# Transfer Learning on fMRI Datasets
# Supplementary Material

**Hejia Zhang**
Department of Electrical Engineering
Princeton University
Princeton, NJ 08544
hejiaz@princeton.edu

**Po-Hsuan Chen**
Department of Electrical Engineering
Princeton University
Princeton, NJ 08544
pohsuan@princeton.edu

**Peter J. Ramadge**
Department of Electrical Engineering
Princeton University
Princeton, NJ 08544
ramadge@princeton.edu

Proceedings of the $21^{st}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. JMLR: W&CP volume 7X. 

## S.1 Notation

| Variable | Description |
|---|---|
| $v_i$ | number of voxels for subject $i$ |
| $n$ | number of datasets |
| $d$ | index for dataset, $d \in \{1, \ldots, n\}$ |
| $\mathcal{M}_d$ | set of subjects in dataset $d$ |
| $m$ | number of subjects across all datasets |
| $m_d$ | number of subjects in dataset $d$ |
| $i$ | index for subject, $i \in \{1, \ldots, m\}$ |
| $t_d$ | number of TRs in dataset $d$ |
| $t$ | index for TR |
| $k$ | dimensionality of the feature space |
| $q$ | index for feature, $q \in \{1, \ldots, k\}$ |
| $x_{dit}$ | $t$-th observation from subject $i$ in dataset $d$, taking values in $\mathbb{R}^{v_i}$ |
| $X_{di}$ | observations from subject $i$ in dataset $d$, $X_{di} \in \mathbb{R}^{v_i \times t_d}$ |
| $x_{dt}$ | $x_{dt}^T = [{x_{d1t}}^T \ldots {x_{dit}}^T]$, $x_{dt} \in \mathbb{R}^{\sum_i v_i}$<br>concatenated observation of $t$-th observations from all subjects in dataset $d$ |
| $X_{(d)}$ | $X_{(d)}^T = [{X_{d1}}^T \ldots {X_{di}}^T]$, $X_{(d)} \in \mathbb{R}^{\sum_i v_i \times t_d}$<br>concatenated observations from all subjects in dataset $d$ |
| $s_{dt}$ | estimated shared response of $t$-th observations in dataset $d$, $s_{dt} \in \mathbb{R}^k$ |
| $S_d$ | estimated shared response in dataset $d$, $S_d \in \mathbb{R}^{k \times t_d}$ |
| $\mu_{di}$ | mean observation from subject $i$ in dataset $d$, $\mu_{di} \in \mathbb{R}^{v_i}$ |
| $\mu_d$ | $\mu_d^T = [\mu_{d1}^T \ldots \mu_{dm_d}^T]$, $\mu_d \in \mathbb{R}^{\sum_i v_i}$<br>concatenated mean observation from all subjects in dataset $d$ |
| $W_i$ | subject specific mapping for subject $i$, $W_i \in \mathbb{R}^{v_i \times k}$ |
| $W_{(d)}$ | $W_{(d)}^T = [W_1^T \ldots W_{m_d}^T]$, $W_{(d)} \in \mathbb{R}^{\sum_i v_i \times k}$<br>concatenated subject specific mappings for all subjects in dataset $d$ |
| $\Sigma_{s_d}$ | covariance for shared response $s_{dt}$, $\Sigma_{s_d} \in \mathbb{R}^{k \times k}$ |
| $\rho_{di}^2 I_{v_i}$ | isotropic covariance for conditional distribution of $x_{it}$ |
| $\Psi_d$ | $\Psi_d = \text{diag}(\rho_1^2 I_{v_1}, \ldots, \rho_{m_d}^2 I_{v_{m_d}})$, $\Psi_d \in \mathbb{R}^{\sum_i v_i \times \sum_i v_i}$<br>joint covariance for condition distribution of $x_{dt}$ in dataset $d$ |

