British Journal of Mathematics & Computer Science
14(4): 1-13, 2016, Article no.BJMCS.23829
ISSN: 2231-0851

SCIENCEDOMAIN *international*
www.sciencedomain.org

Spatial Regularization for Multitask Learning and Application in fMRI Data Analysis

Xin Yang[1], Qiang Wu[1], Jiancheng Zou[1,2] and Don Hong[1,2*]

[1] *Computational Science Program, Middle Tennessee State University, Murfreesboro, TN, USA.*
[2] *College of Sciences, North China University of Technology, Beijing, China.*

Authors' contributions:

This work was carried out in collaboration between all authors. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/BJMCS/2016/23829
Editor(s):
(1) H. M. Srivastava, Department of Mathematics and Statistics, University of Victoria, Canada.
Reviewers:
(1) Loc Nguyen, Vietnam.
(2) Wojciech Salabun, West Pomeranian University of Technology, Szczecin, Poland.
(3) S. Zimeras, University of the Aegean, Greece.
(4) Behnam Sharif, University of Calgary, Alberta, Canada.
Complete Peer review History: http://sciencedomain.org/review-history/13264

Original Research Article

Received: 24th December 2015
Accepted: 28th January 2016
Published: 11th February 2016

Abstract

Functional magnetic resonance imaging (fMRI) has become one of the most widely used techniques in investigating human brain function over the past two decades. However, the analysis of fMRI data is extremely complex due to its difficulties in big data processing, complicated structure of relationship between hemodynamic response and brain activity, and analysis using advanced technology and sophisticated techniques for classification and pattern recognition. Hence, efficient and accurate machine learning models are necessary to interpret fMRI data by incorporating spatial with temporal information. In this paper, we investigate a class of spatial multitask learning models which incorporates spatial information of each task's neighborhood. Simulation and real application results show satisfactory performance from spatial multitask learning algorithms.

Keywords: Spatial regularization; multitask learning; fMRI.

**Corresponding author: E-mail: dhong@mtsu.edu*

2010 Mathematics Subject Classification: 68T10, 68U10, 62H35

1 Introduction

Functional magnetic resonance imaging (fMRI) measures changes in blood oxygenation that are associated to neural activity in a localized brain region. The general purpose of fMRI studies is to detect blood oxygenation level dependent (BOLD) signal in response to a particular stimulus and hence to infer regional neuronal activity by examining BOLD signal contrast in two or more conditions. However, fMRI data has an extremely complicated structure. The subject's 3D volume brain is divided into a grid of volume boxes, or voxels. The BOLD signal is observed at each voxel at each time point, resulting in an enormous amount of data. Hence, powerful models are necessary for accurately detecting neuronal activity.

In addition to high dimensionality and complicated structure, analysis of fMRI data is challenging due to artifacts and variability in the data. The major components of fMRI analysis include, but are not limited to, the processing techniques to deal with these problems, statistical modeling and inference from the data, and applications in medical diagnosis. The initial development of fMRI was driven by cognitive psychology researchers, who are interested in exploring the brain's active responses to external tasks [1]. One of the most important research areas in the fMRI analysis literature has focused on the detection of the active brain regions associated with human activities or diseases; (see [2]-[7] and references therein). This could be modeled from either voxel level [4], [7] or cluster level [2]. The statistical models for active region detection include the general linear models [4] and autoregressive models [7]. In these models, each voxel is associated with a linear regression task. As all the tasks are correlated, considering all voxels together may benefit the modeling and inference. This has driven the use of multitask learning in this area (see [8]-[12]).

Multitask Learning (MTL) refers to a machine learning framework that learns multiple related tasks simultaneously to improve generalization performance. This is especially true when the dataset is small and the performance of single task learning is not as good. The intuition is that learning of one task could benefit from the information of closely related tasks. A more formal explanation is that the learning of related tasks introduces an inductive bias while helping significantly in reducing the variance. MTL has been found successful in the study of many real applications (see [13]-[19] and references therein). A variety of techniques and algorithms have been proposed for MTL for different purposes and different problem domains. The idea of MTL could date back at least to the application of NETtalk to learning both phonemes and their stresses (see [17], [19]), although the concept of MTL was coined much later. In the context of neural network learning, backpropagation has been used to learn multiple related tasks that are drawn from the same domain and share the same hidden units (see [15]-[17]). MTL formulation was also proposed for k-nearest neighborhood, kernel regression, and decision tree in [15].

