SUPPLEMENTARY INFORMATION

1. Supplementary Methods

Here, we provide further details for the four categories of simulated models that we used for validating our data analysis pipeline (see Methods).

Models of decisions

We simulated three distinct models of decisions governed by Equation 42 (see Methods), which are described below

Saddle node (Unstable Integration)

This model implemented binary decision making through a saddle node instability. The latent state ($\mathbf{x}$) of the model was governed by the following non-linear recurrent dynamics (Equation 42, Methods):

$$\mathbf{F}(\mathbf{x}) \quad = \begin{bmatrix} -6x_{(1)}^3 + 20x_{(1)} \\ -10x_{(2)} \end{bmatrix} \tag{S1}$$

where, $\mathrm{x}_{(\mathrm{i})}$ is the *i*th dimension of the 2-dimensional latent state. $\mathbf{F}(.)$ was characterized by 3 distinct fixed points - an unstable fixed point (saddle node) at the origin: (0,0) where the dynamics are transiently stable and, two stable fixed points located symmetrically at (1.825,0) and (-1.825,0), where the dynamics are globally stable. Initial conditions were sampled from a zero-mean bivariate Gaussian distribution with covariance equal to a scaled identity matrix ($= \sigma_i^2.\mathbf{I}$), with σ_i^2 = 0.005. On each trial *k*, the input drive at time *t* was defined as:

$$\mathbf{u}_t(k) = c(k).\begin{bmatrix} 8 \\ -8 \end{bmatrix} \tag{S2}$$

where, $\mathrm{c}(k) \in \{-1,1\}$ determined the sign of the input on each trial (choice 1 trials: $\mathrm{c}(k)$ = +1, choice 2 trials: $\mathrm{c}(k)$ = -1) and was sampled from a binomial distribution. Hence, the input drive was condition dependent but *time-invariant*. The latent noise ($\boldsymbol{\epsilon_t}$) was a zero-mean gaussian disturbance, whose covariance was the scaled identity matrix ($= \sigma_e^2.\mathbf{I}$), with σ_e^2 = 0.0001. The covariance of the observation noise ($\boldsymbol{\eta_t}$) was also set to a scaled identity matrix ($= \sigma_n^2.\mathbf{I}$), with σ_n^2 = 0.000001.

Line attractor (Perfect Integration)

This model integrated sensory evidence along a line attractor, which was implemented using a linear dynamical process:

$$\mathbf{F}(\mathbf{x}) \quad = \mathbf{A}.\mathbf{x} = \begin{bmatrix} 0 & -5 \\ 0 & -5 \end{bmatrix} \mathbf{x} \tag{S3}$$

The dynamics matrix ($\mathbf{A}$) has two distinct eigenvalues equal to 0 and -5 and the corresponding eigenvectors are $[1\ 0]^{\mathrm{T}}$ and $[1/\sqrt{2}\ \ 1/\sqrt{2}]^{\mathrm{T}}$ respectively, the former specifying the local direction of the line attractor. The dynamics were non-normal, as the two eigenvectors were non-orthogonal. The input drive at time *t* ($\mathbf{u}_{\mathrm{t}}$) was chosen independently for each condition choice 1 or choice 2) such that the condition average at the next time step had an exact match to the corresponding condition averaged response simulated from the saddle point model. The covariance of the latent noise ($\boldsymbol{\epsilon}_{\mathbf{t}}$) was a scaled identity matrix ($= \sigma_e^2.\mathbf{I}$), with σ_e^2 = 0.00003. The variability in the initial conditions (parameterized by $\sigma_i^2\mathbf{I}$) and the observation noise covariance ($\sigma_n^2\ \mathbf{I}$) were the same as in the saddle node model.

The augmented line attractor model (Fig 5) was characterized by a 4-dimensional latent state i.e 2 additional latent dimensions (eigenvalues = -7.5, -10), and governed by the following dynamics (similar to Equation 42, Methods):

$$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}) + \mathbf{u}_t + \boldsymbol{\epsilon}_{\mathbf{t}} = \begin{bmatrix} 0 & -5 & 0 & 0 \\ 0 & -5 & 0 & 0 \\ 0 & 0 & -7.5 & 0 \\ 0 & 0 & 0 & -10 \end{bmatrix} \mathbf{x} + \begin{bmatrix} \mathrm{u}_1(\mathrm{t}) \\ \mathrm{u}_2(\mathrm{t}) \\ \mathrm{u}_3(\mathrm{t}) \\ \mathrm{u}_4(\mathrm{t}) \end{bmatrix} + \boldsymbol{\epsilon}_{\mathbf{t}} \tag{S4}$$

where, the input components $\mathrm{u}_1(\mathrm{t})$ and $\mathrm{u}_2(\mathrm{t})$ were the same as defined previously for the line attractor model, and components $\mathrm{u}_3(\mathrm{t})$ and $\mathrm{u}_4(\mathrm{t})$ were defined as sinusoidal functions, $20 \times cos(1.6\pi t)$ and $20 \times sin(1.6\pi t)$ respectively. The noise parameters of the model (σ_i^2, σ_e^2, σ_n^2) were the same as in the line attractor model.

Point Attractor with time-varying inputs (Leaky Integration)

A 'leaky' model of evidence integration was implemented with the following linear dynamics:

$$\mathbf{F}(\mathbf{x}) \quad = \mathbf{A}.\mathbf{x} = \begin{bmatrix} -20 & 0 \\ 0 & -40 \end{bmatrix} \mathbf{x} \tag{S5}$$

The matrix $\mathbf{A}$ is normal (orthogonal eigenvectors) and has two distinct strongly stable eigenvalues equal to -20 and -40. The input drive at time *t* ($\mathbf{u}_{\mathrm{t}}$) was once again chosen such that the condition average at the next time step had an exact match to the corresponding condition averaged

response of the saddle node model. The various noise parameters of the model ($\sigma_i^2, \sigma_e^2, \sigma_n^2$) were the same as in the line attractor model.

Models of movement

We simulated three distinct models of movement (also governed by Equation 42, Methods) described below

Rotational Dynamics

We implemented a model of movement characterized by rotational dynamics. The recurrent dynamics $\mathbf{F}$ was specified in terms of polar coordinates $[\mathrm{r}\ \phi]$:

$$\begin{aligned} r &= \sqrt{x_{(1)}^2 + x_{(2)}^2} \\ \phi &= tan^{-1}\left(\frac{x_{(2)}}{x_{(1)}}\right) \\ \dot{r} &= 0 \\ \dot{\phi} &= 1.2r \end{aligned} \tag{S6}$$

The latent state ($\mathbf{x}_\mathrm{t}$) on each 'trial' (obtained from converting polar into cartesian coordinates), was a result of purely autonomous dynamics set by an initial condition that differed across trials. The initial condition on the k^{th} trial was sampled from a Gaussian distribution as follows:

$$\mathbf{x}_0^k \sim \mathcal{N}\left(\begin{bmatrix} c(k) \\ 0 \end{bmatrix}, \begin{bmatrix} 0.01 & 0 \\ 0 & 0.0001 \end{bmatrix}\right) \tag{S7}$$

where, $\mathrm{c}(k)$ determined the sign of the mean of the random initial condition vector on each trial (similar to $\mathrm{c}(k)$ in the saddle node dynamics model). Thus, differences in initial conditions across trials determined the two types of conditions (choice 1 trials: = $\mathrm{c}(k)$ = +1, choice-2 trials: = $\mathrm{c}(k)$ = -1). The covariances of the latent noise ($\boldsymbol{\epsilon}_\mathbf{t}$) and observation noise ($\boldsymbol{\eta}_t$) were scaled identity matrices $\sigma_e^2.\mathbf{I}$ and $\sigma_n^2.\mathbf{I}$ respectively, with σ_e^2 = 0.00003 and σ_n^2 = 0.000001.

