1 Results

1.1 Response Properties from symmetrical Recurrent Interaction Networks in Correlation with Feedforward Recurrent Alignment

The feedforward recurrent alignment, defined in the method under section **??** quantifies the degree of how much the feedforward input is aligned with the direction that is spanned by eigenvectors of the recurrent network. If the input is well aligned with the eigenvector corresponding to the maximal eigenvalue of the recurrent interaction J, the random noise in the response would be suppressed by the evoked response due to selective amplification. Response amplification determines the steady-state response,

$$r^* = \sum_{i=1}^{n} \frac{(e_i \cdot h) e_i}{1 - \lambda_i}\,, \tag{1.1}$$

where $\cdot$ denotes the dot product between two vectors.

The steady state response is then dominated by the projection of the input vector h along the axis defined by the eigenvector $e_{\max}$ whose eigenvalue $\lambda_{\max}$ is maximal and near one [**?**].

$$r^* \approx \frac{(e_{\max} \cdot h) e_{\max}}{1 - \lambda_{\max}} \tag{1.2}$$

The projection of input h on eigenvector $e_{\max}$ reaches its maximal when h is approximately $e_{\max}$ itself. Therefore, the steady state responses reaches its maximum and the feedforward recurrent alignment equals the maximal eigenvalue $\lambda_{\max}$ of J because of eq.(**??**).

On the other hand, if the input is not well aligned with the dominant eigenvectors, the random noise is large relative to the response. For an extreme example, if the input is aligned with the eigenvector $e_{\min}$ with minimal eigenvalue $\lambda_{\min}$, the response will almost not contribute to the steady state response at all due to the response amplification eq.(1.1). This could also be reflected by the feedforward recurrent alignment, which equals $\lambda_{\min}$ in this case because of eq.(**??**).

We can thus conclude that the feedforward recurrent alignment eq.(**??**) reflects the alignment between the feedforward input and the dominant eigenvectors of the recurrent interaction network J. The better the input h is aligned to the dominant eigenvectors of J, the stronger and more reliable the response, and at the same time the higher the feedforward recurrent alignment score.

In the following sections, the four properties introduced at section **??** will be evaluated in the model and compare to the tendency observed in [**?**].

1.1.1 Trial-to-Trial Correlation increases with larger Alignment

As defined under section **??** in paragraph "Trial-to-trial correlation", the inputs are constructed by multivariate normal distribution eq.(**??**). When aligning the input h with eigenvectors e_i, e_i are the the mean vector for input distribution as defined in eq.(**??**),

$$h_i \sim \mathcal{N}(e_i, \sigma^2_{\text{trial}} I_n)\,. \tag{1.3}$$

The feedforward recurrent alignment score is then determined by corresponding eigenvalue λ_i due to eq.(**??**), and reaches its maximal when align input h to e_{max}.

We want to find out the correlation between the feedforward recurrent and the trial-to-trial correlation β_s defined by eq.(**??**). If sorting the eigenvectors in the order such that their corresponding eigenvalues are in ascending order,

$$e_{\text{min}}, ..., e_i, e_j, ..., e_{\text{max}} \text{ such that } \lambda_{\text{min}} < ... < \lambda_i < \lambda_j < ... < \lambda_{\text{max}}\,, \tag{1.4}$$

with λ_i the corresponding eigenvalue for eigenvector e_i. The inputs that aligned with eigenvectors in this order should have monotonously increasing feedforward recurrent alignments.

Generating the results with eq.(**??**) for N trials. The trial-to-trial correlation can be calculated with eq.(**??**).

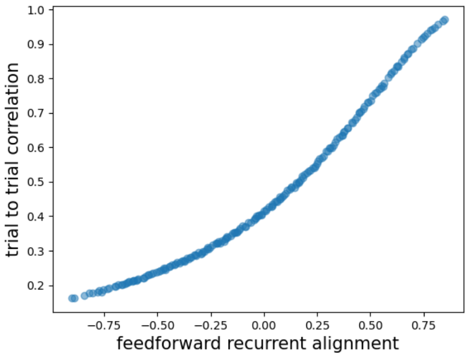


Figure 1.1 Correlation between feedforward recurrent alignment and trial to trial correlation for symmetric RNNs. Inputs aligned to eigenvectors e_i of interaction matrix J in the ascending order of eigenvalues eq.(1.4), resulting the feedforward recurrent alignment varies approximately between λ_{min} and λ_{max}. For each input alignment to an eigenvector, $N = 100$ trials of evoked responses were generated for calculation of the trial-to-trial correlation calculated with eq.(**??**).

We assume feedforward recurrent alignment for visually naive cortex equals zero, which could be interpreted as responses evoked by random inputs. The trial-to-trial correlation with random inputs is smaller than it in the case, when the input is aligned to e_{max}. This coincides with the experimental observations of responses from visually naive and experienced primary visual cortex of ferrets [**?**].

Moreover, the modeling result Figure 1.1 suggests a positive correlation between the feedforward recurrent alignment and the trial-to-trial correlation over the whole alignment range. The result confirms the idea that with the feedforward inputs more and more aligned with the dominant eigenvectors of the recurrent network, the stability between trials increases also simultaneously. The process of reaching higher trial-to-trial stability is therefore a process of becoming more aligned with the dominant eigenvector. This positive correlation also coincides with the experimental results from ferrets [**?**].

1.1.2 Intra-Trial Stability increases with larger Alignment

Now we want to see if our modeling could capture the change in intra-trial stability during the development observed in the primary visual cortex of ferrets [**?**]. The intra-trial stability increased after the eye-opening and a couple of days. The feedforward recurrent alignment hypothesis suggests that the visually experienced cortex should have a better alignment between feedforward inputs and the dominant modes in the recurrent network. To confirm this idea, we would expect the intra-trial stability would be larger with a higher feedforward recurrent alignment score.

Analogous to trial-to-trial correlation, the eigenvectors can be sorted in descending order according to the eigenvalues eq.(1.4). The intra-trial stability is calculated with eq.(**??**).

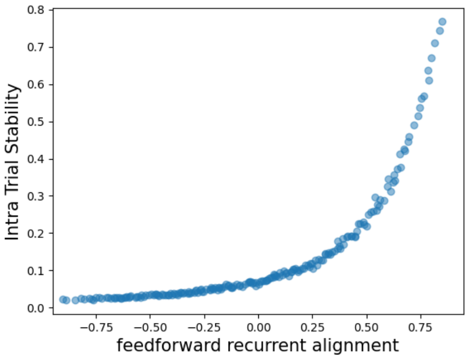


Figure 1.2 Correlation between feedforward recurrent alignment and intra-trial stability for symmetric RNNs. Inputs aligned to eigenvectors e_i of interaction matrix J sorted according to the ascending order of eigenvalues eq.(1.4), resulting the feedforward recurrent alignment varies approximately between λ_{min} and λ_{max}. For one input aligned to an eigenvector, the intra-trial stability is calculated with the evoked steady-state response eq.(**??**).

The result (Figure 1.2) indicates a positive correlation between the feedforward recurrent alignment and the intra-trial stability. With random feedforward inputs, the feedforward recurrent alignment takes the value near zero. So the eye-opening happens somewhere between feedforward recurrent equals zero and reaches maximal. Thus, before the eye-opening, there is already a certain alignment that leads to a certain degree of intra-trial correlation.

Furthermore, the correlation is almost exponential. So, the enhancement of the input alignment to the dominant eigenvector is more rapid after the eye-opening than before. One assumption for this phenomenon could be that after the eye-opening, the environment provides more training data for the network so that the alignment between inputs and the dominant eigenvector could be improved more efficiently. As a result, the responses get more intense and drive a better alignment forward. A positive loop could arise and speed up until the optimum is reached.

1.1.3 Dimensionality decreases with larger Alignment

Dimensionality is a generally important property of neural representations and could help to understand processes for example in learning and controlling [?, ?]. A low-dimensional representation will encode a diverse range of inputs into a small set of common, orthogonal activity patterns. In other words, low-dimensional activity patterns require a small number of basis vectors from the response space to represent themselves. On the contrary, a high-dimensional representation will separate even similar inputs into orthogonal activity patterns. Compared to low-dimensional representation, a high-dimensional activity pattern is represented through a large set of basis vectors [?].

Therefore, we want to take a look at the change of dimensionality during the increase of alignment in the modeling. In ferrets' primary visual cortex, the dimensionality decreased from days before eye opening until days after eye-opening [?]. We would then expect that the model should also suggest a decrease in dimensionality with an increase in alignment between inputs and dominant eigenvectors of recurrent networks.

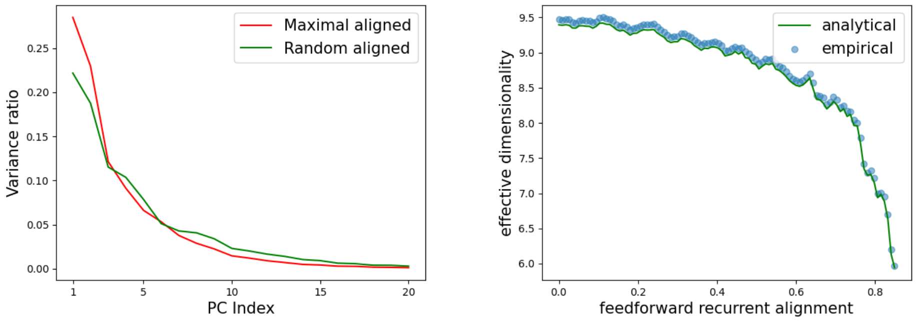
For a certain alignment, the principal component analysis reflects directly the dimensionality of the evoked activity pattern under this alignment. Because the principal components are the eigenvectors of response covariance, they also build up a set of basis vectors for the activity pattern space. Variance ratios reflect the weight that each principal component takes to represent the activity pattern. Thus, if only a small number of principal components contribute the most, the activity pattern is then low dimensional. If a broad set of principal components are similarly important, the activity pattern is highly dimensional.

With the idea of obtaining the linear dimensionality defined as participation ratio based on the principal component analysis, the eigenvectors here are ordered in descending order,

$$e_{\max}, ..., e_i, e_j, ..., e_{\min} \text{ such that } \lambda_{\max} > ... > \lambda_i > \lambda_j > ... > \lambda_{\min}\,. \tag{1.5}$$

For the generation of inputs with covariance matrix Σ^{Dim} defined in eq.(**??**), a subset of eigenvectors $\{e_i\}_{i=L,...,L+M_{\text{dim}}}$ will be chosen for $L = 1, ..., \frac{n}{2}$. M_{dim} then determines how many eigenvectors will contribute to generating inputs and evoked activity. In each such subset of eigenvectors, the leading eigenvector is e_L. Approximate here the feedforward recurrent alignment with the leading eigenvector only. Since L is considered only in the range of the first half of eigenvectors ordered as eq.(1.5), the range of feedforward recurrent alignment is between around 0 and $\lambda_{\max}$.