In recent years, regularization theory was introduced into MTL. Regularized MTL algorithms are usually problem dependent because the penalty term is designed according to prior knowledge of the problem. For instance, by assuming all the tasks share a common component and each task has an additional individual component, the authors in [18] proposed an MTL approach by trading off the size of the common component and the individual components. By adjusting the trade-off parameter, this method allows the data itself to demonstrate how closely the tasks are related and how much improvement can be garnered by learning multiple tasks at the same time. In some applications, not all tasks share the same components, but there is a cluster structure and only tasks belonging to the same cluster share a common component, while the relationship between tasks from different clusters may be weak. This has motivated the structured regularization for MTL (see [20],[21]). Temporal priors were introduced in a study of the progression of Alzheimer's

disease, where each task is the status of patients at a time point, and the temporal relationship arises naturally [22].

In high dimensional data analysis such as fMRI data, feature selection is a natural issue and sparse penalty is required. In order to facilitate sparsity, the adaptive multitask lasso and elastic net were introduced in [8] which utilizes the l_1/l_q mixed matrix norm. In [11], a new procedure called sparse overlapping sets lasso was proposed. In [9], manifold regularization was introduced to multitask feature selection for multi-modality classification in Alzheimer's disease.

In this paper we propose a spatial regularization approach for MTL and apply it to fMRI data analysis. In the problem of active region detection using fMRI data, the tasks (brain voxels) are spatially related. It is natural to code the spatial information into the training process to improve the learning performance. Works on this topic include [23], [24]. However, to the best of our knowledge, the idea of coding spatial information in the regularization theory context is new. The remainder of this paper is organized as follows. The linear regression model for single task is described in Section 2. We develop the spatial regularized multitask learning models and provide algorithms to solving the models in Section 3. In Section 4, the models are tested on both simulated and real fMRI data. We finish with concluding remarks in Section 5.

2 Linear Regression for a Single Task

The most traditional method to solving a linear regression model is the ordinary least squares method (OLS, [25]). In linear regression, a scalar response variable y is assumed to be linearly dependent on a set of p predictors. The data is a sample of n observations subject to noise:

$$y_i = x_i\beta + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{2.1}$$

where $y_i \in \mathbb{R}$, $x_i \in \mathbb{R}^p$ is a row vector, and $\beta \in \mathbb{R}^p$ is an unknown column vector. Denote $Y = (y_1, y_2, \ldots, y_n)^\top \in \mathbb{R}^n$ as a column vector of the response values, $X = [x_1; x_2; \ldots; x_n] \in \mathbb{R}^{n\times p}$ the data matrix, and $E = (\epsilon_1, \ldots, \epsilon_n)^\top$ the error vector. We can rewrite (2.1) as

$$Y = X\beta + E.$$

The OLS estimator minimizes the sum of squared errors (SSE) made by predicting the true response y_i by $x_i\beta$, that is

$$\hat{\beta} = \arg\min \|Y - X\beta\|_2^2 = \arg\min \sum_{i=1}^{n} (y_i - x_i\beta)^2.$$

Here and in the sequel $\|\cdot\|_q$ denotes the q-Euclidean norm for any $1 \le q \le \infty$. If X is of full rank, the OLS estimator can be solved by a linear system

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y. \tag{2.2}$$

The OLS estimator is known as the best linear conditionally unbiased estimator. However, it could be numerically unstable when the matrix $X^\top X$ is singular or has a large conditional number. This is usually the case when $n < p$ or when the predictors are highly correlated. Even when the matrix is well conditioned, it may be beneficial to introduce some bias to facilitate some desired properties (such as sparsity). These considerations have led to the development and application of regularized regression methods such as ridge regression [26], LASSO [27] and Elastic Net [28].

Ridge regression is a method that utilizes Tikhonov regularization of the OLS estimator. It shrinks the coefficients in the estimator by minimizing the penalized SSE where the penalty term $\lambda_2\|\beta\|_2^2$

is determined by a regularization parameter $\lambda_2 > 0$ and the Euclidean 2-norm square of β, that is

$$\hat{\beta}_{Ridge} = \arg\min_{\beta} \left\{ \|Y - X\beta\|_2^2 + \lambda_2 \|\beta\|_2^2 \right\}. \tag{2.3}$$

Ridge regression estimator can also be solved by a linear system, which gives

$$\hat{\beta}_{Ridge} = (X^\top X + \lambda_2 I)^{-1} X^\top Y. \tag{2.4}$$

Here and in the sequel, I denotes an identity matrix (whose dimension is omitted if it is clear from the context or appears as subscript otherwise).

