Sparsistent Learning of Varying-coefficient Models with Structural Changes

Mladen Kolar, Le Song and Eric P. Xing
School of Computer Science, Carnegie Mellon University
{mkolar,lesong,epxing}@cs.cmu.edu

1 Appendix

1.1 TD-transformation

In this Section we detail the TD-transformation used in Section 3, to transform the TD regression

$$\sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i' \beta(t_i)\right)^2 + 2\lambda_2 \sum_{k=1}^{p} ||\beta_k||_{\mathrm{TV}}\,, \tag{1}$$

into an ℓ_1 penalized regression

$$\hat{\beta}^\dagger = \underset{\beta \in \mathbb{R}^{np}}{\operatorname{argmin}} \left|\left|Y^\dagger - \mathbf{X}^\dagger \beta^\dagger\right|\right|_2^2 + 2\lambda_2 \left|\left|\beta^\dagger\right|\right|_1. \tag{2}$$

We start by demonstrating the transformation for the following model with one variable, *i.e.*, $X \in \mathbb{R}$,

$$Y_i = X_i \beta_i + \epsilon_i, \qquad i = 1, \ldots, n. \tag{3}$$

Subtracting two consecutive equations, we have for $i = 2, \ldots, n$,

$$\begin{aligned} Y_i - Y_{i-1} &= X_i\beta_i - X_{i-1}\beta_{i-1} + \epsilon_i - \epsilon_{i-1} \\ &= X_i(\beta_i - \beta_{i-1}) + (X_i - X_{i-1})\beta_{i-1} + \epsilon_i - \epsilon_{i-1} \\ &= X_i(\beta_i - \beta_{i-1}) + \sum_{j=2}^{i} (X_j - X_{j-1})\beta_{j-1} + \epsilon_i - \epsilon_{i-1}, \end{aligned}$$

which after introducing $Y_1^\dagger = Y_1, Y_i^\dagger = Y_i - Y_{i-1}, \beta_1^\dagger = \beta_1, \beta_i^\dagger = \beta_i - \beta_{i-1}, \epsilon_1^\dagger = \epsilon_1, \epsilon_i^\dagger = \epsilon_i - \epsilon_{i-1}, i = 1, \ldots, n$ and

$$\mathbf{X}_1^\dagger = \begin{pmatrix} X_1 & 0 & 0 & \ldots & 0 \\ X_2 - X_1 & X_2 & 0 & \ldots & 0 \\ X_3 - X_2 & X_3 - X_2 & X_3 & \ldots & 0 \\ \vdots & & & \ddots & \vdots \\ X_n - X_{n-1} & X_n - X_{n-1} & X_n - X_{n-1} & \ldots & X_n \end{pmatrix}$$

can be written in a matrix form as

$$Y^\dagger = \mathbf{X}_1^\dagger \beta^\dagger + \epsilon^\dagger. \tag{4}$$

Similarly, for a general multivariate model $Y_i = \mathbf{X}_i \beta_i + \epsilon_i,\ i = 1, \ldots, n,\ \mathbf{X}_i \in \mathbb{R}^p$ we perform the same transformation with the difference that $\mathbf{X}^\dagger = (\mathbf{X}_1^\dagger, \ldots, \mathbf{X}_p^\dagger), \mathbf{X}^\dagger \in \mathbb{R}^{n \times np}$ is obtained by concatenating matrices corresponding to TD features.

1.2 Randomized Lasso

In this paper, we use the randomized Lasso to identify the partition of the interval $[0, 1]$ on which the regression coefficient are constant. The partition could be obtained by finding a minimum in Eq. (2), however, note that the randomized Lasso does not find a solution that minimizes Eq. (2). The randomized Lasso finds a minimum of a related optimization problem