The linear dimensionality analytically and empirically will be calculated according to eq.(**??**) and eq.(**??**).



(a) **(b)**

Figure 1.3 The correlation between dimensionality and feedforward recurrent alignment for symmetric RNNs. Prior experimental observations suggested that the dimensionality decreases from prior until post eye-opening [**?**]. With the feedforward recurrent alignment hypothesis, the dimensionality property of the neural representation during development could be captured. **(a)** Principal component analysis for the evoked activity under input aligned to e_{max} and spontaneous random activity. The red line is for maximal alignment and the green line is for spontaneous alignment. **(b)** Achieve the correlation between dimensionality and feedforward recurrent alignment with analytically eq.(**??**) and empirically eq.(**??**). The green line displays the analytical approximation for dimensionality and the blue dots for empirical.

As explained above, the principal component analysis helps to decode the dimensionality. The curve in Figure 1.3a of the variance ratio reflects the dimensionality. Spontaneous random alignment has a flatter variance ratio curve, indicating a broader range of eigenvector contributions. Therefore, the spontaneous random alignment has a higher dimensionality than under alignment with e_{max}.

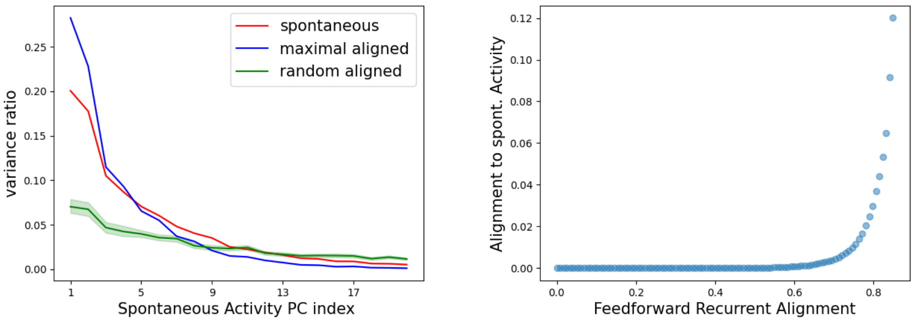
In total, a negative correlation between the dimensionality and feedforward recurrent alignment is shown in Figure 1.3b. The correlation forms nearly a flipped logarithmic function. The error between analytical and empirical results is small, which confirmed the good approximation formulated analytically by eq.(**??**).

Besides, the flipped logarithmic correlation suggests the a similar principle for intra-trial stability (Figure 1.2). After eye-opening, the reduction of dimensionality becomes larger when the feedforward inputs align better with the dominant eigenvector. It could be the case, that after the eye-opening, the recurrent network tries to encode the environment information with a large number of eigenvectors, which is costly for the system. After some time of getting used to the stimuli and the evoked activity becomes more stable, the information is more determined, and fewer eigenvectors are needed to efficiently to encode it.

1.1.4 Alignment to Spontaneous Activity Increases with larger Alignment

Spontaneous activity in neural systems is defined as neural activity that is not driven by an external stimulus. The activity patterns of spontaneous activity are not completely random and have often unique spatiotemporal patterns that instruct neural circuit development in the developing brain. Moreover, normal and aberrant patterns of spontaneous activity underline behavioral states and diseased conditions in adult brains. Therefore, spontaneous activity is essential for the understanding of brain development [**?**]. The alignment between activity patterns and spontaneous activity patterns could show the structural relation between patterns.

In baby ferrets' brains, a transformation from visual responses that are loosely aligned with spontaneous activity in the cortex before eye-opening to reliable and well-aligned responses several days after eye-opening was observed [**?**]. Thus, we expect that the feedforward recurrent hypothesis can reflect the tendency that evoked activity aligned better with spontaneous activity during the development of the brain.



(a) **(b)**

Figure 1.4 Correlation between alignment to spontaneous activity and feedforward recurrent alignment score in symmetric RNNs. The spontaneous activity reflects inputs from a wide range of different sources and is considered to be already aligned to the recurrent network[**?**]. Aligning activity patterns to spontaneous activity is in principle to explain the activity pattern by the principal components of spontaneous activity. **(a)** Variance ratio eq.(**??**) of spontaneous activity, evoked activity by feedforward input maximally aligned to recurrent network, and evoked activity by randomly aligned to recurrent network explained by principal components of spontaneous activity. The red line illustrates the variance ratio of spontaneous activity, the blue line the maximal alignment, and the green line the random alignment. Shadow shows the 95% confidence interval for 50 symmetric RNNs. **(b)** The correlation between final alignment to spontaneous activity and feedforward recurrent alignment eq.(**??**). The eigenvalues are ordered in descending order as eq.(1.5). Only the first half of eigenvectors are taken into account to determine the correlation.

Under the assumption that at eye-opening, the patterns of feedforward inputs are aligned to random activity patterns, thus not as well aligned to the recurrent network as the spontaneous activity. The evoked and spontaneous pattern overlaps only a little (Figure 1.4(a), visually comparing the green line to the red line.). It could be observed here that only the last few principal components have a similar variance ratio, while the first few dominant eigenvectors differ a lot. There is not much overlap between green and red curves. On the contrary, experience-driven changes that optimize the feedforward-recurrent alignment to $e_{\max}$ results in a stronger overlap between distributions of evoked and spontaneous activity patterns (Figure 1.4(a), visually comparing red line to blue line.). Most of the principal components have a similar explained variance ratio. Two curves overlap a lot. In both cases, the theoretical modeling hypothesis matches experimental observations in baby ferrets' brains [**?**].

To visualize and quantify the overlaps between activity patterns from evoked and endogenous patterns, we considered the summarized alignment score eq.(**??**). An exponential correlation between alignment to spontaneous activity and feedforward recurrent alignment (Figure 1.4b) is suggested by the modeling. The strong growth of overlaps between evoked and endogenous activity starts only shortly before the optimal experience-driven alignment, indicating that the alignment could require a large amount of experience and training. The costly process of optimal alignment to spontaneous activity could on the other hand reflect the importance of the connection between evoked and endogenous activity patterns.

1.2 Evaluation of Feedforward Recurrent Alignment Modulations for asymmetric Recurrent Interaction Networks

For symmetric interaction networks, the feedforward recurrent alignment hypothesis and the modeling based on it in section 1.1 can demonstrate the response properties observed in ferrets [**?**]. With better alignment between inputs and recurrent network, key results of the response properties were

- The trial-to-trial correlation increases.
- The intra-trial stability increases.
- The Dimensionality decreases.
- The alignment of evoked activity to spontaneous activity increases.

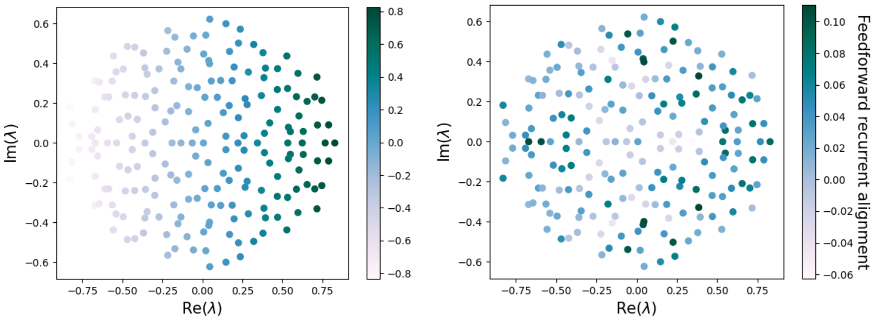
However, since the symmetric interaction matrices simplify the neural connection dramatically, we try to embed the more biology-realistic asymmetric interaction matrices. Considering the modifications listed in methods in section **??**, we want to evaluate the modifications of the feedforward recurrent alignment score based on the key results we got with symmetric RNNs.

Firstly, we check if the feedforward recurrent alignment score keeps the proportionality to eigenvalues. We expect that a suitable modification could keep the monotonously positive correlation with eigenvalues. Then, we go through the four response properties to verify if the tendency above is still kept with increased alignment between inputs and recurrent network.

1.2.1 Monotony of Feedforward Recurrent Alignment Score in dependence of Eigenvalues

During the development, suggested by the feedforward recurrent alignment hypothesis for symmetric interactions, the inputs align better to the RNNs through aligning to dominant eigenvectors. Feedforward recurrent alignment is proportional to the corresponding eigenvalue of the eigenvector that is aligned with input eq.(**??**).

When considering the asymmetric interaction network, keeping the hypothesis that the inputs align more and more to the eigenvector with maximal eigenvalue, we examine if the modified feedforward recurrent alignment score could still keep the proportionality to eigenvalues. If a monotonously positive correlation between the feedforward recurrent alignment and eigenvalues could be kept, the alignment score could at least definitely quantify how well inputs are aligned to the dominant eigenvector.



(a) **(b)**

Figure 1.5 Correlation between eigenvalues and feedforward recurrent alignment of modifications for asymmetric RNNs. To verify the considered modifications from section **??**, a monotonously positive correlation between eigenvalues and feedforward recurrent alignment is necessary. Represent the correlation in the complex plane, since eigenvalues are complex. The color-bar indicates the score of feedforward recurrent alignment. The darker the color, the larger the alignment score. **(a)** Representation of alignment score calculated with modification **??** by eq.(**??**) in complex plane. **(b)** Representation of alignment score calculated with modification **??** by eq.(**??**) in complex plane.

Modification 1 : Apply the real part of complex inputs with eq.(**??**).

With the feedforward recurrent alignment, we mainly consider the correlation between alignment score and corresponding eigenvalues. Inserting inputs h aligned to eigenvectors, for simplicity $h := e_i \in \mathbb{C}^{n\times 1}$ into eq.(**??**), it leads to

$$\nu_{\text{Re}} = \frac{\text{Re}(e_i)^T J \text{Re}(e_i)}{\|\text{Re}(e_i)\|^2} . \tag{1.6}$$

Because $\|e_i\| = c\|\text{Re}(e_i)\|$ with $c \in \mathbb{R}_+$ a general positive constant, we could get the proportionality between alignment score ν_{Re} and the real part of eigenvalues $\text{Re}(\lambda_i)$ of J :

$$\begin{aligned} \nu_{\text{Re}} &= c\text{Re}\left(\frac{e_i}{\|e_i\|}\right)^T J\,\text{Re}\left(\frac{e_i}{\|e_i\|}\right) \\ &= c\text{Re}\left(\frac{e_i^T}{\|e_i\|} J \frac{e_i}{\|e_i\|}\right) \\ &= c\text{Re}(\lambda_i) , \end{aligned} \tag{1.7}$$

with c a general positive constant.

