Supplementary material

S.1 Parametrization of noise covariance

We denote the noise of voxel i at time point t by $\varepsilon_{i,t}$, and denote the temporal covariance of the noise time series ε_i as Σ_{ϵ_i} .

Following AR(1) model,

$$\epsilon_{i,t} = \rho_i \epsilon_{i,t-1} + \eta_{i,t} \tag{s-1},$$

where $\eta_{i,t} \sim N(0, \sigma_i^2)$ is the "new" noise.

Then we have $\epsilon_t - \rho_1 \epsilon_{t-1} = \eta_t$.

Denote $B = I - \rho_i P_1$, where $P_1 = \begin{vmatrix} 0 & 0 & 0 & & \\ 1 & 0 & 0 & \cdots & \\ 0 & 1 & 0 & & \\ & & & 0 & 0 \\ & \cdots & & 1 & 0 \end{vmatrix}$. Multiplying matrix B to the noise ϵ_i

removes the temporal correlation: $B \epsilon_i \sim N(0, \sigma_i^2 \hat{I})$.

$\hat{I}$ is the same as an identity matrix except that its (1,1) element is $\dfrac{1}{1-\rho_i^2}$, because the variance of

the AR(1) noise at any time point without knowing the noise at previous time points is $\dfrac{\sigma_i^2}{1-\rho_i^2}$.

We also know that $B \epsilon_i \sim N(0, B \Sigma_{\epsilon_i} B^T)$

Therefore, $B \Sigma_{\epsilon_i} B^T = \sigma_i^2 \hat{I}$, $\Sigma_{\epsilon_i} = \sigma_i^2 B^{-1} \hat{I} (B^T)^{-1}$

and

$$\Sigma_{\epsilon_i}^{-1} = \frac{1}{\sigma_i^2} B^T \hat{I}^{-1} B \tag{s-2}$$

$\hat{I}^{-1}$ has $1-\rho_i^2$ on element (1,1). In other words, $\hat{I}^{-1} = I - \rho_i^2 \delta_1 \delta_1^T$. Here we use δ_k to denote a unit vector which is 1 at the k-th element and 0 in other elements.

We expand (s-2):

$$\Sigma_{\epsilon_i}^{-1} = \sigma_i^{-2} B^T \hat{I}^{-1} B$$
$$= \sigma_i^{-2} (I - \rho_i P_1)^T (I - \rho_i^2 \delta_1 \delta_1^T)(I - \rho_i P_1)$$
$$= \sigma_i^{-2} [I - \rho_i P_1^T - \rho_i^2 \delta_1 \delta_1^T + \rho_i^3 P_1^T \delta_1 \delta_1^T - \rho_1 P_1 + \rho_1^2 P_1^T P_1 + \rho_i^3 \delta_1 \delta_1^T P_1 - \rho_i^4 P_1^T \delta_1 \delta_1^T P_1]$$
$$= \sigma_i^{-2} [I - \rho_i (P_1^T + P_1) + \rho_i^2 (I - \delta_1 \delta_1^T - \delta_n \delta_n^T)]$$

The last equality is because $P_1^T \delta_1$ is a zero vector, and that $P_1^T P_1 = I - \delta_n \delta_n^T$

We can write $D = P_1^T + P_1$ and $F = I - \delta_1 \delta_1^T - \delta_n \delta_n^T$

Then D and F are two template matrices which do not depend on any free parameters.

So we can write the inverse covariance matrix as

$$\Sigma_{\epsilon_i}^{-1} = \sigma_i^{-2} [I - \rho_i D + \rho_i^2 F] \tag{s-3}$$

We can further denote $A_i = A(\rho_i) = I - \rho_i D + \rho_i^2 F$, so $\Sigma_{\epsilon_i}^{-1} = \sigma_i^{-2} A_i$ $\qquad$ (s-4)

1

Because $|B|=1$ and $|\hat{I}^{-1}|=1-\rho_1^2$, we have $|\Sigma_{\epsilon_i}^{-1}|=\sigma_i^{-2n_T}|B||\hat{I}^{-1}||B|=\sigma_i^{-2n_T}(1-\rho_i^2)$

