# Appendix A: Nonparametric Bayesian Chunk Learner

## 1 Introduction

Learning is about discovering structure in the input and store information related to some subset of this structure that can be used later for altered processing of new input. While in natural context, such as language acquisition, forming object representations or mastering to perform a task, the input structure and the acquired knowledge can be described through abstract semantic concepts (e. g. words or sentences for language, parts and objects for vision), formal treatment of the learning phenomenon focuses on the canonical problem of extracting and representing various lower- and higher-order correlational structures of atomic pieces of information under various constraints. Models of such representational learning range from simple storage of frequent episodes handled adequately by counting or associative learning models focusing on various levels of element-co-occurrences, to more sophisticated "chunk-learners", where the underlying constraints limiting the learning process can be widely different (Perruchet, 2019, Orbán et al. 2008).

Following a long line of earlier work (Ullman et al., 2002, Christiansen et al., 1998), a recent study investigated the question of which of these alternative learning model classes would correspond to human representational learning of new visual structures (Orbán et al., 2008). In this study, various versions of the simple counting models together with a recursive 2nd-order associative learner (essentially, a restricted Boltzman machine) was compared to a normative Bayesian chunk learner (BCL) in a large set of test conditions that followed, in exquisite precision, the stimuli and training of previously published human visual statistical learning experiments (Fiser and Aslin 2001, 2005). In the experiments, humans were passively exposed to a large set of multi-element scenes, where the elements were simple geometric shapes and they were positioned on a visible underlying grid. The structure of the input was defined by a co-occurrence structure defining an inventory, namely the fact that particular pairs, triplets or larger tuples always co-occurred in a fixed spatial configuration together in the scenes (chunks), and these chunks were always connected to other chunks in a given scene but never overlapped. Since all shapes were the same-sized black forms in the middle of their respective grid cell, this arrangement yielded scenes with multiple chunks such that the chunks were never obviously separated in the scenes. Thus to gain any understanding of the un-

derlying structure of the scenes, observers had to extract some representation beyond storing individual shape information (which was useless, as all shapes appeared equal number of times) or full scenes (which was impossible due to the large number of scenes). Two-alternative forced choice tests provided overwhelming evidence that after 10-20 minutes of observation of such streams of scenes without any instruction, human adults as well as 8 month-old infants automatically extracted the underlying chunk structure of such scenes (Fiser and Aslin 2001, 2002, 2005).

In modeling these human results, the BCL model searched for the high-level building blocks based on combinations of lower-level features (chunks) of the scenes in a principled way: by finding chunks that are minimally sufficient to encode the scenes from the environment. The resulting representation of the environment is then formalized as a distribution over the possible chunks. While this BCL model was a parametric chunking model, non-parametric accounts of chunking have also been proposed in the language and in the visual domains (Goldwater et al., 2009, Austerweil and Griffiths, 2011 and Lee et al., 2020). In the BCL model the number of chunks is inferred during learning. Nonparametric models in contrast handle the number of chunks as unbounded, albeit only a finite number of them is observed. Although in theory both methods allow for discovering any number of chunks, the nonparametric version doesn't need a separate sampling method for inferring the number of latents. Moreover, the IBP model enables to build a Gibbs sampler to estimate the inventory's distribution, which makes the computation more effective and flexible enough to address new questions about chunk learning.

## 2 Non-parametric version of the Bayesian chunk learner

As discussed above, here we expand the BCL model into the nonparametric direction, and use the Indian Buffet Process (IBP) as a prior for the link matrix between features and chunks, similarly to Austerweil and Griffiths (2011). This choice is a natural one as in the learning process one feature can belong to more than one chunk, which is enabled in the IBP model. In another attempt of a nonparametric chunk learner model, Lee et al. (2020) use the Chinese Restaurant Process as a prior on the feature-chunk link matrix. In this case however, the problem is formulated as cluster learning, where one feature can only belong to one chunk.

Our model also exploits the spatial invariance of learning, i.e., it marginalizes out the positions of the chunks, making the calculation much more efficient. Hence, it fits to earlier data, presented in Orbán et al. (2008) faster.

More specifically, the BCL model solves the chunking task of shapes presented on a grid. The observed variables (or features) are the shapes' presence or absence in any given trial, and their positions on the grid. The statistical

structure of the shapes is learned through the discovery of a set of latent variables (chunks), which affect the occurrence of the observed variables and their relative positions. As discussed above, we extend this earlier model by setting the number of latent (hidden) variables to be unbounded (as opposed to the earlier approach where the number of latents was sampled from the posterior distribution). We further assume that although the number of latent variables is unbounded, only a finite number of latents have effect on the observed variables. The non-parametric version of the model was implemented using the IBP, as a prior on the link matrix, which defines the latent variable-observed variable connections.

In case of $N$ binary variables (shapes in our experiments) observed in $T$ trials, the observation can be summarized in the $N \times T$ binary matrix, $\mathbf{X}$ with its $x_{nt}$ being 1 if the $n^{\text{th}}$ variable is present in the $t^{\text{th}}$ trial, and 0 otherwise.

In case of $K$ latent variables (chunks) the link matrix is a binary, $N \times K$ matrix, denoted by $\mathbf{Z}$. The matrix element $z_{nk}$ is 1 if the $k^{\text{th}}$ latent variable has an effect on the $n^{\text{th}}$ observed variable. That is, the $k^{\text{th}}$ chunk involves the $n^{\text{th}}$ shape. If not, $z_{nk}$ is 0.

The state of the hidden causes is summarized in the $K \times T$ binary matrix $\mathbf{Y}$. Its $y_{kt}$ element is 1, if the $k^{\text{th}}$ latent is active in trial $t$.

In order to assess the latent variable-observed variable dependency, we need to infer the link matrix $\mathbf{Z}$ given the observation matrix $\mathbf{X}$. In the following we discuss the calculation of this posterior distribution. Based on the generative model depicted in Figure 1A, we first discuss the prior distributions of the random variables describing the latents and then the likelihood of the observed variables given the latents. Then, using these, we discuss how the approximated posterior distribution of the latents can be obtained by constructing samplers (in section *Sampling*).

## 2.1 Prior distributions

### 2.1.1 Latent state-, bias- and position variables: $y_{kt}, r_k, u_{kt}$

Latent variables, indexed by $k$, are described by $y_{kt}$, the binary *state variable* at trial $t$, and $\mathbf{u_{kt}}$, the 2-dimensional position of the chunk.

The state variable $y_{kt}$ is a Bernoulli random variable with parameter $r_k$, where $r_k$ is learned. If $y_{kt} = 1$, then the latent variable is active in trial $t$. The prior distribution of $r_k$ is uniform distribution between 0 and 1.

$$\mathbb{P}(y_{kt}) = \text{Bernoulli}\,(r_k) \tag{1}$$

$$\mathbb{P}(r_k) = \text{Beta}\,(1,1) = \text{Uniform}\,(0,1) \tag{2}$$

The position variable's prior distribution is normal around the center of the grid, marked by $(0,0)$, with variance $\sigma_u^2$. The spatial dimensions can be evaluated independently, therefore here we refer only to one dimension when describing the model, denoted by $u_{kt}$:

$$\mathbb{P}(u_{kt}) = \mathcal{N}\left(0, \sigma_u^2\right). \tag{3}$$

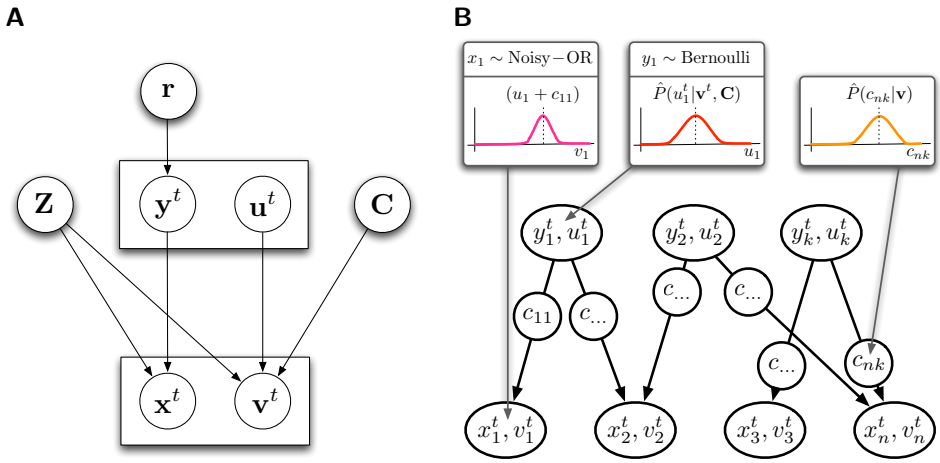


Figure 1: **A**, The generative model. **B**, The graphical model

Note that both dimensions are modeled the same way, using the same descriptive parameters.

#### 2.1.2 Latent link- and relative position matrices: $\mathbf{Z}, \mathbf{C}$

The link matrix $\mathbf{Z}$ has an IBP prior, $\mathbb{P}(\mathbf{Z}) = \text{IBP}(\alpha)$, $\alpha$ being the IBP's model parameter governing the number of latents.

The $N \times K$-sized relative position random variable matrix $\mathbf{C}$ (for only one spatial dimension because of the same reasons we discussed in case of $u_{kt}$) represents the relative position of the observed variables from their (linked) latent variables' geometric center. I.e., $c_{nk}$ is a random variable, representing the $n^{\text{th}}$ observed variable's relative position to the $k^{\text{th}}$ latent variable's geometric center. Its prior is normal distribution around the center of the latent, with $\sigma_C^2$ variance:

$$\mathbb{P}(c_{nk}) = \mathcal{N}(0, \sigma_C^2) \tag{4}$$

### 2.2 Likelihoods of the observed variables given the latents

The observed variables too are characterized by state- and position variables, $x_{nt}$ and $v_{nt}$, respectively (again, $v_{nt}$ only represents one dimension for simplicity). These variables encode the shape's presence or absence and its position in case of being present in trial $t$.

#### 2.2.1 Observed state- and position for one shape and trial: $x_{nt}, v_{nt}$

The likelihood of an observed variable's state in trial $t$ given the latent variables' state and the link matrix $\mathbf{Z}$ is characterized by a noisy-or distribution with baseline and latent effect parameters $\epsilon$ and $\lambda$, respectively.

$$P\left(x_{nt}=1|\mathbf{Z},\mathbf{y}_{:\mathbf{t}}\right)=1-\left(1-\lambda\right)^{\mathbf{z}_{n:}\cdot\mathbf{y}_{:t}}\left(1-\epsilon\right), \tag{5}$$

where $\mathbf{z}_{n:}\cdot\mathbf{y}_{:\mathbf{t}}$ is the number of active latent variables that are linked to the observed variable.

