A Sparsity analysis in detail

A.1 Sparsity analysis of W: likelihood $L(\boldsymbol{\alpha})$

Since it is analytically intractable of solving W from its posterior distribution, we turned to discuss its precision hyperparameter $\boldsymbol{\alpha}$ instead. To estimate $\boldsymbol{\alpha}$, we can compute the likelihood of the weight precision $\boldsymbol{\alpha}$ by marginalizing the model likelihood $P(\boldsymbol{y}|W,\boldsymbol{x})$ over the loading weight W, and we get :

$$P(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\alpha}) = \int P(\boldsymbol{y}|\boldsymbol{x},W)P(W|\boldsymbol{\alpha})dW \tag{51}$$

$$= \int \cdots \int \mathcal{N}(\boldsymbol{y}|\sum_{i=1}^{D} x_i \boldsymbol{w}_i + \boldsymbol{\mu}, L^{-1}) \prod_{i=1}^{D} \mathcal{N}(\boldsymbol{w}_i|0, \alpha_i^{-1} I_N) d\boldsymbol{w}_1 \dots d\boldsymbol{w}_D \tag{52}$$

$$= \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}, L^{-1} + \sum_i x_i^2 \alpha_i^{-1} I) \tag{53}$$

where $L = diag(\boldsymbol{\tau})$.

The log-likelihood function of $\boldsymbol{\alpha}$ is:

$$L(\boldsymbol{\alpha}) : \log P(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\alpha}) = -\frac{1}{2}\log|C| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T C^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) + \text{const.}$$

where $C = L^{-1} + \sum_{i=1}^{N} \alpha_i^{-1} x_i^2 I$, $L = diag(\boldsymbol{\tau})$.

Separating the terms with and without α_i, the covariance matrix C can be rewritten as:

$$C = C_{-i} + \alpha_i^{-1} x_i^2 I$$

where $C_{-i} = L^{-1} + \sum_{d \neq i} \alpha_d^{-1} x_d^2 I_N$, $L = diag(\boldsymbol{\tau})$.

Based on the inverse of the sum matrices [32], the determinate and inverse matrix of C can be rewritten as:

$$|C| = |C_{-i} + \alpha_i^{-1} x_i^2 I| = \alpha_i^{-1} |C_{-i}||\alpha_i I + x_i^2 C_{-i}^{-1}| \tag{54}$$

$$C^{-1} = (C_{-i} + \alpha_i^{-1} x_i^2 I)^{-1} = C_{-i}^{-1} - \frac{1}{\alpha_i + g_i} x_i^2 C_{-i}^{-2}, \text{where } g_i = x_i^2 tr(C_{-i}^{-1}) \tag{55}$$

$L(\boldsymbol{\alpha})$ can be reformed as:

$$L(\boldsymbol{\alpha}) = -\frac{1}{2}\log|C_{-i}| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T C_{-i}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) \tag{56}$$

$$+ \frac{N}{2}\log\alpha_i - \frac{1}{2}\log|\alpha_i I + x_i^2 C_{-i}^{-1}| + \frac{1}{2}\frac{1}{\alpha_i + g_i} x_i^2 (\boldsymbol{y}-\boldsymbol{\mu})^T C_{-i}^{-2}(\boldsymbol{y}-\boldsymbol{\mu}) + \text{const.} \tag{57}$$

$$= L(\boldsymbol{\alpha_{-i}}) + l(\alpha_i) \tag{58}$$

The partial derivative over α_i becomes

$$\frac{\partial L(\boldsymbol{\alpha})}{\partial \alpha_i} = \frac{\partial l(\alpha_i)}{\partial \alpha_i} = \frac{N}{2}\frac{\partial \log\alpha_i}{\partial \alpha_i} - \frac{1}{2}\frac{\partial \log|\alpha_i I + x_i^2 C_{-i}^{-1}|}{\partial \alpha_i} \tag{59}$$

$$= \frac{1}{2}\frac{a_i^{-1} g_i^2 - (q_i^2 - g_i)}{(\alpha_i + g_i)^2} \tag{60}$$

where we have defined

$$g_i = x_i^2 tr(C_{-i}^{-1}), q_i^2 = x_i^2 (\boldsymbol{y}-\boldsymbol{\mu})^T C_{-i}^{-2}(\boldsymbol{y}-\boldsymbol{\mu}), \tag{61}$$

and we have made use of the properties of matrix derivatives:

$$\frac{\partial \log|\alpha_i I + x_i^2 C_{-i}^{-1}|}{\partial \alpha_i} = \frac{N}{\alpha_i} - \frac{g_i}{\alpha_i(\alpha_i + g_i)}$$

Let $\frac{\partial L(\boldsymbol{\alpha}_i, W)}{\partial \alpha_i} = 0$, we can get two stationary points for α_i:

$$\alpha_i = \begin{cases} +\infty, \\ \frac{g_i^2}{q_i^2 - g_i}, \text{ subject to } q_i^2 > g_i \end{cases} \tag{62}$$

Next, we need to discuss whether the stationary points are maximum by taking the second derivative of $L(\boldsymbol{\alpha})$ with respect to α_i:

$$\frac{\partial^2 L(\boldsymbol{\alpha})}{\partial \alpha_i^2} = \frac{-\alpha_i^{-2} g_i^2 (\alpha_i + g_i)^2 - 2(\alpha_i + g_i)[\alpha_i^{-1} g_i^2 - (q_i^2 - g_i)]}{2(\alpha_i + g_i)^4} \tag{63}$$

$$= \begin{cases} \frac{-g_i^2}{2\alpha_i^2(\alpha_i + g_i)^2} < 0, & \text{if } \alpha_i = \frac{g_i^2}{q_i^2 - g_i} \\ 0, & \text{if } \alpha_i = \infty \end{cases} \tag{64}$$

Because $\frac{\partial^2 L(\boldsymbol{\alpha})}{\partial \alpha_i^2}\Big|_{\alpha_i = \frac{g_i^2}{q_i^2 - g_i}} < 0$, $L(\boldsymbol{\alpha})$ at $\alpha_i = \frac{g_i^2}{q_i^2 - g_i}$ has a unique maximum. On the other hand, because $\frac{\partial^2 L(\boldsymbol{\alpha})}{\partial \alpha_i^2}\Big|_{\alpha_i = \infty} = 0$, the condition of the infinity stationary point needs to be determined based on the analysis of the sign of $\frac{\partial L(\boldsymbol{\alpha})}{\partial \alpha_i}$ while α_i approaches to ∞: when $g_i^2 - s_i > 0$, $\frac{\partial L(\boldsymbol{\alpha})}{\partial \alpha_i}$ is negative while α_i approaches to ∞, therefore $\alpha_i = \infty$ is a minimum; when $g_i^2 - s_i \leq 0$, $\frac{\partial L(\boldsymbol{\alpha})}{\partial \alpha_i}$ is positive while α_i approaches to ∞, then $\alpha_i = \infty$ is a maximum.

Thus, we can get the maximum likelihood solution of α_i with conditions as:

$$\alpha_i = \begin{cases} +\infty, \text{ subject to } q_i^2 \leq g_i \\ \frac{s_i^2}{q_i^2 - g_i}, \text{ subject to } q_i^2 > g_i \end{cases}$$

Condition of $\alpha_i = \infty$

To let $w_i = 0$, we need the corresponding $\alpha_i = \infty$. The condition of $\alpha_i = \infty$, $q_i^2 \leq g_i$, requires:

$$x_i^2 (\boldsymbol{y} - \boldsymbol{\mu})^T C_{-i}^{-2} (\boldsymbol{y} - \boldsymbol{\mu}) \leq x_i^2 tr(C_{-i}^{-1})$$

Because C is a diagonal matrix, the inequation above can be expanded explicitly as:

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 c_{-in}^{-2} \leq \sum_{n=1}^{N} c_{-in}^{-1}$$

where $c_{-in} = \tau_n^{-1} + \sum_{d \neq i} \alpha_d^{-1} x_d^2$, y_n, μ_n, τ_n represent the observed neural activity, the mean activity and the estimated Gaussian noise of the n-th neuron, respectively.

Thus, to prune certain w_i, we need

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} \tau_n^{-1} + N \sum_{d \neq i} \alpha_d^{-1} x_d^2 \tag{65}$$

Based on the variational solution of x (as shown in Equation 23), the value of x_d^2 depends on the corresponding β_d. In Bayesian PCA, β_d is fixed at 1, whereas in the Dual ARD formulation, β_d has the opportunity to be estimated to a smaller value. Consequently, x_d^2 has a higher chance of taking a larger value compared to Bayesian PCA. As a result, it is easier for the Dual ARD formulation to satisfy the condition $q_i^2 \leq g_i$, leading to a greater

degree of sparsity in W compared to Bayesian PCA.