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (Y_i^\dagger - \mathbf{X}_i^\dagger \beta)^2 + 2\lambda \sum_{k=1}^{p} \frac{|\beta_k|}{W_k} \tag{5}$$

where $\{W_k\}_{k=1}^p$ are independent and identically distributed uniform random variables on an interval $[\alpha, 1]$ and α is a weakness parameter. It can be seen from Eq. (5) that the randomized Lasso assigns each feature a different penalty, which can also be thought of as rescaling different features. It can be shown [8] that this rescaling weakens the necessary condition for the ordinary Lasso to select the relevant features. Note that the randomized Lasso is a random algorithm and in order to see benefits over the ordinary Lasso, it has to be run multiple times. We use it together with bootstrap, so, instead of running the random Lasso multiple times on a same dataset, it is run on multiple bootstrap replicas of the dataset. For each run of the algorithm, an estimate of the partition $\hat{\mathcal{T}}$ is obtained. Finally, one obtains the estimate of the partition by keeping the jump points that appear in more than τ fraction of the partitions, where τ is a tuning parameter that controls the number of falsely identified jumps. Theorem 1 in [8] provides a way to choose the parameter τ. The weakness parameter α also plays role in falsely choosing jump points. The lower values of α help to reduce the number of false positives, however, this compromises numerical stability. Notice that small values of α can affect the condition number of the design matrix. [8] reports that choosing $\alpha \in (0.2, 0.8)$ gives good results. From our numerical experience, we have noticed that the choice of parameter α does not have a huge impact on the solution. We choose to report experimental results with the value $\alpha = 0.6$.

1.3 Optimization Procedure

In this section, we outline a coordinate descent algorithm for finding the minimum of the following objective

$$\sum_{i=1}^{n} \left(Y_i - \sum_{k=1}^{p} X_{i,k} \beta_{k,i} \right)^2 + 2\lambda_2 \sum_{k=1}^{p} ||\beta_k||_{\mathrm{TV}} \,. \tag{6}$$

The algorithm we are about to describe is motivated by the coordinate descent algorithm for Lasso [2]. Let us introduce $\beta' \in \mathbb{R}^{p \times n}$ such that $\beta_{k,i} = \sum_{j=1}^{i} \beta'_{k,j}$. Using β' Eq. (6) becomes

$$\sum_{i=1}^{n} \left(Y_i - \sum_{k=1}^{p} X_{i,k} \sum_{j=1}^{i} \beta_{k,j} \right)^2 + 2\lambda_2 \sum_{k=1}^{p} \sum_{j=1}^{n} |\beta'_{k,j}|. \tag{7}$$

Keeping all the coefficients but $\beta'_{u,v}$ fixed, one can find a minimizer of Eq. (7) in a closed form

$$\hat{\beta}'_{u,v} = \frac{C}{\sum_{i=v}^{n} X_{i,u}^2} \left(1 - \frac{\lambda_2}{|C|} \right)_+ \tag{8}$$

where

$$C = \sum_{i=v}^{n} Y_i X_{i,u} - \sum_{\substack{k=1 \\ k \neq u}}^{p} \sum_{i=v}^{n} X_{i,u} X_{i,v} \beta_{k,i} - \sum_{i=v}^{n} X_{i,u}^2 \sum_{\substack{j=1 \\ j \neq v}}^{i} \beta'_{u,j}. \tag{9}$$

Now, the minimization algorithm iteratively minimizes coefficients $\beta'_{u,v}$, $u = 1, \ldots, p$, $v = 1, \ldots, n$, while keeping the rest fixed, until convergence. The algorithm is guaranteed to converge a global optimum due to the result of [6].

1.4 Applications and Generalizations

In this section we provide some insight on how to apply TDB-lasso to different models. We consider time-varying Gaussian graphical models [10], generalized varying-coefficient models and time-varying Markov Random Fields.

Consider a time-varying Gaussian graphical model [10], defined with the covariance function $\mathbf{\Sigma}(t) = [\sigma_{jk}(t)]_{j,k\in 1\dots p}$, whose components are functions of time. There are n independent samples, $\mathbf{X}_1, \dots, \mathbf{X}_n,\ \mathbf{X}_i \in \mathbb{R}^p$, each drawn from a different underlying multivariate Gaussian distribution with covariance $\mathbf{\Sigma}(t_i),\ t_i = i/n$, and the problem is to estimate $\mathbf{\Sigma}(\tau)$, for a given time point $\tau \in [0, 1]$. In [10], it is assumed that the coefficient functions $\sigma_{jk}(t)$ are smooth and the rate of convergence in Frobenius norm is established for a nonparametric kernel estimator $\hat{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}(\tau)$. Often, estimating the structure of the graphical model is an important problem to domain experts, who can interpret it more easily than the coefficient values. We address the problem of the structure estimation of a time-varying Gaussian graphical model, under the assumption that components of $\mathbf{\Sigma}(t)$ are piecewise constant functions. The structure of the graphical model is encoded with a non-zero pattern of the concentration matrix $\mathbf{\Omega}(t) := \mathbf{\Sigma}(t)^{-1}$ [4]. Let $\mathbf{\Omega}(t) = [\omega_{jk}(t)]_{j,k\in 1\dots p}$. It is possible to formulate the structure estimation problem as a sequence of regression problems by regressing each variable $X_{i,j}$ to the rest of variables $X_{i,\setminus j}$ [7]:

