Statistica Sinica **18**(2008), 1449-1464

PENALIZED PARAFAC ANALYSIS OF SPONTANEOUS EEG RECORDINGS

Eduardo Martínez-Montes, José M. Sánchez-Bornot and Pedro A. Valdés-Sosa

Cuban Neuroscience Center

Abstract: The multidimensional nature of neuroscience data has made the use of multi-way statistical analysis suitable in this field. Parallel Factor Analysis (PARAFAC) is a multidimensional generalization of PCA with the advantage of offering unique solutions. However, imposing physiologically acceptable constraints would improve the interpretation of this type of analysis. In this work we propose a new algorithm called Alternating Penalized Least Squares to estimate PARAFAC solutions using different kinds of soft penalization. The algorithm relies on the recent generalization of modified Newton-Raphson techniques to estimate a multiple penalized least squares model. Applied to semi-synthetic and real spontaneous EEG time-varying spectra, we show that a wide range of sparse and smooth solutions can be found separately, as well as with these two properties combined. Smoothness is usually desired in spectra, and different sparse scenarios are observed in the temporal evolution of physiological intermittent phenomena. The degree of constraints can be tuned through the weighting parameters, whose optimal values can be chosen by means of the cross-validation and Corcondia measures.

Key words and phrases: Dimensionality reduction, EEG, PARAFAC, penalized regression.

1. Introduction

Tools for the analysis of multidimensional data arrays have recently gained popularity in neuroscience (Miwakeichi, Martínez-Montes, Valdés-Sosa, Nishiyama, Mizuhara and Yamaguchi (2004), Morup, Hansen, Herrmann, Parnas and Arnfred (2006) and Beckmann and Smith (2005)). This multi-way analysis is the natural extension of usual multivariate analysis, and it offers several advantages over the well-known bilinear methods for dimensionality reduction, such Principal Component Analysis (PCA) and Independent Component Analysis (ICA). The first advantage is that of more parsimonious and interpretable data models. Another advantage is the achievement of unique decompositions under very mild conditions, without constraining the solutions to be either orthogonal or statistically independent. Several models and algorithms for multi-way analysis have been developed (Bro (1998)). Of particular interest is the PARAFAC model,

first proposed by Harshman (1970) and recently used for the analysis of spontaneous EEG data (Miwakeichi et al. (2004)). The basic model for a PARAFAC decomposition of a three-way data array $\mathbf{X}_{(I\times J\times K)}$ of elements x_{ijk} is:

$$x_{ijk} = \sum_{f=1}^{N_f} a_{if} b_{jf} c_{kf} + \varepsilon_{ijk}, \tag{1.1}$$

where ε_{ijk} represents an error term. The problem is to find the loading matrices, or signatures, $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, whose elements are a_{if}, b_{jf} and c_{kf}, respectively, with columns corresponding to components (indexed by f) which are also designated as 'atoms' (see Figure 1 of supplemental material online). This model does not suffer from rotational freedom and its only intrinsic indeterminacies are the order of the atoms and the relative scaling of the signatures. These can be solved in practice by choosing the first atom as the one explaining most of the variance, normalizing two of the estimated loadings and scaling the other with the overall explained variance. Therefore, the model is considered to be essentially unique (Stegeman and Sidiropoulos (2007)).

Sufficient conditions for the uniqueness of PARAFAC were given in Harshman (1970), although the most general condition is due to Kruskal (1977). Kruskal´s rank (*k-rank*) of a matrix is the largest number r such that every subset of r columns of the matrix is linearly independent. Uniqueness of the solution is guaranteed when $k\text{-}rank(\mathbf{A})+ k\text{-}rank(\mathbf{B})+ k\text{-}rank(\mathbf{C}) \geq 2N_f + 2$. This is a less-stringent condition than either orthogonality or statistical independence (Sidiropoulos and Bro (2000)). Necessary and sufficient conditions for unique decomposition of higher dimensional arrays are discussed in Stegeman and Sidiropoulos (2007).

Kruskal also showed that if the data conforms to the model, PARAFAC analysis will recover the true underlying phenomena if the correct number of components is used and if the signal-to-noise ratio is appropriate (Kruskal (1977)). To select the appropriate number N_f of components we use the Core Consistency Diagnostic (Corcondia) test (Bro (1998)). This measure takes the value 100% when the data conform exactly to the trilinear model. If Corcondia is lower than 85%, then either too many components have been extracted, the model is misspecified, or gross outliers disturb the model (Bro (1998)). For other details on this issue see Section 1 of the supplemental material (online).

Although other algorithms have been proposed, PARAFAC is most often estimated by Alternating Least Squares (ALS), which offers a good trade-off between computational expense and quality of the solution (Tomasi and Bro (2006)). This consists of simply dividing the parameters into several sets, each being estimated in a least squares sense, conditionally on the remaining parameters. This can be formalized using the following definition.

Definition. Let $\mathbf{A} \in R^{m\times n}$ and $\mathbf{B} \in R^{p\times n}$ be two matrices with columns denoted as $\mathbf{a}_i$ and $\mathbf{b}_i$, $i = 1, \ldots, n$, respectively. Then, the matrix $\mathbf{C} \in R^{mp\times n}$; $\mathbf{C} = \left[\mathbf{a}_1 \otimes \mathbf{b}_1 \cdots \mathbf{a}_n \otimes \mathbf{b}_n\right]$ is called the Khatri-Rao product of $\mathbf{A}$ and $\mathbf{B}$, denoted as $\mathbf{C} = \mathbf{A} \,|\otimes|\, \mathbf{B}$, where $\otimes$ is the Kronecker product.