Although ridge regression is numerically stable, the coefficients are never exactly zero even when the corresponding predictors are irrelevant to the response. To implement variable selection, Tibshirani [27] proposed an alternative regularization approach called least absolute shrinkage and selection operator (LASSO). It minimizes the SSE with an ℓ_1 norm penalty.

$$\hat{\beta}_{LASSO} = \arg\min_{\beta} \left\{ \|Y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 \right\}, \tag{2.5}$$

where $\lambda_1 > 0$ is the regularization parameter. Nowadays it is well known that ℓ_1 penalty leads to sparse solution. Therefore, LASSO is advantageous for sparse models because of its facilitation of variable selection.

The elastic net (EN, [28]) also combines shrinkage and variable selection, and in addition encourages grouping of variables: groups of highly correlated variables tend to be selected together, whereas the LASSO would only select one variable of the group. To implement the grouping effect, EN ultimates both ℓ_2 and ℓ_1 penalty.

$$\hat{\beta}_{EN} = \arg\min_{\beta} \left\{ \|Y - X\beta\| l_2^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1 \right\}. \tag{2.6}$$

EN is particularly useful in the "large p small n" setting where the number of predictors is much bigger than the number of observations. Since ℓ_1 norm is not differentiable at 0, the optimization process to solve LASSO and EN is more complicated than ridge regression. The most commonly used solvers include the LARS [29], cyclical coordinate descent [30], etc.

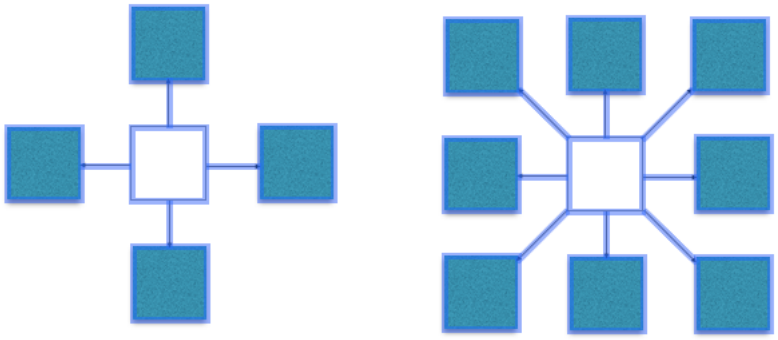
3 Spatial Multitask Learning

In fMRI studies, one of the important problems is detection of a functional region associated with certain brain activities. For each voxel, this can be done by a linear regression model. As the brain contains thousands of voxels, we need to solve thousands of linear regression problems. Of course one can solve these problems voxel by voxel using the single task learning methods. However, this is suboptimal because each functional region contains multiple voxels that are spatially continuous. As a result, if one voxel is active, then its neighbors are very likely to be active as well. Conversely, if one voxel is inactive, its neighbors are unlikely to be active. We expect such spatial information will benefit the learning performance if it is used in the training process. In some applications, Markov random field is used to incorporate the spatial information: image reconstruction [31] and IMS proteomic data analysis in [32], for instance. In this paper, we propose a spatial regularization approach for MTL.

In MTL regression, there are $T \geq 2$ tasks. Assume the t-th task has the data matrix X_t and response vector Y_t which are linked by

$$Y_t = X_t \beta_t + E_t.$$

To code the spatial information, we first define a neighborhood system. It is defined by the user and may be quite data dependent. An example of the 4 or 8 nearest neighborhood system in two dimensional space is shown in Fig. 1.



(a) 4 Neighborhood (b) 8 Neighborhood

Fig. 1. Neighborhood structure for each task

Based on the neighborhood system, we define task similarity coefficient by

$$w_{tk} = \begin{cases} 1, & \text{if task } t \text{ is a neighborhood of task } k; \\ 0, & \text{if task } t \text{ is not a neighborhood of task } k. \end{cases}$$

We assume the neighborhood system is symmetrically defined so that $w_{tk} = w_{kt}$. The penalty term for spatial regularization is defined by

$$\lambda_s \sum_{t,k=1}^{T} w_{tk} \|\beta_t - \beta_k\|_2^2.$$

When λ_s becomes large, it forces the neighboring tasks to become very close, while as λ_s tends to 0, the tasks are treated as independent.