### S.2 Other single-dataset multi-subject models

We briefly introduce some other single-dataset multi-subject models. Independent Vector Analysis (IVA) [1, 2] is a deterministic algorithm that does not assume time synchronized stimulus and learns spatial independent components. Hyperalignment (HA) [3–5] is a deterministic model that uses temporally synchronized data to learn subject specific mappings that are generalizable to other stimulus. Topographic Factor Analysis (TFA) and Hierarchical Topographic Factor Analysis (HTFA) [6, 7] are two probabilistic factor models using a topographic basis composed of spherical Gaussians with different centers and widths. This choice of basis is constraining but since each factor is an "blob" in the brain it has the advantage of providing a simple spatial interpretation. HTFA has a further assumption that topographic bases across subjects are different perturbation from the same group template.

### S.3 Deterministic MDMS

We can model the data as $X_{di} = W_i S_d + E_{di}$ and minimize the cost function $\sum_{d=1}^{n} \sum_{i \in m_d} \|X_{di} - W_i S_d\|_F^2$. Furthermore, to ensure the uniqueness of coordinates it is necessary that $W_{di}$ has linearly independent columns. We make the stronger assumption that each $W_{di}$ has orthonormal columns, $W_i^T W_i = I_k$. Therefore, we formulate the model as

$$\begin{array}{ll} \min_{W_i, S_d} & \sum_{d=1}^{n} \sum_{i \in m_d} \|X_{di} - W_i S_d\|_F^2 \\ \text{s.t.} & W_i^T W_i = I_k, \end{array} \tag{1}$$

The parameters for this model can be estimated with an alternating optimization scheme. Each $W_i$ is first initialized as a random orthonormal matrix. Then we repeat the following steps until a stopping criterion in satisfied. Firstly, with respect to each $S_d$, optimize (1) by setting $S_d = {}^1/_{m_d} \sum_{i \in m_d} W_i^T X_{di}$, where $m_d$ is the number of subjects in dataset $d$. Secondly, with respect to each $W_i$, optimize (1) with solution $W_i = UV^T$, where $U \Sigma V^T$ is an SVD of $\sum_{d: i \in m_d} X_{di} S_d^T$ .

Next, we show how probabilistic MDMS matches deterministic MDMS by explicitly writing out its maximum likelihood estimation (MLE). First we notice the fact that $x_{dit}$ is the $t$'th column of $X_{di}$ and $s_{dt}$ is the $t$'th column of $S_d$. The negative log-likelihood of probabilistic MDMS is $\mathcal{L} = \sum_d \sum_{i \in m_d} \sum_t \frac{v_i}{2} \log 2\pi + \frac{v_i}{2} \log \rho_{di}^2 + \frac{\rho_{di}^{-2}}{2} (x_{dit} - W_i s_{dt} - \mu_{dt})^T (x_{dit_{dt}} - W_i s_{dt})$. Without loss of generality, assume the observations have zero-mean, so the $\mu_{dt}$ terms can be dropped. We also assume identical noise across subjects and datasets, $\rho_{di} = \rho \forall d, i$. The MLE can be computed by minimizing $\mathcal{L}$ with respect to $W_i$ and $S_d$:

$$\min \sum_d \sum_{i \in m_d} \sum_{t=1}^{t_d} (x_{dit} - W_i s_{dt})^T (x_{dit} - W_i s_{dt}) \tag{2}$$

By expanding and combining terms, (2) becomes:

$$\min \sum\nolimits_d \sum\nolimits_{i \in m_d} \|X_{di} - W_i S_d\|_F^2,$$

which is identical to the objective of deterministic MDMS .