A.2 Sparsity analysis of x: likelihood $L(\beta)$

The likelihood function of a precision hyperparameter can be achieved using type-II likelihood [cite Bishop, 2006]. In a linear latent variable model, we marginalize the model likelihood $P(\boldsymbol{y}|\boldsymbol{W},\boldsymbol{x})$ over the latent variable $\boldsymbol{x}$, and we get the log likelihood of its precision $\boldsymbol{\beta}$ as:

$$\begin{aligned} L(\boldsymbol{\beta}) : \log P(\boldsymbol{y}|W,\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\tau}) &= \log \int P(\boldsymbol{y}|W,\boldsymbol{x},\mu,\boldsymbol{\tau})P(\boldsymbol{x}|\boldsymbol{\beta})d\boldsymbol{x} \\ &= -\frac{1}{2}\log|C| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T C^{-1}(\boldsymbol{t}-\boldsymbol{\mu}) + \text{const.} \end{aligned}$$

where $C = L^{-1} + WB^{-1}W^T$, $L = diag(\boldsymbol{\tau})$, $B = diag(\boldsymbol{\beta})$.

The maximum likelihood solution of each individual β_i can be obtained by setting the derivative of $L(\beta)$ with respect to each β_i to zero. If a stationary point of the likelihood occurs at $\beta_i = \infty$, it will lead to the corresponding x_i peaking at zero. To ensure our discussion on each individual β_i is not disturbed by the terms containing the other precision hyperparameters, we rewrote the likelihood function $L(\boldsymbol{\beta})$ by separating the terms with and without β_i. The covariance matrix C can be rewritten as:

$$C = C_{-i} + \beta_i^{-1}\boldsymbol{w}_i\boldsymbol{w}_i^T \tag{66}$$

where $C_{-i} = L^{-1} + \sum_{d\neq i}\beta_d^{-1}\boldsymbol{w}_d\boldsymbol{w}_d^T$.

Based on properties of matrix operations, the determinant and inverse of the covariance matrix C can be re-written as:

$$\begin{aligned} |C| &= |C_{-i}||I + \beta_i^{-1}C_{-i}^{-1}w_i w_i^T| = |C_{-i}|(1+\beta_i^{-1}w_i^T C_{-i}^{-1}w_i) = |C_{-i}|(1+\beta_i^{-1}g_i) \\ C^{-1} &= (C_{-i} + \beta_i^{-1}\boldsymbol{w}_i\boldsymbol{w}_i^T)^{-1} \\ &= C_{-i}^{-1} - \frac{1}{\beta_i + g_i}C_{-i}^{-1}\boldsymbol{w}_i\boldsymbol{w}_i^T C_{-i}^{-1}, \end{aligned}$$

where we have denoted $g_i = tr(\boldsymbol{w}_i\boldsymbol{w}_i^T C_{-i}^{-1}) = \boldsymbol{w}_i^T C_{-i}^{-1}\boldsymbol{w}_i$.

Thus, $L(\boldsymbol{\beta})$ can be rewritten as:

$$\begin{aligned} L(\boldsymbol{\beta},W) =& -\frac{1}{2}\log|C_{-i}| - \frac{1}{2}(y-\mu)^T C_{-i}^{-1}(y-\mu) \\ & - \frac{1}{2}\log(1+\beta_i^{-1}\boldsymbol{w}_i^T C_{-i}^{-1}\boldsymbol{w}_i) - \frac{1}{2}\frac{1}{\beta_i + \boldsymbol{w}_i^T C_{-i}^{-1}\boldsymbol{w}_i}(y-\mu)^T C_{-i}^{-1}\boldsymbol{w}_i\boldsymbol{w}_i^T C_{-i}^{-1}(y-\mu) \\ =& L(\boldsymbol{\beta}_{-i}) + l(\beta_i) \end{aligned}$$

where we have used $L(\beta_{-i})$ to denote the log likelihood with β_i removed from $L(\beta)$, and $l(\beta_i)$ to represent the terms that contains β_i.

The first derivative of $L(\beta)$ over β_i becomes:

$$\frac{\partial L(\beta)}{\partial \beta_i} = \frac{\partial l(\beta_i)}{\partial \beta_i} = \frac{1}{2}\frac{\beta_i^{-1}g_i^2 - (q_i^2 - g_i)}{(\beta_i + g_i)^2} \tag{67}$$

where we have let

$$\begin{aligned} g_i &= \boldsymbol{w}_i^T C_{-i}^{-1}\boldsymbol{w}_i \\ q_i^2 &= (y-\mu)^T C_{-i}^{-1}\boldsymbol{w}_i\boldsymbol{w}_i^T C_{-i}^{-1}(y-\mu) = (\boldsymbol{w}_i^T C_{-i}^{-1}(y-\mu))^2 \end{aligned} \tag{68}$$

Let $\frac{\partial L(\boldsymbol{\beta},W)}{\partial \beta_i} = 0$, we get the stationary points of β_i and corresponding conditions :

$$\beta_i = \begin{cases} +\infty, \text{ subject to } q_i^2 \leq g_i \\ \frac{g_i^2}{q_i^2 - g_i}, \text{ subject to } q_i^2 > g_i \end{cases} \tag{69}$$

Conditions of $\beta_i = \infty$

To achieve sparsity, some of the precision hyperparameter needs to take the infinite stationary point. The corresponding condition of infinite stationary point requires $q_i^2 \leq g_i$:

$$q_i^2 - g_i = \boldsymbol{w}_i^T C_{-i}^{-1} \mathcal{A} C_{-i}^{-1} \boldsymbol{w}_i \leq 0 \tag{70}$$

where we have denoted

$$\begin{aligned} \mathcal{A} &\triangleq (\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})^T - C_{-i} \\ &= (\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})^T - (L^{-1} + \sum_{d \neq i} \beta_d^{-1} w_d w_d^T) \end{aligned} \tag{71}$$

As a sufficient (but not necessary) condition, if $q_i^2 \leq g_i$ holds for any value of w_i, the corresponding β_i will be set to infinite. Since C_{-i}^{-1} is a positive definite matrix, satisfying $q^2 \leq g_i$ for any w_i requires $\mathcal{A}$ to be a negative semi-definite matrix. In practical terms, when pruning out the latent variable x_i, the corresponding prior precision β_i does not need to be estimated as infinitely large. It is sufficient for β_i to take a large enough value that makes the posterior of x_i taking a near-zero mean. Therefore, in addition to being a negative semi-definite matrix, $\mathcal{A}$ can also be a small positive matrix, effectively inducing sparsity in the model.

In the case of Bayesian PCA, where a single ARD prior is assigned to the loading weight W, the prior of latent variables is set as a unit-Gaussian distribution, resulting in a fixed value of $\beta_d = 1$. On the other hand, with Dual ARD formulations, β_d needs to be trained from the data using a Bayesian framework and may be inferred as a value smaller than 1. As a result, it is easier to satisfy the conditions of $q_i^2 \leq g_i$ with Dual ARD formulations compared to single ARD formulations in Bayesian PCA. Thus, our proposed Dual ARD can reduce the dimensionality of the latent space more efficiently than Bayesian PCA.

A.3 Hyperparameter conditions for of $P(\boldsymbol{\alpha})$ and $P(\boldsymbol{\beta})$

A.3.1 Hyperparameter of $P(\boldsymbol{\alpha})$

To ensure that $\boldsymbol{w}$ takes a broad prior, it is important for the prior distribution $P(\boldsymbol{\alpha})$ to favor smaller values of $\boldsymbol{\alpha}$. Additionally, in order to achieve sparsity in the posterior distribution of $\boldsymbol{w}$, we need the posterior distribution of $\boldsymbol{\alpha}$, $Q(\boldsymbol{\alpha})$, to allow for larger values of α_i when corresponding elements w_i approach zero, which will result in w_i peaking at zero and contribute to the desired sparsity.

The prior of $\boldsymbol{\alpha}$ is:

$$P(\boldsymbol{\alpha}) = \prod_{i=1}^{q} \mathrm{Gam}(\alpha_i | a_{\alpha_0}, b_{\alpha_0}) \tag{72}$$

The mode of $P(\boldsymbol{\alpha})$, the most often value of $\boldsymbol{\alpha}$, is:

$$mode(P(\boldsymbol{\alpha})) = \begin{cases} 0, \text{ when } a_{\alpha_0} \leq 1 \\ \frac{a_{\alpha_0} - 1}{b_{\alpha_0}}, \text{ when } a_{\alpha_0} > 1 \end{cases} \tag{73}$$

To ensure that $\boldsymbol{w}$ has a broad prior, $P(\boldsymbol{\alpha})$ should favor small values of $\boldsymbol{\alpha}$. Consequently, the mode of $P(\boldsymbol{\alpha})$ should take a small value. Therefore, we need $a_{\alpha_0} \leq 1$ or $a_{\alpha_0} - 1 \ll b_{\alpha_0}$ (Fig. S4a,d).