$$X_{i,j} = \sum_{k\neq j} \beta_{jk}(t_i) X_{i,k} + \epsilon_j, \quad i \in 1, \dots, n, \tag{10}$$

where $\beta_{jk}(t_i) = -\frac{\omega_{jk}(t_i)}{\omega_{jj}(t_i)}$. Estimating the non-zero pattern of $\mathbf{\Omega}(\tau)$ is equivalent to estimating the non-zero pattern of vector $\beta_j(\tau)$ for all $j = 1, \dots, p$. We fit the parameters of the model (10) using the TDB-lasso procedure.

Now, consider a generalized varying-coefficient model

$$g(m(X_i, t_i)) = \mathbf{X}_i \beta(t_i), \quad i = 1, \dots, n, \tag{11}$$

where $g(\cdot)$ is a given link function and $m(\mathbf{X}_i, t_i) = \mathbb{E}[Y|\mathbf{X} = \mathbf{X}_i, t = t_i]$ is the conditional mean. Again, we assume that the coefficient functions are piecewise constant and formulate the estimate $\hat{\beta}$ as a solution to the following penalized log-likelihood problem

$$\hat{\beta} = \min_{\beta} \{ -\sum_{i=1}^{n} \ell(Y_i, \mathbf{X}_i, t_i) + \lambda_1^n \sum_{i=1}^{n} ||\beta(t_i)||_1 + \lambda_2^n \sum_{k=1}^{p} ||\beta_k||_{\mathrm{TV}} \}, \tag{12}$$

where $\ell(Y_i, \mathbf{X}_i, t_i) = \ell(Y_i, m(\mathbf{X}_i, t_i)$ is the conditional log-likelihood function. Instead of directly minimizing (12) we can again use the two step procedure, in which the first step aims to identify block partitions on which the coefficient functions are constant, and the second step estimates the coefficient functions using the ℓ_1 penalized log-likelihood maximization.

One immediate application of the generalized varying-coefficient model is in estimation of time-varying discrete networks [3]. The problem can be described as the graph structure estimation problem of a Markov Random Field, whose structure is constant on unknown blocks and changes abruptly between blocks. To estimate the varying structure, one decomposes the estimation across different nodes of the graph and separately estimates neighborhoods of these nodes. One neighborhood estimation problem can be modeled using the generalized varying-coefficient model (11) and the estimated neighborhood corresponds to the non-zero coefficients $\hat{\beta}$.

1.5 Additional Real Data Experiments

The TDB-Lasso procedure is run on EEG measurements to infer how the brain interactions form over the time of the experiment. We regress the measurement of one channel onto the rest of channels to estimate the connections with other channels. Repeating this procedure, we estimate a network of interactions between different positions in the brain. We report the estimated interactions at three time points after the visual cues have been presented. Fig. 1 gives visualization of the brain interactions when the subjects were presented visual cues from the class 1 (right) and Fig. 2 for the class 2 (foot).

1.6 Proof of Lemma 1

Lemma 1 *Let $\hat{\gamma}_j$ and $\hat{\mathcal{B}}_j,\ j = 1, \dots, \hat{B}$ be vectors and segments obtained from a minimizer of*

$$\min_{\beta} \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i' \beta(t_i)\right)^2 + 2\lambda_1 \sum_{i=1}^{n} ||\beta(t_i)||_1 + 2\lambda_2 \sum_{k=1}^{p} ||\beta_k||_{\mathrm{TV}}\,. \tag{13}$$