### S.4 Derivation of constrained EM algorithm for MDMS

Let us define $\theta$ as the vector of all parameters, and $\theta^{\text{old}}$ as the initial value or estimated $\theta$ from the previous M-step. In the E-step, given $\theta^{\text{old}}$, we calculate the sufficient statistics by taking expectation with respect to $p(s_{dt} | x_{dit}, \theta^{\text{old}})$:

$$\begin{aligned} \mathbb{E}_{s_d|x_d}[s_{dt}] &= (W_{(d)} \Sigma_{s_d})^T (W_{(d)} \Sigma_{s_d} W_{(d)}^T + \Psi_d)^{-1} (x_{dt} - \mu_d), \\ \mathbb{E}_{s_d|x_d}[s_{dt} s_{dt}^T] &= \text{Var}_{s_d|x_d}[s_{dt}] + \mathbb{E}_{s_d|x_d}[s_{dt}] \mathbb{E}_{s_d|x_d}[s_{dt}]^T \\ &= \Sigma_{s_d} - \Sigma_{s_d}^T W_{(d)}^T (W_{(d)} \Sigma_{s_d} W_{(d)}^T + \Psi_d)^{-1} W_{(d)} \Sigma_{s_d} + \mathbb{E}_{s_d|x_d}[s_{dt}] \mathbb{E}_{s_d|x_d}[s_{dt}]^T. \end{aligned} \tag{3}$$

In the M-step, we first calculate $Q(\theta|\theta^{\text{old}})$:

$$
\begin{aligned}
Q(\theta|\theta^{\text{old}}) =& \sum_d \sum_{t=1}^{t_d} \int p(s_{dt}|x_{dt};\theta^{\text{old}}) \log p(x_{dt}, s_{dt};\theta) ds_{dt} \\
=& \sum_d \sum_{t=1}^{t_d} \int p(s_{dt}|x_{dt};\theta^{\text{old}}) (\log p(x_{dt}|s_{dt};\theta) + \log p(s_{dt};\theta)) ds_{dt} \\
=& \sum_d \sum_{t=1}^{t_d} \mathbb{E}_{s_d|x_d}[\log((2\pi)^{\frac{\sum_{i\in\mathcal{M}_d} v_i}{2}} |\Sigma_{x_d}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x_{dt} - W_{(d)}s_{dt} - \mu_d)^T \Sigma_{x_d}^{-1} \\
& (x_{dt} - W_{(d)}s_{dt} - \mu_d)\}) + \log((2\pi)^{\frac{k}{2}} |\Sigma_{s_d}|^{-\frac{1}{2}} \exp\{-\frac{1}{2} s_{dt}^T \Sigma_{s_d}^{-1} s_{dt}\})] \\
=& \sum_d (-\frac{t_d \sum_{i\in\mathcal{M}_d} v_i}{2} \log(2\pi) - \frac{t_d}{2}|\Sigma_{x_d}| - \frac{1}{2}\sum_{t=1}^{t_d} \mathbb{E}_{s_d|x_d}[(x_{dt} - W_{(d)}s_{dt} - \mu_d)^T \Sigma_{x_d}^{-1} \\
& (x_{dt} - W_{(d)}s_{dt} - \mu_d)] - \frac{t_d k}{2}\log(2\pi) - \frac{t_d}{2}|\Sigma_{s_d}| - \frac{1}{2}\sum_{i\in\mathcal{M}_d} \mathbb{E}_{s_d|x_d}[s_{dt}^T \Sigma_{s_d}^{-1} s_{dt}]).
\end{aligned}
$$

Parameters $\theta^{\text{new}}$ are estimated by maximizing $Q$ with respect to $W_i$, $\mu_{di}$, $\rho_{di}^2$, and $\Sigma_{s_d}$ separately. The orthogonality constraint of $W_i$ is brought in by adding $\text{tr}(\Lambda_i(W_i^T W_i - I))$ to the objective function, where $\Lambda_i$ is a symmetric matrix.