Using variational Bayesian inference, the approximated posterior of $\boldsymbol{\alpha}$, $Q(\boldsymbol{\alpha})$, can be expressed as:

$$Q(\boldsymbol{\alpha}) = \prod_{i=1}^{q} \mathrm{Gam}(\alpha_i | a_{\alpha_N}, b_{\alpha_{N_i}}), \tag{74}$$

$$\text{where } a_{\alpha_N} = a_{\alpha_0} + \frac{d}{2} \tag{75}$$

$$b_{\alpha_{N_i}} = b_{\alpha_0} + \frac{1}{2} \langle ||w_i||^2 \rangle \tag{76}$$

The mode of $Q(\boldsymbol{\alpha})$:

$$mode(Q(\boldsymbol{\alpha})) = \begin{cases} 0, \text{ when } a_0 + \frac{d}{2} \leq 1 \\ \frac{a_0 + \frac{d}{2} - 1}{b_0 + \frac{1}{2}||w_i||^2}, \text{ when } a_0 + \frac{d}{2} > 1 \end{cases} \tag{77}$$

To induce sparsity in the posterior of $\boldsymbol{w}$, $Q(\boldsymbol{\alpha})$ should allow for large values of α_i when corresponding elements w_i approach zero. Thus, the mode of $Q(\boldsymbol{\alpha})$ should have chance to take a large value. Consequently, b_{α_0} need to be sufficiently small. Considering the requirement for a small variance of $P(\boldsymbol{\alpha})$, a_0 need to be comparable small as b_0^2 (Fig. S4b-c, e-f).

In conclusion, in order to achieve a broad prior for $\boldsymbol{w}$ and obtain sparsity in its posterior distribution, both a_{α_0} and that b_{α_0} should be set sufficiently small values.

A.3.2 Hyperparameter of $P(\boldsymbol{\beta})$

The prior of $\boldsymbol{\beta}$ is:

$$P(\boldsymbol{\beta}) = \prod_{i=1}^{q} \text{Gam}(\beta_i | a_{\beta_0}, b_{\beta_0}) \tag{78}$$

Using variational Bayesian method, the posterior of $\boldsymbol{\beta}$, $Q(\boldsymbol{\beta})$, is:

$$Q(\boldsymbol{\beta}) = \prod_{i=1}^{q} \text{Gam}(\beta_i | a_{\beta_N}, b_{\beta_{N_i}}), \tag{79}$$

$$\text{where } a_{\beta_N} = a_{\beta_0} + \frac{N}{2} \tag{80}$$

$$b_{\beta_{N_i}} = b_{\beta_0} + \frac{1}{2}\langle ||x_i||^2 \rangle \tag{81}$$

Similarly as the discussion of $P(\boldsymbol{\alpha})$, the conditions of the hyperparameter of $P(\boldsymbol{\beta})$ also can be constrained. Specifically, both a_{β_0} and b_{β_0} should be sufficiently small values (Fig. S5).

B Property of matrix determinant and inverse

Properties of matrix determinant and inverse used in the manuscript:

$$\begin{aligned} \frac{\partial \log |A|}{\partial x} &= tr(A^{-1}\frac{\partial A}{\partial x}) \\ \frac{\partial A^{-1}}{\partial x} &= -A^{-1}\frac{\partial A}{\partial x}A^{-1} \\ (A+B)^{-1} &= A^{-1} - \frac{1}{1+g}A^{-1}BA^{-1}, \text{ where } g = tr(BA^{-1}) \end{aligned}$$

C Variational Bayesian inference and update rules

C.1 Variational Bayesian inference

In Bayesian modeling, the objective is typically to determine the posterior distribution of the model parameters, given the observed data. However, calculating the exact posterior distribution can be infeasible for many complex models. Variational Inference tackles this challenge by approximating the true posterior distribution using a more manageable, tractable distribution, such as a Gaussian distribution. The aim is to identify the variational distribution that is closest to the true posterior distribution, as quantified by the Kullback-Leibler (KL) divergence.

The marginal log-likelihood (also known as the model evidence) is formed as:

$$\begin{aligned}\ln P(D) = \ln \int P(D,\theta)d\theta &= \ln \int Q(\theta)\frac{P(D,\theta)}{Q(\theta)}d\theta \\ &= \int Q(\theta)\ln\frac{P(D,\theta)}{Q(\theta)}d\theta + \mathrm{KL}(Q||P) \\ &\geq \int Q(\theta)\ln\frac{P(D,\theta)}{Q(\theta)}d\theta = \mathcal{L}(Q)\end{aligned}$$

where $\mathrm{KL}(Q||P)$ represents the KL divergence between the approximating $Q(\theta)$ and the posterior $P(\theta|D)$; $Q(\theta)$ is factorized over the component variables $\{\theta_i\}$ in θ, so that

$$Q(\theta) = \prod_i Q_i(\theta_i)$$

The best approximating distribution $Q(\theta)$ can be achieved by minimizing the KL divergence $\mathbf{KL}(Q||P)$, and the solution to each $Q_i(\theta_i)$ is then set as:

$$Q_i(\theta_i) = \frac{\exp\langle \ln P(D,\theta)\rangle_{k\neq i}}{\int \exp\langle \ln P(D,\theta)\rangle_{k\neq j} d\theta_j}$$

where $\langle\cdot\rangle_{k\neq i}$ denotes an expectation with respect to the distributions $\{Q_k(\theta_k)\}_{k\neq i}$.

C.2 Update rules of Variational Bayesian inference

C.2.1 Dual ARD (with individual noise variance)

In Dual ARD, the true joint distribution of data and parameters is given by

$$P(D,\theta) = \prod_{n=1}^{N} P(t_n|x_n,W,\tau)P(x_n|\beta)P(W|\alpha)P(\alpha)P(\beta)P(\tau)$$

where

$$\begin{aligned}P(t_n|x_n,W,\tau) &= \mathcal{N}(t_n|Wx_n+\mu,\mathrm{diag}(\tau)) \\ P(W|\alpha) &= \prod_{i=1}^{q}\mathcal{N}(w_i|0,\alpha_i^{-1}I) \\ P(x_n|\beta) &= \mathcal{N}(x_n|0,\mathrm{diag}(\beta)_q) \\ P(\alpha) &= \prod_{i=1}^{q}\mathrm{Gam}(\alpha_i|a_{\alpha_i},b_{\alpha_i}) \\ P(\beta) &= \prod_{i=1}^{q}\mathrm{Gam}(\beta_i|a_{\beta_i},b_{\beta_i}) \\ P(\tau) &= \prod_{i=1}^{d}\mathrm{Gam}(\tau_i|a_{\tau_k},b_{\tau_k})\end{aligned} \tag{82}$$

Based the conditional distribution of Gaussian distribution and Gamma distribution, the solution to the

component distributions of Q(·) are:

$$
\begin{aligned}
Q(X) &= \prod_{n=1}^{N} \mathcal{N}(x_n | m_x, \Sigma_x) \\
Q(W) &= \prod_{i=1}^{q} \mathcal{N}(\tilde{w}_k | m_{w_k}, \Sigma_{w_k}) \\
Q(\alpha) &= \prod_{i=1}^{q} \mathrm{Gam}(\alpha_i | a_{\alpha_N}, b_{\alpha_{N_i}}) \\
Q(\beta) &= \prod_{i=1}^{q} \mathrm{Gam}(\beta_i | a_{\beta_N}, b_{\beta_{N_i}}) \\
Q(\tau) &= \prod_{k=1}^{d} \mathrm{Gam}(\tau_k | a_{\tau_N}, b_{\tau_{N_k}})
\end{aligned}
\tag{83}
$$

where $\tilde{w}_k$ represents the k-th row of W, and

$$
\begin{aligned}
m_x &= \Sigma_x \langle W \rangle^T \ \mathrm{diag}\langle \tau \rangle (t_n - \mu) \\
\Sigma_x &= (\mathrm{diag}\langle \beta \rangle + \langle W \rangle^T \ \mathrm{diag}\langle \tau \rangle \langle W \rangle)^{-1} \\
m_{w_k} &= \langle \tau_k \rangle \Sigma_{w_k} \sum_{n=1}^{N} \langle x_n \rangle (t_{nk} - \mu_k) \\
\Sigma_{w_k} &= (\mathrm{diag}\langle \alpha \rangle + \langle \tau_k \rangle \sum_{n=1}^{N} \langle x_n x_n^T \rangle)^{-1} \\
a_{\alpha_N} &= a_{\alpha_0} + \frac{d}{2} \\
b_{\alpha_{N_i}} &= b_{\alpha_0} + \frac{1}{2} \langle ||w_i||^2 \rangle \\
a_{\beta_N} &= a_{\beta_0} + \frac{N}{2} \\
b_{\beta_{N_i}} &= b_{\beta_0} + \frac{1}{2} \langle ||x_i||^2 \rangle \\
a_{\tau_N} &= a_{\tau_0} + \frac{N}{2} \\
b_{\tau_{N_k}} &= b_{\tau_0} + \frac{1}{2} \sum_{n=1}^{N} (t_{nk} - \langle w_k \rangle \langle x_n \rangle - \mu_k)^2
\end{aligned}
\tag{84}
$$