Then each $\hat{\gamma}_j$ can be found as a solution to the subgradient equation:

$$\mathbf{X}'_{\hat{\mathcal{B}}_j}\mathbf{X}_{\hat{\mathcal{B}}_j}\hat{\gamma}_j - \mathbf{X}'_{\hat{\mathcal{B}}_j}Y_{\hat{\mathcal{B}}_j} + \lambda_1|\hat{\mathcal{B}}_j|\hat{s}_j^{(1)} + \lambda_2\hat{s}_j^{(\mathrm{TV})} = 0, \tag{14}$$

where

$$\hat{s}_j^{(1)} \in \partial\,||\hat{\gamma}_j||_1 = \mathrm{sign}(\gamma_j), \tag{15}$$

by convention $\mathrm{sign}(0) \in [-1,1]$*, and* $\hat{s}_j^{(\mathrm{TV})} \in \mathbb{R}^p$ *such that*

$$\hat{s}_{1,k}^{(\mathrm{TV})} = \begin{cases} -1 & \text{if } \hat{\gamma}_{2,k} - \hat{\gamma}_{1,k} > 0 \\ 1 & \text{if } \hat{\gamma}_{2,k} - \hat{\gamma}_{1,k} < 0 \end{cases}, \quad \hat{s}_{\hat{B},k}^{(\mathrm{TV})} = \begin{cases} 1 & \text{if } \hat{\gamma}_{\hat{B},k} - \hat{\gamma}_{\hat{B}-1,k} > 0 \\ -1 & \text{if } \hat{\gamma}_{\hat{B},k} - \hat{\gamma}_{\hat{B}-1,k} < 0 \end{cases} \tag{16}$$

and, for $1 < j < \hat{B}$*,*

$$\hat{s}_{j,k}^{(\mathrm{TV})} = \begin{cases} 2 & \text{if } \hat{\gamma}_{j+1,k} - \hat{\gamma}_{j,k} > 0, \hat{\gamma}_{j,k} - \hat{\gamma}_{j-1,k} < 0 \\ -2 & \text{if } \hat{\gamma}_{j+1,k} - \hat{\gamma}_{j,k} < 0, \hat{\gamma}_{j,k} - \hat{\gamma}_{j-1,k} > 0 \\ 0 & \text{if } (\hat{\gamma}_{j,k} - \hat{\gamma}_{j-1,k})(\hat{\gamma}_{j+1,k} - \hat{\gamma}_{j,k}) = 1. \end{cases} \tag{17}$$

Proof We briefly sketch the proof idea, which is based on the analysis of the subgradient of Eq. (14). The subgradient $\partial\,||x||_1$ of $x \in \mathbb{R}^p$ is the set $\{s \in \mathbb{R}^p \mid s = \mathrm{sign}(x)\}$ where $\mathrm{sign}(0) \in [-1,1]$ by definition. Let $L \in \mathbb{R}^{(p-1)\times p}$ be a matrix with entries $T_{k,k} = -1$ and $T_{k,k+1} = 1$ for $k = 1,\ldots,p-1$ and 0 otherwise. Now $\partial\,||X||_{\mathrm{TV}} = \partial\,||LX||_1$. Using this we compute the subgradient of (13) with respect to γ_j and Lemma follows by grouping variables within estimated segments. ■

1.7 Proof of Theorem 1

Theorem 1 *Let A1 be satisfied. Let the weakness α be given as $\alpha^2 = \nu\varphi_{\min}(CJ^2, \mathbf{X}^\dagger)/(CJ^2)$, for any $\nu \in (7/\kappa, 1/\sqrt{2})$. If the minimum size of the jump is bounded away from zero as*

$$\min_{k\in\mathcal{J}^\dagger} |\beta_k^\dagger| \geq 0.3(CJ)^{3/2}\lambda_{\min}, \tag{18}$$

where $\lambda_{\min} = 2\sigma^\dagger(\sqrt{C}J+1)\sqrt{\frac{\log np}{n}}$ and $\sigma^{\dagger^2} \geq Var(Y_i^\dagger)$, for $np > 10$ and $J \geq 7$, there exists some $\delta = \delta_J \in (0,1)$ such that for all $\tau \geq 1-\delta$, the collection of the estimated jump points $\hat{\mathcal{J}}^\tau$ satisfies,

$$\mathbb{P}(\hat{\mathcal{J}}^\tau = \mathcal{J}^\dagger) \geq 1 - 5/np. \tag{19}$$

Proof

We briefly sketch the main proof ideas here. The proof strategy is based on the approach of [8], which is based on the approach of [9].