The update equations we get by setting the derivation

$$
\begin{aligned}
&\mu_{di}^{\text{new}} = \tfrac{1}{t_d}\textstyle\sum_{t=1}^{t_d} x_{dit}, \\
&W_i^{\text{new}} = A_i(A_i^T A_i)^{-1/2}, \quad A_i = \tfrac{1}{2}\big(\textstyle\sum_{d:i\in\mathcal{M}_d}\sum_{t=1}^{t_d}(x_{dit} - \mu_{di}^{\text{new}})\mathbb{E}_{s_d|x_d}[s_{dt}]^T\big), \\
&\rho_{di}^{2\ \text{new}} = \tfrac{1}{t_d v_i}\textstyle\sum_{t=1}^{t_d}\big(\|x_{dit} - \mu_{di}^{\text{new}}\|^2 - 2(x_{dit} - \mu_{di}^{\text{new}})^T W_i^{\text{new}} \mathbb{E}_{s_d|x_d}[s_{dt}] + \text{tr}(\mathbb{E}_{s_d|x_d}[s_{dt}s_{dt}^T])\big), \\
&\Sigma_{s_d}^{\text{new}} = \tfrac{1}{t_d}\textstyle\sum_{t=1}^{t_d}(\mathbb{E}_{s_d|x_d}[s_{dt}s_{dt}^T]).
\end{aligned}
\tag{4}
$$

### S.5 Tricks used to speed up MDMS

The computational bottleneck of solving MDMS with a constrained EM algorithm is the inversion of $(W_{(d)}\Sigma_{s_d}W_{(d)}^T + \Psi_d)^{-1}$, which is a size $V_d$ by $V_d$ matrix for dataset $d$, where $V_d = \sum_{i\in\mathcal{M}_d} v_i$. We adopt the method described in [8] to avoid computing this inversion directly by applying the matrix inversion lemma and use the facts that $W_i^T W_i = I_k$ and $\Psi_d$ is a diagonal matrix.
In details,

$$
\begin{aligned}
\Sigma_{s_d} - \Sigma_{s_d}^T W_{(d)}^T (W_{(d)}\Sigma_{s_d}W_{(d)}^T + \Psi_d)^{-1} W_{(d)}\Sigma_{s_d} &= (\Sigma_{s_d}^{-1} + W_{(d)}^T \Psi_d^{-1} W_{(d)})^{-1} \\
&= (\Sigma_{s_d}^{-1} + \textstyle\sum_{i\in\mathcal{M}_d} W_i^T(\rho_{di}^{-2} I)W_i)^{-1} \\
&= (\Sigma_{s_d}^{-1} + \rho_d I)^{-1}
\end{aligned}
\tag{5}
$$

, where $\rho_d = \sum_{i\in\mathcal{M}_d} \rho_{di}^{-2}$, and

$$
(W_{(d)}\Sigma_{s_d})^T (W_{(d)}\Sigma_{s_d}W_{(d)}^T + \Psi_d)^{-1} = \Sigma_{s_d}^T [I - \rho_d(\Sigma_{s_d}^{-1} + \rho_d I)^{-1}] W_{(d)}^T \Psi_d^{-1} \tag{6}
$$

In (5) and (6), only inversions of $k$ by $k$ matrices are involved. The computational complexity is reduced from $\mathcal{O}(V_d^2)$ to $\mathcal{O}(k^2)$.

### S.6 Connections of MDMS with other methods

We restate the mathematical formulation for MDMS and deterministic version of MDMS for ease of comparison with related methods. MDMS model:

$$
\begin{aligned}
s_{dt} &\sim \mathcal{N}(0, \Sigma_{s_d}), \\
x_{dit}|s_{dt} &\sim \mathcal{N}(W_i s_{dt} + \mu_{di}, \rho_{di}^2 I), \\
W_i^T W_i &= I_k,
\end{aligned}
\tag{7}
$$

The deterministic version of MDMS model:

$$\begin{array}{ll} \min_{W_i,S_d} & \sum_{d=1}^{n}\sum_{i\in\mathcal{M}_d}\|X_{di}-W_iS_d\|_F^2 \\ \text{s.t.} & W_i^TW_i=I_k, \end{array} \tag{8}$$