We also simulated an augmented rotational dynamics model with two additional latent dimensions. The recurrent dynamics (eigenvalues = -7.5 and -10) and the inputs ($\mathrm{u}_3(\mathrm{t})$ and $\mathrm{u}_4(\mathrm{t})$) associated with these two additional dimensions were the same as defined previously for the augmented line attractor model (Equation S4).

Dynamic Attractor

We constructed a dynamical system with a stable attracting circular trajectory. We parameterized the dynamics in terms of polar coordinates as follows:

$$\begin{aligned} \rho &= 1 - r \\ \dot{\rho} &= -20\rho \\ \dot{\phi} &= 1.2(1-\rho) \end{aligned} \tag{S8}$$

The initial conditions for all trials were the same as the initial conditions generated for the rotational dynamics model. The model was autonomous (no external inputs). The covariances of the latent noise ($\boldsymbol{\epsilon_t}$) and observation noise ($\boldsymbol{\eta_t}$) were scaled identity matrices $\sigma_e^2.\mathbf{I}$ and $\sigma_n^2.\mathbf{I}$ respectively, with σ_e^2 = 0.0001 and σ_n^2 = 0.000001.

Point Attractor with time-varying inputs (input-driven movement dynamics)

The recurrent dynamics ($\mathbf{F}$) was the same as that of the leaky integration model (Equation S5, see models of decisions). However, unlike the rotational dynamics and dynamic attractor models, this model was input driven. The input drive at each time t ($\mathbf{u}_t$) was chosen independently for each condition, such that the condition average at the next time step had an exact match to the corresponding condition averaged response of the rotational dynamics model. The initial conditions were the same as those generated for the rotational dynamics model. The covariances of the latent noise ($\boldsymbol{\epsilon_t}$) and observation noise ($\boldsymbol{\eta_t}$) were scaled identity matrices $\sigma_e^2.\mathbf{I}$ and $\sigma_n^2.\mathbf{I}$ respectively, with σ_e^2 = 0.0003 and σ_n^2 = 0.000001.

Linear state-space models with uncorrelated latent noise

These models were governed by linear time-varying dynamics (Equation 11, Methods) but using a continuous time latent state update (as in Equation 35, Methods). The observation models were either linear gaussian (Equation 11, Methods) or poisson as described below:

Linear-Gaussian Models

The models were characterized by 3 latent dimensions and 20 observed dimensions. For models characterized by *time-invariant latent noise*, the latent noise covariance was a scaled identity matrix ($\mathbf{Q} = 0.5\mathbf{I}$). The observation noise matrix $\mathbf{R}$ was diagonal, with elements sampled from a uniform distribution between $[0,0.01]$. The steady-state initial noise covariance ($\mathbf{Q}_0$) was obtained by solving the continuous-time Lyapunov equation. The above choice of $\mathbf{Q}$, $\mathbf{R}$ and $\mathbf{Q}_0$ resulted in simulated data with variability that matched the level of variability in the neural data (Extended Data Fig. 3e).

The other simulation parameters were the same as those used for the simulations of the time-invariant linear dynamical models used to determine the optimal bin-size (see Methods, Equation 35). We simulated three different types of latent, time-varying dynamics matrices:

1. *Switching eigenvalues (non-rotational)*

The time-varying dynamics matrices of this model were normal (orthogonal eigenvectors) but showed an abrupt switch in the eigenvalue spectrum (Extended Data. Fig. 3a-c, top row). The eigenvectors of $\mathbf{A}_t$ were time-invariant. The elements of the eigenvector matrix were sampled from a standard normal distribution and then orthogonalized with respect to one another. The eigenvalue spectrum was always real and switched from $[-1, -3, -5]$ to $[-2, -4, -6]$ midway through the trial.

2. *Switching eigenvectors*

   The dynamics matrices of this model had eigenvalues $[-1, -3, -5]$ which did not vary in time. Instead the associated eigenvector basis exhibited a switch from an orthogonal ('normal' dynamics) to a non-orthogonal eigenvector basis ('non-normal' dynamics) mid-way through the trial (Extended Data. Fig. 3a-c, middle row).

3. *Switching eigenvalues (rotational)*

   The time-varying dynamics matrices of this model (Extended Data. Fig. 3a-c, bottom row) were normal (orthogonal eigenvectors) but showed an abrupt switch in the eigenvalue spectrum from real - $[-1, -3, -5]$ to complex-valued eigenvalues - $[-2 + \mathrm{i}2\pi, -2 - \mathrm{i}2\pi, -5]$ mid-way through the trial. The complex eigenvalues were associated with 'slow' rotations of 1Hz frequency.

The models characterized by *time-varying* latent noise differed only in a single aspect. They were characterized by latent noise covariance that was itself time-dependent ($\mathbf{Q}_t$ instead of $\mathbf{Q}$, as in Equation 35 of Methods). The latent noise covariance switched from $0.5\mathbf{I}$ to $0.3\mathbf{I}$ midway through the trial, locked to the corresponding change in the latent dynamics ($\mathbf{A}_t$).

Poisson Models

Models described by a poisson observation process shared the same time-varying, latent dynamics processes as described above, but were instead associated with 100 observed dimensions. The poisson observations ($\mathbf{s}_t$), which corresponded to spike counts in 1ms bins were obtained as:

$$\begin{aligned} \mathbf{y}_t &= \mathbf{C}\mathbf{x}_t + \mathbf{d} \\ \mathbf{s}_t &\sim Poisson(exp(\mathbf{y}_t)) \end{aligned} \tag{S9}$$

where, $\mathbf{x}_t$ is the latent state and the observation matrix $\mathbf{C}$ ($\in \mathcal{R}^{100 \times 2}$) was a random (elements sampled from a standard normal distribution), orthogonal matrix. The elements of the baseline input vector $\mathbf{d}$ were chosen uniformly at random in the range [-4.8 -4.1], while the latent noise

covariance was time-invariant and set to a scaled identity matrix - $\mathbf{Q} = 12500\mathbf{I}$. These choice of $\mathbf{Q}$ and $\mathbf{d}$ ensured realistic, over-dispersed fano factors between 1 and 1.5, and single-trial variability that matched the level of variability in the neural data (Extended Data Fig. 3e). The steady-state initial noise covariance ($\mathbf{Q}_0$) was obtained by solving the continuous time lyapunov equation.

To match the overall level of variability in the simulated residuals to those measured in the data, we computed three different measures that specified (i) the total magnitude of instantaneous variability ($l_0 = \langle \left\| \tilde{\mathbf{Z}}_t \tilde{\mathbf{Z}}_t{}' \right\|_F^2 \rangle_t$), (ii) the total magnitude of lagged co-variability ($l_1 = \langle \left\| \tilde{\mathbf{Z}}_{t+1} \tilde{\mathbf{Z}}_t{}' \right\|_F^2 \rangle_t$) and, (iii) the strength of the most dominant mode of lagged co-variability ($\mathrm{pvar} = \langle \sigma_{max}(\tilde{\mathbf{Z}}_{t+1} \tilde{\mathbf{Z}}_t{}') \rangle_t$) in the simulated and measured residuals, where, $\tilde{\mathbf{Z}}_t$ is the aligned residual data matrix at time *t* (see Methods) of dimensionality $20 \times K$, where K is the number of trials, $\| . \|_F$ is the frobenius norm, and $\sigma_{max}(.)$ corresponds to the largest singular value of its matrix-valued argument. The latent noise parameter ($\mathbf{Q}$) of the linear gaussian/poisson models, the observation noise parameter ($\mathbf{R}$) of the linear gaussian models, and the mean offset ($\mathbf{d}$) of the poisson model simulations were chosen such that these measures of variability qualitatively matched to those computed using PFC neural residuals.