If the tensor data is reshaped by joining along the second dimension all slices $\mathbf{X}_k$ $(k = 1, \ldots, K)$, we end up with a matrix $\overline{\mathbf{X}}_{\mathbf{A}}^{(I\times JK)}$, and the model can be rewritten in terms of the loadings matrices as $\overline{\mathbf{X}}_{\mathbf{A}}^{(I\times JK)} = \mathbf{A}\,(\mathbf{C} \,|\otimes|\, \mathbf{B})^T + \mathbf{E}_{\mathbf{A}}^{(I\times JK)}$. Here, $\mathbf{E}_{\mathbf{A}}^{(I\times JK)}$ is the error matrix equally rearranged. Similarly, reshaping the original data in such a way that the second (or third) dimension runs along rows and the other two are joined along columns, leads to the following equivalent forms of (1.1):

$$\overline{\mathbf{X}}_{\mathbf{B}}^{(J\times KI)} = \mathbf{B}\,(\mathbf{A} \,|\otimes|\, \mathbf{C})^T + \mathbf{E}_{\mathbf{B}}^{(J\times KI)}; \quad \overline{\mathbf{X}}_{\mathbf{C}}^{(K\times IJ)} = \mathbf{C}\,(\mathbf{B} \,|\otimes|\, \mathbf{A})^T + \mathbf{E}_{\mathbf{C}}^{(K\times IJ)}.$$

The global or general problem in PARAFAC has the loss function

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \left\| \overline{\mathbf{X}}_{\mathbf{A}}^{(I\times JK)} - \mathbf{A}\,(\mathbf{C} \,|\otimes|\, \mathbf{B})^T \right\|^2,$$

where $\|\mathbf{Y}\|$ denotes the Frobenius ($l2$-) norm of a matrix $\mathbf{Y}$, $\|\mathbf{Y}\| = \sqrt{\text{trace}\,(\mathbf{Y}^T\mathbf{Y})}$. With auxiliary matrices $\mathbf{Z}_{\mathbf{A}}^{(JK\times N_f)} = (\mathbf{C} \,|\otimes|\, \mathbf{B})$, $\mathbf{Z}_{\mathbf{B}}^{(KI\times N_f)} = (\mathbf{A} \,|\otimes|\, \mathbf{C})$, and $\mathbf{Z}_{\mathbf{C}}^{(IJ\times N_f)} = (\mathbf{B} \,|\otimes|\, \mathbf{A})$, the ALS algorithm can be expressed as three ordinary least squares (OLS) regressions.

1. Initialize two of the loadings, say $\mathbf{B}$ and $\mathbf{C}$.
2. $\widehat{\mathbf{A}} = \arg\min \left\|\overline{\mathbf{X}}_{\mathbf{A}} - \mathbf{A}\mathbf{Z}_{\mathbf{A}}^T\right\|^2$.
3. $\widehat{\mathbf{B}} = \arg\min \left\|\overline{\mathbf{X}}_{\mathbf{B}} - \mathbf{B}\mathbf{Z}_{\mathbf{B}}^T\right\|^2$.
4. $\widehat{\mathbf{C}} = \arg\min \left\|\overline{\mathbf{X}}_{\mathbf{C}} - \mathbf{C}\mathbf{Z}_{\mathbf{C}}^T\right\|^2$.
5. Repeat Steps 2, 3 and 4 until relative change in fit is smaller than a specified criterion.

Such an algorithm may only improve the fit or keep it the same, driving the loss function to monotonically decrease. Since the problem is a bounded-cost problem (the loss function cannot be less than zero) convergence follows. This property is very attractive, and one of the reasons for the widespread use of ALS. However, the noisy nature of neuroscience data may lead to difficult-to-interpret solutions and, in the worst case, to solutions without physiological interpretation at all. Therefore, the use of appropriate constraints is usually helpful for obtaining clinically and neurophysiologically sound results.

In this context, constraints can be applied as either approximate or exact. In the available implementation of PARAFAC (Andersson and Bro (2000)), only

exact orthogonality, nonnegativity and unimodality of components can be used as constraints, although in theory many others are possible (Bro (1998)). However, in the study of complex systems such as the brain through noisy data, exact constraints are not suitable. Recently, PARAFAC has been estimated through the Expectation-Maximization (EM) algorithm and the Variational Bayesian EM (Morup (2005)). These algorithms imply the use of prior information (approximate constraints) on some or all of the loadings, reducing to ALS when delta functions are used. The use of the Bayesian approach offers a natural way of imposing constraints through prior information and also allows one to address the evaluation of the optimal number of components to extract (e.g. through Automatic Relevant Detection or the Bayesian Information Criterion). However, the implementation of these methods depends strongly on the assumed prior densities for the loadings.

As an alternative, in this work we propose the use of approximate constraints in the ALS approach for a physiologically valid PARAFAC analysis of neuroscience data. Recent advances in the field of least squares regression allow one for the first time to efficiently constrain one or more signatures to be smooth, or sparse, or even to have both these properties. Some of these constraints would be very difficult to deal with in the EM/VBEM approaches and would lead to very slow algorithms. The next section presents the modifications of the ALS algorithm to include penalizations, as well as other details for efficient implementation and estimation of optimal weights for the constraints. Section 3 gives the results of the application of the new method to the analysis of actual and semi-synthetic EEG data, and Section 4 is devoted to the discussion and conclusions of the study.

2. Alternating Penalized Least Squares

Without loss of generality, we focus on the estimation of one of the loadings to be penalized, say $\mathbf{A}$. For this loading, the OLS solution is given by the second step of the ALS algorithm presented above. For simplicity, we write $\overline{\mathbf{X}}_{\mathbf{A}}^{T}$ and $\mathbf{Z}_{\mathbf{A}}$ as $\mathbf{X}$ and $\mathbf{Z}$, respectively. A constraint is introduced by adding a penalization term $P(\mathbf{A})$:

$$\widehat{\mathbf{A}} = \arg\min \left(\left\| \mathbf{X} - \mathbf{Z}\mathbf{A}^{T} \right\|^{2} + \lambda P(\mathbf{A}) \right).$$

