AN EFFICIENT AUGMENTED LAGRANGIAN ALGORITHM FOR GRAPH REGULARIZED SPARSE CODING IN CLUSTERING

Qiegen Liu[1,2], *Leslie Ying*[3] *and Dong Liang*[1,*]

[1]Paul C. Lauterbur Research Centre for Biomedical Imaging, Shenzhen Institutes of Advanced Technology, China
[2]Department of Electronic Information Engineering, Nanchang University, China
[3]Department of Biomedical Engineering and Department of Electrical Engineering, University at Buffalo, The State University of New York, Buffalo, New York 14260, USA

ABSTRACT

The combination of sparse coding and manifold learning has received much attention recently. However, the computational complexity of the resulting optimization problem hinders its practical application. In this paper, an augmented Lagrangian method is proposed to address this issue, which first transforms the unconstrained problem to an equivalent constrained problem and then an alternating direction method is used to iteratively solve the subproblems. Experimental results validate the effectiveness of the propose algorithm.

Index Terms— Image clustering, augmented Lagrangian, alternating direction method, graph regularized sparse coding.

1. INTRODUCTION

Sparse coding/representation (SC), as a new way to encode signal using only a few active coefficients under overcomplete and adaptive dictionary, has drawn a lot of attention recently in signal processing and machine learning [1, 2].

Given an input data matrix $X \in R^{N\times M}$, with each column corresponding to a signal vector, the matrix factorization methods aim to simultaneously construct two matrices $B \in R^{N\times J}$ and $S \in R^{J\times M}$ such that $X \approx BS$. In this approximate decomposition model, each column of B is a basis vector corresponding to a certain semantic concept and the whole set is named as the dictionary. Meanwhile, each column of S stands for the representation weights of the corresponding signal in this dictionary. Compared with other developed approaches such as sparse PCA [3] and sparse NMF [4], sparse coding exhibits several advantages. First, the target dictionary is usually posed a few constraints such as norm-constraint to enable meaningful result, thus producing more freedom and flexibility on capturing the high-level semantics. Second, sparse coding adds the sparse constraint on S such that the sparse representations allow quick retrieval, thus benefitting image indexing.

Several variants of the sparse coding technique have been proposed recently by adding additional constraints [5, 6]. Among them, the consistence between similar local features is most typically considered. Mairal et al. developed simultaneous sparse coding by adding a group-sparsity regularizer to integrate the self-similarity constraint into dictionary learning for image restoration [5]. Zhang et al. presented a graph regularized sparse coding (GraphSC) method which incorporates a k-nearest neighbor graph into the sparse coding objective function as a regularizer [6]. Superior performance was generally attained.

Although sparse coding has attracted much attention with its wide applications in various settings, the issues of computational complexity and local optimality are yet to be addressed due to the non-convexity and high non-linearity nature of the problem [7], which limits its practical application. Therefore, developing an efficient and robust algorithm is still highly desirable. In this paper, an augmented Lagrangian method is proposed to solve the graph regularized sparse coding problem.

1.1. Problem Formulation

In GraphSC [6], the manifold assumption is adopted. It states that if two data points x_i, x_j are close in the intrinsic geometry of the data distribution, then their representations s_i and s_j in the new dictionary are also close to each other. Specifically, a nearest neighbor graph G with M vertices is constructed, where each vertex represents a data point in X, and W be the weight matrix of G. If x_i is among the k-nearest neighbors of x_j or vice versa, $W_{ij} = 1$, otherwise, $W_{ij} = 0$. Additionally, the degree of x_i is defined as

*Correspondence to: Dong Liang (dong.liang@siat.ac.cn). This work was supported in part by the National Natural Science Foundation of China under Nos. 61102043, 81120108012 and the Basic Research Program of Shenzhen JC201104220219A.