The likelihood of the observed variables' positions in a given trial is given by a mixture of experts distribution, detailed below. The factors in the mixture of experts distribution are the bias position of the observed variables (i.e., the spontaneous occurrence of the observed, without any linked latents being active) and the observed variables' positions relative to their linked latent variables (i.e., the occurrence because of a chunk being present). Note that similarly to the latent variables' positions, we only discuss in detail one of the two dimensions of the observed variables ($v_{nt}$).

$$\begin{aligned}\mathbb{P}(v_{nt}|\mathbf{Z},\mathbf{y}_{:t},\mathbf{u}_{:t},\mathbf{C}) &= \frac{1}{F_n}\mathcal{N}\left(v_{nt};0,\sigma_v^2\right)\prod_{k\in\mathrm{par}(n)}\mathcal{N}\left(v_{nt};u_{kt}+c_{nk},\frac{\sigma_v^2}{\phi}\right)\\ &= \mathcal{N}\left(v_{nt};\Gamma_n\mu_n^v,\sigma_v^2\Gamma_n\right)\end{aligned} \tag{6}$$

where:

$\sigma_v^2$ = model parameter, variance of the observed variables' position along one dimension,
$\mathbf{C}_{N\times K}$ = relative position matrix,
$\phi$ = scale of the variance of the position of a shape within the chunk relative to the prior variance of a shape position,
$F_n$ = normalization factor, making $v_{nt}$ normally distributed. The calculation of Gaussian products is calculated in the Appendix,
$\mathrm{par}(n)$ = set of latent variables that have link to the $n^{\mathrm{th}}$ observed variable (parents of $n$)
$\Gamma_n$ = $\frac{1}{1+\sum_{k\in\mathrm{par}(n)}\phi}$,
$\mu_n^v$ = $\sum_{k\in\mathrm{par}(n)}\phi\left(u_{kt}+c_{nk}\right)$.

### 2.2.2 Observed state- and position variables over all shapes and trials: $\mathcal{D}=\{\mathbf{X},\mathbf{V}\}$

Given the latent variables, the observed variables are conditionally independent, therefore the joint probabilities can be calculated as the products of the conditional probabilities:

$$\mathbb{P}(\mathbf{x}_{:t}|\mathbf{Z},\mathbf{y}_{:t})=\prod_{n=1}^{N}\mathbb{P}(x_{nt}|\mathbf{Z},\mathbf{y}_{:t}) \tag{7}$$

$$\mathbb{P}(\mathbf{v}_{:t}|\mathbf{Z},\mathbf{y}_{:t},\mathbf{u}_{:t},\mathbf{C})=\prod_{n=1}^{N}\mathbb{P}(v_{nt}|\mathbf{Z},\mathbf{y}_{:t},\mathbf{u}_{:t},\mathbf{C}). \tag{8}$$

As the trials are assumed to be independent, the likelihood of the observed data can be written as

$$\mathbb{P}(\mathcal{D}|\mathbf{Z},\mathbf{Y},\mathbf{U},\mathbf{C}) = \prod_{t=1}^{T} \mathbb{P}(\mathbf{x}_{:t}|\mathbf{Z},\mathbf{y}_{:t})\,\mathbb{P}(\mathbf{v}_{:t}|\mathbf{Z},\mathbf{y}_{:t},\mathbf{u}_{:t},\mathbf{C}) \tag{9}$$

## 2.3 Gibbs-sampling

In our implementation the inference and learning are performed by Gibbs-sampling. As an important feature of the implementation, we derive a case where the latent positions ($\mathbf{u}_{:\mathbf{t}}$) and the relative positions ($\mathbf{C}$) are marginalized. Details of the marginalization can be found in the Appendix.

We implement Gibbs-sampling for inferring the latent variables' states $y_k^t$; the latent variables' biases $r_k$ and the link matrix $\mathbf{Z}$. For the calculation details please refer to the Appendix.

### 2.3.1 Gibbs-sampler for the latent states $y_{kt}$

We draw samples from a Bernoulli distribution as follows:

$$\mathbb{P}\left(y_{kt}|\mathbf{X},\mathbf{V},\mathbf{Z},\mathbf{Y}_{-kt},\mathbf{r},\mathcal{M}\right) = \text{Bernoulli}\left(\frac{\pi_1}{\pi_1+\pi_0}\right), \text{with} \tag{10}$$

$$\pi_i = P(y_{kt}=i|\mathbf{X},\mathbf{V},\mathbf{Z},\mathbf{Y}_{-kt},\mathbf{r},\mathcal{M}) \tag{11}$$

$$P(y_{kt}=i|\mathbf{X},\mathbf{V},\mathbf{Z},\mathbf{Y}_{-kt},\mathbf{r},\mathcal{M}) \propto$$
$$\propto \mathbb{P}(\mathbf{x}_{:t}|\mathbf{Z},\mathbf{y}_{-kt},y_{kt}=i,\mathcal{M})\,\mathbb{P}(\mathbf{V}|\mathbf{Z},\mathbf{Y}_{-kt},y_{kt}=i,\mathcal{M})\,r_k^i(1-r_k)^{1-i}, \tag{12}$$

where:

$\mathbf{X},\mathbf{V}$ = observed variables' state and position in all the trials,
$\mathbf{Y}_{-kt}$ = all the latent state variable in the $Y$ matrix, except latent $k$'s state in trial $t$,
$\mathbf{y}_{-kt}$ = $\mathbf{y}_{(:-k)t}$, noted in a simpler format for the sake of readability: vector, all but the $k^{\text{th}}$ latent's state variable in the $t^{\text{th}}$ trial,
$\mathbf{r}$ = all the latent variables' biases,
$\mathcal{M}$ = model defined parameter set of $\{\epsilon,\lambda,\sigma_u,\sigma_v,\sigma_c,\phi\}$.

### 2.3.2 Gibbs-sampler for the latent biases $r_k$

As $r_k$-s are independent when conditioned on the latent states $\mathbf{Y}$, they can be evaluated separately. We draw samples from a Beta distribution as follows - for detailed calculations please refer to the Appendix.

$$\mathbb{P}(r_k|\mathbf{y}_{k:}) \propto \text{Beta}\left(1+\sum_{t=1}^{T} y_{kt}, 1+T-\sum_{t=1}^{T} y_{kt}\right) \tag{13}$$

### 2.3.3 Gibbs-sampler for the link matrix elements $z_{nk}$

Similarly to the sampling of the latent states, we draw samples from a Bernoulli distribution. Here we only expand on the probabilities from which we derive the $\pi_i$ Bernoulli parameters. For detailed calculations please refer to the Appendix.

$$P(z_{nk} = i|\mathbf{X}, \mathbf{V}, \mathbf{Z}_{-nk}, \mathbf{Y}, \mathcal{M}) \propto$$
$$\propto \prod_{t=1}^{T} \mathbb{P}(\mathbf{x}_{:t}|\mathbf{Z}_{-nk}, z_{nk} = i, \mathbf{y}_{:t}, \mathcal{M})\, \mathbb{P}(\mathbf{V}|\mathbf{Z}_{-nk}, z_{nk} = i, \mathbf{Y}, \mathcal{M})\, P(z_{nk} = i|\mathbf{Z}_{-nk}), \tag{14}$$

where:

$\mathbf{Z}_{-nk}$ = all association information in the link matrix $\mathbf{Z}$ except the one between the $n^{\text{th}}$ and the $k^{\text{th}}$ latent,
$\mathbf{Y}$ = all the latents' state variables across all the trials,
$P(z_{nk} = i|\mathbf{Z}_{-nk}) = \frac{\sum_{j=1, j\neq n}^{N} z_{jk}}{N}$, according to the IBP.

Additionally, as the number of latents is unbounded, there can be new links in $\mathbf{Z}$ for new latent variables. The algorithm proposes a set of new links between not yet exploited latent variables and observed variables. The new set of links, $K_n^{\text{new}}$ is sampled from the following distribution, and it is the number of columns in $\mathbf{Z}$ with 1 in row $n$ and 0 everywhere else (then, in the next iteration of the Gibbs sampling, when sampling the existent elements of $\mathbf{Z}$, described above, the values of these elements can be changed).

$$\mathbb{P}(K_n^{\text{new}}|\mathbf{X}, \mathbf{V}, \mathbf{z}_{n(1:K+K_n^{\text{new}})}, \mathbf{Y}, \mathcal{M}) \propto$$
$$\propto \prod_{t=1}^{T} \mathbb{P}\big(x_{nt}|K_n^{\text{new}}, \mathbf{z}_{n(1:K+K_n^{\text{new}})}, \mathbf{y}_{:t}, \mathcal{M}\big)\, \mathbb{P}\big(\mathbf{V}|\mathbf{z}_{n(1:K+K_n^{\text{new}})}, \mathbf{Y}, \mathcal{M}\big)\, \mathbb{P}(K_n^{\text{new}}) \tag{15}$$

where $\mathbb{P}(K_n^{\text{new}}) = \text{Poisson}\,(\alpha/N)$, $\alpha$ being the IBP's model parameter.

# 3 Notations

- Matrices are denoted by bold capital letters ($\mathbf{Z}$) and their elements are denoted by lowercase letters, indexed by the two dimensions of the matrix ($z_{nk}$).
- Vectors are denoted by lowercase bold letters, in case they are rows or columns of a matrix they are indexed by both coordinate, with : signaling all elements in the given row or column ($\mathbf{z}_{:k}$ is the $k^{\text{th}}$ column of $Z$).
- $\mathbb{P}(x)$ denotes the distribution of random variable $x$.
- $P(A)$ denotes the probability of event $A$.

- For the sake of simpler annotations, random variables and their realizations are not distinguished unless it is crucial for the computations, e.g. in case of the Gibbs sampling calculation in the Appendix.

# Appendix

## A.1 Mixture of experts

In Equation (6) we use the the mixture of experts (product of experts) model to obtain the likelihood of the shapes' positions. For more details on the model please refer to Welling(2007). The resulting normal distribution's parameters and normalization factor are coming from the product of Gaussian densities (in this case the product of 1+|par($n$)| Gaussians). For the calculation of the product's parameters we exploit the followings:

$$\mathcal{N}(x; m_1, \sigma_1^2) \cdot \mathcal{N}(x; m_2, \sigma_2^2) = F \cdot \mathcal{N}(x; m_f, \sigma_f^2)$$

$$\begin{aligned} m_f &= \frac{\frac{1}{\sigma_1^2} m_1 + \frac{1}{\sigma_2^2} m_2}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}, \\ \sigma_f &= \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}, \\ F &= \mathcal{N}(m_1; m_2, \sigma_1^2 + \sigma_2^2) \\ &= \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} \exp\left[-\frac{(m_1 - m_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right]. \end{aligned} \tag{16}$$

Analogously, as it will be used in the Marginalization section in case of multivariate Gaussians the product can be written as

$$\mathcal{N}(\mathbf{x}; \mathbf{m}_1, \mathbf{\Sigma}_1) \cdot \mathcal{N}(\mathbf{x}; \mathbf{m}_2, \mathbf{\Sigma}_2) = F \cdot \mathcal{N}(\mathbf{x}; \mathbf{m_f}, \mathbf{\Sigma_f})$$

$$\begin{aligned} \mathbf{m}_f &= (\mathbf{\Sigma}_1^{-1}\mathbf{m}_1 + \mathbf{\Sigma}_2^{-1}\mathbf{m}_2)(\mathbf{\Sigma}_1^{-1} + \mathbf{\Sigma}_2^{-1})^{-1}, \\ \mathbf{\Sigma}_f &= (\mathbf{\Sigma}_1^{-1} + \mathbf{\Sigma}_2^{-1})^{-1}, \\ F &= \mathcal{N}(\mathbf{m}_1; \mathbf{m}_2, \mathbf{\Sigma}_1^2 + \mathbf{\Sigma}_2^2) \\ &= \frac{1}{\sqrt{2\pi \det(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)}} \\ &\quad \cdot \exp\left[-\frac{1}{2}(\mathbf{m}_1 - \mathbf{m}_2)^T(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)^{-1}(\mathbf{m}_1 - \mathbf{m}_2)\right]. \end{aligned} \tag{17}$$

Note that when calculating multidimensional quadratic forms below, in most of the cases we will omit notating transposes for the sake of readability.