The positive correlation between alignment and the real part of eigenvalues λ_i could also be observed in representation in the complex plane (Figure 1.5a). With increasing the real part of the eigenvalue, the size of the alignment score also becomes larger. A better alignment between inputs and a certain dominant activity pattern is thus the only reason for the increase in the feedforward recurrent alignment score. Since the alignment score only depends on the real part, there is no correlation found in the direction of the imaginary part of eigenvalues.

Modification 2 : Apply the magnitude for each neuron input eq.(**??**).

When the inputs are aligned to eigenvectors e of interaction matrix J, the feedforward recurrent alignment could be modified with $|e| := (|e_i|)_{i=1,\dots,n} \in \mathbb{R}^{n\times 1}$. Since the euclidean norm of vector $|e|$ is the same as the norm directly on the complex eigenvector $\|e\|_2$, the final feedforward recurrent alignment score from eq.(**??**) can be formulated as following:

$$\begin{aligned} \nu_{\text{mag}} &= \frac{|e|^T J |e|}{\|e\|^2} \\ &= \frac{\sum_{j=1}^{n} |e_j| \left(\sum_{i=1}^{n} |e_i| J_{ij}\right)}{\|e\|^2}\,, \end{aligned} \tag{1.8}$$

where e_i, e_j are the i-th and j-th element of the vector e, and J_{ij} the matrix element at i-th row and j-th column.

No direct proportionality between the alignment score and corresponding eigenvalues λ can be established. The lack of correlation is also represented in the complex plane (Figure 1.5b).

Modification 3 : Align the inputs to eigenvectors of symmetrized network eq.(**??**).

Instead of aligning the inputs to eigenvectors of original asymmetric RNNs and thinking about modifications of complex eigenvectors to calculate the feedforward recurrent alignment score, we now align the inputs to real eigenvectors $\tilde{e}$ from symmetrized interaction matrix $\tilde{J}$ as an approximation. However, the alignment score is still being projected to the original asymmetric RNNs defined in eq.(**??**). Despite of original asymmetric interaction matrix for the feedforward recurrent alignment score, there is still a proportionality between the eigenvalues of the symmetrized network and alignment score, shown in Figure 1.6.

The kept proportionality can also be obtained analytically. With the formulation of symmetrized interaction matrix through eq.(**??**), it follows with the help of definition for ν_{sym} from eq.(**??**),

$$
\begin{aligned}
&\frac{\tilde{e}^T\tilde{J}\tilde{e}}{\|\tilde{e}\|^2} = \frac{\tilde{e}^T\frac{J+J^T}{2}\tilde{e}}{\|\tilde{e}\|^2} = \frac{1}{2}\left(\frac{\tilde{e}^TJ\tilde{e}}{\|\tilde{e}\|^2} + \frac{\tilde{e}^TJ^T\tilde{e}}{\|\tilde{e}\|^2}\right)\\
\Rightarrow\ & 2\frac{\tilde{e}^T\tilde{J}\tilde{e}}{\|\tilde{e}\|^2} - \frac{\tilde{e}^TJ^T\tilde{e}}{\|\tilde{e}\|^2} = \frac{\tilde{e}^TJ\tilde{e}}{\|\tilde{e}\|^2}\\
\Rightarrow\ & 2\tilde{\lambda} - c = \frac{\tilde{e}^TJ\tilde{e}}{\|\tilde{e}\|^2} \text{ with } c := \frac{\tilde{e}^TJ^T\tilde{e}}{\|\tilde{e}\|^2} \in \mathbb{R} \text{ and } \tilde{\lambda} \text{ the corresponding eigenvalue of } \tilde{e}\\
\Rightarrow\ & \frac{\tilde{e}^TJ\tilde{e}}{\|\tilde{e}\|^2} \propto \tilde{\lambda}\,.
\end{aligned}
\tag{1.9}
$$

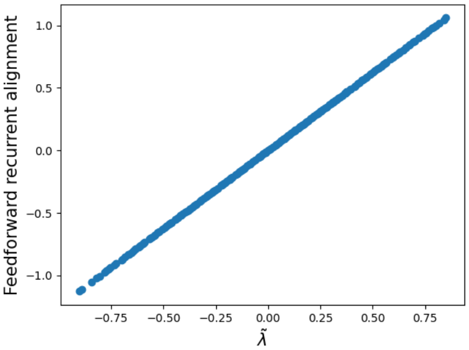


Figure 1.6 Positive correlation between feedforward recurrent alignment score and eigenvalues of symmetrized network as modification for asymmetric RNNs. Align the inputs to the eigenvectors of the symmetrized network while keeping feedforward recurrent alignment obtained by the original asymmetric interaction matrix with eq.(**??**). The correlation between the alignment score (y-axis) to the corresponding eigenvalues of the symmetrized network (x-axis) remains positive.

Conclusion After the analysis above for all modifications, our expectation of a contained positive correlation is kept when considering the real part (modification 1) and alignment with the symmetrized interaction matrix (modification 3). The variant with magnitude (modification 2) fails to fulfill the expected correlation. Thus, we will leave modification 2 out without further analysis of response properties in the following section.

1.2.2 Representing Response Properties with modified Feedforward Recurrent Alignment

After verifying the expected correlation between the feedforward recurrent alignment score and eigenvalues. We will check out if the remaining modifications 1 and 3 could still reflect the four experimental observations listed at the beginning of the section 1.2.

When aligning the inputs to eigenvectors of the asymmetric recurrent network to test the feedforward recurrent alignment hypothesis,

- modification 1 considers only the real part of eigenvectors eq.(**??**).
- modification 3 aligns the inputs to eigenvectors of the symmetrized interaction matrix but calculates the alignment score still with the original asymmetric network eq.(**??**).

For the modeling, the asymmetric RNNs are constructed with a certain degree of symmetry with eq.(**??**). We therefore also controlled how much the degree of symmetry influences the results.

Trial-to-trial Correlation Trial-to-trial correlation is the averaged correlations between trials defined by eq.(**??**). The expectation is a positive correlation between feedforward recurrent alignment and trial-to-trial correlation. Under varies degree of symmetry, the positive correlation should still be kept.

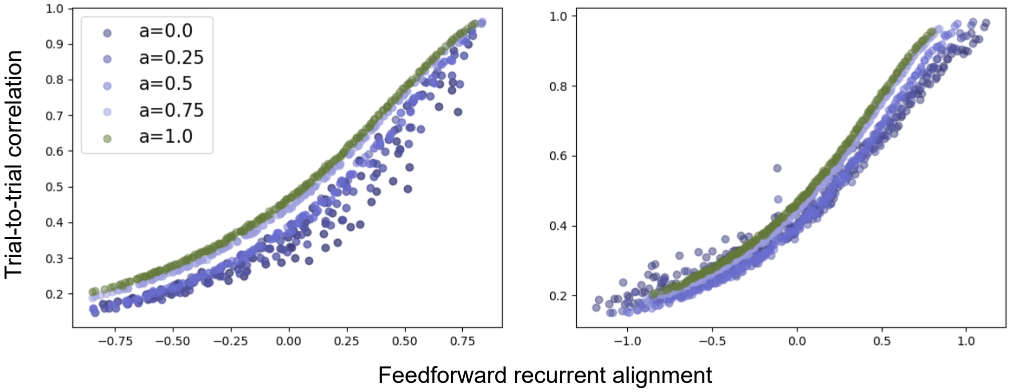


Figure 1.7 Correlation between trial-to-trial correlation and feedforward recurrent alignment with asymmetric RNNs inclusive the influence of symmetry. For different degrees of symmetry a, different color dots are applied shown in the legend. The green dots with $a = 1.0$ illustrate the case of symmetric RNNs. The darkest purple dots $a = 0.0$ represent the asymmetric network without any degree of symmetry. The x-axis shows the alignment score calculated with considered modifications 1 and 3, and the y-axis indicates the trial-to-trial correlation calculated with aligned inputs through eq.(**??**) **Left**: Results generated with modification 1 (only consider real part). **Right**: Result with modification 3 (symmetrized interaction matrix).

The results in Figure 1.7 point out that the trend of positive correlation keeps while the network increases its asymmetry for both modifications 1 and 3.

However, when the network is fully asymmetric, the correlation has a larger dispersion with modification 1. We suspect that the reason is that in the case of full asymmetry, more information got lost if only considering information from the real part of aligned inputs.

If there is a certain degree of symmetry, a part of the information originates from the symmetric structure. For this part, no information gets lost when only taking the real part of aligned inputs. On the other hand, aligning to fully asymmetric RNNs ($a = 0.0$ in Figure 1.8 left panel) and only taking the real part of aligned inputs, all neurons receive as a result incomplete information. Therefore, the correlation between evoked patterns will be disturbed mostly.

In the case of modification 3, the correlation is almost maintained the overall degree of symmetry as expected, shown in Figure 1.8 right panel.

Intra-trial Stability Intra-trial stability is the averaged time-delayed activity correlation inside one single trial quantified by eq.(**??**). As the name indicates, it shows how stable the information is represented inside one trial. According to the result from symmetric RNNs, we expect a similar exponential positive correlation as in Figure 1.2 also with asymmetric RNNs. The degree of symmetry should also not influence the result significantly under the assumption that the hypothesis works well with asymmetric RNNs generally.

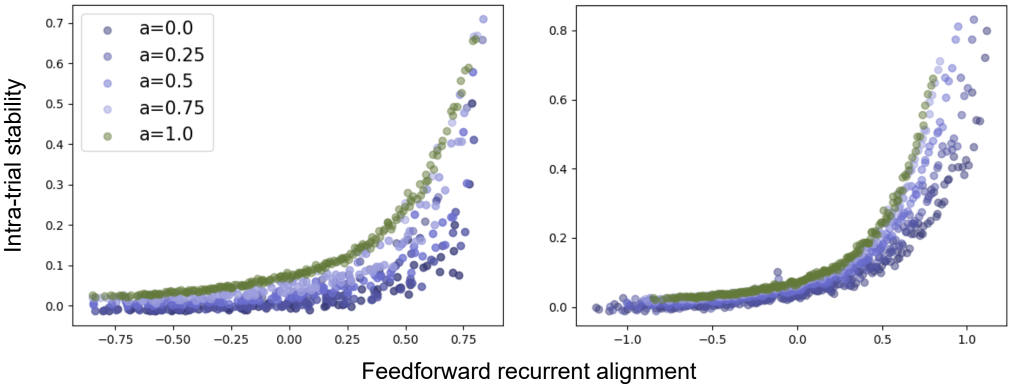


Figure 1.8 Correlation between intra-trial stability and feedforward recurrent alignment with asymmetric RNNs considering the influence from the degree of symmetry. For multiple degrees of symmetry a, different color dots are applied shown in the legend. From complete symmetric ($a = 1.0$ as the control group) to full asymmetric ($a = 0.0$) RNNs, the corresponding feedforward recurrent alignment (x-axis) is plotted against the intra-trial stability (y-axis). **Left**: Results with modification 1 (only real part of aligned inputs). **Right**: Results with modification 3 (align input to symmetrized network).