By applying the spatial penalty to ridge regression, LASSO, and EN, we propose three new MTL algorithms. We discuss their formulation and solution in the next three subsections. In the sequel we will denote $B = [\beta_1; \beta_2; \ldots; \beta_T] \in \mathbb{R}^{pT}$ as the column vector composed of all the task coefficients and $W = [w_{tk}]_{t,k=1}^{T}$ as the matrix of the task similarity coefficients.

3.1 Spatial ridge regression

When we learn all T ridge regression problems simultaneously and apply the spatial penalty, the resulted MTL learning algorithm, called spatial ridge regression algorithm, takes the form

$$\hat{B}_{SR} = \arg\min_{B} \left\{ \sum_{t=1}^{T} \|Y_t - X_t\beta_t\|_2^2 + \lambda_2 \sum_{t=1}^{T} \|\beta_t\|_2^2 + \lambda_s \sum_{t,k=1}^{T} \omega_{tk} \|\beta_t - \beta_k\|_2^2 \right\}. \tag{3.1}$$

It is easy to check that

$$\sum_{t=1}^{T} \|Y_t - X_t\beta_t\|_2^2 = B^\top SB - 2V^\top B + \sum_{i=1}^{T} \|Y_t\|_2^2$$

where $S = \text{diag}(X_1^\top X_1, \ldots, X_T^\top X_T)$ and $V = [X_1^\top Y_1; \ldots; X_T^\top Y_T]$.

Let $d_t = \sum_{k=1}^T w_{tk}$, $D_1 = \text{diag}(d_1 I_p, \ldots, d_T I_p)$, and $D_2 = W \otimes I_p$, (where $\otimes$ denotes the kronecker product of two matrices). Define $D = 2D_1 - 2D_2$. Then we have

$$\sum_{t,k=1}^{T} \omega_{tk} \|\beta_t - \beta_k\|_2^2 = B^\top D B.$$

Let $Q = S + \lambda_s D$. The function in (3.1) that needs to be minimized takes the quadratic form

$$B^\top (Q + \lambda_2 I) B - 2V^\top B + \sum_{i=1}^{T} \|Y_t\|_2^2.$$

It can be solved by a linear system:

$$\hat{B} = (Q + \lambda_2 I)^{-1} V.$$

Noticing that $Q = S + \lambda_S D$, where S is a block diagonal matrix, $D = 2(D_1 - D_2)$ with D_1 a diagonal matrix and D_2 a sparse matrix, we see that Q is a sparse matrix. Therefore, this linear system can be solved quickly by using the conjugate gradient method.

3.2 Spatial lasso

Analogously, the spatial LASSO algorithm takes the form

$$\hat{B}_{SL} = \arg\min_B \left\{ \sum_{t=1}^{T} \|Y_t - X_t \beta_t\|_2^2 + \lambda_1 \sum_{t=1}^{T} \|\beta_t\|_1 + \lambda_s \sum_{t,k=1}^{T} \omega_{tk} \|\beta_t - \beta_k\|_2^2 \right\} \tag{3.2}$$

By ignoring the constant term that does not affect the solution, we need to minimize the ℓ_1 penalized quadratic function:

$$B^\top Q B - 2V^\top B + \lambda_1 \|B\|_1. \tag{3.3}$$

One of the most popular methods for solving (3.3) is in the class of iterative shrinkage-thresholding algorithms (ISTA). In 2009, a Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) with the computational simplicity of ISTA but a significantly better global rate of convergence was proposed in [33]. To this end, we first define a soft thresholding operator on $\mathbb{R}^{pT}$

$$\left(prox_{\lambda_1 \alpha}(z)\right)_i = \begin{cases} z_i - \lambda_1 \alpha, & \text{if } z_i > \lambda_1 \alpha \\ 0, & \text{if } |z_i| \leq \lambda_1 \alpha \\ z_i + \lambda_1 \alpha, & \text{if } z_i < -\lambda_1 \alpha \end{cases}$$

with some $\alpha \in (0, \frac{1}{\|Q\|})$. Then the spatial LASSO can be solved by using the following iterating steps based on FISTA [33]:

- $B_k = prox_{\lambda_1 \alpha}\left(g_k - \alpha(Q g_k - V)\right)$;
- $a_{k+1} = \frac{1+\sqrt{1+4a_k^2}}{2}$;
- $g_{k+1} = B_k + (\frac{a_k - 1}{a_{k+1}})(B_k - B^{k-1})$,

after given suitable initial values of B, a, and g.