### S.6.1 MDMS and Group-ICA (GICA)

GICA [9] is a deterministic algorithm tries to learn independent spatial maps for single-dataset multi-subject analysis. We compare MDMS with a variant of GICA [10] that uses temporally synchronized stimulus since it has a closer connection with the problem we are addressing. We will abbreviate this variant as GICA in our following discussion. The basic goal of GICA is to decompose $X_i = W_iS$ for all subjects $i$ in a single dataset, where $X_i$ is data of subject $i$, such that $S$ contains temporal components that are statistically independent. There are $4$ steps in GICA: Firstly, perform a PCA along the spatial dimension for each subject $i$: $X_i = F_iP_i$. Secondly, concatenate the reduced data and perform a PCA along spatial dimension: $P = [P_1^T \cdots P_m^T]^T$, where $m$ is number of subjects in this dataset, and then $P = GY$. Thirdly, perform an ICA along spatial dimension: $Y = AS$. Lastly, partition $G$ matrix to subject-specific $G_i$: $[G_1^T \cdots G_m^T]^T = G$. The basis of subject $i$ is computed as $W_i = F_iG_iA$.

In a single dataset case, MDMS and GICA both try to approximate $X_i$ with $W_iS$ using different objectives and constraints. MDMS tries to minimize Frobenius norm of $l2$ loss with the constraint of orthonormal subject basis while GICA tries to find statistically independent components.

### S.6.2 MDMS and Dictionary learning (DL)

DL [11] tries to learn subject-specific basis $W_i$ and loadings $U_i$ while regularizing $W_i$ to be similar to a group template $W$. As DL does not require temporally synchronized stimulus, we can concatenate data from different datasets of the same subject together and use the formulation of DL directly in a multi-dataset setting. Here we compare MDMS with DL applied on multiple datasets (MDDL). MDMS and MDDL are closely related. The formulation of MDDL is as follows: Denote $X_i$ a concatenation of subject $i$'s data from different datasets along temporal dimension. More specifically, $X_i = [X_{d_1i} \cdots X_{d_pi}]$, where $\{d_1 \cdots d_p\} = set\{d : i \in \mathcal{M}_d\}$. Plug it into the formulation of DL

$$\min_{U_i,W_i,W} \quad \sum_{i\in m}\tfrac{1}{2}(\|X_i^T - U_iW_i^T\|_F^2 + \mu\|W_i - W\|_F^2) + \mu\alpha\Omega(W) \tag{9}$$

Note that it is a transpose of the data matrix used in [11]. $\Omega$ is a regularization on the group template $W$. We used Total-Variation (TV) as $\Omega$ as in [11].
We now rewrite this formulation by transposing the first term and combining the second and third terms as a general regularization term. Then (9) becomes

$$\min_{U_i,W_i,W} \quad \sum_{i\in m}\tfrac{1}{2}\|X_i - W_iU_i^T\|_F^2 + \beta\Omega(W_i, W) \tag{10}$$

, where $\Omega$ stands for a general regularization term on $W_i$ and $W$.
Expand $X_i$ in the first term, we can further rewrite (10) as

$$\min_{U_i,W_i,W} \quad \sum_{d=1}^{n}\sum_{i\in m_d}\|X_{di} - W_iU_{di}^T\|_F^2 + \beta\Omega(W_i, W) \tag{11}$$

In this case, each data point $X_{di}$ has its own loadings $U_{di}$.
Note that the constraint $W_i^TW_i = I_k$ in (8) can also be written as a general regularization term with a Lagrange multiplier, and (8) would become

$$\min_{W_i,S_d} \quad \sum_{d=1}^{n}\sum_{i\in m_d}\|X_{di} - W_iS_d\|_F^2 + \beta\Omega(W_i) \tag{12}$$

If we enforce loadings in (11) to be the same within each dataset, and denote the transpose of dataset specific loadings as $S_d$, then (11) and (12) have the same objective with different regularizations. In our experiments, we use the formulation of (11) in training, and whenever we need the $k$-dimensional feature of dataset $d$, we set $S_d = {}^1/_{|m_d|}\sum_{i\in\mathcal{M}_d}U_{di}^T$.