The nonnegative parameter λ quantifies the relative importance of the two competing (fit and constraint) terms. Of particular interest is the well-known Ridge regression (Hoerl and Kennard (2000)), where the penalty function is quadratic in $\mathbf{A}$, having the general form $P(\mathbf{A}) = \|\mathbf{L}_1\mathbf{A}\mathbf{L}_2\|^2$, with $\mathbf{L}_1$ and $\mathbf{L}_2$ being two operators that operate on the columns and rows of $\mathbf{A}$ respectively. The choice of the first or second order difference operator for $\mathbf{L}_1$ ($\mathbf{L}_2$) is aimed

at the imposition of smoothness along rows (columns) of the coefficients matrix (Timmerman and Kiers (2002)). Alternatively, $P(\mathbf{A})$ can be a non-convex penalty function characterized by having a singularity at the origin and which leads to sparse solutions (Fan and Li (2001)). In this line, some penalizers are the Least Absolute Shrinkage Selection Operator (Lasso) (Tibshirani (1996)), which uses the *l1*-norm of $\mathbf{A}$, so $P(\mathbf{A}) = \|\mathbf{A}\|_1$; a variant called the "Fusion Lasso" (Land and Friedman (1996)), with $P(\mathbf{A}) = \|\mathbf{L}\mathbf{A}\|_1$, where $\mathbf{L}$ is the first order difference operator; and the Smooth Clipped Absolute Deviation (SCAD) (Fan (1997)), with a more complicated definition for the penalty function. Also, some particular combinations of penalties have been introduced, such as the "Fused Lasso" (Tibshirani, Saunders, Rosset, Zhu and Knight (2005)), which combines typical penalties of Lasso and Fusion Lasso; and Elastic Net (Enet) (Zou and Hastie (2005)) combining *l1*-norm penalties (Lasso) and quadratic penalties (Ridge). A general expression for this penalty is $P(\mathbf{A}) = \mu_1 \|\mathbf{L}_1\mathbf{A}\|_1 + \mu_2 \|\mathbf{L}_2\mathbf{A}\|^2$, where $\mu_i, (i = 1, 2)$ is the weight for the *l(i)*-norm term. These strategies are suitable in problems where group behavior is searched for in some of the coefficients. A compendium of different non-convex penalizers and their application to neuroscience data can be found in Valdés-Sosa, Sánchez-Bornot, Vega-Hernández, Melie-García, Lage-Castellanos and Canales-Rodríguez (2006).

For estimating penalized linear regression models with the use of non-convex penalties, (which are not algebraically treatable), we used the Local Quadratic Approximation (LQA) algorithm (Fan and Li (2001)). It unifies nearly all variable selection techniques into an easy-to-implement iterative application of Ridge regression, and retains the convergence properties of the Newton-Raphson algorithm (Hunter and Li (2005)). Recently, our group has developed a generalized LQA variant to tackle the estimation of a penalized least squares model with combinations of different types of penalties (Sánchez-Bornot, Martínez-Montes, Lage-Castellanos, Vega-Hernández and Valdés-Sosa (2008)). This is called Multiple Penalized Least Squares (MPLS) and, for a PARAFAC loading, is established as

$$\widehat{\mathbf{A}} = \arg\min \left(\left\| \mathbf{X} - \mathbf{Z}\mathbf{A}^T \right\|^2 + \sum \lambda_l P_l(\mathbf{A}) \right), \tag{2.1}$$

where $l = 1, \ldots, N_l$, indexes the penalty functions and corresponding weighting parameters. This loss function cannot be separated into contributions from columns of $\mathbf{A}$ (rows of $\mathbf{A}^T$), thus each column has to be estimated conditionally on the others using a backfitting algorithm (Hastie and Tibshirani (1990)). Mathematically, if we set $\mathbf{T}_f = \mathbf{X} - \sum_{f' \neq f} \mathbf{z}_{f'} \mathbf{a}_{f'}^T$, then the loss function for the f-th atom $\mathbf{a}_f$ can be written as $\|\mathbf{T}_f - \mathbf{z}_f \mathbf{a}_f^T\|^2 + \sum \lambda_l P_l(\mathbf{A})$, where $\mathbf{z}_f$ is the f-th column of $\mathbf{Z}$. The solution to this problem is not necessarily the overall solution of (2.1) in the least squares sense. However, this formulation is very useful in

practice due to the following lemma, whose proof can be found in Section 2 of the supplemental material (online).

Lemma 1. *Consider the minimization subject to any constraint of the loss function of a multiple penalized linear regression model for a row* $\mathbf{a}^T$: $\min(\|\mathbf{T} - \mathbf{z}\mathbf{a}^T\|^2 + \sum \lambda_l P_l(\mathbf{a}))$. *The solution is that of* $\min \|\alpha - \mathbf{a}\|^2 + \sum \overline{\lambda}_l P_l(\mathbf{a}))$,*where* α *is the solution of the unconstrained problem* $\alpha = \arg\min(\|\mathbf{T} - \mathbf{z}\alpha^T\|^2) = \mathbf{T}^T\mathbf{z}/\mathbf{z}^T\mathbf{z}$, *and* $\overline{\lambda}_l = \lambda_l/\mathbf{z}^T\mathbf{z}$.

This result allows for a fast computation of each atom which compensates for the slowness of the iterative backfitting process. Moreover, it is possible to use different constraints for each atom separately. The APLS algorithm can then be summarized as follows.

1. Initialize the loadings: $\mathbf{A}_0$, $\mathbf{B}_0$ and $\mathbf{C}_0$.
2. Iterate until convergence the following steps (iteration t).
3. Estimate $\widehat{\mathbf{A}}_t = backfitting(\overline{\mathbf{X}}_\mathbf{A}, \mathbf{A}_{t-1}, \mathbf{B}_{t-1}, \mathbf{C}_{t-1}, P_\mathbf{A})$.
4. Estimate $\widehat{\mathbf{B}}_t = backfitting(\overline{\mathbf{X}}_\mathbf{B}, \mathbf{A}_{t-1}, \mathbf{B}_{t-1}, \mathbf{C}_{t-1}, P_\mathbf{B})$.
5. Estimate $\widehat{\mathbf{C}}_t = backfitting(\overline{\mathbf{X}}_\mathbf{C}, \mathbf{A}_{t-1}, \mathbf{B}_{t-1}, \mathbf{C}_{t-1}, P_\mathbf{C})$.