$d_i = \sum_{j=1}^{M} W_{ij}$ and $D = diag(d_1, \cdots, d_M)$. A reasonable criterion for properly mapping the weighted graph G to sparse coefficients S is to minimize the following function [6]:

$$\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}(\mathrm{s}_i - \mathrm{s}_j)^2 W_{ij} = Tr(SLS^T) \tag{1}$$

where L=D-W is the Laplacian matrix. Overall, the objective function of GraphSC consists of three terms: the empirical loss function, the Laplacian regularizer, and L_1-based sparse penalty function as following:

$$\begin{aligned} &\min_{B,S} \frac{\lambda}{2}\left\|X - BS\right\|_F^2 + \alpha Tr(SLS^T) + \sum_{i=1}^{M}\left\|\mathrm{s}_i\right\|_1 \\ &s.t. \left\|\mathrm{b}_j\right\|^2 \le 1, \forall j = 1, \cdots, J \end{aligned} \tag{2}$$

where b_j is the j-th vector from matrix B. Particularly, Eq.(2) degrades to the basic sparse coding when $\alpha = 0$.

1.2. Previous Work on Solving (2)

Traditional approaches solve the nonlinear minimization problem (2) with $\alpha = 0$ by iterating the two-step procedure. This procedure consists of a sparse-coding step where the sparse coefficients S are estimated with the dictionary fixed and a dictionary-updating step where B is computed based on the current sparse representation.

At the sparse coding step, the commonly used techniques are the Matching Pursuit (MP) and the Basis Pursuit (BP). At the dictionary updating step, the constrained optimization problem can be solved by using several algorithms, such as the Maximum A Posteriori (MAP) method with iterative projection [8], dual version derived from its Lagrangian (Dual-Feature) [6], and K-Singular Value Decomposition (K-SVD) [1]. Recently, Liu et al. [9] introduced the augmented Lagrangian and alternating direction method to dictionary learning (AL-DL) for image denoising. The numerical comparisons show its faster convergence and better recovery than the predecessors.

For the more complicated graph regularized sparse coding with $\alpha \neq 0$ (e.g. GraphSC), the Dual-Feature method was usually extended to solve (2) [6]. However, new algorithms are still needed with a lower computational complexity and better convergence.

1.3. Proposed Approach

In this paper, we extend the augmented Lagrangian (AL) technique to deal with the general problem (2) for clustering, and illustrate its superiority on computational complexity and convergence behavior over existing algorithms. Specifically, the original unconstrained problem (2) is transformed into a constrained problem, by the application of a variable splitting operation; the resulting problem is then handled using a variant of the alternating direction method (ADM). The proposed method can be seen as an extension of the AL-DL for solving the problem of combing the Graph Regularization and Sparse Coding, and is thus named GRSC-AL.

Compared with the traditional sparse coding and GraphSC, GRSC-AL is experimentally shown to efficiently solve the image clustering problem with formulation (2), in terms of clustering accuracy and computation time.

The outline of the paper is as follows. Section 2 briefly reviews the AL and ADM. Section 3 elaborates on the derivation of GRSC-AL. Section 4 reports the results of numerical experiments, and Section 5 concludes the paper with a few remarks and points to future work.

2. VARIABLE SPLITTING AND ADM

Although AL method is a commonly studied optimization algorithm for solving the constrained problems in mathematical programming community [10], it is enjoying a re-popularization recently due to the work of Osher et al. [12] and has been used in various applications of signal/image processing [9-14]. The AL related methods usually employ the operator splitting first to transform the original unconstrained minimization problem to the equivalent constrained problem, and then alternating-minimization strategy is used to iteratively find solutions of the subproblems. Generally speaking, the AL scheme aims to solve the following problem:

$$\min_{Z,B,S} E(Z) \quad s.t. \quad Z - BS = 0 \tag{3}$$

Problem (3) can be solved via the standard augmented Lagrangian method. Specifically, starting from $Y^0 = 0$, it solves

$$\begin{aligned} (Z^{k+1}, S^{k+1}, B^{k+1}) &= \arg\min_{Z,S,B} L(B,S,Z,Y^k) \\ &= \arg\min_{Z,S,B} E(Z) + \langle Y^k, Z - BS\rangle + \frac{\mu}{2}\left\|Z - BS\right\|_2^2 \end{aligned} \tag{4}$$

at the k-th iteration for $(Z^{k+1}, S^{k+1}, B^{k+1})$, then updates the multiplier Y by the formula