Linear state-space models with correlated latent noise

We explored three distinct sub-categories of models (described below), all specified by Equation 43, but which differed in the nature of $\mathbf{A}$ and $\boldsymbol{\phi}$ (i.e. either stable or rotational). To simplify our analyses, we only considered models with 2-dimensional dynamics (i.e. Dim($\mathbf{x}$) = Dim($\boldsymbol{\varepsilon}$) = 2; Equation 43, Methods), although our analytical derivations (Supplementary Math Note B) were applicable for arbitrary dimensionalities. For each sub-category, we generated multiple instantiations of the model by systematically sweeping the magnitude and phase of the eigenvalues of $\mathbf{A}$ and $\boldsymbol{\phi}$. For all models, the covariance of the latent noise process $\boldsymbol{\zeta}(t)$ was set to a scaled identity matrix ($\mathbf{Q} = \sigma_q^2 . \mathbf{I}$), with σ_q^2 sampled uniformly at random in the range [10^{-5}, 10^{-7}], separately for each model instantiation.

1. *Inflation of eigenvalues (Extended Data Fig. 4a-b)*

   We instantiated 5000 distinct models within this category by varying the eigenvalue magnitudes of $\mathbf{A}$ and $\boldsymbol{\phi}$. For all 5000 models, we set one of the eigenvalues of $\mathbf{A}$ to a fixed constant such that it was associated with a decay time constant of 400ms. We swept the other eigenvalue (λ_2, Extended Data. Fig. 4a-b) on a uniform grid in the range [0.4 0.999], resulting in 1000 distinct $\mathbf{A}$ matrices. Both eigenvalues of $\boldsymbol{\phi}$ were constrained to be equal to one another (λ_1, Extended Data. Fig. 4a-b) and were sampled in 5 discrete steps ($\lambda_1 = [0.40\ 0.75\ 0.91\ 0.97\ 0.99]$), which resulted in 5 distinct $\boldsymbol{\phi}$ matrices. The eigenvectors of

each $\mathbf{A}$ and $\boldsymbol{\phi}$ were the cardinal unit vectors $\mathbf{e}_1$, $\mathbf{e}_2$ of the 2D space. We also simulated 5 additional models, in which, λ_2 (eigenvalue specifying $\mathbf{A}$) was set to zero specifying a scenario, in which, recurrent dynamics only unfolds upstream within the input area (Extended Data Fig. 4b).

2. *Inflation of rotation frequency (Extended Data Fig. 4c-d)*

   We instantiated 5000 distinct models within this category by varying the eigenvalue phase of $\boldsymbol{\phi}$ and the eigenvalue magnitude of $\mathbf{A}$ (Extended Data. Fig. 4c-d). Since $\boldsymbol{\phi}$ was associated with rotational dynamics, it was associated with a pair of complex-conjugate eigenvalues, which could be completely specified using its magnitude and phase. For all 5000 models, we set the magnitude of the complex-conjugate eigenvalue pair to a fixed constant that translated into a decay time constant of 1s. We instead systematically varied the phase of the complex-conjugate eigenvalue pair in 5 discrete steps, resulting in 5 distinct $\boldsymbol{\phi}$ matrices that had associated rotation frequencies (ω_1 , Extended Data. Fig. 4d) equal to [0.5 0.75 1.00 1.25 1.5] Hz. Both eigenvalues of $\mathbf{A}$ were constrained to be equal to one another (λ_2, Extended Data. Fig. 4c-d) and sampled from a uniform grid in the range [0.4 0.999], resulting in 1000 distinct $\mathbf{A}$ matrices. The eigenvectors of $\mathbf{A}$ were the cardinal unit vectors $\mathbf{e}_1$, $\mathbf{e}_2$ of the 2D space. The eigenvectors of $\boldsymbol{\phi}$ were complex-conjugate and were chosen such that they characterized by a circularly symmetric rotational flow-field**.**

3. *Equivalence of upstream and local recurrent dynamics (Extended Data Fig. 4e-f)*

   We instantiated 5000 distinct models within this category by varying the eigenvalue phase of $\mathbf{A}$ and the eigenvalue magnitude of $\boldsymbol{\phi}$, in a manner that was mirror symmetric to the way the parameters were specified above (Extended Data Fig. 4c-d).

In addition to the above analytical steady-state analyses, we also simulated data with a 1-dimensional latent state (i.e. Dim($\mathbf{x}$) = Dim($\boldsymbol{\varepsilon}$) = 1; Equation 43, Methods) in order to provide a graphical explanation for the inflation of the estimates of residual dynamics. We simulated 1000 trials/trajectories in steps of 45ms starting from a base scenario (Extended Data Fig. 5a-b) where the $\mathbf{x}$ (Equation 43, Methods) was driven by input noise that was not temporally autocorrelated (i.e $\boldsymbol{\phi}$ = 0; Equation 43, Methods) and with a fixed variance (Var ($\boldsymbol{\zeta}(t)$) = 1e-6; Equation 43, Methods). We then either switched (t=0 is denoted as time of switch in Extended Data Fig 5) to a condition - (i) where the latent noise ($\boldsymbol{\varepsilon}(t)$; Equation 43, Methods) was autocorrelated ($\boldsymbol{\phi}$ = 0.798, Extended Data Fig. 5c-d), or (ii) where the variance was increased (by a factor of 10) to 1e-5 (Extended Data Fig. 5e-f).

Modular two-area recurrent neural network

Below, we describe the network configurations used in this study, the details of applying and analyzing residual dynamics in these models, and simulated perturbation experiments performed using these models. As described in the Methods, the RNN has two nodes/areas (PPC and PFC) characterized with both intra-areal (within area) and inter-areal (across areas) excitatory and inhibitory connections (see ref. 38 of main text for details of the architecture, and network parameters defined below)

Network Configurations

The intra-areal recurrent connectivity was determined by weight matrices $\mathbf{J}^{ppc}$ and $\mathbf{J}^{pfc}$, whereas the inter-areal connectivity was determined by feed-forward weight matrix $\mathbf{J}^{ppc \rightarrow pfc}$ and feedback weight matrix $\mathbf{J}^{ppc \leftarrow pfc}$. Each of the above four weight matrices were directly determined by 4 scalar parameters $\mathrm{J}_s{}^{ppc}$, $\mathrm{J}_s{}^{pfc}$, $\mathrm{J}_s{}^{ppc \rightarrow pfc}$ and $\mathrm{J}_s{}^{ppc \leftarrow pfc}$, which modulated the coupling strength of the connectivity matrices. We constrained all our analyses to a setting where the within area recurrent weights of each area were the same (i.e $\mathbf{J}^{ppc} = \mathbf{J}^{pfc}$, by setting $\mathrm{J}_s{}^{ppc}$ = $\mathrm{J}_s{}^{pfc}$ = J_{self}). We simulated two different classes of networks, one in which feedback was absent (i.e $\mathbf{J}^{ppc \leftarrow pfc} = \mathbf{0}$, by setting $\mathrm{J}_s{}^{ppc \leftarrow pfc}$ = 0) and, another in which feedback was present and equaled the feed-forward connection strength (i.e $\mathbf{J}^{ppc \rightarrow pfc} = \mathbf{J}^{ppc \leftarrow pfc}$, by setting $\mathrm{J}_s{}^{ppc \rightarrow pfc}$ = $\mathrm{J}_s{}^{ppc \leftarrow pfc}$ = J_{across}).