## A.2 Calculations for the Gibbs-sampling

### A.2.1 Gibbs-sampler for the latent states

In the calculation of the latent states' posterior in Equation (12) we marginalized the latent variables' positions ($\mathbf{U}$) and the relative positions of latent and observed variables ($\mathbf{C}$). The details on this marginalization can be found in the *Marginalization* section below. Note that as we sample from a Bernoulli distribution described in Equation (10), all the terms that are the same in $\pi_0$ and in $\pi_1$ can be disregarded from the sampling point of view as they don't change the Bernoulli parameter. For this reason it is sufficient to calculate only the below proportionalities (instead of equalities). E.g., in case of (18) we can disregard the term $\mathbb{P}(\underset{\sim}{\mathrm{X}}, \mathbf{V}|\mathbf{Z}, \mathbf{Y}_{-kt}, \mathbf{r}, \mathcal{M})$ because it is the same in case of $y_{kt} = 0$ and in case of $y_{kt} = 1$ as well. Also, please refer to (1) when calculating the conditional probabilities to see the dependencies, e.g. conditioned on $\mathbf{Y}$, $\mathbf{X}$ is independent from $r$, which is exploited in (18).

The steps of the following calculations include 1) applying Bayes rule and exploiting the dependencies 2) marginalizing over $\mathbf{U}$ and $\mathbf{C}$ and separating the terms based on the dependencies of $\mathbf{U}$ and $\mathbf{C}$ 3) disregarding the factors that are the same for both cases of $\pi_i$.

$$
\begin{aligned}
&P(y_{kt} = i|\mathbf{X}, \mathbf{V}, \mathbf{Z}, \mathbf{Y}_{-kt}, \mathbf{r}, \mathcal{M}) = \\
&= \frac{\mathbb{P}(\mathbf{X}, \mathbf{V}|\mathbf{Z}, \mathbf{Y}_{-kt}, y_{kt} = i, \mathbf{r}, \mathcal{M})\, \mathbb{P}(\mathbf{Y}_{-kt}, y_{kt} = i|\mathbf{r})}{\mathbb{P}(\mathbf{X}, \mathbf{V}|\mathbf{Z}, \mathbf{Y}_{-kt}, \mathbf{r}, \mathcal{M})} \propto \\
&\propto \mathbb{P}(\mathbf{X}, \mathbf{V}|\mathbf{Z}, \mathbf{Y}_{-kt}, y_{kt} = i, \mathcal{M})\, \mathbb{P}(\mathbf{Y}_{-kt}, y_{kt} = i|\mathbf{r}) = \qquad (18) \\
&= \left[\int \mathbb{P}(\mathbf{X}, \mathbf{V}|\mathbf{Z}, \mathbf{Y}_{-kt}, y_{kt} = i, \mathbf{U}, \mathbf{C}, \mathcal{M})\, \mathbb{P}(\mathbf{U})\, \mathbb{P}(\mathbf{C})\, d\mathbf{U} d\mathbf{C}\right] \mathbb{P}(\mathbf{Y}_{-kt}, y_{kt} = i|\mathbf{r}) = \\
&= \int \mathbb{P}(\mathbf{x}_{:t}|\mathbf{Z}, \mathbf{y}_{-kt}, y_{kt} = i, \mathcal{M}) \left[\int \mathbb{P}(\mathbf{v}_{:t}|\mathbf{Z}, \mathbf{y}_{-kt}, y_{kt} = i, \mathbf{u}_{:t}, \mathbf{C}, \mathcal{M})\, \mathbb{P}(\mathbf{u}_{:t})\, d\mathbf{u}_{:t}\right] \cdot \\
&\quad \cdot \prod_{\tau \neq t} \left\{ \mathbb{P}(\mathbf{x}_{:\tau}|\mathbf{Z}, \mathbf{y}_{:\tau}, \mathcal{M}) \left[\int \mathbb{P}(\mathbf{v}_{:\tau}|\mathbf{Z}, \mathbf{y}_{:\tau}, \mathbf{u}_{:\tau}, \mathbf{C}, \mathcal{M})\, \mathbb{P}(\mathbf{u}_{:\tau})\, d\mathbf{u}_{:\tau}\right] \right\} \mathbb{P}(\mathbf{C})\, d\mathbf{C} \cdot \\
&\quad \cdot \mathbb{P}(\mathbf{Y}_{-kt}, y_{kt} = i|\mathbf{r}) = \\
&= \mathbb{P}(\mathbf{x}_{:t}|\mathbf{Z}, \mathbf{y}_{-kt}, y_{kt} = i, \mathcal{M}) \prod_{\tau \neq t} \mathbb{P}(\mathbf{x}_{:\tau}|\mathbf{Z}, \mathbf{y}_{:\tau}, \mathcal{M}) \cdot \\
&\quad \cdot \int \left[\int \mathbb{P}(\mathbf{v}_{:t}|\mathbf{Z}, \mathbf{y}_{-kt}, y_{kt} = i, \mathbf{u}_{:t}, \mathbf{C}, \mathcal{M})\, \mathbb{P}(\mathbf{u}_{:t})\, d\mathbf{u}_{:t}\right] \\
&\quad \cdot \left[\prod_{\tau \neq t} \int \mathbb{P}(\mathbf{v}_{:\tau}|\mathbf{Z}, \mathbf{y}_{:\tau}, \mathbf{u}_{:\tau}, \mathbf{C}, \mathcal{M})\, \mathbb{P}(\mathbf{u}_{:\tau})\, d\mathbf{u}_{:\tau}\right] \mathbb{P}(\mathbf{C})\, d\mathbf{C} \\
&\quad \cdot \mathbb{P}(\mathbf{Y}_{-kt}, y_{kt} = i|\mathbf{r}) \propto \\
&\propto \mathbb{P}(\mathbf{x}_{:t}|\mathbf{Z}, \mathbf{y}_{-kt}, y_{kt} = i, \mathcal{M})\, \mathbb{P}(\mathbf{V}|\mathbf{Z}, \mathbf{Y}_{-kt}, y_{kt} = i, \mathcal{M})\, r_k^i (1 - r_k)^{1-i}. \qquad (19)
\end{aligned}
$$

The first term is the observed states' likelihood and it's a noisy-or distribution

described in Equation (5). The second term is the observed positions' likelihood marginalized over the latent positions and relative positions, calculated in section *Marginalization*, Equation (48).The third term comes from the latent states' distribution (conditioned on their bias): Bernoulli distribution in Equation (1).

### A.2.2 Gibbs-sampler for the latent biases

When sampling the latent variables' biases we can sample from a Beta distribution because of the conjugate prior property of the Beta distribution to the Binomial distribution, and because the prior Uniform distribution is a special case of Beta ($\mathrm{Beta}(1,1) = \mathrm{Uniform}(0,1)$).

$$\begin{aligned}
\mathbb{P}(r_k|\mathbf{y}_{k:}) &= \frac{\mathbb{P}(\mathbf{y}_{k:}|r_k)\,\mathbb{P}(r_k)}{\mathbb{P}(\mathbf{y}_{k:})} \propto \mathbb{P}(\mathbf{y}_{k:}|r_k)\,\mathbb{P}(r_k) \\
&= \prod_{t=1}^{T} r_k^{y_{kt}} \left(1 - r_k^{1-y_{kt}}\right) \mathrm{Beta}(1,1) \\
&= r_k^{\sum_t y_{kt}} (1-r_k)^{T-\sum_t y_{kt}} \mathrm{Beta}(1,1) \\
&= \mathrm{Beta}\left(1 + \sum_{t=1}^{T} y_{kt}, 1 + T - \sum_{t=1}^{T} y_{kt}\right) \qquad (20)
\end{aligned}$$

### A.2.3 Gibbs-sampler for the link matrix

When calculating the posterior of the elements of the link matrix $\mathbf{Z}$ our calculation follows the same logic as in case of the latent states. Additionally, we need to sample not only the links between the existing latents and observed variables, but we also need to propose $K_n^{\mathrm{new}}$ new links between observed variables and latents that haven't yet been exploited.

$$\begin{aligned}
&P(z_{nk} = i|\mathbf{X}, \mathbf{V}, \mathbf{Z}_{-nk}, \mathbf{Y}, \mathcal{M}) = \\
&\quad = \left[\mathbb{P}(\mathbf{X}, \mathbf{V}|\mathbf{Z}_{-nk}, z_{nk} = i, \mathbf{Y}, \mathcal{M})\,\mathbb{P}(z_{nk}|\mathbf{Z}_{-nk}, \mathbf{Y}, \mathcal{M})\right] / \mathbb{P}(\mathbf{X}, \mathbf{V}|\mathbf{Z}_{-nk}, \mathbf{Y}, \mathcal{M}) \\
&\quad \propto \mathbb{P}(\mathbf{X}, \mathbf{V}|\mathbf{Z}_{-nk}, z_{nk} = i, \mathbf{Y}, \mathcal{M})\,\mathbb{P}(z_{nk}|\mathbf{Z}_{-nk}) = \\
&\quad = \int \mathbb{P}(\mathbf{X}, \mathbf{V}|\mathbf{Z}_{-nk}, z_{nk} = i, \mathbf{Y}, \mathbf{U}, \mathbf{C}, \mathcal{M})\,\mathbb{P}(\mathbf{U})\,\mathbb{P}(\mathbf{C})\,d\mathbf{U}d\mathbf{C} \cdot \mathbb{P}(z_{nk}|\mathbf{Z}_{-nk}) \\
&\quad = \int \prod_{t=1}^{T} \mathbb{P}(\mathbf{x}_{:t}|\mathbf{Z}_{-nk}, z_{nk} = i, \mathbf{y}_{:t}, \mathcal{M}) \cdot \\
&\qquad \cdot \left[\int \mathbb{P}(\mathbf{v}_{:t}|\mathbf{Z}_{-nk}, z_{nk} = i, \mathbf{y}_{:t}, \mathbf{u}_{:t}, \mathbf{C}, \mathcal{M})\,\mathbb{P}(\mathbf{u}_{:t})\,d\mathbf{u}_{:t}\right] \mathbb{P}(\mathbf{C})\,d\mathbf{C} \cdot \mathbb{P}(z_{nk}|\mathbf{Z}_{-nk}) \\
&\quad = \prod_{t=1}^{T} \mathbb{P}(\mathbf{x}_{:t}|\mathbf{Z}_{-nk}, z_{nk} = i, \mathbf{y}_{:t}, \mathcal{M}) \cdot
\end{aligned}$$

$$
\begin{aligned}
&\cdot \int \prod_{t=1}^{T} \left[ \int \mathbb{P}(\mathbf{v}_{:t}|\mathbf{Z}_{-nk}, z_{nk}=i, \mathbf{y}_{:t}, \mathbf{u}_{:t}, \mathbf{C}, \mathcal{M})\, \mathbb{P}(\mathbf{u}_{:t})\, d\mathbf{u}_{:t} \right] \mathbb{P}(\mathbf{C})\, d\mathbf{C} \cdot \mathbb{P}(z_{nk}|\mathbf{Z}_{-nk}) \\
&= \prod_{t=1}^{T} \mathbb{P}(\mathbf{x}_{:t}|\mathbf{Z}_{-nk}, z_{nk}=i, \mathbf{y}_{:t}, \mathcal{M})\, \mathbb{P}(\mathbf{V}|\mathbf{Z}_{-nk}, z_{nk}=i, \mathbf{Y}, \mathcal{M}) \cdot P(z_{nk}=i|\mathbf{Z}_{-nk}).
\end{aligned} \tag{21}
$$