S.2 Derivation of log likelihood

Following the notation in the article, for voxel i, we reparametrize $\beta_i=(s_i\sigma_i)L\alpha_i$, $\alpha_i\sim N(0,I_r)$, where r is the rank of $U=LL^T$, the shared covariance matrix. The likelihood of observing data Y_i is:

$$p(Y_i|L,\alpha_i,s_i\sigma_i,\rho_i)p(\alpha_i)$$

$$=N(Y_i;s_i\sigma_iXL\alpha_i,\rho_i)N(\alpha_i;0,I_r)$$

$$=(2\pi)^{-\frac{n_T}{2}}|\Sigma_{\epsilon_i}|^{-\frac{1}{2}}\exp[-\frac{1}{2}(Y_i-s_i\sigma_iXL\alpha_i)^T\Sigma_{\epsilon_i}^{-1}(Y_i-s_i\sigma_iXL\alpha_i)](2\pi)^{-\frac{r}{2}}\exp[-\frac{1}{2}\alpha_i^T\alpha_i]$$

$$=(2\pi)^{-\frac{(n_T+r)}{2}}|\Sigma_{\epsilon_i}|^{-\frac{1}{2}}\exp[-\frac{1}{2}(s_i^2\sigma_i^2\alpha_i^TL^TX^T\Sigma_{\epsilon_i}^{-1}XL\alpha_i-2s_i\sigma_iY_i^T\Sigma_{\epsilon_i}^{-1}XL\alpha_i+Y_i^T\Sigma_{\epsilon_i}^{-1}Y_i+\alpha_i^T\alpha_i)]$$

$$=(2\pi)^{-\frac{(n_T+r)}{2}}|\Sigma_{\epsilon_i}|^{-\frac{1}{2}}\exp[-\frac{1}{2}(\alpha_i^T(s_i^2\sigma_i^2L^TX^T\Sigma_{\epsilon_i}^{-1}XL+I_r)\alpha_i-2s_i\sigma_iY_i^T\Sigma_{\epsilon_i}^{-1}XL\alpha_i+Y_i^T\Sigma_{\epsilon_i}^{-1}Y_i)]$$

$$=(2\pi)^{-\frac{(n_T+r)}{2}}|\Sigma_{\epsilon_i}|^{-\frac{1}{2}}\exp\{-\frac{1}{2}[\alpha_i^T\Lambda_i^{-1}\alpha_i-2(s_i\sigma_i\Lambda_iL^TX^T\Sigma_{\epsilon_i}^{-1}Y_i)^T\Lambda_i^{-1}\alpha_i+Y_i^T\Sigma_{\epsilon_i}^{-1}Y_i]\}$$

$$=(2\pi)^{-\frac{(n_T+r)}{2}}|\Sigma_{\epsilon_i}|^{-\frac{1}{2}}\exp[-\frac{1}{2}(\alpha_i-\mu_i)^T\Lambda_i^{-1}(\alpha_i-\mu_i)+\frac{1}{2}(s_i^2\sigma_i^2Y_i^T\Sigma_{\epsilon_i}^{-1}XL\Lambda_iL^TX^T\Sigma_{\epsilon_i}^{-1}Y_i-Y_i^T\Sigma_{\epsilon_i}^{-1}Y_i)]$$

$$=(2\pi)^{-\frac{n_T}{2}}|\Sigma_{\epsilon_i}|^{-\frac{1}{2}}|\Lambda_i|^{\frac{1}{2}}\exp[\frac{1}{2}(s_i^2\sigma_i^2Y_i^T\Sigma_{\epsilon_i}^{-1}XL\Lambda_iL^TX^T\Sigma_{\epsilon_i}^{-1}Y_i-Y_i^T\Sigma_{\epsilon_i}^{-1}Y_i)]N(\alpha_i;\mu_i,\Lambda_i)$$

$$=(2\pi)^{-\frac{n_T}{2}}|\Sigma_{\epsilon_i}|^{-\frac{1}{2}}|\Lambda_i|^{\frac{1}{2}}\exp[\frac{1}{2}(\mu_i^T\Lambda_i^{-1}\mu_i-Y_i^T\Sigma_{\epsilon_i}^{-1}Y_i)]N(\alpha_i;\mu_i,\Lambda_i)$$

(s-5)

In the derivation above,
$\mu_i=s_i\sigma_i\Lambda_iL^TX^T\Sigma_{\epsilon_i}^{-1}Y_i$ and $\Lambda_i=(s_i^2\sigma_i^2L^TX^T\Sigma_{\epsilon_i}^{-1}XL+I)^{-1}$ are the posterior mean and variance of α_i .

By plugging in (s-4), we get
$\Lambda_i=(s_i^2L^TX^TA_iXL+I)^{-1}$

$\mu_i=\frac{s_i}{\sigma_i}\Lambda_iL^TX^TA_iY_i$

(s-6)

Marginalizing α_i :
$p(Y_i|L,s_i\sigma_i,\rho_i)$

$=\int d\alpha_i p(Y_i|L,\alpha_i,s_i\sigma_i,\rho_i)p(\alpha_i)$

$=\int d\alpha_i N(\alpha_i;\mu_i,\Lambda_i)(2\pi)^{-\frac{n_T}{2}}|\Sigma_{\epsilon_i}|^{-\frac{1}{2}}|\Lambda_i|^{\frac{1}{2}}\exp[\frac{1}{2}(\mu_i^T\Lambda_i^{-1}\mu_i-Y_i^T\Sigma_{\epsilon_i}^{-1}Y_i)]$