We suspect with a similar reason as in trial-to-trial correlation that for modification 1, the loss of imaginary part information of aligned inputs could lead to increased dispersion with the network becoming more asymmetric (Figure 1.8 left panel).

For modification 3, the expected positive correlation between intra-trial stability and feedforward recurrent alignment score is kept. Besides, we also observe that the degree of symmetry influences here more than in trial-to-trial correlation (Figure 1.7 right panel). It is perhaps because the intra-trial stability is more sensitive to the information lost than trial-to-trial correlation. With the increased asymmetry, more information got lost during symmetrization. Imagine if having a total symmetric network, the result of symmetrization is the network itself, and therefore no information is lost in eigenvectors. However, if having a fully asymmetric recurrent network, after symmetrization, information stored in complex eigenvectors gets transformed to lower dimension real eigenvectors. Thus, when aligning inputs to the symmetrized interactions, transformed information could be the reason for the influence of the degree of symmetry.

Dimensionality Dimensionality reflects the complexity of the information they encoded. A high dimensional activity pattern needs more orthogonal activity patterns for representation than a low dimensional activity pattern and thus indicates larger variability and higher complexity of contained information [**?**, **?**, **?**]. Modifications 1 and 3 change the eigenvectors for the construction of input covariance and the eigenvalues for analytical calculation of effective dimensionality (section **??** eq. (**??**), (**??**)).

If both modifications work well, the results should be similar to results from symmetric RNNs (Figure 1.3b). A low degree of symmetry is allowed to lead to a small range of dispersion, which however should still keep the tendency of decreased dimensionality with increased feedforward recurrent alignment score.

With modification 1 (Figure 1.9 left panel), the empirical dimensionality differs a lot from the analytical calculation. Besides, there is no significant correlation between dimensionality and feedforward recurrent alignment with empirical data as long as asymmetric structure is included. The error between analytical and empirical approximations is significantly large. The reason for those phenomena could be the loss of orthogonality between the real part of eigenvectors, that is

$$e_i \perp e_j \;\not\Rightarrow \mathrm{Re}(e_i) \perp \mathrm{Re}(e_j) \;\forall i, j = 1, ..., n \,. \tag{1.10}$$

As a result, the covariance matrix Σ^{Dim} from eq.(**??**) is not necessarily constructed with orthogonal vectors anymore, which contradicts the assumption of reflecting dimensionality by projecting activity pattern on orthogonal patterns that are also applied to construct covariance matrix.

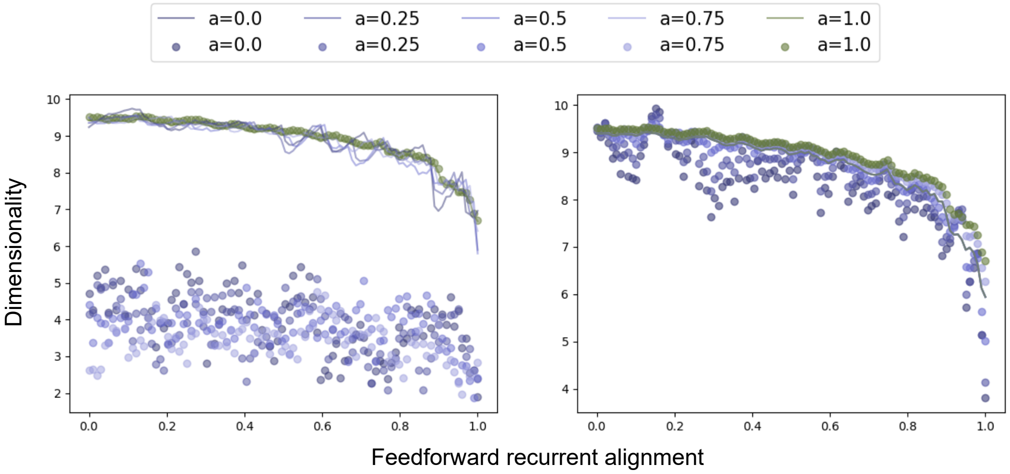


Figure 1.9 Analytical and empirical effective dimensionality in correlation with feedforward recurrent alignment score with asymmetric RNNs including impact from the degree of symmetry. The full symmetric recurrent network ($a = 1.0$) is a control for other cases. When $a = 0.0$, the RNN is fully asymmetric. The dots represent the empirical approximation for effective dimensionality eq.(**??**). The lines are for the analytical calculation for dimensionality adjusted to modification 1 and 3 (section **??** eq.(**??**), (**??**)). **Left**: Results with modification 1 (only real part of aligned inputs and eigenvalues for analytical dimensionality). **Right**: Results with modification 3 (align inputs to symmetrized interaction matrix).

With modification 3, the results fulfill largely our expectations: the negative correlation between dimensionality and feedforward recurrent alignment is kept in both analytical and empirical approximations. Little dispersion could be due to the structural information lost during symmetrization.

Thus, until this step, we would also drop modification 1 and only further consider modification 3.

Alignment to spontaneous activity The alignment of the evoked activity pattern to the endogenous pattern measures the overlap between their variance ratio curves explained by principal components of endogenous pattern eq.(**??**). For example the overlaps between curves in Figure 1.4a. Only modification 3 is now left for the evaluation.

Similar to prior cases, if the modification works, we expect the correlation between alignment to spontaneous activity and feedforward recurrent alignment score would also be similar to symmetric RNNs in Figure 1.4b. The alignment to spontaneous activity is modified with eq.(**??**) and calculated with eq.(**??**).

As shown in Figure 1.10, the dispersion increases with the degree of asymmetry due to the increased information lost, but the general tendency that alignment of evoked activity pattern to spontaneous activity becomes larger when the inputs are aligned to more dominant eigenvectors of symmetrized interaction matrix. With a high degree of symmetry, for example, $a = 0.75$, the correlation is very similar to it with total symmetric recurrent network.

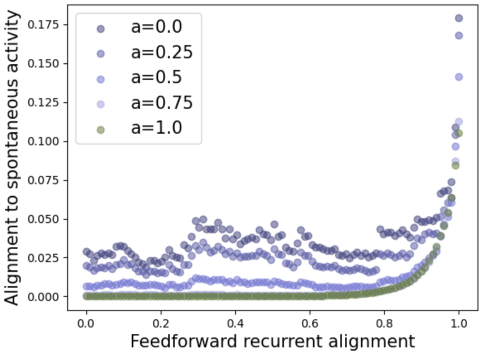


Figure 1.10 The correlation between alignment to spontaneous activity and feedforward recurrent alignment for asymmetric RNNs with symmetrized interactions considering influence from the degree of symmetry. As a control group,a fully symmetric RNN ($a = 1.0$) is represented with dark green dots. For different degrees of symmetry from $a = 0.75$ to 0, the darker the dots' color, the more asymmetry is in the network. Only modification 3 is evaluated for alignment to spontaneous activity.

Conclusion After evaluating the rest of modifications 1 and 3 with four perspectives of response properties, modification 1 is filtered out because of the lack of orthogonality between eigenvectors after modification. This could be the reason for the large inconsistency between analytical and empirical approximations of effective dimensionality.

Modification 3 performs at the end the best through all four response properties, and could generally fulfill the expectations of correlations between the response properties and feedforward recurrent alignment.

Therefore, modification 3 which aligns the inputs to asymmetric RNNs with a symmetrized interaction matrix is considered to be a good candidate for modeling feedforward recurrent alignment hypothesis in asymmetric recurrent networks.

1.3 Modeling Feedforward Recurrent Alignment Hypothesis on Low-rank Recurrent Neural Networks (Low-rank RNNs)

Although fully recurrent connectivity structure is one of the most popular network models for theoretical neuroscience, there are experimental recordings suggesting that the transformation of sensory stimuli into motor outputs relies on low-dimensional dynamics at the population level [**?**]. Therefore, the low-rank connectivity structure can be a good candidate for understanding the neural mechanism from another perspective.

We hence also try to model the feedforward recurrent alignment hypothesis on the low-rank RNNs to discover if the hypothesis modeling adapted from full-rank RNNs also could work with low-rank RNNs. Hereby, we consider both symmetric and asymmetric RNNs with constructions with and without noise described in section **??** by eq.(**??**) and eq.(**??**). Under symmetric or asymmetric conditions, we go through both constructions with the four response properties in correlation with the feedforward recurrent alignment score as for full-rank RNNs before. Those four response properties are trial-to-trial correlation, intra-trial stability, dimensionality, and alignment to spontaneous activity.

1.3.1 Evaluation of Feedforward Recurrent Alignment in symmetric Low-rank RNNs based on response properties

We first consider symmetric low-rank networks in constructions with and without noise. For each case, the results of response property analysis based on the modeling of the feedforward recurrent alignment hypothesis for symmetric networks are evaluated with correlations between them.

Low-rank RNNs without random noise The formulation of low-rank RNNs is followed by eq.**??** with the rank G significantly smaller than the number of neurons n. The eigenvalues of symmetric low-rank RNNs are real numbers.

As shown in Figure **??**, only two real eigenvalues are seen. A number of G eigenvalues equal the normalization factor $R < 1$ that limits the value range by eq.(**??**). The rest of $n - G$ eigenvalues take the value 0.

The phenomenon of bi-polarized eigenvalue distribution is due to the construction of low-rank RNNs here. If considering the case of having the left connectivity vectors as orthonormal basis for construction of RNNs from eq.(**??**), which is

$$J = \frac{1}{n}\sum_{g=1}^{G} l^{(g)} l^{(g)T} = \sum_{g=1}^{G} \frac{1}{n} l^{(g)} l^{(g)T} \, . \tag{1.11}$$

If the rank equals the number of neurons, $G = n$, the formulation of low-rank matrix eq.(1.11) is at the same time a symmetric full rank matrix with eigenvectors $\left\{l^{(g)}\right\}$

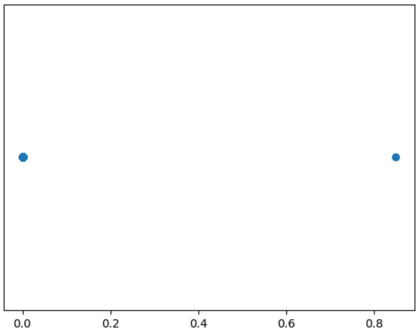


Figure 1.11 Eigenvalue distribution of symmetric low-rank RNNs without random noise. The eigenvalues of symmetric low-rank RNNs are real numbers (x-axis). With rank $G = 1 \ll n = 200$ the number of neurons, 1 eigenvalue takes the value of normalization factor $R = 0.85$, and the rest $n - G = 199$ eigenvalues equal 0.

and all eigenvectors correspond to the same eigenvalue $\frac{1}{n}$. Since the eigenvalues are re-scaled by parameter $R < 1$ to enable the stable steady state, the eigenvalues are equal to R based on eq.(**??**).