C.2.2 variant Dual ARD variant (with common noise variance)

In the variant Dual ARD with common noise term, the true joint distribution of data and parameters is given by

$$
P(D, \theta) = \prod_{n=1}^{N} P(t_n | x_n, W, \tau) P(x_n | \beta) P(W | \alpha) P(\alpha) P(\beta) P(\tau)
$$

where

$$
\begin{aligned}
P(t_n | x_n, W, \tau) &= \mathcal{N}(t_n | W x_n + \mu, \mathrm{diag}(\tau)) \\
P(W | \alpha) &= \prod_{i=1}^{q} \mathcal{N}(w_i | 0, \alpha_i^{-1} I) \\
P(x_n | \beta) &= \mathcal{N}(x_n | 0, \mathrm{diag}(\beta)_q) \\
P(\alpha) &= \prod_{i=1}^{q} \mathrm{Gam}(\alpha_i | a_{\alpha_i}, b_{\alpha_i}) \\
P(\beta) &= \prod_{i=1}^{q} \mathrm{Gam}(\beta_i | a_{\beta_i}, b_{\beta_i}) \\
P(\tau) &= \mathrm{Gam}(\tau | a_\tau, b_\tau)
\end{aligned}
\tag{85}
$$

We assume a Q distribution of the form:

$$Q(X, W, \alpha, \beta, \tau) = Q(X)Q(W)Q(\alpha)Q(\beta)Q(\tau)$$

Based the conditional distribution of Gaussian distribution and Gamma distribution, the solution to the component distributions of Q($\cdot$) are:

$$\begin{aligned}
Q(X) &= \prod_{n=1}^{N} \mathcal{N}(x_n | m_{x_n}, \Sigma_x) \\
Q(W) &= \prod_{k=1}^{d} \mathcal{N}(\tilde{w}_k | m_{w_k}, \Sigma_w) \\
Q(\alpha) &= \prod_{i=1}^{q} \mathrm{Gam}(\alpha_i | a_{\alpha_N}, b_{\alpha_{N_i}}) \\
Q(\beta) &= \prod_{i=1}^{q} \mathrm{Gam}(\beta_i | a_{\beta_N}, b_{\beta_{N_i}}) \\
Q(\tau) &= \mathrm{Gam}(\tau | a_{\tau_N}, b_{\tau_N})
\end{aligned} \tag{86}$$

where $\tilde{w}_k$ represents the k-th row of W, and

$$\begin{aligned}
m_{x_n} &= \langle \tau \rangle \Sigma_x \langle W \rangle^T (t_n - \mu) \\
\Sigma_x &= (\mathrm{diag}\langle \beta \rangle + \langle \tau \rangle \langle W^T W \rangle)^{-1} \\
m_{w_k} &= \langle \tau \rangle \Sigma_w \sum_{n=1}^{N} \langle x_n \rangle (t_{nk} - \mu_k) \\
\Sigma_w &= (\mathrm{diag}\langle \alpha \rangle + \langle \tau \rangle \sum_{n=1}^{N} \langle x_n x_n^T \rangle)^{-1} \\
a_{\alpha_N} &= a_{\alpha_0} + \frac{d}{2} \\
b_{\alpha_{N_i}} &= b_{\alpha_0} + \frac{1}{2} \sum_{k=1}^{d} \langle w_{ik}^2 \rangle = b_{\alpha_0} + \frac{1}{2} \langle ||w_i||^2 \rangle \\
a_{\beta_N} &= a_{\beta_0} + \frac{N}{2} \\
b_{\beta_{N_i}} &= b_{\beta_0} + \frac{1}{2} \sum_{n=1}^{N} \langle x_{ni} \rangle^2 = b_0 \beta_0 + \frac{1}{2} \langle ||x_i||^2 \rangle \\
a_{\tau_N} &= a_{\tau_0} + \frac{Nd}{2} \\
b_{\tau_N} &= b_{\tau_0} + \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{d} (t_{nk} - \langle \tilde{w}_k \rangle \langle x_{nk} \rangle - \mu_k)^2
\end{aligned} \tag{87}$$

C.2.3 Bayesian PCA

In Bayesian PCA, the true joint distribution of data and parameters is given by

$$P(D, \theta) = \prod_{n=1}^{N} P(t_n | x_n, W, \mu, \tau) P(X) P(W | \alpha) P(\alpha) P(\mu) P(\tau)$$

where

$$
\begin{aligned}
P(t_n|x_n, W, \tau) &= \mathcal{N}(t_n|Wx_n + \mu, \tau^{-1}I_d) \\
P(W|\alpha) &= \prod_{i=1}^{q} \mathcal{N}(w_i|0, \alpha_i^{-1}I_d) \\
P(x_n) &= \mathcal{N}(x_n|0, I_q) \\
P(\alpha) &= \prod_{i=1}^{q} \mathrm{Gam}(\alpha_i|a_\alpha, b_\alpha) \\
P(\tau) &= \mathrm{Gam}(\tau|a_\tau, b_\tau)
\end{aligned}
\tag{88}
$$

We assume a Q distribution of the form:

$$
Q(X, W, \alpha, \mu, \tau) = Q(X)Q(W)Q(\alpha)Q(\tau)
$$

Based the conditional distribution of Gaussian distribution and Gamma distribution, the solution to the component distributions of Q($\cdot$) are:

$$
\begin{aligned}
Q(X) &= \prod_{n=1}^{N} \mathcal{N}(x_n|m_{x_n}, \Sigma_x) \\
Q(W) &= \prod_{k=1}^{d} \mathcal{N}(\tilde{w}_k|m_{w_k}, \Sigma_w) \\
Q(\alpha) &= \prod_{i=1}^{q} \mathrm{Gam}(\alpha_i|a_{\alpha_N}, b_{\alpha_{N_i}}) \\
Q(\tau) &= \mathrm{Gam}(\tau|a_{\tau_N}, b_{\tau_N})
\end{aligned}
\tag{89}
$$

where $\tilde{w}_k$ represents the k-th row of W, and

$$
\begin{aligned}
m_{x_n} &= \langle\tau\rangle\Sigma_x\langle W\rangle^T(t_n - \mu) \\
\Sigma_x &= (I + \langle\tau\rangle\langle W^T W\rangle)^{-1} \\
m_{w_k} &= \langle\tau\rangle\Sigma_w \sum_{n=1}^{N}\langle x_n\rangle(t_{nk} - \mu_k) \\
\Sigma_w &= (\mathrm{diag}\langle\alpha\rangle + \langle\tau\rangle \sum_{n=1}^{N}\langle x_n x_n^T\rangle)^{-1} \\
a_{\alpha_N} &= a_{\alpha_0} + \frac{d}{2} \\
b_{\alpha_{N_i}} &= b_{\alpha_0} + \frac{1}{2}\sum_{k=1}^{d}\langle w_{ik}^2\rangle = b_{\alpha_0} + \frac{1}{2}\langle||w_i||^2\rangle \\
a_{\tau_N} &= a_{\tau_0} + \frac{Nd}{2} \\
b_{\tau_N} &= b_{\tau_0} + \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{d}(t_{nk} - \langle\tilde{w}_k\rangle\langle x_{nk}\rangle - \mu_k)^2
\end{aligned}
\tag{90}
$$

C.2.4 Bayesian PCA variant (with individual noise variance)

In factor analysis with ARD (VBFA), the true joint distribution of data and parameters is given by

$$
P(D, \theta) = \prod_{n=1}^{N} P(t_n|x_n, W, \mu, \tau)P(X)P(W|\alpha)P(\alpha)P(\mu)P(\tau)
$$

where

$$
\begin{aligned}
P(t_n|x_n, W, \tau) &= \mathcal{N}(t_n|Wx_n + \mu, \tau^{-1}I_d) \\
P(W|\alpha) &= \prod_{i=1}^{q} \mathcal{N}(w_i|0, \alpha_i^{-1}I_d) \\
P(x_n) &= \mathcal{N}(x_n|0, I_q) \\
P(\alpha) &= \prod_{i=1}^{q} \mathrm{Gam}(\alpha_i|a_\alpha, b_\alpha) \\
P(\tau) &= \prod_{i=1}^{d} \mathrm{Gam}(\tau_i|a_\tau, b_\tau)
\end{aligned}
\tag{91}
$$