We simulated 6000 trials from each network, each trial spanning a total of 1.2s (steps of 0.1ms, Euler integration). Each population within an area also received noisy current inputs determined by an Ornstein-Uhlenbeck (OU) process whose variance σ_{noise} was set to 0.07. The observation matrices specific to each area (Equation 44, Methods; $\mathbf{C}_{\mathrm{ppc}}, \mathbf{C}_{\mathrm{pfc}} \in \mathcal{R}^{10 \times 2}$) were specified as random orthogonal matrices. The covariance of the multivariate, gaussian observation noise process ($\boldsymbol{\eta}_{\mathrm{t}}$) was isotropic (variance = 0.0006).

Residual dynamics estimation

For 'local' estimates of residual dynamics, the cross-validation of the hyper-parameters (see Step 3 of Extended Data Fig. 2) yielded dynamics of dimensionality *d* = 2 for all network configurations, which was consistent with the ground truth. For the 'global' estimates, depending on the underlying network configuration, we recovered residual dynamics with dimensionalities *d* in the range 2-4. As in the neural data (Extended Data Fig. 8), we consistently recovered an optimal lag (l_{opt}) of either 3 or 4 across most model configurations. Therefore, for the final model fits (Figs. 6,7), we used l_{opt} = 3 for all model configurations. The smoothness hyper-parameter (α in Equation 27, Methods) of the second stage of 2SLS (see Step 4 of Extended Data Fig. 2) was uniquely determined for each model using cross-validation.

Overlap of task-relevant modes with residual dynamics

For each model configuration (pair of J_{self} and J_{across} for each network type), we extracted the set of estimated residual eigenvectors at a specific time instant t_{max}^{pfc}, defined as the time instant at which the maximum eigenvalue magnitude was attained 'locally' within PFC (t_{max}^{pfc} = $\underset{t}{\text{argmax}} \max_{j} |\lambda_j^{pfc}(t)|$, where j indexes the j^{th} eigen-mode). The overlap between residual eigenvectors and the task-relevant model dimensions was quantified as the subspace angle between the pair of vectors. To ensure vectors of a consistent dimensionality, we projected the 4-dimensional task-relevant dimensions (Equation 45, Methods) through the ground-truth observation matrix $\mathbf{C}_{\text{model}}$ (Equation 44, Methods); and the estimated 'local' and 'global' residual eigenvectors through the corresponding estimate of $\mathbf{C}_{\text{model}}$ (i.e dynamics subspace).

Simulated perturbation experiments

Perturbations corresponded to small pulse-like current injections within a specific area along one of the two distinct task-relevant, dimensions ("choice" or "time"; Equation 45, Methods) defined for that area. The magnitude of each perturbation was set to a fixed constant of 0.001nA (duration of 0.1ms), chosen such that the resulting perturbed state lay within the realm of single-trial neural variability (in the absence of a perturbative input). Perturbations were applied at six different times in the trial ([270 405 540 675 810 945]ms after trial onset). For each network configuration, we simulated 6000 'unperturbed' trials, and 6000 'perturbed' trials for 24 distinct perturbation conditions (4 directions *x* 6 times). Each 'perturbed' trial was paired with a corresponding 'unperturbed' trial, with which it shared the time course of the latent noise and input noise.

To compute the ground truth impulse response for a specific perturbation condition, we sorted the 'unperturbed' trials into two choice conditions, based on the single-trial readout along $\boldsymbol{u}_{choice}^{pfc}$ (Equation 45, Methods) at the last time of the trial, and computed trial-averaged trajectories specific to each area, for each choice condition (choice-1 and choice-2). Likewise, we sorted the 'perturbed' trials in each area based on the above assignment of choice (determined for the unperturbed trials) and computed 'perturbed residuals' by subtracting from each 'perturbed' single trial the corresponding 'unperturbed', choice-specific trial-averaged trajectory. The ground truth impulse response in a given area was then defined as the mean across trials of the 'perturbed residuals' specific to that area.

The impulse response $\mathbf{v}_{prop}(t')$ predicted by the residual dynamics was defined as:

$$\mathbf{v}_{prop}(t') = \mathbf{U}_{dyn}^{d_{opt}} . (\mathbf{A}_{t'} . \mathbf{A}_{t'-\Delta t} . \dots \mathbf{A}_{t_0+\Delta t}) . \mathbf{U}_{dyn}^{d_{opt}\prime} \mathbf{v} \qquad \text{(S10)}$$

where, Δt corresponds to the temporal bin size, t_0 corresponds to the time at which the perturbation is applied, $\mathbf{U}_{dyn}^{d_{opt}}$ is the estimate of the dynamics subspace and, $\mathbf{v}$ corresponds to the initial condition. We estimated residual dynamics ($\mathbf{A}_t$) using only 'unperturbed' trials, separately for each choice condition. We constructed predicted impulse responses based on 'local' and 'global' residual dynamics. The initial condition $\mathbf{v}$ was the ground truth impulse response computed at the exact time instant of the corresponding perturbation. In other words, here we did not attempt to predict $\mathbf{v}$ itself (i.e. activity during the current injection), but rather how activity in the simulated networks evolved from $\mathbf{v}$ after the end of the current injection.

To summarize the ground-truth and predicted impulse responses, we computed the norm of the impulse response and plotted it as a function of time in the trial (Fig. 8, colored points: ground truth, black/gray curves: predictions). We only reported the two-step look-ahead impulse responses, i.e we only plotted the impulse response at the time-instant of a given perturbation (which corresponds to $\mathbf{v}$ for both the prediction and the ground-truth) and the two time instants that immediately followed the perturbation ($t' = t_0 + 2\Delta t$, in Equation S10).

We also simulated impulse responses in simpler, two-dimensional linear time-invariant dynamical systems (Extended Data Fig 10) that reproduced key features of the estimated global residual dynamics in the two example networks (Extended Data Fig 10). The cardinal directions of the 2D state space that defined these linear dynamical systems corresponded to the choice modes in PPC and PFC respectively, and perturbations were applied along one of these cardinal directions to mimic 'local' perturbations applied to an area. The two models only differed in the arrangement of the two eigenvectors, but not in their eigenvalue magnitudes. For the feedforward model, the unstable eigenvector (EV_1, magnitude = 1.1; see Extended Data Fig 10) projected mostly onto the PPC choice mode (angle with PPC_c = 15 deg, Extended Data Fig 10), while the stable eigenvector (EV_3, magnitude = 0.5; see Extended Data Fig 10) was perfectly aligned with the PFC choice mode (angle with PFC_c = 0, Extended Data Fig 10). For the "feedback" model, both the unstable (EV_1, Extended Data Fig 10) and stable (EV_4, Extended Data Fig 10) eigenvectors were equally shared across PPC and PFC (angle of 45 deg with PPC_c and PFC_c; Extended Data Fig 10).

2. Supplementary Analyses

The table below summarizes the range (across all 4 task configurations, Extended Data Fig 6a) of the largest subspace angle (in degrees) between any pair of task activity subspaces ($\mathbf{U}_{task}^{i}$ and $\mathbf{U}_{task}^{j}$, see Methods) computed using the aligned response patterns (Methods, Extended Data Fig 2) for a specific epoch and monkey.

