Stability-Weighted Matrix Completion of Incomplete Multi-modal Data for Disease Diagnosis

Kim-Han Thung, Ehsan Adeli, Pew-Thian Yap, and Dinggang Shen(✉)

Department of Radiology and BRIC, University of North Carolina, Chapel Hill, USA
dgshen@med.unc.edu

Abstract. Effective utilization of heterogeneous multi-modal data for Alzheimer's Disease (AD) diagnosis and prognosis has always been hampered by incomplete data. One method to deal with this is low-rank matrix completion (LRMC), which simultaneous imputes missing data features and target values of interest. Although LRMC yields reasonable results, it implicitly weights features from all the modalities equally, ignoring the differences in discriminative power of features from different modalities. In this paper, we propose *stability-weighted LRMC* (swLRMC), an LRMC improvement that weights features and modalities according to their *importance* and *reliability*. We introduce a method, called *stability weighting*, to utilize subsampling techniques and outcomes from a range of hyper-parameters of sparse feature learning to obtain a stable set of weights. Incorporating these weights into LRMC, swLRMC can better account for differences in features and modalities for improving diagnosis. Experimental results confirm that the proposed method outperforms the conventional LRMC, feature-selection based LRMC, and other state-of-the-art methods.

1 Introduction

Effective methods to jointly utilize heterogeneous multi-modal and longitudinal data for Alzheimer's Disease (AD) diagnosis and prognosis often need to overcome the problem of incomplete data. Data are incomplete due to various reasons, including cost concerns, poor data quality, and subject dropouts. Most studies deal with this issue by simply discarding incomplete samples, hence significantly reducing the sample size of the study.

A more effective approach to deal with missing data is by imputing them using k-nearest neighbor, expectation maximization, low-rank matrix completion (LRMC) [2], or other methods [8,13]. However, these methods perform well only if a small portion, but not a whole chunk, of the data is missing. To avoid propagation of the imputation error to the diagnosis stage, Goldberg et al. [3] propose to simultaneously impute the missing data and the diagnostic labels

This work was supported in part by NIH grants (NS093842, EB006733, EB008374, EB009634, AG041721, and MH100217).

S. Ourselin et al. (Eds.): MICCAI 2016, Part II, LNCS 9901, pp. 88–96, 2016.
DOI: 10.1007/978-3-319-46723-8_11

using LRMC. This approach, along with other variants [9], however, inherently assumes that the features are equally important. This might not be the case especially when the data are multi-modal and heterogeneous, with some features being more discriminative than others [4,6,10]. For example, in our study involving magnetic resonance imaging (MRI) data, positron emission tomography (PET) data, and cognitive assessment data, we found that clinical scores, though fewer in dimension, are more discriminative than PET data, and within the PET data, only few features are related to the progression of mild cognitive impairment (MCI), a prodromal stage of AD. To address this issue, the method in [9] shrinks the data via selection of the most discriminant features and samples using sparse learning methods and then applies LRMC. Although effective, this approach still neglects the disproportionate discriminative power of different features, when employing LRMC.

In this paper, we explicitly consider the differential discriminative power of features and modalities in our formulation of LRMC by weighting them using a procedure called *stability weighting*. We first explain feature weighting, where each feature is assigned a weight according to its feature-target relationship, i.e., more discriminative features are assigned higher weights, and vice versa. For instance, in sparse feature weighting [14], the feature-target regression coefficients are used as feature weights. Feature weighting like [14] always involves tuning one (or multiple) regularizing hyper-parameter(s), which is (are) normally determined via cross-validation. However, as pointed out in [7], it is difficult to choose a single set of hyper-parameter that is able to retain all the discriminative features while removing the noisy features.