$=(2\pi)^{-\frac{n_T}{2}}|\Sigma_{\epsilon_i}|^{-\frac{1}{2}}|\Lambda_i|^{\frac{1}{2}}\exp[\frac{1}{2}(\mu_i^T\Lambda_i^{-1}\mu_i-Y_i^T\Sigma_{\epsilon_i}^{-1}Y_i)]$

Therefore, the marginal log likelihood is

2

$$\log \left(p \left(Y_i | L, s_i \sigma_i, \rho_i \right) \right)$$

$$= \log \left\{ \left(2\pi \right)^{-\frac{n_T}{2}} |\Sigma_{\epsilon_i}|^{-\frac{1}{2}} |\Lambda_i|^{\frac{1}{2}} \exp \left[\frac{1}{2} \left(\mu_i^T \Lambda_i^{-1} \mu_i - Y_i^T \Sigma_{\epsilon_i}^{-1} Y_i \right) \right] \right\}$$

$$= -\frac{n_T}{2} \log \left(2\pi \right) - \frac{n_T}{2} \log \sigma_i^2 + \frac{1}{2} \log \left(1 - \rho_i^2 \right) - \frac{1}{2} \log \left| s_i^2 L^T X^T A_i X L + I \right| \qquad \textbf{(s-7)}$$

$$+ \frac{1}{2\sigma_i^2} \left[s_i^2 Y_i^T A_i X L \Lambda_i L^T X^T A_i Y_i - Y_i^T A_i Y_i \right]$$

If there are multiple runs, the temporal covariance matrix becomes block diagonal matrix, with each block on the diagonal corresponding to the covariance matrix of one run. Its determinant is the product of the determinant of each block on the diagonal. Therefore, the term $\frac{1}{2} \log \left(1 - \rho_i^2 \right)$ becomes $\frac{n_{run}}{2} \log \left(1 - \rho_i^2 \right)$, where n_{run} is the number of fMRI runs.

$$\log \left(p \left(Y_i | L, s_i \sigma_i, \rho_i \right) \right)$$

$$= -\frac{n_T}{2} \log \left(2\pi \right) - \frac{n_T}{2} \log \sigma_i^2 + \frac{n_{run}}{2} \log \left(1 - \rho_i^2 \right) - \frac{1}{2} \log \left| \Lambda_i^{-1} \right| \qquad \textbf{(s-8)}$$

$$+ \frac{1}{2\sigma_i^2} \left[s_i^2 Y_i^T A_i X L \Lambda_i L^T X^T A_i Y_i - Y_i^T A_i Y_i \right]$$

In our implementation, we parametrize s_i with $\log(s_i^2)$, σ_i with $\log(\sigma_i^2)$, and ρ_i with $a_i = \tan \left(\frac{\pi}{2} \rho_i \right)$. In this way, all free parameters become unbounded and we can use unrestricted gradient based algorithm to maximize the marginal log likelihood.

S.3 Gradient of the log likelihood

S.3.1 Gradient with respect to σ_i^2

$$\frac{\partial \log \left(p \left(Y_i | L, s_i \sigma_i, \rho_i \right) \right)}{\sigma_i^2}$$

$$= -\frac{n_T}{2\sigma_i^2} - \frac{1}{2\sigma_i^4} \left[s_i^2 Y_i^T A_i X L \Lambda_i L^T X^T A_i Y_i - Y_i^T A_i Y_i \right]$$

The maximum likelihood estimate of σ_i^2 given other parameters is achieved when the gradient above equals zero. Therefore,

$$\sigma_{i,ML}^2 = \frac{Y_i^T A_i Y_i - s_i^2 Y_i^T A_i X L \Lambda_i L^T X^T A_i Y_i}{n_T} \qquad \textbf{(s-9)}$$

So we do not need to pursue gradient descent on σ_i^2. This reduces the number of free parameters by close to one third.

3

S.3.2 Gradient with respect to log(s_i^2)

$$\frac{\partial \log\left(p\left(Y_i|L, s_i\,\sigma_i, \rho_i\right)\right)}{\partial s_i^2}$$

$$=-\frac{1}{2}Tr\left[\left(s_i^2 L^T X^T A_i X L+I\right)^{-1}\frac{\partial\left(s_i^2 L^T X^T A_i X L+I\right)}{\partial s_i^2}\right]$$

$$+\frac{1}{2}\sigma_i^{-2}Y_i^T A_i X L\left[s_i^2 L^T X^T A_i X L+I\right]^{-1}L^T X^T A_i Y_i$$

$$-\frac{1}{2}\sigma_i^{-2}s_i^2 Y_i^T A_i X L\left[s_i^2 L^T X^T A_i X L+I\right]^{-1}$$