$\angle i, j$	$\angle jPC_{12}, choice$	$\angle jPC_{12}, time$	$\angle jPC_{34}, choice$	$\angle jPC_{34}, time$	$\angle choice, time$
Monkey 'T' decision epoch	[76.8, 85.0]	[57.3, 64.3]	[34.2, 50.7]	[73.9 88.2]	[76.3 89.2]
Monkey 'T' movement epoch	[58.5, 83.8]	[21.1, 39.1]	[31.1, 41.9]	[62.6, 79.6]	[79.3, 88.9]
Monkey 'V' decision epoch	[79.8, 87.5]	[78.6, 85.3]	[18.2, 49.7]	[74.7, 87.2]	[79.2, 88.2]
Monkey 'V' movement epoch	[65.1, 83.9]	[16.4, 36.1]	[42.6, 61.2]	[61.9, 89.0]	[76.2, 85.4]

3. Supplementary Math Note A

In this section, we provide a brief overview of subspace system identification (SSID) theory as applied to parameter identification of linear time-invariant (LTI) state-space models. We adapt linear, time-invariant SSID methods described below to models characterized by a linear time-varying latent dynamical process, such as the one used in our study to investigate residual dynamics (Equation 11, see Methods). A standard linear, time-invariant state-space model can be described as:

$$\begin{aligned} \mathbf{x}(\mathrm{t}+1) &= \mathbf{A}\mathbf{x}(\mathrm{t}) + \boldsymbol{\varepsilon}(\mathrm{t}) \\ \tilde{\mathbf{z}}(\mathrm{t}) &= \mathbf{C}\mathbf{x}(\mathrm{t}) + \boldsymbol{\eta}(\mathrm{t}) \end{aligned} \qquad (S11)$$

where, $\mathbf{A}$ and $\mathbf{C}$ are the dynamics and observation matrices respectively, $\mathbf{x}(\mathrm{t})$ is a $p \times 1$ latent state at time 't', $\tilde{\mathbf{z}}(\mathrm{t})$is a corresponding $n \times 1$ observation vector, and $\boldsymbol{\varepsilon}(\mathrm{t})$ and $\boldsymbol{\eta}(\mathrm{t})$ are zero-mean white gaussian noise processes. In our study, "observations" $\tilde{\mathbf{z}}(\mathrm{t})$ correspond to aligned neural residuals, computed by subtracting the condition-averaged trajectory from each corresponding aligned single-trial trajectory (see Methods).

A common approach for parameter estimation of models such as those described by Equation S11 involves probabilistic inference techniques. The expectation-maximization (EM) algorithm provides closed-form updates for the parameter estimates of an LTI model through the Kalman filtering and smoothing recursions[1]. However, estimating system parameters for time-varying dynamics, as in our model (Equation 11, see Methods), is intractable using the exact EM. Such models typically require approximate inference techniques[2,3] (e.g. variational inference) that rely on gradient optimization, which are susceptible to local minima in parameter space.

We therefore opted for an alternative approach based on subspace identification techniques[4,5]. These non-probabilistic methods rely on a series of matrix decompositions and linear regressions to estimate the system parameters in closed form. SSID methods are 'non-optimal' in a

probabilistic sense, in that, they do not explicitly model the uncertainty in the underlying latent state. However, these methods do not suffer from problems of local minima and can yield unbiased and consistent parameter estimates for large sample sizes.

The key idea underlying SSID methods is to build a linear predictor of the underlying latent state directly from the observed time series $\tilde{\mathbf{z}}$, by matching the empirical first and second order moments to those predicted by the model (Equation S11). This predictor, serves as a "bottleneck", summarizing all the information in the "past" of the observed time-series that is relevant to predicting the "future" of the time-series. For a given time *t* and trial *k*, we can define the finite "future" and "past" (relative to *t*) of the observed time series as follows:

$$\begin{aligned} \tilde{\mathbf{z}}_{\mathbf{t}}^{+}(k) &= [\tilde{\mathbf{z}}_{\mathrm{t}}(k)' \quad \tilde{\mathbf{z}}_{\mathrm{t+1}}(k)' \ldots \tilde{\mathbf{z}}_{\mathrm{t+q-1}}(k)']' \\ \tilde{\mathbf{z}}_{\mathbf{t}}^{-}(k) &= [\tilde{\mathbf{z}}_{\mathrm{t-1}}(k)' \quad \tilde{\mathbf{z}}_{\mathrm{t-2}}(k)' \ldots \tilde{\mathbf{z}}_{\mathrm{t-q}}(k)']' \end{aligned} \tag{S12}$$

where, $\tilde{\mathbf{z}}_{\mathrm{t}}(k)$ is the observation vector of dimensionality $\mathrm{n} \times 1$ at time *t* on trial *k* that populates the observation data matrix $\tilde{\mathbf{Z}}$ (of dimensionality $\mathrm{n} \times \mathrm{T} \times \mathrm{K}$, where T and K are the total number of time-bins and trials respectively). $\tilde{\mathbf{z}}_{\mathrm{t}}^{+}$ and $\tilde{\mathbf{z}}_{\mathrm{t}}^{-}$ respectively are the 'future' and 'past' observation vectors (each of dimensionality $(\mathrm{n.q}) \times 1$) constructed by stacking together *q* 'future' and, *q* 'past' lags of the observed time series relative to 't'.

An ordinary least squares linear predictor of the "future" ($\tilde{\mathbf{z}}_{\mathbf{t}}^{+}$) based on the past ($\tilde{\mathbf{z}}_{\mathbf{t}}^{-}$) is defined as:

$$\hat{\tilde{\mathbf{z}}}_{\mathbf{t}}^{+} = \left(\mathrm{Cov}(\tilde{\mathbf{z}}_{\mathrm{t}}^{+}, \tilde{\mathbf{z}}_{\mathrm{t}}^{-})\mathrm{Cov}^{-1}(\tilde{\mathbf{z}}_{\mathrm{t}}^{-}, \tilde{\mathbf{z}}_{\mathrm{t}}^{-})\right)\tilde{\mathbf{z}}_{\mathrm{t}}^{-} = \mathbf{H}_{\mathrm{t}}{\mathbf{G}_{\mathrm{t}}}^{-1}\tilde{\mathbf{z}}_{\mathrm{t}}^{-} \tag{$S13$}$$

The "future-past" covariance matrix $\mathbf{H}_t$ specified by $\mathrm{Cov}(\tilde{\mathbf{z}}_{\mathrm{t}}^{+}, \tilde{\mathbf{z}}_{\mathrm{t}}^{-})$ in Equation S13 is referred to as a "*hankel matrix*", and is a key element of SSID methods. The hankel matrix is in fact a block matrix of size $(\mathrm{n} \times \mathrm{q}) \times (\mathrm{n} \times \mathrm{q})$, with the individual blocks representing temporally lagged covariance matrices computed using observation vectors $\tilde{\mathbf{z}}$ at different time-lags (relative to *t*)