Stability weighting avoids the difficulties of proper regularization [7] in feature weighting by going beyond one set of hyper-parameters. It utilizes multiple sets of hyper-parameters and subsampled data to compute a set of aggregated weights for the features. Using random subsampling and aggregation, *stability weighting* estimates the weights based on the "*stability*" of the contribution of a feature. More specifically, we perform a series of logistic regression tasks, involving different hyper-parameters and different data subsets, for each modality. Regression coefficients corresponding to the hyper-parameters that yield higher prediction performance are then aggregated as feature weights. We use the term "*importance*" and "*reliability*" to denote how good a feature and a modality are in the prediction task, respectively. In the context of *stability weighting*, feature importance is quantified by the aggregated weight values while modality reliability is quantified by the performance measures. We then incorporate the feature importance and modality reliability into LRMC, giving us stability-weighted LRMC (swLRMC) for greater prediction accuracy.

The contribution of our work is two-fold. (1) We propose a *stability weighting* procedure to quantify the *importance* of features and the *reliability* of modalities. (2) We incorporate this information into the formulation of the proposed swLRMC for more robust and accurate prediction using incomplete heterogeneous multi-modal data.

2 Materials, Preprocessing and Feature Extraction

In this study, we focus on MCI and use the baseline multi-modal data from ADNI dataset[1], including MRI, PET, and clinical scores (i.e., Mini-Mental State Exam (MMSE), Clinical Dementia Rating (CDR-global, CDR-SOB), and Alzheimer's Disease Assessment Scale (ADAS-11, ADAS-13)). Only MRI data is complete, the other two modalities are incomplete. MCI subjects who progressed to AD within 48 month are retrospectively labeled as pMCI, whereas those who remained stable are labeled as sMCI. MCI subjects who progressed to AD after the 48th month are excluded from this study. Table 1 shows the demographic information of the subjects involved.

Table 1. Demographic information of MCI subjects involved in this study. (Edu.: Education)

	# Subjects	Gender (M/F)	Age (years)	Edu. (years)
pMCI	169	103/66	74.6 ± 6.7	15.8 ± 2.8
sMCI	61	45/16	73.9 ± 7.7	14.9 ± 3.4
Total	230	148/82	-	-

We use region-of-interest (ROI)-based features from the MRI and PET images in this study. The processing steps involved are described as follows. Each MRI image was AC-PC aligned using MIPAV[2], corrected for intensity inhomogeneity using the N3 algorithm, skull stripped, tissue segmented, and registered using a template to obtain subject-labeled image with 93 ROIs [11]. Gray matter (GM) volumes, normalized by the total intracranial volume, were extracted from 93 ROIs as features [9,10]. We also linearly aligned each PET image to its corresponding MRI image, and used the mean intensity values of each ROI as PET features.

3 Method

Figure 1 gives an overview of the proposed swLRMC framework. The main difference between swLRMC and LRMC is the introduction of a stability weight matrix $\mathbf{W}$, which is computed via stability weighting. $\mathbf{W}$ is then used in swLRMC to simultaneously impute the missing feature values and the unknown target values (i.e., diagnostic labels and conversion times). We provide the details of each step in the following.

[1] http://adni.loni.ucla.edu.
[2] http://mipav.cit.nih.gov.

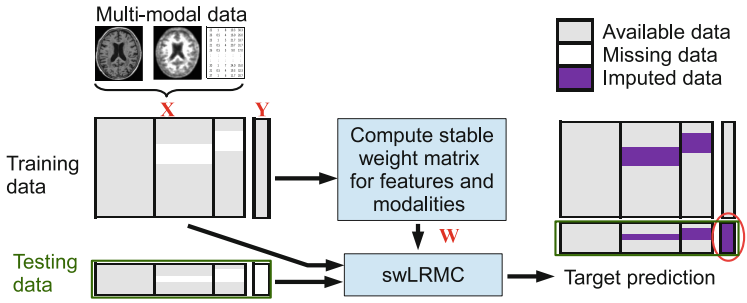

Fig. 1. Stability-weighted low-rank matrix completion (swLRMC).