($P_\mathbf{A}$, $P_\mathbf{B}$ and $P_\mathbf{C}$ summarize multiple penalties on $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, respectively.)

Backfitting Algorithm: $\widehat{\mathbf{A}} = backfitting(\overline{\mathbf{X}}_\mathbf{A}, \mathbf{A}, \mathbf{B}, \mathbf{C}, P_\mathbf{A})$.

(i) For each column $\mathbf{a}_f$, $\mathbf{b}_f$ and $\mathbf{c}_f$ of $\mathbf{A}$,$\mathbf{B}$ and $\mathbf{C}$, respectively, compute $\mathbf{z}_f = \mathbf{c}_f \otimes \mathbf{b}_f$ and $\alpha_f = (\overline{\mathbf{X}}_\mathbf{A} - \sum_{f' \neq f} \mathbf{z}_{f'} \mathbf{a}_{f'}^T)^T \mathbf{z}_f / \mathbf{z}_f^T \mathbf{z}_f$.

(ii) Estimate $\widehat{\mathbf{a}}_f = \arg\min(\|\alpha_f - \mathbf{a}_f\|^2 + P_\mathbf{A})$.

(iii) Repeat (i) and (ii) until convergence.

Finally, two important issues should be mentioned. First, each iterative step (penalized least squares) of the backfitting algorithm is approximated by iterative ridge regressions using LQA to guarantee its global convergence. This ensures the convergence of the backfitting (Ansley and Kohn (1994)), improving the fit or keeping it the same. Therefore, similar to ALS, since the loss function is non-negative, the whole algorithm converges at least to a local minima. On the other hand, in some cases PARAFAC is known to depend strongly on initial loadings. For the ALS algorithm, several options have been used for obtaining initial estimates ranging from random guesses to direct trilinear decomposition (Bro (1998)). A common option has been to use several runs with initial guesses to ensure convergence to a unique solution. We follow this approach in the case of synthetic data, although for real data we always start from the unconstrained PARAFAC solution, which ensures that penalized loadings will resemble the original ones.

Second, we have to set values for the weighting parameters for each penalty function that allow a continuous control over the corresponding constraint. Automatic selection of optimal values can be found by generalized cross-validation (GCV) (Golub, Heath and Wahba (1979)), or information criteria such as Akaike's (Akaike (1974)) or Schwartz's Bayesian Information Criterion (Schwartz (1978)). In this work we compute solutions with different values for the weighting parameters and review corresponding values of the logarithm of GCV (logGCV), and the Corcondia measure, for identifying an 'optimal' solution. In the case of using several penalty functions, this can lead to a computationally expensive approach. Thus, for the case of the Enet penalty, we follow a different approach that consists of using only a few pairs of values for μ_1 and μ_2 such that $\mu_1 + \mu_2 = 1$, and finding the optimal weighting parameter common for both terms through inspection of log GCV and Corcondia. The former is logGCV$= \log(\widehat{\sigma}^2) - \log(1 - df/N)$, where $N = IJK$ is the number of data elements, $\widehat{\sigma}^2 = \|\mathbf{X} - \widehat{\mathbf{X}}\|^2/(N - df)$ is an estimate of the error variance component, and *df* is the effective number of parameters (degrees of freedom), which is very difficult to compute for nonlinear models and non-quadratic penalties. Here, we approximate *df* by the sum of each loading's degrees of freedom. This approximation is also used in the backfitting process for each atom, as proposed by Hastie and Tibshirani (1990), Chapters 2 and 6.

3. Constrained Decomposition of EEG Data

3.1. Ordinary PARAFAC

The data used in this study is the time-varying spectrum of a resting-state EEG recording of 16 bipolar derivations. This is a three-dimensional array of 208 320 elements, indexed by 16 derivations, 124 frequencies and 105 time points, that can be subject to PARAFAC analysis as is schematically shown in Figure 1 of the supplemental material (online). The estimated loadings correspond to spatial, spectral and temporal signatures, respectively. More details about these data set and their preprocessing for PARAFAC can be found in the supplemental material (online), and in Martínez-Montes, Valdés-Sosa, Miwakeichi, Goldman and Cohen (2004)

Unconstrained PARAFAC decomposition via ALS was performed, and examination of Corcondia, residual errors, and explained variance allowed us to determine the appropriate number of components as three. Figure 1 shows the three atoms extracted for the spatial, temporal and spectral loadings. The latter allows the identification of the present rhythms in the data, namely alpha (solid line), theta (dot line), and gamma (dash line) atoms. Note that temporal signatures show different behaviors, being quite constant for the gamma atom, and showing intermittent activity for the alpha and theta atoms. The spatial

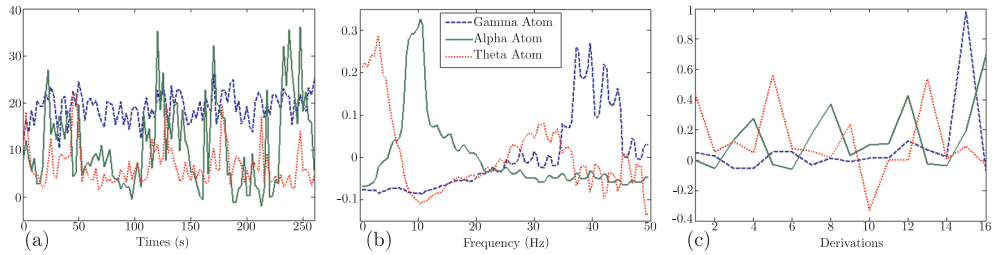

Figure 1. Unconstrained PARAFAC decomposition via ALS of the time-varying spectra of resting-state EEG. (a) Temporal signatures (b) Spectral signatures (c) Spatial signatures. The three atoms extracted are identified according to the classical band classification as alpha (8-12Hz), theta (4-8Hz), and gamma (>30Hz) from the spectral loadings. The spectral and spatial signatures are normalized and the temporal retains the scale of the data. Figure in color in the online version.

loadings are more difficult to interpret in this view and are not of interest in this paper, although the representation on the scalp is shown in Figure 2 of the supplemental material (online).