$$\cdot\frac{\partial\left(s_i^2 L^T X^T A_i X L+I\right)}{\partial s_i^2}\cdot\left[s_i^2 L^T X^T A_i X L+I\right]^{-1}L^T X^T A_i Y_i$$

$$=-\frac{1}{2}Tr\left[\left(s_i^2 L^T X^T A_i X L+I\right)^{-1}L^T X^T A_i X L\right]$$

$$+\frac{1}{2}\sigma_i^{-2}Y_i^T A_i X L \Lambda_i L^T X^T A_i Y_i$$

$$-\frac{1}{2}\sigma_i^{-2}s_i^2 Y_i^T A_i X L \Lambda_i L^T X^T A_i X L \Lambda_i L^T X^T A_i Y_i$$

$$=-\frac{1}{2}Tr\left[\left(s_i^2 L^T X^T A_i X L+I\right)^{-1}\left(s_i^2 L^T X^T A_i X L+I-I\right)s_i^{-2}\right]$$

$$+\frac{1}{2}\sigma_i^{-2}Y_i^T A_i X L \Lambda_i L^T X^T A_i Y_i$$

$$-\frac{1}{2}\sigma_i^{-2}s_i^2 Y_i^T A_i X L \Lambda_i L^T X^T A_i X L \Lambda_i L^T X^T A_i Y_i$$

$$=-\frac{1}{2s_i^2}Tr\left[I-\left(s_i^2 L^T X^T A_i X L+I\right)^{-1}\right]$$

$$+\frac{1}{2}\sigma_i^{-2}Y_i^T A_i X L \Lambda_i L^T X^T A_i Y_i$$

$$-\frac{1}{2}\sigma_i^{-2}s_i^2 Y_i^T A_i X L \Lambda_i L^T X^T A_i X L \Lambda_i L^T X^T A_i Y_i$$

$$=\frac{1}{2s_i^2}\left(-r+Tr\left[\Lambda_i\right]\right)+\frac{1}{2\sigma_i^2}Y_i^T A_i X L \Lambda_i L^T X^T A_i Y_i$$

$$-\frac{s_i^2}{2\sigma_i^2}Y_i^T A_i X L \Lambda_i L^T X^T A_i X L \Lambda_i L^T X^T A_i Y_i$$

Therefore,

4

$$\frac{\partial \log\left(p\left(Y_i|L,s_i\,\sigma_i,\rho_i\right)\right)}{\partial \log s_i^2}$$

$$=\frac{\partial \log\left(p\left(Y_i|L,\sigma_i,\rho_i,s_i\right)\right)}{\partial s_i^2}\frac{\partial s_i^2}{\partial \log s_i^2}$$

$$=\frac{1}{2}\left(-r+Tr\left[\Lambda_i\right]\right)+\frac{s_i^2}{2\,\sigma_i^2}Y_i^T A_i X L \Lambda_i L^T X^T A_i Y_i$$

$$-\frac{s_i^4}{2\,\sigma_i^2}Y_i^T A_i X L \Lambda_i L^T X^T A_i X L \Lambda_i L^T X^T A_i Y_i$$

(s-10)

S.3.3 Gradient with respect to $a_i=\tan\left(\frac{\pi}{2}\rho_i\right)$

$$\frac{\partial a_i}{\partial \rho_i}=\frac{2}{\pi\left(1+a_i^2\right)}$$

The derivative of A_i with respect to ρ_i is:

$$\frac{\partial A_i}{\partial \rho_i}=\frac{\partial\left(I-\rho_i D+\rho_i^2 F\right)}{\partial \rho_i}=-D+2\rho_i F$$

(s-11)

The gradient of the marginal log likelihood with respect to ρ_i is :

$$\frac{\partial \log\left(p\left(Y_i|L,s_i\,\sigma_i,\rho_i\right)\right)}{\partial \rho_i}$$

$$=-\frac{n_{run}\rho_{1i}}{1-\rho_{1i}^2}-\frac{1}{2}Tr\left[\left(s_i^2 L^T X^T A_i X L+I\right)^{-1}\frac{\partial\left(s_i^2 L^T X^T A_i X L+I\right)}{\partial \rho_i}\right]$$

$$+\frac{1}{2\,\sigma_i^2}\left[s_i^2 Y_i^T \frac{\partial A_i}{\partial \rho_i}X L \Lambda_i L^T X^T A_i Y_i+s_i^2 Y_i^T A_i X L \frac{\partial \Lambda_i}{\partial \rho_i}L^T X^T A_i Y_i+s_i^2 Y_i^T A_i X L \Lambda_i L^T X^T \frac{\partial A_i}{\partial \rho_i}Y_i\right]$$