The first difficulty we have to take care of is that after the TD-transformation is done, the samples are not *i.i.d.*any more. After the transformation, the elements of the vector $\epsilon^\dagger$ are not independent. For each $2 \leq i \neq j \leq n$ it holds $\mathbb{E}\epsilon_i^\dagger = 0$, $\mathrm{Var}\,\epsilon_i^\dagger = 2\sigma^2$ and

$$\mathrm{Cov}(\epsilon_i^\dagger, \epsilon_j^\dagger) = \begin{cases} -\sigma^2 & \text{if } |i-j| = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

Let $\sigma^{\dagger^2} = 2\sigma^2$ and for $i = 1,\ldots,n$ let $\epsilon_i^* \sim N(0, \sigma^{\dagger^2})$ be independent, such that

$$\begin{cases} \mathbb{E}(\epsilon_i^\dagger\epsilon_i^\dagger) \leq \mathbb{E}(\epsilon_i^*\epsilon_i^*) & 1 \leq i \neq j \leq n \\ \mathbb{E}(\epsilon_i^\dagger)^2 = \mathbb{E}(\epsilon_i^*)^2 & 1 \leq i \leq n. \end{cases} \tag{21}$$

Using Slepian's inequality (see *e.g.* [5]), we can substitute variables ϵ_i^* in place of $\epsilon_i^\dagger$.

The rest of the proof follows the same strategy as [8], with small modifications due to the use of Slepian's inequality and the fact that we use bootstrap instead of subsampling. We leave out the details to a full version of the paper.

■

1.8 Proof of Theorem 2

Theorem 2 *Let A2 be satisfied. Also, assume that the conditions of Theorem 1 are satisfied. Let* $K = \max_{1\le j\le B} ||\gamma_j||_0$ *be the upper bound on the number of features in segments and let* L *be an upper bound on elements of* $\mathbf{X}$*. Let* $\rho = \min_{1\le j\le B} |\mathcal{B}_j|$ *denote the number of samples in the smallest segment. Then for a sequence* $\delta = \delta_n \to 0$*,*

$$\lambda_1 \ge 4L\sigma\sqrt{\frac{\ln\frac{2Kp}{\delta}}{\rho}} \vee 8L\frac{\ln\frac{4Kp}{\delta}}{\rho} \quad \textit{and} \quad \min_{1\le j\le B}\min_{k\in S_{\mathcal{B}_j}} |\gamma_{j,k}| \ge 2\lambda_1,$$

we have

$$\lim_{n\to\infty} \mathbb{P}(\hat{B} = B) = 1, \tag{22}$$

$$\lim_{n\to\infty} \max_{1\le j\le B} \mathbb{P}(||\hat{\gamma}_j - \gamma_j||_1 = 0) = 1, \tag{23}$$

$$\lim_{n\to\infty} \min_{1\le j\le B} \mathbb{P}(\hat{S}_{\mathcal{B}_j} = S_{\mathcal{B}_j}) = 1. \tag{24}$$

Proof Under the assumptions of Theorem 1, for any $\delta' > 0$ and sufficiently large n, $\hat{\mathcal{J}}^\tau = \mathcal{J}^\dagger$ with probability at least $1 - \delta'$, *i.e.*,, the jump points can be estimated consistently. This implies equation (22).

Now, on each of the estimated segments, coefficient values are estimated using the ordinary Lasso. Under the assumption A1, we can apply the known results of the Lasso procedure on each of the estimated segments. Theorem follows from [1] after some adjustments of constants. ■

References

[1] Florentina Bunea. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electronic Journal of Statistics*, 2:1153, 2008.

[2] J. Friedman, T. Hastie, H. Hofling, and R. Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1:302, 2007.

[3] Mladen Kolar, Le Song, and Eric Xing. Estimating time-varying networks. In *arXiv:0812.5087*, 2008.

[4] S. L. Lauritzen. *Graphical Models (Oxford Statistical Science Series)*. Oxford University Press, USA, July 1996.

[5] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, 1991.

[6] Z. Q. Luo and P. Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *J. Optim. Theory Appl.*, 72(1):7–35, 1992.