We assume a Q distribution of the form:

$$
Q(X, W, \alpha, \mu, \tau) = Q(X)Q(W)Q(\alpha)Q(\tau)
$$

Based the conditional distribution of Gaussian distribution and Gamma distribution, the solution to the component distributions of Q($\cdot$) are:

$$
\begin{aligned}
Q(X) &= \prod_{n=1}^{N} \mathcal{N}(x_n|m_{x_n}, \Sigma_x) \\
Q(W) &= \prod_{k=1}^{d} \mathcal{N}(\tilde{w}_k|m_{w_k}, \Sigma_w) \\
Q(\alpha) &= \prod_{i=1}^{q} \mathrm{Gam}(\alpha_i|a_{\alpha_N}, b_{\alpha_{N_i}}) \\
Q(\tau) &= \prod_{k=1}^{d} \mathrm{Gam}(\tau_k|a_{\tau_N}, b_{\tau_{N_k}})
\end{aligned}
\tag{92}
$$

where $\tilde{w}_k$ represents the k-th row of W, and

$$
\begin{aligned}
m_{x_n} &= \langle\tau\rangle\Sigma_x\langle W\rangle^T(t_n - \mu) \\
\Sigma_x &= (I + \langle\tau\rangle\langle W^T W\rangle)^{-1} \\
m_{w_k} &= \langle\tau\rangle\Sigma_w \sum_{n=1}^{N}\langle x_n\rangle(t_{nk} - \mu_k) \\
\Sigma_w &= (\mathrm{diag}\langle\alpha\rangle + \langle\tau\rangle\sum_{n=1}^{N}\langle x_n x_n^T\rangle)^{-1} \\
a_{\alpha_N} &= a_{\alpha_0} + \frac{d}{2} \\
b_{\alpha_{N_i}} &= b_{\alpha_0} + \frac{1}{2}\sum_{k=1}^{d}\langle w_{ik}^2\rangle = b_{\alpha_0} + \frac{1}{2}\langle||w_i||^2\rangle \\
a_{\tau_N} &= a_{\tau_0} + \frac{N}{2} \\
b_{\tau_{N_k}} &= b_{\tau_0} + \frac{1}{2}\sum_{n=1}^{N}(t_{nk} - \langle\tilde{w}_k\rangle\langle x_{nk}\rangle - \mu_k)^2
\end{aligned}
\tag{93}
$$

Table S1: Scanning depth

Scanning depth	#neurons
278μm	391
300μm	384
332μm	370
360μm	325
394μm	277
426μm	245
470μm	189
496μm	236

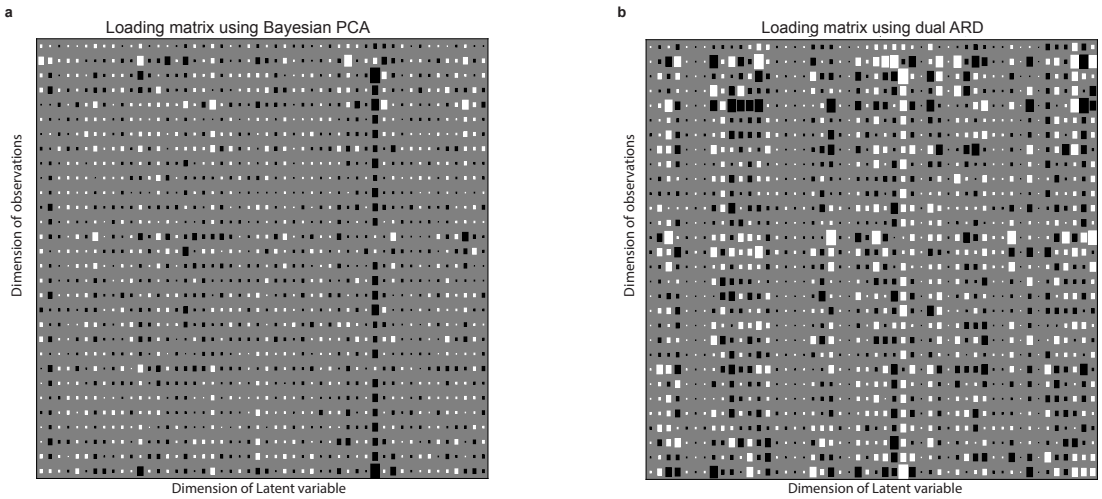

Figure S1: Hinton diagrams of the loading matrix for a two-photon calcium imaging dataset as used in section :
(a) Loading matrix for 30 example observations using Bayesian PCA, with no significant suppression of latent dimensions (columns).
(b) Loading matrix for 30 example observations using Dual ARD, demonstrating clear suppression of certain latent dimensions (columns).

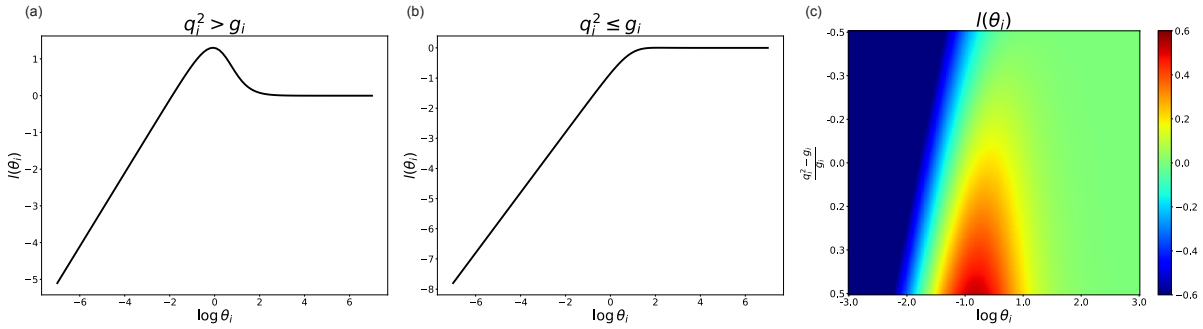

Figure S2: $L(\theta_i)$ under different conditions ($\theta_i \in (\alpha_i, \beta_i)$):
(a-b) Reproducing Fig. 1 in [24]. (a)$l(\theta_i)$ reaches its peak vaule at a finite value when $q_i^2 > g_i$; (b)The maximum of $l(\theta_i)$ is found at $\theta_i = \infty$ when $q_i^2 < g_i$.
(c) The stationary point of $l(\theta_i)$ changes depending on the ratio between q_i^2 and g_i.

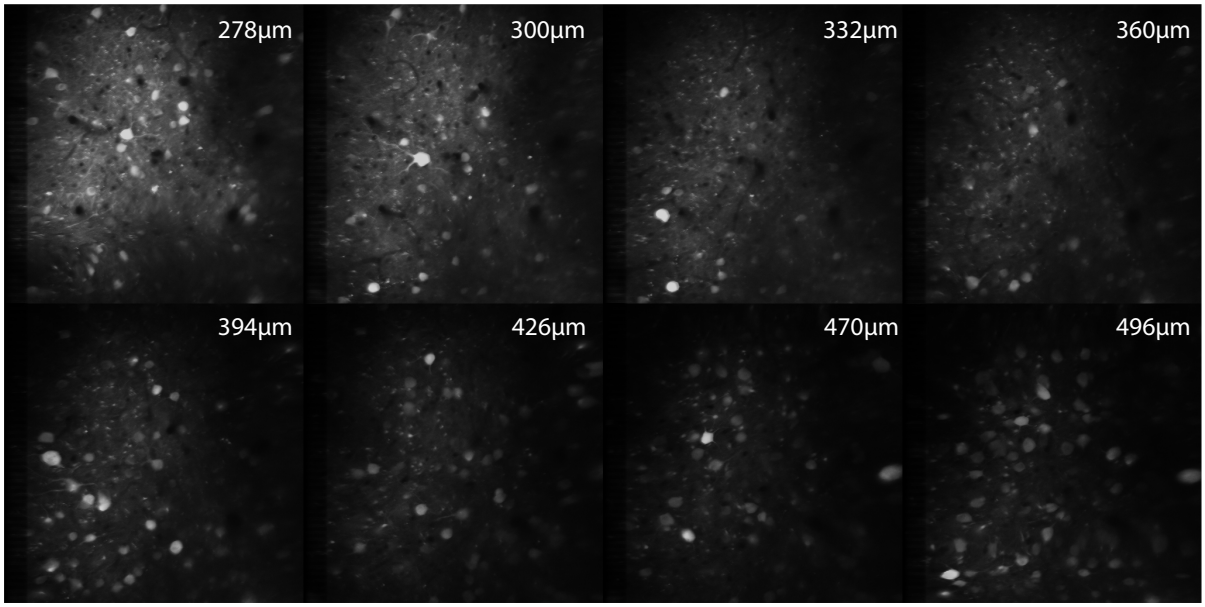

Figure S3: Piezo scanning of two-Photon calcium imaging across 8 cortical depths in the PPC.