3.2. Smoothness

Although the main spectral peaks of the three atoms are clearly distinguished (Figure 1b), this is not always the case, and oscillations or roughness of the spectrum sometimes make it difficult to interpret. To overcome this, we imposed several degrees of smoothness on one spectral loading ($P(\mathbf{B}) = \|\mathbf{L}_2\mathbf{B}\|^2$, $\mathbf{L}_2$ being the second difference operator) while leaving the other two loadings unconstrained. Figure 2a-e show the spectral signatures for different values of the corresponding weighting parameter. As can be seen, the higher the weighting parameter, the smoother the signatures for all atoms. The 'optimal' value for this parameter is $\lambda = 1$ in terms of minimization of the GCV, the residual sum of squares (RSS) and the relative distances to unconstrained solution, as well as maximization of the Corcondia measure, as shown by Table 1 of the supplemental material (online). However, the RSS and logGCV obtained for the unconstrained PARAFAC decomposition are lower, which might be explained by its uniqueness, i.e., the constraint pulls the solution far from the least squares one. On the other hand, Figure 2f shows the spectral signatures obtained by requiring smoothness and non-negativity simultaneously, illustrating the feasibility of combining this kind of soft constraint with the hard constraint already used in PARAFAC (Bro (1998)).

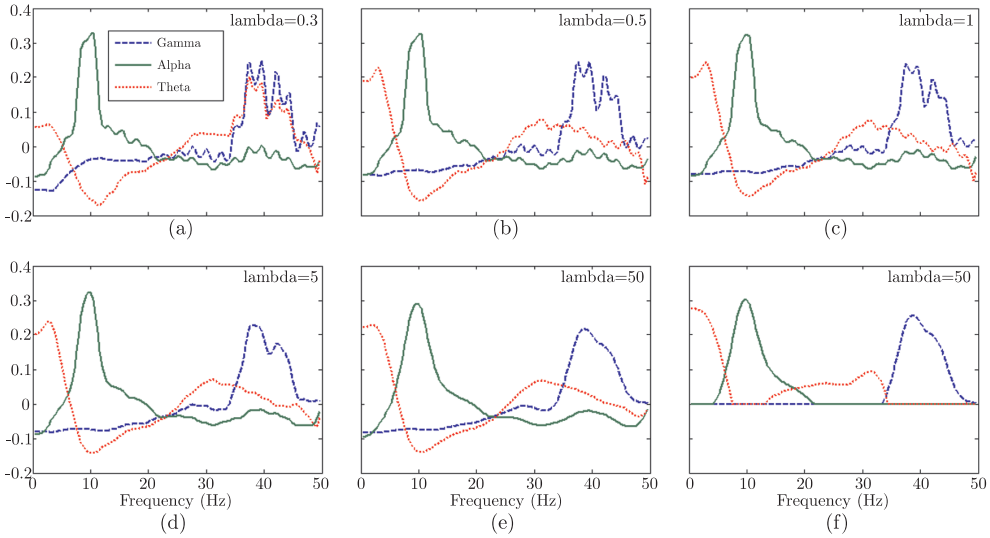

Figure 2. Spectral signatures constrained to be smooth, obtained by PARAFAC via APLS. (a)−(e) Spectral loadings with different degree of smoothness (see value of lambda, the smoothing parameter). (f) Spectral loadings constrained to be smooth and non-negative. Values of logGCV and Corcondia are shown in Table 1 of the supplemental material (online). Figure in color in the online version.

3.3. Sparsity and group behavior

EEG data and other neuroimages often show intermittent activity. For example, epileptic spikes are very localized in time, spontaneous rhythms usually alternates periods of high and low amplitudes, and experimental block designs give the amplitude of oscillations a box-like appearance. Having this in mind, we simulated the three scenarios for a temporal signature that are shown in the top row of Figure 3. The first is theoretically suitable for the use of Lasso penalization since it shows very sparse signatures (Figure 3a top). The second shows non-zero values in groups, within which all points have the same value (Figure 3b top) so, theoretically, this is the ideal situation for applying the Fusion Lasso penalization. Finally, the third also shows signatures with group behavior, but now with smooth variations in values inside a group (Figure 3c top), which is suitably tackled by penalizations combining smoothness and sparsity, such as Elastic Net. With these simulated temporal signatures and the unconstrained spatial and spectral loadings, we recomposed the three-dimensional data and added some white noise (signal-to-noise ratio of 20 dB). The bottom row of Figure 3 shows the unconstrained PARAFAC decomposition of this semi-synthetic data. Note that in all cases there is a good correspondence (Corcondia>99%), but the sparse nature of the real signatures (many zero values) cannot be recovered.