3.3 Spatial EN

Spatial EN solves the problem

$$\hat{B}_{SEN} = \arg\min_{B} \left\{ \sum_{t=1}^{T} \|Y_t - X_t\beta_t\|_2^2 + \lambda_1 \sum_{t=1}^{T} \|\beta_t\|_1 + \lambda_2 \sum_{t=1}^{T} \|\beta_t\|_2^2 + \lambda_s \sum_{t,k=1}^{T} \omega_{tk} \|\beta_t - \beta_k\|_2^2 \right\} \quad (3.4)$$

By ignoring the constant, we need to minimize

$$B^\top (Q + \lambda_2 I) B - 2V^\top B + \lambda_1 \|B\|_1.$$

The solution to this problem can be obtained by the same procedure as spatial LASSO, except we need to replace Q with $Q + \lambda_2 I$.

4 Simulation and Application to fMRI Data Analysis

In this section we illustrate the power of spatial MTL algorithms by simulation and their application to real fMRI data sets. The performance is compared with the single task learning (STL) method and the regularized MTL algorithm proposed in [18]. All the parameters used in this section are selected by cross validation. For spatial MTL algorithm, there are two or three parameters. An extensive but computationally expensive way to cross validate the parameter is using grid search. To speed up the computation, we adopt a simpler way. We first select the non-spatial parameter (e.g. λ_2 for spatial ridge or λ_1 for spatial LASSO) and fix it. Then the spatial parameter λ_s is selected. Both steps are done by cross validation.

4.1 Simulation data

We first verify the effectiveness of spatial MTL algorithms on simulated data. In this case, since we know the true model, it is easy to compare the performance of different algorithms.

The data are generated as follows. We have designed a 10×10 grid to mimic 100 voxels in a slice of the brain. For each grid, there is an associated input variable and an associated response variable. The array of all input variables mimics the design matrix and the response values mimic the fMRI times series. The response of each grid is computed by the average of input variables associated to the grid itself and its left, right, upper, and lower neighbor grids (if they exist). This gives us 100 tasks in the 100 dimensional input space. For the simulation data, we applied 4-neighborhood structure as shown in Fig. 1.

We generate $n = 100$ samples and run spatial MTL algorithms. This process is repeated 20 times and the learning performance is measured by the mean squared error between the estimated model and true model. We compare our algorithms with the STL learning algorithms and the regularized MTL algorithm proposed in [18]. The MSE and the standard deviation (SD) of these algorithms are reported in Table 4.1. It is clear that the MTL is superior to STL. The tasks are related but do not share a common component. The regularized MTL method in [18] is suboptimal. The spatial regularization helps to improve the performance significantly. Since the true model is rather sparse and there is no grouping effect, spatial LASSO performs the best.

4.2 Real data

Neuroscientists have shown that attention to visual motion can increase the activation of certain cortical areas. Decreased or increased activation of specific brain area would lead to the notion that attention is associated with neuronal activity. This study helps understand the brain functional

Table 4.1. Mean squared error on simulated data

STL Algorithm	MSE (SD)	MTL Algorithm	MSE (SD)
Ridge	0.1592 (0.0152)	Spatial Ridge	0.0489 (0.0014)
LASSO	0.1029 (0.0055)	Spatial LASSO	0.0426 (0.0009)
EN	0.0498 (0.0010)	Spatial EN	0.0445 (0.0009)
		RMTL in [18]	0.0742 (0.0010)

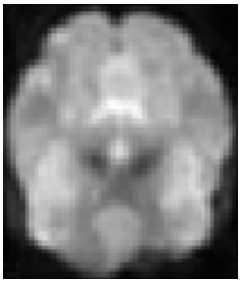
connectivity. In this paper we applied the spatial MTL algorithms to the Attention to Visual motion fMRI data set, which is available on the SPM web site: http://www.fil.ion.ucl.ac.uk/spm/data/attention/. This dataset was collected by Christian Büchel [34] for a study of finding the brain functional connectivity with visual attention. There are four conditions: F, 'fixation', A, 'attention', N, 'no attention', and S, 'stationary' condition. During the 'no attention' and 'attention' conditions, two hundred and fifty white dots were moving radially from a fixation point towards the border of the screen [34]. During the 'fixation' condition, only the fixation mark was visible. The brain is split into 46 slices, with each slice containing 53×63 voxels. For each voxel, data is collected at 360 time points. Thus, the dimension of the whole fMRI dataset is $53 \times 63 \times 46 \times 360$. We obtained the fMRI data for 2 slices of the brain.