Now, if the rank G is smaller than the number of neurons n, the formulation of low-rank RNN eq.(1.11) can be rewritten as

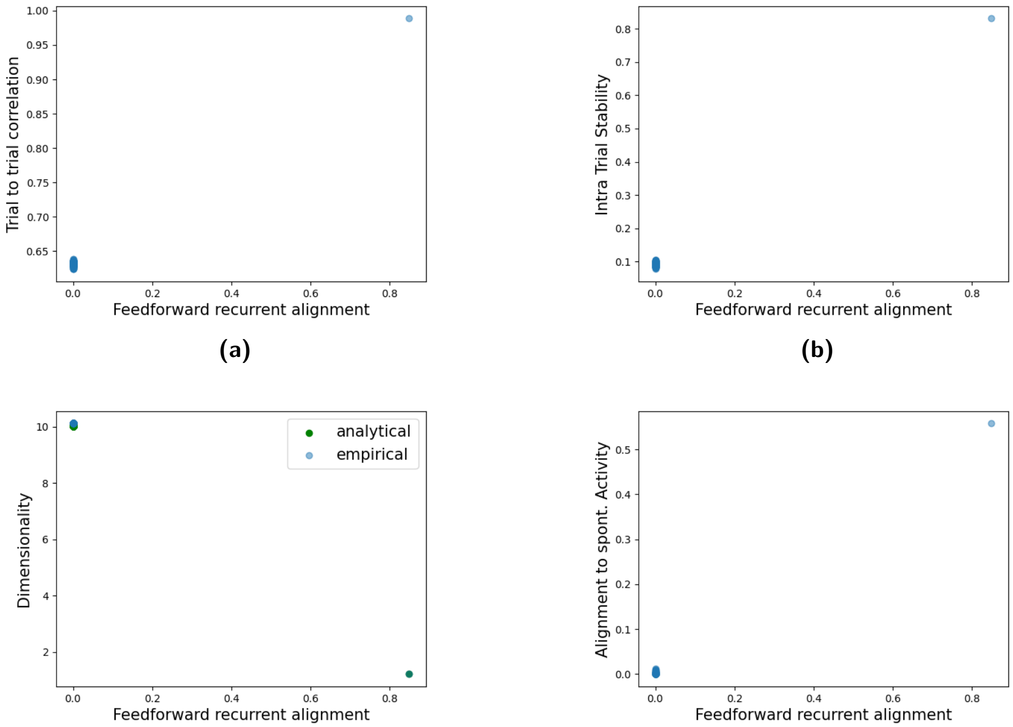
$$J = \sum_{g=1}^{G} \frac{1}{n} l^{(g)} l^{(g)T} + 0 = \sum_{g=1}^{G} \frac{1}{n} l^{(g)} l^{(g)T} + \sum_{g=G}^{n-G} 0 l^{(g)} l^{(g)T} . \tag{1.12}$$

So, there are G basis vectors that have eigenvalue $\frac{1}{n}$, which is further re-scaled to $R < 1$. The rest of $n - G$ eigenvectors have eigenvalue 0. This then results the eigenvalue distribution shown in Figure 1.11.

We first look at the trial-to-trial correlation and expect a positive correlation with the feedforward recurrent alignment. When aligning the inputs to the symmetric recurrent alignment, the inputs-distribution has the mean vector matched to the eigenvectors of the low-rank interaction matrix J. Due to the feedforward recurrent alignment formulation, the alignment scores are equal to the corresponding eigenvalues because of eq.(**??**).

Since there are only two eigenvalues R and 0 for the rank G smaller than number of neurons n, we assume that there is no continuous correlation but two groups of trial-to-trial correlation values. However, low feedforward recurrent alignment should still correlate with a small trial-to-trial correlation value and a large alignment score with a big trial-to-trial correlation. We expect therefore here a discontinuous positive correlation between trial-to-trial correlation and feedforward recurrent alignment.

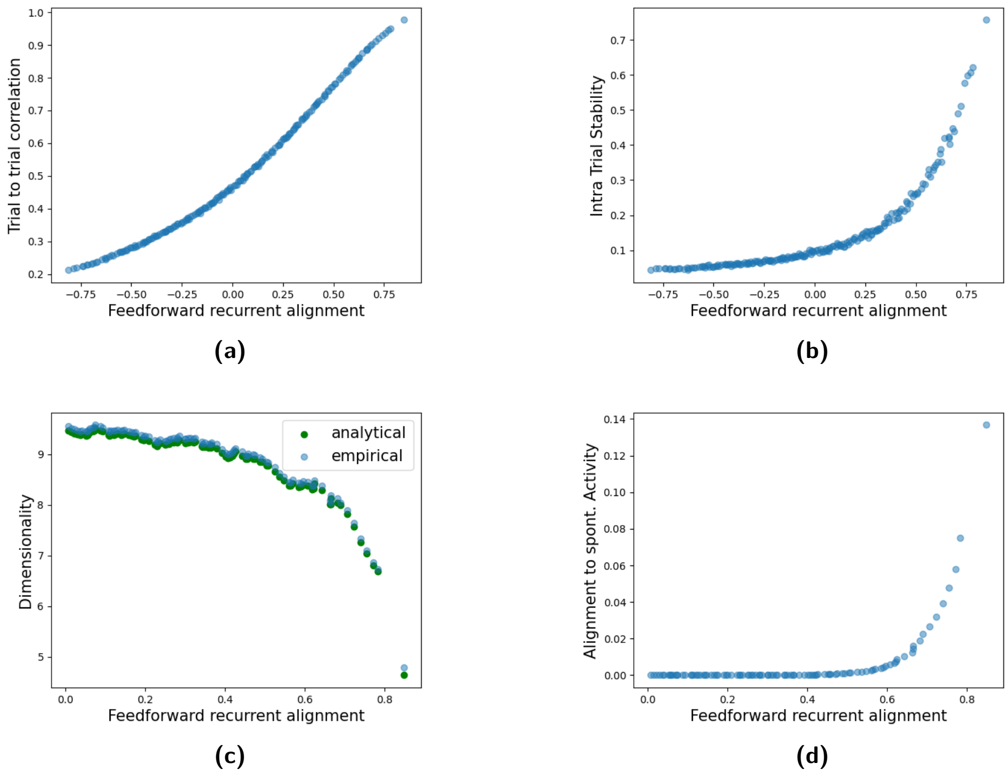
As the result in Figure 1.12a shows, the trial-to-trial correlation distributes separately into two groups due to the distribution of eigenvalues. Also as expected, there is still a positive correlation between feedforward recurrent alignment score and trial-to-trial correlation.



(a) **(b)** **(c)** **(d)**

Figure 1.12 Correlation between response properties and feedforward recurrent alignment considering symmetric low-rank RNNs without random noise. Construct symmetric low-rank RNNs under the assumption of having left connectivity vectors equal to right connectivity vectors eq.(1.11). Rank G is the number of connectivity vectors and is significantly smaller than the number of neurons n. Here $G = 1$ and $n = 200$. To evaluate the feedforward recurrent alignment hypothesis, the correlations between response properties and the modeled feedforward recurrent alignment score are considered.
(a) Trial-to-trial correlation (y-axis) in correlation with feedforward recurrent alignment score (x-axis).
(b) Intra-trial stability (y-axis) in correlation with feedforward recurrent alignment score (x-axis).
(c) Dimensionality (y-axis) calculated analytically (green dots, eq.(**??**)) and empirically (blue dots, eq.(**??**)) in relationship with feedforward recurrent alignment score (x-axis).
(d) Correlation between alignment to spontaneous activity (y-axis) and feedforward recurrent alignment score (x-axis).

Analogous to the trial-to-trial correlation, we receive the results for intra-trial stability (Figure 1.12b) and alignment to spontaneous activity (Figure 1.12c) also be a discontinuous positive correlation to feedforward recurrent alignment, meanwhile a discontinuous negative correlation for dimensionality (Figure 1.12d). Due to the same reason of existing only two groups of eigenvalues at R and 0, the intra-trial stability value, alignment to spontaneous activity, and dimensionality distribute also separately into two groups while keeping the expected correlations with feedforward recurrent alignment.



(a)

(b)

(c)

(d)

Figure 1.13 Correlation between response properties and feedforward recurrent alignment in symmetric low-rank RNNs with random noise. Besides the part of a low-rank matrix with rank G, low-rank RNNs with noise include an extra part of symmetrized random Gaussian matrix from eq.(1.13). $G = 1$ here.
(a) Trial-to-trial correlation (y-axis) against feedforward recurrent alignmnet score (x-axis).
(b) Intra-trial stability (y-axis) against feedforward recurrent alignment score (x-axis).
(c) Dimensionality (y-axis) in dependence of feedforward recurrent alignment score (x-axis). Green dots represent the analytical calculation of dimensionality eq.(**??**) and blue dots for empirical approximation of eq.(**??**)
(d) Alignment to spontaneous activity (y-axis) against feedforward recurrent alignment (x-axis).

Low-rank RNNs with random noise Low-rank symmetric RNNs with random noise include an extra part of symmetrized Gaussian distributed matrix defined as eq.(**??**) in section **??**. If analog to eq.(1.11) considering the right connectivity vectors equal to the left connectivity vectors, the symmetric low-rank RNNs with random noise can be formulated with a part of symmetric low-rank RNN and a part of symmetrized Gaussian distributed matrix J_{rand}.

$$J = \frac{1}{n} \sum_{g=1}^{G} l^{(g)} l^{(g)T} + J_{\text{rand}} \,. \tag{1.13}$$

The symmetric random part J_{rand} should provide more dynamics in the network.

Since J_{rand} is a full-rank symmetric matrix, the final recurrent network J is also a full-rank symmetric matrix despite the low-rank symmetric part with rank G. Therefore, the correlations between the response properties and feedforward recurrent alignment score are similar to the case of full-rank symmetric RNNs (Figure 1.1, 1.2, 1.3b, and 1.4b).

As a result, the feedforward recurrent alignment hypothesis can be modeled. The expected relationships between feedforward recurrent alignment and response properties from section 1.1) are fulfilled, as shown in Figure 1.13. However, the influence of low-rank construction is therefore overwritten by the random noise, which leads to the final dynamics not much different from the case with general random symmetric RNNs.

Conclusion Considering symmetric low-rank RNNs without random noise, the feedforward recurrent alignment keeps positive correlations to trial-to-trial correlation, intra-trial stability, and alignment to spontaneous activity. Besides, the expected negative correlation between dimensionality and feedforward recurrent alignment is also fulfilled. However, the network construction eq.(1.11) leads to a simple distribution of eigenvalues (Figure 1.11) and in turn also the distribution of feedforward recurrent alignment score. As a result, only a simple discontinuous correlation of feedforward recurrent alignment to response properties can be observed (Figure 1.12).