$$-\frac{1}{2\,\sigma_i^2}Y_i^T \frac{\partial A_i}{\partial \rho_i}Y_i$$

$$=-\frac{n_{run}\rho_{1i}}{1-\rho_{1i}^2}-\frac{s_i^2}{2}Tr\left[\Lambda_i L^T X^T \frac{\partial A_i}{\partial \rho_i}X L\right]$$

$$+\frac{s_i^2}{\sigma_i^2}Y_i^T \frac{\partial A_i}{\partial \rho_i}X L \Lambda_i L^T X^T A_i Y_i$$

$$-\frac{s_i^4}{2\,\sigma_i^2}Y_i^T A_i X L \Lambda_i L^T X^T \frac{\partial A_i}{\partial \rho_i}X L \Lambda_i L^T X^T A_i Y_i-\frac{1}{2\,\sigma_i^2}Y_i^T \frac{\partial A_i}{\partial \rho_i}Y_i$$

Therefore,

5

$$\frac{\partial \log\left(p\left(Y_i|L,s_i\,\sigma_i,\rho_i\right)\right)}{\partial a_i}$$

$$=\frac{\partial \log\left(p\left(Y_i|L,s_i\,\sigma_i,\rho_i\right)\right)}{\partial \rho_i}\frac{\partial \rho_i}{\partial a_i}$$

$$=\frac{2}{\pi\left(1+a_i^2\right)}\{-\frac{n_{run}\rho_{1i}}{1-\rho_{1i}^{~2}}-\frac{s_i^2}{2}Tr[\Lambda_i L^T X^T \frac{\partial A_i}{\partial \rho_i}X L] \qquad\qquad \textbf{(s-12)}$$

$$+\frac{s_i^2}{\sigma_i^2}Y_i^T \frac{\partial A_i}{\partial \rho_i}X L \Lambda_i L^T X^T A_i Y_i$$

$$-\frac{s_i^4}{2\sigma_i^2}Y_i^T A_i X L \Lambda_i L^T X^T \frac{\partial A_i}{\partial \rho_i}X L \Lambda_i L^T X^T A_i Y_i-\frac{1}{2\sigma_i^2}Y_i^T \frac{\partial A_i}{\partial \rho_i}Y_i\}$$

S.3.4 Gradient with respect to *L*

We start with the gradient with respect to the (j, k) element L_{jk} of L (j ≤ k).

Note that $\quad\dfrac{\partial L}{L_{jk}}=\delta_j\,\delta_k^T$

$$\frac{\partial \log\left(p\left(Y|L,s_i\,\sigma_i,\rho_i\right)\right)}{\partial L_{jk}}$$

$$=-\frac{1}{2}Tr[\Lambda_i\frac{\partial \Lambda_i^{-1}}{\partial L_{jk}}]+\frac{s_i^2}{2\sigma_i^2}\frac{\partial\left(Y_i^T A_i X L \Lambda_i L^T X^T A_i Y_i\right)}{\partial L_{jk}}$$

$$=-\frac{1}{2}Tr[\Lambda_i\frac{\partial\left(s_i^2 L^T X^T A_i X L+I\right)}{\partial L_{jk}}]+\frac{s_i^2}{2\sigma_i^2}\frac{\partial\left(Y_i^T A_i X L\left(s_i^2 L^T X^T A_i X L+I\right)^{-1}L^T X^T A_i Y_i\right)}{\partial L_{jk}}$$

$$=-\frac{s_i^2}{2}Tr[\Lambda_i\left(\delta_k\delta_j^T X^T A_i X L+L^T X^T A_i X \delta_j\delta_k^T\right)]$$

$$+\frac{s_i^2}{2\sigma_i^2}Y_i^T A_i X \delta_j\delta_k^T\left(s_i^2 L^T X^T A_i X L+I\right)^{-1}X^T A_i Y$$

$$-\frac{s_i^2}{2\sigma_i^2}Y_i^T A_i X L\left(s_i^2 L^T X^T A_i X L+I\right)^{-1}\frac{\partial\left(s_i^2 L^T X^T A_i X L+I\right)}{\partial L_{jk}}\left(s_i^2 L^T X^T A_i X L+I\right)^{-1}L^T X^T A_i Y_i$$

$$+\frac{s_i^2}{2\sigma_i^2}Y_i^T A_i X L\left(s_i^2 L^T X^T A_i X L+I\right)^{-1}\delta_k\delta_j^T X^T A_i Y$$

$$=-\frac{s_i^2}{2}Tr[\delta_j^T X^T A_i X L \Lambda_i \delta_k]-\frac{s_i^2}{2}Tr[\delta_k^T \Lambda_i L^T X^T A_i X \delta_j]$$

$$+\frac{s_i^2}{2\sigma_i^2}Tr[Y_i^T A_i X \delta_j\delta_k^T \Lambda_i X^T A_i Y]$$

$$-\frac{s_i^4}{2\sigma_i^2}Tr[Y_i^T A_i X L \Lambda_i\left(\delta_k\delta_j^T X^T A_i X L+L^T X^T A_i X \delta_j\delta_k^T\right)\Lambda_i L^T X^T A_i Y_i]$$

$$+\frac{s_i^2}{2\sigma_i^2}Tr[Y_i^T A_i X L \Lambda_i \delta_k\delta_j^T X^T A_i Y]$$