3.1 Notation

Let $\mathbf{X} = [\mathbf{X}^{(1)} \cdots \mathbf{X}^{(m)}] \in \mathbb{R}^{N \times d}$ denotes the feature matrix of N samples. The features from m modalities (i.e., MRI, PET and clinical scores (Cli)) are concatenated to give d features per sample. Since, for each sample, not all the modalities are available, $\mathbf{X}$ is incomplete with some missing values. We use $\mathbf{Y} = [\mathbf{y}_1 \cdots \mathbf{y}_t] \in \mathbb{R}^{N \times t}$ to denote the corresponding target matrix with two targets ($t = 2$), i.e., the diagnostic labels (1 for pMCI and -1 for sMCI), and the conversion time (i.e., number of months prior to AD conversion). The conversion time of an sMCI subject should ideally be set to infinity. But for feasibility, we set the conversion time to a large value computed as 12 months plus the maximum conversion time over all pMCI samples. Throughout the paper, we use bold upper-case to denote matrices and bold lower-case to denote column vectors.

3.2 Low-Rank Matrix Completion (LRMC)

Prediction using LRMC is based on several assumptions. First, it assumes linear relationship between $\mathbf{X}$ and $\mathbf{Y}$, i.e., $\mathbf{Y} = [\mathbf{X}\ \mathbf{1}] * \boldsymbol{\beta}$, where $\mathbf{1}$ is a column vector of all 1's, and $\boldsymbol{\beta}$ is the coefficient matrix. Second, it assumes $\mathbf{X}$ is low-rank, i.e., rows (columns) of $\mathbf{X}$ could be represented by other rows (columns). It can be inferred then that the concatenated matrix $\mathbf{M} = [\mathbf{X}\ \mathbf{1}\ \mathbf{Y}]$ is also low-rank [3]. Hence, it follows that LRMC can be applied on $\mathbf{M}$ to impute the missing feature values and the unknown output targets simultaneously, without knowing $\boldsymbol{\beta}$. This is achieved by solving $\min_{\mathbf{Z}}\{\|\mathbf{Z}\|_* \mid \mathcal{P}_{\Omega}(\mathbf{M}) = \mathcal{P}_{\Omega}(\mathbf{Z})\}$ [2], where $\mathbf{Z}$ is the completed version of $\mathbf{M}$, Ω is the set of indices of known values in $\mathbf{M}$, $\mathcal{P}$ is the projection operator, and $\|\cdot\|_*$ is the nuclear norm (i.e., sum of singular values), which is used as a convex surrogate for matrix rank. In the presence of noise, and using different loss functions for $\mathbf{X}$ and $\mathbf{Y}$, this problem is reformulated as [3]:

$$\min_{\mathbf{Z}} \mu\|\mathbf{Z}\|_* + \frac{1}{|\Omega_X|}\|\mathcal{P}_{\Omega_X}(\mathbf{Z} - \mathbf{M})\|_F^2 + \sum_i^t \frac{\lambda_i}{|\Omega_{y_i}|}\mathcal{L}_i(\mathcal{P}_{\Omega_{y_i}}(\mathbf{Z}), \mathcal{P}_{\Omega_{y_i}}(\mathbf{M})), \quad (1)$$

where $\mathcal{L}_i(\cdot,\cdot)$ is the loss function for the i-th column of $\mathbf{Y}$. Since the first target is the diagnostic label (binary) and the second target is the conversion time

(continuous), we use logistic loss ($\mathcal{L}_1(\mathbf{u},\mathbf{v}) = \sum_j \log(1+\exp(-u_j v_j))$) and mean square loss ($\mathcal{L}_2(\mathbf{u},\mathbf{v}) = \sum_j 1/2(u_j - v_j)^2$) functions for the first and second targets, respectively. Ω_X and Ω_{y_i} are the index sets of the known feature values and target outputs in $\mathbf{M}$, respectively. $|\cdot|$ denotes the cardinality of a set, and $\|\cdot\|_F$ is the Frobenius norm. Parameters μ and λ_i are the tuning hyper-parameters that control the effect of each term. The features fitting term (second term) in (1) shows that the conventional LRMC treats all the features equally, without considering the importance of each feature in relation to the target(s). In the following, we propose to modulate this fitting term according to the feature-target relationship.