Similarly to the latent states' sampling in Equation (19), the first term is the observed states' likelihood and it's a noisy-or distribution described in Equation (5). The second term is also the same, the observed positions' likelihood marginalized over the latent positions and relative positions, calculated in section *Marginalization*, Equation (48). The third term is coming from the IBP model and it is proportional to the popularity of the given, $k^{\text{nt}}$ latent.

$$
P(z_{nk}=i|\mathbf{Z}_{-nk}) = \frac{\sum_{j=1, j\neq n}^{N} z_{jk}}{N}
$$

When proposing new links between observed variables and latents we sample from the following distribution.

$$
\begin{aligned}
&\mathbb{P}(K_n^{\text{new}}|\mathbf{X}, \mathbf{V}, \mathbf{z}_{n(1:K+K_n^{\text{new}})}, \mathbf{Y}, \mathcal{M}) = \\
&= \frac{\mathbb{P}\big(\mathbf{X}, \mathbf{V}|K_n^{\text{new}}, \mathbf{z}_{n(1:K+K_n^{\text{new}})}, \mathbf{Y}, \mathcal{M}\big)\, \mathbb{P}\big(K_n^{\text{new}}|\mathbf{z}_{n(1:K)}, \mathbf{Y}, \mathcal{M}\big)}{\mathbb{P}\big(\mathbf{X}, \mathbf{V}|\mathbf{z}_{n(1:K)}, \mathbf{Y}, \mathcal{M}\big)} \\
&\propto \mathbb{P}\big(\mathbf{X}, \mathbf{V}|K_n^{\text{new}}, \mathbf{z}_{n(1:K+K_n^{\text{new}})}, \mathbf{Y}, \mathcal{M}\big)\, \mathbb{P}(K_n^{\text{new}}) \\
&= \int\int \mathbb{P}\big(\mathbf{X}, \mathbf{V}|K_n^{\text{new}}, \mathbf{z}_{n(1:K+K_n^{\text{new}})}, \mathbf{Y}, \mathbf{U}, \mathbf{C}, \mathcal{M}\big)\, \mathbb{P}(\mathbf{U})\, \mathbb{P}(\mathbf{C})\, d\mathbf{U} d\mathbf{C} \cdot \mathbb{P}(K_n^{\text{new}}) \\
&= \left[ \prod_{t=1}^{T} \mathbb{P}\big(\mathbf{x}_{:t}|K_n^{\text{new}}, \mathbf{z}_{n(1:K+K_n^{\text{new}})}, \mathbf{y}_{:t}, \mathcal{M}\big) \right] \mathbb{P}\big(\mathbf{V}|K_n^{\text{new}}, \mathbf{z}_{n(1:K+K_n^{\text{new}})}, \mathbf{Y}, \mathcal{M}\big) \cdot \\
&\quad \cdot \mathbb{P}(K_n^{\text{new}}) \\
&\propto \left[ \prod_{t=1}^{T} \mathbb{P}\big(x_{nt}|K_n^{\text{new}}, \mathbf{z}_{n(1:K+K_n^{\text{new}})}, \mathbf{y}_{:t}, \mathcal{M}\big) \right] \mathbb{P}\big(\mathbf{V}|K_n^{\text{new}}, \mathbf{z}_{n(1:K+K_n^{\text{new}})}, \mathbf{Y}, \mathcal{M}\big) \cdot \\
&\quad \cdot \mathbb{P}(K_n^{\text{new}})\,.
\end{aligned} \tag{22}
$$

Similarly to the sampling of the latent states and the elements of the link matrix above, he first term is the likelihood of the observed states, but as the new links to the $n^{\text{nt}}$ observed don't depend on any other observed state but the $n^{\text{nt}}$, we can disregard the factors that are the same for all $K_n^{\text{new}}$, taking into account only the $n^{\text{nt}}$ observed's states. The second term is also the same as before, calculated in Equation (48). The third term is coming from the IBP:

$$
\mathbb{P}(K_n^{\text{new}}) = \text{Poisson}(\alpha/N),
$$

$\alpha$ being the IBP's model parameter, controlling the number of latent variables.

## A.3 Marginalization

In this section we calculate the likelihood of the observed variables' positions, $\mathbb{P}(\mathbf{V}|\mathbf{Z},\mathbf{Y},\mathcal{M})$ which is needed for the Gibbs sampling in Equations of the latent's sampling (12), (19), and the Equations of the link matrix elements' sampling (14), (21), (15), (22). For this we need to marginalize over $\mathbf{u}_{:t}$ and then over $\mathbf{C}$. For the sake of clarity we will use $\mathcal{S} = \{\mathbf{Z},\mathbf{Y},\mathcal{M}\}$. Note that $\mathcal{M}$ is the model defined parameter set of $\{\epsilon,\lambda,\sigma_u,\sigma_v,\sigma_c,\phi\}$.

For the marginalization over $\mathbf{u}_{:t}$ and $\mathbf{C}$ we will take the following steps.

1. Show that $\mathbb{P}(v_{nt}|\mathbf{u}_{:t},\mathbf{C},\mathcal{S})$ is a Gaussian with $\mathbf{u}_{:t}$ being the dependent variable.
2. Exploiting the Gaussian yielded in the previous point, calculate $\mathbb{P}(\mathbf{v}_{:t}|\mathbf{C},\mathcal{S})$ by marginalizing over $\mathbf{u}_{:t}$.
3. Express $\mathbb{P}(\mathbf{v}_{:t}|\mathbf{C},\mathcal{S})$ as a Gaussian with $\mathbf{C}$ being the dependent variable.
4. From the previous point, express $\mathbb{P}(\mathbf{V}|\mathbf{C},\mathcal{S})$, the likelihood of the whole observed position matrix across trials, as a Gaussian with $\mathbf{C}$ being the dependent variable.
5. Marginalize over $\mathbf{C}$ to have $\mathbb{P}(\mathbf{V}|\mathcal{S})$.

### A.3.1 $\mathbb{P}(v_{nt}|\mathbf{u}_{:t},\mathbf{C},\mathcal{S})$ is a Gaussian with $\mathbf{u}_{:t}$ being the dependent variable

We saw in Equation (6) that

$$\mathbb{P}(v_{nt}|\mathbf{u}_{:t},\mathbf{C},\mathcal{S}) = \mathcal{N}\left(v_{nt};\Gamma_n\mu_n^v,\sigma_v^2\Gamma_n\right). \tag{23}$$

Here we show that with the right $\mu_{\mathbf{n}}^{\mathbf{u}}$ mean, $\mathbf{\Sigma}_{\mathbf{n}}^{\mathbf{u}}$ variance and $G_n$ normalization factor (defined in the box after Equation (26)) the unidimensional Gaussian in (23) can be rewritten as the multidimensional Gaussan:

$$G_n\mathbb{P}(\mathbf{u}_{:\mathbf{t}}|\mu_{\mathbf{n}}^{\mathbf{u}},\mathbf{\Sigma}_{\mathbf{n}}^{\mathbf{u}}) = G_n\mathcal{N}(\mathbf{u}_{:\mathbf{t}};\mu_{\mathbf{n}}^{\mathbf{u}},\mathbf{\Sigma}_{\mathbf{n}}^{\mathbf{u}}). \tag{24}$$

Note that the notations for the below calculations are in the box after Equation (25).

$$\begin{aligned}
\mathcal{N}\left(v_{nt};\Gamma_n\mu_n^v,\sigma_v^2\Gamma_n\right) &= \mathcal{N}\left(v_{nt};\frac{\sum_{k|P_n}\phi\left(u_{kt}+c_{nk}\right)}{1+\sum_{k|P_n}\phi},\frac{\sigma_v^2}{1+\sum_{k|P_n}\phi}\right)\\
&= \left(2\pi\frac{\sigma_v^2}{1+\sum_{k|P_n}\phi}\right)^{-\frac{1}{2}}\\
&\quad\cdot\exp\left[-\frac{1}{2\frac{\sigma_v^2}{1+\sum_{k|P_n}\phi}}\left(v_{nt}-\frac{\sum_{k|P_n}\phi\left(u_{kt}+c_{nk}\right)}{1+\sum_{k|P_n}\phi}\right)^2\right]
\end{aligned}$$

$$= \frac{1}{\sqrt{2\pi s_n^2}} \exp\left[ -\frac{1}{2s_n^2} \left( v_{nt} - \frac{\sum_{k|P_n} \phi \, (u_{kt} + c_{nk})}{1 + \sum_{k|P_n} \phi} \right)^2 \right]$$

$$= \frac{1}{\sqrt{2\pi s_n^2}} \exp\left[ -\frac{1}{2s_n^2} \left( v_{nt} - \sum_{k|P_n} A_n \, u_{kt} - B_n \right)^2 \right]$$

$$= \frac{1}{\sqrt{2\pi s_n^2}} \exp\left[ -\frac{1}{2s_n^2} \left( \left( v_{nt} - B_n \right) - \left( \sum_{k|P_n} A_n \, u_{kt} \right) \right)^2 \right]$$

$$= \frac{1}{\sqrt{2\pi s_n^2}} \exp\left[ -\frac{1}{2} \left( \frac{(v_n - B_n)^2}{s_n^2} - \right. \right.$$

$$\left. \left. - \; 2 \sum_{k|P_n} \frac{(v_n - B_n) A_n u_{kt}}{s_n^2} + \sum_{k,l|P_n} \frac{A_n^2}{s_n^2} \, u_{kt} \, u_{lt} \right) \right] \qquad (25)$$

$$\boxed{\begin{aligned} s_n^2 &= \frac{\sigma_v^2}{1 + \sum_{k|P_n} \phi}, \\ A_n &= \frac{\phi}{1 + \sum_{k|P_n} \phi}, \\ B_n &= \frac{\sum_{k|P_n} \phi \, c_{nk}}{1 + \sum_{k|P_n} \phi}, \\ k|P_n &= \{k : k \in \mathrm{par}(n)\} \end{aligned}}$$

$$G_n \mathbb{P}(\mathbf{u_{:t}} | \mu_{\mathbf{n}}^{\mathbf{u}}, \mathbf{\Sigma_n^u}) = G_n \mathcal{N}(\mathbf{u_{:t}}; \mu_{\mathbf{n}}^{\mathbf{u}}, \mathbf{\Sigma_n^u})$$

$$= G_n (2\pi)^{-\frac{M_n}{2}} \det\left(\mathbf{\Sigma_n^u}\right)^{-\frac{1}{2}} \exp\left[ -\frac{1}{2} \left( (\mathbf{u_{:t}} - \mu_{\mathbf{n}}^{\mathbf{u}})^{\mathrm{T}} \mathbf{\Sigma_n^u}^{-1} (\mathbf{u_{:t}} - \mu_{\mathbf{n}}^{\mathbf{u}}) \right) \right]$$