If adding random noise to disturb the simple dynamics of symmetric low-rank RNNs with eq.(1.11), the final construction eq.(1.13) is a full-rank symmetric RNN. As a result, the correlations of feedforawrd recurrent alignment to response properties are similar to the results observed in general symmetric RNNs from section 1.1. Although the expected phenomena are seen, the effect of low-rank RNNs cannot be significantly detected.

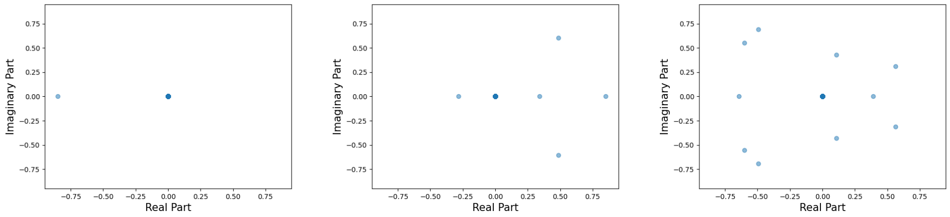
In total, after considering the symmetric low-rank RNNs with and without random noise, the feedforward recurrent alignment hypothesis can be modeled in both cases. The relationships between feedforward recurrent alignment and response properties also meet the expectations.

1.3.2 Different Constructions influence the impact of rank in Asymmetric Low-rank RNNs based on Response Properties

Asymmetric low-rank RNNs can have more complex dynamics than symmetric low-rank RNNs. Section **??** also introduces the construction of asymmetric low-rank RNNs with or without random noise. Generally, if taking the set of left connectivity vectors different from the set of right connectivity vectors, asymmetric matrices can be formulated by eq.(**??**), (**??**).

Since asymmetric low-rank RNNs are only a certain case of asymmetric RNNs with specific constructions, they also have the problem of complex eigenvectors and eigenvalues when modeling the feedforward recurrent alignment hypothesis. Therefore, from the previous results with general asymmetric RNNs, we choose modification 3 from section **??** with aligning inputs to symmetrized RNN because of the overall best performance among all considered modifications as mentioned in section 1.2.

Low-rank RNNs without random noise Asymmetric low-rank RNNs without random noise consist of non-equal left and right connectivity vectors, which are mutually orthogonal illustrated in Figure **??**) and eq.(**??**).



(a) rank= 1 **(b)** rank= 5 **(c)** rank= 10

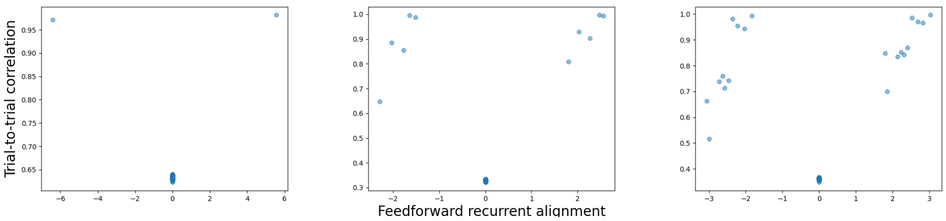
Figure 1.14 Eigenvalue distribution in dependence of rank with asymmetric low-rank RNNs without random noise. The asymmetric low-rank RNNs without random noise are constructed by mutually orthogonal non-equal left and right connectivity vectors defined by eq.(**??**). The number of vector pairs G also determines the rank of the interaction matrix, which should be significantly smaller than the number of neurons n. The rank G can influence the distribution of eigenvalues for asymmetric low-rank RNNs in complex plane. Plotting eigenvalues in the complex plane with x-axis for the real part and y-axis for the imaginary part. Low-rank RNNs with rank: **(a)** $G = 1$. **(b)** $G = 5$. **(c)** $G = 10$.

For symmetric low-rank RNNs without random noise, as long as the rank G is significantly smaller than the number of neurons n as defined, the distribution of eigenvalues is separated into two groups as in Figure 1.11 keeps unchanged independent of the rank. While for asymmetric low-rank RNNs without random noise, the influence of rank G is more significant, as shown in Figure 1.14.

The construction of asymmetric low-rank RNNs without random noise defined by eq.(**??**) can be understood as G pairs of presented connectivity vectors, while the rest of $n - G$ pairs are suppressed, rewritten with

$$J = \frac{1}{n}\sum_{g=1}^{G} l^{(g)} r^{(g)T} = \sum_{g=1}^{G} \frac{1}{n} l^{(g)} r^{(g)T} + \sum_{g=1}^{n-G} 0 l^{(g)} r^{(g)T} . \tag{1.14}$$

Therefore, exactly G number of eigenvalues do not concentrate at 0 but the rest $n - G$ eigenvalues are. Since the mathematical construction eq.(**??**) for asymmetric low-rank RNNs is not an eigendecomposition, the eigenvalues after re-scaling by normalization parameter R (method section **??**) do not exactly have magnitude R as at symmetric low-rank RNNs from eq.(1.12).



(a) rank= 1 **(b)** rank= 5 **(c)** rank= 10

Figure 1.15 Relationship of trial-to-trial correlation and feedforward recurrent alignment in dependence of rank by asymmetric low-rank RNNs without random noise. The rank G in asymmetric low-rank RNNs from eq.(1.14) influence the eigenvalue distribution and thus also the distribution of trial-to-trial correlation. Rank G is significantly smaller than the number of neurons n by definition. $n = 200$. Illustrate the correlation between feedforward recurrent alignment and trial-to-trial correlation under varied ranks: **(a)** $G = 1$. **(b)** $G = 5$. **(c)** $G = 10$.

We expect that the distribution of the correlation between feedforward recurrent alignment and trial-to-trial correlation will be related to the eigenvalue distribution under different ranks. Besides, a positive correlation between feedforward recurrent alignment and trial-to-trial correlation should be maintained independent of the change of rank.

As expected, the rank influence the eigenvalue distribution and thus also the correlation distribution between feedforward recurrent alignment and trial-to-trial correlation. However, the positive correlation is not kept over the whole range of feedforward recurrent alignment. In fact, it can be observed that the distribution is almost identical in the negative and positive range of feedfroward recurrent alignment. The number of dots in each half-range also exactly equal the number of range, which is also the number of non-zero eigenvalues in Figure 1.14. There are now doubled number of non-zero feedforward recurrent alignment scores as expected.

We assume that the duplication could be due to the symmetrization of asymmetric low-rank RNNs with a simple construction of only applying connectivity vectors. With the symmetrization, a part of the reflected network also contributes to eigenvalue expression and therefore increases the number of presented feedforward recurrent alignment scores. Symmetrization of eq.(1.14) leads to

$$J_{\text{sym}} = \frac{J + J^T}{2} = \frac{1}{2n}\left(\sum_{g=1}^{G} l^{(g)} r^{(g)T} + \sum_{g=1}^{G} r^{(g)} l^{(g)T}\right) = \frac{1}{2n}\sum_{g=1}^{G} l^{(g)} r^{(g)T} + \frac{1}{2n}\sum_{g=1}^{G} r^{(g)} l^{(g)T} \,. \tag{1.15}$$

The second part of eq.(1.15) is presumed to be the reason for the distribution in the negative range of feedforward recurrent alignment in Figure 1.15. The construction from eq.(1.15) doubles the network in two sub-networks separately. Each sub-network is spread over the negative or the positive range of feedforward recurrent alignment. A positive correlation between feedforward recurrent alignment and trial-to-trial correlation can be observed in each half-range. But there is in the end no positive correlation overall in the feedforward recurrent alignment score range.

Thus, the modeling of the feedforward recurrent alignment hypothesis in the case of asymmetric low-rank RNNs without noise cannot fulfill the phenomenon over the whole range of alignment between inputs and symmetrized interaction matrix.

Low-rank RNNs with random noise Above without random noise, the simple construction by eq.(1.14) does not perform well with the modification of alignment to symmetrized RNNs. Now, we add a Gaussian distributed asymmetric random part to the simple construction. As a result, the asymmetric recurrent network is composed of a low-rank part from connectivity vectors and a part of full-rank random noise defined by eq. (**??**) and explained in method section **??**. The final interaction matric J is therefore full-rank and asymmetric.

If the feedforward recurrent alignment hypothesis can be modeled in the asymmetric low-rank RNNs with random noise, positive correlations between feedforward recurrent alignment score and trial-to-trial correlation, intra-trial stability, and alignment to spontaneous activity are expected. Moreover, dimensionality should be negatively correlated with feedforward recurrent alignment. Based on observations from symmetric low-rank RNNs with random noise in Figure 1.13, the results can be similar to the case with general asymmetric full-rank RNNs in section 1.2.2. In other words, the dynamics of low-rank can be suppressed by the full-rank noise dynamics.

The results in Figure 1.16 confirmed our expectations of correlations between response properties and feedforward recurrent alignment. Besides, the correlations have high similarity to the results we got from general asymmetric RNNs. Under the strong effect of full-rank random noise, the difference caused by the low-rank part is not significant (comparing blue and orange dots in Figure 1.16). The results in section 1.2 indicate that the dispersion could be increased with an increased proportion of asymmetry in RNNs. Since the final construction for asymmetric low-rank RNNs with noise in eq.(**??**) consists only of asymmetric networks, the dispersion is expected to be large, especially at dimensionality.

Conclusion We explore in this section the modeling of feedforward recurrent alignment in asymmetric low-rank RNNs. Two types of constructions are taken into account: 1) single part of the non-equal left and right connectivity vectors and 2) with additional full-rank random noise defined in method section **??**.

The simple construction without random noise has an eigenvalue distribution depending on the rank. Keeping the rank significantly smaller than the number of neurons, the number of non-zero eigenvalues equals the rank (Figure 1.14). Thus, the number of non-zero feedforward recurrent alignment scores also depends on the rank. Due to the modification of aligning inputs to a symmetrized interaction matrix for calculation of feedforward recurrent alignment score, the correlation between alignment score and trial-to-trial correlation is separated into two sub-groups. As a result, although in each sub-region the positive correlation between alignment score and trial-to-trial correlation is kept, there is no global positive correlation over the total range (Figure 1.15).

Adding the full-rank random noise can avoid the effect of doubling caused by symmetrization. The expected positive correlations between feedforward recurrent alignment and trail-to-trial correlation, intra-trial stability, and alignment to spontaneous activity are kept (Figure 1.16). Large dispersion is caused by the total asymmetry of the interaction matrix. Despite large dispersion, the negative correlation between dimensionality and feedforward recurrent alignment can be observed. However, the expression of low-rank connections is suppressed by the full-rank random noise, such that there is no significant difference between low-rank RNNs with various ranks.