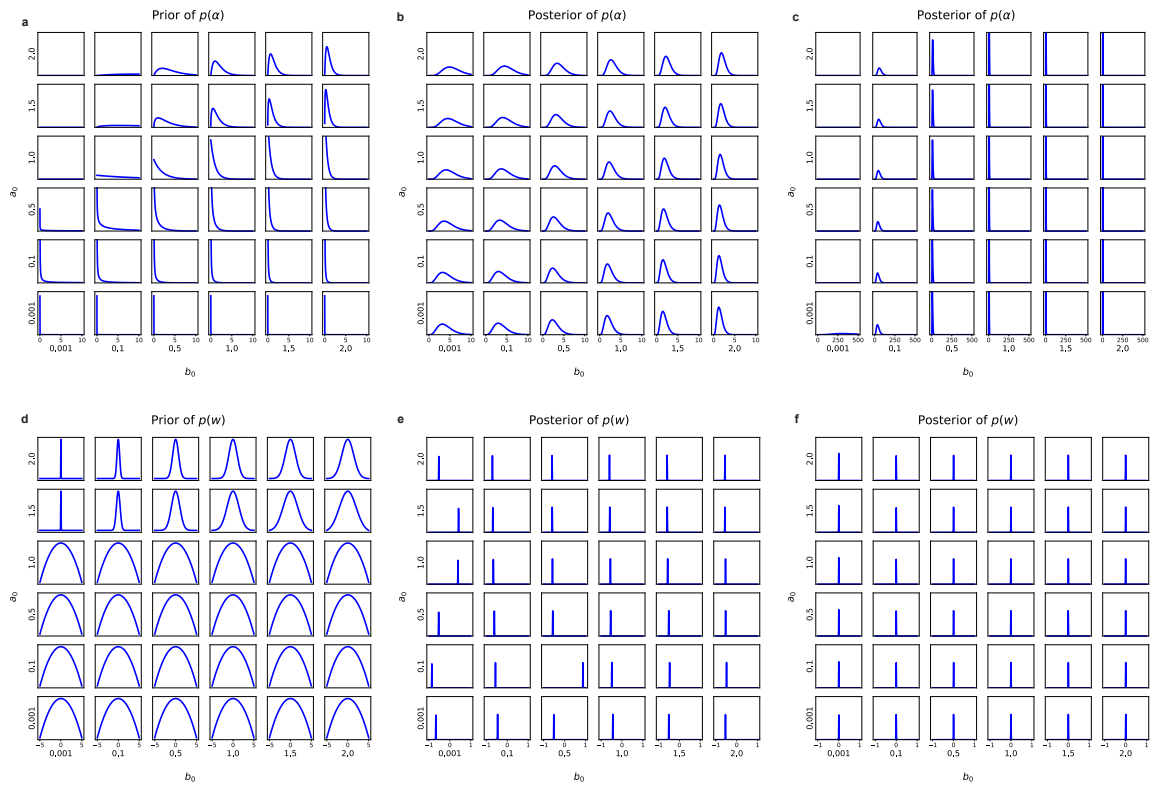

Figure S4: $P(\boldsymbol{\alpha})$ under various hyperparameter conditions. (Data used is the same as in Fig.2)
(a) Prior of $P(\boldsymbol{\alpha})$ when the hyperparameters a_0 and b_0 take different values.
(b-c) Posterior of $P(\alpha_i)$ for high variance x_i and low variance x_i, respectively, given the conditions for a_0 and b_0 in (a).
(d) Prior of $p(\boldsymbol{w})$ when $p(\boldsymbol{\alpha})$ takes different condition in (a).
(e-f) Corresponding posterior of high variance $p(w_i)$ and and low variance $p(w_i)$, respectively.

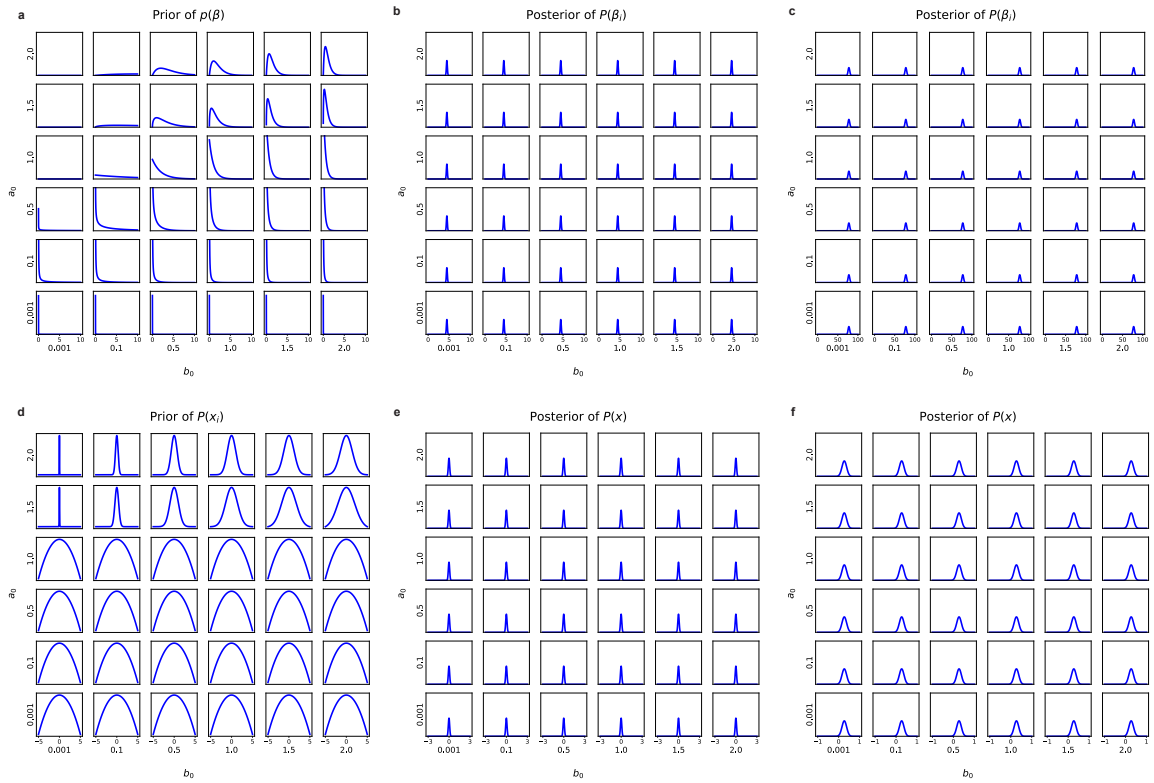

Figure S5: $P(\boldsymbol{\beta})$ under various hyperparameter conditions. (Data used is the same as in Fig.2)
(a) Prior of $P(\boldsymbol{\beta})$ when the hyperparameters a_0 and b_0 take different values.
(b-c) Posterior of $P(\beta_i)$ for high variance x_i and low variance x_i, respectively, given the conditions for a_0 and b_0 in (a)
(d) Prior of $p(\boldsymbol{x})$ when $p(\boldsymbol{\beta})$ takes different condition in (a).
(e-f) Corresponding posterior of high variance $p(x_i)$ and and low variance $p(x_i)$, respectively.

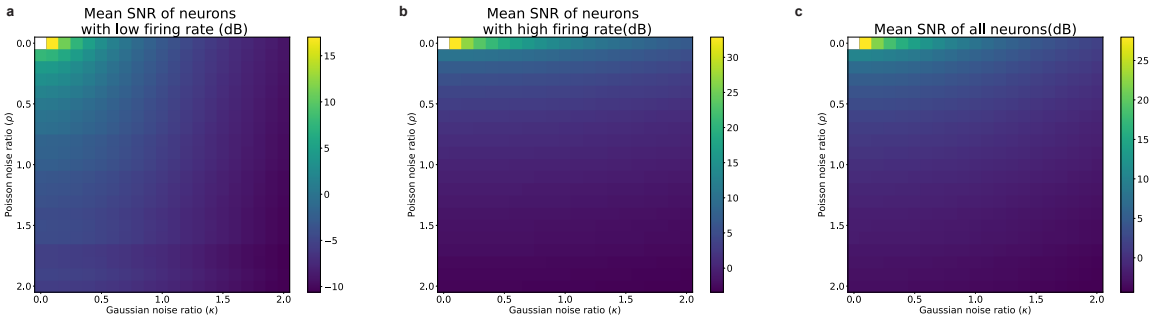

Figure S6: Signal-to-noise ratio (SNR) of simulated calcium data with different noise ratios.
The signal-to-noise ratio was measured by $\mathrm{SNR}_{dB} = 10 \log_{10} \frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}$, where the P_{signal} and P_{noise} represent the power of the signal and noise, respectively, calculated as the mean of their squared sampled values.
(a) Mean SNR of all neurons.
(b) Mean SNR for neurons with high firing rates.
(c) Mean SNR for neurons with low firing rates.