$$= G_n (2\pi)^{-\frac{M_n}{2}} \det\left(\mathbf{\Sigma_n^u}\right)^{-\frac{1}{2}} \exp\left[ -\frac{1}{2} \left( \sum_{k,l|P_n} \mathbf{\Sigma_n^u}^{-1}_{kl} \, u_{kt} \, u_{lt} - \right. \right.$$

$$\left. \left. - \; 2 \sum_{k,l|P_n} \mathbf{\Sigma_n^u}^{-1}_{kl} \, u_{kt} \, \mu_{\mathbf{n}l}^{\mathbf{u}} + \sum_{k,l|P_n} \mathbf{\Sigma_n^u}^{-1}_{kl} \, \mu_{\mathbf{n}k}^{\mathbf{u}} \, \mu_{\mathbf{n}l}^{\mathbf{u}} \right) \right]$$

$$= G_n (2\pi)^{-\frac{M_n}{2}} \det\left(\mathbf{\Sigma_n^u}\right)^{-\frac{1}{2}} \exp\left[ -\frac{1}{2} \left( \sum_{k,l|P_n} \frac{A_n^2}{s_n^2} \, u_{kt} \, u_{lt} - \right. \right.$$

$$\left. \left. - \; 2 \sum_{k,l|P_n} \frac{A_n^2}{s_n^2} \, u_{kt} \, \frac{v_n - B_n}{A_n M_n} + \sum_{k,l|P_n} \frac{A_n^2}{s_n^2} \, \frac{v_n - B_n}{A_n M_n} \, \frac{v_n - B_n}{A_n M_n} \right) \right]$$

$$= \frac{1}{\sqrt{2\pi s_n^2}} \exp\left[ -\frac{1}{2} \left( \sum_{k,l|P_n} \frac{A_n^2}{s_n^2} \, u_{kt} \, u_{lt} - \right. \right.$$

$$- 2 \sum_{k|P_n} \frac{(v_n - B_n)A_n u_{kt}}{s_n^2} + \frac{(v_n - B_n)^2}{s_n^2} \Bigg) \Bigg] \tag{26}$$

$$\boxed{\begin{aligned} M_n &= |\{k : k \in \text{par}(n)\}| \\ \boldsymbol{\Sigma}^{\mathbf{u}\,-1}_{\mathbf{n}kl} &= \frac{A_n^2}{s_n^2} \\ \mu^{\mathbf{u}}_{\mathbf{n}k} &= \frac{v_n - B_n}{A_n M_n} \\ G_n &= (2\pi)^{\frac{M_n}{2} - \frac{1}{2}} \frac{1}{s_n} \det\left(\boldsymbol{\Sigma}^{\mathbf{u}}_{\mathbf{n}}\right)^{\frac{1}{2}} \end{aligned}}$$

With the above defined mean, variance and normalization factor we can see that the two Gaussians indeed can be transformed into each other as the last row of Equation (25) is clearly the same as that of Equation (26), hence

$$\mathcal{N}\left(v_{nt}; \Gamma_n \mu_n^v, \sigma_v^2 \Gamma_n\right) = G_n \mathcal{N}(\mathbf{u}_{:\mathbf{t}}; \mu^{\mathbf{u}}_{\mathbf{n}}, \boldsymbol{\Sigma}^{\mathbf{u}}_{\mathbf{n}}). \tag{27}$$

**A.3.2 Exploiting the Gaussian yielded in the previous point, calculate $\mathbb{P}(\mathbf{v}_{:t}|\mathbf{C}, \mathcal{S})$ by marginalizing over $\mathbf{u}_{:t}$.**
For the marginalization we will use that as per Equation (3), the position variable's prior distribution is normal around the center of the grid, marked by $(0, 0)$, with variance $\sigma_u^2$. Here we merge the $\mathbf{u}_{:t}$-dependent Gaussians to have only one $\mathbf{u}_{:t}$-dependent Gaussian which has the integral of 1. For the calculations of the normalization factors please refer to (17) and (27). Note that the auxiliary calculations and notations for steps (28), (29), (30), (31) are in the boxes at the end of the calculation.

$$\begin{aligned}
\mathbb{P}(\mathbf{v}_{:t}|\mathbf{C}, \mathcal{S}) &= \int \mathbb{P}(\mathbf{v}_{:t}|\mathbf{u}_{:\mathbf{t}}, \mathbf{C}, \mathcal{S})\, \mathbb{P}(\mathbf{u}_{:\mathbf{t}}|\mathcal{S})\, d\mathbf{u}_{:\mathbf{t}} \\
&= \int \prod_n \mathbb{P}(v_{nt}|\mathbf{u}_{:\mathbf{t}}, \mathbf{C}, \mathcal{S})\, \mathbb{P}(\mathbf{u}_{:\mathbf{t}}|\mathcal{S})\, d\mathbf{u}_{:\mathbf{t}} \\
&= \int \prod_n \mathcal{N}\left(v_{nt}; \Gamma_n \mu_n^v, \sigma_v^2 \Gamma_n\right) \mathcal{N}(\mathbf{u}_{:\mathbf{t}}; \mathbf{0}, \sigma_u^2 \mathbf{I}) d\mathbf{u}_{:\mathbf{t}} \\
&= \int \prod_n G_n \mathcal{N}(\mathbf{u}_{:\mathbf{t}}; \mu^{\mathbf{u}}_{\mathbf{n}}, \boldsymbol{\Sigma}^{\mathbf{u}}_{\mathbf{n}}) \mathcal{N}(\mathbf{u}_{:\mathbf{t}}; \mathbf{0}, \sigma_u^2 \mathbf{I}) d\mathbf{u}_{:\mathbf{t}} \qquad \text{(see Eq. 27)} \\
&= \int \mathcal{N}(\mathbf{u}_{:\mathbf{t}}; \mathbf{0}, \sigma_u^2 \mathbf{I})\, \mathcal{N}(\mathbf{u}_{:\mathbf{t}}; \mu^{\mathbf{v}}_{\mathbf{c}}, \boldsymbol{\Sigma}^{\mathbf{v}}_{\mathbf{c}}) d\mathbf{u}_{:\mathbf{t}} \qquad \text{(see Eq. 17)} \\
&\quad \cdot \frac{\prod_n G_n \mathcal{N}(\mu^{\mathbf{u}}_{\mathbf{n}}; 0, \boldsymbol{\Sigma}^{\mathbf{u}}_{\mathbf{n}})}{\mathcal{N}(\mu^{\mathbf{v}}_{\mathbf{c}}; \mathbf{0}, \boldsymbol{\Sigma}^{\mathbf{v}}_{\mathbf{c}})} \qquad (28) \\
&= \int \mathcal{N}(\mathbf{u}_{:\mathbf{t}}; \mu_{\mathbf{c}}, \boldsymbol{\Sigma}_{\mathbf{c}}) d\mathbf{u}_{:\mathbf{t}} \qquad \text{(see Eq. 17)} \\
&\quad \cdot \frac{\prod_n G_n \mathcal{N}(\mu^{\mathbf{u}}_{\mathbf{n}}; \mathbf{0}, \boldsymbol{\Sigma}^{\mathbf{u}}_{\mathbf{n}})}{\mathcal{N}(\mu^{\mathbf{v}}_{\mathbf{c}}; \mathbf{0}, \boldsymbol{\Sigma}^{\mathbf{v}}_{\mathbf{c}})} \mathcal{N}(\mu^{\mathbf{v}}_{\mathbf{c}}; \mathbf{0}, \boldsymbol{\Sigma}^{\mathbf{v}}_{\mathbf{c}} + \sigma_u^2 \mathbf{I}) \qquad (29)
\end{aligned}$$

$$
\begin{aligned}
&= \prod_n G_n \mathcal{N}(\mu_{\mathbf{n}}^{\mathbf{u}}; \mathbf{0}, \mathbf{\Sigma}_{\mathbf{n}}^{\mathbf{u}}) \frac{\mathcal{N}(\mu_{\mathbf{c}}^{\mathbf{v}}; \mathbf{0}, \mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}} + \sigma_u^2 \mathbf{I})}{\mathcal{N}(\mu_{\mathbf{c}}^{\mathbf{v}}; \mathbf{0}, \mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}})} \\
&= (2\pi)^{-\frac{K}{2}} \left| \mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}} + \sigma_u^2 \mathbf{I} \right|^{-\frac{1}{2}} \left| \mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}} \right|^{\frac{1}{2}} \left| \mathbf{\Sigma}_{\mathbf{c}} \right|^{-\frac{1}{2}} \\
&\quad \cdot \prod_n G_n \mathcal{N}(\mu_{\mathbf{n}}^{\mathbf{u}}; \mathbf{0}, \mathbf{\Sigma}_{\mathbf{n}}^{\mathbf{u}}) \frac{1}{\mathcal{N}(\mu_{\mathbf{c}}; \mathbf{0}, \mathbf{\Sigma}_{\mathbf{c}})} && (30) \\
&= (2\pi)^{-\frac{K}{2}} \left| \mathbf{\Sigma_0} \right|^{-\frac{1}{2}} \frac{\prod_n G_n \mathcal{N}(\mu_{\mathbf{n}}^{\mathbf{u}}; \mathbf{0}, \mathbf{\Sigma}_{\mathbf{n}}^{\mathbf{u}})}{\mathcal{N}(\mu_{\mathbf{c}}; \mathbf{0}, \mathbf{\Sigma}_{\mathbf{c}})} && (31) \\
&= (2\pi)^{-\frac{K}{2}} \left| \mathbf{\Sigma_0} \right|^{-\frac{1}{2}} \frac{\prod_n \mathcal{N}(B_n; v_n, s_n^2)}{\mathcal{N}(\mu_{\mathbf{c}}; \mathbf{0}, \mathbf{\Sigma}_{\mathbf{c}})} && (32)
\end{aligned}
$$

In line (28) we use the mean covariance matrix as:

$$
\begin{aligned}
\mu_{\mathbf{c}}^{\mathbf{v}} &= \mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}} \left( \sum_n \mathbf{\Sigma}_{\mathbf{n}}^{\mathbf{u}-1} \mu_{\mathbf{n}}^{\mathbf{u}} \right) \\
\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}} &= \sum_n \mathbf{\Sigma}_{\mathbf{n}}^{\mathbf{u}-1}.
\end{aligned}
$$

In line (29) we use the mean covariance matrix as:

$$
\begin{aligned}
\mu_{\mathbf{c}} &= \mathbf{\Sigma}_{\mathbf{c}} \left( \mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}-1} \mu_{\mathbf{c}}^{\mathbf{v}} + \frac{1}{\sigma_u^2} \mathbf{I} \cdot \mathbf{0} \right) \\
&= \mathbf{\Sigma}_{\mathbf{c}} \left( \sum_n \mathbf{\Sigma}_{\mathbf{n}}^{\mathbf{u}-1} \mu_{\mathbf{n}}^{\mathbf{u}} \right) \\
\mathbf{\Sigma}_{\mathbf{c}} &= \mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}-1} + \frac{1}{\sigma_u^2} \mathbf{I} = \sum_n \mathbf{\Sigma}_{\mathbf{n}}^{\mathbf{u}-1} + \frac{1}{\sigma_u^2} \mathbf{I}.
\end{aligned}
$$