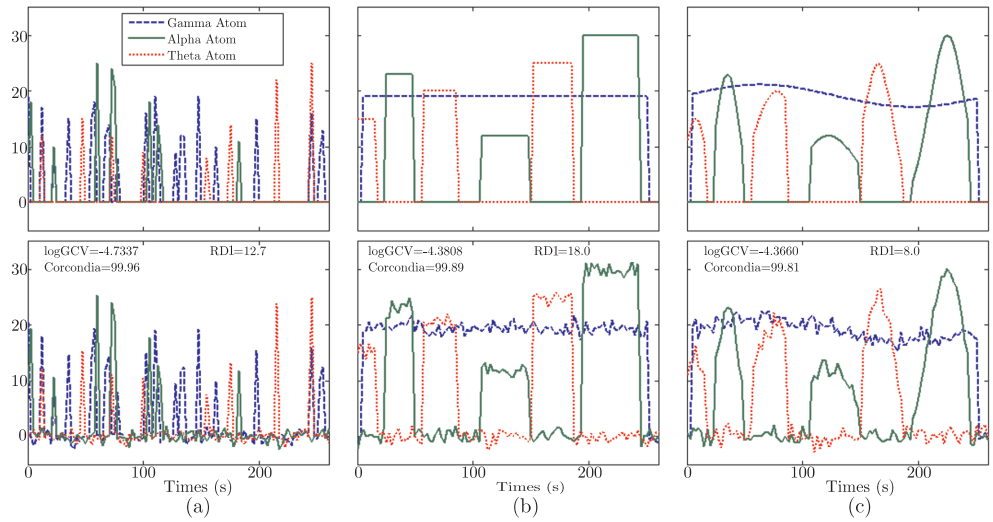

Figure 3. Simulated temporal signatures for the three atoms (top), and those obtained by unconstrained PARAFAC (bottom). Logarithm of Generalized Cross Validation (logGCV) function and relative distance (RD1) to the real loadings are also shown. (a) Lasso scenario; only some time points are activated. (b) Fusion Lasso scenario; each signature is a box-like function. (c) Elastic Net scenario; only a few patches are activated, but this activation is smooth inside the patch. Figure in color in the online version.

Penalized PARAFAC analyses with different constraints on the temporal signature, and without constraining the other two loadings, were performed. Figure 4a shows the temporal loading obtained by using a Lasso penalization ($P(\mathbf{C}) = \|\mathbf{C}\|_1$) for the first scenario (Figure 3a). Here, the optimum value (minimum logGCV) for the weighting parameter is 0.1, which also produces the lowest relative distance to the real loading, i.e., the one resembling the real loading most accurately. This plot for other values of the weighting parameter and corresponding logGCV, Corcondia, and relative distances are shown in the top row of Figure 3 of the supplemental material (online). Similarly, Figure 4b shows the temporal signatures obtained by using Fusion Lasso penalization ($P(\mathbf{C}) = \|\mathbf{L}_1\mathbf{C}\|_1$, $\mathbf{L}_1$ being the first difference operator) on the second simulated data set (Figure 3b). The value $\lambda = 0.9$ seems to be optimal, having the highest Corcondia, the lowest logGCV, and the lowest relative distance to the real loading. The temporal loadings estimated for different values of λ are shown in the middle row of Figure 3 of the supplemental material (online). Finally, in Figure 4c the temporal signatures estimated with the use of Enet penalization ($P(\mathbf{C}) = \mu_1 \|\mathbf{L}_1\mathbf{C}\|_1 + \mu_2 \|\mathbf{L}_2\mathbf{C}\|^2$) on the third simulated data set (Figure 3c) are shown. Enet solutions were found using a first order difference operator for the

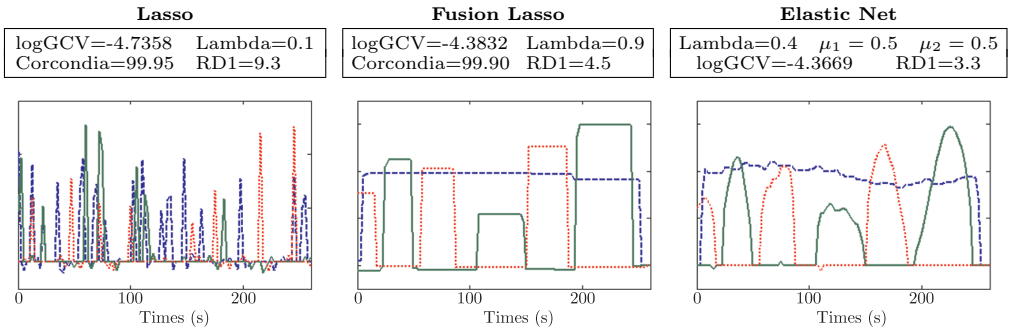

Figure 4. Estimated temporal signatures for the three atoms using constrained PARAFAC with corresponding penalization for the three simulated scenarios. Values of optimum lambda, logarithm of GCV, Corcondia (except for Enet), and relative distance to real temporal signature (in percent) are shown. Solid line represents the Alpha atom, dotted line represents the Theta atom, and dashed line the Gamma atom. Figure in color in the online version.

l1-norm term and a second order difference operator for the *l2*-norm term. Different values of the parameter λ were explored for three different pairs of weights $(\mu_1, \mu_2) = \{(0.9, 0.1);\ (0.5, 0.5);\ (0.1, 0.9)\}$. Solutions with the lowest logGCV in each case are shown in the bottom row of Figure 3 of the supplemental material (online). Since values of logGCV and Corcondia (not shown) are almost the same in the three cases, in Figure 4c we present the solution with $\mu_1 = \mu_2 = 0.5$ as best, based only on the relative distance to the real loading. The slowest computed solution took around 2.5 minutes to converge.

3.4. Combining smoothness and sparsity

Finally, we explored the three types of sparse constraints on the temporal signature of the real data. Additionally, smoothness was required for the spectral loading in order to test the ability of the proposed algorithm to simultaneously impose different types of constraints to different loadings. Figure 5 shows three PARAFAC decompositions corresponding to the use of the Lasso, the Fusion Lasso and the Enet penalizations on the temporal loading, and Ridge (with a second order difference operator) on the spectral loading. The 'optimal' solutions were selected as those with minimum logGCV, also taking into account the Corcondia measure. All decomposition converged in less than 4 minutes.