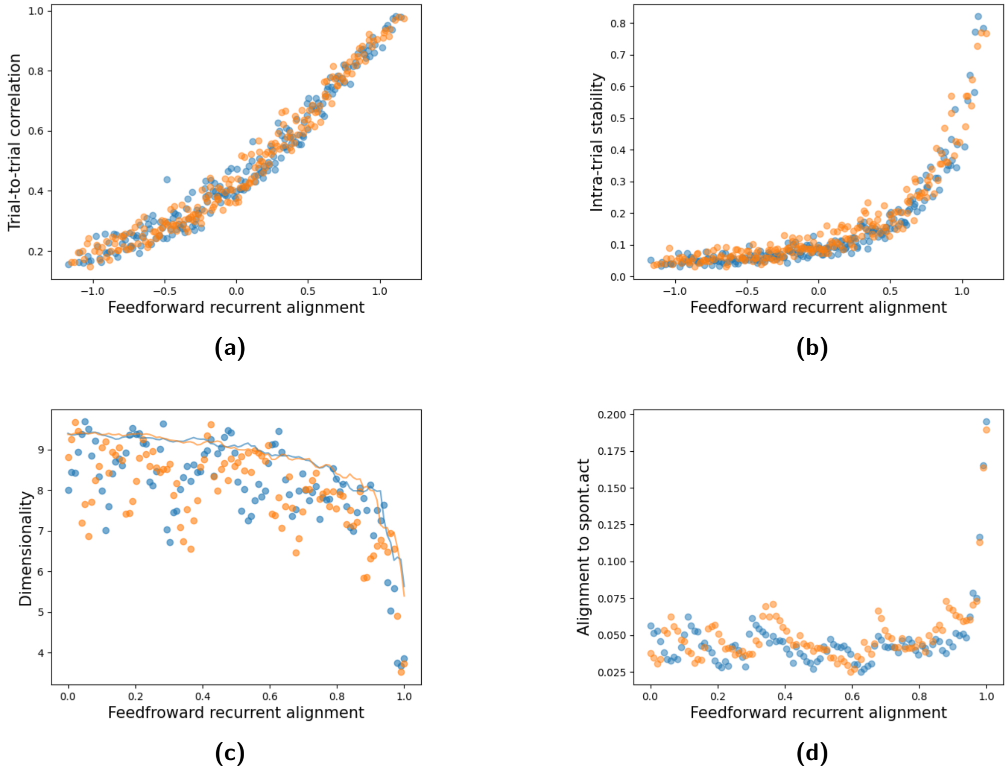


Figure 1.16 Correlations between feedforward recurrent alignment and selected response properties for asymmetric low-rank RNNs with random noise. In addition to the low-rank part constructed only by left and right connectivity vectors, a Gaussian distributed full-rank asymmetric random noise is included in the network dynamic eq.(**??**). Due to the full-rank random noise, the results have a large similarity to results from general asymmetric RNNs in section 1.2.2. Two different ranks are taken to assess the influence of ranks on correlations. The rank G should be significantly smaller than the number of neurons n. With $n = 200$, comparing ranks $G = 1$ in blue dots with $G = 5$ in orange:
(a) Correlation between feedforward recurrent alignment (x-axis) and trial-to-trial correlation (y-axis). **(b)** Correlation between feedforward recurrent alignment (x-axis) and intra-trial stability (y-axis). **(c)** Dimensionality (y-axis) calculated analytically (lines, blue line for $G = 1$ and orange line for $G = 5$) and empirically in correlation with feedforward recurrent alignment (x-axis). **(d)** Alignment of evoked patterns to spontaneous activity (y-axis) in relationship with feedforward recurrent alignment (x-axis).

1.4 White Noise Evoked Activity Can Help to Approximate Dominant Activity Direction in Response Space for Unknown Asymmetric Recurrent Networks

Normally during experimental procedures, the complete recurrent network structure of the laboratory animal is difficult to access. Therefore, even if the feedforward recurrent alignment hypothesis can be theoretically underpinned well in both symmetric and asymmetric recurrent networks (section 1.1 and section 1.2.2), the hypothesis cannot be well undertaken in experimental environments.

Motivated by some predictions from a series of theoretical frameworks [?, ?] that the matching between inputs and spontaneous activity can lead to more reliable evoked responses and more efficient transmission across cortical networks, we consider the idea of generating spontaneous-like activity pattern for alignment as an approximation of the original recurrent network.

Our goal here is to achieve a theoretical framework for the feedforward recurrent alignment that is compatible experimentally without knowing the exact recurrent network structure. Besides, the framework could also support the previous founding that better alignment between input and spontaneous activity leads to reliable response activity.

To generate spontaneous-like activity patterns, an experimental method with white noise was suggested by the lab of H.Mulholland [?]. For this case, we model the white noise and its evoked activity for the construction of a modified feedforward recurrent alignment score. The inputs are aligned to the principal components of the white-noise-evoked activity pattern. Evaluation of our framework covers mainly four perspectives of response properties, namely trial-to-trial correlation, intra-trial stability, dimensionality, and alignment of evoked activity to spontaneous activity. They are introduced in the method section ??.

1.4.1 Input Alignment with White-noise-evoked Activity Pattern Support Previous Theoretical Frameworks

Positive Monotonic Correlation between Feedforward Recurrent Alignment Score and Eigenvalues of White-noise-evoked Activity Pattern

The feedforward recurrent alignment score should reflect how well the input is aligned with the dominant direction of activity patterns generated by RNNs. Since the original recurrent network is unknown, we apply the white-noise-evoked activity pattern for the alignment defined by eq.(??). Here aligning the input to principal components of white-noise-evoked activity pattern instead of the original recurrent network.

A positive monotony correlation between the feedforward recurrent alignment score and variance ratio of the aligned white-noise-evoked activity pattern is essential for guaranteeing the functionality of the feedforward recurrent alignment hypothesis. Due to the selective response amplification eq.(1.1), the dominant variance ratios of the white-noise-evoked pattern should determine the strength and reliability of responses. Thus, the feedforward recurrent alignment should have a high value when aligning to the corresponding principal components of dominant variance ratios, indicating that the evoked response would be reliable. The monotony guarantees that the alignment to dominant principal components is the only source for the increase of the feedforward recurrent alignment score.

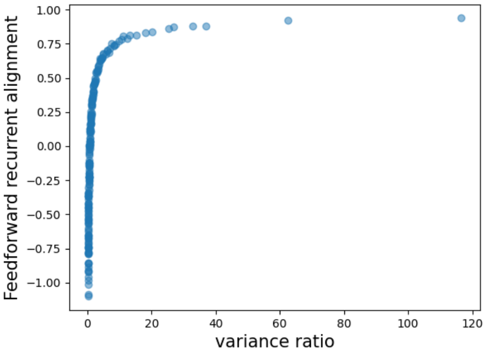


Figure 1.17 Correlation between feedforward recurrent alignment and variance ratio of white-noise-evoked activity patterns. With an unknown recurrent network structure, a white-noise-evoked spontaneous-like activity pattern is applied for aligning inputs. The Feedforward recurrent alignment score is formulated with principal components of the covariance matrix from white-noise-evoked activity eq.(**??**). The alignment score should be positive and monotonously correlated with the corresponding variance ratio.

The result in Figure 1.17 illustrates the correlation between the feedforward recurrent alignment score calculated with principal components and its corresponding variance ratios of white-noise-evoked activity. The dominant variance ratios are also associated with higher feedforward recurrent alignment. Besides, a significant positive monotony can be observed.

Thus, the feedforward recurrent alignment measured with white-noise-evoked activity through eq.(**??**) can be a good candidate measurement for supporting the predictions that inputs aligning well to spontaneous activity can generate more reliable responses.

Evaluation of Approximation with White Noise through Response Properties

To evaluate if the candidate measurement through aligning inputs to white-noise-evoked activity eq.(**??**) can quantify reliable responses without knowing the recurrent structure, four response properties are taken into account: 1) trial-to-trial correlation, 2) intra-trial stability, 3) dimensionality, and 4) alignment of evoked activity to spontaneous activity. Those four properties are observed in experienced ferrets' primary visual cortex [?] for characterizing the reliability of neural responses. Both in symmetric and asymmetric RNNs, the theoretical modeling of the feedforward recurrent alignment hypothesis also supports the experimental observations in ferrets.

If the newly constructed feedforward recurrent alignment with white-noise-evoked activity can well quantify the feedforward recurrent alignment hypothesis for aligning to white-noise-evoked activity pattern eq.(**??**), the correlations between feedforward recurrent alignment score and four response properties that mentioned above should coincide with prior results from feedforward recurrent alignment hypothesis for full-rank RNNs in sections 1.1 and 1.2.

Therefore, we expect at least that high feedforward recurrent alignment corresponds with high trial-to-trial correlation, intra-trial stability, and alignment to spontaneous activity, while with low dimensionality. The results in Figure 1.18 fulfill our expectations even under different degrees of symmetry in the original networks.

However, relatively high trial-to-trial correlation, low dimensionality, and high alignment to spontaneous activity occur when the feedforward recurrent alignment score is small. In other words, a large discrepancy exists in a range of low feedforward recurrent alignment.

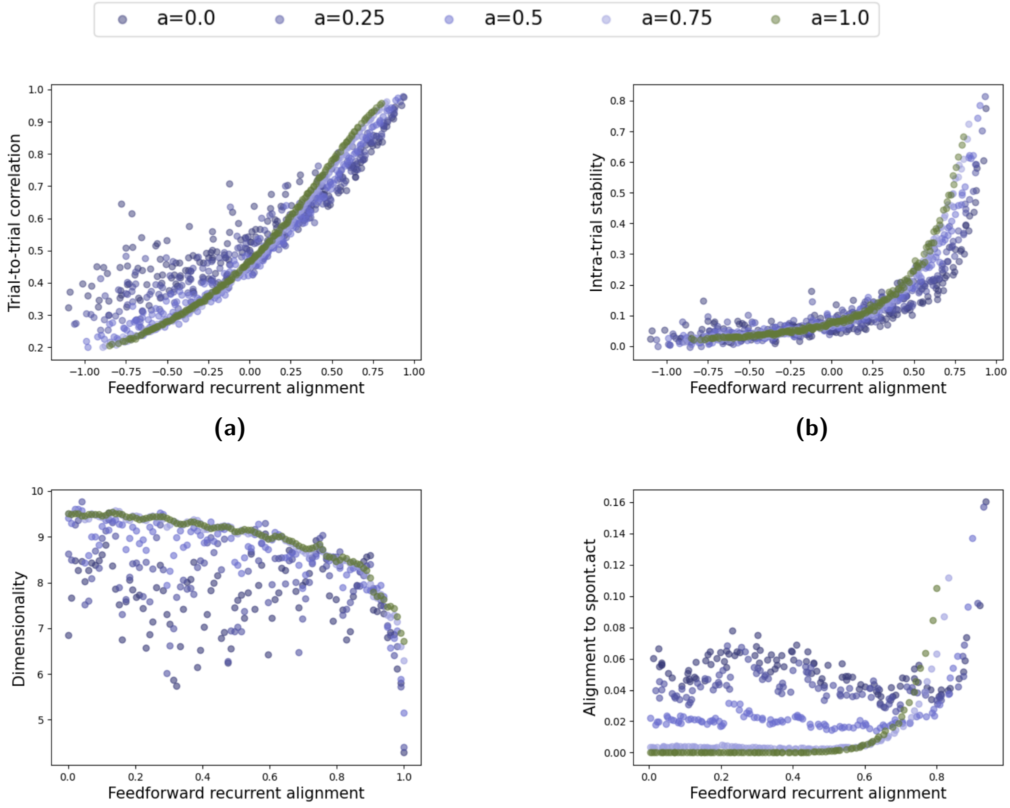
Multiple perspectives could lead to the discrepancy. One assumption for the discrepancy is that the influence of increased asymmetry in the network. Another reason could be that for principal components with small variance ratios, a single principal component cannot approximate the original eigenvectors as well as dominant principal components.