6

$$= -\frac{s_i^2}{2}\{X^T A_i X L \Lambda_i\}_{jk} - \frac{s_i^2}{2}\{\Lambda_i L^T X^T A_i X\}_{kj}$$

$$+\frac{s_i^2}{2\sigma_i^2}Tr[\delta_k^T \Lambda_i X^T A_i Y Y_i^T A_i X \delta_j] + \frac{s_i^2}{2\sigma_i^2}Tr[\delta_j^T X^T A_i Y Y_i^T A_i X L \Lambda_i \delta_k]$$

$$-\frac{s_i^4}{2\sigma_i^2}Tr[\delta_j^T X^T A_i X L \Lambda_i L^T X^T A_i Y_i Y_i^T A_i X L \Lambda_i \delta_k] - \frac{s_i^4}{2}Tr[\delta_k^T \Lambda_i L^T X^T A_i Y_i Y_i^T A_i X L \Lambda_i L^T X^T A_i X \delta_j]$$

$$= -\frac{s_i^2}{2}\{X^T A_i X L \Lambda_i\}_{jk} - \frac{s_i^2}{2}\{\Lambda_i L^T X^T A_i X\}_{kj}$$

$$+\frac{s_i^2}{2\sigma_i^2}\{\Lambda_i X^T A_i Y Y_i^T A_i X\}_{kj} + \frac{s_i^2}{2\sigma_i^2}\{X^T A_i Y Y_i^T A_i X L \Lambda_i\}_{jk}$$

$$-\frac{s_i^4}{2\sigma_i^2}\{X^T A_i X L \Lambda_i L^T X^T A_i Y_i Y_i^T A_i X L \Lambda_i\}_{jk} - \frac{s_i^4}{2}\{\Lambda_i L^T X^T A_i Y_i Y_i^T A_i X L \Lambda_i L^T X^T A_i X\}_{kj}$$

$$= -s_i^2\{X^T A_i X L \Lambda_i\}_{jk} + \frac{s_i^2}{\sigma_i^2}\{X^T A_i Y Y_i^T A_i X L \Lambda_i\}_{jk}$$

$$-\frac{s_i^4}{2\sigma_i^2}\{X^T A_i X L \Lambda_i L^T X^T A_i Y_i Y_i^T A_i X L \Lambda_i\}_{jk}$$

(s-13)

In the above derivation, we used the properties of $Tr[AB]=Tr[BA]$ and that a number is the trace of itself considered as a one-element matrix. $\{M\}_{jk}$ means the (j,k) element of matrix M.

Notice that the derivative with respect to the (j,k) element L is equal to the (j,k) element of the matrix in the last equality. Therefore, we can see that the derivative with respect to L in the matrix form is:

$$\frac{\partial \log(p(Y|L,s_i \sigma_i, \rho_i))}{\partial L}$$

$$= -s_i^2 X^T A_i X L \Lambda_i + \frac{s_i^2}{\sigma_i^2} X^T A_i Y Y_i^T A_i X L \Lambda_i$$

(s-14),

$$-\frac{s_i^4}{2\sigma_i^2} X^T A_i X L \Lambda_i L^T X^T A_i Y_i Y_i^T A_i X L \Lambda_i$$

for the lower-triangular part.

S.4 Gaussian Process prior on log(s)

The kernel of the Gaussain Process (GP) we use is the product of a squared exponential kernel defined on the spatial coordinates and one defined on the mean intensity of each voxel.

We denote W_{space} as the matrix of squared spatial distance between voxels, and W_{inten} as the matrix of squared intensity difference between voxels.

Then, in our GP prior, the random vector $\log(s) \in \mathbb{R}^{n_s}$ follows multivariate Gaussian distribution of

$$\log(s) \sim N(0, \tau^2 K)$$ **(s-15)**,

where

$$K = \exp\left[-\frac{1}{2}\left(\frac{W_{space}}{l_{space}^2} + \frac{W_{inten}}{l_{inten}^2}\right)\right] + \eta I$$ **(s-16)**

η is a small number added to the diagonal of the covariance kernel, in order to guarantee matrix K is

7

invertible. In our implementation, we set it to 0.0001

l_{space} and l_{inten} are length scales of the squared exponential kernel for spatial distance and intensity difference, respectively. τ^2 is the variance of the GP.