$$Y^{k+1} = Y^k + \mu(Z^{k+1} - B^{k+1}S^{k+1}) \ . \tag{5}$$

Since solving (4) for Z, S and B simultaneously can be difficult, an alternative choice is to minimize the augmented Lagrangian function with respect to each block variable Z, S and B one at a time while fixing the other two blocks at their latest values, and then update the Lagrange multiplier using

$$(S^{k+1}, Z^{k+1}) = \arg\min_{Z,S} L(B^k, S, Z, Y^k) \ , \tag{6}$$

$$B^{k+1} = \arg\min_{B} L(B, S^{k+1}, Z^{k+1}, Y^k) \ . \tag{7}$$

3. PROPOSED METHOD

3.1. Framework

By employing the operator splitting to problem (2), the unconstrained minimization problem is transformed to an equivalent constrained problem:

$$\min_{B,S} \frac{\lambda}{2}\|X-Z\|_F^2 + \alpha Tr(SLS^T) + \sum_{i=1}^{M}\|\mathrm{s}_i\|_1 \quad \text{s.t.} \quad Z = BS; \quad \|\mathrm{b}_j\|^2 \le 1, \forall j = 1,\cdots,J \tag{8}$$

The augmented Lagrangian function of problem (8) is

$$L(B,S,Z,Y) = \frac{\lambda}{2}\|X-Z\|_F^2 + \alpha Tr(SLS^T) + \sum_{i=1}^{M}\|s_i\|_1 - < Y, BS-Z > + \frac{\mu}{2}\|BS-Z\|_F^2 . \tag{9}$$

Directly finding the saddle point of the augmented Lagrangian function $L(B,S,Z,Y)$ is difficult, hence the alternating direction method (ADM) is used to solve the following sub-problems iteratively:

$$(S^{k+1}, Z^{k+1}) = \arg\min_{S,Z} \frac{\lambda}{2}\|X-Z\|_F^2 + \alpha Tr(SLS^T) + \sum_{i=1}^{M}\|\mathrm{s}_i\|_1 + \frac{\mu}{2}\|B^k S - Z - Y^k/\mu\|_F^2 \tag{10}$$

$$B^{k+1} = \arg\min_{B} \frac{\mu}{2}\|BS^{k+1} - Z^{k+1} - Y^k/\mu\|_F^2 , \tag{11}$$

$$Y^{k+1} = Y^k + \mu(-B^{k+1}S^{k+1} + Z^{k+1}) . \tag{12}$$

3.2. S -and Z -subproblems

Firstly, the minimization of Eq. (10) with respect to Z can be computed analytically. Specially, with the first and fourth terms of Eq. (10), we obtain the optimal solution:

$$Z = (\lambda X + \mu(B^k S - Y^k/\mu))/(\lambda+\mu) . \tag{13}$$

Moreover, it follows that

$$Y^{k+1} = Y^k + \mu(-B^k S^{k+1} + Z^{k+1}) = \frac{\lambda\mu(X - B^k S^{k+1} + Y^k/\mu)}{\lambda+\mu} . \tag{14}$$

In the following, the determination of S is a crucial problem. Here a proximal operator and the threshold technique are employed to find the approximate solution, and subsequently an iterative procedure is developed. Concretely, substituting Z of Eq. (13) into Eq. (10) yields

$$S^{k+1} = \arg\min_{S} \frac{\lambda\mu}{2(\lambda+\mu)}\|B^k S - X - Y^k/\mu\|_F^2 + \alpha Tr(SLS^T) + \sum_{i=1}^{M}\|\mathrm{s}_i\|_1 . \tag{15}$$

When updating s_i, the other vectors $\{\mathrm{s}_j\}_{j\neq i}$ are fixed. The optimization problem for each s_i:

$$\mathrm{s}_i^{k+1} = \arg\min_{\mathrm{s}_i} \frac{\lambda\mu}{2(\lambda+\mu)}\|B^k \mathrm{s}_i - \mathrm{x}_i - \mathrm{y}_i^k/\mu\|_F^2 + \alpha L_{ii}\mathrm{s}_i^T\mathrm{s}_i + \mathrm{s}_i^T\mathrm{h}_i + \|\mathrm{s}_i\|_1 . \tag{16}$$

where $\mathrm{h}_i = 2\alpha\sum_{j\neq i} L_{ij}\mathrm{s}_j$.