For the real fMRI data analysis, we do not know the true model. We adopt the cross validation error to evaluate the performance of different algorithms. Cross validation error is an unbiased estimator of the mean squared prediction error. Small cross validation error usually leads to small prediction error and thus is a relatively reliable metric with which to compare regression algorithms.

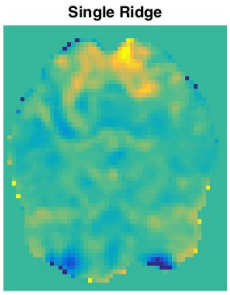
In this real data set, there are 4 contiguous blocked image sets: (0016-0105), (0116-0205), (0316-0405), (0416-0505). Each block has 90 time points, so there are 360 data points in the time series. It is natural to use 4 fold cross validation, considering the special property of the fMRI data. Both 4-neighborhood and 8-neighborhood structures are shown in Fig. 1. Though no significant difference was noticed between these two structures for the cross validation error result in this real data analysis, it's possible to have a better contrast in other applications. The time complexity increase for 8-neighborhood is minimal because only the sparsity of D_2 is slightly increased. All results for real data given here are based on 8-neighborhood structure.

Applying the single task learning and multiple task learning algorithms to the two slices of fMRI data, the cross validation errors are compared in Table 4.2. On one hand we see spatial MTL algorithms slightly improve the result. This indicates that the spatial information does help in the multiple task learning process. On the other hand, we see the improvement is very small. A possible explanation is that, since the design matrix in this study is very simple, the signal is very clear and easy to detect. At the same time, because the noise level is high, the prediction error cannot decrease significantly even if the spatial regularization helps to improve the model estimation.

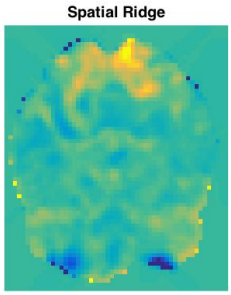
In this paper, we have run 2 slices of the whole brain: slice16, and slice 20. Fig. 2 (a) and Fig. 3 (a) show the functional EPI image for slice 16 and slice 20. Fig. 2 (b)-(g) shows the active area of the brain (slice 16) under attention condition by using the estimated $\hat{\beta}$ learned from both STL and MTL algorithms. Fig. 3 (b)-(g) shows the active area of the brain (slice 20) under attention condition correspondingly. The activity of voxels is indicated by the $\hat{\beta}$ values in the regression model – the larger and more positive the values, the more active the voxels are. With the naked eye, it is hard to see the difference between the six algorithms. But the numerical values of the $\hat{\beta}$ coefficients do have some small differences. Since the spatial MTL algorithms provides slightly better cross validation error, it is reasonable to assume the active area detection by spatial MTL algorithms is more accurate.



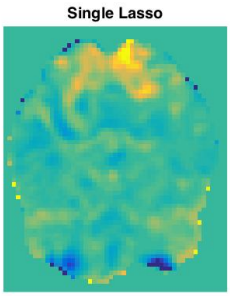
(a) Anatomy Slice 16



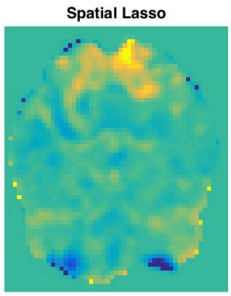
(b) Single Ridge



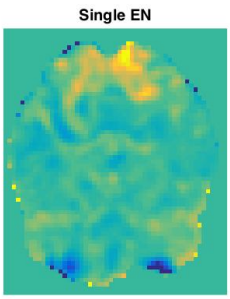
(c) Spatial Ridge



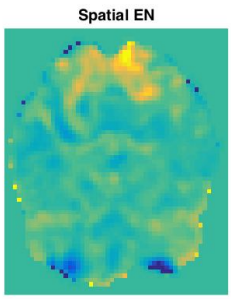
(d) Single Lasso



(e) Spatial Lasso

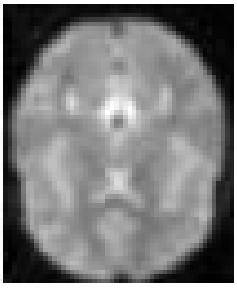


(f) Single EN

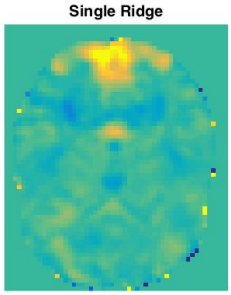


(g) Spatial EN

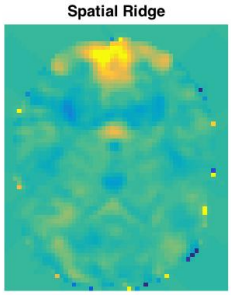
Fig. 2. Attention activation of slice 16