The discussed properties of each penalty used can be easily distinguished. In the first case (Figure 5a, top), the signatures are sparser, since more coefficients are set to zero. In the second case (Figure 5b, top), there are some flat periods, and in the third (Figure 5c, top), the groups of coefficients with nonzero values

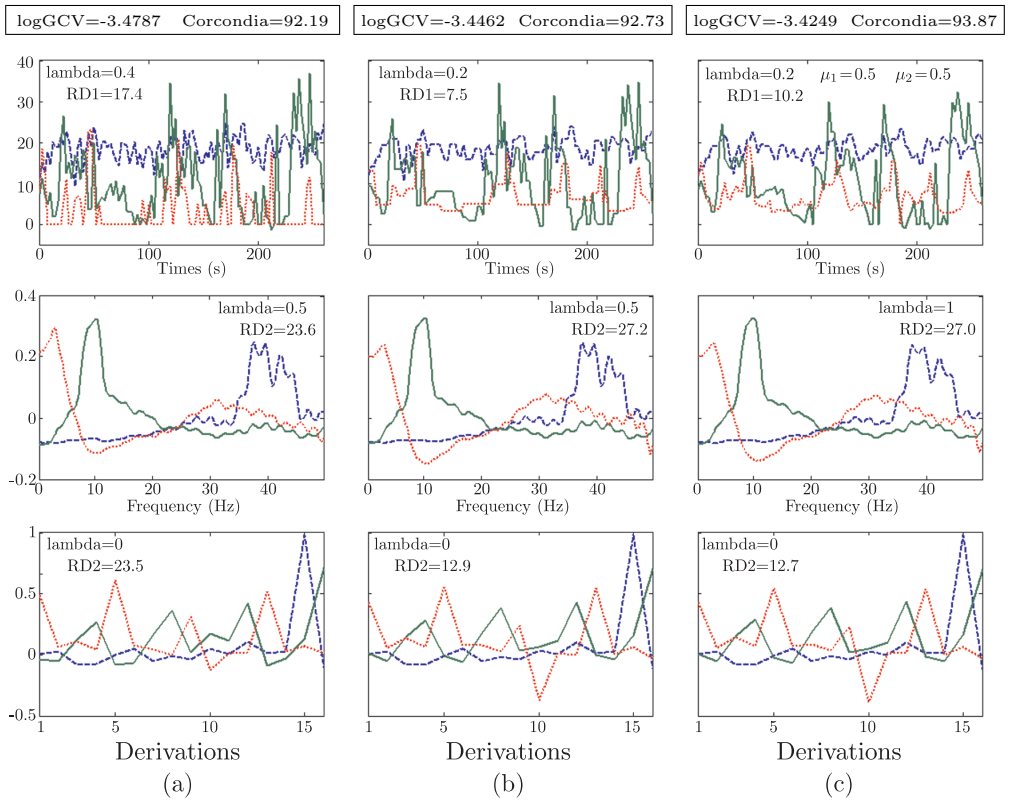

Figure 5. Estimated temporal (top), spectral (middle), and spatial (bottom) signatures for the three atoms using constraints for temporal and spectral loadings in real data. a) Lasso penalization (sparsity) on the temporal loading; b) Fusion Lasso penalization (sparsity on the first differences); c) Enet penalization (combination of sparsity and smoothness) with first and second order difference operators in the *l1*-norm and *l2*-norm terms, respectively. In all cases the smoothness constraint (using the second order difference operator) was required on the spectral loading. Corresponding optimal values for lambda, logGCV, Corcondia, and relative distances to unconstrained loadings (in percent) are shown. Solid line represents the Alpha atom, dotted line represents the Theta atom, and dashed line the Gamma atom. Figure in color in the online version.

show a smoother behavior. In all cases, the theta atom seems to be the one most reactive to the imposed constraint, and the gamma atom the one least reactive. Spectral loadings are almost the same as presented in Figure 2 with a corresponding smoothing parameter, and the spatial loadings closely resemble the unconstrained one shown in Figure 1c. This can be considered as evidence of a small influence of penalization in one loading on the remaining loadings.

Moreover, it can be of help in reducing the time needed for performing this kind of analysis, since one can first explore the optimal values for the weighting parameters in separate constrained analysis for the loadings, and then look for the optimal set of parameters for the conjoint analysis in a small neighborhood.

4. Conclusions

In this work we have proposed a methodology for imposing constraints on loadings in a PARAFAC decomposition. The combination of a multiple penalized linear regression algorithm and the alternating least squares philosophy has given rise to what we have called the Alternating Penalized Least Squares algorithm.

Although the idea of constraining the loading matrices in a PARAFAC regression is not new, to our knowledge, this is the first time that such a general algorithm is proposed, allowing the use (together with the usual constraints of orthogonality, nonnegativity, and others) of a wide range of unexplored penalties and combinations of penalties. This is particularly important in neuroscience, when the complex and noisy nature of the data makes the use of prior information unavoidable.

In our exploration, PARAFAC via APLS was useful for imposing smoothness on the spectral loading of the time-varying spectrum of real spontaneous EEG recording. It was equally successful in estimating different kinds of temporal evolutions that are common in neuroscience experimental designs. They range from very sparse signatures with only a few nonzero 'appearances' in time, to other group behavior such as box-like and piece-wise smooth functions. We found that the true simulated loadings are better recovered with the use of appropriate constraints than with the unconstrained solution. On the other hand, though the degree of constraint can be tuned by hand, we found that the use of GCV and the Corcondia measure can help in selecting an optimal solution. Constraining different loadings simultaneously did not affect the optimal values of the weighting parameters. This can reduce time of computation if they are selected in faster, separated analysis.

The proposed approach inherits some of the virtues and drawbacks of unconstrained PARAFAC. Among the former, the most attractive is the uniqueness of solution under very mild conditions. In this sense, the use of constraints can even help in those cases in which the noise level of the data restricts the convergence. Among the latter, we can mention the strong dependency on initial estimates, as well as the appearance of highly correlated atoms known as degeneracy. Again the use of constraints, when needed, can be helpful in avoiding degeneracy, and initial estimates for loadings can be obtained from the unconstrained solution. We conjecture that APLS provides more robust solutions than the ordinary ALS, although a more thorough study on this issue should be carried out in the future.

On the other hand, the use of a backfitting procedure can make the overall algorithm slower, although the efficient implementation through LQA and the use of Lemma 1 (Section 2) compensate for this effect. The computational time of the algorithm proposed depends on the chosen weighting parameters, usually being lower when optimum values are used. In our analysis, the slowest case converged in no more than 10 minutes, although the average computational time was around 2-3 minutes for actual data, and about a minute for synthetic data. Some approaches developed for speeding up the ALS algorithm in PARAFAC, such as Candelinc (Carroll, Pruzansky and Kruskal (1980)) and the use of QR-decompositions, could also be implemented in the context of the proposed algorithm.