Let $f(\mathrm{s}_i) = \frac{\lambda\mu}{2(\lambda+\mu)}\|B^k \mathrm{s}_i - \mathrm{x}_i - \mathrm{y}_i^k/\mu\|_F^2 + \alpha L_{ii}\mathrm{s}_i^T\mathrm{s}_i + \mathrm{s}_i^T\mathrm{h}_i$, we give

$$\nabla f(\mathrm{s}_i) = \frac{\lambda\mu}{(\lambda+\mu)}(B^k)^T(B^k\mathrm{s}_i - \mathrm{x}_i - \mathrm{y}_i^k/\mu) + 2\alpha L_{ii}\mathrm{s}_i + \mathrm{h}_i .$$

Then following the iterative shrinkage/thresholding algorithm (ISTA) [12, 14], it yields

$$\begin{aligned}\mathrm{s}_i^{m+1} &= \arg\min_{\mathrm{s}_i}\left\{\gamma\|\mathrm{s}_i - [\mathrm{s}_i^m - \nabla f(\mathrm{s}_i^m)/2\gamma]\|_2^2 + \|\mathrm{s}_i\|_1\right\} \\ &= \arg\min_{\mathrm{s}_i}\left\{\gamma\|\mathrm{s}_i - [\mathrm{s}_i^m + (-2\alpha L_{ii}\mathrm{s}_i^m - \mathrm{h}_i + (B^k)^T\mathrm{y}_i^m)/2\gamma]\|_2^2 + \|\mathrm{s}_i\|_1\right\} \\ &= Shrink(\mathrm{s}_i^m + (-2\alpha L_{ii}\mathrm{s}_i^m - \mathrm{h}_i^m + (B^k)^T\mathrm{y}_i^m)/2\gamma, 1/2\gamma)\end{aligned} \tag{17}$$

where $\gamma \ge [\lambda\mu/2(\lambda+\mu)]eig((B^k)^T B^k) + \alpha L_{ii}$.

3.3. B -subproblem

By taking the derivative of Eq. (11) with respect to B and setting it to zero, we can get the following update rule:

$$\begin{aligned}B^{k+1} &= B^k - \zeta\left[-Y^k + \mu(B^k S^{k+1} - Z^{k+1})\right](S^{k+1})^T \\ &= B^k + \zeta Y^{k+1}(S^{k+1})^T .\end{aligned} \tag{18}$$

Then, the normalization on dictionary columns is required such that the dictionary B^{k+1} is a matrix whose columns are unit l_2 -norm.

3.4. Parameters and Algorithm Convergence

The proposed algorithm involves three important parameter: λ , α and μ . The computation time of the algorithm is mainly controlled by Line 4-5 in **Algorithm** GRSC-AL.

Algorithm GRSC-AL

1: initiation: $S^0 = 0$; $C^0 = 0$; B^0 ; λ ; α ; μ

2: while stop-criterion not satisfied (loop in k):

3: while stop-criterion not satisfied (loop in m):

4: $Y^{m+1} = \frac{\lambda\mu}{\lambda+\mu}(-B^k S^{k,m} + X + C^k/\mu)$; update H^m

5: $S^{k,m+1} = Shrink(S^{k,m} + \frac{[-2\alpha S^{k,m}diag(L) - H^m + (B^k)^T Y^m]}{2\gamma^k}, \frac{1}{2\gamma^k})$

6: end while of loop m

7: $C^{k+1} = Y^{m+1}$; $S^{k+1,0} = S^{k,m+1}$

8: $B^{k+1} = B^k + \zeta C^{k+1}(S^{k+1,0})^T$; $\mathrm{b}_j^{k+1} = \mathrm{b}_j^{k+1}/\|\mathrm{b}_j^{k+1}\|_2$, $\forall j$

9: $\gamma^k = [\lambda\mu/2(\lambda+\mu)]eig((B^k)^T B^k) + \alpha diag(L)$

10: end while of loop k

It is worth noting that when the augmented Lagrangian method was employed in non-convex problems, the authors in Refs.[9] and [13] stated that only a weak convergence is observed in their algorithms, i.e., under mild conditions any limit point of the iteration sequence generated by the algorithm is a Karush-Kuhn-Tucker (KKT) point. In our work, we also give a result with regard to the convergence of the **Algorithm** GRSC-AL for this new non-convex

problem**.** It should be emphasized that although the following convergence result is far from being satisfactory, it provides an assurance for the behavior of the algorithm. Moreover, empirical evidence suggests that the proposed algorithm has a very strong convergence behavior.