3.3 Stability-Weighted LRMC (swLRMC)

Due to missing feature values for some modalities, conventional feature selection methods cannot be applied to the whole data. Thus, we compute the weights separately for each modality. Denoting the importance of features in the j-th modality as vector $\mathbf{w}^{(j)}$ and the reliability of the j-th modality as $s^{(j)}$, we reformulate the second term of (1) as follows:

$$\frac{1}{|\Omega_X|}\sum_{j=1}^{m} s^{(j)} \|\mathcal{P}_{\Omega_{X^{(j)}}}(\mathrm{diag}(\mathbf{w}^{(j)})(\mathbf{Z}_{\mathbf{X}^{(j)}} - \mathbf{X}^{(j)}))\|_F^2, \tag{2}$$

where $\mathbf{Z}_{\mathbf{X}^{(j)}}$ is the j-th modality feature part of $\mathbf{Z}$, $\Omega_{X^{(j)}}$ is the known value indices of $\mathbf{X}^{(j)}$, and $\mathrm{diag}(\cdot)$ is the diagonal operator. Each element in $\mathbf{w}^{(j)}$ quantifies the importance of the corresponding feature in $\mathbf{X}^{(j)}$ in terms of discriminative power. More important features are given higher values, so that they are less affected by the smoothing effect of the low rank constraint (first term of (1)), and play more dominant roles in the optimization process. In the following, we explain how $\mathbf{w}^{(j)}$ and $s^{(j)}$ are obtained via stability weighting.

Stability Weighting: Stability weighting uses data subsampling and sparse feature weighting with multiple hyper-parameters (similar to stability selection [7]), to improve robustness in feature weighting. Any feature weighting method can be used for stability weighting. In this paper, we choose logistic elastic net [14]. First, we use elastic net to compute a weight vector for each modality:

$$\min_{\boldsymbol{\beta}^{(i)}} \|\log(1+\exp(-\mathbf{y}_1 \odot (\mathbf{X}^{(i)}\boldsymbol{\beta}^{(i)})))\|_1 + \alpha_1\|\boldsymbol{\beta}^{(i)}\|_1 + \alpha_2\|\boldsymbol{\beta}^{(i)}\|_2^2 \ , \tag{3}$$

where $\mathbf{y}_1$ is a column vector of diagnostic labels, $\odot$ is element-wise multiplication, α_1 and α_2 are the tuning hyper-parameters, and $\boldsymbol{\beta}^{(i)}$ is a sparse coefficient vector. The magnitude of each element in $\boldsymbol{\beta}^{(i)}$ can be seen as an indicator of the importance of the corresponding feature in $\mathbf{X}^{(i)}$. Note that, in this process one needs to determine the hyper-parameter $\boldsymbol{\alpha} = [\alpha_1\ \alpha_2]$, which is normally done through cross-validation. However, instead of limiting ourselves to just one hyper-parameter and one set of data, we use a range of hyper-parameters and