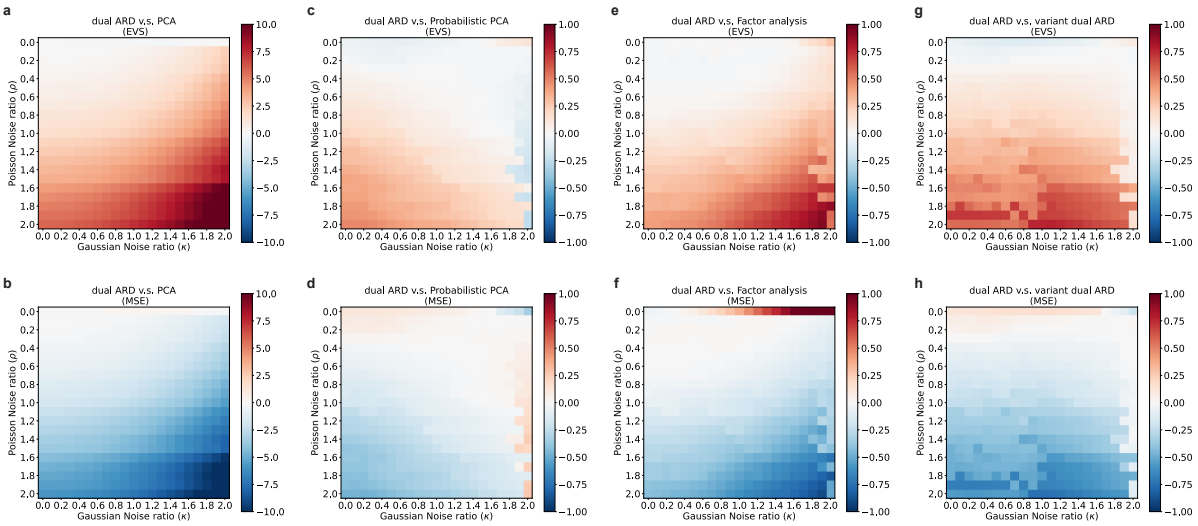

Figure S7: Comparing the performance of Dual ARD with PCA, probabilistic PCA, factor analysis, and variant Dual ARD (with common noise variance) when applied to simulated calcium imaging data featuring varying Poisson and Gaussian noise ratios:
(a-g) Model performance difference in terms of explained variance scores.
(h-n) Model performance difference in terms of mean squared error.

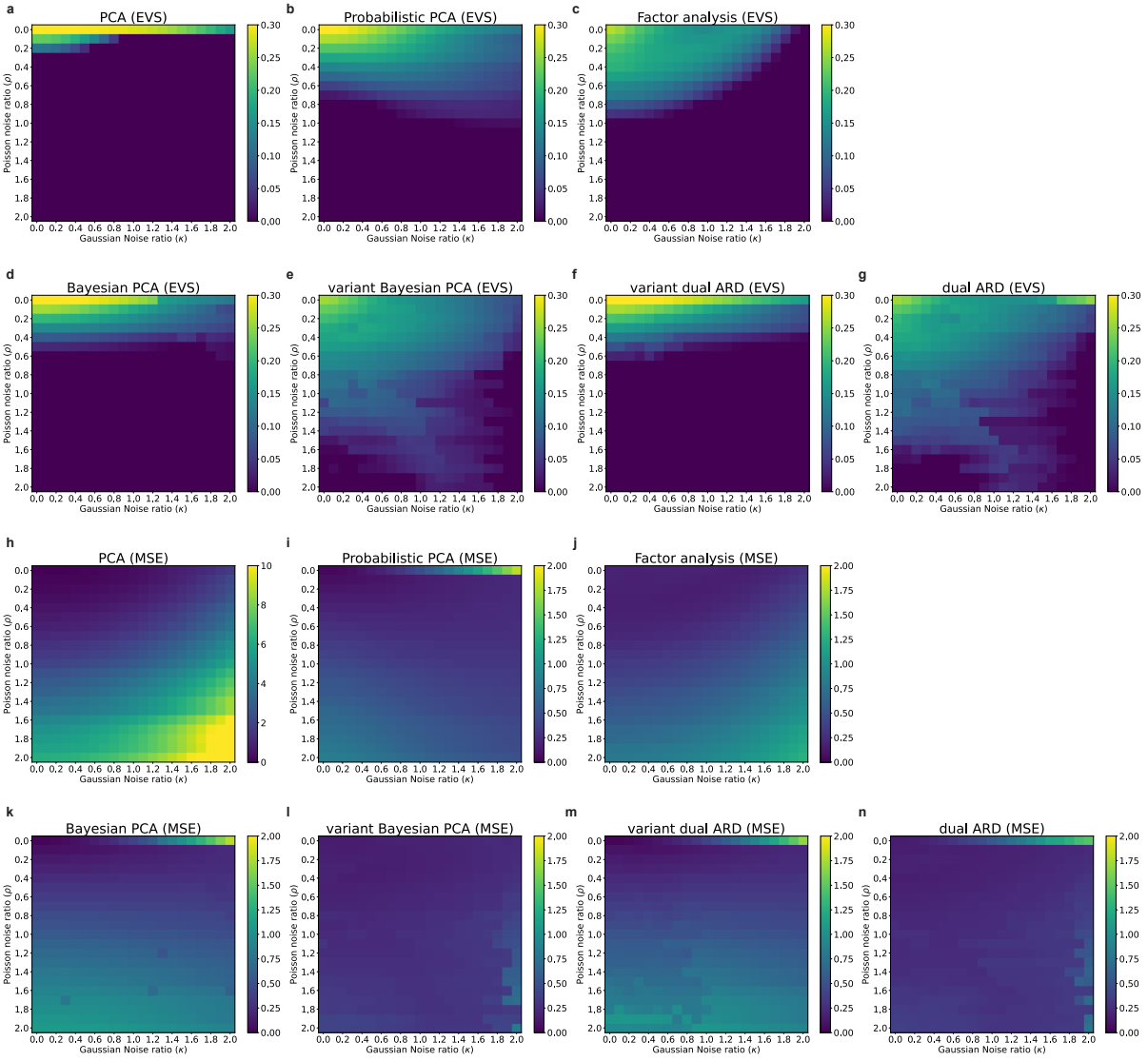

Figure S8: Explained variance scores (EVS) and Mean squared error scores (MSE) of different models with varying Poisson and Gaussian noise ratios.
(a-g) Explained variance scores of models on simulated calcium imaging data with varying Poisson and Gaussian noise ratios. The models include PCA, Probabilistic PCA, Factor Analysis, Bayesian PCA, variant Bayesian PCA, variant Dual ARD, and Dual ARD, respectively.
(h-n) Mean squared error scores of the aforementioned models on simulated calcium imaging data with varying Poisson and Gaussian noise ratios, respectively.

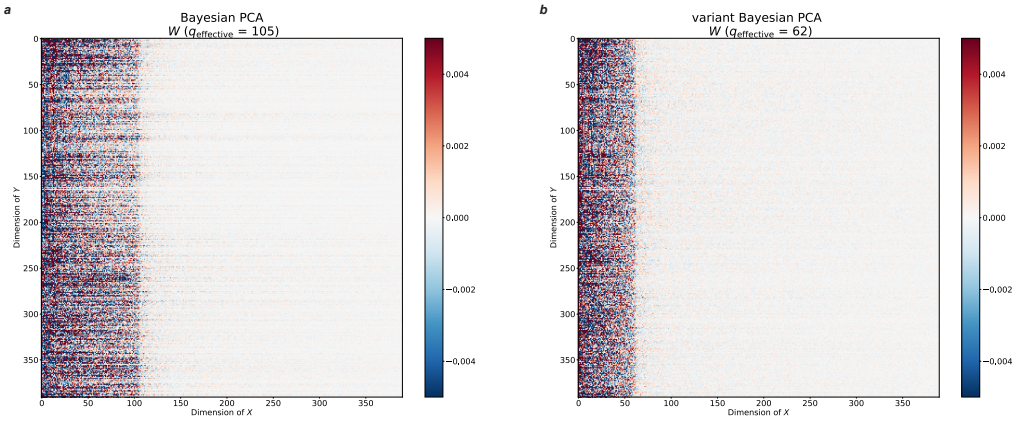

Figure S9: Maximum effective dimensions using Bayesian PCA and variant Bayesian PCA.
(a) The loading matrix (W) using Bayesian PCA with setting $q_{\max} = d-1$, where d is the dimensionality of the observation y. The resulting effective dimensionality of the latent variable is $q_{\text{effective}} = 105$.
(b) The loading matrix (W) achieved using variant Bayesian PCA with setting $q_{\max} = d - 1$. The resulting effective dimensionality of the latent variable is $q_{\text{effective}} = 62$.

