Y}$ is drawn independently as $y_{i,k} \sim \text{Bernoulli}(\phi)$. The entries of $\mathbf{X}$ are conditionally independent given the latents, and are assumed to be generated from a *noisy-OR* distribution

$$P(x_{i,t} = 1|\mathbf{Z}, \mathbf{Y}) = 1 - (1 - \lambda)^{\mathbf{z}_{i,:}\mathbf{y}_{:,t}}(1 - \epsilon) \tag{2}$$

where $\mathbf{z}_{i,:}\mathbf{y}_{:,t} = \sum_{k=1}^{K} \mathbf{z}_{i,k}\mathbf{y}_{k,t}$ is the number of latent causes that have the t-th pixel as 1. ϵ is the baseline probability that $x_{i,t} = 1$, and λ is the probability with which any of the hidden causes is effective in eliciting a visual signal.

While the number of rows for $\mathbf{Z}$ is fixed by the number of observations, the number of distinct visuo-motor contexts (K) required to effectively represent the set of experiences during the trial depend on the statistics of the data and the beliefs of the observer. To account for this, our prior distribution on $\mathbf{Z}$ utilises the Indian Buffet Process [IBP]. This allows us to sample from a distribution of sparse binary matrices with a fixed number of rows and a potentially infinite number of columns. The hyperparameter α controls our propensity to add new features (columns of $\mathbf{Z}$) during the learning process.

Given a dataset, we utilise a Gibbs sampler to collect samples from $\mathbf{Z}$, $\mathbf{Y}$, and $\mathbf{A}$. Details of this sampling procedure, which iteratively grows the number of latent features, are provided below:

2 Algorithm for Gibbs Sampler

An outline of the sampling procedure is:

Algorithm 1 Outline for the Gibbs sampler $P(\mathbf{Z}, \mathbf{Y}, \mathbf{A} | \mathbf{X}, \mathbf{F})$

Require: Force observations $\mathbf{F}$, Visual observations $\mathbf{X}$, initial number of features K.

1: initialise $\mathbf{Z}$, $\mathbf{A}$ and $\mathbf{Y}$.

2: **while** gathering posterior samples **do**

3:   resample $\mathbf{A}$

4:   resample $\mathbf{Y}$

5:   **for** $i = 1$ to N **do**

6:     resample $\mathbf{Z}$

7:     sample the number of new features K_i^{new}

8:     **if** $K_i^{new} > 0$ **then**
9:       expand $\mathbf{Z}$, $\mathbf{A}$, $\mathbf{Y}$
10:     **end if**

11:   **end for**
12: **end while**

The complete algorithm is

Algorithm 2 Gibbs sampler for $P(\mathbf{Z}, \mathbf{Y}, \mathbf{A} | \mathbf{X}, \mathbf{F})$

Require: Force observations $\mathbf{F}$, Visual observations $\mathbf{X}$, initial number of features K.

1: $\mathbf{Z} \leftarrow P(\mathbf{Z})$ {Initialise $N \times K$ feature ownership matrix $\mathbf{Z}$ from the prior}
2: $\mathbf{A} \leftarrow P(\mathbf{A})$ {Initialise $K \times D$ force basis matrix $\mathbf{A}$ from the prior}
3: $\mathbf{Y} \leftarrow P(\mathbf{Y})$ {Initialise $K \times T$ image filter basis matrix $\mathbf{Y}$ from the prior}

4: **while** gathering posterior samples **do**

5:   $\mathbf{A} \leftarrow p(\mathbf{A} | \mathbf{F}, \mathbf{Z})$ {Resample $\mathbf{A}$ according to Eq. 3}

6:   **for all** $y_{k,t} \in \mathbf{Y}$ **do**
7:     $y_{k,t} \leftarrow P(y_{kt} | \mathbf{Z}, \mathbf{X}, \mathbf{Y}_{-k,t})$ {Resample each element of $\mathbf{Y}$ according to Eq. 4}
8:   **end for**