Several issues remain unexplored and will be the subject of future work. First, the extension of the algorithm to use different penalization for each atom might allow for the extraction of, e.g., temporal evolutions with different properties for different rhythms in the same decomposition. Second, the use of statistical techniques such as bootstrapping for assessing the significance of findings is needed. Third, other approaches, such as the use of the Variational Bayesian framework, can be of help for selecting the optimal penalized decompositions. Finally, it should be mentioned that the APLS algorithm can also be applied in the context of other multidimensional models, such as Tucker, Parafac2, and multi-way Partial Least Squares (Bro (1998)).

Acknowledgement

Authors gratefully acknowledge Mayrim Vega-Hernández, Agustín Lage-Castellanos and Lester Melie-García for their useful comments and for detailed revision of the manuscript.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **19**, 716-723.

Andersson, C. A. and Bro, R. (2000). The N-way Toolbox for MATLAB. *Chemometrics Intell. Lab. Syst.* **52**, 1-4.

Ansley, C. F. and Kohn, R. (1994). Convergence of the backfitting algorithm for additive models. *J. Austral. Math. Soc. Ser. A* **57**, 316-329.

Beckmann, C. F. and Smith, S. M. (2005). Tensorial extensions of independent component analysis for multisubject FMRI analysis. *NeuroImage* **25**, 294-311.

Bro, R. (1998). Multi-way Analysis in the Food Industry: Models, Algorithms and Applications. Ph.D. Thesis, University of Amsterdam and Royal Veterinary and Agricultural University, Denmark.

Carroll, J. D., Pruzansky, S. and Kruskal, J. B. (1980). Candelinc: A general approach to multidimensional analysis of many-ways arrays with linear constraints on parameters. *Psychometrika* **45**, 3-24.

Fan, J. (1997). Comments on 'Wavelet in Statistics: A Review.' by Antoniadis. J. Italian Statist. Assoc. **6**, 131-138.

Fan, J. Q. and Li, R. Z. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 456, 1348-1360.

Golub, G., Heath, M. and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215-223.

Harshman, R. A. (1970). Foundations of the PARAFAC procedure: models and conditions for an 'explanatory' multi-modal factor analysis. *UCLA Work. Pap. Phon.* **16**, 1-84.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Hoerl, A. E. and Kennard, R. W. (2000). Ridge Regression: biased estimation for nonorthogonal problems. *Technometrics* **42**, 80-86.

Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. Ann. Statist. **33**, 1617-1642.

Kruskal, J. B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra Appl.* **18**, 95-138.

Land, S. and Friedman, J. (1996). Variable fusion: a new method of adaptive signal regression. Technical Report. Department of Statistics, Stanford University, Stanford.

Martínez-Montes, E., Valdés-Sosa, P. A., Miwakeichi, F., Goldman, R. and Cohen, M. (2004). Concurrent EEG/fMRI analysis by multi-way partial least squares. Neuroimage **22**, 1023-1034.

Miwakeichi, F., Martinez-Montes, E., Valdés-Sosa, P. A., Nishiyama, N., Mizuhara, H. and Yamaguchi, Y. (2004). Decomposing EEG data into space-time-frequency components using Parallel Factor Analysis. Neuroimage **22**, 1035-1045.

Morup, M. (2005). Analysis of brain data using multi-way array models on the EEG. Msc Thesis, Technical University of Denmark.

Morup, M., Hansen, L. K., Herrmann, C. S., Parnas, J. and Arnfred, S. M. (2006). Parallel Factor Analysis as an exploratory tool for wavelet transformed event-related EEG. *NeuroImage* **29**, 938-947.

Sánchez-Bornot, J. M., Martínez-Montes, E., Lage-Castellanos, E., Vega-Hernández, M. and Valdés-Sosa, P. A. (2008). Uncovering sparse brain effective connectivity: a voxel-based approach using penalized regression. *Statist. Sinica* **18**, 1501-1518.

Schwartz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Sidiropoulos, N. D. and Bro, R. (2000). On the uniqueness of multilinear decomposition of N-way arrays. *J. Chemometr.* **14**, 229-239.

Stegeman, A. and Sidiropoulos, N. D. (2007). On Kruskal's uniqueness condition for the Candecomp/Parafac decomposition. *Linear Algebra Appl.* **420**, 540-552.

Tibshirani, R. (1996). Regression shrinkage and variable selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *J. Roy. Statist. Soc. Ser. B* **67**, 91-108.

Timmerman, M. E. and Kiers, H. A. L. (2002). Three-way component analysis with smoothness constraints. *Comput. Statist. Data Anal.* **40**, 447-470.

Tomasi, G. and Bro, R. (2006). A comparison of algorithms for fitting the PARAFAC model. *Comput. Statist. Data Anal.* **50**, 1700-1734.

Valdés-Sosa, P. A., Sánchez-Bornot, J. M., Vega-Hernández, M., Melie-García, L., Lage-Castellanos A. and Canales-Rodríguez, E. (2006). Granger Causality on Spatial Manifolds: applications to Neuroimaging. Chapter 18 in *Handbook of Time Series Analysis: Recent Theoretical Developments and Application*. (Edited by Björn Schelter, Matthias Winterhalder and Jens Timmer). Wiley-VCH, Weinheim.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67**, 301-320.

Neurostatistics Department, Cuban Neuroscience Center, Havana, Cuba.

E-mail: eduardo@cneuro.edu.cu

Neurostatistics Department, Cuban Neuroscience Center, Havana, Cuba.

E-mail: bornot@cneuro.edu.cu

Neurostatistics Department, Cuban Neuroscience Center, Havana, Cuba.

E-mail: peter@cneuro.edu.cu

(Received April 2007; accepted March 2008)