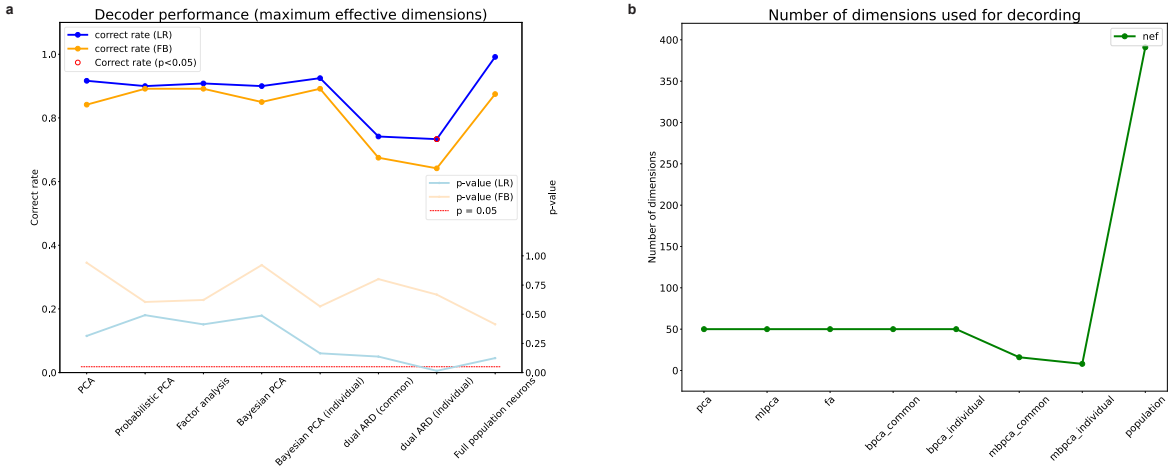

Figure S10: Decoder using maximum effective dimensions of each model.
(a) Decoding performance of all models: the dark blue curve represents the correct prediction rate for left-right categorization, while the dark orange curve indicates the correct prediction rate for front-back categorization. The light blue and light orange curves display the p-values for left-right and front-back categorizations, respectively. The red circles represent the correct rate for instances where $p < 0.05$, and the red dotted line signifies the threshold of p-value for $p = 0.05$.
(b) The effective dimensions of the latent variable obtained used in the decoder: $D_{\text{effective}} = 50$ for PCA, Probabilistic PCA, factor analysis, Bayesian PCA, and variant Bayesian PCA; $D_{\text{effective}} = 12$ for Dual ARD variant with common noise; $D_{\text{effective}} = 8$ for Dual ARD with individual noise; $D_{\text{effective}} = 391$ for full population neurons .

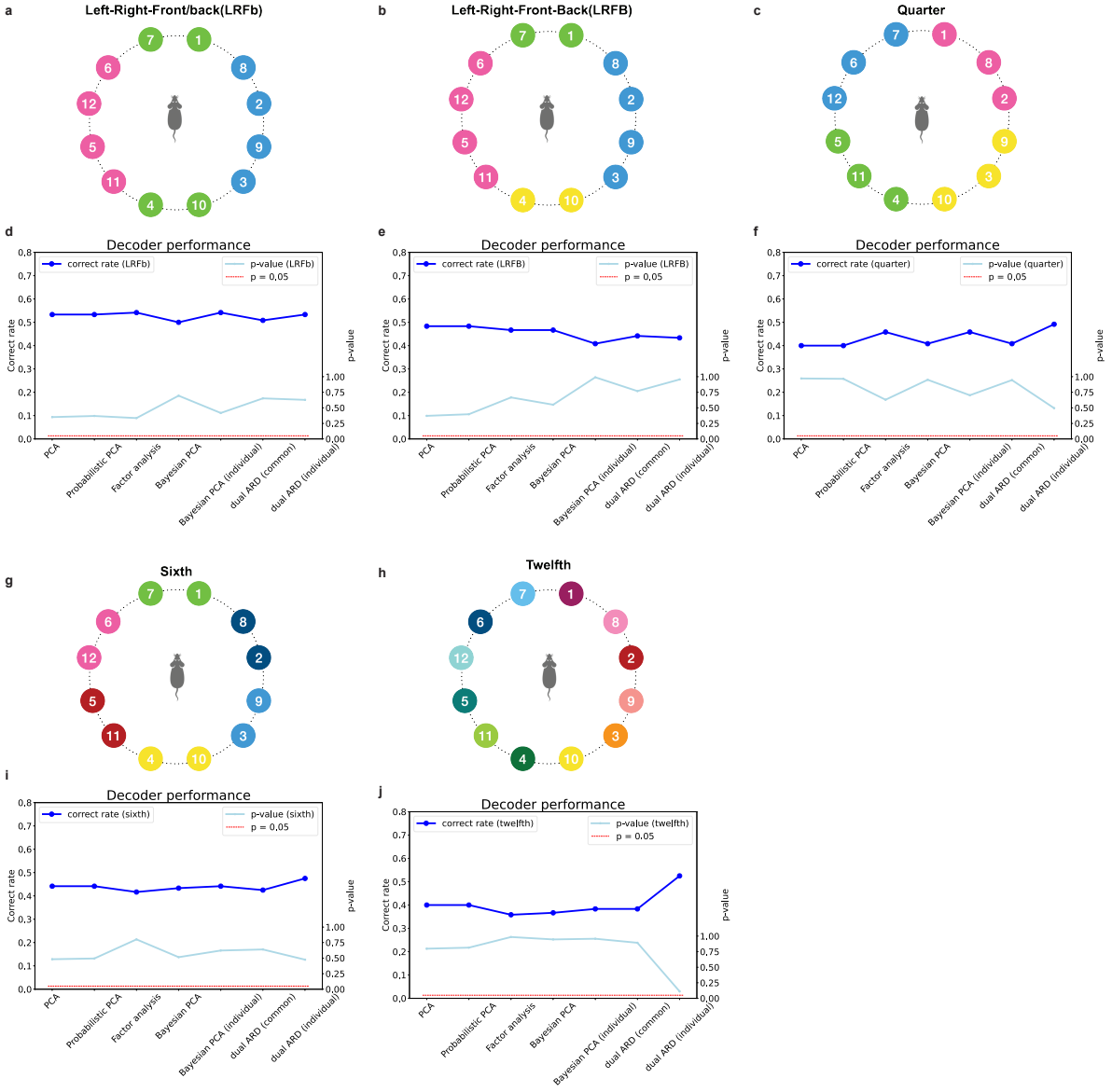

Figure S11: Decoder performance for alternative categorizations of speaker locations.
(a-c, g-h) Layout of alternative location categorizations. (d-f, i-j) Decoder performance for the corresponding categorizations. The dark blue curve represents the correct rate, the light blue curve denotes the p-value, and the red dotted line indicates the p-value threshold of p = 0.05.

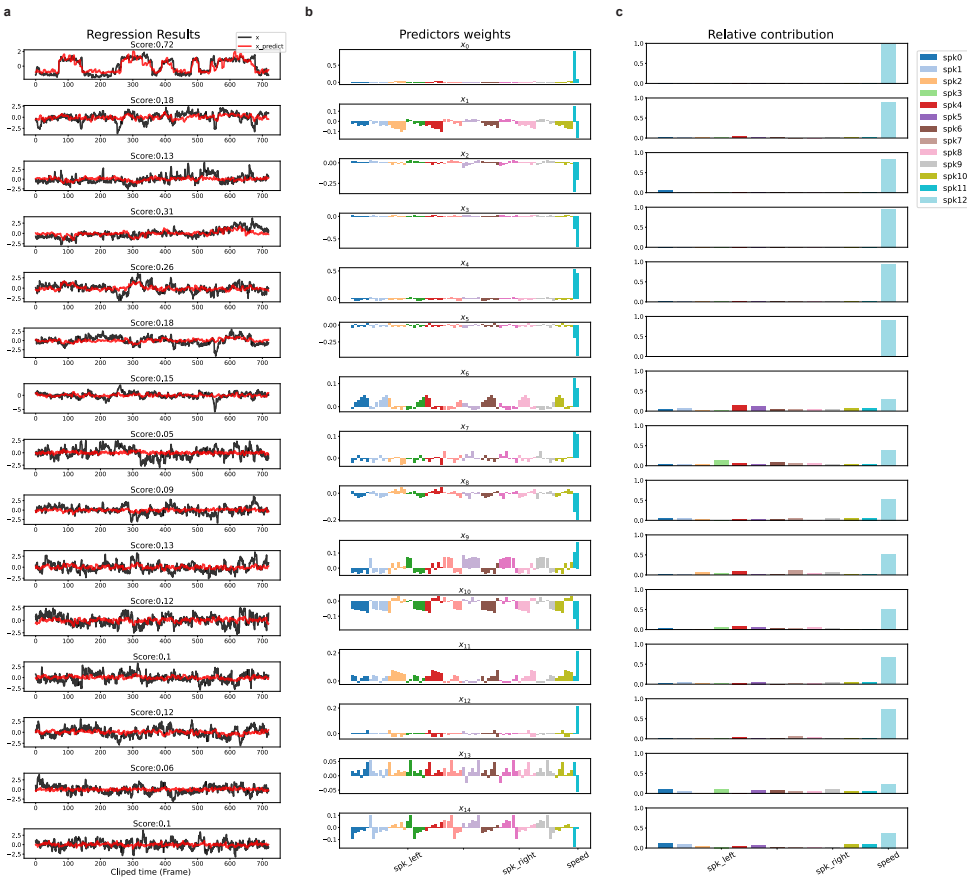

Figure S12: Encoding Results, Parameters, and Evaluations.
(a) Encoding results of each latent variable by behavioral variables.
(b) Predictor weights of all predictors in the encoding model.
(c) Relative contribution of each predictor in encoding each latent variable.