The main difference between MDMS and MDDL is that MDMS utilizes temporally synchrony during optimization, so it would be more useful when temporally synchronized data is available. MDDL does not assume temporally synchronized data, but assumes spatial patterns of subject basis, so it would be more useful when finding spatial maps in resting-state data.

### S.6.3 MDMS and SRM

We show that MDMS degenerates to SRM [12] when number of datasets $n = 1$. Set $n = 1$, and ask data of all subjects to be available, then (7) becomes

$$\begin{aligned} s_t &\sim \mathcal{N}(0, \Sigma_s), \\ x_{it}|s_t &\sim \mathcal{N}(W_i s_t + \mu_i, \rho_i^2 I), \\ W_i^T W_i &= I_k, \end{aligned} \tag{13}$$

, where $x_{it} \in \mathbb{R}^{v_i}$ are the observations from subject $i$ in this single dataset at time $t$, $s_t$ is the latent variable of this dataset. The SRM has the identical formulation as (13) in this degenerated case. SRM cannot be applied to multiple datasets directly because SRM requires data from all subjects have the same stimulus, which is not the case in a multiple dataset analysis in general.

## S.7 Subject list of each dataset

There are 5 datasets and 85 subjects in total. We number the subjects from 1 to 85, and list here indices of subjects in each dataset.
*greeneyes*: [1-40]
*milky*: [1,8,13,17,18,19,22,27,28,29,31,33,35,37,40,41,42,43]
*vodka*: [4,5,6,7,9,11,12,15,21,23,24,25,26,32,34,36,38,39]
*sherlock* (and *sherlock-recall*): [8,20,31,35,44-55]
*schema*: [38,56-85]

## S.8 More experiment results

### S.8.1 Do secondary datasets help learning?

Here we show time segment matching results on PT and EAC in Fig. 1. We see MDMS has a robust performance across ROIs we tested, while MDDL is not as robust as MDMS . It suggests MDMS enables leveraging secondary datasets robustly.

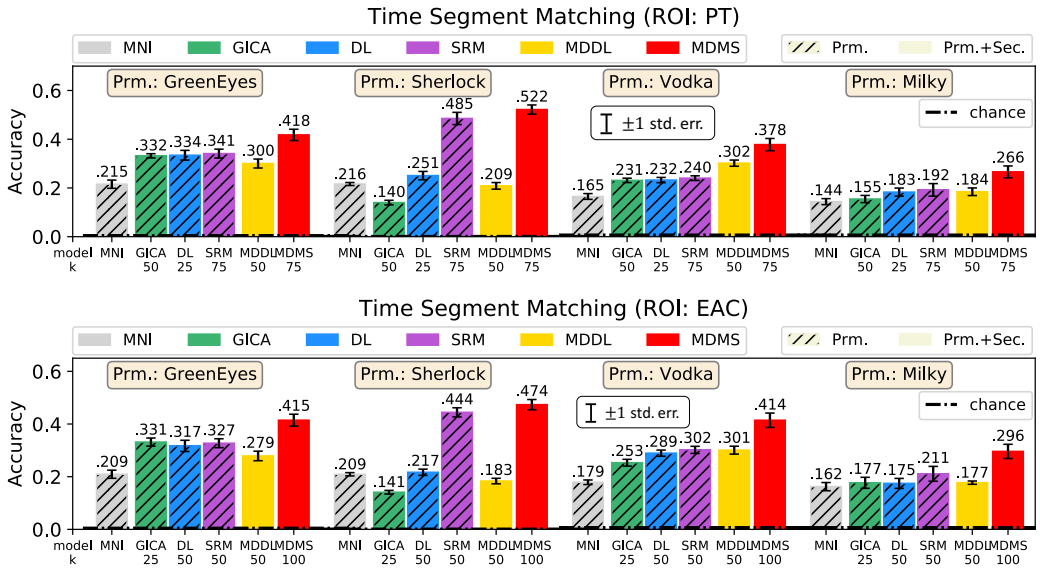


Figure 1: Results of time segment matching on PT and EAC. Chance accuracy: *greeneyes*: 0.005; *sherlock*: 0.001; *vodka*: 0.008; *milky*: 0.008. $k$ values selected based on cross-validation.