Proposition 1. *Let* $V \triangleq (B,S,Z)$ *and* $\{V^k\}_{k=1}^{\infty}$ *be generated by algorithm GRSC-AL, Assume that* $\{V^k\}_{k=1}^{\infty}$ *is bounded and* $\lim_{k\to\infty}(V^{k+1}-V^k)=0$. *Then any accumulation of* $\{V^k\}_{k=1}^{\infty}$ *satisfies the KKT conditions. In particular, whenever* $\{V^k\}_{k=1}^{\infty}$ *converges, it converges to a KKT point of (8).*

4. EXPERIMENTS

The following three algorithms for image clustering: the proposed GRSC-AL + K-means, Sparse coding (SC) + K-means, and GraphSC (implemented by Dual-Feature method) + K-means, were compared on two real world image datasets, i.e., CMU-PIE face database and COIL20 image database. The CMU-PIE face database contains 68 subjects and each has 21 images under different lighting conditions. The COIL20 image database contains 20 subjects and each has 72 images under different rotated orientations. All algorithms were implemented in MATLAB on a Window XP laptop with 3.1GHz processor and 4GB of RAM.

Specifically, all algorithms firstly applied PCA to reduce the data dimensionality and then performed in the subspace. Finally, the K-means algorithm was applied on the new representations to obtain the clustering result. The dimensionality after PCA projection and the dictionary size used in the experiments were the same as those in ref. [6]. The parameters in SC, GraphSC, and GRSC-AL (e.g. λ, α) were determined by cross validation [6]. Besides, the parameter μ in GRSC-AL was set to be 10. The clustering results were evaluated by accuracy (AC) and normalized mutual information (NMI) [6], where AC was defined by comparing the clustered label of each sample with the label provided by the dataset, and NMI measured the dependence between the sets of clusters obtained from the compared algorithm and the ground truth.

Fig. 1 and Fig.2 show the plots of clustering AC and NMI versus the number of tested cluster (c) ranging from 4 to 68 on the CMU-PIE database and from 2 to 20 on the COIL20 database, respectively. It can be observed that our proposed algorithm generally outperforms the GraphSC algorithm, especially when c is larger. Fig. 3 displays one run of these two algorithms, in terms of the evolution of the NMI when $c=68$. As can be seen, GraphSC needs almost 40 iterations to reach the convergence zone of NMI and our proposed algorithm needs only about 10 iterations. The average computation times per iteration of GraphSC and GRSC-AL on the two datasets are shown in Table 1. Obviously, the average computation time per iteration of GRSC-AL is much shorter than that of GraphSC.