[7] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34:1436, 2006.

[8] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Preprint*, 2008.

[9] Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36(4):1567–1594, 2008.

[10] Shuheng Zhou, John Lafferty, and Larry Wasserman. Time varying undirected graphs. In Rocco A. Servedio and Tong Zhang, editors, *COLT*, pages 455–466. Omnipress, 2008.

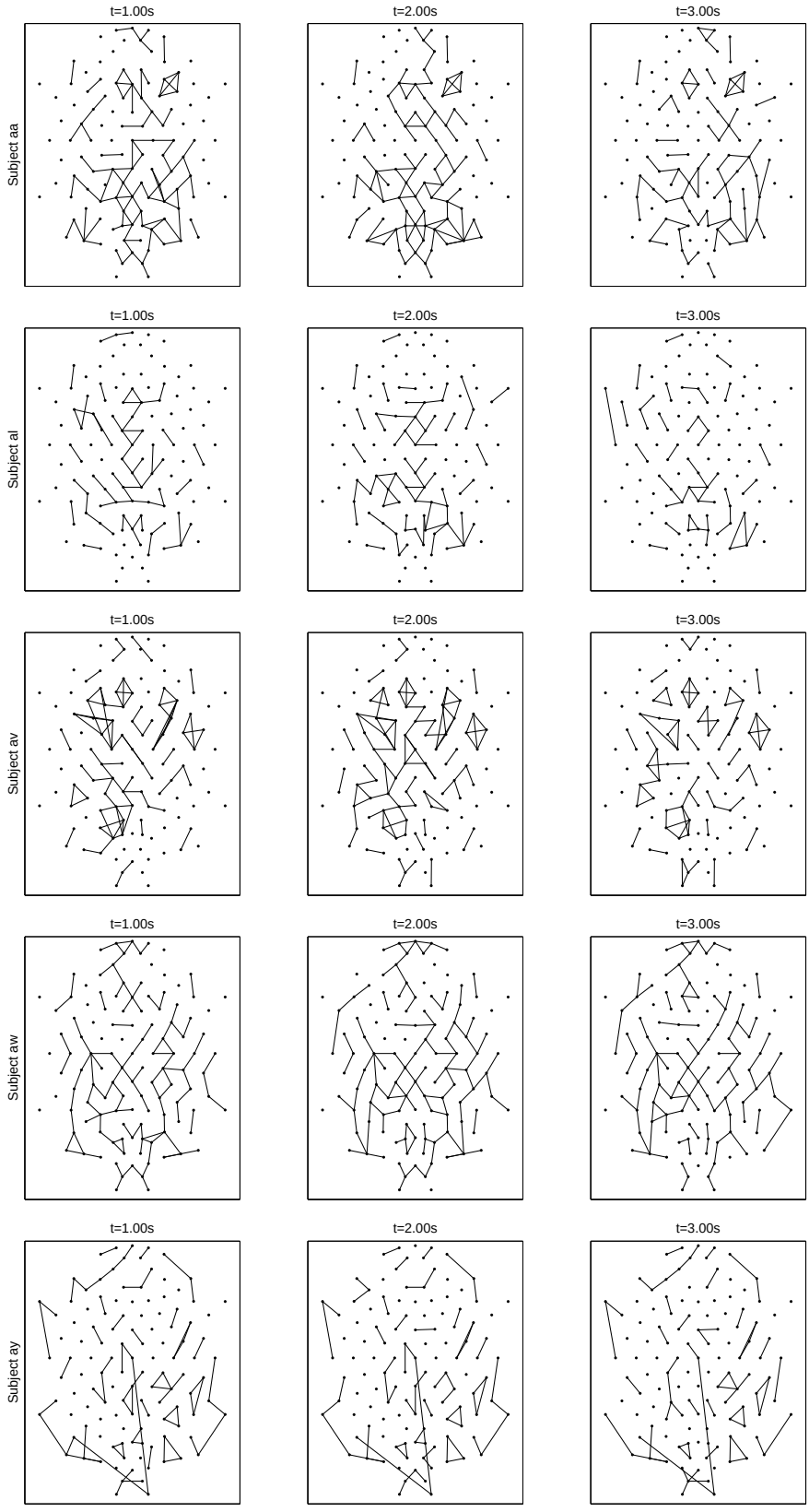

Figure 1: Brain interactions of different subjects estimated at different times after the visual cues for Class 1 (right hand) had been presented.