### S.8.2 leveraging secondary datasets for semantic embedding

We show the classification accuracy of fMRI data to text semantic embedding mapping experiment on EAC and DMN in Fig. 2. We see MDMS and MDDL are the two methods that have the highest accuracy in this experiment among methods we tested here.

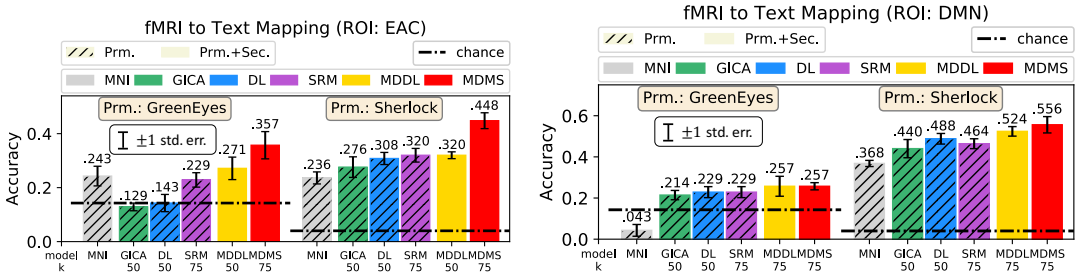


Figure 2: fMRI data to text embedding mapping classification accuracy. Results on EAC and DMN. Chance accuracy: *greeneyes*:0.14; *sherlock*: 0.04. $k$ selected based on cross-validation.

### S.8.3 Transfer learning to unseen datasets

We show the results from experiment 3 on PT and EAC in Fig. 3. It suggests we can use small secondary datasets as anchors to transfer information from big secondary datasets to the primary dataset.

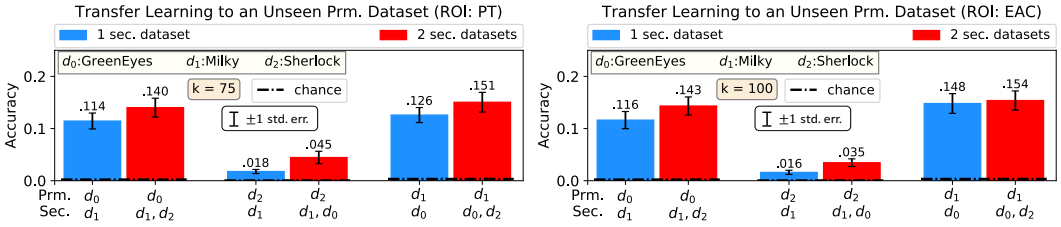


Figure 3: Time segment matching accuracy on prm. dataset using subject basis learned from 1 or 2 sec. datasets. Results on PT and EAC. Chance accuracy: *greeneyes*: 0.0025; *milky*: 0.004; *sherlock*: 0.0005. $k$ value same as experiment 1.

### S.8.4 Effect of adding independent subjects in secondary datasets

We evaluate the effect of different number of independent subjects in the secondary dataset. We start with only one shared subject in the secondary dataset, and then add in independent subjects one by one. No clear pattern on accuracies can be observed. Result from a pair of primary and secondary dataset is shown in Fig. 4 as an example. This suggests that independent subjects in a single secondary dataset do not necessarily help the primary dataset directly. We would then like to explore if independent subjects can help the primary dataset indirectly, and notice that dataset *schema* only has one shared subject with other datasets, and has no direct shared subject with dataset *milky* and *sherlock*, but has many independent subjects. We redo time segment matching in experiment 1 without dataset *schema* and see how the accuracies change. Results on all tested ROIs shown in Fig. 4. We notice that small primary datasets, including *milky*, still get improvement after adding *schema* as a secondary dataset. These results imply a possibility that independent subjects are used as anchors between secondary datasets so that information from all connected secondary datasets can propagate to the primary dataset.