1.4.2 Iterative Feedforward Recurrent Alignment from Low-dimensional Inputs Indicates Alignment Improvement

With experimental producible low-dimensional inputs, we aim to discover how the RNNs amplify those inputs and if the amplification can provide new insights about alignment development.

As described by the method section **??**, low-dimensional inputs are modeled with eq.(**??**). Under the feedforward recurrent alignment hypothesis, the feedforward recurrent alignment score defined by eq.(**??**) should reflect how well the input is aligned to dominant activity patterns for generating reliable neural responses. Repeatedly applying the prior evoked activity pattern as inputs could reflect possible structural change after recurrent amplification thus indicating a possible plasticity adaptation of response activity patterns.

The updated dynamic of alignment in dependence of times for repeatedly applying prior response as inputs is an oscillation shown in Figure 1.19. At the n-th times using the prior response r_{n-1} as inputs, the feedforward recurrent alignment score ν_n defined in eq.(**??**) only depends on r_{n-1}. The oscillation can only be a consequence of the oscillation of evoked responses. So, repeatedly applying the prior response as feedforward input can lead to a stable oscillation of response activities r_n and therefore also the feedforward recurrent alignment score.



(a)

(b)

(c)

(d)

Figure 1.18 Correlation between feedforward recurrent alignment and selected response properties under aligning inputs to white-noise-evoked activity pattern. With inputs aligned to white-noise-evoked spontaneous-like activity, the modified feedforward recurrent alignment eq.(**??**) aligns inputs to principal components of white-noise-evoked activity pattern. The correlations between response properties and feedforward recurrent alignment are modeled with varied degrees of symmetry a for RNNs eq.(**??**) from $a = 0$ complete asymmetric to $a = 1.0$ complete symmetric.

(a) Correlation between feedforward recurrent alignment (x-axis) and trial-to-trial correlation (y-axis). **(b)** Correlation between feedforward recurrent alignment (x-axis) and intra-trial stability (y-axis). **(c)** Empirical approximation of effective dimensionality (y-axis) in correlation with feedforward recurrent alignment (x-axis). **(d)** Relationship between feedforward recurrent alignment (x-axis) and alignment to spontaneous activity (y-axis).

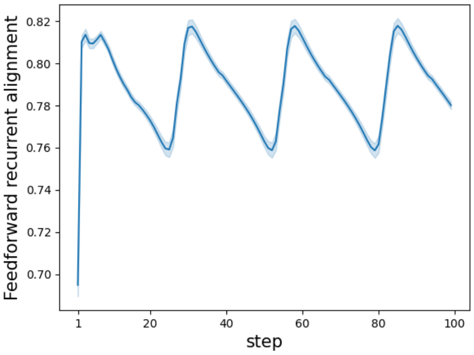


Figure 1.19 Dynamic of iterative feedforward recurrent alignment through applying prior response as input. Starting with low-dimensional input eq.(**??**) with $\beta_{\text{Low}} = 5$, the evoked activity pattern is applied as input for updating feedforward recurrent alignment. Iteratively repeating this procedure results in successive updates of feedforward recurrent alignment score eq.(**??**). For statistics 500 response samples are considered. The shadow indicates the 95% confidence interval.

1.5 Hebbian Learning of Feedforward Network Leads to Better Alignment between Feedforward Input and Recurrent Network

Now considering the different connection structures inside neural layers in the brain, we extend the focus from only on the recurrent network to also take the feedforward network structure into account, which delivers the feedforward inputs to the recurrent network illustrated in Figure **??**.

The main difference compared to prior theoretical experiments is, that the feedforward recurrent network allows plasticity in the feedforward network. Furthermore, we want to verify the feedforward recurrent alignment hypothesis during the network learning process. That is the change of alignment between dynamic feedforward input and recurrent network.

For firstly a basic understanding of the influence of plasticity, we consider the feedforward network with single neuron input and fixed symmetric recurrent network illustrated in Figure **??**). Only the feedforward network is updated with the Hebbian rule, which is the classic rule for activity-dependent synaptic plasticity [**?**]. The Hebbian rule describes the dynamics of feedforward weights through an ordinary differential equation eq. **??**. Moreover, the rule reflects the hypothesis for a principle: neurons that fire together wire together. With the help of the Euler scheme, the time development of feedforward weight can be approximated eq.(**??**).

To track the alignment between feedforward input and recurrent network, two alternatives were regarded:

- Based on the time-dependent feedforward weight dynamics, project weights to the space spanned by eigenvectors of recurrent networks. Observe the projection coefficients distribution for dominant eigenvectors of recurrent networks.

- Update the feedforward recurrent alignment score simultaneously with a time-related update of feedforward input. Observe the development of feedforward recurrent alignment score in dependence on time.

1.5.1 Feedforward Weights are Determined by Dominant Eigenvectors after Learning

Projecting the feedforward weights to the space spanned by eigenvectors of recurrent networks results in a linear combination as the presentation of feedforward weights through eigenvectors with suitable coefficients eq. **??**. The coefficients reflect the influence of corresponding eigenvectors on feedforward weights. A larger absolute value of one projection coefficient indicates a stronger influence of the corresponding eigenvector on feedforward weights.

According to the feedforward recurrent alignment hypothesis, a large alignment between feedforward input and recurrent network can be obtained by aligning feedforward input proportional to dominant eigenvectors. Therefore, feedforward weights with large coefficients for dominant eigenvectors would quantify a good alignment between feedforward input and recurrent network. Since the projection of feedforward weights could consist of multiple dominant eigenvectors, the projection ratio defined by eq.**??** takes the projection coefficients of the first twenty most dominant eigenvectors into account and could thus capture the dominance from more patterns.

The modeling of the projection ratio follows the update of feedforward weights according to the eq.**??**. If during the learning, the feedforward network can generate inputs fitting better to the dominant eigenvectors of the recurrent network, the feedforward weight should concentrate on dominant eigenvectors. Thus, the coefficients for dominant eigenvectors would be expected to be larger during learning. As a result, the projection ratio should increase with the development of time. According to the feedforward recurrent hypothesis, this could lead to more reliable recurrent responses.

The results illustrated by Figure 1.20 support the assumption that during learning, the feedforward input aligns better with the recurrent network by strengthening the influence of dominant eigenvectors of the recurrent network. At least the first twenty most dominant eigenvectors gain more weights in the linear combination for the projection during the Hebbian learning process of the feedforward network. Until the stable state, almost all projection weights concentrate at the first twenty dominant eigenvectors since the projection ratio reaches almost 1.

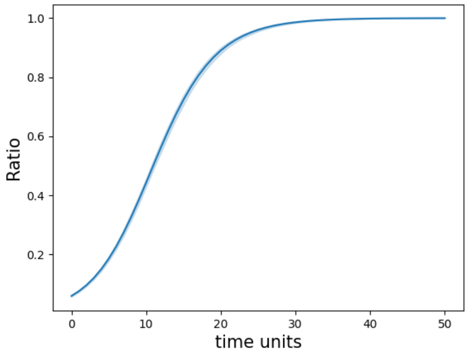


Figure 1.20 Time related development of projection ratio. Projection ratio defined by eq.**??** quantifies the strength of linear dependency between feedforward weights and the first twenty most dominant eigenvectors of the recurrent network. With the time-dependent update of feedforward weights eq.(**??**), the projection coefficients are also updated synchronously. Step width Δt for the Euler scheme is 0.1 over the total duration of $T = 50$ time units. For statistics, 50 repeats with different initial feedforward weights were implemented. The shadow indicates the 95% confidence interval.

1.5.2 Feedforward Recurrent Alignment Score Increases through Learning

Another alternative considers the change of feedforward recurrent alignment score directly over time simultaneously with the update of feedforward weights according to the eq.(**??**). Besides, the derivative of the feedforward recurrent alignment score dependent on time can also be explicitly formulated with the help of the Hebbian learning dynamic eq. (**??**), resulting the final derivative described by eq.(**??**).

According to the results from the first alternative, the feedforward weights improve their alignment to the recurrent network over time. The feedforward recurrent alignment score should increase if the feedforward inputs align better with the recurrent network. Since feedforward inputs are proportional to the feedforward weights due to the single input rate eq. **??**, the development of the feedforward recurrent alignment score should be synchronized to the dynamic of projection ratio shown in Figure 1.20.

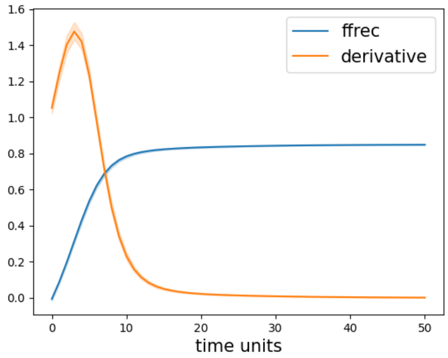


Figure 1.21 Dynamics of feedforward recurrent alignment score and its derivative. Simultaneously with the update of feedforward weight, the feedforward recurrent alignment score (shown with blue line) can also be updated simultaneously with eq.(**??**). The derivative of the feedforward recurrent alignment score (shown with orange line) is determined by eq.(**??**). Step width Δt for the Euler scheme is 0.1 over a total duration of $T = 50$ time units. For statistics, 50 repeats with different initial feedforward weights were implemented.

Implementation of feedforward recurrent alignment score eq.(**??**) and its derivative eq. (**??**) along time coincide with the results from the first alternative in Figure 1.20. As shown in Figure 1.21, the feedforward recurrent alignment score also increases over time until stable state. This fulfills our expectations that the two alternatives reflect both the phenomenon of increased alignment between feedforward input and recurrent network during the Hebbian learning considering a single input rate.