$$\mathbf{H}_{\mathrm{t}} = \mathrm{Cov}(\tilde{\mathbf{z}}_{\mathrm{t}}^{+}, \tilde{\mathbf{z}}_{\mathrm{t}}^{-}) = \begin{bmatrix} \tilde{\mathbf{Z}}_{\mathrm{t}}.\tilde{\mathbf{Z}}'_{\mathrm{t-1}} & \tilde{\mathbf{Z}}_{\mathrm{t}}.\tilde{\mathbf{Z}}'_{\mathrm{t-2}} & \ldots & \tilde{\mathbf{Z}}_{\mathrm{t}}.\tilde{\mathbf{Z}}'_{\mathrm{t-q}} \\ \tilde{\mathbf{Z}}_{\mathrm{t+1}}.\tilde{\mathbf{Z}}'_{\mathrm{t-1}} & \tilde{\mathbf{Z}}_{\mathbf{t+1}}.\tilde{\mathbf{Z}}'_{\mathrm{t-2}} & \ldots & \tilde{\mathbf{Z}}_{\mathrm{t+1}}.\tilde{\mathbf{Z}}'_{\mathrm{t-q}} \\ \vdots & & \ddots & \vdots \\ \tilde{\mathbf{Z}}_{\mathrm{t+q-1}}.\tilde{\mathbf{Z}}'_{\mathrm{t-1}} & \tilde{\mathbf{Z}}_{\mathrm{t+q-1}}.\tilde{\mathbf{Z}}'_{\mathrm{t-2}} & \ldots & \tilde{\mathbf{Z}}_{\mathrm{t+q-1}}.\tilde{\mathbf{Z}}'_{\mathrm{t-q}} \end{bmatrix} \tag{S14}$$

where, $\tilde{\mathbf{Z}}_{\mathrm{t}}$ is an individual slice of the observation data matrix $\tilde{\mathbf{Z}}$ indexed by time 't'. For LTI systems in steady state, stationarity implies that the "future-past" hankel covariance matrix in

Equation S14 is *time independent*, since the individual blocks of the hankel matrix are covariance matrices that only depend on the relative lag between temporal observations (see Equations S14 and S15) under stationary conditions.

The parameter '*q*' (referred to as the order of the hankel matrix) is a critical parameter required for accurate model identification of an LTI model. SSID theory specifies that *q* has to be chosen to be at least as large as the dimensionality of the true underlying latent state. An alternative interpretation of the above statement is that a given choice of *q* sets the upper bound for the dimensionality of the latent state and dynamics that one can recover based on the hankel matrix constructed out of the observations.

For an LTI system in steady-state, the time independent hankel matrix ($\mathbf{H}$) can be written as:

$$\mathbf{H} \quad = \mathrm{Cov}(\tilde{\mathbf{z}}_{\mathrm{t}}^{+}, \tilde{\mathbf{z}}_{\mathrm{t}}^{-}) = \begin{bmatrix} \boldsymbol{\Lambda}_1 & \boldsymbol{\Lambda}_2 & \dots & \boldsymbol{\Lambda}_q \\ \boldsymbol{\Lambda}_2 & \boldsymbol{\Lambda}_3 & \dots & \boldsymbol{\Lambda}_{q+1} \\ \vdots & & \ddots & \vdots \\ \boldsymbol{\Lambda}_q & \boldsymbol{\Lambda}_{q+1} & \dots & \boldsymbol{\Lambda}_{2q-1} \end{bmatrix} \tag{S15}$$

where, $\boldsymbol{\Lambda}_l$ = Cov ($\tilde{\mathbf{Z}}_{\mathrm{t}+l}$, $\tilde{\mathbf{Z}}_{\mathrm{t}}$). These lagged covariance matrices can then be written as a function of the underlying model parameters. Specifically, $\boldsymbol{\Lambda}_l$ has the following functional form:

$$\boldsymbol{\Lambda}_l = \mathrm{Cov}\left(\tilde{\mathbf{Z}}_{\mathrm{t}+l}, \tilde{\mathbf{Z}}_{\mathrm{t}}\right) = \mathbf{C}\mathbf{A}^{l}\mathbf{P}_t\mathbf{C}' \tag{S16}$$

where, $\mathbf{P}_t = \mathrm{Cov}(\mathbf{x}_t, \mathbf{x}_t)$ is the zero-lag covariance of the latent state. For an LTI system in steady state, $\mathbf{P}_t$ is typically independent of 't' (stationarity). Given the form of the lagged covariance $\boldsymbol{\Lambda}_l$ in Equation S16 and assuming steady-state, we can factorize $\mathbf{H}$ as shown below:

$$\mathbf{H} \quad = \underbrace{\begin{bmatrix} \mathbf{C} \\ \mathbf{C}\mathbf{A} \\ \mathbf{C}\mathbf{A}^2 \\ \vdots \\ \mathbf{C}\mathbf{A}^{q-1} \end{bmatrix}}_{\boldsymbol{\mathcal{O}}} \underbrace{\begin{bmatrix} \mathbf{A}\mathbf{P}\mathbf{C}' & \mathbf{A}^2\mathbf{P}\mathbf{C}' & \cdots & \mathbf{A}^q\mathbf{P}\mathbf{C}' \end{bmatrix}}_{\boldsymbol{\Psi}} \tag{S17}$$

The matrix $\boldsymbol{\mathcal{O}}$ in Equation S17 is referred to as the *observability* matrix, and is a critical concept in linear systems theory. When $\boldsymbol{\mathcal{O}}$ is full rank, latent states $\mathbf{x}_t$ are uniquely determinable based on observations $\tilde{\mathbf{z}}_{\mathrm{t}}$ alone. By construction, the column space of $\boldsymbol{\mathcal{O}}$ spans the same subspace as the column space of $\mathbf{C}$, which we have referred to alternately as the dynamics subspace (see Methods). Therefore, estimating the column space of $\boldsymbol{\mathcal{O}}$ allows us to obtain an estimate of the dynamics subspace. Based on the relationship between $\boldsymbol{\mathcal{O}}$ and $\mathbf{H}$ in Equation S17, if $\boldsymbol{\mathcal{O}}$ is full rank

(i.e latent states are uniquely determinable), then $\boldsymbol{\mathcal{O}}$ and $\mathbf{H}$ span the same column space. Therefore, $\boldsymbol{\mathcal{O}}$ can be obtained empirically through a singular value decomposition of $\mathbf{H}$ because the left singular vectors of $\mathbf{H}$ specify its column space.

$$\mathbf{H} = \mathbf{U}.\mathbf{S}.\mathbf{V}' \qquad \text{(S18)}$$
$$\widehat{\boldsymbol{\mathcal{O}}} = \mathbf{U}_{(\mathrm{r})}.\mathbf{S}_{(\mathrm{r})}{}^{1/2}$$

The particular definition of $\widehat{\boldsymbol{\mathcal{O}}}$ in Equation S18 i.e the first 'r' left singular vectors ($\mathbf{U}_{(\mathrm{r})}$) weighted by the square root of their respective singular values ($\mathbf{S}_{(\mathrm{r})}$) is often referred to as the "balanced realization". This choice of weighting has no bearing on the directions spanned by the dynamics subspace ($\hat{\mathbf{C}}$).

Obtaining an accurate estimate of $\widehat{\boldsymbol{\mathcal{O}}}$ requires determining the rank of $\mathbf{H}$. Typically, for appropriate choices of q ($q \geq p$), the Cayley-Hamilton theorem requires that $\mathbf{H}$ is low rank and that the rank of $\mathbf{H}$ matches the dimensionality of the latent state (p). Therefore, the parameter r in Equation S18 specifies the estimated rank of $\mathbf{H}$.