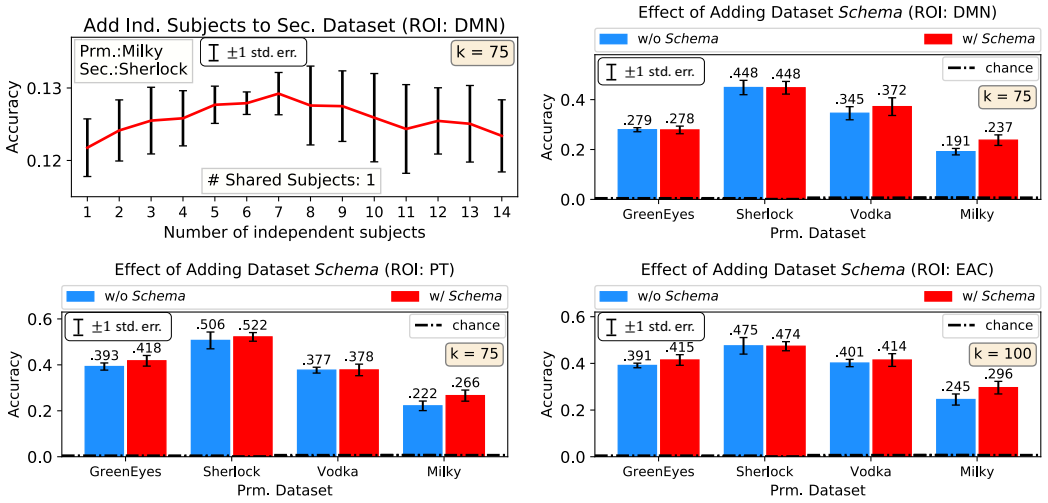


Figure 4: **Top(left)**: Time segment matching accuracy on the prm. dataset when the sec. dataset has 1 shared subject and different number of independent subjects. **Top(right) and Bottom**: Time segment matching accuracy before and after adding the *schema* dataset on DMN, PT and EAC. Chance accuracy and $k$ values same as experiment 1.

## References


[1] J. Lee et al. Independent vector analysis: multivariate approach for fMRI group study. *Neuroimage*, 2008.

[2] A. M. Michael et al. Preserving subject variability in group fMRI analysis: performance evaluation of GICA vs. IVA. *Frontiers in systems neuroscience*, 8, 2014.

[3] J. V. Haxby et al. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011.

[4] J. V. Haxby, A. C. Connolly, and J. S. Guntupalli. Decoding neural representational spaces using multivariate pattern analysis. *Annual review of neuroscience*, 37:435–456, 2014.

[5] J. S. Guntupalli et al. A model of representational spaces in human cortex. *Cerebral Cortex*, 2016.

[6] J. R. Manning, R. Ranganath, K. A. Norman, and D. M. Blei. Topographic factor analysis: a bayesian model for inferring brain networks from neural data. *PloS one*, 2014.

[7] J. R. Manning et al. Hierarchical topographic factor analysis. In *IEEE Pattern Rec. in Neuroimaging*, 2014.

[8] M. J. Anderson et al. Enabling factor analysis on thousand-subject neuroimaging datasets. *arXiv*, 2016.

[9] V. D. Calhoun, J. Liu, and T. Adalı. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *Neuroimage*, 45(1):S163–S172, 2009.

[10] H. Zhang et al. A searchlight factor model approach for locating shared information in multi-subject fMRI analysis. *arXiv:1609.09432*, 2016.

[11] A. Abraham et al. Extracting brain regions from rest fMRI with total-variation constrained dictionary learning. In *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2013.

[12] P.-H. Chen et al. A reduced-dimension fMRI shared response model. In *NIPS*, 2015.