9:   **for** $i = 1$ to N **do**

10:     **for** $k = 1$ to K **do**
11:       $m_{-i,k} \leftarrow \sum_{j \neq k}^{N} z_{j,k}$ {Count objects that own feature k without considering object i}
12:       **if** $m_{-i,k} > 0$ **then**
13:         $z_{i,k} \leftarrow P(z_{i,k} | \mathbf{f}_{i,:}, \mathbf{x}_{i,:}, \mathbf{A}, \mathbf{Y}, \mathbf{Z}_{-i,k})$ {Resample $z_{i,k}$ according to Eq. 7}
14:       **else**
15:         mark $z_{i,k}$ to be zeroed
16:       **end if**
17:     **end for**
18:     set all marked $z_{i,k}$ to 0

19:     $K_i^{new} \leftarrow P(K_i^{new} | \mathbf{x}_{i,:}, \mathbf{f}_{i,:}, \mathbf{z}_{i,:}, \mathbf{Y}, \mathbf{A})$ {Sample K_i^{new} according to Eq. 11}

20:     **if** $K_i^{new} > 0$ **then**
21:       $\mathbf{Z}_{new} \leftarrow Z_{new}[K_i^{new}, i]$ {Get K_i^{new} new columns for $\mathbf{Z}$, each with a 1 at row i}
22:       $\mathbf{A}_{new} \leftarrow p(\mathbf{A}_{new} | \mathbf{F}, \mathbf{Z}_{new}, \mathbf{Z}, \mathbf{A})$ {Get K_i^{new} new rows for $\mathbf{A}$, sampled from Eq. 15}
23:       $\mathbf{Z} \leftarrow [\mathbf{Z}, \mathbf{Z}_{new}]$ {Expand $\mathbf{Z}$ with the new columns}
24:       $\mathbf{A} \leftarrow [\mathbf{A}; \mathbf{A}_{new}]$ {Expand $\mathbf{A}$ with the new rows}
25:       $\mathbf{Y}_{new} \leftarrow 0_{k_i^{new} \times T}$ {Get K_i^{new} new empty rows for $\mathbf{Y}$}
26:       $\mathbf{Y} \leftarrow [\mathbf{Y}; \mathbf{Y}_{new}]$ {Expand $\mathbf{Y}$ with the new rows}
27:       **for all** $y_{k,t} \in \mathbf{Y}_{K+1:K+Knew_i,:}$ **do**
28:         $y_{k,t} \leftarrow P(y_{kt} | \mathbf{Z}, \mathbf{X}, \mathbf{Y}_{-k,t})$ {Resample the new parts of $\mathbf{Y}$ according to Eq. 4}
29:       **end for**
30:       $K \leftarrow K + K_i^{new}$
31:     **end if**

32:   **end for**
33: **end while**

---

3 Posterior Terms Used During Sampling

3.1 Posterior $p(\mathbf{A}|\mathbf{F}, \mathbf{Z})$

Given the observed forces $\mathbf{F}$, and the binary feature ownership matrix $\mathbf{Z}$, the posterior $\mathbf{A}$ is sampled from from the following Gaussian

$$p(\mathbf{A}|\mathbf{F}, \mathbf{Z}) \sim \mathcal{N}\left(\left(\mathbf{Z}^T\mathbf{Z} + \frac{\sigma_n^2}{\sigma_A^2}\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{F}, \sigma_n^2\left(\mathbf{Z}^T\mathbf{Z} + \frac{\sigma_n^2}{\sigma_A^2}\mathbf{I}\right)^{-1}\right) \tag{3}$$

3.2 Posterior $P(y_{k,t}|\mathbf{Z}, \mathbf{X}, \mathbf{Y}_{-k,t})$

When performing updates to $\mathbf{y}_{:,t}$, the t-th column of the image filter basis matrix $\mathbf{Y}$, this impacts the t-th column of the reconstructed images. Accordingly, the posterior from proportional is

$$\begin{aligned} &P(y_{k,t} = a|\mathbf{Z}, \mathbf{X}, \mathbf{Y}_{-k,t}) \\ &\propto P(y_{k,t} = a)P(\mathbf{x}_{:,t}|\mathbf{Z}, \mathbf{y}_{:,t})|_{y_{k,t}=a} \end{aligned} \tag{4}$$