Based on the definition of $\boldsymbol{\mathcal{O}}$ in Equation S18, the estimate of the observation matrix/dynamics subspace ($\hat{\mathbf{C}}$) can then be read off as the first block-row of $\widehat{\boldsymbol{\mathcal{O}}}$.

$$\hat{\mathbf{C}} = \widehat{\boldsymbol{\mathcal{O}}}(1{:}\,\mathrm{n}, :) \qquad \text{(S19)}$$

We adapted and extended the above framework described for LTI systems to a model characterized with a linear time-varying (LTV) latent dynamical process, by explicitly considering the dependence of the hankel and observability matrices on time. Therefore, in our data analysis pipeline, we build time-dependent hankel matrices $\mathbf{H}_{\mathrm{t}}$, by considering a window of observations (residuals) centred around time 't'. These time-dependent hankel matrices then specify time-dependent observability matrices $\boldsymbol{\mathcal{O}}_{\mathrm{t}}$ in a manner similar to Equation S18. However, the explicit dependence of observability matrices on time adds a complication in the estimation of a time-invariant observation matrix, as in the model used to estimate residual dynamics (Equation 11, Methods). Therefore, we develop a set of alternate procedures, wherein, we use $\boldsymbol{\mathcal{O}}_{\mathrm{t}}$ to first estimate a 'momentary' dynamics subspace using the same definitions as in Equations S17-S20 ($\hat{\mathbf{C}}_t$; Equations 18 and 19, Methods) that is time-dependent. We then use the sequence of time-dependent momentary dynamics subspaces estimated across all time to construct a single time-invariant dynamics subspace (Equation 20, Methods).

4. Supplementary Math Note B

In this section, we derive the effect of temporally correlated latent noise (Fig 1b, complex input; Methods & Supplementary Methods) on estimates of residual dynamics. We assumed the linear time-invariant state-space model specified by Equation 43 (see Methods). We assume that we only have access to latent residual states $\mathbf{x}(\mathrm{t})$ through observed residuals $\mathbf{z}(\mathrm{t})$, and that the model operates under steady-state conditions. By defining an augmented latent state, $\mathbf{p}(\mathrm{t})$, we can rewrite the model (Equation 43, Methods) as follows:

$$\begin{aligned}\underbrace{\begin{bmatrix}\mathbf{x}(\mathrm{t}+1)\\ \boldsymbol{\varepsilon}(\mathrm{t}+1)\end{bmatrix}}_{\mathbf{p}(\mathrm{t}+1)} &= \underbrace{\begin{bmatrix}\mathbf{A} & \mathbf{I}\\ \mathbf{0} & \boldsymbol{\phi}\end{bmatrix}}_{\tilde{\mathbf{A}}}\underbrace{\begin{bmatrix}\mathbf{x}(\mathrm{t})\\ \boldsymbol{\varepsilon}(\mathrm{t})\end{bmatrix}}_{\mathbf{p}(\mathrm{t})} + \underbrace{\begin{bmatrix}\mathbf{0}\\ \boldsymbol{\zeta}(t)\end{bmatrix}}_{\boldsymbol{\vartheta}(\mathrm{t})}\\ \begin{bmatrix}\mathbf{z}(\mathbf{t})\\ \mathbf{0}\end{bmatrix} &= \begin{bmatrix}\mathbf{C} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{x}(\mathbf{t})\\ \boldsymbol{\varepsilon}(\mathrm{t})\end{bmatrix} + \begin{bmatrix}\boldsymbol{\eta}(\mathrm{t})\\ \mathbf{0}\end{bmatrix}\end{aligned} \tag{S20}$$

where,

$$\boldsymbol{\vartheta}(\mathrm{t}) \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{Q}\end{bmatrix}\right) \tag{S21}$$

We define the *l*-lag covariance between random variables **u** and **v** as follows:

$$\mathbf{C}_{\mathrm{ss}}^{(\mathrm{l})}(\mathbf{u},\mathbf{v}) = \mathrm{Cov}(\mathbf{u}(\mathrm{t}+\mathrm{l}),\mathbf{v}(\mathrm{t})) \tag{S22}$$

The steady-state zero-lag covariance of the augmented latent state $\mathbf{p}(\mathrm{t})$ can be computed using the discrete-time lyapunov equation for a specified $\tilde{\mathbf{A}}$ and $\mathbf{Q}$. It corresponds to a block matrix $\mathbf{C}_{\mathrm{ss}}^{(0)}$ which can be written as follows:

$$\begin{aligned}\mathbf{C}_{\mathrm{ss}}^{(0)}(\mathbf{p},\mathbf{p}) = \mathrm{Cov}\big(\mathbf{p}(\mathrm{t}),\mathbf{p}(\mathrm{t})\big) &= \begin{bmatrix}\mathrm{Cov}\big(\mathbf{x}(\mathrm{t}),\mathbf{x}(\mathrm{t})\big) & \mathrm{Cov}\big(\mathbf{x}(\mathrm{t}),\boldsymbol{\varepsilon}(\mathrm{t})\big)\\ \mathrm{Cov}\big(\boldsymbol{\varepsilon}(\mathrm{t}),\mathbf{x}(\mathrm{t})\big) & \mathrm{Cov}\big(\boldsymbol{\varepsilon}(\mathrm{t}),\boldsymbol{\varepsilon}(\mathrm{t})\big)\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{C}_{\mathrm{ss}}^{(0)}(\mathbf{x},\mathbf{x}) & \mathbf{C}_{\mathrm{ss}}^{(0)}(\mathbf{x},\boldsymbol{\varepsilon})\\ \mathbf{C}_{\mathrm{ss}}^{(0)}(\boldsymbol{\varepsilon},\mathbf{x}) & \mathbf{C}_{\mathrm{ss}}^{(0)}(\boldsymbol{\varepsilon},\boldsymbol{\varepsilon})\end{bmatrix}\end{aligned} \tag{S23}$$

The steady-state *lag-l* covariance of the augmented state $\mathbf{p}(\mathrm{t})$ is therefore defined as:

$$\mathbf{C}_{\mathrm{ss}}^{(\mathrm{l})}(\mathbf{p},\mathbf{p}) = \mathrm{Cov}(\mathbf{p}(\mathrm{t}+\mathrm{l}),\mathbf{p}(\mathrm{t})) = \tilde{\mathbf{A}}^{l}\mathbf{C}_{\mathrm{ss}}^{(0)}(\mathbf{p},\mathbf{p}) \tag{S24}$$

Now, the exponentiated dynamics matrix $\widetilde{\mathbf{A}}^{l}$ in the above equation has the following general form:

$$\widetilde{\mathbf{A}}^{l} = \begin{bmatrix} \mathbf{A}^{l} & \mathbf{U}(l) \\ \mathbf{0} & \boldsymbol{\phi}^{l} \end{bmatrix} \tag{S25}$$

where,

$$\mathbf{U}(l) = \sum_{k=0}^{l-1} \mathbf{A}^{k} \boldsymbol{\phi}^{\,l-1-k} \tag{S26}$$

Using Equations S23, S25 and S26, Equation S24 can therefore be expanded as follows:

$$\begin{aligned} \mathbf{C}_{\mathrm{ss}}^{(l)}(\mathbf{p},\mathbf{p}) = \widetilde{\mathbf{A}}^{\boldsymbol{l}}\mathbf{C}_{\mathrm{ss}}^{(0)}(\mathbf{p},\mathbf{p}) &= \begin{bmatrix} \mathbf{A}^{l} & \mathbf{U}(l) \\ \mathbf{0} & \boldsymbol{\phi}^{l} \end{bmatrix} \begin{bmatrix} \mathbf{C}_{\mathrm{ss}}^{(0)}(\mathbf{x},\mathbf{x}) & \mathbf{C}_{\mathrm{ss}}^{(0)}(\mathbf{x},\boldsymbol{\varepsilon}) \\ \mathbf{C}_{\mathrm{ss}}^{(0)}(\boldsymbol{\varepsilon},\mathbf{x}) & \mathbf{C}_{\mathrm{ss}}^{(0)}(\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}) \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{A}^{l}\mathbf{C}_{\mathrm{ss}}^{(0)}(\mathbf{x},\mathbf{x}) + \mathbf{U}(l)\mathbf{C}_{\mathrm{ss}}^{(0)}(\mathbf{x},\boldsymbol{\varepsilon}) & \mathbf{A}^{l}\mathbf{C}_{\mathrm{ss}}^{(0)}(\mathbf{x},\boldsymbol{\varepsilon}) + \mathbf{U}(l)\mathbf{C}_{\mathrm{ss}}^{(0)}(\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}) \\ \boldsymbol{\phi}^{l}\mathbf{C}_{\mathrm{ss}}^{(0)}(\boldsymbol{\varepsilon},\mathbf{x}) & \boldsymbol{\phi}^{l}\mathbf{C}_{\mathrm{ss}}^{(0)}(\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}) \end{bmatrix} \end{aligned} \tag{S27}$$