For the Gaussian in the denominator of (30) we show here that

$$\mu_{\mathbf{c}}^{\mathbf{v}}(\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}+\mathbf{\Sigma_0})^{-1}\mu_{\mathbf{c}}^{\mathbf{v}}-\mu_{\mathbf{c}}^{\mathbf{v}}{\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}}^{-1}{\sigma_{\mathbf{c}}}^{-}1\mu_{\mathbf{c}}^{\mathbf{v}}=\mu_{\mathbf{c}}{\mathbf{\Sigma}_{\mathbf{c}}}^{-1}\mu_{\mathbf{c}}.$$

We use notations $\mathbf{\Sigma_0}=\sigma_u^2\mathbf{I}$ and $\hat{\mu}_{\mathbf{c}}^{\mathbf{v}}=\sum_n {\mathbf{\Sigma}_{\mathbf{n}}^{\mathbf{u}}}^{-1}\mu_{\mathbf{n}}^{\mathbf{u}}$. For better readability we omit notating transposes here.

$$\begin{aligned}
&\mu_{\mathbf{c}}^{\mathbf{v}}(\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}+\mathbf{\Sigma_0})^{-1}\mu_{\mathbf{c}}^{\mathbf{v}}-\mu_{\mathbf{c}}^{\mathbf{v}}{\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}}^{-1}{\sigma_{\mathbf{c}}}^{-}1\mu_{\mathbf{c}}^{\mathbf{v}}\\
&\qquad=\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}\hat{\mu}_{\mathbf{c}}^{\mathbf{v}}\left[\left(\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}+\mathbf{\Sigma_0}\right)^{-1}-{\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}}^{-1}\right]\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}\hat{\mu}_{\mathbf{c}}^{\mathbf{v}}\\
&\qquad=\hat{\mu}_{\mathbf{c}}^{\mathbf{v}}\left[\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}\left(\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}+\mathbf{\Sigma_0}\right)^{-1}\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}-\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}\right]\hat{\mu}_{\mathbf{c}}^{\mathbf{v}}\\
&\qquad=\hat{\mu}_{\mathbf{c}}^{\mathbf{v}}\left[\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}\left(\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}+\mathbf{\Sigma_0}\right)^{-1}\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}-(\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}+\mathbf{\Sigma_0})(\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}+\mathbf{\Sigma_0})^{-1}\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}\right]\hat{\mu}_{\mathbf{c}}^{\mathbf{v}}\\
&\qquad=-\hat{\mu}_{\mathbf{c}}^{\mathbf{v}}\mathbf{\Sigma_0}\left(\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}+\mathbf{\Sigma_0}\right)^{-1}\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}\hat{\mu}_{\mathbf{c}}^{\mathbf{v}}\\
&\qquad=-\hat{\mu}_{\mathbf{c}}^{\mathbf{v}}\left({\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}}^{-1}+{\mathbf{\Sigma_0}}^{-1}\right)^{-1}\hat{\mu}_{\mathbf{c}}^{\mathbf{v}} \qquad\qquad \mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}=(\mathbf{A}+\mathbf{B})^{-1}\\
&\qquad=-\hat{\mu}_{\mathbf{c}}\mathbf{\Sigma_c}\hat{\mu}_{\mathbf{c}}
\end{aligned}$$

For the normalizing factor in (31) we show here that

$$\left|\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}+\sigma_u^2\mathbf{I}\right|^{-\frac{1}{2}}\left|\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}\right|^{\frac{1}{2}}\left|\mathbf{\Sigma_c}\right|^{-\frac{1}{2}}=\left|\mathbf{\Sigma_0}\right|^{-\frac{1}{2}}.$$

Similarly to the calculation above, we use notation $\mathbf{\Sigma_0}=\sigma_u^2\mathbf{I}$.

$$\begin{aligned}
&|\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}+\mathbf{\Sigma_0}|^{-\frac{1}{2}}|\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}|^{\frac{1}{2}}\left|\mathbf{\Sigma_c}\right|^{-\frac{1}{2}}\\
&\qquad=\left|\left(\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}+\mathbf{\Sigma_0}\right){\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}}^{-1}\left({\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}}^{-1}+{\mathbf{\Sigma_0}}^{-1}\right)^{-1}\right|^{-\frac{1}{2}}\\
&\qquad=\left|\left(\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}\mathbf{\Sigma_0}\right)\left({\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}}^{-1}{\mathbf{\Sigma_0}}^{-1}\right)\left(\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}+\mathbf{\Sigma_0}\right){\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}}^{-1}\left({\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}}^{-1}+{\mathbf{\Sigma_0}}^{-1}\right)^{-1}\right|^{-\frac{1}{2}}\\
&\qquad=\left|\left(\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}\mathbf{\Sigma_0}\right)\left({\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}}^{-1}+{\mathbf{\Sigma_0}}^{-1}\right){\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}}^{-1}\left({\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}}^{-1}+{\mathbf{\Sigma_0}}^{-1}\right)^{-1}\right|^{-\frac{1}{2}}\\
&\qquad=\left|\mathbf{\Sigma_0}\right|^{-\frac{1}{2}}.
\end{aligned}$$

In the last step we explioted that
$\left({\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}}^{-1}+{\mathbf{\Sigma_0}}^{-1}\right){\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}}^{-1}={\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}}^{-1}\left({\mathbf{\Sigma}_{\mathbf{c}}^{\mathbf{v}}}^{-1}+{\mathbf{\Sigma_0}}^{-1}\right)$.

For (32) we rewrite the (multidimensional) factors of the production with the original variables. This leads to the 1-dimensional Gaussian: $\mathcal{N}(B_n; v_n, s_n^2)$.

$$\begin{aligned} G_n\mathcal{N}(\mu_{\mathbf{n}}^{\mathbf{u}}; \mathbf{0}, \mathbf{\Sigma}_{\mathbf{n}}^{\mathbf{u}}) = \\ &= \left((2\pi)^{\frac{M_n}{2}-\frac{1}{2}} \frac{1}{s_n} \det\left(\mathbf{\Sigma}_{\mathbf{n}}^{\mathbf{u}-1}\right)^{-\frac{1}{2}}\right)\left((2\pi)^{\frac{M_n}{2}} \frac{1}{s_n} \det\left(\mathbf{\Sigma}_{\mathbf{n}}^{\mathbf{u}}\right)^{-\frac{1}{2}}\right) \\ &\quad \cdot \exp\left[-\frac{1}{2}\sum_{k,l} \frac{v_n - B_n}{A_n M_n} \frac{A_n^2}{s_n^2} \frac{v_n - B_n}{A_n M_n}\right] \\ &= \mathcal{N}(B_n; v_n, s_n^2). \end{aligned}$$

**A.3.3 Express $\mathbb{P}(\mathbf{v}_{:t}|\mathbf{C}, \mathcal{S})$ as a Gaussian with C being the dependent variable.**
Here we reiterare the result from the previous point (in Equation 32).

$$\mathbb{P}(\mathbf{v}_{:t}|\mathbf{C}, \mathcal{S}) = (2\pi)^{-\frac{K}{2}} |\mathbf{\Sigma_0}|^{-\frac{1}{2}} \frac{\prod_n \mathcal{N}(B_n; v_n, s_n^2)}{\mathcal{N}(\mu_{\mathbf{c}}; \mathbf{0}, \mathbf{\Sigma_c})} \tag{33}$$

Now we will express (33) as a **C**-dependent Gaussian in the following steps. First, we focus on the quadratic form of the Gaussian in the denominator (i). Then we merge the quadratic forms of the Gaussian in the nominator and that in the denominator (ii). That merged quadratic form is then expressed as a quadratic form of a **C**-dependent Gaussian and a **C**-independent contstans (iii). At the end we express $\mathbb{P}(\mathbf{v}_{:t}|\mathbf{C}, \mathcal{S})$ as a Gaussian, dependent on **C**, multiplied by a normalizing factor (iv).

**A.3.3.i Quadratic form of $\mathcal{N}(\mu_{\mathbf{c}}; \mathbf{0}, \mathbf{\Sigma_c})$**
Here we will use that

$$\begin{aligned} \mu_{\mathbf{c}} &= \mathbf{\Sigma_c}\left[\sum_n \mathbf{\Sigma}_{\mathbf{n}}^{\mathbf{u}-1}\mu_{\mathbf{n}}^{\mathbf{u}}\right] = \mathbf{\Sigma_c}\left[\sum_n \left(\sum_l \mathbf{\Sigma}_{\mathbf{n}kl}^{\mathbf{u}-1}\mu_{\mathbf{n}l}^{\mathbf{u}}\right)\right] \\ &= \mathbf{\Sigma_c}\left[\sum_n \frac{A_n^2}{s_n^2}\frac{v_n - B_n}{A_n M_n}\right] = \mathbf{\Sigma_c}\left[\sum_n \frac{M_n}{M_n}\frac{A_n(v_n - B_n)}{s_n^2}\right] \\ &= \mathbf{\Sigma_c}\left[\sum_n \frac{(v_n - B_n)\phi}{\sigma_v^2}\right]. \end{aligned} \tag{34}$$

We can now calculate the quadratic form as

$$\mu_{\mathbf{c}}\mathbf{\Sigma_c}^{-1}\mu_{\mathbf{c}} = \mathbf{\Sigma_c}\left[\sum_n \frac{(v_n - B_n)\phi}{\sigma_v^2}\right]\mathbf{\Sigma_c}^{-1}\mathbf{\Sigma_c}\left[\sum_m \frac{(v_m - B_m)\phi}{\sigma_v^2}\right]$$

$$= \sum_{k,l} \mathbf{\Sigma_c}_{kl} \left[ \sum_{n,m} \frac{(v_n - B_n)(v_m - B_m)\phi^2}{\sigma_v^4} \right]$$
$$= \sum_{n,m} \left[ (v_n - B_n)(v_m - B_m) \sum_{k,l} \frac{\mathbf{\Sigma_c}_{kl}\phi^2}{\sigma_v^4} \right]. \tag{35}$$

**A.3.3.ii Quadratic form of $\frac{\prod_n \mathcal{N}(B_n; v_n, s_n^2)}{\mathcal{N}(\mu_\mathbf{c}; \mathbf{0}, \mathbf{\Sigma_c})}$**

We merge the quadratic form of the production in the nominator with that of the denominator, calculated in the previous point.