This prior on log(s) adds additional term to the posterior of the parameters.

$$p(L,\sigma,\rho,s,\tau,l_{space},l_{inten}|Y) \propto p(Y|L,\sigma,\rho,s)\,P(s|\tau,l_{space},l_{inten})\,P(\tau)\,P(l_{space})\,P(l_{inten})$$

We assume uniform prior for other parameters: L, σ and ρ, because empirically σ and ρ can be recovered well, and we do not have a good principle to regularize L.

To fit the model, we optimize the joint probability of both the parameters L, σ, ρ and s, and the hyper-parameters τ, l_{space} and l_{inten}.

We parametrize l_{space} and l_{inten} with $\log(l_{space}^2)$ and $\log(l_{inten}^2)$ to keep all parameters unconstrained. In practice, the optimization of parameters can be instable without regularizing τ, l_{space} and l_{inten}. Therefore, we introduced weakly informative half-Cauchy prior on these hyper-parameters:

$$p(\tau)=\frac{2\,\gamma_\tau}{\pi}\frac{1}{\tau^2+\gamma_\tau^2} \quad , \quad p(l_{space})=\frac{2\,\gamma_{l_{space}}}{\pi}\frac{1}{l_{space}^2+\gamma_{l_{space}}^2} \quad , \quad p(l_{inten})=\frac{2\,\gamma_{l_{inten}}}{\pi}\frac{1}{l_{inten}^2+\gamma_{l_{inten}}^2}$$

The scale parameters γ are set with reasonably large values which the user believe covers the plausible range of τ, l_{space} and l_{inten}. Since the GP is defined for $\log(s)$, $\gamma_\tau = 5$ is a reasonable scale parameter ($e^5 \sim 150$). We set $\gamma_{l_{spacec}}$ and $\gamma_{l_{inten}}$ as half of the maximal distance in space or intensity between all voxels in the ROI

Therefore, the following term is added to the log of joint probability of all parameters (neglecting constant terms):

$$-\frac{n_S}{2}\log(2\pi\tau^2)-\frac{1}{2}\log|K|-\frac{1}{2\tau^2}\log(s)^T K^{-1}\log(s)-\log(\tau^2+\gamma_\tau^2)-\log(l_{space}^2+\gamma_{l_{space}}^2)-\log(l_{inten}^2+\gamma_{l_{inten}}^2)$$

(s-17).

$-\frac{1}{2\tau^2}K^{-1}\log(s)$ is added to the gradient with respect to $\log(s^2)$.

$$\frac{\partial\,p(L,s_i\sigma_i,\rho_i,\tau,l_{space},l_{inten}|Y)}{\tau^2}$$

$$=-\frac{n_s}{2\tau^2}+\frac{\log(s)^T K^{-1}\log(s)}{2\tau^4}-\frac{1}{\tau^2+\gamma_\tau^2}$$

The positive solution of the above term being equal to zero provides the maximum a posterior estimate of τ^2 given other parameters:

$$\tau_{ML}^2=\frac{\log(s)^T K^{-1}\log(s)-n_s\gamma_\tau^2+\sqrt{n_S^2\gamma_\tau^4-(2n+8)\gamma_\tau^2\log(s)^T K^{-1}\log(s)+[\log(s)^T K^{-1}\log(s)]^2}}{2(n_S+2)}$$

(s-18)

The gradient with respect to $\log(l_{space}^2)$ is:

$$\frac{\partial\,p(L,s,\sigma,\rho,\tau,l_{space},l_{inten}|Y)}{\log(l_{space}^2)}$$

(s-19)

$$=l_{space}^2\{\frac{1}{2\tau^2}\log(s)^T K^{-1}\frac{\partial K}{\partial l_{space}^2}K^{-1}\log(s)-\frac{1}{2}Tr[K^{-1}\frac{\partial K}{\partial l_{space}^2}]-\frac{1}{l_{space}^2+\gamma_{l_{space}}}\}$$

where the (j,k) element of $\frac{\partial K}{\partial l_{space}^2}$ is

8

$$\{\frac{\partial K}{\partial l_{space}^2}\}_{jk} = \frac{1}{2l_{space}^4}\{W_{space}\}_{jk} \cdot \{\exp[-\frac{1}{2}(\frac{W_{space}}{l_{space}^2}+\frac{W_{inten}}{l_{inten}^2})]\}_{jk} \tag{s-20}$$