Therefore, based on Equation S27, the lag-*l* steady state covariance of the model latent state $\mathbf{x}$ (Equation 43, Methods) can be read off as the upper left block of $\mathbf{C}_{\mathrm{ss}}^{(l)}(\mathbf{p},\mathbf{p})$ as below:

$$\mathbf{C}_{\mathrm{ss}}^{(l)}(\mathbf{x},\mathbf{x}) = \mathbf{A}^{l}\mathbf{C}_{\mathrm{ss}}^{(0)}(\mathbf{x},\mathbf{x}) + \mathbf{U}(l)\mathbf{C}_{\mathrm{ss}}^{(0)}(\mathbf{x},\boldsymbol{\varepsilon}) \tag{S28}$$

The expression in Equation S28 can be used to build a two stage least squares (2SLS) estimator. As outlined in the Methods, the first stage of the 2SLS regression, involves regressing the latent state $\mathbf{x}$(t) against its past in order to build a predictor $\hat{\mathbf{x}}$(t). In the second stage, the predictor $\hat{\mathbf{x}}$(t) is regressed against $\mathbf{x}$(t+1) in order to estimate the dynamics matrix. Assuming, that we use *m* past lags for the first-stage of the 2SLS, we constructed the following three 'stacked' multi-lag covariance matrices using $\mathbf{C}_{\mathrm{ss}}^{(\mathrm{l})}(\mathbf{x},\mathbf{x})$

$$\mathbf{P}_0 = \begin{bmatrix} \mathbf{C}_{ss}^{(0)}(\mathbf{x},\mathbf{x}) & \mathbf{C}_{ss}^{(1)}(\mathbf{x},\mathbf{x}) & \dots & \mathbf{C}_{ss}^{(m-1)}(\mathbf{x},\mathbf{x}) \\ \mathbf{C}_{ss}^{(-1)}(\mathbf{x},\mathbf{x}) & \mathbf{C}_{ss}^{(0)}(\mathbf{x},\mathbf{x}) & \dots & \mathbf{C}_{ss}^{(m-2)}(\mathbf{x},\mathbf{x}) \\ \vdots & & \ddots & \vdots \\ \mathbf{C}_{ss}^{(-m-1)}(\mathbf{x},\mathbf{x}) & \mathbf{C}_{ss}^{(-m-2)}(\mathbf{x},\mathbf{x}) & \dots & \mathbf{C}_{ss}^{(0)}(\mathbf{x},\mathbf{x}) \end{bmatrix} \quad \text{(S29)}$$

$$\mathbf{P}_1 = \begin{bmatrix} \mathbf{C}_{ss}^{(1)}(\mathbf{x},\mathbf{x}) & \mathbf{C}_{ss}^{(2)}(\mathbf{x},\mathbf{x}) & \dots & \mathbf{C}_{ss}^{(m)}(\mathbf{x},\mathbf{x}) \end{bmatrix}$$

$$\mathbf{P}_2 = \begin{bmatrix} \mathbf{C}_{ss}^{(2)}(\mathbf{x},\mathbf{x}) & \mathbf{C}_{ss}^{(3)}(\mathbf{x},\mathbf{x}) & \dots & \mathbf{C}_{ss}^{(m+1)}(\mathbf{x},\mathbf{x}) \end{bmatrix}$$

where, $\mathbf{C}_{ss}^{(-k)}(\mathbf{x},\mathbf{x}) = \mathbf{C}_{ss}^{(k)}(\mathbf{x},\mathbf{x})^{T}$. The two-stage least squares estimator based on *m* past lags is then given by:

$$\widehat{\mathbf{A}}_{2sls}^{(m)} = \left(\mathbf{P}_2 . (\mathbf{P}_0{}^{-1})^{T} \mathbf{P}_1{}^{T}\right)\left(\mathbf{P}_1 . (\mathbf{P}_0{}^{-1})^{T} \mathbf{P}_1{}^{T}\right)^{-1} \quad \text{(S30)}$$

As in the data, we use a lag of m = 3, in order to construct the 2SLS analytical estimator of the residual dynamics in Extended Data Fig. 4. The pseudocode for constructing the estimator is given below:

Algorithm 1: Compute analytical 2SLS estimator for linear-SSM with correlated input noise

INPUT: $\mathbf{A}, \boldsymbol{\phi}, \mathbf{Q}, m$

OUTPUT: $\widehat{\mathbf{A}}_{2sls}^{(m)}$

1. Construct $\widetilde{\mathbf{A}} = \begin{bmatrix} \mathbf{A} & \mathbf{I} \\ \mathbf{0} & \boldsymbol{\phi} \end{bmatrix}$ and $\widetilde{\mathbf{Q}} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} \end{bmatrix}$
2. Solve for $\mathbf{C}_{ss}^{(0)}(\mathbf{p},\mathbf{p})$: $\widetilde{\mathbf{A}}\,\mathbf{C}_{ss}^{(0)}\widetilde{\mathbf{A}}^{T} - \mathbf{C}_{ss}^{(0)} + \widetilde{\mathbf{Q}} = \mathbf{0}$ (discrete-time lyapunov equation)
3. Use individual block entries of $\mathbf{C}_{ss}^{(0)}(\mathbf{p},\mathbf{p})$, $\mathbf{A}$ and $\boldsymbol{\phi}$ to define $\mathbf{C}_{ss}^{(l)}(\mathbf{x},\mathbf{x})$ for l = 0,1,2 … m+1 (Equation S28)
4. Define $\mathbf{P}_0$, $\mathbf{P}_1$ and $\mathbf{P}_2$ using $\mathbf{C}_{ss}^{(l)}(\mathbf{x},\mathbf{x})$ as in Equation S29
5. Compute and return $\widehat{\mathbf{A}}_{2sls}^{(m)}$ using output of step 4 in Equation S30

Supplementary References

1. Roweis, S. & Ghahramani, Z. A Unifying Review of Linear Gaussian Models. *Neural Comput.* **11**, 305–345 (1999).
2. Ghahramani, Z. & Hinton, G. E. Variational learning for switching state-space models. *Neural Comput.* **12**, 831–864 (2000).
3. Fox, E. B., Sudderth, E. B., Jordan, M. I. & Willsky, A. S. Nonparametric Bayesian Learning of Switching Linear Dynamical Systems. in *Advances in Neural Information Processing Systems* vol. 21 457–464 (2009).
4. Ljung, L. *System identification: theory for the user*. (Prentice Hall PTR, 1999).
5. Katayama, T. *Subspace Methods for System Identification*. (Springer London, 2005). doi:10.1007/1-84628-158-X.