the subsamples of training data to determine the feature weights. More specifically, we solve (3) using a range of $\boldsymbol{\alpha}$ values using 5-fold cross-validation on the training data with 10 repetitions. For each $\boldsymbol{\alpha}$, we therefore have 50 versions of $\boldsymbol{\beta}^{(i)}$, and one average F-score[3]. We choose three $\boldsymbol{\alpha}$ values that give us highest F-score values, and compute the weight vector for the i-th modality as $\mathbf{w}^{(i)} = \bar{\boldsymbol{\beta}}^{(i)}/\max(\bar{\boldsymbol{\beta}}^{(i)}) + \epsilon$, where ϵ is a small constant and $\bar{\boldsymbol{\beta}}^{(i)}$ is the mean absolute vector of all $(50 \times 3 = 150)$ $\boldsymbol{\beta}^{(i)}$'s that correspond to the $\boldsymbol{\alpha}$'s with the highest average F-scores. We then use the best average F-score to quantify the reliability of using $\mathbf{X}^{(i)}$ in predicting target $\mathbf{y}_1$, which is denoted as $s^{(i)}$. Note that $s^{(i)}$ and $\mathbf{w}^{(i)}$ in (2) can be combined into a single weight matrix as $\mathbf{W} = \text{diag}([s^{(1)}\mathbf{w}^{(1)}; \cdots ; s^{(m)}\mathbf{w}^{(m)}])$. Finally, the compact equivalent form of swLRMC is given as

$$\min_{\mathbf{Z}} \mu\|\mathbf{Z}\|_* + \frac{1}{|\Omega_X|}\|\mathcal{P}_{\Omega_X}(\mathbf{W}(\mathbf{Z}-\mathbf{M}))\|_F^2 + \sum_i^t \frac{\lambda_i}{|\Omega_{y_i}|}\mathcal{L}_i(\mathcal{P}_{\Omega_{y_i}}(\mathbf{Z}), \mathcal{P}_{\Omega_{y_i}}(\mathbf{M})). \quad (4)$$

Optimization: Equation (4) can be solved to obtain matrix $\mathbf{Z}$ by iterating through l in the two steps below until convergence [3]:

1. Gradient Step: $\mathbf{G}^l = \mathbf{Z}^l - \tau g(\mathbf{Z}^l)$, where $\mathbf{G}$ is a intermediate matrix, τ is the step size, and $g(\mathbf{Z}^l)$ is the matrix gradient defined as

$$g(Z_{ij}) = \begin{cases} \frac{\lambda_1}{|\Omega_{y_1}|}\frac{-M_{ij}}{1+\exp(M_{ij}Z_{ij})}, & (i,j) \in \Omega_{y_1} \\ \frac{W_{jj}}{|\Omega_X|}(M_{ij} - Z_{ij}), & (i,j) \in \Omega_X \\ \frac{\lambda_2}{|\Omega_{y_2}|}(M_{ij} - Z_{ij}), & (i,j) \in \Omega_{y_2} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

2. Shrinkage Step [2]: $\mathbf{Z}^{l+1} = S_{\tau\mu}(\mathbf{G}^l) = \mathbf{P}\left(\max(\boldsymbol{\Lambda} - \tau\mu, 0)\right)\mathbf{Q}^T$, where $S(\cdot)$ is the matrix shrinkage operator, $\mathbf{P}\boldsymbol{\Lambda}\mathbf{Q}^T$ is the SVD of $\mathbf{G}^l$, and $\max(\cdot)$ is the element-wise maximum operator.

4 Results and Discussions

We evaluated the proposed method, swLRMC, using multi-modal data for the ADNI database. We evaluated two versions of swLRMC: (1) swLRMC on the original feature matrix without removing any features, and (2) swLRMC on feature-selected matrix (fs-swLRMC) by discarding the features that were selected less than 50 % of the time in stability selection. We compared our methods with two baseline LRMC methods: (1) LRMC without feature selection, and (2) LRMC with sparse feature selection (fs-LRMC). The hyper-parameters $\mu, \lambda_1, \lambda_2$ for all methods were selected automatically using Bayesian hyper-parameter optimization [1] in the ranges of $\{10^{-6}, \cdots, 10^{-2}\}$, $\{10^{-4}, \cdots, 10^{-1}\}$,

[3] We use F-score as performance measure as our dataset is unbalanced.

and $\{10^{-4}, \cdots, 10^{-1}\}$, respectively. For sparse feature selection, we used the SLEP package[4] and performed 5-fold cross validation on the training data to select the best hyper-parameter.