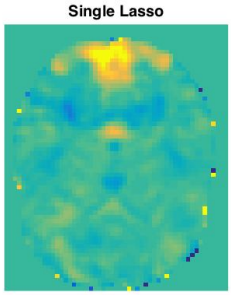
(a) Anatomy Slice 20



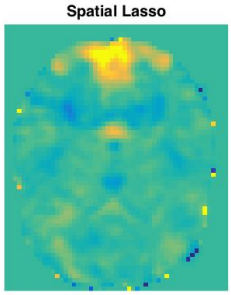
(b) Single Ridge



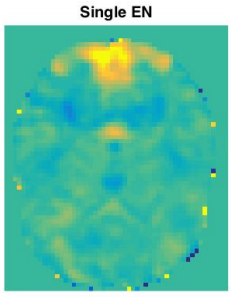
(c) Spatial Ridge



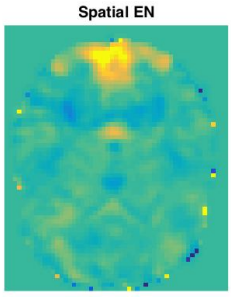
(d) Single Lasso



(e) Spatial Lasso



(f) Single EN



(g) Spatial EN

Fig. 3. Attention activation of slice 20

Table 4.2. The cross validation error of regression algorithms on the fMRI data

	Algorithm	Slice 16	Slice 20
STL	Ridge	43.5825	33.9385
	LASSO	43.2791	33.7631
	EN	43.3004	33.7600
MTL	Spatial Ridge	43.1141	33.8207
	Spatial LASSO	43.0957	33.7068
	Spatial EN	43.1211	33.7066
	RMTL in [18]	43.2320	33.9384

5 Conclusion

Motivated by the fMRI data analysis where spatial information is available between voxels, we proposed a class of spatial multiple task learning algorithms for regression. In these methods, we assume the spatially adjacent regression tasks are close. This leads to a natural spatial regularization approach to code the spatial information by using a user-defined neighborhood system. The spatial regularization multiple task learning is shown to be effective in simulated data and real data analysis.

The spatial regularization approach is not necessarily limited to the fMRI data analysis. Instead, it may potentially be useful in many fields where spatial information is available, for instance, in environment data from multiple geographical sites. For multiple task learning where no spatial information is available, if soft clustering structure or neighborhood systems could be defined, spatial regularization formulation may also be used, although the regularization term does not code spatial information in this situation. Thus, it would be interesting to further investigate the application domains of spatial regularization in future research. In fMRI data analysis, the real challenges are related to the direction and application of the study. In this paper, we only developed an MTL approach for analyzing brain activity with visual attention based on 2-dimensional spatial information. MTL scheme(s) using 3-dimensional spacial information and tasks associated with more general regions of interest (ROIs) of the study could be considered.

Acknowledgements

The authors would like to thank the anonymous referees for their valuable suggestions on the paper. D. Hong's research was partially supported by Beijing Overseas Talents Program. We are also grateful to Drs. Baxter Rogers and Carissa Cascio at Vanderbilt Medical Center for valuable discussions in the study.

Competing Interests

The authors declare that no competing interests exist.

References

[1] Poldrack RA, Mumford JA, Nichols TE. Handbook of functional MRI data analysis. Cambridge University Press; 2011.

[2] Dove A, Pollmann S, Schubert T, Wiggins CJ, von Cramon DY. Prefrontal cortex activation in task switching: An event-related fMRI study. Cognitive Brain Research. 2000;9(1):103-109.

[3] Friston KJ, Holmes A, Poline JB, Price CJ, Frith CD. Detecting activations in PET and fMRI: Levels of inference and power. Neuroimage. 1996;4(3):223-235.

[4] Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RS. Statistical parametric maps in functional imaging: A general linear approach. Human Brain Mapping. 1994;2(4):189-210.

[5] Gallagher HL, Happé F, Brunswick N, Fletcher PC, Frith U, Frith CD. Reading the mind in cartoons and stories: An fMRI study of theory of mind' in verbal and nonverbal tasks. Neuropsychologia. 2000;38(1):11-21.

[6] Kimberg DY, Aguirre GK, D'Esposito M. Modulation of task-related neural activity in task-switching: An fMRI study. Cognitive Brain Research. 2000;10(1):189-196.