And

$$\frac{\partial\, p(L,s,\sigma,\rho,\tau,l_{space},l_{inten}|Y)}{\log(l_{inten}^2)}$$

$$=l_{inten}^2\{\frac{1}{2\,\tau^2}\log(s)^T K^{-1}\frac{\partial K}{\partial l_{inten}^2}K^{-1}\log(s)-\frac{1}{2}Tr[K^{-1}\frac{\partial K}{\partial l_{inten}^2}]-\frac{1}{l_{inten}^2+\gamma_{l_{inten}}}\} \tag{s-21}$$

$$\{\frac{\partial K}{\partial l_{inten}^2}\}_{jk} = \frac{1}{2l_{inten}^4}\{W_{inten}\}_{jk} \cdot \{\exp[-\frac{1}{2}(\frac{W_{space}}{l_{space}^2}+\frac{W_{inten}}{l_{inten}^2})]\}_{jk} \tag{s-22}$$

S.5 Bias introduced by applying standard RSA on raw fMRI data

Instead of building design matrix that reflects hypothetical hemodynamic response, researchers sometimes take the raw fMRI signal (sometimes after pre-processing such as despiking and detrending) approximately 6 seconds after an event as the neural response "pattern" to that event. As mentioned in the main text at the end of **Section 2**, such approach also suffers from bias. The expected covariance matrix between such patterns is

$$(X^T X)^{-1}X^T X_{true}\,U\,X_{true}^T X(X^T X)^{-1}+(X^T X)^{-1}X^T \Sigma_{\epsilon} X(X^T X)^{-1} \tag{s-23}.$$

Where X_{true} is the design matrix reflecting the true HRF in fMRI, and X is a design matrix which has a single-pulse 6 sec after each event. Such a design matrix with single-pulses is an implicit assumption when treating the raw fMRI data as neural pattern. It is easy to observe that averaging patterns of the same condition following this approach is equivalent as calculating $\hat{\beta}$ with equation (2) but assuming single-pulse HRF.

The first term in (s-23) is introduced by implicitly assuming an unrealistic HRF. Because of this assumption, the "patterns" obtained are in fact mixtures of the true neural patterns of adjacent neural events. Therefore, it alters the correlation between estimated patterns of temporally proximate neural events. To illustrate this bias, we simulate data which is generated by weighting the true design matrix reflecting realistic HRF with β's following the covariance structure in Fig 2A, but without adding any noise. If there were no noise, standard RSA on the estimates of β's would not introduce bias. However, with this approach on raw simulated data even without noise, we observe structured bias unrelated to the true covariance matrix shown in **Fig S-1**, which is due to the first term in (s-23).
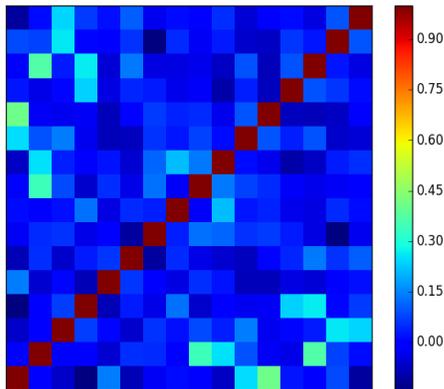
Figure S-1. A bias structure is obtained by applying standard RSA on the raw patterns of simulated data without noise added.

9

The second term in (s-23) is introduced because of the autocorrelation structure of fMRI noise, together with the event timing. To illustrate this bias, we applied the raw pattern approach to preprocessed resting state fMRI data of 30 participants in the Human Connectome Project (HCP), pretending the participants took part in the task in **Fig 1A**. These data have no relation with the task, so there is no chance of measuring any real neural activity related to the task. However, we observe a similarity structure with many high correlations, as shown in **Fig S-2**. This clearly illustrates the bias introduced by treating raw fMRI data as neural patterns of events.
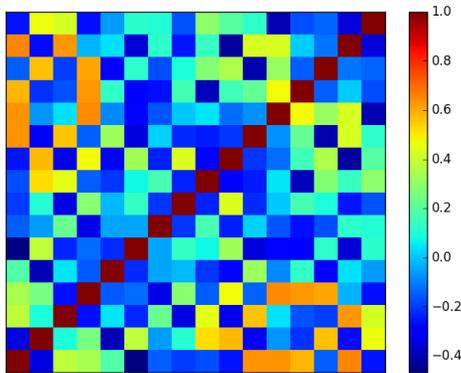
Figure S-2. A bias structure is obtained by applying standard RSA on the raw patterns of an unrelated resting state fMRI dataset.

Data used for the illustration in **S.5** were provided by the MGH-USC Human Connectome Project (HCP; Principal Investigators: Bruce Rosen, M.D., Ph.D., Arthur W. Toga, Ph.D., Van J. Weeden, MD). HCP funding was provided by the National Institute of Dental and Craniofacial Research (NIDCR), the National Institute of Mental Health (NIMH), and the National Institute of Neurological Disorders and Stroke (NINDS). HCP data are disseminated by the Laboratory of Neuro Imaging at the University of California, Los Angeles.

10