Since the dataset we used was unbalanced, we used the F-score and the area under the ROC curve (AUC) to measure the classification performance, and correlation coefficient (CC) to measure the accuracy of conversion time prediction. All the results reported are the averages of 10 repetitions of 10-fold cross validation. The results shown in Fig. 2 indicate that swLRMC (blue bars) performs consistently better than baseline LRMC (orange bar), for all the performance metrics and modality combinations. It is worth noting that swLRMC and fs-swLRMC seem to be performing almost equally well, but fs-swLRMC is faster in computation, due to its smaller matrix size during imputation. It is also interesting to see that swLRMC performs better than fs-LRMC in terms of F-score and CC values, indicating that penalizing less discriminative features is better than removing them. Another encouraging observation is that swLRMC is less sensitive to "noisy" features in the multi-modal data. This can be seen in MRI+PET combination, where performance of LRMC drops, compared to the case where only MRI is used, whereas the performance of swLRMC improves. A similar pattern can be observed for MRI+PET+Cli, where LRMC performs poorer than MRI+Cli case, whereas swLRMC maintains its performance.

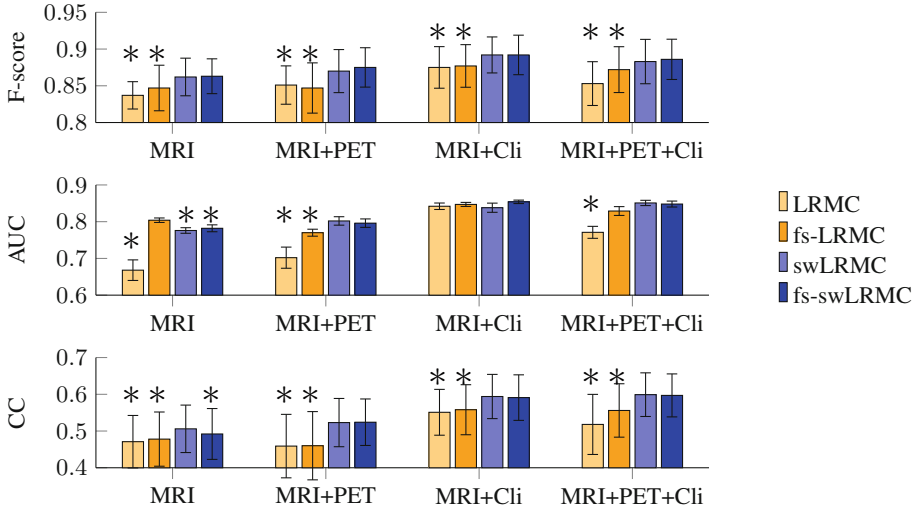

Fig. 2. Comparisons between the baseline LRMC and the proposed swLRMC methods using multi-modal data. The first two plots: pMCI/sMCI classification results (first target), the last plot: conversion time prediction results (second target). Error bars: standard deviations, *: statistically significant.

[4] http://www.yelab.net/software/SLEP/.

Table 2. Comparison with [12] and [5]. **Bold**: Best results, *: Statistically significant.

Data	F-score			AUC			CC		
	swLRMC	[12]	[5]	swLRMC	[12]	[5]	swLRMC	[12]	[5]
MRI	**0.862**	0.853*	0.777*	**0.776**	0.758*	0.755*	**0.506**	0.427*	0.464*
MRI+PET	**0.870**	0.853*	0.798*	0.802	0.786*	**0.806**	**0.523**	0.469*	0.492*
MRI+Cli	**0.892**	0.853*	0.809*	0.838	0.829*	**0.841**	**0.594**	0.567*	0.560*
MRI+PET+Cli	**0.883**	0.859*	0.805*	**0.851**	0.827*	0.842	**0.599**	0.568*	0.553*

We also show in Table 2 a comparison of swLRMC with two methods that works with incomplete dataset: (1) incomplete data multi-task learning [12], and (2) Ingalhalikar's ensemble method [5]. We selected the best hyper-parameters for these methods using 5-fold cross validation. We used logistic loss and mean-square loss function for classification and regression, respectively, for [12]. The highest score for each category is highlighted in bold. The results show that swLRMC outperforms both methods in F-score and CC for all the combinations of modalities. In terms of AUC, swLRMC gives comparable performance.