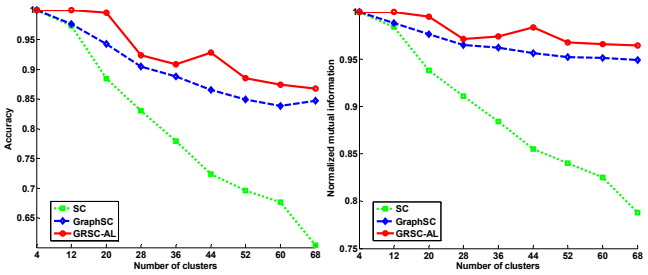

Fig.1. AC (left) and NMI (right) versus the number of clusters on CMU-PIE

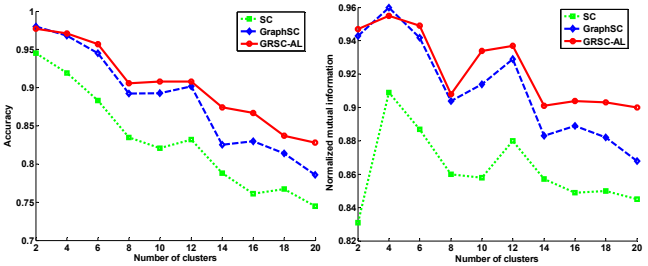

Fig.2. AC (left) and NMI (right) versus the number of clusters on COIL20

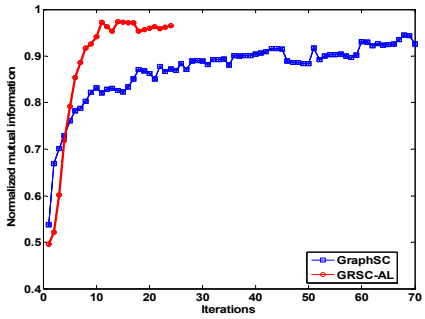

Fig.3. One run of the evolution of the NMI on CMU-PIE using GraphSC(left) and GRSC-AL(right) when $c=68$

Table 1. Average computation time per iteration

Algorithm	CMU-PIE, $c=68$	COIL20, $c=20$
GraphSC	31.2s	68.7s
GRSC-AL	0.74s	1.19s

5. CONCLUSIONS

An efficient algorithm for solving graph regularized sparse coding is proposed. The algorithm extends the augmented Lagrangian framework to non-convex problems involving more than two blocks of variables. Preliminary experiments in clustering demonstrate that the proposed algorithm outperforms the existing algorithms in terms of computation-al time and discriminating power. Ongoing research includes a thoroughly experimental evaluation of GRSC-AL in clustering, classification [6] and restoration [15].

6. REFERENCES

[1] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, pp. 4311-4322, 2006.

[2] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Advances in Neural Information Processing Systems*, vol. 20, pp. 801-808, 2007.

[3] A. d'Aspremont, L. El Ghaoui, M. Jordan, and G. Lanckriet, "A direct formulation for sparse PCA using semidefinite programming," *SIAM review*, vol. 49, p. 434, 2007.

[4] P. Hoyer, "Non-negative matrix factorization with sparseness constraints," *The Journal of Machine Learning Research*, vol. 5, pp. 1457-1469, 2004.

[5] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2272-2279, 2009.

[6] M. Zheng, J. Bu, C.A. Chen, C. Wang, L. Zhang, G. Qiu, D. Cai, "Graph regularized sparse coding for image representation," *IEEE Trans. Image Process.*, vol. 20, pp. 1327-1336, 2011.

[7] K. Skretting and K. Engan, "Recursive least squares dictionary learning algorithm," *IEEE Trans. Image Process.*, vol. 58, pp. 2121-2130, 2010.

[8] K.K-Delgado, J. Murray, B. Rao, K. Engan, T. Lee, T. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Comp.*, vol. 15, pp. 349-396, 2003.

[9] Q. Liu, S. Wang, and J. Luo et al, "A multi-scale augmented Lagrangian approach to general dictionary learning for image denoising," *Journal of Visual Communication and Image Representation*, vol. 23, pp. 753-766, 2012..

[10] RT Rockafellar, "Augmented Lagrangians and applications of the proximal point algorithm in convex programming," *Math. Oper. Res.*, vol. 1, pp. 97-116, 1976.

[11] M. Afonso, J.B-Dias, and M. Figueiredo, "An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems," *IEEE Trans. Image Process.*, vol. 20, 2011.

[12] W. Yin, S. Osher, D. Goldfarb, and J Darbon, "Bregman iterative algorithms for l1-minimization with applications to compressed sensing," *SIAM J. Imaging Sci.*, vol. 1, pp. 142-168, 2008.

[13] Y. Shen, Z. Wen, and Y. Zhang, "Augmented Lagrangian alternating direction method for matrix separation based on low-rank factorization," Rice University CAAM Technical Report TR11-02.

[14] X. Zhang, M. Burger, X. Bresson, and S. Osher, "Bregmanized nonlocal regularization for deconvolution and sparse reconstruction," *SIAM J. Imag. Sci.*, vol. 3, pp. 253-276, 2010.

[15] X. Lu, H. Yuan, P. Yan, Y. Yuan, and X. Li, "Geometry constrained sparse coding for single image super-resolution" :in *Proc. CVPR*, pp.1648-1655, 2012.