$$\begin{aligned}
&\sum_n \frac{(B_n - v_n)^2}{s_n^2} - \mu_\mathbf{c} {\mathbf{\Sigma_c}}^{-1} \mu_\mathbf{c} \\
&= \sum_{n,m} \frac{(B_n - v_n)(B_m - v_m)}{s_n s_m} \mathbf{I}_{nm} - \mu_\mathbf{c} {\mathbf{\Sigma_c}}^{-1} \mu_\mathbf{c} \\
&= \sum_{n,m} \left[ (B_n - v_n)(B_m - v_m) \left( \frac{\mathbf{I}_{nm}}{s_n s_m} - \sum_{k,l} \frac{\mathbf{\Sigma_c}_{kl}\phi^2}{\sigma_v^4} \right) \right] \\
&= \sum_{n,m} \left[ \left( \frac{\sum_{k|P_n} \phi\, c_{nk}}{1 + \sum_{k|P_n} \phi} - v_n \right) \left( \frac{\sum_{k|P_m} \phi\, c_{mk}}{1 + \sum_{k|P_m} \phi} - v_m \right) \left( \frac{\mathbf{I}_{nm}}{s_n s_m} - \sum_{k,l} \frac{\mathbf{\Sigma_c}_{kl}\phi^2}{\sigma_v^4} \right) \right] \\
&= \sum_{n,m} \left[ \left( \sum_{k|P_n} \phi\, c_{nk} - \left( 1 + \sum_{k|P_n} \phi \right) v_n \right) \left( \sum_{k|P_m} \phi\, c_{mk} - \left( 1 + \sum_{k|P_m} \phi \right) v_m \right) \cdot \right. \\
&\quad \left. \cdot \frac{1}{1 + \sum_{k|P_n} \phi} \cdot \frac{1}{1 + \sum_{k|P_m} \phi} \left( \frac{\delta_{nm}}{s_n^2} - \sum_{k,l} \frac{\mathbf{\Sigma_c}_{kl}\phi^2}{\sigma_v^4} \right) \right] \\
&= \sum_{n,m} \left[ \left( \sum_{k|P_n} \phi\, c_{nk} - \left( 1 + \sum_{k|P_n} \phi \right) v_n \right) \left( \sum_{k|P_m} \phi\, c_{mk} - \left( 1 + \sum_{k|P_m} \phi \right) v_m \right) \cdot \right. \\
&\quad \left. \cdot \left( \delta_{nm} \frac{1 + \sum_{k|P_n} \phi}{\sigma_v^2 \left( 1 + \sum_{k|P_n} \phi \right) \left( 1 + \sum_{k|P_m} \phi \right)} - \sum_{k,l} \frac{\mathbf{\Sigma_c}_{kl}\phi^2}{\sigma_v^4 \left( 1 + \sum_{k|P_n} \phi \right) \left( 1 + \sum_{k|P_m} \phi \right)} \right) \right] \\
&= \sum_{n,m} \left[ \left( \sum_{k|P_n} \phi\, c_{nk} - \left( 1 + \sum_{k|P_n} \phi \right) v_n \right) \left( \sum_{k|P_m} \phi\, c_{mk} - \left( 1 + \sum_{k|P_m} \phi \right) v_m \right) \cdot \right. \\
&\quad \left. \cdot \frac{1}{\sigma_v^4} \left( \delta_{nm} s_n^2 - \sum_{k,l} A_n \mathbf{\Sigma_c}_{kl} A_m \right) \right]
\end{aligned}$$

$$= \sum_{n,m} \Omega_{nm} \left[ \left( \sum_{k|P_n} \phi \, c_{nk} - \left( 1 + \sum_{k|P_n} \phi \right) v_n \right) \left( \sum_{k|P_m} \phi \, c_{mk} - \left( 1 + \sum_{k|P_m} \phi \right) v_m \right) \right] \tag{36}$$

$$\boxed{\Omega_{nm} = \frac{1}{\sigma_v^4} \left( \delta_{nm} s_n^2 - \sum_{k,l} A_n \mathbf{\Sigma_c}_{kl} A_m \right)}$$

**A.3.3.iii Express the quadratic from in point (ii) as a quadratic form of a C-dependent Gaussian plus C-independent constants**
We show that the quadratic form in point (ii) is $(\mathbf{C}^{vec} - \mu_{v^t}) \, \mathbf{\Sigma}_{v^t}^{-1} \, (\mathbf{C}^{vec} - \mu_{v^t})$ plus a C-independent constant. The auxiliary calculations and notations are in the box below.

$$\sum_{n,m} \Omega_{nm} \left[ \left( \sum_{k|P_n} \phi \, c_{nk} - \left( 1 + \sum_{k|P_n} \phi \right) v_n \right) \left( \sum_{k|P_m} \phi \, c_{mk} - \left( 1 + \sum_{k|P_m} \phi \right) v_m \right) \right]$$

$$= \sum_{n,m} \Omega_{nm} \left( 1 + \sum_{k|P_n} \phi \right) v_n \left( 1 + \sum_{k|P_m} \phi \right) v_m$$

$$+ \sum_{n,m} \Omega_{nm} \left( \sum_{k|P_n} \phi \, c_{nk} \right) \left( \sum_{k|P_m} \phi \, c_{mk} \right) \tag{37}$$

$$- \sum_{n,m} \Omega_{nm} \left( \sum_{k|P_n} \phi \, c_{nk} \right) \left( 1 + \sum_{k|P_m} \phi \right) v_m \tag{38}$$

$$- \sum_{n,m} \Omega_{nm} \left( \sum_{k|P_m} \phi \, c_{mk} \right) \left( 1 + \sum_{k|P_n} \phi \right) v_n \tag{39}$$

$$= \sum_{n,m} \Omega_{nm} \left( 1 + \sum_{k|P_n} \phi \right) v_n \left( 1 + \sum_{k|P_m} \phi \right) v_m$$

$$+ \mathbf{C}^{vec} \mathbf{\Sigma}_{v^t}^{-1} \mathbf{C}^{vec} - 2 \mathbf{C}^{vec} \mathbf{\Sigma}_{v^t}^{-1} \mu_{v^t} \tag{40}$$

$$= \sum_{n,m} \Omega_{nm} \left( 1 + \sum_{k|P_n} \phi \right) v_n \left( 1 + \sum_{k|P_m} \phi \right) v_m$$

$$- \mu_{v^t} \mathbf{\Sigma}_{v^t}^{-1} \mu_{v^t} + (\mathbf{C}^{vec} - \mu_{v^t}) \, \mathbf{\Sigma}_{v^t}^{-1} \, (\mathbf{C}^{vec} - \mu_{v^t}) \tag{41}$$

For (40) we need to calculate the covariance matrix for the distribution in matrix $\mathbf{C}$. For this, we will handle matrix $\mathbf{C}$ as a $\sum_{i,j}\mathbf{Z}_{ij}$-dimensional vector (as will be exploited later, in (43)), $\mathbf{Z}$ being the link matrix. For clarity, we will use the annotation for the number of links

$$L = \sum_{i,j}\mathbf{Z}_{ij}.$$

The covariance matrix of such-defined $L$-dimensional $\mathbf{C}_{\text{vec}}$ vector will have joint indices of $\{nk\}$ and $\{ml\}$. The covariance matrix can be derived from lines (37) and (40) as:

$$\mathbf{\Sigma}^{-1}_{v^t\,\{nk\}\{ml\}} = \Omega_{nm}\phi^2.$$

Also, based on lines (38), (39) and (40), we show below that the mean is

$$\mu_{v^t\,\{ml\}} = \sum_{\{nk\}}\mathbf{\Sigma}_{v^t\,\{nk\}\{ml\}}\mathbf{w}^t_{\{nk\}},$$

where

$$\mathbf{w}^t_{\{nk\}} = \phi\left(\sum_m \Omega_{nm}\left(1+\sum_{k|P_m}\phi\right)v_m\right). \tag{42}$$

For this we need to show the equivalence of the expressions in line (38) and (40): as $\Omega_{nm} = \Omega_{mn}$, we have to show that

$$\mathbf{C}^{vec}\mathbf{\Sigma}^{-1}_{v^t}\mu_{v^t} = \sum_{n,m}\Omega_{nm}\left(\sum_{k|P_n}\phi\,c_{nk}\right)\left(1+\sum_{k|P_m}\phi\right)v_m.$$

Indeed,

$$\begin{aligned}
\mathbf{C}^{vec}\mathbf{\Sigma}^{-1}_{v^t}\mu_{v^t} &= \sum_{\{nk\}}\mathbf{C}^{vec}_{\{nk\}}\sum_{\{ml\}}\mathbf{\Sigma}^{-1}_{v^t\,\{nk\}\{ml\}}\mu_{v^t\{ml\}} \\
&= \sum_{\{nk\}}\mathbf{C}^{vec}_{\{nk\}}\sum_{\{ml\}}\mathbf{\Sigma}^{-1}_{v^t\,\{nk\}\{ml\}}\sum_{\{nk\}}\mathbf{\Sigma}_{v^t\,\{nk\}\{ml\}}\mathbf{w}^t_{\{nk\}} \\
&= \sum_{\{nk\}}\mathbf{C}^{vec}_{\{nk\}}\phi\left(\sum_m\Omega_{nm}\left(1+\sum_{k|P_m}\phi\right)v_m\right) \\
&= \sum_n\left(\sum_{k|P_n}c_{nk}\phi\right)\left(\sum_m\Omega_{nm}\left(1+\sum_{k|P_m}\phi\right)v_m\right).
\end{aligned}$$

We will se later that in the implementation we will not need to calculate $\mu_{v^t}$, hence $\mathbf{\Sigma}^{-1}_{v^t}\mu_{v^t}$ won't need to be inverted.

**A.3.3.iv Express $\mathbb{P}(\mathbf{v}_{:t}|\mathbf{C},\mathcal{S})$ as a Gaussian, dependent on C (multiplied by a normalizing factor)**
Note that the auxiliary calculation for line (44) is in the box below.

$$
\begin{aligned}
\mathbb{P}(\mathbf{v}_{:t}|\mathbf{C},\mathcal{S}) &= (2\pi)^{-\frac{K}{2}} \left|\mathbf{\Sigma_0}\right|^{-\frac{1}{2}} \frac{\prod_n \mathcal{N}(B_n; v_n, s_n^2)}{\mathcal{N}(\mu_\mathbf{c}; \mathbf{0}, \mathbf{\Sigma_c})} \\
&= (2\pi)^{-\frac{K}{2}} \left|\mathbf{\Sigma_0}\right|^{-\frac{1}{2}} \frac{(2\pi)^{-\frac{N}{2}} \prod_n s_n^{-1} \exp\left[-\frac{1}{2}\sum_n \frac{(B_n - v_n)^2}{s_n^2}\right]}{(2\pi)^{-\frac{N}{2}} \left|\mathbf{\Sigma_c}\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\mu_\mathbf{c}{\mathbf{\Sigma_c}}^{-1}\mu_\mathbf{c}\right]} \\
&= (2\pi)^{-\frac{K}{2}} \left|\mathbf{\Sigma_0}\right|^{-\frac{1}{2}} \prod_n s_n^{-1} \left|\mathbf{\Sigma_c}\right|^{\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\sum_n \frac{(B_n - v_n)^2}{s_n^2} - \mu_\mathbf{c}{\mathbf{\Sigma_c}}^{-1}\mu_\mathbf{c}\right)\right] \\
&= (2\pi)^{-\frac{K}{2}} \left|\mathbf{\Sigma_0}\right|^{-\frac{1}{2}} \prod_n s_n^{-1} \left|\mathbf{\Sigma_c}\right|^{\frac{1}{2}} \\
&\quad \cdot \exp\left[-\frac{1}{2}\sum_{n,m} \Omega_{nm} \left[\left(\sum_{k|P_n} \phi\, c_{nk} - \left(1 + \sum_{k|P_n} \phi\right) v_n\right) \right.\right. \\
&\qquad \left.\left. \left(\sum_{k|P_m} \phi\, c_{mk} - \left(1 + \sum_{k|P_m} \phi\right) v_m\right)\right]\right] \\
&= (2\pi)^{-\frac{K}{2}} \left|\mathbf{\Sigma_0}\right|^{-\frac{1}{2}} \prod_n s_n^{-1} \left|\mathbf{\Sigma_c}\right|^{\frac{1}{2}} \\
&\quad \cdot \exp\left[-\frac{1}{2}\left[\sum_{n,m} \Omega_{nm} \left(1 + \sum_{k|P_n} \phi\right) v_n \left(1 + \sum_{k|P_m} \phi\right) v_m - \mu_{v^t}{\mathbf{\Sigma}_{v^t}^{-1}}\mu_{v^t}\right]\right] \\
&\quad \cdot \exp\left[-\frac{1}{2}\left(\mathbf{C}^{vec} - \mu_{v^t}\right)\mathbf{\Sigma}_{v^t}^{-1}\left(\mathbf{C}^{vec} - \mu_{v^t}\right)\right] \\
&= (2\pi)^{-\frac{K}{2}} \left|\mathbf{\Sigma_0}\right|^{-\frac{1}{2}} \prod_n s_n^{-1} \left|\mathbf{\Sigma_c}\right|^{\frac{1}{2}} (2\pi)^{\frac{L}{2}} \left|\mathbf{\Sigma}_{v^t}\right|^{\frac{1}{2}} \qquad (43) \\
&\quad \cdot \exp\left[-\frac{1}{2}\left[\sum_{n,m} \Omega_{nm} \left(1 + \sum_{k|P_n} \phi\right) v_n \left(1 + \sum_{k|P_m} \phi\right) v_m - \mu_{v^t}{\mathbf{\Sigma}_{v^t}^{-1}}\mu_{v^t}\right]\right] \\
&\quad \cdot \mathcal{N}\left(\mathbf{C}^{vec}; \mu_{v^t}, \mathbf{\Sigma}_{v^t}\right) \\
&= \Psi_t (2\pi)^{\frac{L}{2}} \left|\mathbf{\Sigma}_{v^t}\right|^{\frac{1}{2}} \exp\left[\frac{1}{2}\mu_{v^t}{\mathbf{\Sigma}_{v^t}^{-1}}\mu_{v^t}\right] \mathcal{N}\left(\mathbf{C}^{vec}; \mu_{v^t}, \mathbf{\Sigma}_{v^t}\right) \qquad (44)
\end{aligned}
$$