The prior is defined using the hyperparameter ϕ as

$$P(y_{k,t} = a) = \phi^a(1-\phi)^{1-a} \tag{5}$$

And the likelihood is:

$$P(\mathbf{x}_{:,t}|\mathbf{Z}, \mathbf{y}_{:,t}) = \prod_{i=1}^{N}[1 - (1-\lambda)^{\mathbf{z}_{i,:}\mathbf{y}_{:,t}}(1-\epsilon)]^{x_{i,t}} \times [(1-\lambda)^{\mathbf{z}_{i,:}\mathbf{y}_{:,t}}(1-\epsilon)]^{1-x_{i,t}} \tag{6}$$

$|_{y_{k,t}=a}$ denotes carrying out the calculation with $y_{k,t}$ temporarily set to a.

3.3 Posterior $P(z_{i,k}|\mathbf{f}_{i,:}, \mathbf{x}_{i,:}, \mathbf{A}, \mathbf{Y}, \mathbf{Z}_{-i,k})$

As our force and image observations are conditionally independent given $\mathbf{Z}$, our posterior proportional is

$$\begin{aligned} &P(z_{i,k} = a|\mathbf{f}_{i,:}, \mathbf{x}_{i,:}, \mathbf{A}, \mathbf{Y}, \mathbf{Z}_{-i,k}) \\ &\propto P(z_{i,k} = a|\mathbf{Z}_{-i,k})P(\mathbf{x}_{i,:}|\mathbf{Y}, \mathbf{z}_{i,:})p(\mathbf{f}_{i,:}|\mathbf{A}, \mathbf{z}_{i,:})|_{z_{i,k}=a} \end{aligned} \tag{7}$$

The prior $P(z_{i,k} = a|\mathbf{Z}_{-i,k})$ is defined with $m_{-i,k} = \sum_{j \neq i}^{N} z_{j,k}$ as

$$P(z_{i,k} = a|\mathbf{Z}_{-i,k}) = \left(\frac{m_{-i,k}}{N}\right)^a\left(1 - \frac{m_{-i,k}}{N}\right)^{1-a} \tag{8}$$

The likelihood term for the visual observation is $P(\mathbf{x}_{i,:}|\mathbf{Y}, \mathbf{z}_{i,:})$ is

$$P(\mathbf{x}_{i,:}|\mathbf{Y}, \mathbf{z}_{i,:}) = \prod_{t=1}^{T}[1 - (1-\lambda)^{\mathbf{z}_{i,:}\mathbf{y}_{:,t}}(1-\epsilon)]^{x_{i,t}} \times [(1-\lambda)^{\mathbf{z}_{i,:}\mathbf{y}_{:,t}}(1-\epsilon)]^{1-x_{i,t}} \tag{9}$$

and the likelihood term for the force observation is $p(\mathbf{f}_{i,:}|\mathbf{A}, \mathbf{z}_{i,:})$ is simply

$$p(\mathbf{f}_{i,:}|\mathbf{A}, \mathbf{z}_{i,:}) = \frac{1}{(2\pi\sigma_n^2)^{D/2}}\exp\left(-\frac{1}{2\sigma_n^2}(\mathbf{f}_{i,:} - \mathbf{z}_{i,:}\mathbf{A})^T(\mathbf{f}_{i,:} - \mathbf{z}_{i,:}\mathbf{A})\right) \tag{10}$$

As previously, The symbol $|_{z_{i,k}=a}$ denotes carrying out the calculation with $z_{i,k}$ temporarily set to a.