[7] Worsley K, Liao C, Aston J, Petre V, Duncan G, Morales F, Evans A. A general statistical analysis for fMRI data. Neuroimage. 2002;15(1):1-15.

[8] Chen X, He J, Lawrence R, Carbonell JG. Adaptive multi-task sparse learning with an application to fMRI study. SIAM International Conference on Data Mining. 2012;212-223.

[9] Jie B, Zhang D, Cheng B, Shen D. Manifold regularized multi-task feature selection for multi-modality classification in alzheimer's disease. Medical Image Computing and Computer-Assisted InterventionMICCAI 2013. Springer. 2013;275-283.

[10] Lee K, Jones GL, Caffo BS, Bassett SS. Spatial Bayesian variable selection models on functional magnetic resonance imaging time-series data. Bayesian Analysis (Online). 2014;9(3):699-732.

[11] Rao N, Cox C, Nowak R, Rogers TT. Sparse overlapping sets lasso for multitask learning and its application to fMRI analysis. Advances in Neural Information Processing Systems. 2013;2202-2210.

[12] Wan J, Zhang Z, Yan J, Li T, Rao BD, Fang S, Kim S, Risacher SL, Saykin AJ, Shen L. Sparse bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in alzheimer's disease. 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2012;940-947:IEEE.

[13] Bakker B, Heskes T. Task clustering and gating for Bayesian multitask Learning. The Journal of Machine Learning Research. 2003;4:83-99.

[14] Bi J, Xiong T, Yu S, Dundar M, Rao RB. An improved multi-task learning approach with applications in medical diagnosis. Machine Learning and Knowledge Discovery in Databases. Springer. 2008;117-132.

[15] Caruana RA. Multitask learning. Machine Learning. 1997;28(1):41-75.

[16] Caruana RA. Multitask connectionist learning. Proceedings of the 1993 Connectionist Models Summer School. 1993;372-79.

[17] Dietterich TG, Hild H, Bakiri G. A comparative study of ID3 and backpropagation for english text-to-speech mapping. Machine Learning. 1995;18(1):51-80.

[18] Evgeniou T, Pontil M. Regularized multi–task learning. Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM. 2004;109-117.

[19] Sejnowski TJ, Rosenberg CR. Nettalk: A parallel network that learns to read aloud. Neurocomputing: Foundations of Research. 1988;661-672.

[20] Agarwal A, Gerber S, Daume H. Learning multiple tasks using manifold Regularization. Advances in Neural Information Processing Systems. 2010;46-54.

[21] Zhou J, Chen J, Ye J. Clustered multi-task learning via alternating structure optimization. Advances in Neural Information Processing Systems. 2011;702-710.

[22] Zhou J, Yuan L, Liu J, Ye J. A multi-task learning formulation for predicting disease progression. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM. 2011;814-822.

[23] Norman KA, Polyn SM, Detre GJ, Haxby JV. Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. Trends in Cognitive Sciences. 2006;10(9):424-430.

[24] Smith M, Fahrmeir L. Spatial bayesian variable selection with application to functional magnetic resonance imaging. Journal of the American Statistical Association. 2007;102(478):417-431.

[25] Whittle P. Prediction and regulation by linear least-square methods. Applied Mathematics Series, English Univ. Press; 1963.

[26] Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970;12(1)55-67.

[27] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological). 1996;267-288.

[28] Zou H, Hastie T. Regularization and variable selection via the elastic Net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005;67(2):301-320.

[29] Efron B, Hastie T, Johnstone I, Tibshirani R. et al. Least angle Regression. The Annals of Statistics. 2004;32(2):407-499.

[30] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software. 2010;33(1):1-22.

[31] Aykroyd RG, Zimeras S. Inhomogeneous prior models for image Reconstruction. Journal of the American Statistical Association. 1999;94(447):934-946.

[32] Xiong L, Hong D, Incorporating spatial information in IMS data analysis to optimize classification accuracy using Markov Random Field and MCMC method. Statistical Analysis of Spectrometry Based Proteomics and Metabolomics Data. Frontiers in Probability and Statistics series. Springer, New York (to appear).

[33] Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences. 2009;2(1):183-202.

[34] Büchel C, Friston K. Modulation of connectivity in visual pathways by attention: Cortical interactions evaluated with structural equation modelling and fMRI. Cerebral Cortex. 1997;7(8):768-778.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://sciencedomain.org/review-history/13264