$$\Psi_t = (2\pi)^{-\frac{K}{2}} \left|\mathbf{\Sigma_0}\right|^{-\frac{1}{2}} \prod_n s_n^{-1} \left|\mathbf{\Sigma_c}\right|^{\frac{1}{2}} \cdot \exp\left[-\frac{1}{2}\left[\sum_{n,m} \Omega_{nm} \left(1 + \sum_{k|P_n} \phi\right) v_n \left(1 + \sum_{k|P_m} \phi\right) v_m\right]\right]$$

**A.3.4 From the previous point, express $\mathbb{P}(\mathbf{V}|\mathbf{C},\mathcal{S})$, the likelihood of the whole observed position matrix across trials, as a Gaussian with C being the dependent variable.**

We show that

$$\mathbb{P}(\mathbf{V}|\mathbf{C},\mathcal{S}) = H \cdot \mathcal{N}(\mathbf{C}^{vec}; \mu_{v^T}, \mathbf{\Sigma}_{v^T}),$$

where $H$ is a normalizing factor, and the mean and covariance are defined in the box below.

$$\begin{aligned}
\mathbb{P}(\mathbf{V}|\mathbf{C},\mathcal{S}) &= \prod_t \mathbb{P}(\mathbf{v}_{:t}|\mathbf{C},\mathcal{S}) \\
&= \prod_t \left[\Psi_t (2\pi)^{\frac{L}{2}} \left|\mathbf{\Sigma}_{v^t}\right|^{\frac{1}{2}} \exp\left[\frac{1}{2}\mu_{v^t}\mathbf{\Sigma}_{v^t}^{-1}\mu_{v^t}\right] \mathcal{N}\left(\mathbf{C}^{vec}; \mu_{v^t}, \mathbf{\Sigma}_{v^t}\right)\right] \\
&= \left[\prod_t \Psi_t (2\pi)^{\frac{L}{2}} \left|\mathbf{\Sigma}_{v^t}\right|^{\frac{1}{2}} \exp\left[\frac{1}{2}\mu_{v^t}\mathbf{\Sigma}_{v^t}^{-1}\mu_{v^t}\right]\right] \\
&\quad \cdot \left[\prod_t (2\pi)^{-\frac{L}{2}} \left|\mathbf{\Sigma}_{v^t}\right|^{-\frac{1}{2}}\right] \\
&\quad \cdot \exp\left[-\frac{1}{2}\sum_t \left(\left(\mathbf{C}^{vec} - \mu_{v^t}\right)\mathbf{\Sigma}_{v^t}^{-1}\left(\mathbf{C}^{vec} - \mu_{v^t}\right)\right)\right] \\
&= \left[\prod_t \Psi_t\right] \exp\left[\frac{1}{2}\sum_t \mu_{v^t}\mathbf{\Sigma}_{v^t}^{-1}\mu_{v^t}\right] \\
&\quad \cdot \exp\left[-\frac{1}{2}\left(\left(\mathbf{C}^{vec} - \mu_{v^T}\right)\mathbf{\Sigma}_{v^T}^{-1}\left(\mathbf{C}^{vec} - \mu_{v^T}\right)\right)\right. \\
&\qquad \left. -\frac{1}{2}\sum_t \mu_{v^t}\mathbf{\Sigma}_{v^t}^{-1}\mu_{v^t} + \frac{1}{2}\mu_{v^T}\mathbf{\Sigma}_{v^T}^{-1}\mu_{v^T}\right] \\
&= \frac{\prod_t \Psi_t}{\mathcal{N}(\mu_{v^T}; \mathbf{0}, \mathbf{\Sigma}_{v^T})} \mathcal{N}(\mathbf{C}^{vec}; \mu_{v^T}, \mathbf{\Sigma}_{v^T}) \qquad (45)
\end{aligned}$$

For the mean and covariance matrix of the full likelihood over all trials are calculated below, using (17).

$$\begin{aligned}
\mathbf{\Sigma}_{v^T}^{-1} &= \sum_t \mathbf{\Sigma}_{v^t}^{-1} \\
\mu_{v^T} &= \mathbf{\Sigma}_{v^T} \left( \sum_t \mathbf{\Sigma}_{v^t}^{-1} \mu_{v^t} \right) \\
&= \mathbf{\Sigma}_{v^T} \left( \sum_t \mathbf{\Sigma}_{v^t}^{-1} \left( \mathbf{\Sigma}_{v^t} \mathbf{w}^t \right) \right) \\
&= \mathbf{\Sigma}_{v^T} \left( \sum_t \mathbf{w}^t \right) \qquad (46)
\end{aligned}$$

For the formula of $\mathbf{w}^t$ please refer to (42).

**A.3.5 Marginalize over C to have $\mathbb{P}(\mathbf{V}|\mathcal{S})$.**
We use equations (17) in line (47). The auxiliary calculations and notations are in the box below.

$$\begin{aligned}
\mathbb{P}(\mathbf{V}|\mathcal{S}) &= \int \mathbb{P}(\mathbf{V}|\mathbf{C},\mathcal{S})\, \mathbb{P}(\mathbf{C}|\mathcal{S})\, d\mathbf{C} \\
&= \frac{\prod_t \Psi_t}{\mathcal{N}(\mu_{v^T}; \mathbf{0}, \mathbf{\Sigma}_{v^T})} \int \mathcal{N}(\mathbf{C}^{vec}; \mu_{v^T}, \mathbf{\Sigma}_{v^T})\, \mathcal{N}(\mathbf{C}^{vec}; \mathbf{0}, \sigma_C \mathbf{I})\, d\mathbf{C} \\
&= \frac{\prod_t \Psi_t}{\mathcal{N}(\mu_{v^T}; \mathbf{0}, \mathbf{\Sigma}_{v^T})} \mathcal{N}(\mu_{v^T}; \mathbf{0}, \mathbf{\Sigma}_{v^T} + \sigma_C \mathbf{I}) \int \mathcal{N}\Big(\mathbf{C}^{vec}; \mu_{C^T_{\text{post}}}, \mathbf{\Sigma}_{C^T_{\text{post}}}\Big)\, d\mathbf{C} \qquad (47) \\
&= \prod_t \Psi_t \frac{\mathcal{N}(\mu_{v^T}; \mathbf{0}, \mathbf{\Sigma}_{v^T} + \sigma_C \mathbf{I})}{\mathcal{N}(\mu_{v^T}; \mathbf{0}, \mathbf{\Sigma}_{v^T})} \\
&= (2\pi)^{-\frac{L}{2}} \sigma_C^{-L} \prod_t \Psi_t \cdot \frac{1}{\mathcal{N}\Big(\mu_{C^T_{\text{post}}}; \mathbf{0}, \mathbf{\Sigma}_{C^T_{\text{post}}}\Big)} \qquad (48)
\end{aligned}$$

For the mean and covariance matrix of the posterior of $\mathbf{C}$, used above, we use (17).

$$\begin{aligned}
\mathbf{\Sigma}_{C^T_{\text{post}}}^{-1} &= \mathbf{\Sigma}_{v^T}^{-1} + \frac{1}{\sigma_C^2} \\
\mu_{C^T_{\text{post}}} &= \mathbf{\Sigma}_{C^T_{\text{post}}} \left( \mathbf{\Sigma}_{v^T}^{-1} \mu_{v^T} + \frac{1}{\sigma_C^2} \mathbf{0} \right) \\
&= \mathbf{\Sigma}_{C^T_{\text{post}}} \sum_t \mathbf{w}^t \qquad (49)
\end{aligned}$$

# References


[1] Austerweil JL, and Griffiths TL (2011). A rational model of the effects of distributional information on feature learning. Cognitive psychology, 63(4), 173-209.

[2] Christiansen MH, Allen J, Seidenberg MS (1998) Learning to segment speech using multiple cues: a connectionist model. Lang Cognit Proc 13:221–268.

[3] Fiser J, Aslin RN (2001) Unsupervised statistical learning of higher-order spatial structures from visual scenes. Psychol Sci 12:499–504.

[4] Fiser J, Aslin RN (2002) Statistical learning of new visual feature combinations by infants. Proc Natl Acad Sci USA 99:15822–15826.

[5] Fiser J, Aslin RN (2005) Statistical learning of higher-order temporal structure from visual shape sequences. J Exp Psychol Gen 134:521–537.

[6] Goldwater S, Griffiths TL, Johnson M (2009). A Bayesian framework for word segmentation: Exploring the effects of context. Cognition, 112(1), 21–54

[7] Lee AL, Liu Z, Lu H (2020). Parts beget parts: Bootstrapping hierarchical object representations through visual statistical learning. Cognition, 104515.

[8] Perruchet P (2019). What mechanisms underlie implicit statistical learning? Transitional probabilities versus chunks in language learning. Topics in cognitive science, 11(3), 520-535.

[9] Ullman S, Vidal-Naquet M, Sali E (2002) Visual features of intermediate complexity and their use in classification. Nat Neurosci 5:682–687.

[10] Welling M (2007) Product of experts. Scholarpedia, 2(10):3879.

[11] Wood F, Griffith T, Ghahramani Z (2012) A Non-Parametric Bayesian Method for Inferring Hidden Causes. arXiv preprint arXiv:1206.6865.