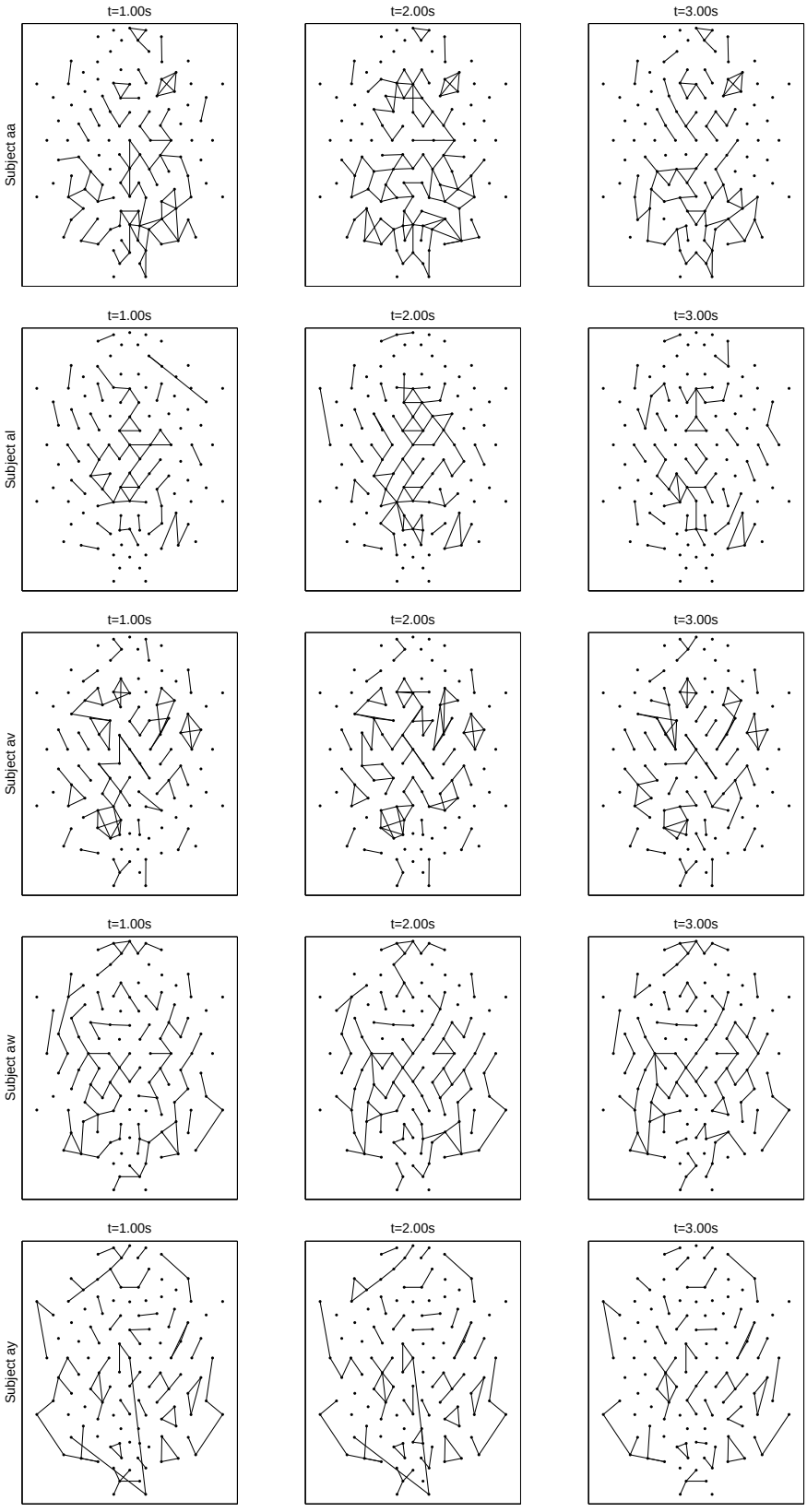

Figure 2: Brain interactions of different subjects estimated at different times after the visual cues for Class 2 (right foot) had been presented.