3.4 Posterior $P(K_i^{new}|\mathbf{x}_{i,:}, \mathbf{f}_{i,:}, \mathbf{z}_{i,:}, \mathbf{Y}, \mathbf{A})$

During the process of sampling, it might be necessary to increase the number of latent features. As new features is initially only active for the i-th row of $\mathbf{z}$

$$\begin{aligned} &P(K_i^{new}|\mathbf{x}_{i,:}, \mathbf{f}_{i,:}, \mathbf{z}_{i,:}, \mathbf{Y}, \mathbf{A}) \\ &\propto P(k_i^{new})P(\mathbf{x}_{i,:}|\mathbf{z}_{i,:}, \mathbf{Y}, K_i^{new})p(\mathbf{f}_{i,:}|\mathbf{z}_{i,:}, \mathbf{A}, K_i^{new}) \end{aligned} \tag{11}$$

The prior $P(k_i^{new})$ is simply a Poisson:

$$P(k_i^{new}) = \text{Pois}\left(\frac{\alpha}{N}\right) \tag{12}$$

The likelihood term for the visual observations given K_i^{new} is

$$\begin{aligned} &P(\mathbf{x}_{i,:}|\mathbf{z}_{i,:}, \mathbf{Y}, K_i^{new}) \\ &= \prod_{t=1}^{T}\left[1 - (1-\epsilon)(1-\lambda)^{\mathbf{z}_{i,:}\mathbf{y}_{:,t}}(1-\lambda\phi)^{K_i^{new}}\right]^{x_{i,t}} \times \left[(1-\epsilon)(1-\lambda)^{\mathbf{z}_{i,:}\mathbf{y}_{:,t}}(1-\lambda\phi)^{K_i^{new}}\right]^{1-x_{i,t}} \end{aligned} \tag{13}$$

The likelihood for the force observations given K_i^{new} is

$$\begin{aligned} p(\mathbf{f}_{i,:}|\mathbf{z}_{i,:}, \mathbf{A}, K_i^{new}) &= \int p(\mathbf{f}_{i,:}|\mathbf{z}_{i,:}, \mathbf{A}, \mathbf{A}_{new}, K_i^{new})p(\mathbf{A}_{new})\mathrm{d}\mathbf{A}_{new} \\ &= \frac{1}{[2\pi(\sigma_n^2 + K_i^{new}\sigma_a^2)]^{D/2}}\exp\left(-\frac{1}{2(\sigma_n^2 + K_i^{new}\sigma_a^2)}(\mathbf{f}_{i,:} - \mathbf{z}_{i,:}\mathbf{A})^T(\mathbf{f}_{i,:} - \mathbf{z}_{i,:}\mathbf{A})\right) \end{aligned} \tag{14}$$

Note how this reduces to the basic likelihood if $K_i^{new} = 0$.

3.5 Posterior $p(\mathbf{A}_{new}|\mathbf{F}, \mathbf{Z}_{new}, \mathbf{Z}, \mathbf{A})$

When k_{new} features were added to $\mathbf{Z}$ in the form of $\mathbf{Z}_{new}$, $\mathbf{A}$ must also be expanded with k_{new} rows. These rows are drawn from the following Gaussian

$$
\begin{aligned}
&p(\mathbf{A}_{new}|\mathbf{F}, \mathbf{Z}_{new}, \mathbf{Z}, \mathbf{A}) \\
&\propto p(A_{new})p(\mathbf{F}|\mathbf{Z}_{new}, \mathbf{Z}, \mathbf{A}, \mathbf{A}_{new}) \\
&\sim \mathcal{N}\left(\left(1_{k_{new} \times k_{new}} + \frac{\sigma_n^2}{\sigma_A^2}\mathbf{I}\right)^{-1}\mathbf{Z}_{new}^T(\mathbf{F} - \mathbf{Z}\mathbf{A}), \sigma_n^2\left(1_{k_{new} \times k_{new}} + \frac{\sigma_n^2}{\sigma_A^2}\mathbf{I}\right)^{-1}\right)
\end{aligned}
\tag{15}
$$