To test the significance of the results, we perform paired t-test between the best result and the other results in each category. The outcomes of the paired t-test are included in Fig. 2 and Table 2, where statistically significantly difference results in comparison with the best method, at 95 % confidence level, are marked with asterisks. The results show that the improvement of the proposed method is statistically significant in terms of F-score and CC values, in all the combinations of multi-modal data.

5 Conclusion

We have demonstrated that the proposed method, swLRMC, which explicitly considers feature importance and modality reliability using stability weighting procedure, outperforms conventional LRMC, fs-LRMC, and two state-of-the-art methods that were designed for incomplete multi-modal data. Experimental results show that our proposed method is effective when dealing with incomplete multi-modal data, where not all the feature values are equally important.

References

1. Bergstra, J.S., et al.: Algorithms for hyper-parameter optimization. In: Proceedings of Advances in Neural Information Processing Systems, pp. 2546–2554 (2011)
2. Candès, E.J., et al.: Exact matrix completion via convex optimization. Found. Comput. Math. **9**(6), 717–772 (2009)
3. Goldberg, A., et al.: Transduction with matrix completion: three birds with one stone. In: Proceedings of Advances in Neural Information Processing Systems, vol. 23, pp. 757–765 (2010)

4. Huang, L., Gao, Y., Jin, Y., Thung, K.-H., Shen, D.: Soft-split sparse regression based random forest for predicting future clinical scores of Alzheimer's disease. In: Zhou, L., Wang, L., Wang, Q., Shi, Y. (eds.) MLMI 2015. LNCS, vol. 9352, pp. 246–254. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24888-2_30
5. Ingalhalikar, M., Parker, W.A., Bloy, L., Roberts, T.P.L., Verma, R.: Using multiparametric data with missing features for learning patterns of pathology. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012. LNCS, vol. 7512, pp. 468–475. Springer, Heidelberg (2012)
6. Jin, Y., et al.: Identification of infants at high-risk for autism spectrum disorder using multiparameter multiscale white matter connectivity networks. Hum. Brain Mapp. **36**(12), 4880–4896 (2015)
7. Meinshausen, N., et al.: Stability selection. J. Roy. Stat. Soc. Ser. B (Stat. Methodol.) **72**(4), 417–473 (2010)
8. Qin, Y., et al.: Semi-parametric optimization for missing data imputation. Appl. Intell. **27**(1), 79–88 (2007)
9. Thung, K.H., et al.: Neurodegenerative disease diagnosis using incomplete multimodality data via matrix shrinkage and completion. Neuroimage **91**, 386–400 (2014)
10. Thung, K.H., et al.: Identification of progressive mild cognitive impairment patients using incomplete longitudinal MRI scans. Brain Struct. Funct. pp. 1–17 (2015)
11. Wang, Y., Nie, J., Yap, P.-T., Shi, F., Guo, L., Shen, D.: Robust deformable-surface-based skull-stripping for large-scale studies. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011. LNCS, vol. 6893, pp. 635–642. Springer, Heidelberg (2011). doi:10.1007/978-3-642-23626-6_78
12. Yuan, L., et al.: Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. NeuroImage **61**(3), 622–632 (2012)
13. Zhu, X., et al.: Missing value estimation for mixed-attribute data sets. IEEE Trans. Knowl. Data Eng. **23**(1), 110–121 (2011)
14. Zou, H., et al.: Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. Ser. B **67**(2), 301–320 (2005)